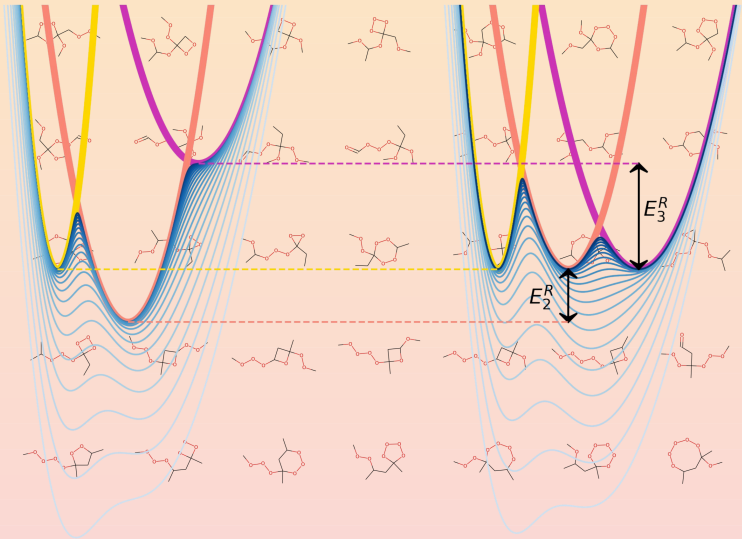
Diss. ETH No. 28572

Development and Application of Novel Force-Field Design and Free-Energy Calculation Approaches



Salomé R. Rieder-Walthard

2022

DISS. ETH NO. 28572

Development and Application of Novel Force-Field Design and Free-Energy Calculation Approaches

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

Salomé Ronja Rieder-Walthard

MSc. ETH in Computational Science and Engineering

born on 07.04.1993
citizen of Switzerland

accepted on the recommendation of

Prof. Dr. Philippe H. Hünenberger, examiner

Prof. Dr. Sereina Riniker, co-examiner
Prof. Dr. Daan P. Geerke, co-examiner

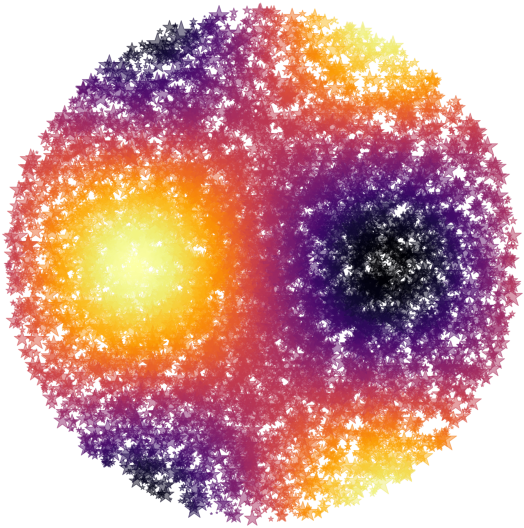
2022

“L’amor che move il sole e l’altre stelle”

“The love that moves the sun and the other stars”

Paradiso, XXXIII, 145

Dante Alighieri



Für

Sara, Hansruedi, Elin

& Alexander

Acknowledgements

This thesis was not created in a vacuum; many people accompanied and supported me on this journey, for which I am immensely grateful.

First of all, I want to thank Philippe Hünenberger. Ever since I first attended your lecture on *Computer Simulations in Chemistry, Biology and Physics*, I have been inspired by your genuine fascination with science and your passion for teaching and sharing your vast knowledge. After completing both my Bachelor's and my Master's thesis under your supervision, I enjoyed our collaboration so much I applied for a PhD position in your group. Thank you for the opportunity to conduct my PhD in the CSMS group and for your trust and support through the past years.

I am deeply grateful to Sereina Riniker for the co-supervision of my PhD. I admire your drive for science, your proficient and concise writing style, and your creativity. Thank you for your patience and support during the past years.

I also want to thank Daan Geerke for the nice conversations in Ausserberg and for agreeing to co-examine my defense, and Gunnar Jeschke for chairing my defense.

A big thank you to all the former and current members of the CSMS, CCG, and IGC groups for the supportive and friendly atmosphere, the enriching conversations, and the many (online and offline) game evenings. I am especially grateful to my fellow CSMS members Alžbeta Kubincová, David Hahn, Marina P. Oliveira, and Sadra Kashef Ol Gheta, as well as the "RE-EDS family", Benjamin Ries, Candide Champion, and Emilia P. Barros, for the fruitful collaborations.

One of my favorite tasks during my employment at ETH was the teaching. Thanks to the students I had the privilege of supervising, Eriks Kletnieks and Friedrich Ginnold, and to the students of the *Informatik I*

and *Statistische Physik und Computer Simulationen* exercise classes I got to teach, for their trust, patience (especially during the online classes), and interest.

I would like to acknowledge my *Gymnasium* maths teacher, Lukas Fischer. Thank you for teaching me the basics and sparking my interest in mathematics and science.

Thank you to my fellow CSE students, Giuseppe Accaputo, Pascal Iselin, Julia Pfund, and Tobias Wicky for making the many days of studying in the ETZ computer room during the winter and summer exam preparations more bearable, and to Thomas Häner for the many joint lunch breaks during the times we both worked at Hönggerberg.

A heartfelt thank you to Rhiannon Zarotiadis for the regular lunches, for listening to my sorrows and joys, and for your friendship.

I am genuinely grateful to my friends and family for their love and support that always kept me going. In particular, thank you Olivia for the many happy memories and for entrusting me with Gian's godparenthood, Stefanie for your long-lasting friendship, Martina and Fabian for accompanying me as my godparents, Pascal for the many wonderful excursions with Elin and Alex (and for being my favorite *Brändi Dog* partner), and a sincere thank you to my grandparents for their encouragement.

Finally, I am indebted to you, Mami, Papi, Elin, and Alex. Thank you for standing by me through the highs and lows, for loving and supporting me, and for believing in me.

> *"But I believe in love, you know; love is a uniquely portable magic. I don't think it's in the stars, but I do believe that blood calls to blood and mind calls to mind and heart to heart."*
>
> Jake Epping in 11/22/63
> Stephen King[1]

Contents

Summary

Molecular dynamics (MD) simulation has nowadays become a standard tool complementary to experiment for investigating molecular systems relevant in chemistry, biology, pharmacology, and material sciences. As a result, the need for high-throughput calculations and robust automation has increased. Broadly speaking, there are two main tasks to automate: (*i*) the preparation, *i.e.*, force-field parameterization and topology generation; and (*ii*) the running and post-processing of MD simulations, *i.e.*, simulation and analysis pipelines. In the present thesis, several contributions to the development of the combinatorial force-field parameterization (CombiFF) scheme and to the development and application of the replica-exchange enveloping distribution sampling (RE-EDS) free-energy method are presented. Chapter 1 provides a brief outline of MD simulation, classical force fields, and free-energy calculations.

CombiFF is a fragment-based force-field parameterization scheme for condensed-phase MD simulations. The scheme relies on a predefined fragment library and achieves a high observable-to-parameter ratio by considering large families of molecules simultaneously during the optimization. To this end, relevant molecules have to be generated and the corresponding molecular topologies have to be prepared for MD simulations. The two tasks are achieved by the C++ programs *enu* (Chapter 2) and *tbl* (Chapter 3). The *enu* program systematically enumerates the constitutional and spatial isomers of a given molecular formula. The *tbl* program automatically creates molecular topologies by assembling molecular fragments. The complementary C++ program *cnv* is also outlined (Chapter 4). The *cnv* program enables the conversion and canonicalization of different molecular identifiers in accordance with the conventions of CombiFF. The codes of *enu*, *tbl*, and *cnv* can be compiled with *cmake*

and are freely available at `https://github.com/csms-ethz/CombiFF`.

RE-EDS is a pathway-independent multistate free-energy method. The program *RestraintMaker* (Chapter 5) is used to assign (locally) optimal distance restraints for relative hydration free-energy calculations with RE-EDS. The obtained free energies are compared to experimental and calculated reference data, and the influence of the distance restraints on the conformational sampling is assessed. The program *amber2gromos* (Chapter 6) is developed to enable the conversion of AMBER topologies to GROMOS-compatible topologies. The use of converted topologies in the GROMOS MD simulation package is validated with single-molecule simulations and RE-EDS relative hydration free-energy calculations. Further, the performance of RE-EDS is assessed for systems with many end-states. The use of a shifted reaction-field correction with an atom-based cutoff for the electrostatic interactions is tested for RE-EDS calculations *via* relative solvation free-energy calculations in GROMOS (Chapter 7). A proof-of-concept implementation of RE-EDS in the OpenMM MD engine is also presented and validated.

Finally, the thesis is concluded by a brief overview of possible future developments (Chapter 8).

Zusammenfassung

Die Simulation der Molekulardynamik (MD) ist heutzutage ein Standardwerkzeug, welches Experimente bei der Untersuchung molekularer Systeme in der Chemie, Biologie, Pharmakologie und den Materialwissenschaften ergänzt. Infolgedessen ist der Bedarf an Berechnungen mit hohem Durchsatz sowie an robuster Automatisierung gestiegen. Grundsätzlich sind zwei wesentliche Aspekte zu automatisieren: (*i*) die Vorbereitung, *d.h.* die Parametrisierung von Kraftfeldern und Erzeugung von Topologien und (*ii*) die Ausführung und Nachbearbeitung von Simulationen, *d.h.* die Simulations- und Analyse-Pipelines. In der vorliegenden Arbeit werden mehrere Beiträge zu der Entwicklung des kombinatorischen Kraftfeld-Parametrisierungsverfahren (CombiFF) und zur Entwicklung und Anwendung der *replica-exchange enveloping distribution sampling* (RE-EDS) Methode zur Berechnung der freien Energie vorgestellt. Kapitel 1 gibt einen kurzen Überblick über MD-Simulationen, klassische Kraftfelder und Freie-Energie-Berechnungen.

CombiFF ist ein auf molekularen Fragmenten basierendes Kraftfeld-Parametrisierungsverfahren für MD-Simulationen in kondensierter Phase. Das Verfahren verwendet eine vordefinierte Fragmentenbibliothek und erreicht ein hohes Verhältnis von Beobachtungswerten zu Parametern, indem es während der Optimierung grosse Molekülfamilien berücksichtigt. Zu diesen Zweck müssen relevante Moleküle erzeugt und die entsprechenden molekularen Topologien für MD-Simulationen vorbereitet werden. Diese beiden Aufgaben werden durch die C++-Programme *enu* (Kapitel 2) und *tbl* (Kapitel 3) erfüllt. Das Programm *enu* zählt auf systematische Weise die Konstitutions- und Konfigurationsisomere einer bestimmten Summenformel auf. Das Programm *tbl* erstellt automatisch molekulare Topologien, indem es molekulare Fragmente zusammensetzt. Das ergänzende C++-

Programm *cnv* wird ebenfalls beschrieben (Kapitel 4). Das Programm *cnv* ermöglicht die Umformung und Kanonisierung verschiedener molekularer Kennzeichnungen gemäss den Konventionen von CombiFF. Die Quelltexte von *enu*, *tbl* und *cnv* können mit *cmake* kompiliert werden und sind unter `https://github.com/csms-ethz/CombiFF` frei zugänglich.

RE-EDS ist eine pfadunabhängige Methode zur Berechnung freier Energien mehrerer End-Zustände. Das Programm *RestraintMaker* (Kapitel 5) wird verwendet, um (lokal) optimale Abstandsbeschränkungen für relative Freie-Hydratationsenergie-Berechnungen mit RE-EDS zuzuweisen. Die erhaltenen freien Energien werden mit experimentellen und berechneten Referenzwerten verglichen, und der Einfluss der Abstandsbeschränkungen auf das Konformationssampling wird untersucht. Das Programm *amber2gromos* (Kapitel 6) dient der Umformung von AMBER Topologien zu mit GROMOS kompatiblen Topologien. Die Verwendung solcher konvertierter Topologien wird mit Ein-Molekül-Simulationen sowie RE-EDS-Berechnungen der relativen freien Hydratationsenergie im MD-Simulationsprogramm GROMOS validiert. Ausserdem wird die Leistungsfähigkeit von RE-EDS für Systeme mit vielen End-Zuständen bewertet. Die Verwendung einer verschobenen Reaktionsfeld-Korrektur mit einer atombasierten Abschaltung der elektrostatischen Wechselwirkungen wird für RE-EDS Berechnungen getestet mittels Berechnungen der relativen freien Solvatationsenergie in GROMOS (Kapitel 7). Eine Proof-of-Concept-Implementierung von RE-EDS im OpenMM MD-Simulationsprogramm wird ebenfalls vorgestellt und überprüft.

Abschliessend wird die Arbeit mit einem kurzen Überblick über mögliche zukünftige Entwicklungen abgerundet (Kapitel 8).

Publications

Articles in peer-reviewed journals:

1. Ries, B.[†]; Rieder, S. R.[†]; Rhiner, C.; Hünenberger, P. H.; Riniker, S. RestraintMaker: A Graph-Based Approach to Select Distance Restraints in Free-Energy Calculations With Dual Topology. *J. Comput.-Aided Mol. Des.* **2022**, *36*, 175–192.
2. Rieder, S. R.; Ries, B.; Schaller, K.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Replica-Exchange Enveloping Distribution Sampling Using Generalized AMBER Force-Field Topologies: Application to Relative Hydration Free-Energy Calculations for Large Sets of Molecules, *J. Chem. Inf. Model.* **2022**, *62*, 3043–3056.
3. Rieder, S. R.; Ries, B.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Replica-Exchange Enveloping Distribution Sampling: Calculation of Relative Free Energies in GROMOS, *CHIMIA*. **2022**, *76*, 327–330.
4. Rieder, S. R.; Ries, B.; Kubincová, A.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Leveraging the Sampling Efficiency of RE-EDS in OpenMM Using a Shifted Reaction-Field With an Atom-Based Cutoff, *J. Chem. Phys.* **2022**, *157*, 104117.
5. Rieder, S. R.; Oliveira, M. P.; Riniker, S.; Hünenberger, P. H. Development of an Open-Source Software for Isomer Enumeration, *J. Cheminform.* **2022**, submitted.

[†] These authors contributed equally.

Related publications:

1. Oliveira, M. P.; Andrey, M.; Rieder, S. R.; Kern, L.; Hahn, D. F.; Riniker, S.; Horta, B. A. C.; Hünenberger, P. H. Systematic Optimization of a Fragment-Based Force Field Against Experimental Pure-Liquid Properties Considering Large Compound Families: Application to Saturated Haloalkanes, *J. Chem. Theory Comput.* **2020**, *16*, 7525–7555.

2. Kashefolgheta, S.; Oliveira, M. P.; Rieder, S. R.; Horta, B. A. C.; Acree Jr, W. E.; Hünenberger, P. H. Evaluating Classical Force Fields Against Experimental Cross-Solvation Free Energies, *J. Chem. Theory Comput.* **2020**, *16*, 7556–7580.

3. Ries, B.; Normak, K.; Weiss, R. G.; Rieder, S. R.; Barros, E. P.; Champion, C.; König, G.; Riniker, S. Relative Free-Energy Calculations for Scaffold Hopping-Type Transformations With an Automated RE-EDS Sampling Procedure, *J. Comput. Aided Mol. Des.* **2022**, *36*, 117–130.

4. Oliveira, M. P.; Gonçalves, Y. M. H.; Kashefolgheta, S.; Rieder, S. R.; Horta, B. A. C.; Hünenberger, P. H. Comparison of the United- and All-Atom Representation of (Halo)alkanes Based on Two Condensed-Phase Force Fields Optimized Against the Same Experimental Data Set, *J. Chem. Theory Comput.* **2022**, online (DOI: 10.1021/acs.jctc.2c00524).

1 Introduction

"I believe that scientific knowledge has fractal properties; that no matter how much we learn, whatever is left, however small it may seem, is just as infinitely complex as the whole was to start with. That, I think, is the secret of the Universe."

Isaac Asimov[2]

1.1 MOLECULAR DYNAMICS SIMULATION

The discipline of *molecular dynamics* (MD) simulation[3–7] concerns itself with the emulation of molecular motions under the action of intra- and intermolecular forces.[8] The underlying principles are based on the classical equations of motion of Sir I. Newton, and the connection to thermodynamics is provided by the statistical mechanics of L. Boltzmann and J. C. Maxwell. Following the first Monte Carlo sampling reported in 1953 by Metropolis *et al.*, the first reported MD simulation dates back to B. J. Alder and T. E. Wainwright in 1957.[3] In 1964, the first Lennard-Jones fluid, consisting of 864 Argon particles, was simulated at a fixed temperature and pressure by A. Rahman.[9] Shortly thereafter, L. Verlet reported the simulation of 864 Argon particles as a Lennard-Jones fluid for various values of the system temperature and density.[10] In 1971, A. Rahman and H. Stillinger presented their work on simulating 216 rigid

water molecules using MD simulation.[11] The first reported simulation of a protein, namely bovine pancreatic trypsin inhibitor (BPTI), goes back to the late 1970s.[12–14] In the decades since then, MD simulation has become a well-established tool, complementary to experiments, to investigate molecular systems *in silico*.[15–19] Nowadays there exist many different MD simulation packages such as AMBER,[20,21] CHARMM,[22,23] GROMACS,[24,25] GROMOS,[26,27] or OpenMM.[28,29]

A *molecular model* is defined by four basic choices: (*i*) the resolution of the model, *i.e.*, which degrees of freedom are explicitly considered as dynamical variables in the simulation; (*ii*) the functional forms and parameters describing the interactions between the considered particles, *i.e.*, the force field; (*iii*) how configurations are generated; and (*iv*) the boundary conditions, such as the system temperature or pressure.[6,18,30–32]

In classical mechanics, as formulated in a Cartesian coordinate system, a particle is characterized by its position vector $\mathbf{r}_i = (x_i, y_i, z_i)$ and its momentum vector

$$\mathbf{p}_i = m_i \mathbf{v}_i \,, \tag{1.1}$$

where m_i is the particle mass and $\mathbf{v}_i$ is its velocity vector. In classical MD simulation, the particles of the system correspond to atoms, united-atoms, or small groups of atoms, depending on the level of resolution.[30] In addition to the coordinates, momenta, and masses, the particles are also assigned a charge and van der Waals parameters.[33] For a system containing N particles, the associated phase space consists of the two $3N$-dimensional vectors[34]

$$\mathbf{r}^N = \{\mathbf{r}_i \,|\, i = 0, \ldots, N-1\} \tag{1.2}$$

$$\mathbf{p}^N = \{\mathbf{p}_i \,|\, i = 0, \ldots, N-1\} \,. \tag{1.3}$$

The energy of the system is then expressed as[34]

$$\mathcal{H}(\mathbf{r}^N, \mathbf{p}^N) = \mathcal{K}(\mathbf{p}^N) + \mathcal{V}(\mathbf{r}^N)\,, \tag{1.4}$$

where $\mathcal{H}$ is the *Hamiltonian function*, $\mathcal{K}$ is the *kinetic energy* of the system (dependent on the momenta of the particles), and $\mathcal{V}$ is the *potential energy* of the system (dependent on the coordinates of the particles). The kinetic energy is given by

$$\mathcal{K}(\mathbf{p}^N) = \sum_{i=0}^{N-1} \frac{\mathbf{p}_i^2}{2m_i} \overset{(1.1)}{=} \sum_{i=0}^{N-1} \frac{1}{2} m_i \mathbf{v}_i^2\,. \tag{1.5}$$

The functional form of the potential energy, on the other hand, depends on the specific system under investigation (see Sec. 1.2).

Newton's equations of motion describe the relationship between the acceleration $\dot{\mathbf{v}}_i$ of a particle i (where a dot over a symbol indicates a time-derivative) and the force $\mathbf{f}_i$ acting on it in a Cartesian coordinate system as

$$\mathbf{f}_i(t) = m_i \dot{\mathbf{v}}_i(t) \overset{(1.1)}{=} \dot{\mathbf{p}}_i(t)\,. \tag{1.6}$$

In Hamiltonian mechanics, which represents a generalization of Newtonian mechanics to arbitrary coordinate systems, a particle i is propagated *via* the two equations

$$\dot{\mathbf{r}}_i(t) = \frac{\partial \mathcal{H}(\mathbf{r}_i, \mathbf{p}_i)}{\partial \mathbf{p}_i} \tag{1.7}$$

$$\dot{\mathbf{p}}_i(t) = -\frac{\partial \mathcal{H}(\mathbf{r}_i, \mathbf{p}_i)}{\partial \mathbf{r}_i}\,. \tag{1.8}$$

Recalling that the Hamiltonian is the sum of the kinetic and potential energies, where the kinetic energy only depends on the momentum of the particle and the potential energy only depends on the coordinates of the particle, and taking into account Eq. (1.6), Eqs. (1.7) and (1.8) can be

rewritten as

$$\dot{\mathbf{r}}_i(t) = \frac{\partial \mathcal{K}(\mathbf{p}_i)}{\partial \mathbf{p}_i} \overset{(1.5)}{=} \frac{\mathbf{p}_i}{m_i} \overset{(1.1)}{=} \mathbf{v}_i \tag{1.9}$$

$$\dot{\mathbf{p}}_i(t) = -\frac{\partial \mathcal{V}(\mathbf{r}_i)}{\partial \mathbf{r}_i} \overset{(1.6)}{=} \mathbf{f}_i(t)\,. \tag{1.10}$$

In plain words, the coordinates of the particle are propagated in time according to the particle velocity, and the momentum (and therefore, the velocity) of the particle changes according to the negative derivative of the potential energy with respect to the coordinates (*i.e.*, the force).

Eqs. (1.9) and (1.10) can be integrated numerically using, *e.g.*, the leap-frog scheme[15]

$$\mathbf{r}_i(t_n + \Delta t) = \mathbf{r}_i(t_n) + \mathbf{v}_i(t_n + \frac{\Delta t}{2})\Delta t + \mathcal{O}(\Delta t^3) \tag{1.11}$$

$$\mathbf{v}_i(t_n + \frac{\Delta t}{2}) = \mathbf{v}_i(t_n - \frac{\Delta t}{2}) + \frac{\mathbf{f}_i\left(\mathbf{r}_i(t_n)\right)}{m_i}\Delta t + \mathcal{O}(\Delta t^3) \tag{1.12}$$

to simulate molecular dynamics.

An extension of MD is *stochastic dynamics* (SD), which corresponds to applying the Langevin equation of motion[35,36]

$$m_i \dot{\mathbf{v}}_i(t) = \bar{\mathbf{f}}_i(t) + \mathbf{f}_i^{\mathrm{fr}}(t) + \mathbf{f}_i^{\mathrm{st}}(t) \tag{1.13}$$

instead of the Newtonian one. Here, $\bar{\mathbf{f}}_i$ is the mean force, $\mathbf{f}_i^{\mathrm{st}}$ is the stochastic force, and $\mathbf{f}_i^{\mathrm{fr}} = -m_i\gamma_i\mathbf{v}_i$ is the frictional force, where γ_i is the atomic friction coefficient of particle i. Eq. (1.13) can be integrated numerically using, *e.g.*, the leap-frog SD scheme,[37] the BAOAB integrator,[38,39] or the "middle" scheme.[40]

A third option to calculate the properties of molecular systems is to rely on *Monte Carlo* (MC) methods, such as Metropolis MC.[41,42] In each step of a MC simulation, one or several particles in the system experience a random displacement. The new configuration is accepted

with a probability of

$$\min\left(1, e^{-\beta(V^{\text{new}}-V^{\text{old}})}\right), \tag{1.14}$$

where V^{new} is the potential energy of the system in the new configuration, V^{old} is the potential energy of the system in the old configuration, and $\beta = 1/(k_B\ T)$, where k_B is the Boltzmann constant and T the reference temperature of the system.

Boundary conditions (BC) enforce global restrictions on the system during the simulation and can be classified as *hard* BCs (*i.e.*, satisfied exactly, affecting all configurations), or as *soft* BCs (*i.e.*, satisfied on average).[32] There exist three types of BCs: (*i*) *spatial* BCs, typically *vacuum*, *fixed*, or *periodic*; (*ii*) *geometric* BCs, *i.e.*, constraints or restraints on the coordinates of the particles; and (*iii*) *thermodynamic* BCs, such as constant volume or pressure, or constant energy or temperature.[32] By default, MD simulations conserve the number of particles, the volume, and the energy of the system, sampling a *microcanonical* (NVE) ensemble. By enforcing (on average) constant temperature or constant pressure, a *canonical* (NVT) or an *isothermal-isobaric* (NPT) ensemble, respectively, can be sampled instead. Simulating at constant temperature or pressure is useful to reproduce experimental conditions. Constant temperature can be achieved by employing a *thermostat*, while constant pressure is obtained through the use of a *barostat*.[43–49] Note that SD and MC simulations keep the temperature constant by default.

1.2 CLASSICAL FORCE FIELDS

The quality of an MD, SD, or MC simulation depends critically on the quality of the underlying force field, *i.e.*, the functional form and the parameters describing the potential energy of the system.[33,50,51] Usually, force fields for classical MD simulations rely on the following functional

form for the potential energy[52,53]

$$V^{\mathrm{pot}}(\mathbf{r}) = \underbrace{V^{\mathrm{bond}}(\mathbf{r}) + V^{\mathrm{angle}}(\mathbf{r}) + V^{\mathrm{torsional}}(\mathbf{r}) + V^{\mathrm{improper}}(\mathbf{r})}_{\text{bonded}} + \underbrace{V^{\mathrm{ele}}(\mathbf{r}) + V^{\mathrm{vdW}}(\mathbf{r})}_{\text{nonbonded}} . \tag{1.15}$$

Here, the potential-energy contribution of the covalent interactions is represented as the sum of the bond, bond-angle, torsional-dihedral, and improper-dihedral terms. The potential-energy contribution of the nonbonded interactions is represented as the sum of the electrostatic (Coulomb) and van der Waals (vdW) energy (Figure 1.1).

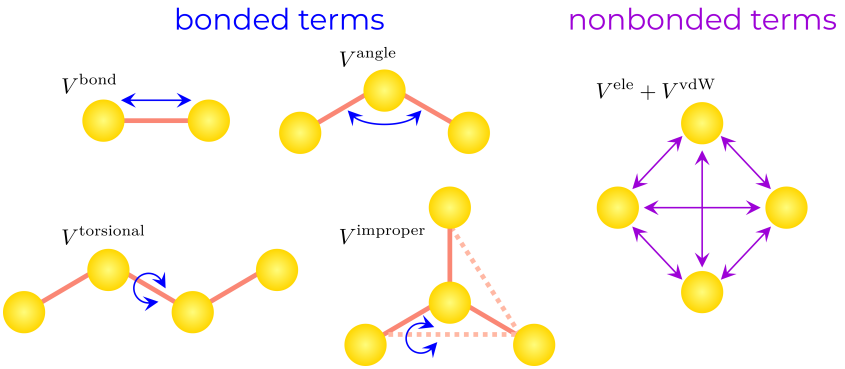


Figure 1.1: Force-field terms. In classical MD simulations, the interactions between the particles of a system are calculated as the sum of the depicted covalent and nonbonded potential-energy terms.

Popular force fields include the AMBER force fields,[54–56] the CHARMM force fields,[22,57–61] the GROMOS (compatible) force fields,[62–74] the OpenFF force fields,[75,76] and the OPLS force fields.[77–82]

1.2.1 Functional Forms

An overview of typical functional forms for the different terms of Eq. (1.15) is provided below. The different functions are illustrated graphically in Figure 1.2.

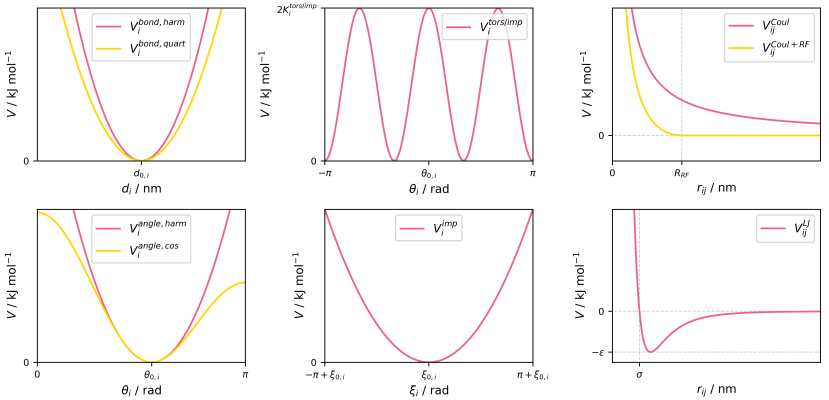


Figure 1.2: Schematic illustration of the functional forms of the different V^{pot} terms. The corresponding equations are provided in Eqs. (1.16) - (1.25). (Top left): Harmonic (pink) and quartic (yellow) bond-stretching. (Bottom left): Harmonic (pink) and cosine-harmonic (yellow) bond-angle bending. (Top middle): Proper dihedral-angle torsions with $m = 3$. (Bottom middle): Improper dihedral-angle bending. (Top right): Coulomb potential (pink) and Coulomb potential with a reaction-field correction (yellow), considering charges of like signs. (Bottom right): Lennard-Jones potential.

Covalent Interactions

Bond-stretching between two atoms is typically modeled *via* a harmonic or quartic potential[33,34]

$$V_i^{\mathrm{bond,harm}}(d_i) = \frac{1}{2} K_i^{b,\mathrm{harm}} \left(d_i - d_{0,i}\right)^2 \tag{1.16}$$

$$V_i^{\mathrm{bond,quart}}(d_i) = \frac{1}{4} K_i^{b,\mathrm{quart}} \left(d_i^2 - d_{0,i}^2\right)^2 , \tag{1.17}$$

where d_i is the distance between the two bonded atoms, $K_i^{b,\mathrm{harm}}$ and $K_i^{b,\mathrm{quart}}$ are the harmonic and quartic force constants, respectively, and

$d_{0,i}$ is the reference distance between the two bonded atoms. The quartic force constant is related to the harmonic one as $K_i^{b,\mathrm{quart}} = 0.5 K_i^{b,\mathrm{harm}} / d_{0,i}^2$. Close to the equilibrium distance, the two functional forms become very similar. However, the quartic form avoids the evaluation of a square root during the calculation of the force, making it computationally more efficient. Since the maximum time step permitting a stable integration of the equations of motion is bounded by the frequency of the fastest motions in the systems (*i.e.*, usually bond vibrations), bond distances are often constrained during MD simulations to permit the use of a larger time step.[83] Most constraint algorithms are based on minimizing the constraint forces using Lagrange multipliers. Possible constraint algorithms include SHAKE,[84] SETTLE,[85] LINCS,[86] M-SHAKE,[83] CCMA,[87] or SHAPE.[88]

Bond-angle bending between three atoms is usually modeled as a harmonic or cosine-harmonic potential[33,34]

$$V_i^{\mathrm{angle,harm}}(\theta_i) = \frac{1}{2} K_i^{a,\mathrm{harm}} \left(\theta_i - \theta_{0,i}\right)^2 \tag{1.18}$$

$$V_i^{\mathrm{angle,cos}}(\theta_i) = \frac{1}{2} K_i^{a,\mathrm{cos}} \left(\cos\theta_i - \cos\theta_{0,i}\right)^2 , \tag{1.19}$$

where θ_i ($0 \leq \theta_i \leq \pi$) is the angle formed by the three atoms, $K_i^{a,\mathrm{harm}}$ and $K_i^{a,\mathrm{cos}}$ are the harmonic and cosine-harmonic force constants, respectively, and $\theta_{0,i}$ is the reference bond angle between the three atoms. The cosine-harmonic form avoids the calculation of an arccosine during the force calculation, making it computationally more efficient. However, the drawback of this form is that it is not well suited for linear molecules, as the slope of the potential-energy function (*i.e.*, the force) becomes very small for $\theta_i \to 0$ and $\theta_i \to \pi$. While employing bond-angle constraints would permit a further increase of the time step, they introduce considerable artifacts in the dynamics of flexible molecules.[89,90] For completely rigid molecules, on the other hand, bond-angle constraints are common practice, *e.g.*, for solvent molecules.[83]

Proper and improper dihedral-angle changes between four atoms are

typically represented as[33]

$$V_i^{\text{tors/imp}}(\theta_i) = K_i^{\text{tors/imp}} \left[1 + \cos(m\theta_i - \theta_{0,i})\right] , \tag{1.20}$$

where θ_i ($-\pi \leq \theta_i \leq \pi$) is the torsion angle between the four atoms, $K_i^{\text{tors/imp}}$ is the torsion force constant, and m is the multiplicity. In the CHARMM and GROMOS force fields, Eq. (1.20) is typically only used to model proper dihedral-angle torsions, and there is a distinct functional form for improper dihedral-angle bending[33]

$$V_i^{\text{imp}}(\xi_i) = \frac{1}{2} K_i^{\text{imp}} \left(\xi_i - \xi_{0,i}\right)^2 , \tag{1.21}$$

where ξ_i ($-\pi + \xi_{0,i} \leq \xi_i \leq \pi + \xi_{0,i}$) is the improper torsional angle formed by the four atoms, K_i^{imp} is the improper dihedral angle force constant, and $\xi_{0,i}$ is the reference angle.

Nonbonded Interactions

The covalent interactions typically only involve a small subset of pairs (*i.e.*, bonds), triplets (*i.e.*, bond angles), or quadruplets (*i.e.*, dihedrals) of atoms. Pairwise nonbonded interactions, on the other hand, have in principle to be calculated for all pairs of atoms, *i.e.*, for a system with N atoms there are $\mathcal{O}(N^2)$ pairwise nonbonded interactions. For this reason, the calculation of the nonbonded interactions is typically the most time-consuming task of an MD simulation. For a typical biomolecular simulation, the number of particles is around 10^4 to 10^6, and accessible simulation time scales are on the order of nano- to milliseconds.[91] Accounting for all pairwise nonbonded interactions would make such simulations computationally too expensive. To mitigate this, a cutoff is usually applied to the vdW and electrostatic interactions.

The vdW interactions are typically modeled *via* the Lennard-Jones

(LJ) potential[34]

$$V_{ij}^{\mathrm{LJ}} = \frac{C_{12,ij}}{r_{ij}^{12}} - \frac{C_{6,ij}}{r_{ij}^{6}} = 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{1.22}$$

where $C_{12,ij}$ is the repulsion coefficient and $C_{6,ij}$ is the dispersion coefficient for atoms i and j, r_{ij} is the (minimum-image) distance between atoms i and j, ε_{ij} is the depth of the potential well, and σ_{ij} is the distance at which V_{ij}^{LJ} becomes zero. As the magnitude of the LJ potential becomes very small for long-range interactions, a straight truncation is typically used, *i.e.*, the interactions of atom pairs beyond the cutoff distance are neglected entirely. Due to the missing attractive interactions beyond the cutoff, this can lead to an underestimation of, *e.g.*, the simulated densities and vaporization enthalpies.[92] To mitigate this, the cutoff has to be sufficiently large. Alternatively, the interactions beyond the cutoff can be approximated using a long-range component, *e.g.*, an analytical tail correction,[5] LJ isotropic periodic sum approaches,[93–97] or a lattice sum formulation.[98–106]

The electrostatic interactions are represented by a Coulomb potential[33]

$$V_{ij}^{\mathrm{Coul}} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_{cs}}\frac{1}{r_{ij}}\,, \tag{1.23}$$

where q_i and q_j are the charges of atoms i and j, ε_0 is the permittivity of vacuum, and ε_{cs} is the relative permittivity of the medium (usually set to one, as all charges are treated explicitly). For the electrostatic interactions, a straight truncation leads to serious artifacts in the simulated properties.[107–111] Instead of using a straight truncation, the long-range interactions beyond the cutoff can be approximated by employing a mean-field scheme, or a lattice-sum scheme that approximates the environment beyond the cutoff as a periodic medium. A possible mean-field approach

is the reaction-field (RF) scheme[34,112–114]

$$V_{ij}^{\text{Coul+RF}} = \frac{q_i q_j}{4\pi\varepsilon_0\varepsilon_{\text{cs}}}\left[\frac{1}{r_{ij}} - \frac{C_{\text{RF}} r_{ij}^2}{2R_{\text{RF}}^3} - \frac{1-0.5C_{\text{RF}}}{R_{\text{RF}}}\right], \tag{1.24}$$

with[34,115]

$$C_{\text{RF}} = \frac{(2\varepsilon_{\text{cs}} - 2\varepsilon_{\text{RF}})(1+\kappa_{\text{RF}}R_{\text{RF}}) - \varepsilon_{\text{RF}}(\kappa_{\text{RF}}\ R_{\text{RF}})^2}{(\varepsilon_{\text{cs}} + 2\varepsilon_{\text{RF}})(1+\kappa_{\text{RF}}R_{\text{RF}}) + \varepsilon_{\text{RF}}(\kappa_{\text{RF}}\ R_{\text{RF}})^2}, \tag{1.25}$$

where ε_{RF} is the RF permittivity, and κ_{RF} the inverse Debye screening length (usually set to zero).[34] Lattice-sum schemes to approximate the long-range electrostatic interactions include Ewald summation,[116] particle-particle particle-mesh (P3M),[117] or particle-mesh Ewald (PME).[99,118]

1.2.2 Force-Field Parameterization

Broadly speaking, there exist three main strategies to parameterize condensed-phase force fields:[73] fragment-based,[20,22,62,119,120] hybrid,[121,122] and QM-derived.[123–132] Various tools exist to automate the laborious task of topology generation for small molecules, such as Antechamber,[55,133,134] the Automated Topology Builder (ATB),[68,135] General Automated Atomic Model Parameterization (GAAMP),[136,137] LigParGen,[138] Open Force Field (SMIRNOFF and OpenFF),[75,76] ParamChem,[59–61] PRODRG,[139] R.E.D.,[140] or SwissParam.[141]

CombiFF

Recently, the fragment-based *CombiFF* condensed-phase force-field parameterization scheme was published by Oliveira *et al.*[73,74] It relies on an assumption of *transferability*, *i.e.*, force-field parameters for (large) molecules are transferred from force-field parameters of (small) molecular fragments.

The CombiFF workflow consists of five main steps (Figure 1.3):[73] (*i*) definition of a molecule family; (*ii*) combinatorial enumeration of all

isomers of the family; (*iii*) search for experimental reference data in the in-house experimental database *DBS*; (*iv*) automated generation of molecular topologies based on a library of fragments; and (*v*) automated refinement of the force-field parameters by minimizing an objective function quantifying the discrepancy between simulated and experimental condensed-phase properties.

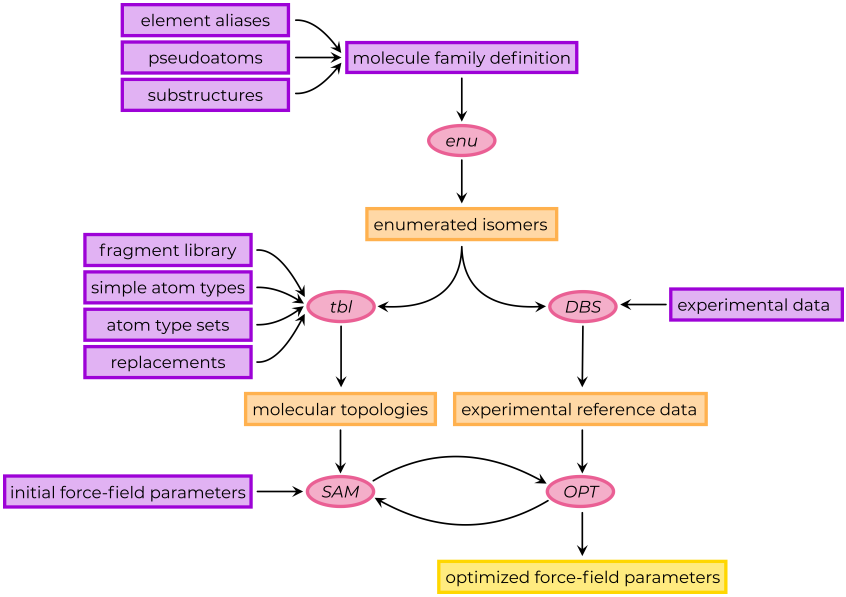


Figure 1.3: Schematic illustration of the CombiFF force-field parameterization workflow. For a specified family of molecules, all isomers are enumerated, experimental reference data is gathered, and molecular topologies are automatically assembled. In an iterative procedure, the initial force-field parameters are optimized based on the deviation of simulated properties from the experimental reference data. The program *enu*, as well as the input files (*i.e.*, element aliases, pseudoatoms, substructures, and molecule family definition) and generation of the output files (*i.e.*, enumerated isomers) is discussed in Chapter 2. The program *tbl*, as well as the input files (*i.e.*, fragment library, simple atom types, atom type sets, and replacements) and the generation of the output files (*i.e.*, molecular topologies) is presented in Chapter 3.

While, in general, the experimental data curation for a given target family still remains time-consuming, the refinement of a new set of parameters is automated and typically takes on the order of a few days. In

particular, this allows to investigate the influence of different choices for the simulation protocol (*e.g.*, united-atom or all-atom resolution, choice of combination rules, *etc.*) by evaluating simulated properties obtained from simulations with parameters optimized for the respective choice.[142]

Chapters 2 and 3 present the C++ programs *enu* and *tbl*, which have been written for step (*ii*) and step (*iv*) of the CombiFF scheme, respectively. Chapter 4 outlines the additional program *cnv*, which allows the conversion and canonicalization of different molecular identifiers and properties in accordance with the conventions of CombiFF.

1.3 FREE-ENERGY CALCULATIONS

Within the field of computational chemistry, free-energy calculations represent a particularly important (yet challenging) task. In particular, free-energy calculations are an increasingly well-established tool in computer-aided drug design,[143–151] where binding free-energy calculations enable the prediction of protein-ligand binding affinities.[146]

The free energy F of a system in the canonical (NVT) ensemble at thermodynamic equilibrium (expressed in Cartesian coordinates) is given by[152]

$$F = -\frac{1}{\beta} \ln Q = -\frac{1}{\beta} \ln \left[\frac{1}{h^{3N} N!} \int \int e^{-\beta H(\mathbf{p},\mathbf{r})} \mathrm{d}\mathbf{p} \mathrm{d}\mathbf{r} \right] , \tag{1.26}$$

where Q is the partition function of the system, h is Planck's constant, and H is the Hamiltonian of the system. Note that the factor $1/N!$ is only relevant for indistinguishable particles. Due to convergence issues, the calculation of absolute free energies is, in general, computationally expensive or even not achievable.[146,152] The calculation of free-energy differences, on the other hand, is accessible more easily. The free-energy difference between two states A and B (*e.g.*, two different molecules or

configurations) is given by

$$\Delta F_{BA} = F_B - F_A \overset{(1.26)}{=} -\frac{1}{\beta}\ln Q_B + \frac{1}{\beta}\ln Q_A = -\frac{1}{\beta}\ln\frac{Q_B}{Q_A}\,. \quad (1.27)$$

To obtain well-converged free-energy differences from a simulation, it is essential that all the relevant configurations of states A and B are visited during the simulation.[152]

The free-energy difference between two states can be estimated, for example, *via* free-energy perturbation (FEP) as[153]

$$\Delta F_{BA} = -\frac{1}{\beta}\ln\langle e^{-\beta(V_B - V_A)}\rangle_A\,. \quad (1.28)$$

FEP is exact in the limit of infinite samples. However, if there is a low phase-space overlap between the states, FEP becomes inefficient.[154,155] Other free-energy methods include thermodynamic integration (TI),[156] Bennett's acceptance ratio (BAR),[157] multistate BAR (MBAR),[158] λ-dynamics,[159–161] enveloping distribution sampling (EDS),[162,163] replica-exchange EDS (RE-EDS),[164–166] accelerated EDS (A-EDS),[167,168] and λ-EDS.[169]

1.3.1 Solvation Free-Energy Calculations

The solvation free energy represents the change in free-energy when a molecule is transferred from gas to the solvent.[170,171] Due to the considerably smaller system sizes, solvation free-energy calculations are much faster than binding free-energy calculations, making them a good test case to compare the quality of different free-energy methods and force fields.[170,172] In addition, there is ample experimental and calculated reference data available in databases such as FreeSolv,[155,173] the Minnesota solvation database,[174] or the ATB server.[68,135]

Solvation free-energy differences between end-states i and j obtained from simulation can be compared to experimental (or calculated) absolute

solvation free energies *via* the relative solvation free energy $\Delta\Delta G^{ji}_{\text{solv}}$ as[146]

$$\Delta\Delta G^{ji}_{\text{solv}} = \Delta G^{j}_{\text{solv}} - \Delta G^{i}_{\text{solv}} = \Delta G^{ji}_{\text{solvent}} - \Delta G^{ji}_{\text{vac}}\,, \tag{1.29}$$

where $\Delta G^{i}_{\text{solv}}$ and $\Delta G^{j}_{\text{solv}}$ are the solvation free energies of end-state i and j, respectively, $\Delta G^{ji}_{\text{solvent}}$ is the free-energy difference between end-states i and j in the solvent, and $\Delta G^{ji}_{\text{vac}}$ is the free-energy difference between end-states i and j in vacuum (Figure 1.4).

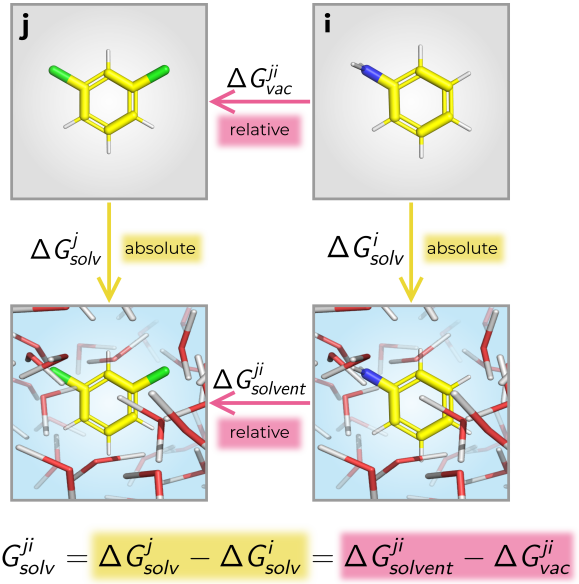


Figure 1.4: Schematic illustration of the thermodynamic cycle to calculate the relative solvation free energy $\Delta\Delta G^{ji}_{\text{solv}}$ for a pair i and j of molecules. $\Delta\Delta G^{ji}_{\text{solv}}$ can be equivalently calculated *via* the solvation free energy of molecule i ($\Delta G^{i}_{\text{solv}}$) and of molecule j ($\Delta G^{j}_{\text{solv}}$), or *via* the free-energy difference of molecules i and j in the solvent ($\Delta G^{ji}_{\text{solvent}}$) and the free-energy difference of molecules i and j in vacuum ($\Delta G^{ji}_{\text{vac}}$). Figure adapted from Rieder *et al.*[175] (see Chapter 6).

Chapters 5, 6, and 7 investigate various aspects of solvation free-energy calculations with the RE-EDS free-energy method.

2

Enu: Development of an Open-Source Software for Isomer Enumeration*

"One, two, three, four, five, six, seven, eight forks. One, two, three, four, five, six, seven, eight plates. One, two, three serving spoons. [...] I always knew how many of everything there were. Things were there and could be counted and accounted for, and so that's what I did."

Katherine Johnson [176]

The present chapter documents *enu*, a freely-downloadable, open-source and stand-alone program written in C++ for the enumeration of the constitutional isomers and stereoisomers of a molecular formula. The program relies on graph theory to enumerate all the constitutional isomers of a given formula on the basis of their canonical adjacency matrix. The stereoisomers of a given constitutional isomer are enumerated as well, on the basis of the automorphism group of the canonical adjacency matrix. The isomer list is then reported in the form of canonical SMILES strings within files in XML format. The specification of the molecule family of interest is very flexible and the code is optimized for computational

∗ This chapter is reproduced in part from Rieder, S. R.; Oliveira, M. P.; Riniker, S.; Hünenberger, P. H. Development of an Open-Source Software for Isomer Enumeration, *J. Cheminform.* **2022**, submitted.

efficiency. The algorithms and implementations underlying *enu* are described, and simple illustrative applications are presented. The *enu* code is freely available on GitHub at `https://github.com/csms-ethz/CombiFF`.

2.1 INTRODUCTION

Chemistry is the science of molecular transformations, *i.e.*, the recombination of sets of atoms in different molecules. Therefore, the concepts of molecular formula (atom content), structure (connectivity and stereochemistry), and geometry (conformation) are central to the field of chemistry.

Except for the smallest compounds, there generally exist many molecular structures compatible with a given formula. The corresponding molecules are referred to as *isomers*, and their number typically increases exponentially with the number of atoms in the molecule. Among these isomers, one may further distinguish between *constitutional isomers* and *stereoisomers*. Constitutional isomers differ exclusively in terms of the connectivity of the atoms, disregarding any spatial considerations. Given a specific constitutional isomer, the associated stereoisomers typically differ by the *chirality* or the *cis-trans isomery* of specific groups in the molecule. These spatial differences are ascribed to structure (topology) rather than geometry (conformation) because the interconversion between stereoisomers does not occur spontaneously under usual conditions. As a result, the individual stereoisomers can be isolated, and their physicochemical properties are generally distinct.

The determination of isomer sets has been of interest in the fields of chemistry, mathematics, and computer science for a long time.[177,178] Isomer counting is in itself already a very challenging mathematical problem in the field of graph theory, that has been addressed since nearly a century.[179] In view of the large isomer counts for all but the smallest

compounds, the explicit enumeration of the isomers of a given molecular formula was essentially impossible before the development of sufficiently powerful computers. In this context, one may mention the pioneering DENDRAL project, going back to the 1960s.[177,180] Since then, many algorithms have been developed for performing isomer enumeration in an efficient way.[181–188]

From a fundamental point of view, isomer counting and enumeration are important tools to improve our knowledge of chemical space,[187] and to analyze the effective coverage of chemical databases in terms of this space.[177] Isomer enumeration can also be used as a starting point for structure elucidation (by generating structures fulfilling certain restrictions obtained from spectroscopy) and virtual screening (by generating candidate structures).[177]

Recently, our group has introduced a new scheme called CombiFF for the design of classical force fields for molecular simulation,[73,74] in which isomer enumeration plays a central role. More specifically, CombiFF performs the automated calibration of force-field parameters against experimental condensed-phase data, considering entire classes of organic molecules constructed using a fragment library *via* combinatorial isomer enumeration. The main steps of the scheme are: (*i*) definition of a molecule family; (*ii*) enumeration of all isomers; (*iii*) query for experimental data; (*iv*) automatic construction of the molecular topologies by fragment assembly; and (*v*) iterative refinement of the force-field parameters considering the entire family.

The goal of the present chapter is to document the isomer enumerator of the CombiFF workflow, a C++[189] program called *enu*. Although the motivation for the development of *enu* was the CombiFF scheme, the program is an open-source and stand-alone software that can be used and further developed independently of CombiFF for any other purpose in cheminformatics.

The main features of the *enu* program are the following:

1. The constitutional isomers of a given molecular formula are enumerated on the basis of their adjacency matrix, given the constraint of fixed valences for the different atom types and the application of a canonicalization by lexicographical matrix ordering.

2. The stereoisomers of a given constitutional isomer are enumerated on the basis of the automorphism group of the canonical adjacency matrix.

3. The generated constitutional isomers and stereoisomers are reported in the form of canonical SMILES strings within files following an XML format.

4. The specification of the molecule family of interest is very flexible, including count ranges for the atoms in the molecular formula, selectors for specified substructures, and values of basic properties such as the number of cycles, unsaturations or multiple bonds.

5. The C++ code was optimized for computational efficiency, which is essential due to the combinatorial explosion of isomer counts with the number of atoms.

6. The code is freely available on GitHub at `https://github.com/csms-ethz/CombiFF`.

The algorithm used in *enu* for the enumeration of constitutional isomers is largely inspired from the PhD thesis of R. Grund at the University of Bayreuth in 1994,[185] which is also the approach underlying the structure generator MOLGEN.[177,186,190] The enumeration of stereoisomers based on the automorphism group of the adjacency matrix, on the other hand, was developed independently. The generation of canonical SMILES strings is based on the works of Weininger *et al.*[191,192] and Schneider *et al.*[193] The present chapter describes the algorithms and features of *enu* in terms of the six points above.

The first version of *enu* (capable of enumerating constitutional isomers only) was implemented as part of the author's Master thesis.[194] Since this first version, the code has been extensively refactored, optimized and extended. The current chapter focuses principally on the novel features of *enu*, namely the stereoisomer generation, the efficient implementation of the algorithms, and the flexibility of the input specifications. For completeness, more details on the other points (including the generation and canonicalization of constitutional isomers) are provided in Appendix Secs. 2.A - 2.C.

2.2 THEORY

This section consists of four parts. First, it provides an overview of the basic principles underlying (molecular) graph theory. Second, it describes the enumeration algorithm for constitutional isomers developed by Grund.[185] Third, it briefly explains the concept of canonical SMILES strings. Fourth, it describes the procedure developed here for the enumeration of stereoisomers.

2.2.1 Graph Theory

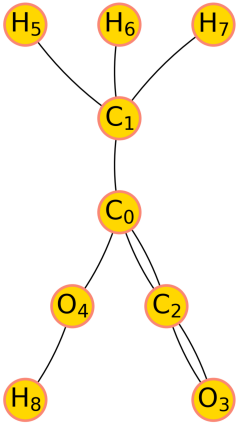
The isomer enumerator relies on graph theory. This discipline goes back to the first half of the 18^{th} century when the Swiss mathematician Leonhard Euler published his famous article on the *Problem of the Königsberg Bridges*.[195] Since then, graph theory has become increasingly relevant, with applications in fields such as the social sciences, economics, electrical, and industrial engineering, as well as all branches of the natural sciences, namely physics, chemistry, and biology.[196]

Molecular Graphs

A molecular graph is a connected labeled multigraph, *i.e.*, a graph in which there exists a path from each node to every other node, the nodes

are labeled, and there can be multiple edges between two nodes. The vertices represent atoms and the edges account for covalent bonds between the atoms.[197] The graph describes the topology of a molecule, but does not provide any information on its geometry.

A molecular graph can be described by the combination of a *label vector* $\boldsymbol{\alpha}$, a *valence vector* $\boldsymbol{\delta}$, a *partition vector* $\boldsymbol{\lambda}$, and a symmetric *adjacency matrix* $\boldsymbol{A} \in \mathbb{N}_0^{+\,N\times N}$.[185] An example is provided in Figure 2.1 and the terminology is explained in more detail in Appendix Sec. 2.B.1. The combination of the label vector (atom-type names) and partition vector (number of atoms of a given type) provide the molecular formula. The valence vector contains the fixed valences of the atom types listed in the label vector. A matrix element $A_{i,j}$ of $\boldsymbol{A}$ describes the order of the bond possibly connecting the atom at position i to the atom at position j in the atom vector (or is set to zero in the absence of a bond).



$$\boldsymbol{\alpha} = (\mathrm{C}, \mathrm{O}, \mathrm{H})$$
$$\boldsymbol{\delta} = (4, 2, 1)$$
$$\boldsymbol{\lambda} = (3, 2, 4)$$
$$\mathbf{a} = (\mathrm{C}, \mathrm{C}, \mathrm{C}, \mathrm{O}, \mathrm{O}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H})$$
$$\mathbf{d} = (4, 4, 4, 2, 2, 1, 1, 1, 1)$$
$$p_0 = \{0, 1, 2\},\ p_1 = \{3, 4\},\ p_2 = \{5, 6, 7, 8\}$$

$$\mathbf{A} = \begin{array}{c} \begin{matrix} \mathrm{C}_0 & \mathrm{C}_1 & \mathrm{C}_2 & \mathrm{O}_3 & \mathrm{O}_4 & \mathrm{H}_5 & \mathrm{H}_6 & \mathrm{H}_7 & \mathrm{H}_8 \end{matrix} \\ \begin{pmatrix} 0 & 1 & 2 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{matrix} \mathrm{C}_0 \\ \mathrm{C}_1 \\ \mathrm{C}_2 \\ \mathrm{O}_3 \\ \mathrm{O}_4 \\ \mathrm{H}_5 \\ \mathrm{H}_6 \\ \mathrm{H}_7 \\ \mathrm{H}_8 \end{matrix}$$

Figure 2.1: Illustrative example for a molecular graph. The graph corresponds to an isomer of the chemical formula $C_3O_2H_4$ with the label vector $\boldsymbol{\alpha}$, the valence vector $\boldsymbol{\delta}$, the partition vector $\boldsymbol{\lambda}$, the atom vector $\boldsymbol{a}$, the degree vector $\boldsymbol{d}$, the partitions p_0, p_1 and p_2, and the adjacency matrix $\boldsymbol{A}$.

For a given choice of $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\lambda}$ (*i.e.*, of a molecular formula and of atom-type valences), the specification of an adjacency matrix $\boldsymbol{A}$ (*i.e.*, of a covalent connectivity between the atoms) defines a unique labeled molecular graph. However, since the atoms of a common type in a molecule are physically indistinguishable, two labeled graphs that are directly related by a permutation in the indices of these atoms actually describe the same molecule (merely with a different atom numbering). In other words, for a given choice of $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\lambda}$, the same molecule can generally be represented by many different adjacency matrices $\boldsymbol{A}$. This observation is fundamentally important to the problem of isomer enumeration. It is known as (molecular) graph isomorphism, and discussed in more detail in Appendix Sec. 2.B.2.

Adjacency Matrix Canonicity In order to have a unique representation of a molecular topology in the form of a labeled multigraph, a *lexicographical ordering* can be used as canonicity criterion for the adjacency matrix. An adjacency matrix $\boldsymbol{A}$ is lexicographically larger than an adjacency matrix $\boldsymbol{A'}$ (noted $\boldsymbol{A} > \boldsymbol{A'}$) provided that[185]

$$\exists i_0, j_0 : \left(A_{i,j} = A'_{i,j} \ \forall (i,j) < (i_0, j_0)\right) \ \wedge \ \left(A_{i_0,j_0} > A'_{i_0,j_0}\right) , \tag{2.1}$$

with the definition[185]

$$(i,j) < (k,l) \ \Leftrightarrow \ (i < k) \ \vee \ (i = k \wedge j < l) . \tag{2.2}$$

In plain words, when the two matrices are read row-by-row from the top left to the bottom right, the first difference encountered determines the lexicographical ordering. The canonical adjacency matrix used to represent a molecule is then defined as the lexicographically largest among all possible adjacency matrices, which in turn defines a canonical labeling of the atoms in the molecular graph. Note that for a unique representation of molecules, the canonicity criterion for $\boldsymbol{A}$ must be accompanied by a canonicity criterion for the ordering of the atom types in the vector $\boldsymbol{\alpha}$.

More details on canonicity can be found in Appendix Sec. 2.B.3.

2.2.2 Enumeration of Constitutional Isomers

The following sections present the algorithm that is used in the *enu* isomer enumerator. It is based on the PhD thesis of R. Grund,[185] which proposes a solution to both the problem of finding all possible adjacency matrices as well as testing these adjacency matrices for canonicity.

Enumerating all the unique constitutional isomers of a given molecular formula amounts to finding all the canonical adjacency matrices associated with this formula. This could be achieved in a brute-force way by exhaustively enumerating all possible adjacency matrices compatible with the given molecular formula and the fixed valences of the different atom types, and filtering out those that are not canonical. The algorithm implemented in *enu* is based on this principle, but relies on an effective pruning mechanism that drastically limits the number of adjacency matrices to be generated and tested for canonicity, leading to far superior performance compared with a brute-force approach.

Orderly Enumeration

The first task to be performed is the systematic construction of possible adjacency matrices for a given choice of the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\lambda}$. The algorithm of Grund[185] proceeds by creating these matrices in lexicographically decreasing order (Figure 2.2). Note that this principle of orderly generation was proposed earlier by Read and Faradzev.[198–200] Since the adjacency matrix is symmetric and has only zeros along its diagonal, the algorithm only needs to find valid entries for the upper triangle of the matrix. Starting with an empty adjacency matrix, it proceeds through the matrix from the top left element to the bottom right one (with the line number as a primary index and, within each line, the column number as a secondary index), and fills it using two main subroutines. This filling order is particularly convenient, as it matches that in which the

elements are checked to determine if a matrix is lexicographically larger or smaller than another one. In the *forward step*, the current matrix position is incremented and the maximum possible entry for the new position is determined based on the specified atom valences and the bonds already listed in the matrix. If a compatible value is found, the matrix element at the current position is filled, and another forward step is called. Otherwise, the matrix is not amenable to completion and a *backward step* is performed. The current matrix position is decremented and it is checked whether the matrix element at the new position can be decreased by one. If this is possible, the algorithm continues with a forward step. Otherwise it continues with another backward step. The two routines are outlined in Appendix Sec. 2.B.4.

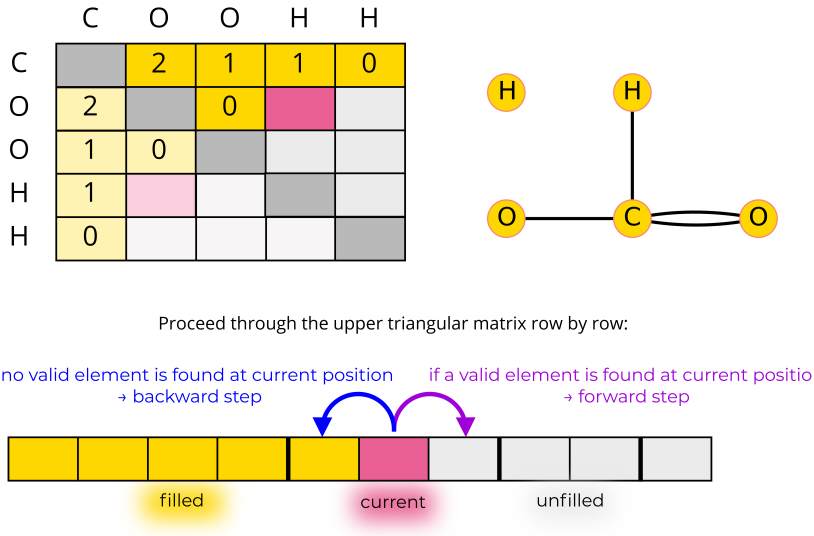


Figure 2.2: Illustrative example of the orderly enumeration algorithm. (Top): A step of the filling algorithm for $\boldsymbol{\alpha}$ =(C,O,H), $\boldsymbol{\delta} = (4, 2, 1)$, and $\boldsymbol{\lambda} = (1, 2, 2)$ with an intermediate adjacency matrix and the corresponding molecular graph. (Bottom): Schematic illustration describing how the algorithm proceeds through the adjacency matrix.

Connectivity Test

While the filling algorithm generates all possible adjacency matrices compatible with the valences of the different atoms in decreasing lexicographical order, it does not guarantee that these adjacency matrices describe a connected graph.[185] Therefore, a potential adjacency matrix has to be tested for connectivity to ensure that it is a viable isomer of the given molecular formula (rather than a collection of two or more molecules). The connectivity test implemented in *enu* is an adapted depth-first search[201] of the graph, as described in Appendix Sec. 2.B.5. The algorithm uses a last-in-first-out stack to go through the connected parts of the graph, and marks the vertices it encounters as visited. If all vertices have been visited once the stack is empty, the graph is connected.

Canonicity Test

The most important (and most difficult) part of the enumeration process is the testing of the generated adjacency matrices for canonicity, *i.e.*, assessing whether there is no isomorphic adjacency matrix that is lexicographically larger. The routines to perform this canonicity test are described in Appendix Sec. 2.B.6.

Special Treatment of Hydrogen Atoms

Hydrogen atoms typically represent the most numerous atom type in a molecule. Therefore, it is advantageous to rely on a special treatment for this atom type during the enumeration process.[185] To that end, following the suggestion of Grund,[185] the hydrogen atoms are associated to the heavy atom bearing them in a pre-processing step, thereby producing united-atoms with accordingly decreased valences. These are used in the filling algorithm. In this way, the hydrogen atoms are "implicit" during the orderly enumeration, resulting in a significant speed gain. This is described in more detail in Appendix Sec. 2.B.7.

2.2.3 Canonical SMILES

The *enu* program enumerates isomers based on lexicographically canonical adjacency matrices. However, for convenience, it reports the generated isomers as canonical *Simplified Molecular Input Line Entry System* (SMILES) strings.[202] The generation of a unique SMILES representation requires the specification of a canonical atom ordering.[192,193] Over the past decades, various algorithms have been developed to achieve such a canonicalization.[192,193,203–206] While the matrix canonicity criterion of *enu* already leads to a canonical ordering of the atoms, this order is not necessarily suitable to create elegant (*i.e.*, easily readable) SMILES strings. Thus, once a new canonical adjacency matrix is found, a different canonicalization algorithm is used as a post-processing step in *enu* to create the corresponding canonical SMILES strings. It is based on a combination of the schemes proposed by Weininger *et al.* in 1989[192] and by Schneider *et al.* in 2015.[193] The applied algorithm is explained in more detail in Appendix Sec. 2.C.

Note that during both the enumeration of the adjacency matrices and the generation of canonical SMILES strings, the hydrogen atoms are treated implicitly.

2.2.4 Enumeration of Stereoisomers

When going from a two- to a three-dimensional representation of a molecule, constitutionally identical molecules can have a different spatial arrangement of their atoms, leading to *stereoisomerism*.[197] In *enu*, the enumeration of the stereoisomers associated with a given constitutional isomer is performed as a postprocessing step to the constitutional isomer generation. Two kinds of stereoisomerism are considered: (*i*) chirality is considered for tetravalent atoms that have four singly-bonded neighbors; and (*ii*) cis/trans stereoisomerism is considered for double bonds connecting two tetravalent atoms that have two singly-bonded neighbors

(in addition to the doubly-bonded one) and that are not part of a cycle. Currently, handling of stereochemistry for possible centers of valence higher than four is not implemented. In addition, double bonds within cycles or cummulene systems are at present not considered in the search of cis/trans stereocenters.

In practice, the tetravalent atoms are typically carbon atoms. The neighbors (substituents) can differ either in constitution, *i.e.*, different atom types or connectivities, or in their spatial arrangements, *i.e.*, different stereo configurations. Stereocenters with neighbors differing in their constitutions are called *true stereocenters*, while stereocenters with neighbors differing only in their stereo configurations are called *para stereocenters*.[207] The considered types of stereocenters (*i.e.*, tetrahedral and cis/trans) as well as the distinction between true and para stereocenters is illustrated in Figure 2.3.

Here, stereoisomers will refer to the isomers of a given constitutional isomer corresponding to different spatial arrangements around tetrahedral centers and double bonds. The entire collection of all stereoisomers for all constitutional isomers of a given formula will be referred to as the spatial isomers of the molecule. The *enu* program thus enumerates all constitutional and spatial isomers of a chemical formula.

The procedure to enumerate stereoisomers consists of two steps. In a first step, all unique true stereoisomers are determined. In a second step, the para stereoisomers are generated. This division is necessary because the para stereocenters may be active or not depending on the stereo configuration of the other stereocenters in the molecule. The enumerator uses canonical SMILES strings to represent the enumerated stereoisomers. These strings describe the *local* stereo configuration of the stereocenters in the molecule, which implies that the specified configuration depends on the order in which the atoms appear in the string.[202]

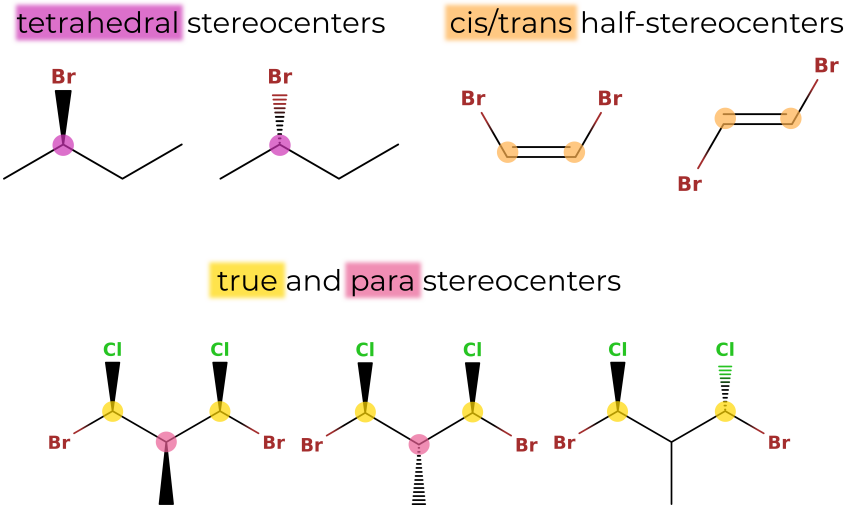


Figure 2.3: Illustration of types of stereocenters. Note that the hydrogen atoms are treated implicitly. (Top): Tetrahedral stereocenters (purple) are bonded to four different substituents, and cis/trans stereocenters (orange) consist of two cis/trans half-stereocenters which are bonded by a double bond and each connected to two different substituents by single bonds. (Bottom): Distinction between true (yellow) and para (pink) stereocenters. For a true stereocenter, the substituents differ in their constitution, whereas for a para stereocenter, the substituents only differ in their spatial arrangement.

Finding the True Stereoisomers of a Molecule

The problem of finding all true stereoisomers of a molecule involves two sequential tasks: (*i*) identifying the true stereocenters; and (*ii*) generating all unique true stereoisomers that arise from these stereocenters.

Recognizing True Stereocenters The procedure to detect true *tetrahedral* stereocenters is shown in Figure 2.4. The process consists of checking all atoms with four neighbors (or three neighbors and one implicitly connected hydrogen). If the four first-neighbor atoms are all different, a true tetrahedral stereocenter is directly detected. If at least two first neighbors are identical and have valence one, the atom cannot be a true stereocenter. If two or more first neighbors are identical but have a valence larger than one, the algorithm relies on the automorphism group of the canonical adjacency matrix. The automorphism group $Aut(\mathbf{A}) \subseteq S_{\boldsymbol{\lambda}}$

is generated as a by-product of the canonicity test in the enumeration algorithm (see Appendix Sec. 2.B.6). It contains the atom index permutations which leave the adjacency matrix $\boldsymbol{A}$ unchanged. If there is at least one permutation $\pi \in Aut(\mathbf{A})$ which leaves the potential stereocenter unchanged (*i.e.*, does not swap it with another atom) but swaps two of its first neighbors, the atom is not a true stereocenter. If such a permutation does not exist, the potential stereocenter is a true tetrahedral stereocenter.

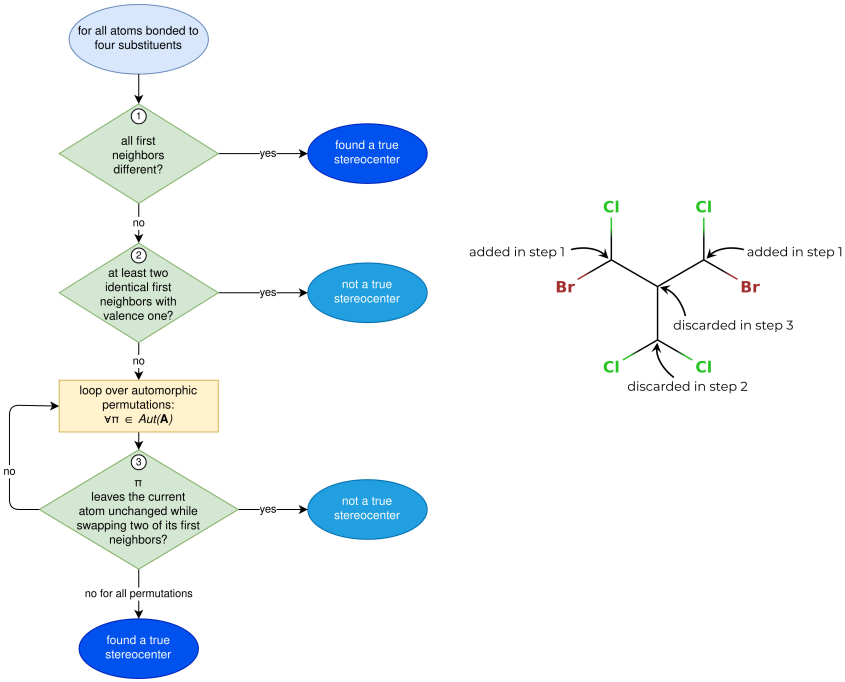


Figure 2.4: (Left): Procedure followed to find all true tetrahedral stereocenters of a molecule. It makes use of the automorphism group $Aut(\mathbf{A})$ of the canonical adjacency matrix which is created during the orderly enumeration process. (Right): Illustrative example how the procedure detects the true tetrahedral stereocenters of a molecule. Note that the hydrogen atoms are treated implicitly.

A cis/trans stereocenter is defined as a pair of atoms connected by a double bond. The term (cis/trans) *half-stereocenter* will be used here

for these two atoms. The procedure to identify true cis/trans half-stereocenters is analogous to the one for tetrahedral stereocenters. It is outlined in Figure 2.5. Here, one considers all the atoms with three neighbors (or two neighbors and one implicitly connected hydrogen atom) that are connected by exactly one double bond to another atom. If the two singly-bonded first-neighbor atoms are identical and have valence one, the considered atom is not a true half-stereocenter. If the two singly-bonded first neighbors are different or if they are identical but there is no permutation $\pi \in Aut(\mathbf{A})$ that leaves the considered atom identical while swapping the two singly-bonded first neighbors, the considered atom is a potential half-stereocenter. All potential half-stereocenters which are connected by a double bond to another potential half-stereocenter are true half-stereocenters.

Generating All Unique True Stereoisomers

Once the n_{tet} true tetrahedral stereocenters and the n_{ct} true cis/trans stereocenters of a molecule have been identified, it is straightforward to enumerate the SMILES strings corresponding to the associated $2^{n_{\text{tet}}+n_{\text{ct}}}$ true stereoisomers. A binary *configuration vector* of size $n_{\text{tet}} + n_{\text{ct}}$ is used for this purpose. For the tetrahedral centers, a `0` specifies a clockwise direction (encoded by @@ in the SMILES string), and a `1` specifies a counter-clockwise direction (encoded by @ in the SMILES string). For double bonds, a `0` specifies a trans configuration (encoded by '/' and '/', or '\' and '\' in the SMILES string) and a `1` specifies a cis configuration (encoded by '/' and '\', or '\' and '/' in the SMILES string). The vector is initially filled with zeros, and binary counting is then used to find all possible configurations of the true stereocenters. However, not all stereoisomers constructed using this approach are unique, as can be seen by considering the examples in Figure 2.6.

To make sure that only unique stereoisomers are reported, the automorphism group $Aut(\boldsymbol{A})$ is used to filter the results of the binary counting procedure. Here, the convention is used that the smallest equivalent

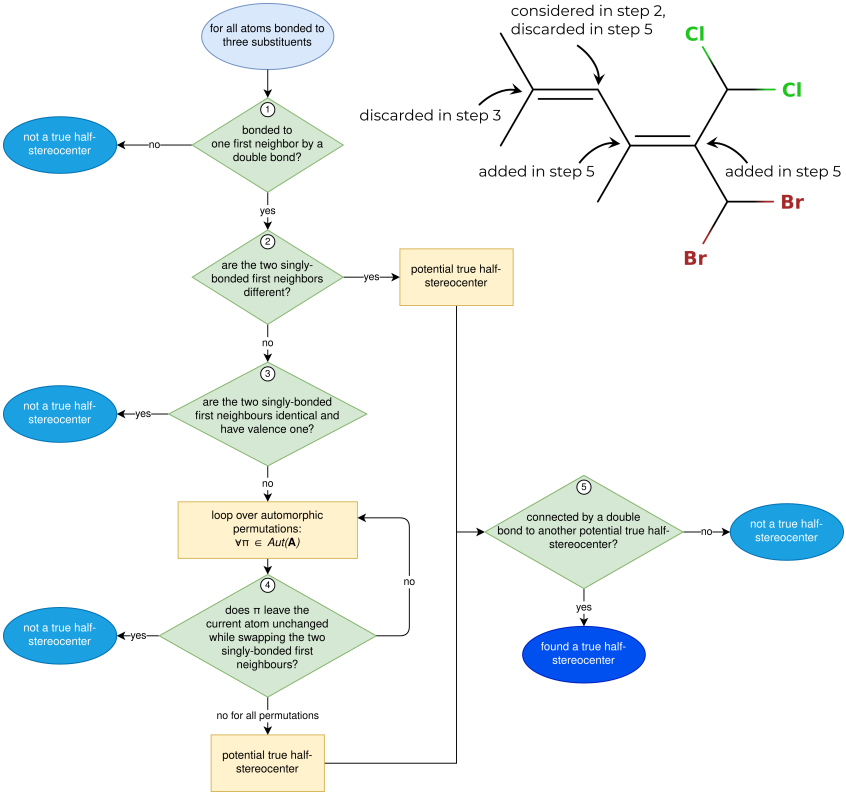


Figure 2.5: (Left): Procedure followed to find all true cis/trans half-stereocenters of a molecule. It makes use of the automorphism group $Aut(\mathbf{A})$ of the canonical adjacency matrix which is created during the orderly enumeration process. (Top right): Illustrative example how the procedure detects the true cis/trans stereocenters of a molecule. Note that the hydrogen atoms are treated implicitly.

stereochemical configuration vector is reported as the canonical one. For each configuration vector, it has to be determined whether the current stereoisomer is equivalent to one of the previously generated stereoisomers, *i.e.*, one with a lexicographically smaller configuration vector. For all permutations $\pi \in Aut(\boldsymbol{A})$, the true stereocenters are checked. By definition, two stereocenters can only be swapped by a permutation of the automorphism group if they are configurationally indistinguishable. If such a swap occurs in a given permutation, the corresponding encoding

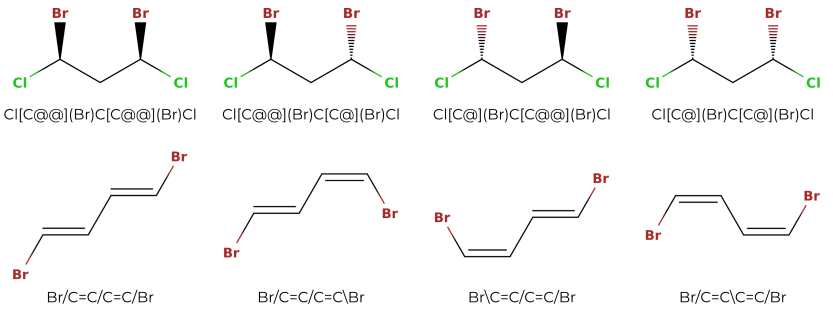


Figure 2.6: Non-uniqueness of the true stereoisomers generated by binary enumeration. Both of the molecules depicted, ClC(Br)CC(Br)Cl and BrC=CC=CBr, posses two true stereocenters. There are thus $2^2 = 4$ possible binary configuration vectors. The corresponding stereoisomers are shown for the two molecules. In both cases, however, only three stereoisomers are unique. (Top): For ClC(Br)CC(Br)Cl, the stereoisomer with configuration vector `[0,0]` (leftmost) is identical to that with configuration vector `[1,1]` (rightmost). (Bottom): For BrC=CC=CBr, the stereoisomer with configuration vector `[0,1]` (left of center) is identical to that with configuration vector `[1,0]` (right of center).

of `0` or `1` in the configuration vector has to be changed accordingly. If the resulting configuration vector is smaller than the original one, the stereoisomer has already been encountered, and is not counted again. When using a notation that specifies absolute stereo configurations, the step of adapting the encoding in the configuration vector is trivial, as the configurations of the two swapped stereocenters are simply swapped as well (Figure 2.7). However, since SMILES strings only specify the stereo configuration locally, the step of adapting the encoding in the configuration vector is not as trivial.

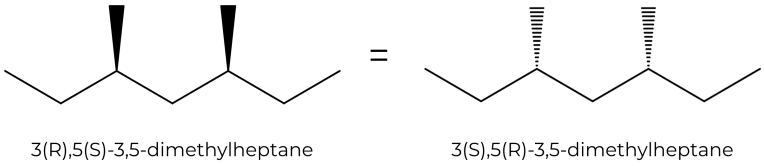


Figure 2.7: Advantage of absolute chirality notations. This example illustrates that when stereocenters with absolute stereo configuration are swapped, the stereo configuration is simply swapped as well.

The local stereo configuration depends on the order in which the first

neighbors of a stereocenter appear in the SMILES string, as can be seen in Figure 2.8. If a tetrahedral stereocenter is connected to four neighboring atoms (or three atoms and one implicit hydrogen), there are $4! = 24$ possible orders (or $3! = 6$ in the case of a center with an implicit hydrogen). When a permutation $\pi \in Aut(\boldsymbol{A})$ is applied to a stereoisomer, it is possible that a stereocenter is permuted with a different stereocenter and the configuration vector has to be adapted accordingly. For this, the order of the neighboring atoms of the original stereocenter in the SMILES string has to be compared to the corresponding order after applying the permutation. If the stereocenter a_i with neighbors $\boldsymbol{n}_i = [n_{i,1}, n_{i,2}, n_{i,3}, n_{i,4}]$ (order in the SMILES string) and configuration $c_i \in \{0, 1\}$ is permuted with the stereocenter a_j with neighbors $\boldsymbol{n}_j = [n_{j,1}, n_{j,2}, n_{j,3}, n_{j,4}]$ and configuration $c_j \in \{0, 1\}$, one determines for each of the neighboring atoms in $\boldsymbol{n}_i$ the neighboring atom in $\boldsymbol{n}_j$ it is permuted with. If the new ordering is an even permutation of the old ordering, the configuration c_i^{perm} of stereocenter a_i in the new configuration vector will be encoded with the old configuration of a_j, *i.e.*, $c_i^{perm} = c_j$. Conversely, if the new ordering is an odd permutation, the configuration c_i^{perm} of a_i will be encoded with the opposite of the old configuration of a_j, *i.e.*, $c_i^{perm} = \neg c_j$. Once the permuted configuration vector is generated, it can be decided whether the current stereoisomer is equivalent to a stereoisomer with a smaller configuration vector. If it is the case, the stereoisomer is not counted again. An example how non-unique stereoisomers are detected is shown in Figure 2.9.

Unlike a tetrahedral stereocenter, a cis/trans half-stereocenter is only connected to three neighboring atoms (or two atoms and one implicit hydrogen). The configuration of a cis/trans stereocenter is determined by the value of `0` or `1` in the stereochemical configuration vector, as well as by which of the two singly-bonded neighbors are considered for the directionality. Here, the convention is used that the directionality is always specified considering the two singly-bonded neighbors that are encountered first in the SMILES string. If two atoms $a_{i,1}$ and $a_{i,2}$ forming

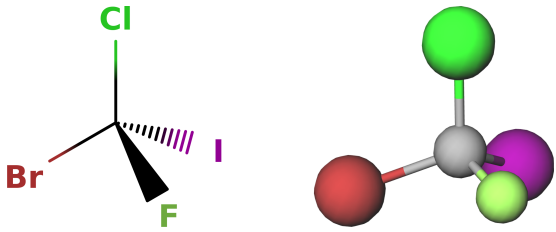


Figure 2.8: Influence of the order of the atoms in the SMILES strings on the notation for the local stereo configuration. The displayed molecule corresponds to, *e.g.*, the SMILES string Br[C@](Cl)(I)F. If the order of the first-neighbor atoms of the central carbon in the string is changed by an odd permutation (odd number of swaps), the stereo configuration notation of the carbon changes from '@' to '@@'. For example, if the iodine and the chlorine are swapped, this results in the string Br[C@@](I)(Cl)F. For an even permutation (even number of swaps), the stereo configuration notation stays the same. For example, if the bromine and the chlorine are swapped and then the chlorine and the iodine are swapped, this results in the string I[C@](Br)(Cl)F.

a cis/trans stereocenter s_i with configuration $c_i \in \{0, 1\}$ (where the singly-bonded neighbors considered for the directionality are the atoms $n_{i,1}$ and $n_{i,2}$) are swapped with two other atoms $a_{j,1}$ and $a_{j,2}$ forming a cis/trans stereocenter s_j with configuration c_j (where the singly-bonded neighbors considered for the directionality are the atoms $n_{j,1}$ and $n_{j,2}$), the new configuration of s_i is determined as follows. If $n_{i,1}$ is swapped with $n_{j,1}$ and $n_{i,2}$ is swapped with $n_{j,2}$, this corresponds to an even permutation, and the configuration of s_i in the new configuration vector will be equal to the configuration of s_j, *i.e.*, $c_i^{perm} = c_j$. If $n_{i,1}$ is not swapped with $n_{j,1}$, but $n_{i,2}$ is swapped with $n_{j,2}$ (or vice versa) this means that, due to the ordering of the atoms in the SMILES string, a different pair of singly-bonded first neighbors is considered for the directionality of s_i than for the directionality of s_j, and the configuration of s_i in the new vector becomes the opposite of the configuration of s_j, *i.e.*, $c_i^{perm} = \neg c_j$. If neither $n_{i,1}$ is swapped with $n_{j,1}$ nor $n_{i,2}$ with $n_{j,2}$, it means that the opposite pair of first neighbors is considered for the directionality of the configuration of s_i than it was for s_j, and the configuration of s_i in the new configuration vector is the same as the configuration of s_j in the old one, *i.e.*, $c_i^{perm} = c_j$.

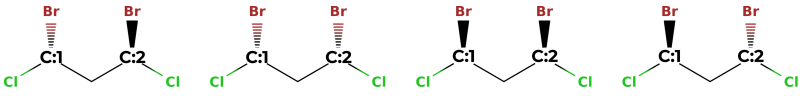


Figure 2.9: Procedure to detect duplicate stereoisomers. The depicted molecule has two true stereocenters a_1 (C:1) and a_2 (C:2). Considering the SMILES string Br[C:1](Cl)C[C:2](Br)Cl, there are four possible configuration vectors `[0,0]`, `[0,1]`, `[1,0]` and `[1,1]`, which correspond to the stereoisomers depicted (from left to right). In a SMILES string, an '@' indicates a counter-clockwise direction and an '@@' a clockwise one (when looking from the atom that comes before the stereocenter in the string towards the stereocenter and then looking at the other three neighbors in the order in which they appear in the string. If there is an implicit hydrogen attached to the stereocenter, it is treated as the first neighbor visited after the stereocenter, or if the stereocenter is the first atom in the SMILES string as the neighbor visited before the stereocenter[202]). Given the atom ordering retained in the SMILES string, the neighbors of a_1 are $\boldsymbol{n}_1$ =`[Br,H,Cl,C]`, and the neighbors of a_2 are $\boldsymbol{n}_2$ =`[C,H,Br,Cl]`. To go from the neighbor order of $\boldsymbol{n}_1$ to the one of $\boldsymbol{n}_2$, one has to swap `Br` and `C` and then `Br` and `Cl`. This corresponds to an even permutation. Thus, if the two stereocenters of Br[C:1@@](Cl)C[C:2@@](Br)Cl are swapped, this results in the string Br[C:2@@](Cl)C[C:1@@](Br)Cl, which is identical to the original one. The same is true for Br[C:1@](Cl)C[C:2@](Br)Cl. On the other hand, if the two stereocenters of Br[C:1@@](Cl)C[C:2@](Br)Cl are swapped, this results in the string Br[C:2@](Cl)C[C:1@@](Br)Cl, which means that the configuration vectors `[0,1]` and `[1,0]` correspond to the same stereoisomer. Thus, only the one with the smaller configuration vector `[0,1]`, *i.e.* Br[C:1@@](Cl)C[C:2@](Br)Cl, will be counted. Finally, the three unique stereoisomers are Br[C:1@@](Cl)C[C:2@@](Br)Cl, Br[C:1@@](Cl)C[C:2@](Br)Cl and Br[C:1@](Cl)C[C:2@](Br)Cl.

Recognizing Para Stereocenters

Once the true stereocenters have been found and all unique stereoisomers stemming from these centers have been enumerated, the next step is to complete the list of stereoisomers by adding the para stereoisomers. In the following, the term *potential para stereocenter* is used to denote a center that can possibly be a para stereocenter. Whether this possibility is realized depends on the actual configuration of the true stereocenters (and of the other potential para stereocenters) in the molecule.

An atom is a potential tetrahedral para stereocenter if it was omitted from the list of true stereocenters in the third test of the algorithm in Figure 2.4. This corresponds to the situation where at least one permutation $\pi \in Aut(\boldsymbol{A})$ leaves the center unchanged while swapping two of its first neighbors. This indicates that the substituents of the center starting with these first neighbors are *constitutionally* identical, but may potentially be *stereochemically* distinct if they encompass true or other

para stereocenters in specific configurations.

The same logic applies to potential cis/trans para half-stereocenters. If an atom was discarded in the fourth or fifth test of the algorithm in Figure 2.5 it might be a potential cis/trans para half-stereocenter. This is the case if the atom has two identical singly-bonded first neighbors which are swapped by at least one permutation $\pi \in Aut(\boldsymbol{A})$, indicating that the two neighboring substituents are *constitutionally* identical, but may be *stereochemically* distinct. The actual para cis/trans stereocenter is defined as a pair of half-stereocenters, so it still has to be checked whether a potential cis/trans para half-stereocenter is connected by a double bond to another potential para cis/trans half-stereocenter.

An additional criterion that can be used to reduce the number of potential para stereocenters to be tested is that a para stereocenter lies "in the middle" of at least one pair (or, possibly, multiple pairs) of configurationally symmetrical true stereocenters. Here, symmetrical means that there is at least one permutation $\pi \in Aut(\boldsymbol{A})$ which swaps the two true stereocenters. If two symmetrical true stereocenters are not part of a cycle, a simple shortest-path algorithm[208] can be used to determine which atom lies in the middle of the two. If an atom is part of one or more cycles, there is no simple shortest-path algorithm to check all potential paths between two true stereocenters, and this filter cannot be employed. This is illustrated in Figure 2.10. Further, note that within a cycle there can also exist para stereocenters that do not depend on true stereocenters, but only on other para stereocenters, *e.g.*, *cis*-1,4-dimethylcyclohexane and *trans*-1,4-dimethylcyclohexane. Such para stereocenters are currently not yet considered in *enu*.

To summarize, the potential para stereocenters that are retained consist of the ones that were discarded as true stereocenters and lie either in a cycle or in the middle of the shortest path between at least one pair of symmetrical true stereocenters. If there are no potential para stereocenters in a molecule, the list of stereoisomers is already complete after considering the true stereoisomers. Otherwise, the list is processed

anew to generate the para stereoisomers.

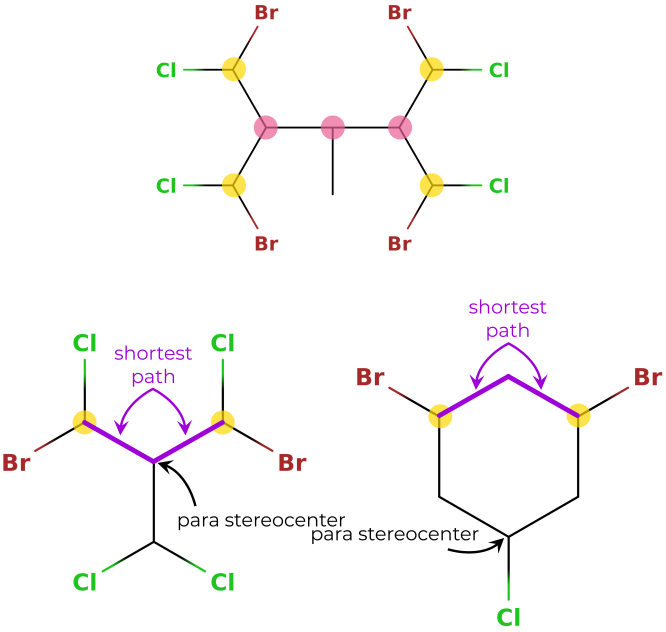


Figure 2.10: Illustration of para stereocenters. (Top): The true stereocenters are marked with a yellow dot and the potential para stereocenters are marked with a pink dot. The hydrogen atoms are not shown explicitly. The two outer potential para stereocenters lie in the middle between the two symmetrical true stereocenters on the left and right, respectively. The inner potential para stereocenter lies in the middle of the two outer potential para stereocenters, as well as of the four combinations of true stereocenters on the two opposite sides of the molecule. (Bottom): Illustration of the shortest path and ring criteria for detecting potential para stereocenters.

Generating All Unique Para Stereoisomers

The process to generate all unique para stereoisomers has to be performed for each true stereoisomer. Whether a potential para stereocenter actually *is* a stereocenter can only be determined once the stereo configuration in a true stereoisomer is specified.

The algorithm to determine all unique para stereoisomers of a molecule goes as follows. For each true stereoisomer, the automorphism group $Aut^{\text{true}}(\boldsymbol{A}) \subseteq Aut(\boldsymbol{A})$ is considered, which contains the subset of permutations in the automorphism group that only swap true stereocenters if they have the same absolute stereo configuration. Next, a para configuration vector is created and binary counting is used to find all possible para stereoisomers. For each of these, it first has to be determined which of the potential para stereocenters are actually stereocenters in the current configuration. Then it has to be determined whether this stereoisomer has already been found before, *i.e.*, with a smaller para configuration vector.

In order to determine which potential para stereocenters are actually stereocenters, the automorphism group $Aut^{\text{para}}(\boldsymbol{A}) \subseteq Aut^{\text{true}}(\boldsymbol{A})$ is considered, which contains the permutations of $Aut^{\text{true}}(\boldsymbol{A})$ that only swap para stereocenters if they have the same absolute stereo configuration in the current para stereoisomer (*i.e.*, the true stereoisomer with the para stereo configuration specified by the current para configuration vector). All potential para stereocenters for which there exists a permutation $\pi \in Aut^{\text{para}}(\boldsymbol{A})$ that leaves the current para stereocenter identical but swaps at least two of its immediate neighbors are not stereocenters in the current para stereoisomer. This is indicated by setting the corresponding entry in the para configuration vector to `-1` (technically making the vector ternary instead of binary). These steps are repeated as long as at least one para stereocenter was determined to be "not active" in the current configuration, until no new para stereocenter is discarded in an iteration.

In order to check whether para stereocenters have the same absolute configuration, the same logic can be used as for true stereocenters. Two symmetrical tetrahedral para stereocenters have the same absolute configuration if (*i*) the order of the first neighbors of the first stereocenter in the SMILES string is an even permutation of the order of the first neighbors of the second stereocenter in the string and they have the same encoding in the para configuration vector; or (*ii*) the order of the first neighbors is an odd permutation and they have the opposite encoding

in the para configuration vector. Similarly, two symmetrical cis/trans para stereocenters have the same absolute configuration if the two singly-bonded neighbors (and the connected substituents) encountered first in the SMILES string (*i*) are both the same for the two stereocenters and the stereocenters have the same encoding in the para configuration vector; (*ii*) are both not the same for the two stereocenters and the stereocenters have the same encoding in the para configuration vector; or (*iii*) if only one of the singly-bonded neighbors of the first stereocenter is the same as the singly-bonded neighbors of the second stereocenter and the stereocenters have the opposite encoding in the para configuration vector.

The process to eliminate equivalent but smaller para configurations is the same as for true stereocenters. For the lexicographical comparison, the values of `-1` in the para configuration vector are treated as `0`. Figure 2.11 shows the stereoisomers of a small molecule that has four true stereocenters and three potential para stereocenters.

Creating SMILES Strings for Stereoisomers

Once the list of stereoisomers of a constitutional isomer is available as a list of configuration vectors for the true and para stereocenters, this information can be included into the corresponding SMILES strings. For the tetrahedral stereocenters, the encoding is straightforward. In the SMILES string, the element symbol of the stereocenter is enclosed by rectangular brackets and either '@@' or '@' is added if the corresponding element in the configuration vector has the value `0` or `1`, respectively. Additionally, if the stereocenter is connected to an implicit hydrogen atom, a 'H' is added before the closing rectangular bracket.

The handling of cis/trans stereocenters is slightly more complicated. In the general case, one simply adds '/' before the half-stereocenter that is visited first in the SMILES string, and '/' or '\' before the first visited singly-bonded neighbor of the second half-stereocenter if the corresponding element in the configuration vector has the value `0` or `1`, respectively. However, the situation is more complicated if a half-stereocenter is also

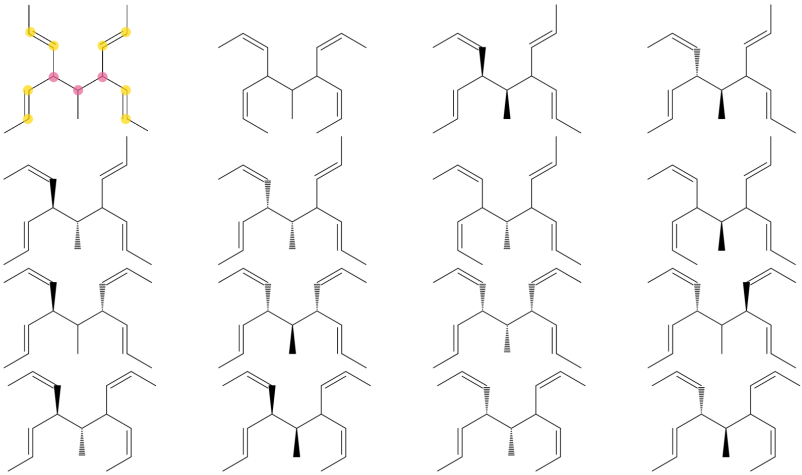


Figure 2.11: Illustration of para stereoisomers. This relatively small molecule contains four true and three potential para stereocenters, with a total of 16 different stereoisomers. In the molecule on the top left, the true (half-)stereocenters are marked in yellow and the potential para stereocenters are marked in pink. With a total of seven stereocenters there could in principle be $2^7 = 128$ stereoisomers. However, in many cases one or more of the potential para stereocenters is not a stereocenter, and multiple stereo configurations actually correspond to the same stereoisomer.

the singly-bonded neighbor of another half-stereocenter (Figure 2.12). A simple way to handle this situation is to go through the cis/trans stereocenters in the order in which the half-stereocenters are visited in the SMILES string, checking if the first half-stereocenter already contains an encoding, and then adapting the encoding of the first visited singly-bonded neighbor of the second half-stereocenter accordingly (*i.e.*, cis or trans).

In the XML output, the number of tetrahedral and cis/trans stereocenters is reported for each stereoisomer. Additionally, for stereoisomers with at least one tetrahedral stereocenter, the enantiomer of each stereoisomer is reported (if it exists).

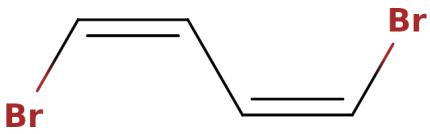


Figure 2.12: Illustration of a potential issue with generating cis/trans SMILES strings. The depicted molecule, described by the SMILES string BrC=CC=CBr contains two true cis/trans stereocenters. The configuration vector is `[1, 1]`, where `1` corresponds to a cis configuration. If the configuration is included in the SMILES string by adding a '/' before the first visited half-stereocenters, and adding a '\' before the first visited singly-bonded neighbors of the two second half-stereocenters, the SMILES string would become Br/C=C\/C=C\Br, which is not valid. Instead, the cis/trans half-stereocenters are processed in the order in which they appear in the string. For each of the half-stereocenters, it is then checked whether the corresponding first visited half-stereocenter already possesses an encoding. In this example, the string would be built as Br/C=C, then before the singly-bonded neighbor of the second carbon atom, the encoding '\' is added, leading to Br/C=C\ (*i.e.*, cis). Then, the string continues to be processed until the next cis/trans half-stereocenter is encountered, leading to Br/C=C\C=C. Since the singly-bonded neighbor of the third carbon atom is already assigned the encoding '\', the corresponding encoding before the singly-bonded neighbor of the fourth carbon atom is chosen accordingly, leading to the final string Br/C=C\C=C/Br (*i.e.*, cis and cis).

2.3 IMPLEMENTATION

The functionalities described in the previous sections are implemented in a C++ program called the *isomer enumerator*, in short *enu*. The current version of this program can be found on GitHub at `https://github.com/csms-ethz/CombiFF`. This repository also contains input and output files, as well as other programs related to the CombiFF scheme.[73,74] The code can be compiled with *cmake*.[209]

The following sections provide an overview over some of the functionalities implemented into the enumerator.

2.3.1 Specifying a Molecular Formula

There is some flexibility to define how many times an element type should occur in the molecular formula. For each element type, this number can be given as a single integer (*e.g.*, `H5`), as a list of integers (*e.g.*, `H[0,2,4,5]`), or as a range (*e.g.*, `H[0-5]`).

2.3.2 Implicit Hydrogen Atoms

As described previously, hydrogen atoms are treated implicitly (*i.e.*, distributed among the other atom types before the enumeration algorithm starts). In order to specify the types of molecules of interest more distinctly, the implicit hydrogen atoms can also be specified directly in the chemical formula. For example, `{CH1}1{CH2}2{OH1}3` will be translated to the formula $C_3H_8O_3$ with the restriction that it includes one carbon atom bonded to exactly one hydrogen atom, two carbon atoms bonded to exactly two hydrogen atoms, and three oxygen atoms bonded to exactly one hydrogen atom. This will generate the two constitutional isomers OCC(O)CO and OCCC(O)O (no stereoisomers), whereas there exist 36 constitutional and spatial isomers for the unrestricted formula $C_3H_8O_3$.

2.3.3 Filtering for Properties

Currently, the user may restrict the following molecular properties: the maximum bond order, the number of unsaturations (summing one for double bonds and cycles, and two for triple bonds), the total number of bonds (irrespective if single or multiple), the number of single bonds, the number of double bonds, the number of triple bonds, the number of quadruple bonds, as well as the number of cycles in the molecule. These restrictions can be set either as an integer, as a list of integers, or as a range of integers.

The filtering for the number of unsaturations is performed before the enumeration starts, by calculating the number of unsaturations for a molecular formula using Eq. (2.19) (Appendix Sec. 2.B.4). The filtering for the other properties is done whenever a new isomer is found. If the restrictions are not met, the isomer is not reported.

For example, enumerating all straight-chain alkane isomers C_nH_{2n+2} from C_1H_4 to $C_{20}H_{42}$ could be achieved with the formula specification `C[1-20]H*` and the restriction that the number of unsaturations should be zero.

2.3.4 Aromaticity

Currently, there is only a very basic implementation to recognize aromatic rings of size six using a substructure search for alternating single and double bonds. This procedure is able to recognize structures like benzene and pyridine, which is sufficient to eliminate duplicate isomers where the ordering of the single and double bonds is different. However, this functionality is still very limited. In the future, it will be extended to recognize aromatics during the enumeration procedure. In the meantime it is possible to post-process the *enu* output using a suitable cheminformatics library such as the RDKit [210] to recognize aromaticity for more complicated cases. Thanks to the convenient XML output format and the SMILES notation, such a post-processing is easy to implement.

2.3.5 Visualization

To visualize the output of *enu*, a small Python script is provided in the GitHub repository. It uses the Python3 [211] *xml.etree.ElementTree* module to parse the XML list of constitutional and spatial (if present) isomers, the Python library *pdfrw* [212] to concatenate PDFs, as well as the RDKit [210] cheminformatics library to create the visualizations (for an example, see Figure 2.13).

2.3.6 Family Definitions

The most straightforward way to use *enu* is *via* the command line by specifying a chemical formula (potentially including atoms with implicit hydrogens and filtering criteria, as described in the previous sections). However, one can also define a so-called *family* which offers more flexibility, *i.e.*, the use of element aliases, of a filtering for substructures, and of pseudoatoms. A brief overview is provided in the following sections.

Figure 2.13: Visualization of the 106 acyclic constitutional and spatial isomers of $C_5H_8Br_2$. The molecules were enumerated with *enu* and the depiction was generated with the RDKit [210] cheminformatics library.

Element Aliases

In order to provide more flexibility in the definition of the chemical formulas for which the isomers are to be enumerated, it is possible to define so-called *element aliases*. An element alias has a name and contains a set of element types. For example, an element alias for the halogen element types could be called "`Hal`" and contain the four types "`Br`", "`Cl`", "`F`" and "`I`". When two or more of the same element alias occur in a chemical formula, AND, XOR, and OR logic can be used to specify if they should be of the same element type, of a different element type, or whether both is allowed. The notation `Hal3` (AND) specifies that the halogen atoms have to be the same type (*e.g.*, `Br3`). `^Hal1^Hal2` (XOR) specifies that the first halogen atom has to be of a different type than the two other halogen atoms (*e.g*, `Br1Cl2`). Finally, `Hal1Hal1Hal1` (OR) specifies that any combination of three halogen atoms is allowed (*e.g.*, `Br1Cl1F1`,

`Br3`, or `Br1Cl2`). The formula to enumerate all straight-chain haloalkanes with ten carbon atoms and two halogen atoms of the same type can then be expressed as `C10H20Hal2` and is equivalent to enumerating the four chemical formulae $C_{10}H_{20}Br_2$, $C_{10}H_{20}Cl_2$, $C_{10}H_{20}F_2$ and $C_{10}H_{20}I_2$.

Filtering for Substructures

It is also possible to filter for the occurrence of substructures. The number of occurrences can be specified by a single integer, a list of numbers, or a range of numbers. By setting the occurrence to zero, one may also prevent the occurrence of a substructure. The implemented substructure search algorithm is the Ullmann algorithm[213] (see also Chapter 3, Sec. 3.2.2). As the enumerator is aimed for relatively small molecules, the performance of the Ullmann algorithm is not a bottleneck for the overall runtime. If this became an issue in the future, it could be replaced by a more modern algorithm such as VF2.[214,215]

Here, a substructure is defined by a name, a list of atoms, and an adjacency matrix stack (*i.e.*, the upper triangle of an adjacency matrix, written row-wise as a one-dimensional vector). Each of the atoms can be either an element type, an element type with a number of implicit hydrogens, an element alias, or a wildcard. If there is more than one element alias of the same type, it is also possible to use the AND, XOR and OR logic to specify if they should be of the same element types, of a different element type, or whether both are allowed. When element aliases are used and multiple occurrences of the substructure are requested, it is also possible to specify how the element types should occur across substructures, also using AND, XOR and OR logic.

When multiple substructures are required, the implemented convention is that there can be a maximum overlap of one atom between the matched substructures. For example, in the molecule CCCC, the substructure CC is found three times, but the substructure CCC is only found once.

Pseudoatoms

Applying substructure matching in the form of a post-processing step as described in the previous section can become very inefficient if a chemical formula is given with many potential isomers, where only a small subset of them contain the desired substructure(s).

Consider, for example, the formula `C[2-10]H*O4`, where the number of hydrogens is chosen such that there are two unsaturations. In total, there exist more than 560 million constitutional and spatial isomers. However, if one requires exactly two occurrences of the substructure COC(=O)H, only 1484 of these isomers contain the desired substructures. The enumeration of these isomers takes about 12 minutes on a laptop with an i7-8565U CPU. A major part of the computation time is thus wasted on constitutional isomers which are not reported. Note that there is no time wasted on enumerating redundant stereoisomers, as the stereoisomers are only generated for the constitutional isomers that are compatible with the given restrictions.

The number of redundant constitutional isomers can be reduced by using implicit hydrogens specifying the formula `C[O-8]{CH1}2H*{OH0}4`, such that isomers containing, *e.g.*, an oxygen-hydrogen bond are not generated during the enumeration. With this more specific formula, the enumeration time reduces to about 1 minute. However, for the larger example `C[3-10]H*O6`, where the number of hydrogens is chosen such that there are three unsaturations and exactly three occurrences of the substructure COC(=O)H are required, even with this trick of using the formula `C[O-7]{CH1}3H*{OH0}6`, the enumeration of the 1328 constitutional and spatial isomers takes about 50 minutes.

To solve this problem in a more general fashion, so-called *pseudoatoms* are introduced. A pseudoatom is a molecular substructure which only contains one atom that is not fully bonded, *i.e.*, can be connected to the atoms of the rest of the molecule. A pseudoatom is defined by a name, a list of atoms, and an adjacency matrix stack. The pseudoatom behaves like a normal atom during the enumeration process. The valence of the

pseudoatom corresponds to the valence of the not-fully-bonded atom minus the number of bonds that this atom forms with the other atoms within the pseudoatom. Whenever a new isomer is found, the pseudoatom is explicited in terms of its atom content, *i.e.*, the adjacency matrix is extended and canonicalized, the SMILES string is generated, and the stereoisomers are listed (if requested). Using a pseudoatom OC(=O)H, it takes less than a second to enumerate the 1328 constitutional and spatial isomers of `C[0-7]H*'OC(=O)H'3` with three occurrences of the substructure COC(=O)H, *i.e.*, much less than the above 50 minutes.

A caveat of this approach is that the canonicity test does not recognize if a pseudoatom can also be constructed using the atoms available in the rest of the molecule. When this is possible, there will be duplicate isomers in the enumeration list. However, based on to the canonical SMILES strings these duplicates can easily be removed by the user in a post-processing step. Thanks to the convenient XML output format and the SMILES notation, such a post-processing is easy to implement if required.

2.4 ILLUSTRATIVE RESULTS

The performance of the isomer enumerator is illustrated in the context of straight-chain alkanes from C_1H_4 to $C_{24}H_{50}$. All time measurements are taken from runs performed using a laptop with an i7-8565U CPU and are averaged over five runs. Table 2.1 lists the run times and number of enumerated constitutional and spatial isomers for the alkanes. Figure 2.14 shows the number of constitutional and spatial alkane isomers as a function of the number of carbon atoms, the time to enumerate these isomers, and the time spent per constitutional/spatial isomer during the enumeration process depending on the number of carbon atoms. The number of existing constitutional/spatial isomers, and thus also the wall-clock time, increases exponentially upon increasing the number of carbon atoms. The time

spent per constitutional isomer also increases exponentially, though with a much smaller slope. The time spent per spatial isomer remains relatively constant up to at least 24 carbon atoms.

Table 2.1: Run times of the isomer enumerator for alkane isomers. The table shows the number n_{consti} of constitutional isomers and the number n_{spatial} of spatial isomers of the straight-chain alkanes from C_1H_4 to $C_{24}H_{50}$, as well as the wall-clock time spent on their enumeration. The time t_{consti} is the wall-clock time in seconds it takes to enumerate the constitutional isomers, and the time t_{spatial} is the wall-clock time in seconds it takes to enumerate the spatial isomers. All calculations were performed on a laptop with an i7-8565U CPU and were averaged over five runs. The results are visualized in Figure 2.14.

molecule	n_{consti}	n_{spatial}	t_{consti} [s]	t_{spatial} [s]
C_1H_4	1	1	< 0.001	< 0.001
C_2H_6	1	1	< 0.001	< 0.001
C_3H_8	1	1	< 0.001	< 0.001
C_4H_{10}	2	2	< 0.001	< 0.001
C_5H_{12}	3	3	< 0.001	< 0.001
C_6H_{14}	5	5	< 0.001	< 0.001
C_7H_{16}	9	11	< 0.001	< 0.001
C_8H_{18}	18	24	< 0.001	< 0.001
C_9H_{20}	35	55	< 0.001	0.001
$C_{10}H_{22}$	75	136	0.001	0.002
$C_{11}H_{24}$	159	345	0.004	0.005
$C_{12}H_{26}$	355	900	0.007	0.011
$C_{13}H_{28}$	802	2412	0.020	0.027
$C_{14}H_{30}$	1858	6563	0.052	0.072
$C_{15}H_{32}$	4347	18127	0.139	0.199
$C_{16}H_{34}$	10359	50699	0.387	0.545
$C_{17}H_{36}$	24894	143255	1.118	1.539
$C_{18}H_{38}$	60523	408429	3.359	4.489
$C_{19}H_{40}$	148284	1173770	9.871	13.201
$C_{20}H_{42}$	366319	3396844	29.713	38.938
$C_{21}H_{44}$	910726	9892302	92.019	118.974
$C_{22}H_{46}$	2278658	28972080	285.502	368.725
$C_{23}H_{48}$	5731580	85289390	945.897	1188.105
$C_{24}H_{50}$	14490245	252260276	3183.896	3847.299

2.5 CONCLUSION

The goal of this chapter was to document the algorithms and implementation underlying the program *enu* for the enumeration of the constitutional isomers and stereoisomers of a molecular formula. Although the motivation for the development of this program was its integration into the CombiFF workflow,[73,74] *enu* is a stand-alone, freely-downloadable and

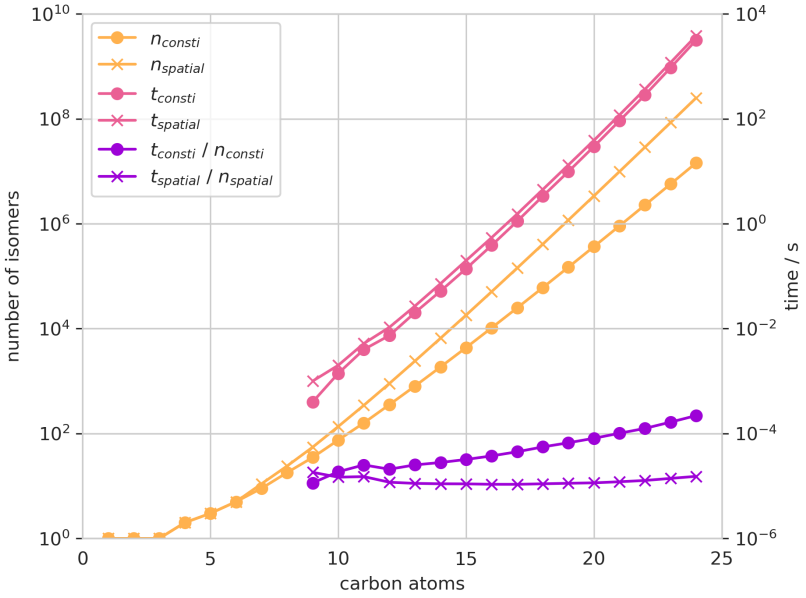


Figure 2.14: Illustration of the performance of the enumerator. The plot exemplifies the performance of the isomer enumerator in the context of the straight-chain alkanes C_nH_{2n+2} from C_8H_{18} to $C_{22}H_{46}$. The plot shows the number of alkane (stereo)isomers depending on the number of carbon atoms, the wall-clock time to enumerate these (stereo)isomers and the wall-clock time spent per (stereo)isomer for this enumeration. The left vertical axis shows the number of (stereo)isomers and the right vertical axis shows the elapsed wall-clock time. All calculations were performed on a laptop with an i7-8565U CPU and were averaged over five runs. The numerical values are shown in Figure 2.1

open-source program, which can be used for any other purpose in chem-informatics. An integration into other workflows can be easily achieved thanks to the convenient XML format and the reporting of isomers *via* canonical SMILES strings.

The illustrative example of the alkane isomers shows that the computational cost grows exponentially with the number of carbon atoms, just as the number of isomers. However, while the time spent per constitutional isomer also tends to increase exponentially, the time spent per spatial isomer stays relatively constant up to at least 24 carbon atoms.

Further development of the *enu* program will include: (*i*) complete

handling the stereochemical properties (chirality, double bonds) within cyclic systems; (*ii*) identifying aromaticity more comprehensively in the SMILES string generation; (*iii*) simplifying the input mechanism for atoms with variable valences (*e.g.*, sulfur or phosphorous); and (*iv*) extending of the special treatment of hydrogen atoms to all singly-connected entities (halogens, pseudoatoms, methyl groups) for computational efficiency.

2.A APPENDIX

In the following sections, the generation and canonicalization of constitutional isomers is described in detail. Some parts of the text were reproduced, or adapted and refined from the author's Master thesis.[194] For the convenience of the reader and for the sake of completeness, we felt it was important to include them in the present chapter.

2.B ENUMERATION OF CONSTITUTIONAL ISOMERS

2.B.1 Molecular Graphs

The definitions and notations adopted here are largely inspired from those of Ref. 185. A molecular graph is a connected labeled multigraph in which the vertices represent atoms and the edges account for the (single or multiple) covalent bonds between the atoms.[197] An example is shown in Figure 2.1 in Sec. 2.2.1. The *label vector* $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_K)$ of a molecular graph describing a molecule with $K+1$ atom types contains a list of these types (element symbols), each of them appearing only once. The *valence vector* $\boldsymbol{\delta} = (\delta_0, \delta_1, \ldots, \delta_K)$ describes the fixed valences of the atom types in $\boldsymbol{\alpha}$. For elements capable of presenting different valences (*e.g.*, S and P), the corresponding atom type can be split across different valences by associating it to different entries in the label and valence vectors. The *partition vector* $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_K)$ of the graph is defined such that λ_k corresponds to the number of occurrences of the element type α_k in the molecule.[185] Consequently, the total number N of atoms is given by

$$N = \sum_{k=0}^{K} \lambda_k \,. \tag{2.3}$$

The combination of a label vector $\boldsymbol{\alpha}$ and a partition vector $\boldsymbol{\lambda}$ corresponds to a molecular formula, which can be written as $\alpha_{0_{\lambda_0}}\alpha_{1_{\lambda_1}}\cdots\alpha_{K_{\lambda_K}}$. Such a combination can be used to create an *atom vector*

$$\begin{aligned}\boldsymbol{a} &= (a_0, a_1, \ldots, a_{N-1}) \\ &= (\underbrace{\alpha_0, \ldots, \alpha_0}_{\lambda_0\times}, \underbrace{\alpha_1, \ldots, \alpha_1}_{\lambda_1\times}, \cdots, \underbrace{\alpha_K, \ldots, \alpha_K}_{\lambda_K\times}) .\end{aligned} \tag{2.4}$$

The corresponding *degree vector*

$$\boldsymbol{d} = (d_0, d_1, \ldots, d_{N-1}) \tag{2.5}$$

$$= (\underbrace{\delta_0, \ldots, \delta_0}_{\lambda_0\times}, \underbrace{\delta_1, \ldots, \delta_1}_{\lambda_1\times}, \cdots, \underbrace{\delta_K, \ldots, \delta_K}_{\lambda_K\times}) \tag{2.6}$$

contains the valence of each atom in the molecule, *i.e.*, the number of covalent bonds it can form. The atoms in the atom vector are numbered with consecutive indices $i = 0, 1, 2, \ldots, N-1$. The indices of the λ_k atoms with the same label α_k are collected in the so-called *partition* p_k. There are $K+1$ such partitions

$$\begin{aligned} p_0 &= \{0, 1, \ldots, \lambda_0 - 1\} \\ p_1 &= \{\lambda_0, \lambda_0 + 1, \ldots, \lambda_0 + \lambda_1 - 1\} \\ &\ldots \\ p_k &= \left\{ \sum_{l=0}^{k-1} \lambda_l, \sum_{l=0}^{k-1} \lambda_l + 1, \ldots, \sum_{l=0}^{k} \lambda_l - 1 \right\} \\ &\ldots \\ p_K &= \left\{ \sum_{l=0}^{K-1} \lambda_l, \sum_{l=0}^{K-1} \lambda_l + 1, \ldots, N-1 \right\} , \end{aligned} \tag{2.7}$$

and the number of elements in partition p_k is equal to λ_k.

A molecular graph can be described by the combination of a label

vector $\boldsymbol{\alpha}$, a valence vector $\boldsymbol{\delta}$, a partition vector $\boldsymbol{\lambda}$, and an *adjacency matrix* $\boldsymbol{A} \in \mathbb{N}_0^{+\,N\times N}$. A matrix element $A_{i,j}$ of $\boldsymbol{A}$ describes the order of the bond possibly connecting the atom at position i to the atom at position j in the atom vector (or is set to zero in the absence of a bond). To be compatible with the degree vector, the adjacency matrix must satisfy

$$\sum_{j=0}^{N-1} A_{i,j} = \sum_{j=0}^{N-1} A_{j,i} = d_i \;\; \forall i \,. \tag{2.8}$$

2.B.2 Molecular Graph Isomorphism

For a given choice of $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\lambda}$ (*i.e.*, of a molecular formula and of atom-type valences), the specification of an adjacency matrix $\boldsymbol{A}$ (*i.e.*, of a covalent connectivity between the atoms) defines a unique labeled molecular graph. However, since the atoms of a common type in a molecule are physically indistinguishable, two labeled graphs that are directly related by a permutation in the labels of these atoms actually describe the same molecule (merely with a different atom numbering). In other words, for a given choice of $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\lambda}$, the same molecule can generally be represented by many different adjacency matrices $\boldsymbol{A}$. This observation is fundamentally important to the problem of isomer enumeration and is known as (molecular) graph isomorphism.

A *permutation* π is a linear ordering of the elements of a set, *i.e.*, any list of all the elements of the set in which each element appears exactly once.[216] The set of all permutations of a set with N elements is called the *symmetric group* S_N and contains $N!$ elements.[216] Here, the relevant permutations operate on the set $\{0, 1, \ldots, N-1\}$ of indices of the atom vector. Permutations can be formulated as a chain of successive transpositions (swaps), denoted by corresponding index tuples (pairs), *i.e.*, as

$$(0, j_0)(1, j_1)(2, j_2)\cdots(N-2, j_{N-2})(N-1, N-1)\,, \tag{2.9}$$

where a tuple (i, j_i) indicates that the atom at index i is to be swapped with the atom at index j_i in the atom vector. The transpositions are performed in sequence from left to right, and the restriction $j_i \geq i$ is imposed for each tuple (i, j_i). In addition, the two indices in a tuple must be contained in the same index partition p_k (*i.e.*, they must swap atoms of the same type). Tuples of the form (i, i) leave the position of index i identical and can thus (but do not have to) be left out of the chain of index tuples.

For a molecular graph with a partition vector $\boldsymbol{\lambda} = (\lambda_0, \lambda_1, \ldots, \lambda_K)$, there are $\lambda_k!$ ways to arrange the indices of the atoms in the k-*th* partition. Consequently, there are $\lambda_0! \cdot \lambda_1! \cdot \ldots \cdot \lambda_K!$ possibilities to arrange all the indices of the atom vector within their respective index partitions.[217] The set of all these permutations is noted by $S_{\boldsymbol{\lambda}}$ and is a subset of the symmetric group S_N. By definition, the permutations in $S_{\boldsymbol{\lambda}}$ leave the vectors $\boldsymbol{a}$ and $\boldsymbol{d}$ unchanged, but they do affect the adjacency matrix $\boldsymbol{A}$. Applying an index transposition (i, j_i) to a matrix corresponds to swapping rows i and j_i as well as columns i and j_i. The new adjacency matrix may differ from the original one, but it still describes the same molecule, just with a different ordering of the atom indices.

Considering two labeled molecular graphs defined by $(\boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{A})$ and by $(\boldsymbol{\alpha}', \boldsymbol{\delta}', \boldsymbol{\lambda}', \boldsymbol{A}')$, the graphs are called *compatible* if and only if

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}' \text{ and } \boldsymbol{\delta} = \boldsymbol{\delta}' \text{ and } \boldsymbol{\lambda} = \boldsymbol{\lambda}' , \tag{2.10}$$

i.e., they correspond to the same molecular formula. Two graphs are called *isomorphic* if and only if they are compatible and[185]

$$\exists \pi \in S_{\boldsymbol{\lambda}} : \boldsymbol{A}\pi = \boldsymbol{A}' . \tag{2.11}$$

In this case, one also says that the corresponding adjacency matrices are isomorphic (noted $\boldsymbol{A}' \sim \boldsymbol{A}$). Two isomorphic graphs are equivalent representations of the same molecules, as illustrated in Figure 2.15. Note

that identity ($\boldsymbol{A}' = \boldsymbol{A}$) is a special case of isomorphism. Finally, two graphs are called *isomeric* if they are compatible but not isomorphic. Two isomeric graphs describe molecules with the same chemical formulas but that are structurally different.[218]

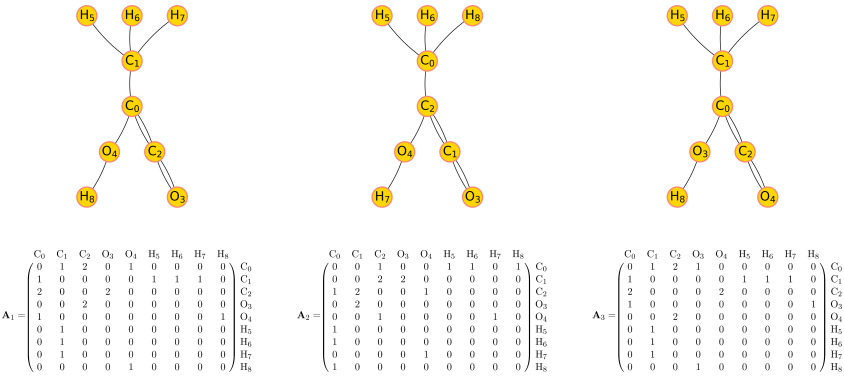


Figure 2.15: Illustration of molecular graph isomorphism. Three isomorphic molecular graphs and their corresponding adjacency matrix are shown including atom indices, representing an isomer of $C_3O_2H_4$.

2.B.3 Adjacency Matrix canonicity

In order to have a unique representation of the molecular topology in the form of a labeled multigraph, a *lexicographical ordering* can be used as canonicity criterion for the adjacency matrix. An adjacency matrix $\boldsymbol{A}$ is lexicographically larger than an adjacency matrix $\boldsymbol{A}'$ (noted $\boldsymbol{A} > \boldsymbol{A}'$) provided that[185]

$$\exists i_0, j_0 : \left(A_{i,j} = A'_{i,j} \ \wedge \ \left(A_{i_0,j_0} > A'_{i_0,j_0}\right) \ \forall\, (i,j) < (i_0,j_0)\right), \tag{2.12}$$

with the definition[185]

$$(i,j) < (k,l) \ \Leftrightarrow \ (i < k) \ \vee \ (i = k \wedge j < l)\,. \tag{2.13}$$

In plain words, when the two matrices are read row-by-row from the top left to the bottom right, the first difference encountered determines the lexicographical ordering.

The canonical adjacency matrix of a molecular graph is then defined as the lexicographically largest among all possible adjacency matrices, which in turn defines a canonical labeling of the atoms in the molecular graph. Thus, for a given choice of $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$, and $\boldsymbol{\lambda}$, an adjacency matrix $\boldsymbol{A}$ is canonical if and only if

$$\nexists \pi \in S_{\boldsymbol{\lambda}} : \; \boldsymbol{A}\pi > \boldsymbol{A} \,. \tag{2.14}$$

For a unique representation of molecules, the canonicity criterion for $\boldsymbol{A}$ must be accompanied by a canonicity criterion for the ordering of the atom types in the vector $\boldsymbol{\alpha}$. The ordering adopted in the *enu* program is as follows. The atom types are sorted from highest to lowest valence δ_k. If multiple atom types have the same valence, they are sorted according to the size of their partitions (*i.e.*, the number λ_k of occurrences of the atom in the molecule) in increasing order. If multiple isovalent atom types have the same partition sizes, they are sorted by their atomic number in increasing order. For example, for $\mathrm{C_1O_1N_1Cl_1Br_3F_1}$, one would order C, N, O, F, Cl, Br. This specific choice of ordering in terms of valence and occurrence can lead to considerable performance increases during the enumeration process (see Sec. 2.B.6).

With these definitions, a molecule can be uniquely represented as a labeled molecular graph with a canonical adjacency matrix. An example is provided for the molecule depicted in Figure 2.16 with

$$\boldsymbol{\alpha} = (\mathrm{C}, \mathrm{H}) \tag{2.15}$$

$$\boldsymbol{\delta} = (4, 1) \tag{2.16}$$

$$\boldsymbol{\lambda} = (3, 2) \,. \tag{2.17}$$

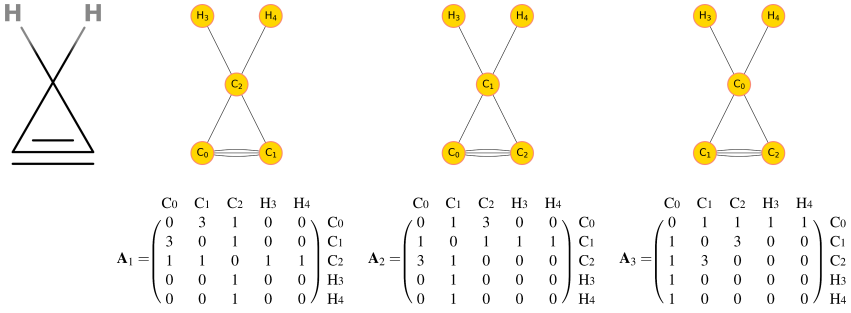


Figure 2.16: Illustration of molecular graph isomorphism. (Left): A constitutional isomer of C_3H_2. (Right): The three possible isomorphic molecular graphs for the molecule on the left, including the corresponding adjacency matrices. The permutations to show isomorphism between the matrix pairs are $\boldsymbol{A}_2 = \boldsymbol{A}_1(1,2)$, $\boldsymbol{A}_3 = \boldsymbol{A}_1(0,1)(1,2)$, and $\boldsymbol{A}_3 = \boldsymbol{A}_2(0,1)$. Using the lexicographical ordering, it can be seen that $\mathbf{A}_1 > \mathbf{A}_2 > \mathbf{A}_3$. Consequently, $\mathbf{A}_1$ is canonical, whereas $\mathbf{A}_2$ and $\mathbf{A}_3$ are not.

2.B.4 Filling Algorithm to Enumerate Adjacency Matrices

Enumerating all the unique constitutional isomers of a given molecular formula amounts to finding all the canonical adjacency matrices associated with this formula. An outline of the orderly enumeration scheme proposed by Grund[185] is provided in Algorithm 1.

In order to define the two functions `FindMaxEntry` and `DecreasePossible` used in Algorithm 1, three matrices are introduced. The matrix $\mathbf{M}$ is defined as[185]

$$\mathbf{M} := (M_{i,j})_{0 \leq i,j \leq N-1}\,, \qquad M_{i,j} = \begin{cases} \min(d_i, d_j) & d_i \neq d_j \\ d_i - 1 & d_i = d_j, i \neq j \\ 0 & i = j \end{cases}. \tag{2.18}$$

The entry $M_{i,j}$ corresponds to the maximum value that $A_{i,j}$ can have, such that the valences of the atoms a_i and a_j are not exceeded, and at least one of the two atoms can be connected to one or more other atoms. Note that diatomic molecules represent an exception, and are handled explicitly in *enu* (with $M_{0,1} = d_0 = d_1$).

In *enu*, a slightly modified version of this matrix is used. For a given

Algorithm 1: Filling Algorithm

```
/* The two functions IncreaseIndex and DecreaseIndex can be found in Algorithm 2 */
/* FindMaxEntry and DecreasePossible are defined in Algorithm 3 and Algorithm 4 */

/* start orderly enumeration with a ForwardStep at the first element of the matrix */
ForwardStep (0,0)

Function ForwardStep(i,j):
    IncreaseIndex (i,j)
    if FindMaxEntry (i,j,x) then
        set A_{i,j} to x
        if i == N − 2 then
            /* last index of the matrix is reached ⇒ potential adjacency matrix is complete */
            Print(current matrix)
            /* continue search for next potential adjacency matrix by back stepping */
            BackwardStep (i,j)
        else
            /* continue filling the matrix */
            ForwardStep (i,j)
    else
        /* no viable entry was found at the current matrix position */
        BackwardStep (i,j)

Function BackwardStep(i,j):
    if j == 1 then
        /* algorithm has terminated */
        return
    else
        DecreaseIndex (i,j)
        if DecreasePossible (i,j) then
            /* decrease current matrix entry by 1 and continue with a ForwardStep */
            set A_{i,j} to A_{i,j} − 1
            ForwardStep(i,j)
        else
            /* continue backstepping until an entry is found that can be decreased */
            BackwardStep(i,j)
```

atom vector, the *degree of unsaturation* can be calculated as[219]

$$d_{\text{unsat}} = 1 + \frac{1}{2}\left(\sum_{k=0}^{K} \lambda_k(\delta_k - 2)\right) , \tag{2.19}$$

where δ_k is the valence of atom type α_k. It also holds that[219]

$$d_{\text{unsat}} = n_{\text{db}} + 2 \cdot n_{\text{tb}} + n_{\text{ring}} , \tag{2.20}$$

where n_{db} is the number of double bonds, n_{tb} the number of triple bonds, and n_{ring} the number of rings in the molecule. Consequently, the maximum possible bond degree in a molecule is equal to $d_{\max} = 1 + d_{\text{unsat}}$. For this

Algorithm 2: Methods to increase and decrease indices

```
Function IncreaseIndex(i,j):
    if j == (N − 1) then
        i ++
        j = i +1
    else
        j ++
Function DecreaseIndex(i,j):
    if j == i +1 then
        i −−
        j = N − 1
    else
        j −−
```

reason, an additional restriction on **M** can be introduced as

$$M_{i,j} = \begin{cases} \min(d_i, d_j, d_{\max}) & d_i \neq d_j \\ \min(d_i - 1, d_{\max}) & d_i = d_j, i \neq j \\ 0 & i = j \end{cases} . \tag{2.21}$$

This restriction is useful for molecules with low degrees of unsaturation; considering atoms with valences ≤ 4, it results in no gain as soon as $d_{\max} \geq 3$.

The two upper triangular matrices **L** and **C** are defined as[185]

$$L := (L_{i,j})_{0 \leq i < j \leq N-1} , \qquad L_{i,j} := \min \left(d_i, \sum_{s=j+1}^{N-1} M_{i,s} \right) \tag{2.22}$$

$$C := (C_{i,j})_{0 \leq i < j \leq N-1} , \qquad C_{i,j} := \min \left(d_j, \sum_{s=i+1}^{N-1} M_{s,j} \right) \tag{2.23}$$

An entry $L_{i,j}$ corresponds to the maximum possible *row capacity* after position (i, j), *i.e.*, the maximum number of potential bonds atom a_i can form with the atoms a_{j+1}, a_{j+2}, ..., a_{N-1}. Analogously, an entry $C_{i,j}$ corresponds to the maximum possible *column capacity* after position (i, j).

Additionally, the values $\hat{L}_{i,j}$ and $\hat{C}_{i,j}$ are defined as[185]

$$\hat{L}_{i,j} := d_i - \sum_{s=0}^{j-1} A_{i,s} \tag{2.24}$$

$$\hat{C}_{i,j} := d_j - \sum_{s=0}^{i-1} A_{s,j} \,. \tag{2.25}$$

$\hat{L}_{i,j}$ corresponds to the number of bonds that still have to be formed by atom a_i (including and after position (i,j)) such that it is fully connected, and $\hat{C}_{i,j}$ corresponds to the number of bonds that still have to be formed by atom a_j (including and after position (i,j)).

Using these definitions, one may now determine what constitutes a viable matrix element at a position (i,j). Four conditions have to be met by a potential matrix element x at (i,j) in the forward step[185]

$$x \geq 0 \tag{2.26}$$

$$x \leq \min\{\hat{L}_{i,j}, \hat{C}_{i,j}, M_{i,j}\} \tag{2.27}$$

$$x \geq \hat{L}_{i,j} - L_{i,j} \tag{2.28}$$

$$x \geq \hat{C}_{i,j} - C_{i,j} \,. \tag{2.29}$$

The two last condition ensure that x is sufficiently large for the atoms a_i and a_j to be both saturated in their respective valences once the corresponding row/column is filled. For example, for $i = N-2$ and $j = N-1$, if $a_i = a_{N-2}$ (*i.e.*, the second to last atom in the atom vector) is a carbon atom with valence four that is already singly-bonded to one of the atoms a_0 to a_{j-1}, then $\hat{L}_{i,j} = 3$ (*i.e.*, a_i still needs to form three bonds to be saturated in its valence). $L_{i,j} = 0$ (*i.e.*, the row capacity after $(i,j) = (N-2, N-1)$ is zero, since the row is complete). Then we have $\hat{L}_{i,j} - L_{i,j} = 3 - 0 = 3$, *i.e.*, in order to saturate a_i in its valence, there would need to be a bond of at least degree three between atoms a_i and a_j. If the last atom in the atom vector, $a_j = a_{N-1}$, is, *e.g.*, a

chlorine atom with valence one, this is not possible due to the restriction $M_{N-1,N-2} = 1$.

Analogously, during the backward step, an entry may be decreased by one if

$$A_{i,j} - 1 \geq 0 \tag{2.30}$$

$$(A_{i,j} - 1) \geq \hat{L}_{i,j} - L_{i,j} \tag{2.31}$$

$$(A_{i,j} - 1) \geq \hat{C}_{i,j} - C_{i,j}\,. \tag{2.32}$$

Note that $A_{i,j} - 1$ is the new potential matrix element at position (i, j) and that Eqs. (2.28) and (2.29) are identical to Eqs. (2.31) and (2.32), respectively, when setting x to $A_{i,j} - 1$. This ensures that the current row/column can still be saturated after the backward step.

Given the above restrictions, the routines `FindMaxEntry` and `DecreasePossible` used by the forward and backward steps, respectively, are outlined in Algorithm 3 and Algorithm 4.

Algorithm 3: FindMaxEntry

```
Function FindMaxEntry(i,j,x):
    x = min(L̂_{i,j}, Ĉ_{i,j}, M_{i,j})
    if x ≥ max(0, L̂_{i,j} − L_{i,j}, Ĉ_{i,j} − C_{i,j}) then
        return true
    else
        return false
```

Algorithm 4: DecreasePossible

```
Function DecreasePossible(i,j):
    x = A_{i,j} − 1
    if x ≥ max(0, L̂_{i,j} − L_{i,j}, Ĉ_{i,j} − C_{i,j}) then
        return true
    else
        return false
```

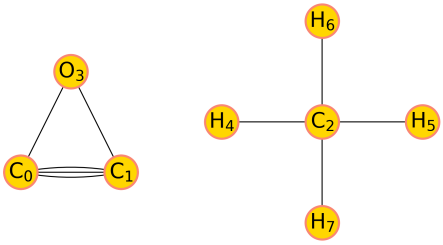
2.B.5 Connectivity Test

Algorithm 5 shows the connectivity test employed in the isomer enumerator.

Algorithm 5: Connectivity Test

```
create a last-in-first-out stack that contains the first vertex
create a boolean vector visited, where each entry corresponds to a vertex
set the first entry of visited to true, all others to false
while stack not empty do
    pop top vertex in stack
    for each neighbor of current vertex do
        if neighbor has not yet been visited then
            set visited at position of neighbor to true
            push neighbor to the stack
for each vertex do
    if vertex has not been visited then
        return false
return true
```

Its application can be illustrated with a simple example for the disconnected molecular graphs depicted in Figure 2.17. The connectivity test proceeds as follows



$$\mathbf{A} = \begin{pmatrix} 0 & 3 & 0 & 1 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Figure 2.17: Example of a disconnected molecular graph for $\boldsymbol{\alpha} = (\mathrm{C}, \mathrm{O}, \mathrm{H})$, $\boldsymbol{\delta} = (4, 2, 1)$, and $\boldsymbol{\lambda} = (3, 1, 4)$ with the corresponding adjacency matrix $\boldsymbol{A}$.

- initialize: `visited = [1,0,0,0,0,0,0,0]`, `stack={0}`
- pop top vertex in `stack`: `0`
  - push unvisited neighbors of `0` to `stack` and set corresponding entry in `visited` to `1`

  - unvisited neighbors are `1` and `3`
  - `stack={1,3}`, `visited=[1,1,0,1,0,0,0,0]`
- pop top vertex in `stack`: `3`
  - push unvisited neighbors of `3` to `stack` and set corresponding entry in `visited` to `1`
  - `3` does not have any unvisited neighbors
  - `stack = {1}`, `visited=[1,1,0,1,0,0,0,0]`
- pop top vertex in `stack`: `1`
  - push unvisited neighbors of `1` to `stack` and set corresponding entry in `visited` to `1`
  - `1` does not have any unvisited neighbors
  - `stack = {}`, `visited=[1,1,0,1,0,0,0,0]`
- stack is empty and some entries in `visited` are still `0` ⇒ the graph is not connected

2.B.6 Canonicity Test

Permutation Trees

To test a matrix for canonicity, the permutations $\pi \in S_{\boldsymbol{\lambda}}$ need to be available. These permutations can be systematically represented using a *permutation tree*.[185,220] The permutation tree for a given molecular formula with N atoms is a tree with height N and $\lambda_0! \cdot \lambda_1! \cdot \ldots \cdot \lambda_K!$ leaves. The nodes of the tree are transpositions (swaps) of index pairs (i, j) $(0 \leq i \leq j \leq N-1)$. Recalling that the permutations are sorted, an index can only be swapped with a higher index in the same partition (see Eq. (2.7)). Thus, for a given index $i \in p_k$, the possible swaps are

$$(i,\, i),\, (i,\, i+1),\, (i,\, i+2),\, \ldots,\, (i,\, i+\lambda_k-1)\,. \tag{2.33}$$

These tuples with the first element set to i are the children of the nodes with depth i in the permutation tree. Consequently, the subtree of a node (i, j) is identical to the subtree of a node (i, j'). For example, the nodes with depth 1 (*i.e.*, the children of the root) are all of the form $(0, j)$, and the nodes with depth N (*i.e.*, the leaves of the tree) are all of the form $(N-1, N-1)$. Figure 2.18 provides an example of such a permutation tree.

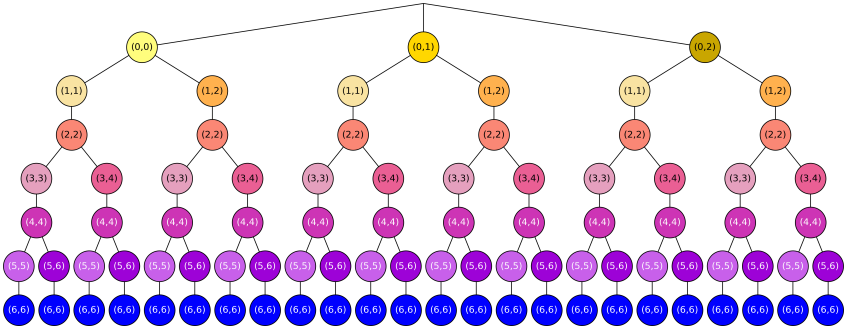


Figure 2.18: Example of a permutation tree. This figure shows the permutation tree for a molecular graph with $\boldsymbol{\lambda} = (3, 2, 2)$. Nodes that contain the same transposition are displayed in the same color.

A path from the root of the tree to a leaf defines a unique permutation $\pi \in S_{\boldsymbol{\lambda}}$ by combining the encountered nodes into a chain of index transpositions that are applied successively. The set of all possible paths from the root to the leaves spans the entire set $S_{\boldsymbol{\lambda}}$. Consequently, a depth-first traversal of the tree is a systematic way of listing all possible permutations in $S_{\boldsymbol{\lambda}}$.

Representation Systems

The permutation trees introduced above permit to visualize the group $S_{\boldsymbol{\lambda}}$. However, even for a small $S_{\boldsymbol{\lambda}}$ such a tree would quickly become too large to store explicitly. A more convenient method for storing permutation trees relies on the use of a *representation system.*[185] For a molecular graph

with N atoms, such a system consists of N ordered sets, where the i^{th} set contains the allowed permutations of atom a_i.

For example, the representation system of the permutation tree shown in Figure 2.18 reads:

$(0,0)$, $(0,1)$, $(0,2)$
$(1,1)$, $(1,2)$
$(2,2)$
$(3,3)$, $(3,4)$
$(4,4)$
$(5,5)$, $(5,6)$
$(6,6)$

The algorithm used in *enu* to achieve a depth-first traversal of the permutation tree given its representation system is shown in Algorithm 6. The vector `current_permutation` is used to to indicate the current permutation in the traversal. This vector of integers has the same size N as the representation system. For example, for the above representation system, the vector `current_permutation = (2,0,0,0,0,1,0)` corresponds to the permutation $(0,2)(1,1)(2,2)(3,3)(4,4)(5,6)(6,6)$. The next permutation found by Algorithm 6 will be
$(0,2)(1,1)(2,2)(3,4)(4,4)(5,5)(6,6)$, represented by
`current_permutation = (2,0,0,1,0,0,0)`. A key advantage of this implementation is that it is very easy to skip subtrees of the permutation tree by changing the value of `cur_index` before the while-loop as desired (instead of setting it to $N-1$). This becomes important when pruning permutation trees (see Sec. 2.B.6).

Naïve Canonicity Test

Given the representation system for the permutation tree of a molecular graph, the canonicity test for a given adjacency matrix $\mathbf{A}$ could be carried out in a brute-force fashion by performing a depth-first traversal of the tree and applying every possible permutation $\pi \in S_{\boldsymbol{\lambda}}$ to $\mathbf{A}$. An adjacency

Algorithm 6: Traversal of Representation System

```
/* current_permutation represents the current permutation. */
/* cur_index stores the index of the last element that was changed in current_permutation  */
/* N is the size of the representation system */

/* The function returns true if a next permutation was found (i.e. we are not yet at the last permutation
   of the traversal) and stores the next permutation using current_permutation. */
/* The function returns false if there are no more permutations. */
Function GetNextPermutation(const representation_system, current_permutation, N):
    cur_index = N − 1
    while cur_index ≥ 0 do
        if current_permutation [ cur_index ] + 1 < size (representation_system [ cur_index ]) then
            current_permutation [ cur_index ]++
            return true
        else
            current_permutation [ cur_index ] = 0
            cur_index − −
    return false
```

matrix $\mathbf{A}$ is rejected if at some point, a permutation $\pi \in S_{\boldsymbol{\lambda}}$ is found for which $\mathbf{A}\pi > \mathbf{A}$. However, even for molecular formulas with only a small number of atoms and few isomers, the filling algorithm generates many potential adjacency matrices that have to be tested for canonicity before they are rejected. Additionally, the number of permutations in $S_{\boldsymbol{\lambda}}$ can become huge even for small molecules. Taking C_3H_8 as an example, the filling algorithm creates 80 potential adjacency matrices. Among these, 40 represent connected molecular graphs and have to be tested for canonicity. There are $|S_{\{3,8\}}| = 3! \cdot 8! = 241\,920$ possible permutations that have to be applied to each of these matrices. At the end, only one of these adjacency matrices is canonical, the one corresponding to the canonical representation of propane. In the case of C_4H_{10}, there are 317 potential adjacency matrices, 81 of which are connected, and $|S_{\{4,10\}}| = 4! \cdot 10! = 87\,091\,200$ possible permutations. Here, two matrices are canonical, the ones corresponding to the canonical representations of *n*-butane and of isobutane. Both the number of adjacency matrices and the number of possible permutations increase exponentially with the number of atoms, making such a brute force canonicity test unusable.[185]

Blockwise Canonicity Test

Grund proposed a more efficient canonicity test.[185] For a given label vector $\boldsymbol{\alpha}$ and a corresponding partition vector $\boldsymbol{\lambda}$ of size $K + 1$, the potential

adjacency matrices can be subdivided into $K+1$ blocks of sizes $\lambda_0, \ldots, \lambda_K$, respectively. The different blocks of an adjacency matrix $\mathbf{A}$ are denoted by $\mathbf{A}^{(0)}, \mathbf{A}^{(1)}, \ldots, \mathbf{A}^{(K)}$, where the *k-th* block is defined as,[185]

$$\begin{aligned} \mathbf{A}^{(k)} := (A_{i,j}^{(k)}) \;\; & i,j \in p_k, p_{k+1}, \ldots, p_K \\ & \wedge \;\; (i \in p_k \;\vee\; j \in p_k)\,, \;\; 0 \leq k \leq K \end{aligned} \tag{2.34}$$

In plain words, the *k-th* block describes the bonds of the atoms of the same type α_k with each other, as well as with the atoms of types α_l with $l > k$, *i.e.*, that come after α_k in the label vector. As an example, the three blocks of the adjacency matrix in Figure 2.1 in Sec. 2.2.1 for $C_3O_2H_4$ can be illustrated as

$$\mathbf{A} = \left(\begin{array}{ccccccccc} 0 & 1 & 2 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 2 & 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right) . \tag{2.35}$$

Since the filling algorithm constructs the matrices row by row, and the symmetric entries are set simultaneously, it also fills them blockwise.[185] This means that a partially filled adjacency matrix can already be tested for canonicity within the filled blocks. Since the definition of the lexicographical order of two matrices depends on the first unequal element, if the already filled block can be permuted to a lexicographically larger form, the current filling of the blocks can be immediately stopped as it will never lead to a canonical matrix.

This approach is included in the filling algorithm by adjusting the `ForwardStep` function. Whenever a viable matrix entry x is found in the

forward step, and the current index (i, j) is the final index of a block, the matrix can already be tested for canonicity. If the matrix is canonical at that point, the next forward step is called. Otherwise, the algorithm continues with a backward step.

In addition to detecting non-canonical matrices early, a canonicity test on a partially filled matrix is less time consuming, since the lexicographical ordering only has to be checked for the current block (since the previous blocks have already been tested). Nevertheless, it still involves processing all permutations of $S_{\boldsymbol{\lambda}}$ (or until a lexicographically larger adjacency matrix is found).

Pruning the Permutation Tree: Stabilizers

To further improve the performance of the canonicity test, the concept of *stabilizers* is introduced. The stabilizer of the blocks $0, \ldots, k$ is defined as the set [185]

$$Aut^{(k)}(\mathbf{A}) := \left\{ \pi \in S_{\boldsymbol{\lambda}} : \mathbf{A}^{(0)} = \mathbf{A}^{(0)}\pi, \ldots, \mathbf{A}^{(k)} = \mathbf{A}^{(k)}\pi \right\} \subseteq S_{\boldsymbol{\lambda}} \,. \tag{2.36}$$

In plain words, the representation system of the stabilizer set $Aut^{(k)}(\mathbf{A})$ consists of all permutations in $S_{\boldsymbol{\lambda}}$ which leave the blocks $0, \ldots, k$ of $\mathbf{A}$ identical upon application. The set $Aut(\mathbf{A}) := Aut^{(K)}(\mathbf{A})$ is the so-called *automorphism group* of the entire matrix.

Clearly, it must hold that, for a given matrix $\mathbf{A}$ [185]

$$Aut \equiv Aut^{(K)} \subseteq Aut^{(K-1)} \subseteq \cdots \subseteq Aut^{(k)} \subseteq \cdots \subseteq Aut^{(0)} \subseteq S_{\boldsymbol{\lambda}} \tag{2.37}$$

As previously discussed, whenever a new block $k+1$ is filled, the previous blocks up to k were already checked to be canonical. Thus, it holds that for $0 \leq l \leq k$,

$$\mathbf{A}^{(l)} \geq \mathbf{A}^{(l)}\pi \quad \forall\, \pi \in S_{\boldsymbol{\lambda}} \,. \tag{2.38}$$

Due to the restrictions on the permutations in $S_{\boldsymbol{\lambda}}$, this must remain true after block $k+1$ is filled. Obviously, the only permutations that still have the potential to produce a lexicographically larger adjacency matrix in the blocks

$0, \ldots, k+1$ are the ones that leave the blocks up to k identical, as all the other permutations lead to an adjacency matrix which is smaller within the blocks $0, \ldots, k$. These permutations correspond to the ones contained in the stabilizer $Aut^{(k)}$.[185]

Taking this observation into account, whenever a new block $k+1$ is filled and the matrix is to be tested for canonicity, it is sufficient to test the new block $k+1$, and the canonicity test only needs to consider the permutations in $\pi \in Aut^{(k)}$. A block $k+1$ is thus canonical if[185]

$$\mathbf{A}^{(k+1)} \geq \mathbf{A}^{(k+1)}\pi \quad \forall\, \pi \in Aut^{(k)}(\mathbf{A})\,. \tag{2.39}$$

In order to perform a canonicity test of block $(k+1)$, $Aut^{(k)}$ is required. The straightforward approach for obtaining this stabilizer is to produce it during the canonicity test of block k. Since the permutations $\pi \in Aut^{(k-1)}$ are used for the canonicity test of this block, and a lexicographical comparison between the matrices $\mathbf{A}$ and $\mathbf{A}\pi$ has to be performed, one simply has to find all the permutations $\pi \in Aut^{(k-1)}$ for which

$$\mathbf{A}^{(k)} = \mathbf{A}^{(k)}\pi\,. \tag{2.40}$$

Once a stabilizing permutation π is found at a node (i, j) during the depth-first traversal of the permutation tree, it holds that[185]

$$\mathbf{A}^{(k)} \geq \mathbf{A}^{(k)}\pi' \tag{2.41}$$

for all permutations π' found by traversing the subtree of node (i, j). Due to the depth-first traversal of the permutation tree, and since the current permutation is a stabilizing one, the permutations found by traversing this subtree were all already tested during the current canonicity check. Since the canonicity test is still ongoing, the current matrix was not (yet) found to be non-canonical. Thus, any permutation that is encountered in the subtree of the current stabilizing permutation cannot lead to a lexicographically larger adjacency matrix. Finding the first stabilizing permutation during every part of the depth-first traversal is thus sufficient to test the canonicity of a matrix block. These first encountered stabilizing permutations form the representation system of the stabilizer $Aut^{(k)}$. Once such a permutation is found, it is added to the representation system

and further traversal of the current subtree can be stopped. When block $k+1$ is checked for canonicity, only the permutations formed by the representation system of $Aut^{(k)}$ have to be considered.

Consider the permutation tree in Figure 2.18. Let us assume we find that the permutation $(0,1)$ is a stabilizer of block zero. When we encounter this permutation, we already checked the permutations $(5,6)$, $(3,4)$, $(3,4)(5,6)$, $(1,2)$, $(1,2)(5,6)$, $(1,2)(3,4)$, and $(1,2)(3,4)(5,6)$. All of these permutations must have lead to an adjacency matrix that is either smaller or equal to the current one (otherwise the current matrix would already have been rejected). Since the permutation $(0,1)$ is a stabilizer and thus leaves the adjacency matrix unchanged, checking the subtree of permutation $(0,1)$, *i.e.*, the permutations $(0,1)(5,6)$, $(0,1)(3,4)$, $(0,1)(3,4)(5,6)$, $(0,1)(1,2)$, $(0,1)(1,2)(5,6)$, $(0,1)(1,2)(3,4)$, and $(0,1)(1,2)(3,4)(5,6)$ is guaranteed to also produce only adjacency matrices that are either smaller or equal to the current one. Finally, if the encountered stabilizers are, for example, $(3,4)(5,6)$ and $(0,1)$, the representation system for the stabilizers of block zero will be

$(0,0)$, $(0,1)$
$(1,1)$
$(2,2)$
$(3,3)$, $(3,4)(5,6)$
$(4,4)$
$(5,5)$
$(6,6)$.

Therefore, the only permutations that need to be considered for the canonicity test of block one will be $(3,4)(5,6)$, $(0,1)$, and $(0,1)(3,4)(5,6)$.

Another helpful observation is that a stabilizer of a previous block is guaranteed to be a stabilizer of the current block, provided that it only affects rows/columns that are part of the previous block. For example, if $(0,1)$ is a stabilizer of block zero, this permutation does not actually have to be checked when we test block one for canonicity. It will swap rows zero and one as well as columns zero and one, which only has an effect on block zero. Since the permutation is a stabilizer of block zero, this leaves the matrix unchanged. The permutation $(0,1)$ can thus directly be added to the representation system of the stabilizers of block one and its subtree can be skipped.

Semi-Canonicity

In addition to the central concept of canonicity, Grund uses the idea of *semi-canonicity*[185] as a weak canonicity criterion. If an adjacency matrix is not semi-canonical, it cannot be canonical, but if it is semi-canonical it is not necessarily canonical. It is based on determining a refined partitioning of the atom vector according to the bonds they form in **A**. A row of the matrix is semi-canonical if the matrix elements within each of these refined partitions are decreasing. The major advantage of the semi-canonicity test is its relatively low computational cost compared to the full blockwise canonicity test described above. The semi-canonicity criterion can be used directly at every iteration of the forward step, instead of just blockwise. It introduces an additional restriction on the current matrix element (*i.e.*, it cannot be larger than the previous element if that element is in the same refined partition), potentially skipping many partially filled matrices that would otherwise be discarded at a later stage.

This provides an additional constraint on the choice of a new matrix entry in the forward step. The value set at a new entry cannot be larger than the previous value if they are in the same refined partition. This restriction is added to the function `FindMatrixEntry` in the `ForwardStep`.

Index Jumps

A fast canonicity test represents one opportunity to make the enumeration algorithm more efficient. Another useful trick is to skip as many non-canonical matrices as possible. Grund introduced a *canonical learning criterion*, which allows to skip the orderly generation of other non-canonical adjacency matrices when a non-canonical matrix is generated.[185] It is based on identifying the index of the first different element in the current non-canonical matrix and the lexicographically larger matrix that was found during the canonicity test.

2.B.7 Special Treatment of Hydrogen Atoms

For a given label vector $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{K-1}, \mathrm{H})$ and partition vector $\boldsymbol{\lambda} = (\lambda_0, \ldots, \lambda_{K-1}, \lambda_K)$, λ_K describes the number of hydrogen atoms in the molecular

formula. The hydrogen vector can be defined as[185]

$$\mathbf{h} = (h_0, ..., h_{\hat{N}-1}) \,, \tag{2.42}$$

where

$$\hat{N} := \sum_{k=0}^{K-1} \lambda_k \tag{2.43}$$

is the number of non-hydrogen atoms and

$$\sum_{i=0}^{\hat{N}-1} h_i = \lambda_K \,. \tag{2.44}$$

In plain words, the length of the hydrogen vector is equal to the number of non-hydrogen atoms in the molecule, and its entries sum up to the number λ_K of hydrogen atoms in the molecule. This hydrogen vector will be used to describe a possible distribution of the hydrogen atoms in terms of their binding to the non-hydrogen atoms of the molecule.

Given the partition vector

$$\bar{\boldsymbol{\lambda}} = (\lambda_0, \lambda_1, \ldots, \lambda_{K-1}) \tag{2.45}$$

of the non-hydrogen atoms, two hydrogen vectors $\mathbf{h}$ and $\mathbf{h}'$ are equivalent if there exists a $\pi \in S_{\bar{\boldsymbol{\lambda}}}$ such that[185]

$$\mathbf{h}' = \mathbf{h}\pi \,. \tag{2.46}$$

Analogous to the definition for adjacency matrices, a hydrogen vector $\mathbf{h}$ is said to be *canonical* if

$$\mathbf{h} \geq \mathbf{h}\pi \quad \forall\, \pi \in S_{\bar{\boldsymbol{\lambda}}} \,, \tag{2.47}$$

which is equivalent to the condition that the entries of $\mathbf{h}$ are decreasing within the partitions described by $\bar{\boldsymbol{\lambda}}$.[185] For a given canonical hydrogen vector, an adapted atom vector

$$\hat{\boldsymbol{a}} = (a_0\mathrm{H}_{\mathrm{h}_0}, a_1\mathrm{H}_{\mathrm{h}_1}, ..., a_{\hat{N}-1}\mathrm{H}_{\mathrm{h}_{\hat{\mathrm{N}}-1}}) \tag{2.48}$$

can be defined, with a corresponding adapted degree vector

$$\hat{\mathbf{d}} = (d_0 - h_0, d_1 - h_1, ..., d_{\hat{N}-1} - h_{\hat{N}-1}) \,. \tag{2.49}$$

Due to the definition of canonicity for the hydrogen vector, atoms with the same label and the same number of implicit hydrogen atoms are always listed next to each other. A new partition vector

$$\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_0, \hat{\lambda}_1, \ldots, \hat{\lambda}_{\hat{K}}) \tag{2.50}$$

can be defined, which counts the number of occurrences of each group of atoms with the same label and the same number of implicit hydrogen atoms. For the newly introduced partition vectors it must hold that

$$\prod_{k=0}^{K} \lambda_k! \geq \prod_{i=0}^{\hat{K}} \hat{\lambda}_k! \tag{2.51}$$

and thus the number of permutations in $S_{\hat{\boldsymbol{\lambda}}}$ is (often considerably) smaller than the number of permutations in $S_{\boldsymbol{\lambda}}$. Here, $\hat{K}$ is the number of different atom types in the atom vector when taking into account the distribution of the hydrogen atoms.

With this, the enumeration process can be made more efficient. Instead of using the filling algorithm directly, the algorithm performs an orderly enumeration of canonical hydrogen vectors. The filling algorithm is then used to create canonical adjacency matrices for each of these hydrogen vectors using the corresponding input vectors $\hat{\boldsymbol{a}}$ and $\hat{\boldsymbol{d}}$.

The algorithm that is used to produce canonical hydrogen vectors in *enu* is adapted from the orderly enumeration algorithm in Algorithm 1, and also consists of a forward step, a backward step and a canonicity test. Analogous to the matrix $\mathbf{M}$, defined in Eq. (2.21), a vector $\mathbf{m}$ is defined as

$$\mathbf{m} = (m_0, m_1, \ldots, m_{\hat{N}-1}) \,, \qquad\qquad m_i = \min(d_i - 1, \lambda_K) \,, \tag{2.52}$$

such that each entry corresponds to the maximum number of hydrogen atoms that can be connected to the corresponding atom in the atom vector. A non-hydrogen atom must have at least one bond left to connect to another

non-hydrogen atom, and it cannot be connected to more hydrogen atoms than the total number of hydrogen atoms λ_K in the molecular formula. The algorithm is shown in Algorithm 7. Note that the case that only one non-hydrogen atom is present (*e.g.*, C_1H_4) is an exception that must be handled explicitly in the program.

Algorithm 7: Filling Algorithm for the Hydrogen Vector

```
/* start the algorithm by calling a HForwardStep at index -1 */
HForwardStep (-1)

Function HForwardStep(i):
    i ++
    if i == N̂ then
        if sum (h) == λ_K then
            /* h is completely filled, and the sum over h is equal to the number of hydrogen atoms in
               the given molecular formula */
            HCanonicityTest(h)
        else
            /* h is completely filled, but the sum over h is not equal to the number of hydrogen atoms
               in the given molecular formula, and thus h is not a viable hydrogen vector */
            HBackwardStep(i)
    else
        /* potential entry at i in h is the minimum of the maximum entry possible, given by entry i of m,
           and the number of hydrogen atoms that are not yet distributed */
        x = min(m_i, λ_K - sum (h))
        set h_i to x
        HForwardStep(i)


Function HBackwardStep(i,j):
    if i == 0 then
        /* algorithm has terminated */
        return
    i − −
    if h_i > 0 then
        /* decrease current entry of h and test if the corresponding hydrogen vector h is now canonical
           up to i, i.e., if the entries are decreasing within their partitions up to i  */
        set h_i to h_i − 1
        if h is canonical up to index i then
            HForwardStep(i)
        else
            i ++
            HBackwardStep (i)
    else
        HBackwardStep(i)
```

The following example illustrates the dramatic impact of this approach on the computational cost even for small molecules. For the molecular formula $C_3O_2H_4$, the following canonical hydrogen vectors can be constructed

$$\begin{aligned}
&\mathbf{h}_0 = (3,1,0,0,0) \qquad \mathbf{h}_1 = (3,0,0,1,0) \qquad \mathbf{h}_2 = (2,2,0,0,0)\\
&\mathbf{h}_3 = (2,1,1,0,0) \qquad \mathbf{h}_4 = (2,1,0,1,0) \qquad \mathbf{h}_5 = (2,0,0,1,1)\\
&\mathbf{h}_6 = (1,1,1,1,0) \qquad \mathbf{h}_7 = (1,1,0,1,1)\,,
\end{aligned}$$

with the corresponding atom vectors

$$\begin{aligned}
&\hat{\boldsymbol{a}}_0 = (\mathrm{CH_3},\mathrm{CH_1},\mathrm{C},\mathrm{O},\mathrm{O}) \qquad \hat{\boldsymbol{a}}_1 = (\mathrm{CH_3},\mathrm{C},\mathrm{C},\mathrm{OH_1},\mathrm{O})\\
&\hat{\boldsymbol{a}}_2 = (\mathrm{CH_2},\mathrm{CH_2},\mathrm{C},\mathrm{O},\mathrm{O}) \qquad \hat{\boldsymbol{a}}_3 = (\mathrm{CH_2},\mathrm{CH_1},\mathrm{CH_1},\mathrm{O},\mathrm{O})\\
&\hat{\boldsymbol{a}}_4 = (\mathrm{CH_2},\mathrm{CH_1},\mathrm{C},\mathrm{OH_1},\mathrm{O}) \qquad \hat{\boldsymbol{a}}_5 = (\mathrm{CH_2},\mathrm{C},\mathrm{C},\mathrm{OH_1},\mathrm{OH_1})\\
&\hat{\boldsymbol{a}}_6 = (\mathrm{CH_1},\mathrm{CH_1},\mathrm{CH_1},\mathrm{OH_1},\mathrm{O}) \qquad \hat{\boldsymbol{a}}_7 = (\mathrm{CH_1},\mathrm{CH_1},\mathrm{C},\mathrm{OH_1},\mathrm{OH_1})\,,
\end{aligned}$$

and degree vectors

$$\begin{aligned}
&\hat{\mathbf{d}}_0 = (1,3,4,2,2) \qquad \hat{\mathbf{d}}_1 = (1,4,4,1,2) \qquad \hat{\mathbf{d}}_2 = (2,2,4,2,2)\\
&\hat{\mathbf{d}}_3 = (2,3,3,2,2) \qquad \hat{\mathbf{d}}_4 = (2,3,1,1,2) \qquad \hat{\mathbf{d}}_5 = (2,4,4,1,1)\\
&\hat{\mathbf{d}}_6 = (3,3,3,1,2) \qquad \hat{\mathbf{d}}_0 = (3,3,4,1,1)\,.
\end{aligned}$$

The corresponding label vectors are

$$\begin{aligned}
&\hat{\boldsymbol{\alpha}}_0 = (\mathrm{CH_3},\mathrm{CH_1},\mathrm{C},\mathrm{O}) \qquad \hat{\boldsymbol{\alpha}}_1 = (\mathrm{CH_3},\mathrm{C},\mathrm{OH_1},\mathrm{O})\\
&\hat{\boldsymbol{\alpha}}_2 = (\mathrm{CH_2},\mathrm{C},\mathrm{O}) \qquad \hat{\boldsymbol{\alpha}}_3 = (\mathrm{CH_2},\mathrm{CH_1},\mathrm{O})\\
&\hat{\boldsymbol{\alpha}}_4 = (\mathrm{CH_2},\mathrm{CH_1},\mathrm{C},\mathrm{OH_1},\mathrm{O}) \qquad \hat{\boldsymbol{\alpha}}_5 = (\mathrm{CH_2},\mathrm{C},\mathrm{OH_1},\mathrm{OH_1})\\
&\hat{\boldsymbol{\alpha}}_6 = (\mathrm{CH_1},\mathrm{OH_1},\mathrm{O}) \qquad \hat{\boldsymbol{\alpha}}_7 = (\mathrm{CH_1},\mathrm{C},\mathrm{OH_1})\,,
\end{aligned}$$

with the corresponding partition vectors

$$\begin{aligned}
&\hat{\boldsymbol{\lambda}}_0 = (1,1,1,2) \qquad \hat{\boldsymbol{\lambda}}_1 = (1,2,1,1) \qquad \hat{\boldsymbol{\lambda}}_2 = (2,1,2)\\
&\hat{\boldsymbol{\lambda}}_3 = (1,2,2) \qquad \hat{\boldsymbol{\lambda}}_4 = (1,1,1,1,1) \qquad \hat{\boldsymbol{\lambda}}_5 = (1,2,2)\\
&\hat{\boldsymbol{\lambda}}_6 = (3,1,1) \qquad \hat{\boldsymbol{\lambda}}_7 = (2,1,2)\,.
\end{aligned}$$

The number of elements in the symmetric groups of the new partition vectors are

$$\begin{aligned}
|S_{\hat{\boldsymbol{\lambda}}_0}| &= 1! \cdot 1! \cdot 1! \cdot 2! = 2 \\
|S_{\hat{\boldsymbol{\lambda}}_1}| &= 1! \cdot 2! \cdot 1! \cdot 1! = 2 \\
|S_{\hat{\boldsymbol{\lambda}}_2}| &= 2! \cdot 1! \cdot 2! = 4 \\
|S_{\hat{\boldsymbol{\lambda}}_3}| &= 1! \cdot 2! \cdot 2! = 4 \\
|S_{\hat{\boldsymbol{\lambda}}_4}| &= 1! \cdot 1! \cdot 1! \cdot 1! \cdot 1! = 1 \\
|S_{\hat{\boldsymbol{\lambda}}_5}| &= 1! \cdot 2! \cdot 2! = 4 \\
|S_{\hat{\boldsymbol{\lambda}}_6}| &= 3! \cdot 1! \cdot 1! = 6 \\
|S_{\hat{\boldsymbol{\lambda}}_7}| &= 2! \cdot 1! \cdot 2! = 4\,,
\end{aligned} \tag{2.53}$$

which are all *considerably* smaller than the number of elements in the original symmetric group

$$|S_{\boldsymbol{\lambda}}| = 3! \cdot 2! \cdot 4! = 288\,. \tag{2.54}$$

Note that the canonical ordering of the atom vector (according to the valence, size of the partition, and atomic number) is generated for each distribution of the hydrogens prior to starting the enumeration algorithm. For example, an oxygen atom without implicit hydrogen and with a valence of two would come before a carbon atom with three implicit hydrogens and a valence of one.

The specification of implicit hydrogens from user input, as described in Sec. 2.3.2, is simply a restriction of the allowed distributions of the hydrogens. If an atom is already assigned a specific number of implicit hydrogen atoms, this number remains constant in the hydrogen vector.

2.C SMILES CANONICALIZATION

Both the algorithm proposed by Weininger [192] and the one proposed by Schneider [193] for canonicalization of SMILES strings start by assigning an initial atom ordering, considering atom *invariants*, which are based on the properties of the atoms in the molecule. The invariants used for the implementation of the

canonicalization algorithm in *enu* are the ones proposed by Weininger *et al.* In order to create the atomic invariants, the following atomic properties are used: the number n_C of connections (*i.e.*, single or multiple bonds) to non-hydrogen atoms, the number n_B of bonds to non-hydrogen atoms, the atomic number Z, the sign of charge s (0 for zero or positive, 1 for negative), the net charge c (in units of e), and the number n_H of attached hydrogens.[192] Given these properties, the initial atom invariant is created by combining them into an eight digit number

$$10^7 \cdot n_C + 10^5 \cdot n_B + 10^3 \cdot Z + 10^2 \cdot s + 10 \cdot c + n_H \,. \tag{2.55}$$

Considering the molecule shown in Figure 2.1 in Sec. 2.2.1 as an example, the initial atom invariants are provided in Table 2.2.

Table 2.2: Initial invariants. Here, the initial invariants for the atoms of the molecule shown in Figure 2.1 in Sec. 2.2.1. Further, the third and fourth rows contain the initial atom indices according to the algorithms proposed by Weininger *et al.* and Schneider *et al.*, respectively.

atoms	C_0	C_1	C_2	O_3	C_4	H_5	H_6	H_7	H_8
invariant	30406000	10106003	20406000	10208000	10108001	10101000	10101000	10101000	10101000
initial index (W)	5	1	4	3	2	0	0	0	0
initial index (S)	8	4	7	6	5	0	0	0	0

Given these initial invariants, an initial ordering of the atoms is achieved by assigning indices in order of increasing value of the invariant. On one hand, Weininger *et al.* propose to directly assign consecutive indices. On the other hand, Schneider *et al.* propose a more elaborate and stable indexing that leaves sufficient space between the indices such that when an index is reassigned, no other index needs change. This is achieved by assigning the same index i_1 to all of the n_1 atoms with the same initial invariant, and then using the next higher index $i_2 = i_1 + n_1$.[193] The two kinds of initial indices are shown in the last two rows of Table 2.2, respectively, and the one that is used in the *enu* program is the one developed by Schneider *et al.* The initial index assignment divides the atoms into different partitions, or *equivalence classes*, with all atoms that have the same index belonging to the same equivalence class.[193] Once the initial equivalence classes are created, they are continuously refined until each equivalence class consists of just one atom.

For two atoms a and a' possessing the same number of first neighbors n_C with corresponding indices $i_0 > \cdots > i_{n_C-1}$ and $i'_0 > \cdots > i'_{n_C-1}$, a lexicographical ordering can be defined as

$$a > a' \Leftrightarrow \exists j : i_j > i'_j \ \wedge \ i_k = i'_k \quad \forall\, k < j \tag{2.56}$$

and

$$a = a' \Leftrightarrow i_k = i'_k \quad \forall\, 0 \leq k \leq n_C - 1\,. \tag{2.57}$$

Due to the definition of the initial invariants, only atoms that have the same number n_C of neighbors can be in the same equivalence class. Thus, when refining an equivalence class, all the atoms in the corresponding partition are compared in a pairwise fashion and the index of the lexicographically larger atom is increased by one. If the two atoms are lexicographically equal, their respective indices are left unchanged. Thus, the partitions that are newly created due to the index reassignments can still contain more than one atom.

The algorithm proposed by Schneider *et al.* keeps a list of the partitions that need to be refined, initially containing all equivalence classes that consist of more than one atom. It proceeds by always refining the equivalence class in the list that corresponds to the highest atom index. Once the refinement is finished, the corresponding partition is erased from the list. At the same time, all the equivalence classes that containing atoms whose neighbors are affected by the index reassignment of the last refinement step are added back to the list for reevaluation if they still contain more than one atom. The algorithm then continues with the refinement of the equivalence class in the list that corresponds to the currently highest atom index.

At some point, the list of partitions that need to be refined may become empty, although there are still equivalence classes containing more than one element. For this scenario, Schneider *et al.* propose a tie-breaking step. The tiebreaking is performed in the equivalence class that corresponds to the highest atom index that still consists of more than one atom. The atom with the largest original index within that equivalence class is assigned the highest possible index within this class. After this tie-breaking step, the refinement process is started once more. The refinement process and the tie-breaking step are used

in alternation, until all equivalence classes consist of just one atom.

Note that while this algorithm produces a canonical atomic ordering in almost all cases, if the molecule is highly symmetrical, it is still possible that the final ordering is not unique but depends on the initial atom ordering. For this, Schneider *et al.* introduce two new invariants. In the *enu* program, these two additional invariants are not used since the canonical adjacency matrix provides a canonical initial atom ordering. Thus, the SMILES string is guaranteed to be canonical.

3

Tbl: A Fragment-Based Topology Builder

"The task of the theorist is to bring order into the chaos of the phenomena of nature, to invent a language by which a class of these phenomena can be described efficiently and simply. [...] Without theoretical concepts one would neither know what experiments to perform nor be able to interpret their outcome."

Walter Noll[221]

The C++ program *tbl* constitutes one of several programs underlying the CombiFF scheme to automatically parameterize force fields for condensed-phase molecular dynamics (MD) simulations. CombiFF is based on a transferability principle, stating that force-field parameters for small molecules can be "transferred" to larger molecules. Accordingly, the force-field parameters are optimized in terms of molecular fragments by alternating simulations and comparison of the calculated condensed-phased properties against experimental data. To obtain robust force-field parameters and avoid overfitting, the CombiFF scheme relies on a high observable-to-parameter ratio, *i.e.*, it requires simulations involving many different molecules. Manual topology creation would therefore present a considerable bottleneck for the scheme. To mitigate this issue, the *tbl* program was developed to assemble molecular

topologies in an automated fashion based on a predefined library of molecular fragments. The *tbl* code is freely available on GitHub at `https://github.com/csms-ethz/CombiFF`.

3.1 INTRODUCTION

Chapters 1 and 2 introduced the CombiFF force-field parameterization scheme, as well as one of the key components of this workflow, the isomer enumerator program *enu*. Like *enu*, the fragment-based topology builder *tbl* is an essential building block of the scheme. Because CombiFF relies on a high observable-to-parameter ratio, it requires to perform molecular dynamics (MD) simulations of many different molecules for both the automated parameterization as well as the subsequent validation of the force field.[73] Persistent methodological, software and hardware advances continue to increase the efficiency of MD simulations,[16,222–226] making such a high-throughput approach possible.[73] However, manual topology building is not only tedious, but also error-prone and non-robust. In recent years, several automated topology builders have been developed (often also in combination with parameterization workflows), such as Antechamber,[55,133,134] the Automated Topology Builder (ATB),[68,135] General Automated Atomic Model Parameterization (GAAMP),[136,137] LigParGen,[138] Open Force Field (SMIRNOFF and OpenFF),[75,76] ParamChem,[59–61] PRODRG,[139] R.E.D.,[140] or SwissParam.[141]

The C++[189] program *tbl*, developed for CombiFF and described in the present chapter, is a fragment-based topology builder. It allows for the automated creation of the molecular topologies necessary for performing simulations with the GROMOS MD engine[27] or the in-house SAMOS MD engine.[73] The *tbl* program relies on a library of predefined *fragments* to assemble the topology of a given molecule based on its SMILES string.[191,202] After decomposing the given molecule into a set of fragments, the corresponding force-field parameters are either determined

from the properties of a single fragment, or from the properties of two or more fragments (*via* atom, bond, bond-angle, improper-dihedral, and torsional-dihedral types).

The topology creation process consists of five successive steps:

1. Reading a list of molecules, *e.g.*, generated by *enu* (see Chapter 2), and reading a list of predefined molecular fragments.

2. For each compound, creating a decomposition of the molecule in terms of the given fragments, using substructure matching.

3. Creating a general topology file specifying all the force-field terms as string macros.

4. Creating a GROMOS molecular topology building block (mtb)[227] file for the entire molecule set with the terms still represented by string macros.

5. Replacing the string macros in the mtb file with concrete force-field parameters.

Except for the GROMOS mtb files, all input and output files follow a well-defined XML format, allowing for an easy parsing and potential integration into other workflows. The *tbl* program enables the fast and robust creation of the many topologies required for the automated CombiFF force-field parameterization scheme. The code is open-source and freely available on GitHub at `https://github.com/csms-ethz/CombiFF`. It can be conveniently compiled using *cmake*.[209]

In the following, we describe in turn the underlying principles and the implementation of *tbl*. This description is followed by a short conclusion.

3.2 UNDERLYING PRINCIPLES

3.2.1 Fragments

In *tbl*, each fragment is assigned a unique identifier, the *fragment code* (by convention starting with a '~'). A fragment is defined *via* covalently-bonded *core atoms* and *link atoms*. Within a fragment, the core atoms can be bonded to other core atoms as well as to link atoms. Each link atom, on the other hand, must be bonded to exactly one core atom by a single or, possibly, multiple bond. Note that a fragment can also consist of only core atoms, without any link atoms (in this case, it represents a separate molecule). Each atom in a fragment is assigned either a *simple atom type* or an *atom-type set*. An atom-type set consists of a list of simple atom types (as a limiting case, this list may consist of a single simple atom type). Each simple atom type corresponds to a unique chemical element, but multiple simple atom types may be associated with the same element (corresponding to different chemical environments). A core atom is always assigned a simple atom type, *i.e.*, the atom type of a core atom within a fragment is fully determined, independently of the other fragments to which this fragment may be connected. In contrast, a link atom is always assigned an atom-type set, implying that the atom type of a link atom within an assembled molecule generally depends on the connections of the fragments. Each atom in the fragment is also assigned a unique identifier. Note that this identifier is only unique within the fragment, *i.e.*, the atoms of two different fragments may have the same identifier. The atom identification within a molecule is then made unique by combining the atom identifier with a fragment identifier.

Two fragments can be *linked* (*i.e.*, connected) using so-called *bond-linking* (Figure 3.1). For linking, bonds in each of the of two fragments are overlaid onto each other, which is allowed if: (*i*) they both involve a core atom and a link atom; (*ii*) the respective bond degrees (*i.e.*, single, double, *etc.*) are identical; and (*iii*) the core atom of the first fragment

is *compatible* with the link atom of the second fragment and *vice versa*. Here, a core atom is compatible with a link atom if its simple atom type is contained in the atom-type set of the link atom. Once two fragments are linked, the atom type of the link atom of the first fragment is set to the simple atom type of the core atom of the second fragment and *vice versa*. When two fragments *A* and *B* are linked *via* the link atoms with the identifiers *a* and *b*, respectively, this is denoted as *A[a:b]B*.

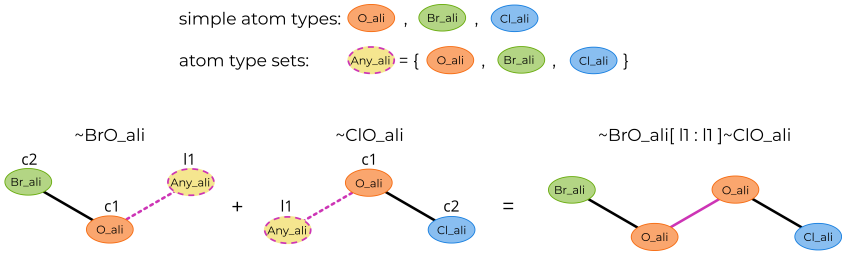


Figure 3.1: Illustrative example for the linking of two fragments. The fragment with fragment code *~BrO_ali* is linked to the fragment with fragment code *~ClO_ali*. There are three simple atom types, *O_ali*, *Br_ali*, and *Cl_ali*, as well as an atom-type set, *Any_ali*, which contains these three simple atom types. The atom identifiers are shown above the atom types. *~BrO_ali* and *~ClO_ali* can be bond-linked by overlaying the bonds between their respective *O_ali* core atom and *Any_ali* link atom. This is allowed since both the bond degrees and the atom types are compatible. Here, the core atoms (orange, green, blue) are displayed with a thin solid border and the link atoms (yellow) with a dashed purple border. The bonds between core atoms are displayed as solid black lines and the bonds between core and link atoms are displayed as dashed purple lines. The overlayed bond is also colored in purple. The atom types of the link atoms in the assembled molecule are determined by the simple atom types of the core atoms with which they are matched. In this example, if the simple atom types *O_ali*, *Br_ali*, and *Cl_ali* correspond to the chemical elements O, Br, and Cl, respectively, then the assembled molecule is BrOOCl.

For a minimal definition of a fragment, only the fragment code, a list of atoms, and a list of bonds is required. Each atom in the list is assigned a unique identifier, classified as a core atom or a link atom, and assigned a simple atom type (for a core atom) or an atom-type set (for a link atom). Each bond is defined *via* the involved pair of atoms (specified by their identifiers) and the bond degree (*i.e.*, single, double, *etc.*). By default, other properties of the fragments, namely bond angles and torsional dihedrals, are detected automatically based on the bonds.

However, this behavior can be overridden by explicitly specifying such a property. For example, to define a special angle in a fragment, a list of the three involved atoms (specified by their identifiers), and a term name (macro later mapped to parameters) are required. Note that improper dihedrals are not detected automatically and must always be specified explicitly, if required. When two fragments are linked, the assembled structure may form new bond angles or torsional dihedrals, which are also detected automatically. As an illustrative example, the force-field terms associated with the linked fragments of Figure 3.1 are listed in Table 3.1.

Table 3.1: Illustrative example for the terms in linked fragments. Here, the force-field terms (*i.e.*, atoms, bonds, bond angles, and torsional dihedrals) of the linked fragments of Figure 3.1 are provided. Note that after linking, only core atoms (*i.e.*, simple atom types) are involved in specifying these properties. The syntax used here to specify the terms is given only for illustration purposes, and is shorter than the syntax that is actually used in the output of *tbl* (see Sec. 3.3.1).

type	involved core atoms	term type
atom	~BrO_ali.c2	typ_Br_ali
atom	~ClO_ali.c2	typ_Cl_ali
atom	~BrO_ali.c1	typ_O_ali
atom	~ClO_ali.c1	typ_O_ali
bond	~BrO_ali.c1, ~BrO_ali.c2	bnd_Br_ali-O_ali
bond	~ClO_ali.c1, ~ClO_ali.c2	bnd_Cl_ali-O_ali
bond	~BrO_ali.c1, ~ClO_ali.c1	bnd_O_ali-O_ali
angle	~BrO_ali.c1, ~BrO_ali.c2, ~ClO_ali.c1	ang_Br_ali-O_ali-O_ali
angle	~BrO_ali.c1, ~ClO_ali.c1, ~ClO_ali.c2	ang_Cl_ali-O_ali-O_ali
torsional dihedral	~BrO_ali.c1, ~BrO_ali.c2, ~ClO_ali.c1, ~ClO_ali.c2	dih_Br_ali-O_ali-O_ali-Cl_ali

3.2.2 Molecule Decomposition

To assemble a topology for a given molecule based on the library of predefined fragments, it is necessary to decompose the molecule into a set of these fragments, linked together *via* bond-linking. An example of such a decomposition is shown in Figure 3.2. The decomposition is achieved in an iterative process. The fragments are considered in order of decreasing priority and matched (as many times as possible) to the parts of the molecule that are yet unmatched. This procedure requires the selection of: (*i*) a priority ordering for the fragments; and (*ii*) a substructure matching algorithm. These two features are discussed in the following sections.

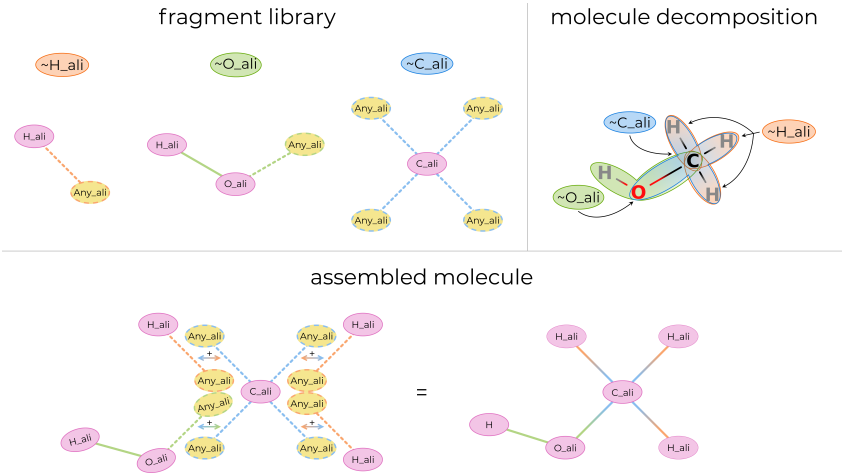


Figure 3.2: Decomposition of the molecule methanol into fragments. The methanol molecule displayed on the top right can be decomposed into the fragments of the fragment library displayed on the top left. Three fragment types are needed: a fragment with a hydrogen core atom connected to a link atom (*~H_ali*), a fragment with an oxygen core atom connected to a hydrogen core atom and a link atom (*~O_ali*), and a fragment with a carbon core atom connected to four link atoms (*~C_ali*). On the bottom, the assembly of the fragments to build the molecule is shown. Here, core atoms are displayed in pink, and link atoms in yellow. Bonds between core atoms are displayed as solid lines, and bonds between core and link atoms as dashed lines.

Fragment Priorities

If desired, the user may pass a custom list of fragment priorities to *tbl* by specifying the fragment codes in decreasing order of priority. Otherwise, the fragments are ordered by default in decreasing order of priority according to the following criteria: fragments that contain cycles, number of triple bonds, number of aromatic bonds, number of double bonds, number of single bonds, number of atoms, and number of core atoms. This sequence ensures that fragments with a lower priority cannot be assembled from fragments with a higher priority. Consider, for example, a fragment with four carbon atoms connected to form a cycle, and a linear fragment where a central carbon atom is connected to the two other carbon atoms. If a molecule contains a cycle of four carbon atoms, this cycle could be decomposed into a single occurrence of the cyclical fragment, or

four occurrences of the linear fragment. Generally, the desired behavior is to match the cyclical fragment (otherwise there would be no need to define such a fragment), reason for which it should be given a higher priority in the matching order.

Substructure Matching

In *tbl*, both the fragments and the molecules are represented by adjacency matrices (Sec. 3.3.1, and Chapter 2, Sec. 2.2.1). The adjacency matrices of both the fragments and the molecules include hydrogen atoms (if present), so that the substructure search is always performed with explicit hydrogens.

The substructure search algorithm implemented in *tbl* is based on the Ullmann algorithm,[213,214] with additional matching criteria to account for the conditions imposed by bond linking. To perform a substructure search of a fragment with n atoms in a molecule with m atoms, a boolean matching matrix $\boldsymbol{M}$ of dimension $n \times m$ is defined. An element M_{ij} of this matrix is set to true if atom i of the fragment is *compatible* with atom j of the molecule (otherwise it is set to false). To assess compatibility, two cases must be distinguished, depending on whether the currently considered fragment atom is a core atom or a link atom. If it is a core atom, it corresponds to a given simple atom type, and thus to a specific chemical element (although different simple atom types may correspond to the same element). A core atom of the fragment is compatible with any atom in the molecule that is: (*i*) of this chemical element type; (*ii*) not yet matched to another core atom; and (*iii*) either not yet matched to a link atom, or only matched to link atom(s) that are compatible with the simple atom type (*i.e.*, this type is contained in the atom-type sets of the link atoms). On the other hand, if the currently considered fragment atom is a link atom, it is associated with an atom-type set. A link atom is compatible with any atom in the molecule that is: (*i*) of a chemical element type compatible with those contained in its atom-type set; and (*ii*) either not yet matched to a core atom, or matched to a core

atom whose simple atom type is contained in the atom-type set of the considered link atom.

For a given fragment, molecule and, possibly, partial molecule decomposition (*i.e.*, some parts of the molecule being already matched to fragments), the matching matrix is generated by iterating over all pairs of atoms with one atom in the fragment and one atom in the molecule, checking the above compatibility criteria, and setting the matrix entries accordingly. After generating the matching matrix, if there is a row with all entries set to false, this indicates that at least one atom in the fragment cannot be matched to any of the atoms in the molecule, *i.e.*, the substructure search is unsuccessful for the current fragment. Otherwise, the algorithm proceeds with the matching algorithm.

The corresponding code consists of the recursive function `UllmannMatch` (Algorithm 8), which in turn relies on the refinement procedure `Refine` (Algorithm 9).[213,214] For a given fragment atom k, the matching algorithm iterates over all molecule atoms l. If the two atoms are compatible (*i.e.*, $\mathbf{M}_{kl}$ is true), they are matched and all other elements in row k and in column l of $\mathbf{M}$ are set to false, as the corresponding atoms are no longer available for other matches. Next, the algorithm refines the matching matrix based on the newly added match. For all yet unmatched fragment atoms (*i.e.*, $i > k$), it is checked whether there is a potential match for each first neighbor of the fragment atom i with at least one first neighbor of a compatible molecule atom j. Note that for the first neighbors to be compatible, the bond multiplicity between the fragment atom and its first neighbor also has to be identical to the bond multiplicity between the molecule atom and its first neighbor. If at least one first neighbor of the considered fragment atom i is not compatible with any of the first neighbors of the considered molecule atom j, the two atoms are not compatible and the corresponding element M_{ij} of $\mathbf{M}$ is reset to false. If the refinement does not lead to any unmatchable fragment atoms (*i.e.*, no rows of $\mathbf{M}$ with all entries set to false), the matching algorithm proceeds by choosing a match for the next fragment atom $k + 1$. Otherwise, the

algorithm backtracks (*i.e.*, resets **M** to the state before the current match) and tries to match fragment atom k with the next molecule atom $l + 1$. If any fragment atom k cannot successfully be matched with any of the molecule atoms, no matching can be found and the algorithm returns false. On the other hand, if there is a match for all fragment atoms (*i.e.*, index n is reached in the iteration over the fragment atoms), the algorithm returns true. After the successful termination, each row of **M** contains exactly one element set to true, corresponding to the detected substructure match.

Algorithm 8: Ullmann Match

```
/* A is the adjacency matrix of the molecule */
/* F is the adjacency matrix of the fragment */
/* M is the matching matrix */
/* n is the number of atoms in the fragment */
/* m is the number of atoms in the molecule */
/* k is the current row index */
/* the function Refine is outlined in Algorithm 9 */

/* assign elements of the matching matrix M  */
for k = 0; k < n; k ++ do                                              // loop over fragment atoms
    for l = 0; l < m; l ++ do                                          // loop over molecule atoms
        if fragment atom k and molecule atom l are compatible then
            M[ k, l ] = true              // fragment atom k and molecule atom l are a potential match
        else
            M[ k, l ] = false                  // fragment atom k and molecule atom l are not a match

k = 0
UllmannMatch(A, F, M, k, m, n)

Function UllmannMatch(A, F, M, k, m, n):
    if k == n then                                  // final row index (i.e., final fragment atom) reached
        return true                                                    // all fragment atoms are matched

    /* iterate over molecule atoms to find a match for the next fragment atom */
    for l = 0; l < m; l ++ do
        /* check if next fragment atom and current molecule atom are compatible */
        if M [ k, l ] then
            M_save = M                          // store current matching for potential backtracking

            /* match fragment atom k with molecule atom l by setting all elements of M in row k and in
               column l to false, except at position (k, l) */
            for j = 0; j < m; j++ do
                M [ k, j ] = false
            for i = 0; j < n; i++ do
                M [ i, l ] = false
            M [ k, l ] = true

            if Refine(A, F, M, k, m, n) then                // check refinement of current matching
                if UllmannMatch(A, F, M, k +1, m, n) then
                    return true                  // valid match found for all fragment atoms > k


            /* refinement not possible for current matching, or no valid match was found for a
               fragment atom > k → backtrack and try to match fragment atom k with the next
               molecule atom l + 1 */
            M = M_save
    /* no match found for fragment atom k → backtrack and try to match the previous fragment atom k - 1
       with the next molecule atom l + 1 */
    return false
```

Algorithm 9: Refinement Procedure

```
/* A is the adjacency matrix of the molecule */
/* F is the adjacency matrix of the fragment */
/* M is the matching matrix */
/* n is the number of atoms in the fragment */
/* m is the number of atoms in the molecule */
/* k is the current row index */

Function Refine(A, F, M, k, m, n):
    changed = true
    while changed do
        changed = false
        /* iterate over fragment atoms >k, i.e., rows of M  */
        for i = k +1; i < n; i ++ do
            /* iterate over molecule atoms, i.e., columns of M  */
            for j = 0; j < m; j ++ do
                if M [ i, j ] then
                    /* check if there is a potential match for each neighbor nbr_i of fragment atom
                       i with at least one of the neighbors nbr_j of the molecule atom j, i.e. nbr_i
                       and nbr_j are compatible and the bond degree between nbr_i and fragment atom
                       i is the same as the bond degree between nbr_j and molecule atom j  */
                    valid = true
                    for each nbr_i in F.GetNeighbors(i) do
                        found = false
                        for each nbr_j in A.GetNeighbors(j) do
                            if M [nbr_i, nbr_j ] and F [ i, nbr_i ] == A [ j, nbr_j ] then
                                found = true
                                break
                        if !found then
                            valid = false
                            break
                    if !valid then
                        /* fragment atom i and molecule atom j are no longer compatible due to
                           unmatchable neighbors */
                        M [ i, j ] = false
                        changed = true

                        /* check if fragment atom i can still be matched with another molecule
                           atom h  */
                        nonzero = false
                        for h = 0; h < m; h ++ do
                            if M [ i, h ] then
                                nonzero = true
                                break
                        /* reject M if fragment atom i cannot be matched */
                        if !nonzero then
                            return false
    return true
```

3.3 IMPLEMENTATION

3.3.1 Input and Output Formats

In the following sections, the formats of the input and output files of *tbl* are outlined. In general, these files follow an *Extensible Markup Language*

(XML) format, to allow for convenient parsing and compatibility with independent programs. For each type of input/output file, there is a corresponding *Document Type Definition* (DTD) specifying the format. This allows to easily validate the format of a given input/output file. Examples of input files can be found in `https://github.com/csms-ethz/CombiFF/tree/main/use/input_files`. This repository contains a subdirectory for each type of input file, including a `README` and the corresponding DTD file. When running *tbl* using the scripts provided in the GitHub repository, the output files will be generated by default in the corresponding subdirectories of `https://github.com/csms-ethz/CombiFF/tree/main/use/output_files`.

Input Files

Fragment Files As outlined in Sec. 3.2.1, a fragment is defined by a fragment code, a list of atoms, a list of bonds, and (optionally) further properties such as bond angles, torsional dihedrals, and improper dihedrals. By convention, fragment codes begin with a '~' character. Each atom in the list of atoms is defined by an identifier (unique within the fragment) and is assigned a link type (*i.e.*, whether it is a core atom or a link atom), and an atom type (*i.e.*, either a simple atom type or an atom-type set). Each bond in the list of bonds is defined by a bond degree (*i.e.*, single, aromatic, double, or triple), along with the two involved atoms (specified *via* their unique identifier). Bond angles and torsional/improper dihedrals can be specified analogously. An example is provided in Listing 3.1.

Listing 3.1: Example of a fragment file. The fragment file contains a single fragment with a central aliphatic carbon core atom and four aliphatic link atoms. This corresponds to the *~C_ali* fragment of Figure 3.2.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE fragments SYSTEM "fragments.dtd">
<fragments version="1.0">
  <!-- fragment ~C_ali: C(*)(*)(*)(*) -->
  <fragment code="~C_ali">
    <atoms>
      <atom id="c">
        <atomtype>C_ali</atomtype>
        <linktype type="core"/>
      </atom>
      <atom id="l1">
```

```
        <atomtype>Any_ali</atomtype>
        <linktype type="link"/>
      </atom>
      <atom id="12">
        <atomtype>Any_ali</atomtype>
        <linktype type="link"/>
      </atom>
      <atom id="13">
        <atomtype>Any_ali</atomtype>
        <linktype type="link"/>
      </atom>
      <atom id="14">
        <atomtype>Any_ali</atomtype>
        <linktype type="link"/>
      </atom>
    </atoms>
    <bonds>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l1</involved_atom>
        </involved_atoms>
      </bond>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l2</involved_atom>
        </involved_atoms>
      </bond>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l3</involved_atom>
        </involved_atoms>
      </bond>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l4</involved_atom>
        </involved_atoms>
      </bond>
    </bonds>
  </fragment>
</fragments>
```

Atom-Type Files The simple atom types and the atom-type sets are defined in so-called *atom-type* files. First, a list of `simpleatomtypes` is expected, where each simple atom type is defined by a unique identifier and a chemical-element symbol. Second, a list of `atomtypesets` is expected, where each atom-type set is defined by an identifier and the list of simple atom types that are contained in the set. Note that an atom-type set can also contain another atom-type set which was defined previously. This results in the simple atom types contained in the inner atom-type set

to be added to the list of simple atom types of the outer atom-type set. When reading in the atom-type sets, each simple atom type and atom type-set is assigned a unique index, starting at zero and increasing by one with each consecutive type. An example is provided in Listing 3.2.

Listing 3.2: Example of a atom-type file. The atom-type file contains aliphatic simple atom types and atom-type sets.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE atomtypes SYSTEM "atomtypes.dtd">
<atomtypes version="1.0">
  <simpleatomtypes>
    <!--
      aliphatic types
    -->
    <atomtype id="C_ali">
      <element_type>C</element_type>
    </atomtype>
    <atomtype id="Br_ali">
      <element_type>Br</element_type>
    </atomtype>
    <atomtype id="Cl_ali">
      <element_type>Cl</element_type>
    </atomtype>
    <atomtype id="F_ali">
      <element_type>F</element_type>
    </atomtype>
    <atomtype id="I_ali">
      <element_type>I</element_type>
    </atomtype>
    <atomtype id="H_ali">
      <element_type>H</element_type>
    </atomtype>
    <atomtype id="O_ali">
      <element_type>O</element_type>
    </atomtype>
    <atomtype id="N_ali">
      <element_type>N</element_type>
    </atomtype>
  </simpleatomtypes>
  <atomtypesets>
    <atomtypeset name="Halo_ali">
      <member id="Br_ali"/>
      <member id="Cl_ali"/>
      <member id="F_ali"/>
      <member id="I_ali"/>
    </atomtypeset>
    <atomtypeset name="Deg1_ali">
      <member id="Halo_ali"/>
      <member id="H_ali"/>
    </atomtypeset>
    <atomtypeset name="Any_ali">
      <member id="C_ali"/>
      <member id="Deg1_ali"/>
      <member id="O_ali"/>
      <member id="N_ali"/>
    </atomtypeset>
  </atomtypesets>
</atomtypes>
```

Replacement Files To replace the macro strings generated by *tbl* by concrete force-field parameters, so-called *replacement* files are used. A replacement file contains a list of replacement specifications, where each specification consists of a macro string and a force-field parameter. The macro types may use *Regular Expression* (RegEx)[228,229] patterns to allow the user the flexibility of being as generic or specific as they wish when defining a replacement rule. The most specific rules should be listed first in the file, followed by increasingly general rules (to avoid replacing a macro that fits a more specific rule according to a more generic one). To process the RegEx, the C++ *regex*[230] Standard Library header is used. An example is provided in Listing 3.3.

Listing 3.3: Example of a small replacement file. The file contains replacement rules for bond macro strings involving carbon and hydrogen atoms with parameters from the GROMOS-compatible 2016H66[72] parameter set.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE replacements SYSTEM "replacements.dtd">
<replacements version="1.0">
  <replacement>
    <macro>bnd_H_ali-C_ali.*</macro>
    <parameter>3</parameter>
  </replacement>
  <replacement>
    <macro>bnd_C_ali-C_ali.*</macro>
    <parameter>27</parameter>
  </replacement>
  <replacement>
    <macro>bnd_C_ali=C_ali.*</macro>
    <parameter>16</parameter>
  </replacement>
  <replacement>
    <macro>bnd_C_aro:C_aro.*</macro>
    <parameter>16</parameter>
  </replacement>
  <replacement>
    <macro>bnd_C.*C.*</macro>
    <parameter>27</parameter>
  </replacement>
</replacements>
```

Family Isomer-Enumeration File The input files specifying the molecules for which the topology should be generated are expected to follow the format of a *family isomer-enumeration* file as generated by *enu* (Listing 3.4). Upon parsing the input file, the SMILES strings defining the molecules are translated to adjacency matrices (see Chapter 4, Sec. 4.2.1).

Listing 3.4: Example of a family isomer-enumeration file as generated by *enu*. It lists straight-chain alkanes from C_1H_4 to C_5H_{12}.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE family_isomer_enumeration SYSTEM "family_isomer_enumeration.dtd">
<family_isomer_enumeration enu_version="1.0" family_code="alkanes" family_version="1.0">
  <enumeration_type type="constitutional"/>
  <number_of_isomers>8</number_of_isomers>
  <enumeration_time>0 h 0 min 0 s 0 ms</enumeration_time>
  <isomer_lists>
    <isomer_list formula="C1H4">
      <isomer isomer_id="alkanes_000000001">
        <constitutional_SMILES>C</constitutional_SMILES>
      </isomer>
    </isomer_list>
    <isomer_list formula="C2H6">
      <isomer isomer_id="alkanes_000000002">
        <constitutional_SMILES>CC</constitutional_SMILES>
      </isomer>
    </isomer_list>
    <isomer_list formula="C3H8">
      <isomer isomer_id="alkanes_000000003">
        <constitutional_SMILES>CCC</constitutional_SMILES>
      </isomer>
    </isomer_list>
    <isomer_list formula="C4H10">
      <isomer isomer_id="alkanes_000000004">
        <constitutional_SMILES>CC(C)C</constitutional_SMILES>
      </isomer>
      <isomer isomer_id="alkanes_000000005">
        <constitutional_SMILES>CCCC</constitutional_SMILES>
      </isomer>
    </isomer_list>
    <isomer_list formula="C5H12">
      <isomer isomer_id="alkanes_000000006">
        <constitutional_SMILES>CC(C)(C)C</constitutional_SMILES>
      </isomer>
      <isomer isomer_id="alkanes_000000007">
        <constitutional_SMILES>CCC(C)C</constitutional_SMILES>
      </isomer>
      <isomer isomer_id="alkanes_000000008">
        <constitutional_SMILES>CCCCC</constitutional_SMILES>
      </isomer>
    </isomer_list>
  </isomer_lists>
</family_isomer_enumeration>
```

Output Files

The creation of molecular topologies for a given list of molecules by *tbl* is executed in three successive steps: (*i*) the creation of a molecule decomposition into fragments; (*ii*) the creation of a molecular topology with string macros (instead of explicit force-field parameters); and (*iii*) the creation of a GROMOS mtb[227] file. These steps and the corresponding file formats are outlined in the following sections.

Molecule Decomposition Files After a given molecule is read from input (in the form of a SMILES string) and translated to an adjacency matrix, *tbl* tries to create a decomposition of the molecule into fragments. The program iterates over the given fragment types in decreasing order of priority (Sec. 3.2.2). For each fragment type, the modified Ullmann substructure search is repeated as long as a new match is found. Once no new match is found, the program moves on to the next fragment type.

If a complete decomposition is found, *i.e.*, every molecule atom is matched to exactly one core atom, the corresponding decomposition is added to the current *molecule decomposition* file. A molecule decomposition is defined by the isomer identifier (as specified in the family isomer-enumeration input file), the corresponding SMILES string, as well as a list of fragments and bond links. Each fragment is assigned an identifier (unique within the molecule), and the fragment code is recorded. Each linkage is defined by the fragment identifiers and the *linksites* of the two involved fragments. Here, the linksite corresponds to the identifier of the link atom, as defined in the fragment. An example of a molecule decomposition file containing a single molecule decomposition is provided in Listing 3.5.

Listing 3.5: Example of a molecule decomposition into fragments. In this decomposition, united-atom fragments are used (see Sec. 3.3.2). A propane molecule is assembled from two *~CH3* fragments and one *~CH2* fragment.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE molecule_decompositions SYSTEM "molecule_decompositions.dtd">
<molecule_decompositions enu_version="1.0" family_code="intro" family_version="1.0" tbl_version="1.0">
  <molecule_decomposition isomer_id="alkanes_000000003" smiles="CCC">
    <fragments>
      <fragment id="f1">~CH3</fragment>
      <fragment id="f2">~CH3</fragment>
      <fragment id="f3">~CH2</fragment>
    </fragments>
    <linkages>
      <linkage>
        <involved_fragment>
          <fragment_id>f1</fragment_id>
          <linksite>1</linksite>
        </involved_fragment>
        <involved_fragment>
          <fragment_id>f3</fragment_id>
          <linksite>11</linksite>
        </involved_fragment>
      </linkage>
      <linkage>
```

```
      <involved_fragment>
        <fragment_id>f2</fragment_id>
        <linksite>1</linksite>
      </involved_fragment>
      <involved_fragment>
        <fragment_id>f3</fragment_id>
        <linksite>12</linksite>
      </involved_fragment>
    </linkage>
  </linkages>
 </molecule_decomposition>
</molecule_decompositions>
```

Molecules with Macros Files From the generated (or a user provided) molecule decomposition file, a general topology file is created, with string macros (instead of force-field parameters), *i.e.*, a so-called molecule with macros file. For each molecule, the topological properties are listed, as defined by the fragments, and macro strings are assigned (as placeholders for the future parameters). These strings are defined with redundancy in mind, such that the user can be as general or as specific as they wish when later assigning the actual force-field parameters. Currently, the defined topological properties are atoms, bonds, bond angles, torsional dihedrals, and improper dihedrals. For each property, the number of atoms involved is specified. For each atom, an identifier (unique within the molecule) is created by combining the chemical element and the atom index in the molecule. The atom type macro for a given atom is set to "`<simple atom type>/<fragment code 1>.<linksite 1>,<fragment code 2>.<linksite 2>,...`", where the first part is defined by the unique core atom matched to the current molecule atom, and the second part (*i.e.*, after the '`/`') lists the fragments and linksites of the matched link atoms. For each atom, a list of excluded atoms is also given. By default, these are first and second neighbors (determined *via* bonds and bond angles, respectively). Additionally, the user can choose to also exclude third neighbors (determined *via* torsional dihedrals). For each bond, a macro string is created following the pattern "`bnd_<simple atom type 1><bond><simple atom type 2>/<fragment code 1><linksite 1>,<fragment code 2><linksite`

`2>,...%<fragment code i><linksite i>,<fragment code i+1, linksite i+1>, ...`". Here, `<bond>` can be one of the characters '`-`' (single), '`:`' (aromatic), '`=`' (double), or '`#`' (triple). `<simple atom type 1>` is the simple atom type of the core atom matched to the first molecule atom forming the bond and `<simple atom type 2>` is the simple atom type of the core atom matched to the second molecule atom forming the bond. The second part of the string (*i.e.*, after '`/`') first lists the fragments and linksites of the link atoms matched to the first atom of the bond and then (after the '`%`') the fragments and linksites of the link atoms matched to the second atom in the bond. Finally, the atoms forming the bond are specified by listing the corresponding unique atom identifiers. Bond angles, torsional dihedrals, and improper dihedrals are specified analogously.

Each list of topological properties is sorted according to the indices of the simple atom types (*i.e.*, the order in which they appear in the atom-type file). Note that by default, all possible torsional dihedrals are generated for a given central bond. However, the user can specify that only one torsional dihedral should be reported instead. In this case, only the torsional dihedral whose matched simple atom types have the smallest indices are kept. For example, if there are two torsional dihedrals $x-b-c-d$ and $y-b-c-d$ (where x, y, b, c, and d are simple atom types), and simple atom type x is listed before simple atom type y in the atom-type file, only the first dihedral is reported. An example of a molecule with macros file for a single molecule is provided in Listing 3.6.

Listing 3.6: Example of a molecule with macros file. The file contains the molecules with macros specification for the molecule decomposition provided in Listing 3.5.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE molecules_with_macros SYSTEM "molecules_with_macros.dtd">
<molecules_with_macros enu_version="1.0" family_code="intro" family_version="1.0" tbl_version="1.0">
  <molecule_with_macros isomer_id="alkanes_000000003" smiles="CCC">
    <atoms number="3">
      <atom atom_id="C1" atom_type="CH2_ali/-CH3.1,-CH3.1">
        <excluded_atoms>
          <excluded_atom>C2</excluded_atom>
          <excluded_atom>C3</excluded_atom>
        </excluded_atoms>
      </atom>
      <atom atom_id="C2" atom_type="CH3_ali/-CH2.11">
```

```
      <excluded_atoms>
        <excluded_atom>C1</excluded_atom>
        <excluded_atom>C3</excluded_atom>
      </excluded_atoms>
    </atom>
    <atom atom_id="C3" atom_type="CH3_ali/~CH2.l2">
      <excluded_atoms>
        <excluded_atom>C1</excluded_atom>
        <excluded_atom>C2</excluded_atom>
      </excluded_atoms>
    </atom>
  </atoms>
  <bonds number="2">
    <bond parameter="bnd_CH2_ali-CH3_ali/~CH3.l,~CH3.l%~CH2.l1">
      <involved_atoms>
        <involved_atom>C1</involved_atom>
        <involved_atom>C2</involved_atom>
      </involved_atoms>
    </bond>
    <bond parameter="bnd_CH2_ali-CH3_ali/~CH3.l,~CH3.l%~CH2.l2">
      <involved_atoms>
        <involved_atom>C1</involved_atom>
        <involved_atom>C3</involved_atom>
      </involved_atoms>
    </bond>
  </bonds>
  <angles number="1">
    <angle parameter="ang_CH3_ali-CH2_ali-CH3_ali/~CH2.l1%~CH3.l,~CH3.l%~CH2.l2">
      <involved_atoms>
        <involved_atom>C2</involved_atom>
        <involved_atom>C1</involved_atom>
        <involved_atom>C3</involved_atom>
      </involved_atoms>
    </angle>
  </angles>
 </molecule_with_macros>
</molecules_with_macros>
```

MTB files To run simulations with the GROMOS MD engine,[231] molecular topologies are specified using so-called *interaction function parameter* (ifp) and *molecular topology building block* (mtb) files.[227] The final step of *tbl* is the creation of an mtb file from a molecule with macros file. The mtb file relies on a block format to specify, *e.g.*, the atoms or bonds of a molecule. In addition to the information that is already contained in the molecules with macros file, each atom is assigned an integer atom code, a mass, a charge, and an atomic charge group code.[227] For the integer atom code, the mass and the charge, a string macro is created by combining the prefix `typ_`, `mass_`, or `chg_`, respectively, with the `atom_type` parameter of the atom. The default behavior is to assign each atom to its own charge group. This behavior could be

extended by adding charge group information to the input fragments (see Sec. 3.3.3). Note that two mtb files are created. The first one contains string macros, whereas the second one contains force-field parameters (assigned according to the provided replacement file(s)). Example of a `MTBUILDBLSOLUTE` block within an mtb file generated by *tbl* are provided in Listing 3.7 (with string macros) and in Listing 3.8 (with force-field parameters). For the future development of CombiFF, the goal is to directly pass the mtb file containing string macros to the MD engine, and let it handle the assignment (and, possibly, optimization within CombiFF) of the corresponding force-field parameters.

Listing 3.7: Example of an `MTBUILDBLSOLUTE` block with string macros. The block was created based on the molecule with macros file provided in Listing 3.6. The corresponding `MTBUILDBLSOLUTE` block with concrete force-field parameters is provided in Listing 3.8.

```
MTBUILDBLSOLUTE
# smiles: CCC
# RNME
alkanes_000000003
# number of atoms, number of preceding exclusion
# NMAT,NLIN
    3    0
# atoms
#ATOM ANM  IACM MASS        CGM ICGM MAE MSAE
    1 C1 typ_CH2_ali/~CH3.l,~CH3.l mass_CH2_ali/~CH3.l,~CH3.l chg_CH2_ali/~CH3.l,~CH3.l 1 2 2 3
    2 C2 typ_CH3_ali/~CH2.l1 mass_CH3_ali/~CH2.l1 chg_CH3_ali/~CH2.l1 1 1 3
    3 C3 typ_CH3_ali/~CH2.l2 mass_CH3_ali/~CH2.l2 chg_CH3_ali/~CH2.l2 1 0
#  NB
    2
#  IB   JB  MCB
    1    2    bnd_CH2_ali~CH3_ali/~CH3.l,~CH3.l%~CH2.l1
    1    3    bnd_CH2_ali~CH3_ali/~CH3.l,~CH3.l%~CH2.l2
# NBA
    1
#  IB   JB   KB  MCB
    2    1    3    ang_CH3_ali~CH2_ali~CH3_ali/~CH2.l1%~CH3.l,~CH3.l%~CH2.l2
#NIDA
    0
# NDA
    0
# NEX
    0
END
```

Listing 3.8: Example of an `MTBUILDBLSOLUTE` block with force-field parameters from the GROMOS-compatible 2016H66[72] parameter set. The block was created based on the molecule with macros file provided in Listing 3.6. The corresponding `MTBUILDBLSOLUTE` block with string macros is provided in Listing 3.7.

```
MTBUILDBLSOLUTE
# smiles: CCC
# RNME
alkanes_000000003
# number of atoms, number of preceding exclusion
# NMAT,NLIN
    3    0
# atoms
#ATOM ANM  IACM MASS        CGM ICGM MAE MSAE
    1 C1     15    4    0.00000 1    2   2 3
    2 C2     16    5    0.00000 1    1   3
    3 C3     16    5    0.00000 1    0
#  NB
    2
#  IB   JB  MCB
    1    2   27
    1    3   27
# NBA
    1
#  IB   JB   KB  MCB
    2    1    3   15
#NIDA
    0
# NDA
    0
# NEX
    0
END
```

3.3.2 All-Atom and United-Atom Types

For computational efficiency, aliphatic CH_n groups may be treated as *united atoms*, a choice that is still generally made for the GROMOS (and GROMOS-compatible) force fields.[33,72–74,232] In *tbl*, the user can choose whether the mtb file should be created for a united- or an all-atom force field. During the molecule decomposition into fragments (*i.e.*, substructure search), hydrogen atoms are always treated explicitly and thus, must be explicitly present in the definitions of fragments. However, if a united-atom representation is desired, all the atoms that were matched with a core atom that contains the keyword "`united`" in their `atom_type` are set to become implicit. These atoms are removed during the generation of the molecules with macros file. All topological properties involving a combination of regular core atoms and implicit core atoms are removed as

well. An example corresponding to a fragment file that contains a single united-atom CH_1 fragment is provided in Listing 3.9. Note the presence of an improper dihedral between the central carbon atom and the three link atoms.

Listing 3.9: Example of a united-atom CH_1 fragment.

```
<?xml version='1.0' encoding='UTF-8'?>
<!DOCTYPE fragments SYSTEM "fragments.dtd">
<fragments version="1.0">
  <!-- fragment CH1: C(H)(*)(*)(*) -->
  <fragment code="-CH1">
    <atoms>
      <atom id="c">
        <atomtype>CH1_ali</atomtype>
        <linktype type="core"/>
      </atom>
      <atom id="h">
        <atomtype>H_ali_united</atomtype>
        <linktype type="core"/>
      </atom>
      <atom id="l1">
        <atomtype>any</atomtype>
        <linktype type="link"/>
      </atom>
      <atom id="l2">
        <atomtype>any</atomtype>
        <linktype type="link"/>
      </atom>
      <atom id="l3">
        <atomtype>any</atomtype>
        <linktype type="link"/>
      </atom>
    </atoms>
    <bonds>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>h</involved_atom>
        </involved_atoms>
      </bond>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l1</involved_atom>
        </involved_atoms>
      </bond>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l2</involved_atom>
        </involved_atoms>
      </bond>
      <bond degree="single">
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l3</involved_atom>
        </involved_atoms>
      </bond>
    </bonds>
```

```
    <improper_dihedrals>
      <improper_dihedral>
        <involved_atoms>
          <involved_atom>c</involved_atom>
          <involved_atom>l1</involved_atom>
          <involved_atom>l2</involved_atom>
          <involved_atom>l3</involved_atom>
        </involved_atoms>
      </improper_dihedral>
    </improper_dihedrals>
  </fragment>
</fragments>
```

3.3.3 Future Developments

Currently, only bonds, bond angles, and torsional/improper dihedrals are recognized as topological properties by *tbl*. However, the underlying C++ code is written such that it can easily be extended. It contains a `TopologicalPropertyBase` class that is defined by a property name, the number of involved atoms, the list of involved atoms, and an unordered map of attribute names and attribute values (*e.g.*, for a bond instance, the unordered map contains the key "`parameter`" with the value of the string macro defining the bond term (see Listing 3.6)). The existing topological properties are then derived from this base class as instances of the template `TopologicalProperty` class, where the template parameter is chosen from the `enum` type `PropertyType` (*e.g.*, `bond` or `angle`). To give an example, the implementation of the constructor for the `improper_dihedral` property is provided in Listing 3.10.

Listing 3.10: Constructor for the `improper_dihedral` class.

```
template <>
TopologicalProperty<improper_dihedral>::TopologicalProperty() {
  type = improper_dihedral;
  property_name = "improper_dihedral";
  property_abbreviation = "imp";
  num_involved_atoms = 4;
}
```

In addition to getter and setter functions, the `TopologicalPropertyBase` class possesses a member function `SortInvolvedAtoms`. By default, atoms are sorted by their indices in the molecule. However, the default behavior can be overridden

such that, *e.g.*, for a bond angle formed by the atoms with indices $i - j - k$, the order i, j, k is used, where $i < k$. To give an example, the implementation of the `SortInvolvedAtoms` function for bond angles is provided in Listing 3.11.

Listing 3.11: Sorting function for the atoms in an angle.

```
template <>
void TopologicalProperty<angle>::SortInvolvedAtoms() {
  if (involved_atoms.front() > involved_atoms.back())
    std::swap(involved_atoms.front(), involved_atoms.back());
}
```

Thanks to this flexible implementation, new properties can easily be added. For example, a charge group index could be added as a `TopologicalProperty` with a single involved atom. Additionally, owing to the multi-step process of *tbl*, it is also possible to implement the creation of topology files for other MD engines based on a molecule with macros file.

3.4 CONCLUSION

The *tbl* program is a newly developed C++ program that automatically assembles the molecular topologies needed for MD simulations based on the corresponding SMILES strings, along with a library of predefined fragments. These fragments are made of core atoms and link atoms. The simple atom type assigned to a core atom corresponds to a unique chemical element, whereas the atom-type set assigned to a link atom may correspond to several different chemical elements. Two fragments can be connected *via* bond-linking if they involve compatible core-link bonds, which are then overlayed. The decomposition of a molecule into a set of fragments is achieved based on the Ullmann substructure search algorithm, using an extended list of matching criteria to account for the requirements imposed by bond-linking. The force-field parameters for the molecule are then determined either based on the properties of a single fragment or on the properties of a combination of two or more fragments.

There are two intermediate files created by the execution of *tbl*, *i.e.*, a molecule decomposition file and a molecule with macros file. Finally, two GROMOS mtb files are generated, one still containing string macros and one containing concrete force-field parameters. This provides the user with a lot of flexibility to interfere in the process and include specific requirements. The code itself was written with flexibility and extensibility in mind.

The *tbl* program allows the fast and robust creation of the many topologies required for the force-field parameterization (and validation) with the CombiFF scheme. As such it is an essential building block of the scheme. The source code can be compiled with *cmake* and is freely available in the CombiFF GitHub repository at `https://github.com/csms-ethz/CombiFF`.

4 Cnv: Canonicalizer and Converter

"Ordnung ist die primitivste Form von Chaos."
"Order is the most primitive form of chaos."

Hans-Jürgen Quadbeck-Seeger[233]

The *cnv* program is a standalone C++ utility program, complementary to the CombiFF programs *enu* and *tbl*, that can be used as a command-line tool or integrated within more complex scripts. It allows the user to canonicalize molecular formulas, SMILES strings, and adjacency matrices in accordance with the conventions used in CombiFF. This canonicalization is useful in particular when searching chemical databases for experimental (and calculated) reference data *via* canonical SMILES strings. Further, *cnv* can be used to rapidly calculate molecular properties such as the mass, the total number of bonds, or the number of cycles of a given molecule. The *cnv* code is freely available on GitHub at `https://github.com/csms-ethz/CombiFF`.

4.1 INTRODUCTION

In Chapters 2 and 3, the programs *enu* and *tbl* were outlined. They represent two key components of the fragment-based *CombiFF* force-field

parameterization scheme (Chapter 1, Sec. 1.2.2).[73,74] In CombiFF, force-field parameters are optimized against experimental condensed-phase reference values. The scheme involves four main steps. First, relevant molecules are enumerated using *enu*. Second, the molecular topologies necessary for performing simulations are generated based on a library of predefined fragments using *tbl*. Third, the home-maintained database *DBS* is searched for experimental reference values for the enumerated molecules and the relevant target properties. Finally, the force-field parameters are optimized in an iterative procedure by comparing the calculated properties (at present, densities and vaporization enthalpies) to the experimental reference data.[73]

As detailed in Chapter 2, the molecules enumerated with *enu* are reported as canonical SMILES strings.[191,202] SMILES strings encode the topological (and, possibly, stereochemical) information of a molecule in a human-readable format, while also being easy to store and manipulate computationally. In general, SMILES strings are not a unique representation of a molecule, *i.e.*, for all but the simplest molecules, there exist many different strings to describe the same molecule. However, there are several approaches to canonicalize SMILES strings so as to generate a unique descriptor for a given molecule.[192,193,203–206] The SMILES canonicalization algorithm implemented in *enu* is described in detail in Chapter 2, Appendix Sec. 2.C. To search an experimental database such as *DBS* for reference data *via* SMILES strings, the same canonicalization algorithm has to be employed for both the input SMILES strings as well as the SMILES strings listed in the database. The *cnv* program was originally developed for this purpose, *i.e.*, to canonicalize SMILES strings according to the conventions used in *enu*.

The *cnv* program can also be used to canonicalize molecular formulas (sum formulas), or to generate the canonical molecular formula of a given molecule provided as a SMILES string or adjacency matrix. In addition, the following properties can be calculated for a given molecule: its mass (in atomic mass unit u), the number of unsaturations, the total number

of bonds, the number of single bonds, the number of multiple bonds, the number of double bonds, the number of triple bonds, and the number of cycles. SMILES strings can be converted to the corresponding canonical adjacency matrix (see Chapter 2, Appendix Sec. 2.B.3) and *vice versa*. Finally, provided a SMILES string or adjacency matrix, the stack of the corresponding canonical adjacency matrix (*i.e.*, the upper triangular matrix, stored as a vector) as well as the canonical atom vector (see Chapter 2, Appendix Sec. 2.B.3) can be generated.

The program *cnv* is written in C++[189] and can be compiled with *cmake*.[209] The code is open source and freely available in the CombiFF GitHub repository at `https://github.com/csms-ethz/CombiFF`.

4.2 IMPLEMENTATION

4.2.1 Usage

There are three types of input formats that can be handled by *cnv*: molecular formulas (`frm`), SMILES strings (`smi`), and adjacency matrices (`mat`). The input can be passed either directly *via* command line arguments, or provided in an input file. The user can choose either a single output option or multiple ones. As for the input, the output can either be written directly to the terminal, or redirected to an output file. An overview of the possible output options is provided in Table 4.1.

The specification for the input/output file can be provided in the command line after a `-i` and/or `-o` argument (both optional), respectively. For example, calling *cnv* with the argument `-i smiles.txt` tells the program to read the input from the file *smiles.txt*. Multiple input files can be provided separated by commas or, if the list of files is surrounded by single quotes, by spaces. Further, the user can specify the type of input/output after the `-I` and `-O` arguments (both optional), respectively, using the keywords listed in Table 4.1. The default input type is `smi`,

Table 4.1: Output options for *cnv*. An overview of the keywords and carried out operations for the various *cnv* output options, as well as the compatible input options.

keyword	functionality	compatible input options
smi	print the canonical SMILES string	smi, mat
frm	print the canonical molecular formula	frm, smi, mat
mass	print the molecular mass in u	frm, smi, mat
num_sat	print the number of unsaturations	frm, smi, mat
num_bnd	print the number of total bonds	smi, mat
num_sin	print the number of single bonds	smi, mat
num_mul	print the number of multiple bonds	smi, mat
num_dbl	print the number of double bonds	smi, mat
num_arm	print the number of aromatic bonds	smi, mat
num_tri	print the number of triple bonds	smi, mat
num_cyc	print the number of cycles	smi, mat
stack	print the stack of the canonical adjacency matrix	smi, mat
atmV	print the atom vector of the canonical adjacency matrix	smi, mat
mat	print the canonical adjacency matrix	smi, mat

and the default output type is also `smi`, except if the input type is `frm`, then the default output type becomes `frm`. Multiple output arguments can be specified in the same manner as for the input files. For example, calling *cnv* with the arguments `-i list.txt -I smi -O 'smi frm mass' -o out.txt` tells the program to read in SMILES strings from the file *smiles.txt*, generate the corresponding canonical SMILES strings, molecular formulas, and molecular masses, and then write them to the output file *out.txt*

The following sections give an overview how the different input formats are processed and how the different output formats are generated.

Input

Molecular Formulas The parsing of a molecular formula (sum formula) is straightforward. For a given string, chemical element symbols and numbers are read in alternation. For each pair of an element symbol and a number, the count of the corresponding atom type is added to an atom vector. If no number is found after an element symbol, the corresponding count is one. Note that also molecular formulas containing parentheses (*e.g.* `CO(CH3)2`) are handled. In this case, the atom types of

the "subformula" within the parentheses are added to the atom vector as many times as the number after the closing parenthesis (or once, if there is no number after the closing parenthesis).

Multiple formulas can be passed *via* the command line by separating the individual strings with commas or spaces, *e.g.*, as `-I frm C10H5O2Br3 H4Br3O2C10` or `-I frm C10H5O2Br3,H4Br3O2C10`. If the input is read from a file, the formulas can be separated by a comma or any number of spaces, tabs, or newlines. For example, for the arguments `-I frm -i formulas.txt`, the file *formulas.txt* could contain, *e.g.*,

```
C10H5O2Br3,H4Br3O2C10
```

or

```
C10H5O2Br3
H4Br3O2C10
```

SMILES strings Upon input, a given SMILES string is converted to an adjacency matrix (see Chapter 2, Sec. 2.2.1). The processing of the string is divided into two main steps: (*i*) gathering basic information about the molecule, such as the number and types of atoms, in order to generate the atom vector; and (*ii*) encoding the connectivity between the atoms by filling the adjacency matrix. Note that aromatic bonds are converted to alternating single and double bonds, so as to adhere to the conventions of *enu.*

In several shells, such as the Bash shell[234] and the C-shell,[235] parentheses and brackets are reserved as special characters. In order to pass SMILES strings that contain such characters (*e.g.*, parentheses to indicate branching in the molecule) *via* the command line, the strings need to be enclosed by single quotes. For example, trying to use the argument `-I smi CC(C)C` will produce the following error in a Bash shell: `bash: syntax error near unexpected token `('`. On the other hand, the argument `-I smi 'CC(C)C'` will work fine.

When using the command line, multiple strings can be separated by a period or by spaces. When reading from an input file, multiple strings

can be separated by any number of spaces, tabs, or newlines. Note that a period character could also be used. The period is a part of the SMILES syntax, indicating disconnected compounds,[202] and the canonicalized SMILES strings for the individual molecules will also be separated by a period in the output. For example, the molecules 5-ethyl-2-methylheptane and 3-ethyl-5-methylheptane can be passed to *cnv* as SMILES strings *via* the command line, *e.g.*, as `-I smi 'CCC(CC)CCC(C)C.CCC(C)CC(CC)CC'` or `-I smi 'CCC(CC)CCC(C)C CCC(C)CC(CC)CC'`. Alternatively, an input file can be specified as `-i smiles.txt -I smi`, where the file *smiles.txt* contains for example

```
CCC(CC)CCC(C)C
CCC(C)CC(CC)CC
```

or

```
CCC(CC)CCC(C)C CCC(C)CC(CC)CC
```

or

```
CCC(CC)CCC(C)C.CCC(C)CC(CC)CC
```

In the last case, the canonical SMILES strings will also be separated by a period.

Adjacency Matrices Similarly to the molecular formulas and the SMILES strings, adjacency matrices can be passed to *cnv* either *via* the command line or from an input file. The expected format for the command line consists of a list of the atom types, separated by spaces, followed by the adjacency matrix elements listed row-by-row, also separated by spaces. For example, to pass an adjacency matrix describing a water molecule,

$$\begin{array}{ccc} \mathrm{O} & \mathrm{H} & \mathrm{H} \\ \end{array} \\ \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{array}{c} \mathrm{O} \\ \mathrm{H} \\ \mathrm{H} \end{array} , \tag{4.1}$$

via the command line, the expected syntax is `-I mat O H H 0 1 1 1 0 0 1 0 0`. Several adjacency matrices can be specified by listing them one after the other, separated by spaces. On the other hand, if the adjacency matrix (or matrices) is (are) to be read in from a file, the expected format is as follows. The atom types are listed one one line, separated by whitespaces, followed by one line for each row of the adjacency matrix, where each line contains the matrix elements for the corresponding row, separated by whitespaces. Several adjacency matrices can be specified by listing them consecutively, separated by a line break. For example, if the adjacency matrices for a water molecule and a methane molecule are to be read from a file, the expected syntax is `-I mat -i matrix.txt`, where *matrix.txt* contains the following:

```
O H H
0 1 1
1 0 0
1 0 0
C H H H H
0 1 1 1 1
1 0 0 0 0
1 0 0 0 0
1 0 0 0 0
```

Output

Canonical Adjacency Matrix The canonical adjacency matrix can be generated for a given SMILES string or a given adjacency matrix. For a SMILES string, the input is first converted to a (generally non-canonical) adjacency matrix as described in the previous section. Once the adjacency matrix is available, the hydrogen atoms (if present) are distributed among the heavy atoms, and the atom vector is canonicalized according to the rules specified in Chapter 2, Appendix Sec. 2.B.3. Next, the adjacency matrix is canonicalized by traversing the permutations $\pi \in S_{\boldsymbol{\lambda}}$ of the symmetric group (see Chapter 2, Appendix Sec. 2.B.2). Whenever a lexicographically larger adjacency matrix is encountered upon application of a permutation, the current matrix is replaced by the larger one. Similarly to the canonicity test implemented in *enu*, not all permutations have to be considered, as the permutation tree is pruned

analogously (see Chapter 2, Appendix Sec. 2.B.6). Nevertheless, for large molecules, the canonicalization process can still involve checking many permutations. To avoid that the program possibly hangs during the canonicalization of a molecule, an iteration limit is specified for the number of permutations to be considered (by default, 100 000). Whenever this limit is reached, the current (partially canonicalized) adjacency matrix is returned, a warning is printed to the standard error, and the next molecule is processed. The user can adapt this limit as desired.

Canonical SMILES String For a given input SMILES string or adjacency matrix, the canonical SMILES string is generated based on the canonical adjacency matrix, *i.e.*, once the adjacency matrix is canonicalized, the canonical SMILES string is generated. The SMILES canonicalization algorithm is based on the procedure of Weininger *et. al.*[192] and Schneider *et. al.*,[193] and is explained in detail in Chapter 2, Appendix Sec. 2.C. The two programs *enu* and *cnv* share a common library to generate the canonical SMILES string for a given adjacency matrices (namely, the `SmilesGenerator` class). This ensures that the resulting SMILES string will be identical irrespective of the used program. However, if the iteration limit for the adjacency matrix canonicalization (see above) is reached, the resulting SMILES may not be canonical if the given molecule is highly symmetrical.[193,206]

Canonical Atom Vector As described in Chapter 2, Appendix Sec. 2.B.3, a canonical atom ordering is defined in *enu*. It is obtained during the first step of the adjacency matrix canonicalization when a SMILES string or adjacency matrix is read in. As for the canonical SMILES generation, *enu* and *cnv* share a common `Atom` class that ensures that the same conventions are used (and avoids code duplication).

Canonical Adjacency Matrix Stack Once an adjacency matrix is canonicalized, obtaining the corresponding adjacency matrix stack is

straightforward. The stack contains the elements of the upper triangular adjacency matrix, listed row-wise in a single vector. For example, for the adjacency matrix

$$\begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}, \tag{4.2}$$

the corresponding stack is 210010.

Canonical Molecular Formula The canonical chemical formula (sum formula) can be generated from a given molecular formula, a SMILES string, or an adjacency matrix. For all three input types, an atom vector is generated. To canonicalize the molecular formula, the atom vector is sorted according to the Hill system, *i.e.*, first carbon, then hydrogen, followed by the rest of the atoms, sorted alphabetically by the respective element symbol.[236] If no carbon atom is present, all atoms are listed in alphabetical order (*i.e.*, also hydrogen).[236] The list of recognized atom types can easily be adapted by the user by modifying the `element_property_map` shared by *enu* and *cnv*. Next, the number of each atom type in the atom vector is counted. Finally, the canonical formula is assembled by listing each atom type and its count, one after the other. Note that if the count of an atom type is one, no number is written (*e.g.*, `H2O` instead of `H2O1`).

Molecular Mass In the `element_property_map`, each atom type is assigned an atomic mass. As for the canonical molecular formula, for all three input types, an atom vector is assembled and sorted. The molecular mass is then obtained by summing up the individual masses of all the atoms in the formula.

Number of Unsaturations For a given molecular formula or molecule, the degree of unsaturation is calculated as[219]

$$d_{\text{unsat}} = 1 + \frac{1}{2}\left(\sum_{k=0}^{K} \lambda_k(\delta_k - 2)\right) , \tag{4.3}$$

where λ_k is the number of times the atom type α_k with valence δ_k occurs in the formula. As a simple example, acethylene (C_2H_2) has $d_{\text{unsat}} = 1 + 0.5 \cdot (2 \cdot (4-2) + 2 \cdot (1-2)) = 2$ unsaturations.

Bonds The following number of bonds can be calculated by *cnv*: the total number of bonds, the number of single bonds, the total number of multiple (*i.e.*, double and triple) bonds, the number of double bonds, and the number of triple bonds. Each of these properties is calculated by considering the elements of the upper triangular adjacency matrix.

Number of Cycles The number of cycles in a given molecule is calculated as[219]

$$n_{\text{ring}} = d_{\text{unsat}} - n_{\text{db}} - 2 \cdot n_{\text{tb}} , \tag{4.4}$$

where d_{unsat} is the degree of unsaturations (Eq. (4.3)), n_{db} is the number of double bonds, n_{tb} is the number of triple bonds, and n_{ring} is the number of cycles. As described above, the number of double and triple bonds can be easily deduced from the adjacency matrix.

4.3 CONCLUSION

The C++ program *cnv* is a tool complementary to the CombiFF programs *enu* and *tbl*. It relies on shared libraries to avoid code duplication and ensure that the applied conventions are consistent throughout the whole CombiFF scheme. The *cnv* utility allows the user to canonicalize molecular

formulas, SMILES strings, and adjacency matrices, as well as calculate properties such as the total number of bonds or the number of cycles in a given molecule. The canonicalization is especially useful when searching chemical databases for experimental (and calculated) reference values *via* canonical SMILES strings. The source code can be compiled with *cmake* and is open source and freely available in the CombiFF GitHub repository at `https://github.com/csms-ethz/CombiFF`.

5

Relative Hydration Free-Energy Calculations With RE-EDS Employing Distance Restraints From RestraintMaker*

"Different constraints are decisive for different situations, but the most fundamental constraint is limited time."

Gary S. Becker[237]

There exist three main choices for the representation of the end-state coordinates in pairwise and multistate free-energy calculations: (*i*) the "single-topology" approach, in which the coordinate space is formed by the union of all coordinates; (*ii*) the "dual-topology" approach, in which all coordinates are explicitly present; and (*iii*) the intermediate "hybrid-topology" approach. The single- and hybrid-topology approach have in principle a higher sampling efficiency compared to the dual-topology approach due to the smaller number of coordinates. However, creating single and hybrid topologies requires the search of a (maximum) common substructure. Dual topologies, on the other hand, are easier to set up automatically as they do

∗ This chapter is reproduced in part from Ries, B.†; Rieder, S. R.†; Rhiner, C.; Hünenberger, P. H.; Riniker, S. RestraintMaker: A Graph-Based Approach to Select Distance Restraints in Free-Energy Calculations With Dual Topology. *J. Comput.-Aided Mol. Des.* **2022**, *36*, 175–192. †These authors contributed equally.

not require any substructure searches. To prevent the end-state molecules from drifting apart during a simulation with dual topologies, atom-atom distance restraints can be employed. In the present chapter, the program *RestraintMaker* is used to generate distance restraints for multistate relative hydration free-energy calculations with replica-exchange enveloping distribution sampling (RE-EDS). *RestraintMaker* relies on a greedy algorithm to find (locally) optimal distance restraints between pairs of atoms based on geometric measures.

5.1 INTRODUCTION

Relative binding free-energy calculations have become an important tool for computer-aided drug discovery,[146,148,149,160,238–246] in particular in the context of predicting binding affinities of ligands to a protein.[152,164,242,247–253] Well-established pathway-dependent pairwise free-energy methods include thermodynamic integration (TI),[156] free-energy perturbation (FEP),[153] Bennett's acceptance ratio (BAR),[157] and multistate BAR (MBAR).[158] Further, multisite λ-dynamics[159–161] is a pathway-dependent multistate free-energy method, and enveloping distribution sampling (EDS)[162,163] is a pathway-independent multistate free-energy method. More recently, replica-exchange EDS (RE-EDS)[164–166] and accelerated EDS (A-EDS)[167,168] were introduced as extensions of EDS that make use of the high sampling efficiency of EDS and aim to increase the accuracy and robustness of the obtained free-energy estimates.

Additionally to the free-energy method, the choice of an end-state representation is an essential ingredient of any free-energy calculation. Recently, Ries, Rieder *et al.*[254] proposed a categorization of three commonly used (and sometimes ambiguously termed[255–257]) approaches.[254] First, in a *single-topology*[258,259] approach, only one set of coordinates is used to represent all the end-states. If the number of coordinates differs between

the molecules, dummy atoms are used to represent the coordinates of the larger molecule(s) that do not have a corresponding coordinate in the smaller one(s). Second, in a *hybrid-topology*[252,260] approach, the common core of the two molecules is represented by a single set of coordinates, whereas the coordinates that do not belong to the common core are represented separately, and perturbed to dummy atoms if they are inactive. Third, in a *dual-topology*[258,261] approach, the coordinates of all molecules are represented explicitly, and perturbed to dummy atoms if they are inactive. In this approach, the molecules could in principle drift away from each other during the simulation. In Ref. 254, three sub-categories of dual-topology approaches are distinguished: linked, separated, and unconstrained.[166] With a linked dual topology-approach, the molecules are prevented from drifting apart *via* direct spatial restraints, such as atom-atom distance restraints. In the separated approach the molecules are positionally restrained to the same area, whereas they are not restrained at all in the unconstrained approach.

Tools to automatically set up single-topology free-energy calculations include BioSimSpace,[262] FESetup,[263] LOMAP,[264] ProtoCaller,[265] or SMArt.[266] FEPrepare is a web-based tool to automate the setup of relative binding free-energy calculations using a hybrid-topology approach.[267] The packages FEW[268] and pyAutoFEP[269] allow the automated set-up of unconstrained dual-topology free-energy calculations. The QligFEP pipeline provides the set-up of free-energy calculations with the linked dual-topology approach, placing distance restraints in the common substructure of the perturbed end-states.[242] Ries, Rieder *et al.*[254] introduced the Python program *RestraintMaker*.[254] It uses a greedy algorithm to assign (locally) optimal distance restraints for pairwise or multistate linked dual-topology free-energy calculations. Notably, the algorithm does not require the molecules to involve a common core or substructure. *RestraintMaker* can be used as a scripting library or within the GUI of PyMOL.[270]

In the present chapter, pairwise relative hydration free energies $\Delta\Delta G^{ji}_{\mathrm{hyd}}$ are calculated with RE-EDS for two sets of molecules, employing distance

restraints generated using *RestraintMaker*. Due to the much smaller system sizes, hydration free-energy calculations are considerably faster than binding free-energy calculations. Therefore, they represent a useful test case for assessing the quality of different methods or force-fields, as well as novel tools such as *RestraintMaker*.[170,172] The obtained relative hydration free energies are compared to experimental values as well as TI calculation results as reported on the ATB server.[135] For two molecules i and j, the pairwise relative hydration free energies are calculated either from free-energy differences in vacuum/water (for the RE-EDS results) or from hydration free energies (for the experimental values and the TI results) as

$$\Delta\Delta G^{ji}_{\mathrm{hyd}} = \Delta G^{ji}_{\mathrm{wat}} - \Delta G^{ji}_{\mathrm{vac}} = \Delta G^{j}_{\mathrm{hyd}} - \Delta G^{i}_{\mathrm{hyd}}\,, \tag{5.1}$$

where $\Delta G^{ji}_{\mathrm{wat}}$ is the free-energy difference between molecules i and j in water, $\Delta G^{ji}_{\mathrm{vac}}$ is the corresponding free-energy difference in vacuum, and $\Delta G^{i}_{\mathrm{hyd}}$ is the hydration free energy of molecule i.

The pairwise relative translational and rotational motions of the molecules during the RE-EDS simulations are also investigated, to assess whether the assigned distance restraints are successful in keeping the molecules well-aligned. Finally, it is verified that the conformational sampling behavior of the molecules is not negatively affected by the distance restraints, even when a flexible cyclohexene ring is restrained to rigid benzene rings.

5.2 THEORY

5.2.1 Free-Energy Methods

In the present chapter, relative hydration free energies calculated with two different free-energy methods are compared. First, the pathway-dependent

TI method, and second, the multistate pathway-independent RE-EDS method.

Thermodynamic Integration

TI is a well-established free-energy method.[156] A linear coupling scheme is used to define a pathway from an end-state A to an end-state B. The end-states correspond, *e.g.*, to a molecule in two different environments (for absolute free-energy calculations) or to two different molecules in the same environment (for relative free-energy calculations). The potential energy of the system is defined as

$$V(\mathbf{r};\lambda) = (1-\lambda)\ V_A(\mathbf{r}) + \lambda\ V_B(\mathbf{r})\,, \tag{5.2}$$

where λ is the coupling parameter. The potential energy at $\lambda = 0$ corresponds to that of end-state A and at $\lambda = 1$ to that of end-state B. The system is simulated at different λ-values between 0 and 1, and the free-energy difference between end-states A and B is estimated as[156]

$$\Delta G_{BA} = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda\,. \tag{5.3}$$

Replica-Exchange Enveloping Distribution Sampling

RE-EDS[164–166] combines Hamiltonian replica exchange (RE)[271,272] with enveloping distribution sampling (EDS).[162,163] EDS is a multistate free-energy method that uses a reference potential V_R combining N end-states as

$$V_R\left(\mathbf{r};s,\mathbf{E}^R\right) = -\frac{1}{\beta s}\ln\left[\sum_{i=1}^{N} e^{-\beta s\left(V_i(\mathbf{r})-E_i^R\right)}\right], \tag{5.4}$$

where s is the smoothness parameter, $\mathbf{E}^R$ is a vector of energy offsets and $\beta = 1/(k_B\ T)$, where k_B is the Boltzmann constant and T the absolute

temperature. At high s-values ($s \sim 1$), due to the negative exponent, the sampling of the reference state is dominated by the end-state with the lowest value of $V_i(\mathbf{r}) - E_i^R$. When the s-value is decreased, more end-states start to contribute to the reference potential. Finally, at low s-values (close to zero), all end-states contribute (roughly) equally to V_R. This situation is referred to as "undersampling".[247]

From a single EDS calculation, the free-energy difference between any pair of end-states in the system is estimated as

$$\Delta G_{BA} = -\frac{1}{\beta} \ln \frac{\langle e^{-\beta(V_B - V_R)} \rangle_R}{\langle e^{-\beta(V_A - V_R)} \rangle_R} . \tag{5.5}$$

To obtain adequate sampling behavior and accurate free-energy estimates from an EDS simulation, an optimal choice of the s-value and of the energy offsets is required in practice.[247] The RE-EDS free-energy method aims at mitigating the need for an optimal s-value, and to enhance the sampling of EDS. To this end, multiple replicas of the system are simulated at different s-values, and Hamiltonian replica exchanges of the s-values are attempted at fixed intervals.[164,165]

5.3 METHODS

5.3.1 RestraintMaker

The algorithm implemented in *RestraintMaker* to assign atom-atom distance restraints for both pairwise and multistate free-energy calculations is briefly outlined here, and described in detail in Ref. 254. It is a graph-based greedy algorithm to select (locally) optimal atomic distance restraints for pre-aligned molecules. The three criteria to define an optimal placement for two molecules are: (*i*) the distance of the restrained atom pairs is maximally distributed over the two molecules; (*ii*) the distance between the restrained atoms of the aligned molecules is small; and (*iii*)

the conformational sampling is not influenced by the restraints.[254] To satisfy condition (*iii*), the distance restraints are usually only placed in relatively rigid regions of the molecules, *e.g.*, only in rings. The only required user inputs for *RestraintMaker* are the number of restraints and a cutoff distance d_{res}. Atoms that are further apart than d_{res} are not considered as potential restraint pairs. Note that, in general, the desired number of restraints is much smaller than the number of atoms in the aligned molecules.

For multistate simulations (*i.e.*, simulations with multiple end-state molecules in the system), the algorithm generates a cycle of pairwise distance restraints between the molecules (Figure 5.1). It initially generates all pairwise optimal distance restraints between the molecules. Next, each pair of molecules is assigned a weight, namely the convex hull volume (CHV) of the midpoints of the atom-atom distance restraints. Following that, the (locally) optimal chain of restrained molecules is created by greedily selecting the molecule pairs with the largest CHVs, without generating cycles or branches.[254,273] Finally, the chain is closed by connecting its ends.

5.3.2 Datasets

The appropriateness of the distance restraints assigned automatically by *RestraintMaker* was investigated in the context of multistate free-energy calculations by estimating relative hydration free energies for two sets of molecules with RE-EDS. Both sets contain molecules assembled from the ATB server[135] with available experimental[170,173,274–277] and calculated[135] reference values.

The first set, labeled set E, contains six molecules with the same benzene core and R-group changes (Figure 5.2, top, Table 5.1, top). The second set, labeled set F, consists of ten molecules without a common core, involving more complex transformations such as ring-flexibility changes (Figure 5.2, bottom, Table 5.1, bottom). Note that the two sets were

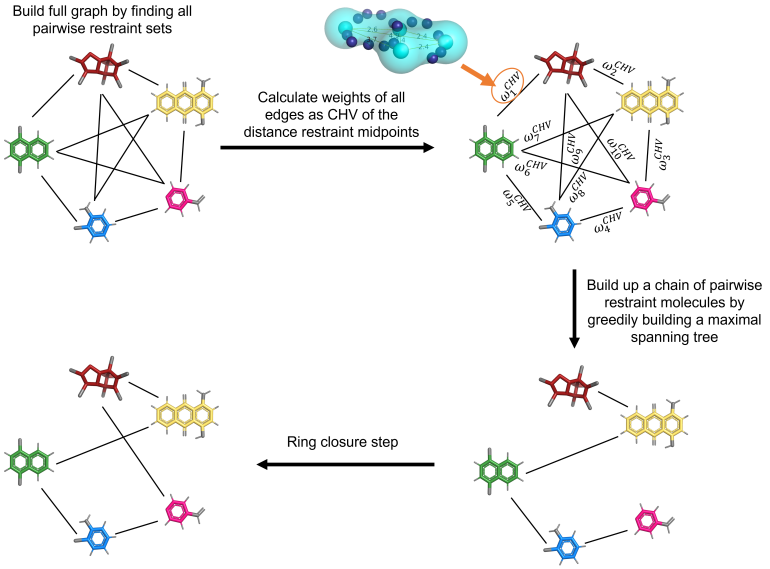


Figure 5.1: Schematic illustration of the algorithm to create distance restraints for a multistate relative free-energy calculation. The selection is carried out in four steps: (*i*) optimal restraints are calculated for all possible molecule pairs, building up a fully connected graph; (*ii*) the weights ω_i^{CHV} of the edges are calculated as the convex hull volume (CHV) formed by the selected restraints; (*iii*) a maximum spanning tree without branching is greedily constructed by selecting the edges with maximal weights; and (*iv*) the ring is closed by connecting the ends of the chain. Figure from Ries, Rieder *et al.*,[254] created by Benjamin Ries.

originally labeled set A and set B, respectively, in Ref. 254. They were relabeled here to avoid confusion with sets A and B of Chapter 6.

5.3.3 Simulation Details

The RE-EDS simulations were carried out using a modified version of the GROMOS MD package[27] version 1.5.0, the RE-EDS pipeline,[166,278] and PyGromosTools.[279]

The simulation parameters were analogous to the ones used to calculate the hydration free energies reported on the ATB server.[135] The simple

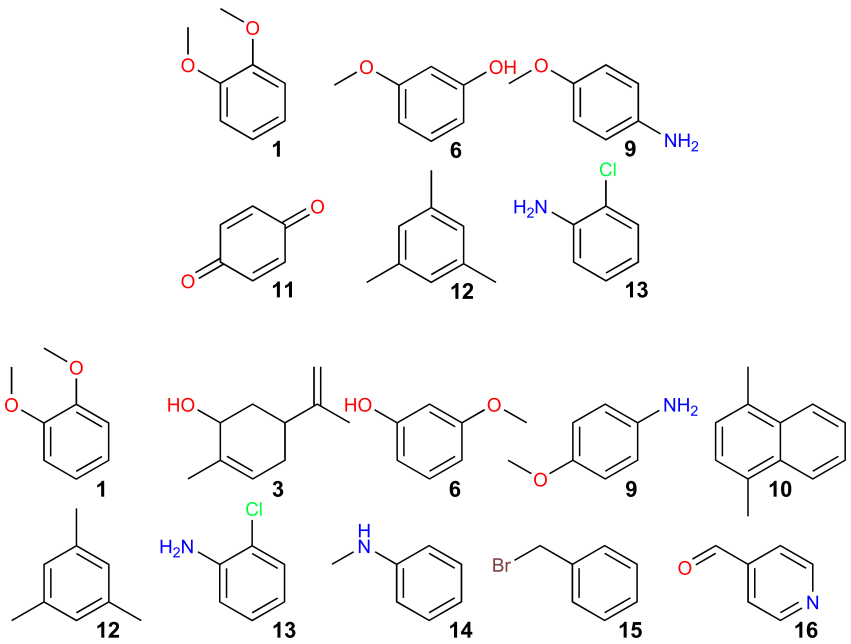


Figure 5.2: Sets E and F used for the RE-EDS simulations. (Top): Set E consists of six molecules with the same benzene core and R-group changes. The distance restraints selected by *RestraintMaker* are shown in Figure 5.3. (Bottom): Set F consists of ten molecules without a common core, involving ring-flexibility changes. The distance restraints selected by *RestraintMaker* are shown in Figure 5.4.

point-charge (SPC) model[280] was used for the simulations in water. The long-range electrostatic interactions were modeled with a reaction-field (RF) correction[114] with the RF permittivity $\varepsilon_{\mathrm{RF}}$ set to one (vacuum) and 61 (water), respectively.[281] For the nonbonded interactions, a single cutoff radius of 1.2 nm was used. The time step was set to 2 fs and the pairlist was updated every five steps. The force constant for the distance restraints was set to 5000 kJ/(mol·nm^2).

The merged topologies for the RE-EDS calculations were prepared using PyGromosTools,[279] and the simulations were set-up using the RE-EDS pipeline.[166] The individual steps of the pipeline are described in detail in Ref. 166. The coordinates of the molecules were aligned to reference molecule **1** (Figure 5.2). The RDKit[210] was used to determine

Table 5.1: Sets E and F: Identifier of the ATB server,[135] IUPAC name, and canonical SMILES strings for the six molecules of set E and the ten molecules of set F.

Set	Molecule	Identifier	IUPAC name	Canonical SMILES (RDKit[210])
E	1	_O6T	1,2-dimethoxybenzene	COc1ccccc1OC
	6	6KET	3-methoxyphenol	COc1cccc(O)c1
	9	F313	4-methoxyaniline	COc1ccc(N)cc1
	11	G277	cyclohexa-2,5-diene-1,4-dione	O=C1C=CC(=O)C=C1
	12	M030	1,3,5-trimethylbenzene	Cc1cc(C)cc(C)c1
	13	M097	2-chloroaniline	Nc1ccccc1Cl
F	1	_O6T	1,2-dimethoxybenzene	COc1ccccc1OC
	3	_O71	(1S,5R)-2-methyl-5-prop-1-en-2-ylcyclohex-2-en-1-ol	C=C(C)[C@@H]1CC=C(C)[C@@H](O)C1
	6	6KET	3-methoxyphenol	COc1cccc(O)c1
	9	F313	4-methoxyaniline	COc1ccc(N)cc1
	10	G078	1,4-dimethylnaphthalene	Cc1ccc(C)c2ccccc12
	12	M030	1,3,5-trimethylbenzene	Cc1cc(C)cc(C)c1
	13	M097	2-chloroaniline	Nc1ccccc1Cl
	14	M218	N-methylaniline	CNc1ccccc1
	15	S002	bromomethylbenzene	BrCc1ccccc1
	16	TVVS	pyridine-4-carbaldehyde	O=Cc1ccncc1

the pairwise maximum common substructures (MCSs), using the molecular skeleton of the rings only, and align the molecules based on it. For some molecules, manual modifications were applied to ensure an optimal overlap of the ring atoms and substituents. The corresponding script is available in the example folder on GitHub (`https://github.com/rinikerlab/restraintmaker/blob/main/examples/publication/b_ATB_solvationFreeEnergies/sets/multistate/prepare_distance_restraints.py`). *RestraintMaker* was then used to select the distance restraints to connect the molecules in a chain using a cutoff distance of 0.1 nm and four pairs of atoms to be restrained per molecule pair (Figures 5.3 and 5.4).

For set E, six EDS simulations of 2 ns length were carried out with $s = 1.0$ in vacuum/water to generate optimized configurations for the starting state mixing (SSM).[166] Each of the six simulations was biased towards one of the end-states by setting the energy offset of that end-state to 500 kJ mol^{-1} and the energy offsets of the other end-states in the same simulation to -500 kJ mol^{-1}. Subsequently, 21 EDS simulations were carried out for 0.2 ns with s-values distributed logarithmically between 1 and 10^{-5} to determine the lower bound for the RE-EDS simulations (0.00316 in vacuum and 0.001 in water). Next, the energy offsets were estimated from a 0.8 ns RE-EDS simulation, with 12 replicas in vacuum, and 14 replicas in water. In vacuum, one s-optimization step of 0.5 ns length adding four replicas was sufficient to achieve frequent round trips

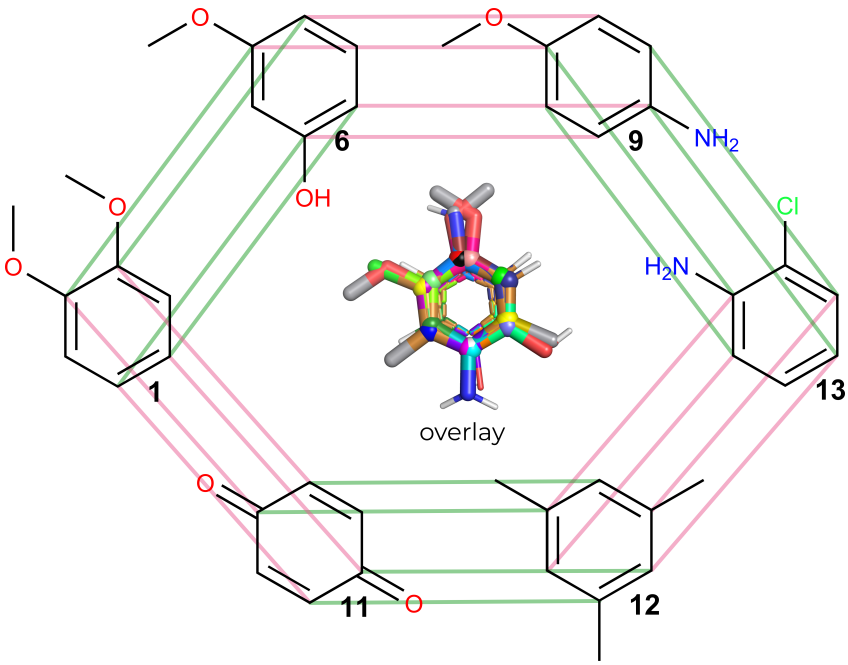


Figure 5.3: Selected distance restraints for the RE-EDS simulations of set E (Figure 5.2, top). For each pair of molecules, four distance restraints were determined with *RestraintMaker*. The lines indicate which atoms were restrained to each other.

and good sampling of all end-states. In water, three *s*-optimization steps of 0.5 ns, 1.0 ns and 1.5 ns, respectively, were carried out to achieve frequent round trips. At each step, five replicas were added. Additionally, in water, the energy offsets were rebalanced over three 0.5 ns simulations to optimize the sampling of all end-states. Energy-offset rebalancing was not necessary in vacuum as all end-states were already sampled well after the *s*-optimization. The production run was 0.5 ns long in vacuum and 1.0 ns in water.

For set F, ten EDS simulations of 2 ns were performed in vacuum/water to generate optimized configurations analogously to set E. The determination of the lower bounds was carried out as above (0.00178 in vacuum and 0.001 in water). For the energy offset estimation, a 0.8 ns RE-EDS

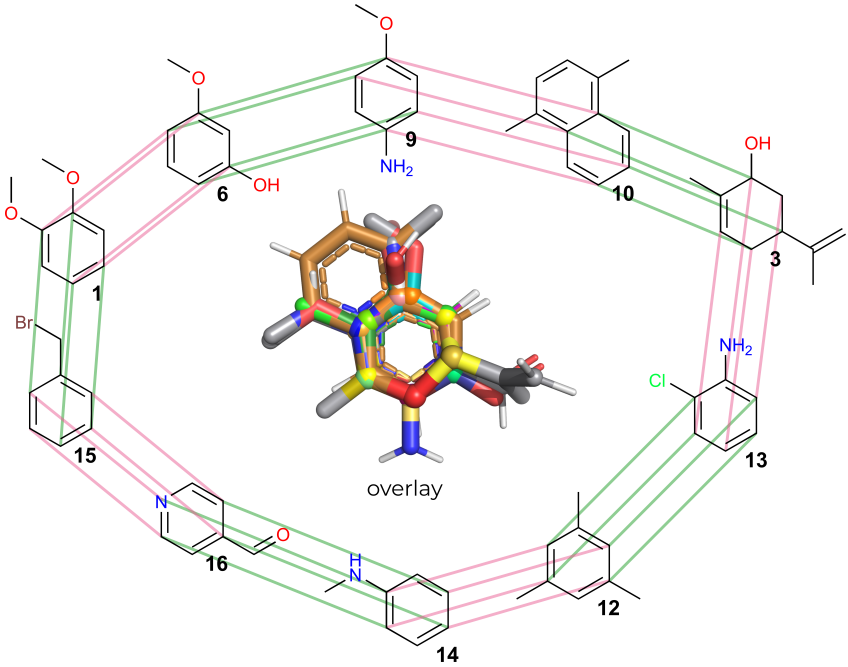


Figure 5.4: Selected distance restraints for the RE-EDS simulations of set F (Figure 5.2, bottom). For each pair of molecules, four distance restraints were determined with *RestraintMaker*. The lines indicate which atoms were restrained to each other.

simulation was also used, with 17 replicas in vacuum and 18 replicas in water. In vacuum, one *s*-optimization step of 0.5 ns length adding four replicas was again sufficient. In water, five *s*-optimization steps of 0.5 ns, 1.0 ns, 1.5 ns, 1.5 ns, and 1.5 ns, respectively, were carried out to achieve frequent round trips. At each step, five replicas were added. In water, the energy offsets were rebalanced over five 0.5 ns simulations. The production run was 1.0 ns long in vacuum and 5.0 ns in water.

5.3.4 Analysis

For the analysis, the GROMOS++[282] and PyGromosTools[279] packages were used. Further, the following Python packages were used for statistical analysis and visualization: Pandas,[283] Matplotlib,[284] NumPy,[285]

SciPy,[286] mpmath,[287] and Jupyter notebooks.[288] For the $\Delta\Delta G_{\text{hyd}}$ values obtained with TI (as reported on the ATB server[135]) as well as RE-EDS, the root-mean-square error (RMSE), the mean absolute error (MAE) and the Spearman[289] correlation coefficient were calculated between the different simulation methods and the experimental values.

5.4 RESULTS

5.4.1 Calculation of Relative Hydration Free Energies

For set E, the results obtained with RE-EDS agreed well with the experimental values (Figure 5.5), with an RMSE of 3.6 kJ mol^{-1} and a MAE of 2.9 kJ mol^{-1}. The Spearman correlation coefficient was 0.93, indicating high correlation with experiment. The highest deviations from experiment were observed for the molecule pair **6** - **11** with a deviation of 7.5 kJ mol^{-1}. The numerical values are reported in Table 5.2. For comparison, $\Delta\Delta G_{\text{hyd}}^{\text{TI}}$ values were derived from the absolute hydration free energies reported on the ATB server,[135] giving an RMSE of 6.2 kJ mol^{-1} and a MAE of 5.3 kJ mol^{-1}. The Spearman correlation coefficient was almost identical with a value of 0.92. In this case, the highest deviations from experiment were observed for the molecule pairs **1** - **11** and **6** - **11**, with deviations of 11.1 kJ mol^{-1} and 11.3 kJ mol^{-1}, respectively. The convergence of the RE-EDS calculations is shown in Figure 5.6.

For set F, the results obtained with RE-EDS also agreed well with the experimental values (Figure 5.7) with an RMSE of 3.3 kJ mol^{-1} and a MAE of 2.7 kJ mol^{-1}. The Spearman correlation coefficient with experiment was 0.96. The highest deviations from experiment were observed for the molecule pairs **14** - **9**, **14** - **10**, **15** - **10**, and **16** - **14** with absolute deviations between 6.1 and 6.9 kJ mol^{-1}. The numerical values are reported in Table 5.3. For comparison, the $\Delta\Delta G_{\text{hyd}}^{\text{TI}}$ values were again

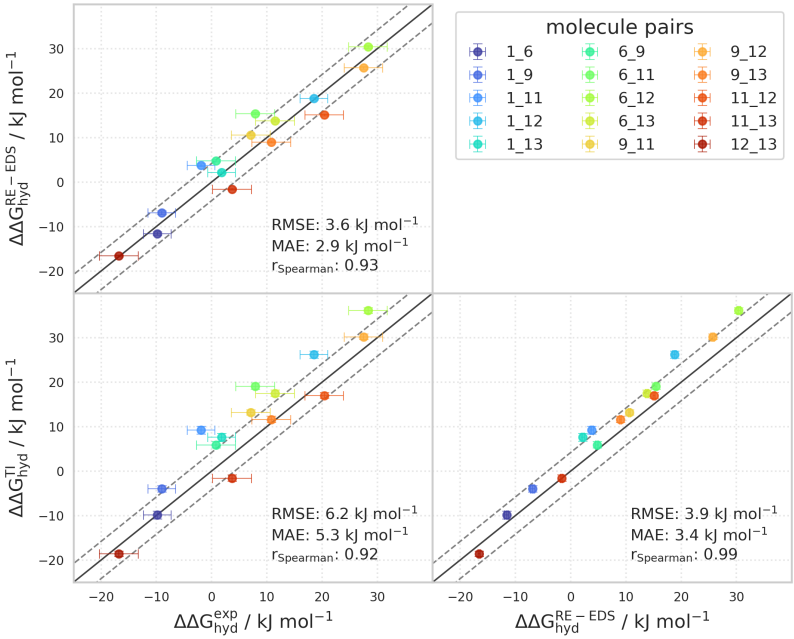


Figure 5.5: Comparison of the relative hydration free energies $\Delta\Delta G_{\mathrm{hyd}}$ for the six molecules in set E between experiment ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$), the multistate relative free-energy calculations with RE-EDS and linked dual topology ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$), and the absolute free-energy calculations with TI taken from the ATB server[135] ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$). The numerical values are provided in Table 5.2.

derived from the absolute hydration free energies reported on the ATB server,[135] giving an RMSE of 5.8 kJ mol^{-1} and a MAE of 4.8 kJ mol^{-1}. The Spearman correlation coefficient was almost identical with a value of 0.93. In this case, the highest deviations from experiment were observed for the molecule pairs **10 - 1**, **10 - 3**, **10 - 6**, **14 - 10**, **14 - 12**, and **15 - 10**, with absolute deviations between 9.1 and 10.7 kJ mol^{-1}. The convergence of the RE-EDS calculations is shown in Figure 5.8.

Table 5.2: $\Delta\Delta G_{\mathrm{hyd}}$ for the six molecules in set E from experiment ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$), the absolute free-energy calculations with TI taken from the ATB server [135] ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$), and the multistate relative free-energy calculations with RE-EDS and linked dual topology ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$). The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The experimental uncertainty for molecule 11 was set to a default value of 2.51 kJ mol^{-1}, [173] as the uncertainty was not reported in the original source. [274] The RMSE and its uncertainty were estimated with a 100 fold bootstrap approach. The accumulated simulation time is split into preparation (pre-processing, equilibration) and production run. The data is displayed graphically in Figure 5.5.

Molecules		Experiment	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$ [135]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
1	6	$-9.8 \pm 2.5^{275,277}$	-9.9 ± 0.9	-11.6 ± 0.4
1	9	$-9.0 \pm 2.5^{275,277}$	-4.0 ± 0.7	-6.9 ± 0.4
1	11	$-1.9 \pm 2.5^{274,277}$	9.2 ± 0.9	3.8 ± 0.3
1	12	$18.5 \pm 2.5^{275,277}$	26.2 ± 0.8	18.8 ± 0.3
1	13	$1.8 \pm 2.5^{275,277}$	7.6 ± 0.9	2.2 ± 0.3
6	9	0.8 ± 3.5^{275}	5.9 ± 0.7	4.8 ± 0.3
6	11	$7.9 \pm 3.5^{274,275}$	19.1 ± 0.7	15.4 ± 0.2
6	12	28.3 ± 3.5^{275}	36.1 ± 0.6	30.4 ± 0.3
6	13	11.5 ± 3.5^{275}	17.5 ± 0.7	13.8 ± 0.3
9	11	$7.1 \pm 3.5^{274,275}$	13.2 ± 0.6	10.6 ± 0.2
9	12	27.5 ± 3.5^{275}	30.2 ± 0.6	25.7 ± 0.3
9	13	10.8 ± 3.5^{275}	11.6 ± 0.7	9.0 ± 0.2
11	12	$20.4 \pm 3.5^{274,275}$	17.0 ± 0.6	15.1 ± 0.2
11	13	$3.7 \pm 3.5^{274,275}$	-1.6 ± 0.7	-1.6 ± 0.2
12	13	-16.8 ± 3.5^{275}	-18.6 ± 0.6	-16.6 ± 0.2
RMSE			6.2 ± 0.3	3.6 ± 0.3
MAE			5.3 ± 3.2	2.9 ± 2.1
r_{Spearman}			0.92	0.93
$t_{\mathrm{preparation}}$				222 ns
$t_{\mathrm{production}}$			42 − 102 ns	36 ns

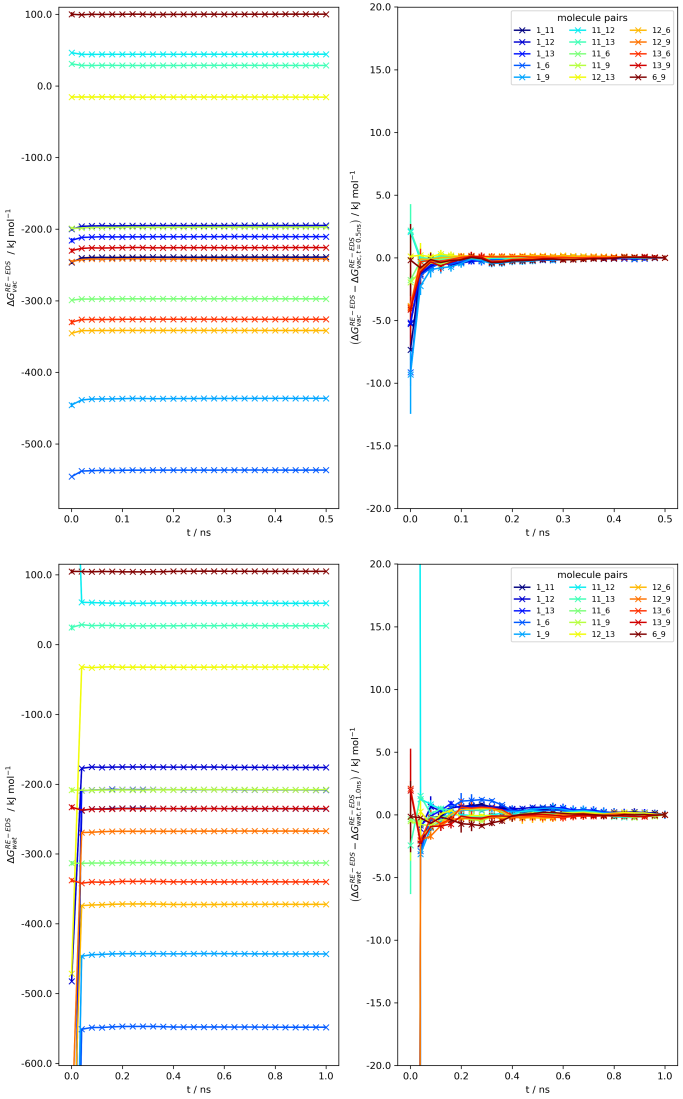


Figure 5.6: Convergence of $\Delta G_{\mathrm{vac}}^{\mathrm{RE\text{-}EDS}}$ and $\Delta G_{\mathrm{wat}}^{\mathrm{RE\text{-}EDS}}$ as a function of the simulation time for the RE-EDS simulation of set E ($s = 1.0$) in vacuum (top) and in water (bottom).

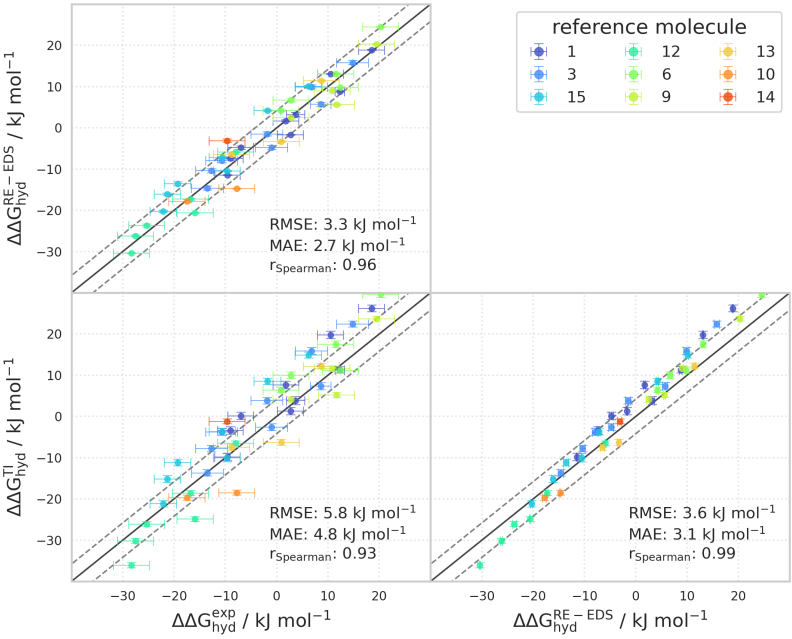


Figure 5.7: Comparison of the relative hydration free energies $\Delta\Delta G_{\mathrm{hyd}}$ for the ten molecules in set F between experiment ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$), the multistate relative free-energy calculations with RE-EDS and linked dual topology ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$), and the absolute free-energy calculations with TI taken from the ATB server[135] ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$). The numerical values are provided in Table 5.3.

Table 5.3: $\Delta\Delta G_{\mathrm{hyd}}$ for the ten molecules in set F from experiment ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$), the absolute free-energy calculations with TI taken from the ATB server [135] ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$), and the multistate relative free-energy calculations with RE-EDS and linked dual topology ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$). The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The RMSE and its uncertainty were estimated with a 100 fold bootstrap approach. The accumulated simulation time is split into preparation (pre-processing, equilibration) and production run. The data is displayed graphically in Figure 5.7.

Molecules		Experiment	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$ [135]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
1	3	3.7 ± 1.8^{277}	3.8 ± 1.0	3.2 ± 0.5
1	15	$12.3 \pm 0.9^{276,277}$	11.3 ± 0.8	8.8 ± 0.2
1	12	$18.5 \pm 2.5^{275,277}$	26.2 ± 0.8	18.9 ± 0.1
1	6	$-9.6 \pm 2.5^{275,277}$	-9.8 ± 0.9	-11.5 ± 0.1
1	9	$-9.0 \pm 2.5^{275,277}$	-3.4 ± 0.9	-7.4 ± 0.2
1	13	$1.8 \pm 2.5^{275,277}$	7.6 ± 0.9	1.7 ± 0.2
1	10	$10.5 \pm 2.5^{275,277}$	19.8 ± 0.9	13.1 ± 0.1
1	14	$2.7 \pm 2.5^{275,277}$	1.3 ± 0.9	-1.7 ± 0.1
1	16	-7.0 ± 2.5^{277}	0.1 ± 0.9	-4.8 ± 0.4
3	15	$8.6 \pm 2.0^{276,277}$	7.4 ± 0.9	5.7 ± 0.5
3	12	$14.8 \pm 3.1^{275,277}$	22.4 ± 0.7	15.8 ± 0.5
3	6	$-13.5 \pm 3.1^{275,277}$	-13.7 ± 0.8	-14.6 ± 0.5
3	9	$-12.7 \pm 3.1^{275,277}$	-7.8 ± 0.8	-10.4 ± 0.5
3	13	$-1.9 \pm 3.1^{275,277}$	3.8 ± 0.8	-1.5 ± 0.5
3	10	$6.8 \pm 3.1^{275,277}$	15.9 ± 0.8	9.9 ± 0.5
3	14	$-1.0 \pm 3.1^{275,277}$	-2.6 ± 0.8	-4.8 ± 0.5
3	16	$-10.7 \pm 3.1^{275,277}$	-3.8 ± 0.8	-7.9 ± 0.6
15	12	$6.2 \pm 2.6^{275,276}$	14.9 ± 0.7	10.1 ± 0.2
15	6	$-22.1 \pm 2.6^{275,276}$	-21.2 ± 0.8	-20.3 ± 0.2
15	9	$-21.3 \pm 2.6^{275,276}$	-15.2 ± 0.8	-16.1 ± 0.3
15	13	$-10.6 \pm 2.6^{275,276}$	-3.7 ± 0.8	-7.2 ± 0.3
15	10	$-1.8 \pm 2.6^{275,276}$	8.5 ± 0.8	4.2 ± 0.2
15	14	$-9.7 \pm 2.6^{275,276}$	-10.0 ± 1.0	-10.5 ± 0.2
15	16	$-19.3 \pm 2.6^{275,276}$	-11.2 ± 0.8	-13.6 ± 0.5
12	6	-28.3 ± 3.5^{275}	-36.1 ± 0.6	-30.4 ± 0.1
12	9	-27.5 ± 3.5^{275}	-30.2 ± 0.6	-26.2 ± 0.2
12	13	-16.8 ± 3.5^{275}	-18.6 ± 0.6	-17.3 ± 0.2
12	10	-8.0 ± 3.5^{275}	-6.5 ± 0.6	-5.9 ± 0.1
12	14	-15.9 ± 3.5^{275}	-24.9 ± 0.6	-20.6 ± 0.1
12	16	-25.3 ± 3.5^{275}	-26.1 ± 0.6	-23.7 ± 0.4
6	9	0.8 ± 3.5^{275}	6.4 ± 0.7	4.2 ± 0.2
6	13	11.5 ± 3.5^{275}	17.5 ± 0.8	13.1 ± 0.2
6	10	20.3 ± 3.5^{275}	29.6 ± 0.8	24.5 ± 0.1
6	14	12.4 ± 3.5^{275}	11.2 ± 0.8	9.8 ± 0.1
6	16	2.8 ± 3.5^{275}	10.0 ± 0.8	6.7 ± 0.4
9	13	10.8 ± 3.5^{275}	11.6 ± 0.7	9.0 ± 0.3
9	10	19.5 ± 3.5^{275}	23.7 ± 0.6	20.3 ± 0.2
9	14	11.7 ± 3.5^{275}	5.2 ± 0.6	5.6 ± 0.2
9	16	2.8 ± 3.5^{275}	4.1 ± 0.6	2.5 ± 0.4
13	10	8.7 ± 3.5^{275}	12.2 ± 0.7	11.4 ± 0.2
13	14	0.9 ± 3.5^{275}	-6.3 ± 0.7	-3.3 ± 0.2
13	16	-8.8 ± 3.5^{275}	-7.5 ± 0.7	-6.5 ± 0.4
10	14	-7.8 ± 3.5^{275}	-18.5 ± 0.6	-14.7 ± 0.1
10	16	-17.5 ± 3.5^{275}	-19.7 ± 0.6	-17.8 ± 0.4
14	16	-9.7 ± 3.5^{275}	-1.2 ± 0.6	-3.1 ± 0.4
RMSE			5.8 ± 0.1	3.3 ± 0.1
MAE			4.8 ± 3.3	2.7 ± 1.8
r_{Spearman}			0.93	0.96
$t_{\mathrm{preparation}}$				418 ns
$t_{\mathrm{production}}$			70 – 170 ns	212 ns

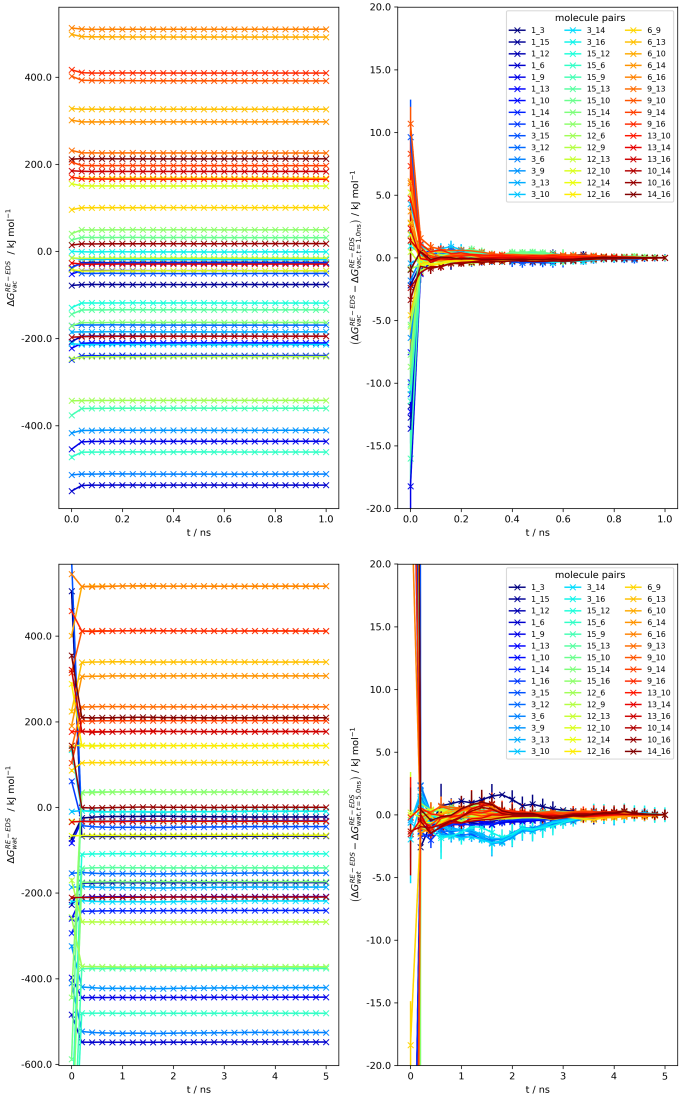


Figure 5.8: Convergence of $\Delta G_{\text{vac}}^{\text{RE-EDS}}$ and $\Delta G_{\text{wat}}^{\text{RE-EDS}}$ as a function of the simulation time for the RE-EDS simulation of set F ($s = 1.0$) in vacuum (top) and in water (bottom).

5.4.2 Sampling

It is essential to verify that (*i*) the employed distance restraints are successful in keeping the molecules well aligned during the simulation; and (*ii*) the conformational sampling of the molecules is not negatively affected by the employed distance restraints. For this, the pairwise relative translational motion of the molecules in the two sets was analyzed for all *s*-values. For a given pair of molecules, this relative motion was calculated as the fluctuations of the distance between the centers of geometry (COGs) of the restrained atoms in the central rings of the two molecules. The pairwise relative rotational motions of the molecules in the sets were also analyzed for all *s*-values by considering the three Euler angles. Further, the substituent or pseudo torsional angle distributions of a selection of molecules were compared to plain MD simulations.

For set E, both the translational fluctuations of the COGs of the central rings as well as the rotational fluctuations were reasonably small. The maximum deviation between the COGs of the molecules was 0.05 nm and the standard deviation of the distance distributions was close to zero for all pairs (Figure 5.9). The relative rotational motions around the z-axis were almost non-existent with a maximum value of 0.7°. The maximum relative rotation was 6.28° around both the x- and the y-axis. The fraction of rotations above 5° was maximally 3.2 % around the x-axis and 3.7 % around the y-axis (Figure 5.10).

To investigate the influence of the restraints on the conformational sampling, the torsional angle distributions of the substituents of molecules **1**, **6**, and **9** were analyzed. The torsional angle distributions obtained during the RE-EDS simulation at $s = 1$ in vacuum/water were compared to the distributions obtained in a plain MD simulation in vacuum/water. No major differences were observed between the respective distributions (Figure 5.11).

The analyses of the translational and rotational motions in the RE-EDS simulations of set F are shown in Figures 5.12 and 5.13. Again, the

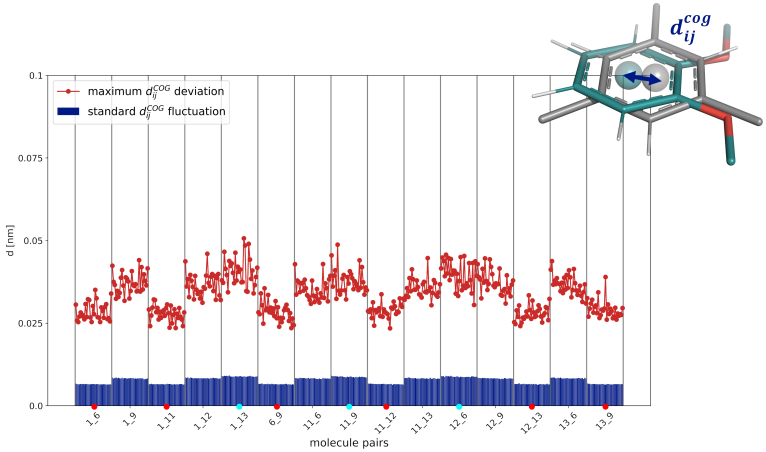


Figure 5.9: Standard deviation of the distance distribution (blue) and maximum distance (red) between the COGs of the central rings of the molecule pairs in the RE-EDS simulations of set E in water ($s = 1.0$). The COG was calculated for the restrained atoms in the rings. The horizontal axis shows, for each molecule pair, the different s-values between 1.0 and 0.001. The molecule pairs that were directly restrained to each other are marked with a red dot on the horizontal axis, and the pairs on opposite sides of the chain of restraints are marked with a cyan dot.

distance restraints kept the molecules well-aligned. Molecule **3** is the only molecule in set F that contains a flexible cyclohexene core instead of a rigid benzene core. Despite being restrained to benzene rings, the reweighted distributions of the three pseudo torsional angles of the cyclohexene ring of molecule **3** during the RE-EDS simulation at $s = 1$ agreed very well with the distributions from plain MD simulations (Figure 5.14).

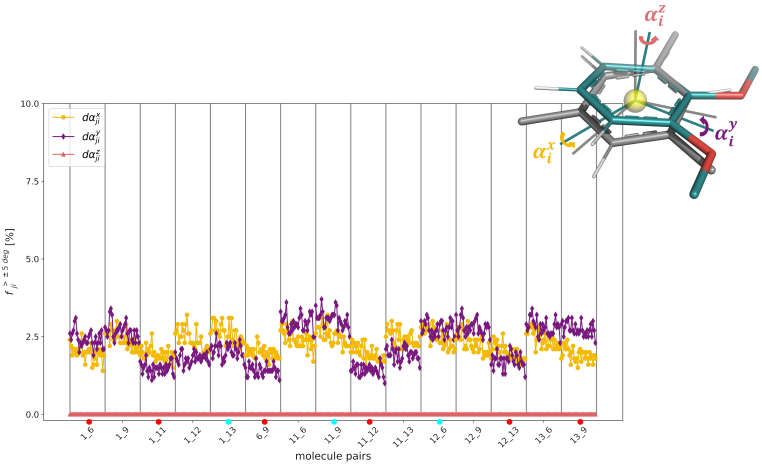


Figure 5.10: Fraction of frames in the RE-EDS simulations of set E in water ($s = 1.0$), in which the relative rotation around the x-axis (yellow), y-axis (purple), and z-axis (red) of the central rings of the molecule pair exceeds 5°. The horizontal axis shows, for each molecule pair, the different s-values between 1.0 and 0.001. The molecule pairs that were directly restrained to each other are marked with a red dot on the horizontal axis, and the pairs on opposite sides of the chain of restraints are marked with a cyan dot.

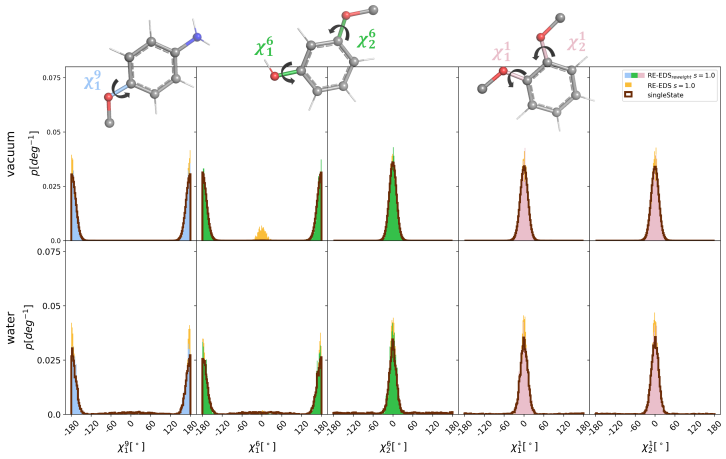


Figure 5.11: Comparison of the normalized torsional angle distributions of the substituents of molecules 9 (blue), 6 (green), and 1 (pink) in the RE-EDS simulation of set E at $s = 1.0$ (filled) and in plain MD simulations (dark red line) in vacuum (top) and in water (bottom). Both the raw (yellow) and the reweighted (with $e^{\beta(V_R - V_i)}$, blue, green, and pink) distributions are shown for RE-EDS.

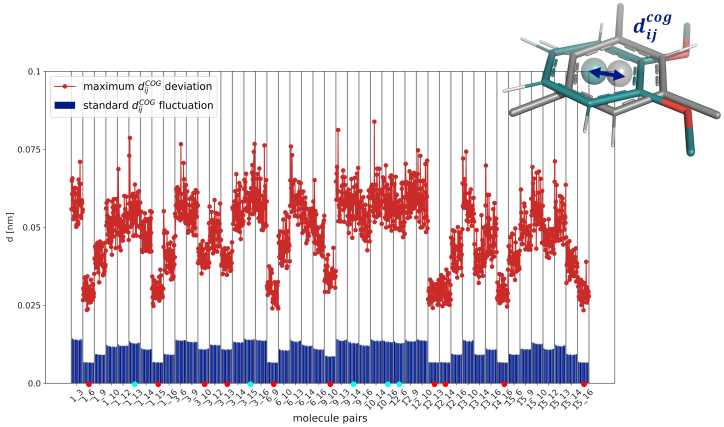


Figure 5.12: Standard deviation of the distance distribution (blue) and maximum distance (red) between the COGs of the central rings of the molecule pairs in the RE-EDS simulations of set F in water ($s = 1.0$). The COG was calculated for the restrained atoms in the rings. The horizontal axis shows, for each molecule pair, the different s-values between 1.0 and 0.001. The molecule pairs that were directly restrained to each other are marked with a red dot on the horizontal axis, and the pairs on opposite sides of the chain of restraints are marked with a cyan dot.

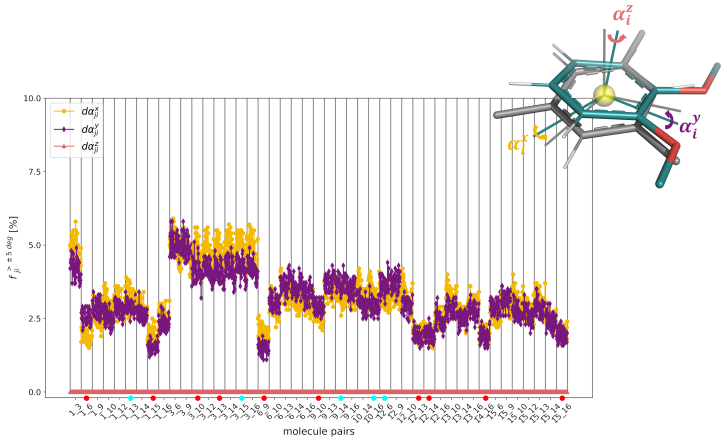


Figure 5.13: Fraction of frames in the RE-EDS simulations of set F in water ($s = 1.0$), in which the relative rotation around the x-axis (yellow), y-axis (purple), and z-axis (red) of the central rings of the molecule pair exceeds 5°. The horizontal axis shows, for each molecule pair, the different s-values between 1.0 and 0.001. The molecule pairs that were directly restrained to each other are marked with a red dot on the horizontal axis, and the pairs on opposite sides of the chain of restraints are marked with a cyan dot.

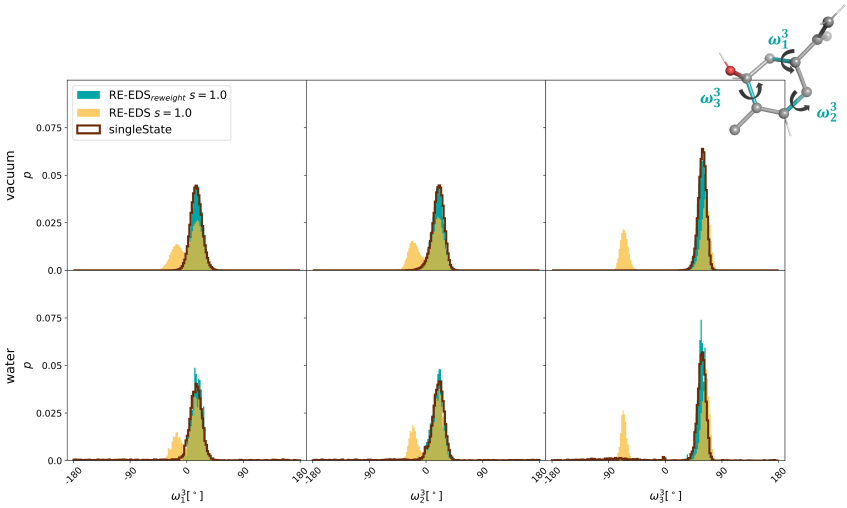


Figure 5.14: Comparison of the normalized torsional angle distributions of the three pseudo torsional angles of the cyclohexene ring of molecule **3** in the RE-EDS simulation at $s = 1.0$ (filled) and in plain MD simulations (dark red line) in vacuum (top) and in water (bottom). Both the raw (yellow) and the reweighted (with $e^{\beta(V_R - V_i)}$, cyan) distributions are shown for RE-EDS. The second peak visible in the raw RE-EDS simulations (*i.e.*, not reweighted) comes from the frames where molecule **3** is in the dummy state (Figure 5.15).

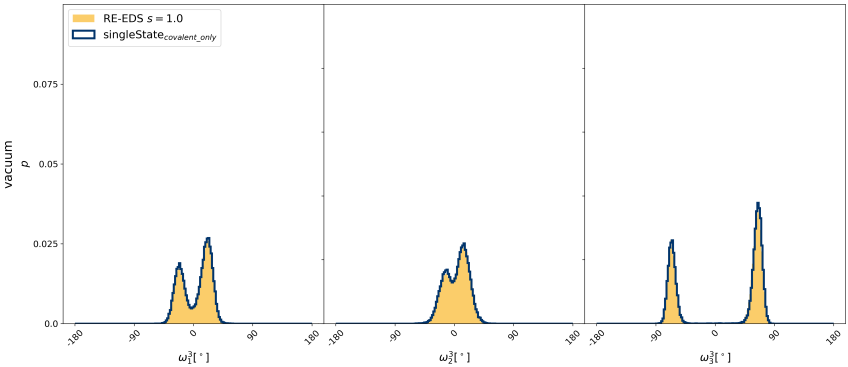


Figure 5.15: Comparison of the normalized torsional angle distributions of the three pseudo torsional angles of the cyclohexene ring of molecule **3** in a RE-EDS simulation with force constant 0 kJ/(mol·nm^2) at $s = 1.0$ (filled yellow) and in plain MD simulations in vacuum with only the covalent interactions turned on (dark blue lines). The distributions show that during the RE-EDS simulation, the same ring torsional configurations occur as when the nonbonded interactions are turned off in a single state simulation, even when the distance restraints are turned off.

5.5 CONCLUSION

The Python program *RestraintMaker* uses a graph-based greedy algorithm to automatically assign (locally) optimal atom-atom distance restraints for dual-topology free-energy calculations. In the present chapter, *RestraintMaker* was tested and validated for multistate free-energy calculations with RE-EDS. To this end, relative hydration free energies were calculated with RE-EDS for two sets of molecules. Set E contained six molecules with the same benzene core and R-group changes, whereas set F also contained a molecule with a cyclohexene core instead of a benzene core. The obtained relative hydration free energies were compared to the relative hydration free energies obtained from TI hydration free-energy calculations as reported on the ATB server, as well as from experimental hydration free-energy values. For both sets, there was good agreement between the RE-EDS results and the TI results, as well as with the experimental ones.

It was also shown that the distance restraints assigned by *RestraintMaker* are successful in keeping the molecules well-aligned during the RE-EDS simulations. For this, the pairwise relative translational motions as well as the pairwise relative rotational motions of the molecules were analyzed. Further, it was verified that the assigned distance restraints do not significantly affect the conformational sampling behavior. For this, the distributions of substituent torsional angles and pseudo torsional angles obtained from the RE-EDS simulations were compared to plain MD simulations. Since this first validation of *RestraintMaker*, the program has been used to generate distance restraints for other RE-EDS solvation free-energy calculations, including the ones described in Chapters 6 and 7.

6

RE-EDS Using GAFF Topologies: Application to Relative Hydration Free-Energy Calculations for Large Sets of Molecules*

"Now energy is a very subtle concept. It is very, very difficult to get right."

Richard Feynman [290]

Free-energy differences between pairs of end-states can be estimated based on molecular dynamics (MD) simulations using standard pathway-dependent methods such as thermodynamic integration (TI), free-energy perturbation, or Bennett's acceptance ratio. Replica-exchange enveloping distribution sampling (RE-EDS), on the other hand, allows for the sampling of multiple end-states in a single simulation without the specification of any pathways. In the present chapter, we use the RE-EDS method as implemented in GROMOS together with generalized AMBER force field (GAFF) topologies, converted

∗ This chapter is reproduced in part from Rieder, S. R.; Ries, B.; Schaller, K.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Replica-Exchange Enveloping Distribution Sampling Using Generalized AMBER Force-Field Topologies: Application to Relative Hydration Free-Energy Calculations for Large Sets of Molecules, *J. Chem. Inf. Model.* **2022**, *62*, 3043–3056.

to a GROMOS-compatible format with a newly developed GROMOS++ program *amber2gromos*, to compute relative hydration free energies for a series of benzene derivatives. The results obtained with RE-EDS are compared to the experimental data as well as calculated values from the literature. In addition, the estimated free-energy differences in water and in vacuum are compared to values from TI calculations carried out with GROMACS. The hydration free energies obtained using RE-EDS for multiple molecules are found to be in good agreement with both the experimental data and the results calculated using other free-energy methods. While all considered free-energy methods delivered accurate results, the RE-EDS calculations required the least amount of total simulation time. The present chapter serves as a validation for the use of GAFF topologies with the GROMOS simulation package and the RE-EDS approach. Furthermore, the performance of RE-EDS for a large set of 28 end-states is assessed with promising results.

6.1 INTRODUCTION

In recent years, free-energy calculations (either absolute or relative) based on classical molecular dynamics (MD) simulations have started to play an increasingly important role in the field of computer-aided drug design.[143–151] There exist many well-established pairwise free-energy methods such as thermodynamic integration (TI),[156] free-energy perturbation (FEP),[153] Bennett's acceptance ratio (BAR),[157] and multistate BAR (MBAR).[158] Approaches such as multi-site λ-dynamics[159–161] and enveloping distribution sampling (EDS)[162,163] enable the calculation of pairwise free-energy differences for multiple end-states from a single simulation. While λ-dynamics uses the coupling parameter λ as a dynamic variable

to connect the end-states, EDS is a pathway-independent method that samples a reference state "enveloping" all end-states. Recently, replica-exchange EDS (RE-EDS)[164–166] and accelerated EDS (A-EDS)[167,168] have been developed as extensions of EDS to simplify the parameter optimization and improve the performance of EDS. Both methods are implemented in the GROMOS software package.[27]

Apart from the free-energy method, the quality of the underlying force field is crucial for the accuracy of free-energy calculations, and of MD simulations in general.[33,50,51,291] In the past years, various tools have been developed to automate the otherwise laborious task of topology generation for small molecule ligands, such as Antechamber,[55,133,134] the Automated Topology Builder (ATB),[68,135] the fragment-based CombiFF,[73,74] General Automated Atomic Model Parameterization (GAAMP),[136,137] LigParGen,[138] Open Force Field (SMIRNOFF and OpenFF),[75,76] ParamChem,[59–61] PRODRG,[139] R.E.D.,[140] or SwissParam.[141] MD simulation engines such as AMBER,[20,21,292] CHARMM,[22,23] GROMACS,[24,25] GROMOS,[26,27] or OpenMM[28,29] require specific file formats to describe the system topology and coordinates. In addition, there are also small differences in the functional form of the force fields or in the units used by different MD engines.[33] In many cases, tools are already available to translate between some of the different file formats, enabling, *e.g.*, the use of AMBER topologies with GROMACS.[262,293–296]

The calculation of (absolute or relative) hydration free energies serves as a straightforward test case to assess and compare the quality of different free-energy methods and to validate force fields.[170,172] Thanks to databases such as FreeSolv,[155,173] the Minnesota solvation database,[174] or the ATB server,[68,135] ample reference data is available for both experimental results as well as calculated values obtained with different force fields and free-energy methods. Furthermore, the calculation of hydration free energies is computationally far less expensive than, for example, that of binding free energies.

In the present chapter the GROMOS++[282] program *amber2gromos*

is described. It translates a topology from the AMBER prmtop[297] file format to a GROMOS topology, enabling users of the GROMOS MD engine to simulate systems with the AMBER or generalized AMBER (GAFF[55]) force fields. Extension to the OpenFF[75,76] family of force fields is straightforward. In the following, the underlying differences between the AMBER and GROMOS force fields are discussed, and the necessary conversions are described in detail. The correctness of the topology conversion is validated by comparison of single-molecule simulations in vacuum using GROMACS or GROMOS. Furthermore, two sets of small benzene derivatives are assembled from the FreeSolv[155,173] database: a small set of 6 molecules, labeled A, and a larger set of 28 molecules, labeled B. For these molecules, relative hydration free energies are calculated with TI[156] in GROMACS (set A only) and with RE-EDS[164–166] in GROMOS (sets A and B). The results are compared to each other and to the experimental and calculated values reported in the FreeSolv[155,173] database.

6.2 THEORY

6.2.1 Differences Between the AMBER and GROMOS Force Fields

The use of an automation tool such as AmberTools[292] (*i.e.*, *antechamber*[55,133,134] and *tleap*) simplifies the process of topology generation for small organic molecules considerably. In order to use GAFF[55] topologies in GROMOS, they have to be converted to a GROMOS-compatible file format. There are several differences between the AMBER/GAFF and GROMOS[65,66,72] force fields.

First, GAFF is an all-atom force field, whereas most GROMOS (compatible) force fields use united atoms (*i.e.*, implicit hydrogens) for the aliphatic CH_n groups, to reduce the computational cost.[33] The GROMOS force fields are usually parameterized with the simple point-charge

(SPC)[280] water model, whereas the AMBER force-field family is parameterized with the TIP3P[298] water model.[33] A minor difference is the use of different units, *e.g.*, nm, degrees, and kJ mol^{-1} in GROMOS versus Å, radians, and kcal mol^{-1} in AMBER.[227,292]

Second, there are several differences in the potential-energy function, *i.e.*, the functional form of the force fields. Here, the subscripts "A" (AMBER) and "G" (GROMOS) are used to distinguish the terms/parameters of the two force-field families. In AMBER, harmonic bond stretching and bond-angle bending terms are used,[33,55,292]

$$V_{A,i}^{\text{bond,harm}}(d_i) = K_{A,i}^{b,\text{harm}} \left(d_i - d_{0,i}\right)^2 \tag{6.1}$$

$$V_{A,i}^{\text{angle,harm}}(\theta_i) = K_{A,i}^{a,\text{harm}} \left(\theta_i - \theta_{0,i}\right)^2 , \tag{6.2}$$

where d_i is the distance between two bonded atoms, $K_{A,i}^{b,\text{harm}}$ the harmonic bond force constant, $d_{0,i}$ the equilibrium distance, θ_i the angle formed by three bonded atoms, $K_{A,i}$ the harmonic angle force constant, and $\theta_{0,i}$ is the reference bond angle. In GROMOS, harmonic bond stretching and bond-angle bending are also implemented. However, quartic bond stretching and cosine-harmonic bond-angle bending are used by default to increase computational efficiency.[33] They are defined as follows[34]

$$V_{G,i}^{\text{bond,harm}}(d_i) = \frac{1}{2} K_{G,i}^{b,\text{harm}} \left(d_i - d_{0,i}\right)^2 \tag{6.3}$$

$$V_{G,i}^{\text{angle,harm}}(\theta_i) = \frac{1}{2} K_{G,i}^{a,\text{harm}} \left(\theta_i - \theta_{0,i}\right)^2 \tag{6.4}$$

$$V_{G,i}^{\text{bond,quart}}(d_i) = \frac{1}{4} K_{G,i}^{b,\text{quart}} \left(d_i^2 - d_{0,i}^2\right)^2 \tag{6.5}$$

$$V_{G,i}^{\text{angle,cos}}(\theta_i) = \frac{1}{2} K_{G,i}^{a,\text{cos}} \left(\cos(\theta_i) - \cos(\theta_{0,i})\right)^2 , \tag{6.6}$$

with the parameters being defined analogously to the AMBER parameters. It is important to note that for AMBER/GAFF, the factor 1/2 in the harmonic bond stretching and bond-angle bending equations is already included in the force constants $K_{A,i}^{b,\text{harm}}$ and $K_{A,i}^{a,\text{harm}}$ (compare Eqs. (6.1)

and (6.2) with Eqs. (6.3) and (6.4)).[55,292] The harmonic force constants can be converted to the quartic and cosine-harmonic force constants, respectively, as[34]

$$K_{G,i}^{b,\mathrm{quart}} = \frac{K_{G,i}^{b,\mathrm{harm}}}{2d_{0,i}^2} \tag{6.7}$$

$$\begin{aligned} K_{G,i}^{a,\cos} = & \frac{2k_B T_{\mathrm{eff}}}{\left[\cos\left(\theta_{0,i} + \left(\frac{k_B T_{\mathrm{eff}}}{K_{G,i}^{a,\mathrm{harm}}}\right)^{\frac{1}{2}}\right) - \cos\theta_{0,i}\right]^2} \\ & + \frac{2k_B T_{\mathrm{eff}}}{\left[\cos\left(\theta_{0,i} - \left(\frac{k_B T_{\mathrm{eff}}}{K_{G,i}^{a,\mathrm{harm}}}\right)^{\frac{1}{2}}\right) - \cos\theta_{0,i}\right]^2}, \end{aligned} \tag{6.8}$$

where k_B is the Boltzmann constant and T_{eff} is an effective absolute temperature for the conversion (*e.g.*, 300 K).[34] Another difference between AMBER and GROMOS is the potential-energy function used for the out-of-plane distortions. In AMBER, the same function is used for both proper and improper dihedral changes,[33,55,292]

$$V_{A,i}^{\mathrm{tors/imp}}(\theta_i) = K_{A,i}^{\mathrm{tors/imp}}\left[1 + \cos(m\theta_i - \theta_{0,i})\right] . \tag{6.9}$$

In contrast, GROMOS uses different functional forms for proper and improper dihedral changes,[33,34]

$$V_{G,i}^{\mathrm{tors}}(\theta_i) = K_{G,i}^{\mathrm{tors}}\left[1 + \cos(m\theta_i - \theta_{0,i})\right] \tag{6.10}$$

$$V_{G,i}^{\mathrm{imp}}(\xi_i) = \frac{1}{2}K_{G,i}^{\mathrm{imp}}\left(\xi_i - \xi_{0,i}\right)^2 . \tag{6.11}$$

The improper term is also used for out-of-tetrahedron distortions around the CH_1 united atom. Both force-field families use the Lennard-Jones

functional form for the van der Waals interactions[34,55,292]

$$V_{A,ij}^{\mathrm{vdW}} = \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \tag{6.12}$$

$$V_{G,ij}^{\mathrm{vdW}} = \left(\frac{C_{12,ij}}{r_{ij}^{12}} - \frac{C_{6,ij}}{r_{ij}^{6}} \right), \tag{6.13}$$

where r_{ij} is the (minimum-image) distance between atoms i and j, A_{ij} and $C_{12,ij}$ are the repulsion coefficients, and B_{ij} and $C_{6,ij}$ are the dispersion coefficients. There are, however, differences in the combination rules used (geometric in GROMOS and Lorentz-Berthelot in AMBER[33]) and the handling of third-neighbor interactions. In AMBER/GAFF, the Lennard-Jones 1,4-interactions are scaled by a factor 1/2,[55,292] whereas GROMOS force fields contain a special set of parameters[299] (CS_{12} and CS_6) for such interactions, typically involving a reduced repulsion coefficient. Using a scaling factor or reduced interaction parameters for third-neighbor interactions avoids having a too large repulsion in *gauche* conformations (relative to *trans*).[34] In AMBER/GAFF, electrostatic 1,4-interactions are also scaled by a factor 1/1.2.[55,292] Such a scaling is not applied in GROMOS, although in some cases, third neighbors are excluded completely, for example for atoms that are in or attached to an aromatic ring.[299] Furthermore, in AMBER/GAFF, the factor[297]

$$k_e^{1/2} = \left(\frac{1}{4\pi\varepsilon_0} \right)^{1/2} = 18.2223 \text{ [kcal Å mol}^{-1} e^{-2}]^{1/2} \tag{6.14}$$

is included in the atomic charges for computational efficiency. Here, k_e is Coulomb's constant and ε_0 is the permittivity of vacuum.

6.2.2 Relative Hydration Free Energies

The hydration free energy quantifies the free-energy change when a molecule is transferred from gas to water.[170,171] In the present chapter, relative hydration free energies $\Delta\Delta G_{\mathrm{hyd}}$ are used to compare different

free-energy methods against each other and against experimental values (Figure 6.1). For three molecules i, j, and k, it holds that[300]

$$\Delta\Delta G^{ji}_{\mathrm{hyd}} = \Delta G^{j}_{\mathrm{hyd}} - \Delta G^{i}_{\mathrm{hyd}} = \Delta G^{ji}_{\mathrm{wat}} - \Delta G^{ji}_{\mathrm{vac}} \tag{6.15}$$

$$\Delta\Delta G^{ki}_{\mathrm{hyd}} = \Delta G^{k}_{\mathrm{hyd}} - \Delta G^{i}_{\mathrm{hyd}} = \Delta G^{ki}_{\mathrm{wat}} - \Delta G^{ki}_{\mathrm{vac}} \tag{6.16}$$

$$\Delta\Delta G^{kj}_{\mathrm{hyd}} = \Delta\Delta G^{ki}_{\mathrm{hyd}} - \Delta\Delta G^{ji}_{\mathrm{hyd}}\,, \tag{6.17}$$

where $\Delta G^{i}_{\mathrm{hyd}}$ is the hydration free energy of molecule i, $\Delta G^{ji}_{\mathrm{vac}}$ is the free-energy difference between molecules i and j in vacuum, $\Delta G^{ji}_{\mathrm{wat}}$ is the free-energy difference between the two molecules in water, and $\Delta\Delta G^{ji}_{hyd}$ is the hydration free-energy difference between the two molecules (relative hydration free energy).

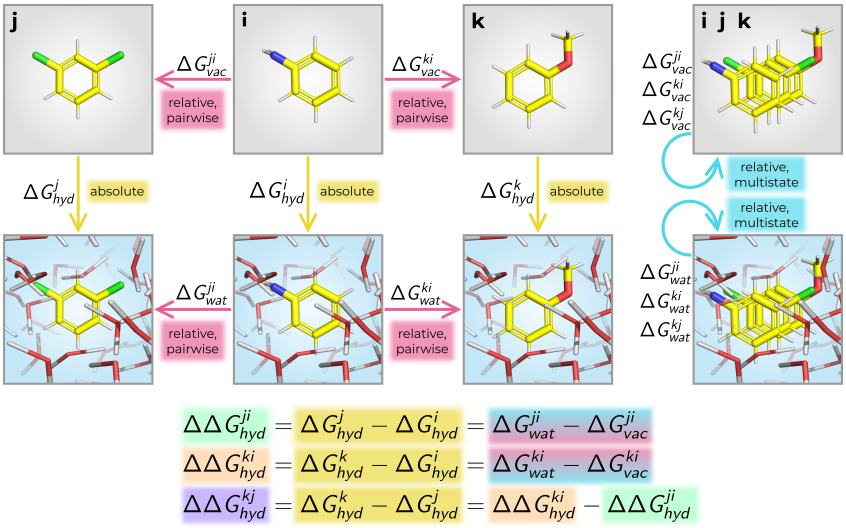


Figure 6.1: Thermodynamic cycle to calculate relative hydration free energies $\Delta\Delta G_{\mathrm{hyd}}$ for three small molecules i (aniline), j (1,3-dichlorobenzene), and k (anisole). Here, $\Delta G^{i}_{\mathrm{hyd}}$ is the hydration free energy of molecule i, $\Delta G^{ji}_{\mathrm{vac}}$ is the free-energy difference between molecules i and j in vacuum, $\Delta G^{ji}_{\mathrm{wat}}$ is the free-energy difference between the two molecules in water, and $\Delta\Delta G^{ji}_{hyd}$ is the hydration free-energy difference between the two molecules (relative hydration free energy). The free-energy difference between two molecules can be calculated from multiple pairwise simulations (as shown on the left, *e.g.*, with TI) or from one simulation with multiple molecules (as shown on the right, *e.g.*, with (RE-)EDS).

In classical MD simulations, hydration free energies are typically calculated with so-called alchemical free-energy methods.[152,154,173,301] Such methods transform a molecule (or its interaction with the environment) into another one *via* nonphysical pathways.[173] The following sections give a brief overview of the two free-energy methods used in the present chapter.

Thermodynamic Integration (TI)

TI is a well-established method to calculate free-energy differences.[156] For two end-states A and B, and using a linear coupling scheme, the potential energy of the system is defined as,

$$V(\mathbf{r};\lambda) = (1-\lambda)\ V_A(\mathbf{r}) + \lambda\ V_B(\mathbf{r})\,. \tag{6.18}$$

At $\lambda = 0$ and $\lambda = 1$, the potential energy corresponds to that of end-state A and end-state B, respectively. This defines a λ-dependent path between the two end-states. After carrying out independent simulations at discrete λ-values between 0 and 1, the free-energy difference between states A and B can be estimated as[156]

$$\Delta G_{BA} = \int_0^1 \left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda\,. \tag{6.19}$$

Replica-Exchange Enveloping Distribution Sampling (RE-EDS)

RE-EDS is a multistate free-energy method,[164–166] which combines Hamiltonian replica exchange (RE)[271,272] with enveloping distribution sampling (EDS).[162,163] In EDS, a reference state V_R is defined based on N end-states as

$$V_R\left(\mathbf{r}; s, \mathbf{E}^R\right) = -\frac{1}{\beta s} \ln\left[\sum_{i=1}^{N} e^{-\beta s\left(V_i(\mathbf{r}) - E_i^R\right)}\right], \tag{6.20}$$

where s is the smoothness parameter ($s > 0$), $\mathbf{E}^R$ is a vector of energy offsets, and $\beta = 1/(k_B\ T)$.

At high s-values (~ 1.0), due to the negative exponent, the end-state with the lowest value of $V_i(\mathbf{r}) - E_i^R$ contributes the most to the sampling of V_R. As s decreases, it "smooths" the potential-energy landscape such that all end-states contribute to the reference state, leading to an unphysical intermediate situation,[164] referred to as "undersampling".[247] The energy offsets control the contribution of the end-states to the reference-state potential. Optimal energy offsets ensure equal weights of all end-states in the reference state.

The free-energy difference between any pair of end-states can then be obtained from a single simulation as[162,163]

$$\Delta G_{BA} = -\frac{1}{\beta} \ln \frac{\langle e^{-\beta(V_B - V_R)} \rangle_R}{\langle e^{-\beta(V_A - V_R} \rangle_R} \,. \tag{6.21}$$

For EDS simulations, it is essential to choose an optimal s-value along with optimal energy offsets to achieve adequate sampling of all end-states, which is non-trivial for more than two end-states.[247] RE-EDS enhances the sampling and simplifies the parameter optimization by simulating several replicas with different s-values in parallel, and performing replica exchanges between them.[164]

In recent studies, RE-EDS was successfully used to calculate relative binding and hydration free energies for molecules containing relatively large structural changes (*i.e.*, R-group modifications and core-hopping transformations such as ring opening/closing and ring size changes).[166,254]

6.3 METHODS

6.3.1 Implementation of amber2gromos

The *amber2gromos* program is a novel C++[189] tool integrated into GROMOS++.[282] It converts AMBER topologies to a GROMOS-compatible

file format. For a given AMBER topology, the program parses the input topology file, converts the force-field parameters, and outputs a GROMOS topology. The following section gives an overview of the conversion steps of the program.

First, the atom names are read and used to initialize the list of atoms in the topology. The atom properties are assigned by parsing the AMBER charges, masses, atom-type indices, residue labels, and residue pointers. The units of the AMBER charges q_A are converted to those in GROMOS according to Eq. (6.14) as[292]

$$q_G = \frac{q_A}{k_e^{1/2}} \,. \tag{6.22}$$

For the masses, no conversion is necessary, as both topology types store them in atomic mass units.[227,292]

Next, the covalent terms are converted. The equilibrium distance of the bond-stretching equation has to be converted from Å to nm.[227,292] The harmonic force constant for the GROMOS topology is calculated as[34]

$$K_G^{b,\mathrm{harm}} = 4.184 \cdot 100 \cdot 2 \cdot K_A^{b,\mathrm{harm}}, \tag{6.23}$$

where the factor $4.184 \cdot 100$ converts kcal mol^{-1} Å^{-2} to kJ mol^{-1} nm^{-2}. The factor 2 compensates for the factor 0.5 included into the force constant in Eq. (6.1). The quartic force constant can then be calculated from the harmonic one *via* Eq. (6.7). The equilibrium angle of the bond-angle bending equation has to be converted from radians to degrees.[227,292] The harmonic bond-angle force constant is converted with[34]

$$K_G^{a,\mathrm{harm}} = 4.184 \cdot \frac{\pi^2}{180^2} \cdot 2 \cdot K_A^{a,\mathrm{harm}} \,. \tag{6.24}$$

Here, the factor $4.184\ \pi^2/180^2$ is used to convert from kcal mol^{-1} rad^{-2} to kJ mol^{-1} deg^{-2}. The factor 2 accounts again for the factor 0.5 included

in the force constant of Eq. (6.2). The cosine-harmonic force constant can then be calculated from the harmonic one *via* Eq. (6.8) using $T_{\text{eff}} = 300$ K. The torsional dihedral periodicities can be used directly, the dihedral phases simply have to be converted from radians to degrees, and the dihedral force constants have to be converted from kcal mol^{-1} to kJ mol^{-1} by multiplication with the factor 4.184. If the fourth atom of a dihedral is negative, the torsion is improper.[302] All improper and torsional dihedrals are converted to torsional dihedrals in the GROMOS topology.

Third, the parameters for the nonbonded interactions are converted. As GROMOS and AMBER use the same Lennard-Jones functional form for the van der Waals interactions, the GROMOS C_6 and C_{12} parameters can be calculated as,

$$C_{12,ij} = 4.184 \cdot 10^{-12} \cdot A_{ij} \tag{6.25}$$

$$C_{6,ij} = 4.184 \cdot 10^{-6} \cdot B_{ij} \tag{6.26}$$

with factors $4.184 \cdot 10^{-12}$ and $4.184 \cdot 10^{-6}$ to convert Å^{12} kcal mol^{-1} to nm^{12} kJ mol^{-1} and Å^{6} kcal mol^{-1} to nm^{6} kJ mol^{-1}, respectively. In AMBER, third-neighbor van der Waals interactions are by default scaled by a factor 1/2,[55,292] whereas GROMOS uses a set of modified CS_6 and CS_{12} parameters that are listed explicitly in the topology file.[34] The modified parameters for the GROMOS topology can therefore be calculated simply as

$$CS_{12,ij} = 0.5 \cdot C_{12,ij} \tag{6.27}$$

$$CS_{6,ij} = 0.5 \cdot C_{6,ij}\,. \tag{6.28}$$

Subsequently, the information about the covalent terms is processed (*i.e.*, bonds with/without hydrogens, angles with/without hydrogens, and dihedrals with/without hydrogens). The order of the atoms in the system is identical in both topologies. However, for run-time efficiency, the listed atom indices correspond to indices into a coordinate array in AMBER

topologies.[297] Therefore, the indices need to be divided by a factor 3 upon conversion to a GROMOS topology. Furthermore, the atom indices are ordered according to GROMOS convention: for bonds, the two atom indices are ordered (*i.e.*, $i < j$ for a bond $i - j$), for angles the first and third indices are ordered (*i.e.*, $i < k$ for an angle $i - j - k$), and for dihedrals the second and third indices are ordered (*i.e.*, $j < k$ for a dihedral $i-j-k-l$).[227] The 1-2, 1-3, and 1-4 exclusion lists are generated *via* the bond, angle, and dihedral relationships of the atoms, respectively.

To enable free-energy calculations, a dummy atom type is added automatically to the list of atom types ($C_{6,i-dummy} = C_{12,i-dummy} = 0$ for all atom types i). The GROMOS++[282] program *prep_eds* was slightly modified to recognize this atom type when creating a perturbation topology for (RE-)EDS simulations.

The process described above generates a valid GROMOS topology from a given AMBER topology. However, there is still a difference between AMBER and GROMOS in the handling of the 1,4-electrostatic interactions (*i.e.*, scaling by a factor 1/1.2[55,292] in AMBER). As 1,4-electrostatic interactions are not scaled in GROMOS, the scaling option is not supported within a GROMOS topology. Therefore, some changes had to be made to the GROMOS[27] source code. A new block type "AMBER" was added to the GROMOS input file, which contains a switch (0 = off, 1 = on) for the scaling of the electrostatic third-neighbor interactions together with the scaling parameter (*e.g.*, 1.2, for scaling by 1/1.2). When the switch is off, the scaling parameter is set to 1.0 so that no scaling is applied. When a reaction-field (RF) correction[114] is used in GROMOS to account for long-range electrostatic interactions, these interactions are calculated as,[33,34]

$$V^{\text{ele}} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\varepsilon_0} \left[\frac{1}{r_{ij}} - \frac{C_{RF} r_{ij}^2}{2R_{RF}^3} - \frac{1 - 0.5 C_{RF}}{R_{RF}} \right], \tag{6.29}$$

where q_i and q_j are the charges of atoms i and j, ε_0 is the permittivity

of vacuum, and R_{RF} is the cutoff distance.[33,34,114,299] Here, C_{RF} is a constant characterizing the effect of the RF continuum, given by[34,114]

$$C_{RF} = \frac{(2\varepsilon_{cs} - 2\varepsilon_{RF})(1 + \kappa_{RF}R_{RF}) - \varepsilon_{RF}(\kappa_{RF}\ R_{RF})^2}{(\varepsilon_{cs} + 2\varepsilon_{RF})(1 + \kappa_{RF}R_{RF}) + \varepsilon_{RF}(\kappa_{RF}\ R_{RF})^2}\,, \tag{6.30}$$

where ε_{cs} is the relative permittivity of the medium in which the simulation is performed, ε_{RF} is the RF permittivity, and κ_{RF} is the inverse Debye screening length.[34] The electrostatic interactions in the GROMOS[27,231] source code were changed to

$$V^{\text{ele}} = \sum_i \sum_{j>i} \frac{q_i q_j}{4\pi\epsilon_0} \left[\frac{s_{ij}}{r_{ij}} - \frac{C_{RF} r_{ij}^2}{2R_{RF}^3} - \frac{1 - 0.5C_{RF}}{R_{RF}} \right], \tag{6.31}$$

where s_{ij} corresponds to the AMBER scaling factor when i and j are third neighbors, and 1.0 otherwise. Note that in this expression, only the direct Coulombic interactions are scaled.[231,303]

As the AMBER force fields do not use charge groups, by default, all the atoms in a molecule/residue are assigned to the same charge group for the simulations with reaction-field electrostatics[114] in GROMOS (which is only appropriate for small molecules). In addition, there is also the option to consider each atom as defining its own charge group (resulting in an atom-based cutoff). The user can also prepare a so-called charge-group file, defining which (consecutive) atoms should be assigned to the same charge group. There are currently ongoing efforts to combine RE-EDS with the shifting-function based scheme developed by Kubincová *et al.*[304] such that an atom-based cutoff could be employed in the reaction-field electrostatics without cutoff artifacts (see Chapter 7). In the near future, this will become the default choice when using the AMBER (and OpenFF) force fields in GROMOS, mitigating the requirement to define charge groups altogether.

An example of an AMBER topology for aniline, and the resulting GROMOS topology translated with *amber2gromos* can be found in Appendix

Sec. 6.A.1 in Listings 6.1 and 6.2, respectively.

6.3.2 RE-EDS Pipeline

Recently, Ries *et al.*[166] presented an improved pipeline to perform relative free-energy calculations with RE-EDS. The RE-EDS pipeline can be divided into three main phases: parameter exploration, parameter optimization, and production. The parameter exploration phase is used to generate relevant configurations for all end-states, to determine a lower bound for the s-values that ensures undersampling, and to obtain initial estimates for the energy offsets. During the parameter optimization, the distribution of s-values and the energy offsets E_i^R are refined such that all end-states are sampled nearly equally (at $s = 1$). Finally, a production run is performed to calculate the relative free-energy differences between all pairs of end-states simultaneously. The entire workflow can be executed using the Python3[211] *reeds* module.[278] In the present chapter, the RE-EDS pipeline was applied to the calculation of relative hydration free energies of benzene derivatives.

6.3.3 Datasets

The main goals of the present chapter were to validate the topology conversion with *amber2gromos* and the accompanying changes in the GROMOS MD engine, as well as to investigate the performance of RE-EDS for systems with many end-states. To this end, two sets of benzene derivatives with experimental hydration free energies in the FreeSolv[155,173] database were selected. This test system was chosen due to the common benzene core, the relatively small size (which limits deviations due to sampling issues), and the availability of calculated and experimental reference data. The potential of RE-EDS to handle larger perturbations has already been demonstrated in previous publications.[166,254]

Set A: Six Benzene Derivatives

As a first test set, six benzene derivatives were selected (Figure 6.2). The number of end-states was deliberately kept small to efficiently test the implementation. The mol2 and frcmod files provided by FreeSolv[155,173] were used to generate AMBER topologies using *tleap* (AmberTools16).[292] The topologies were then converted to GROMOS format using *amber2gromos*, and to GROMACS[25,305] format using ParmEd.[293] The force-field parameters of the original AMBER topologies and of the generated GROMOS/GROMACS topologies were compared manually. In addition, a single energy evaluation was performed for the individual molecules in vacuum/water using both GROMOS and GROMACS. It was verified that the covalent and nonbonded energy terms calculated by the two different MD engines were nearly identical. Longer MD simulations of 5 ns were then carried out for each molecule in vacuum to compare properties such as the system temperature and the different energy terms.

Next, the free-energy differences in vacuum/water were calculated for all 15 molecule pairs from RE-EDS simulations containing six end-states. Complementary single-topology pairwise TI calculations were performed using GROMACS. The relative hydration free energies $\Delta\Delta G^{ji}_{\mathrm{hyd}} = \Delta G^{ji}_{\mathrm{wat}} - \Delta G^{ji}_{\mathrm{vac}}$ were calculated from the RE-EDS and TI calculations. They were compared to the relative hydration free energies $\Delta\Delta G^{ji}_{\mathrm{hyd}} = \Delta G^{j}_{\mathrm{hyd}} - \Delta G^{i}_{\mathrm{hyd}}$ obtained from the experimental and calculated hydration free energies reported in the FreeSolv[155,173] database.

Set B: 28 Benzene Derivatives

To investigate the performance of RE-EDS for a larger number of end-states, the set of benzene derivatives was increased to 28 (Figure 6.3). This extended test set serves as a proof of principle that RE-EDS can be used to calculate free-energy differences for systems with larger numbers of end-states. It is currently the largest set of end-states considered in a RE-EDS simulation. Previous studies involved systems with five,[166] six,[254] nine,[165] and ten[164,254] end-states. The free-energy differences

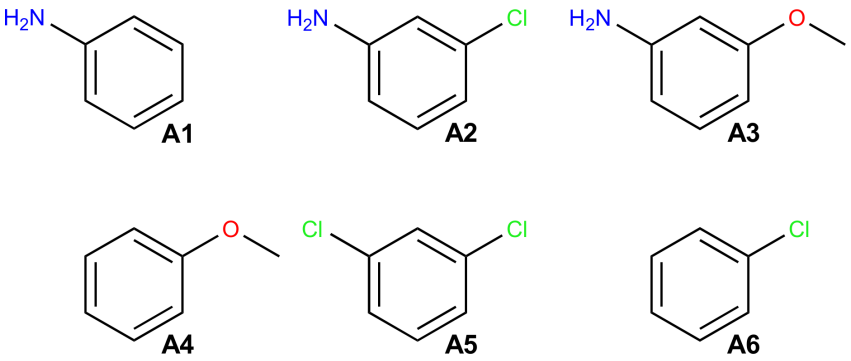


Figure 6.2: Set A consists of six benzene derivatives, selected from the FreeSolv[155,173] database. A table with the molecule indices, the FreeSolv identifiers, the SMILES strings, and the names of the molecules can be found in Table 6.1, top.

were calculated for all 378 molecule pairs in vacuum/water from RE-EDS simulations containing 28 end-states. Analogously to set A, the relative hydration free energies from the RE-EDS calculation were compared to the hydration free energies reported in the FreeSolv[155,173] database.

To further investigate the performance of RE-EDS for such a large set of end-states, set B was subdivided into two smaller subsets, Ba and Bb. Subset Ba consisted of molecules B1 - B14, and subset Bb of molecule B1 together with molecules B15 - B28. For the two subsets, RE-EDS simulations were carried out in vacuum/water to calculate the relative hydration free energies for all end-state pairs in both sets. The relative hydration free energies between the molecule pairs *j*-*k* that were not in the same subset were calculated *via* molecule B1, which was present in both subsets, as

$$\Delta\Delta G_{\mathrm{hyd}}^{kj} = \Delta\Delta G_{\mathrm{hyd}}^{k,B1} - \Delta\Delta G_{\mathrm{hyd}}^{j,B1} \tag{6.32}$$

6.3.4 Simulation Details

The AMBER topologies were generated using AmberTools16[292] and the GAFF 1.7[55] force field with the mol2 and frcmod files provided in the

Table 6.1: Sets A and B. Molecule ID, molecule identifier in the FreeSolv database, SMILES string, and name for the six benzene derivatives of set A (top) and the 28 benzene derivatives of set B (bottom).[155,173]

Set	Molecule	Identifier	SMILES	Name
A	A1	mobley_4883284	c1ccc(cc1)N	aniline
	A2	mobley_6257907	c1cc(cc(c1)Cl)N	3-chloroaniline
	A3	mobley_755351	COc1cccc(c1)N	3-methoxyaniline
	A4	mobley_7295828	COc1ccccc1	anisole
	A5	mobley_1079207	c1cc(cc(c1)Cl)Cl	1,3-dichlorobenzene
	A6	mobley_7608462	c1ccc(cc1)Cl	chlorobenzene
B	B1	mobley_4883284	c1ccc(cc1)N	aniline
	B2	mobley_6257907	c1cc(cc(c1)Cl)N	3-chloroaniline
	B3	mobley_1079207	c1cc(cc(c1)Cl)Cl	1,3-dichlorobenzene
	B4	mobley_2681549	c1c(cc(cc1Cl)Cl)Cl	1,3,5-trichlorobenzene
	B5	mobley_3980099	c1cc(c(c(c1)Cl)Cl)Cl	1,2,3-trichlorobenzene
	B6	mobley_7814642	c1cc(c(cc1Cl)Cl)Cl	1,2,4-trichlorobenzene
	B7	mobley_4553008	c1cc(ccc1Cl)Cl	1,4-dichlorobenzene
	B8	mobley_3183805	Cc1ccc(c(c1)C)C	1,2,4-trimethylbenzene
	B9	mobley_3452749	Cc1cccc(c1C)C	1,2,3-trimethylbenzene
	B10	mobley_1987439	Cc1cc(cc(c1)C)C	mesitylene
	B11	mobley_1424265	Cc1cccc(c1)C	m-xylene
	B12	mobley_3234716	Cc1ccc(cc1)C	p-xylene
	B13	mobley_2925352	Cc1ccc(cc1)O	p-cresol
	B14	mobley_20524	c1ccc(cc1)O	phenol
	B15	mobley_3398536	c1ccc(cc1)I	iodobenzene
	B16	mobley_7608462	c1ccc(cc1)Cl	chlorobenzene
	B17	mobley_4483973	c1ccc(cc1)F	fluorobenzene
	B18	mobley_7599023	c1ccc(cc1)Br	bromobenzene
	B19	mobley_4494568	Cc1ccc(cc1)Br	1-bromo-4-methyl-benzene
	B20	mobley_1873346	Cc1ccccc1	toluene
	B21	mobley_1905088	c1ccc(cc1)CBr	benzyl bromide
	B22	mobley_2484519	c1ccc(cc1)CCl	benzyl chloride
	B23	mobley_8127829	CCc1ccccc1	ethylbenzene
	B24	mobley_1760914	Cc1ccccc1Cl	1-chloro-2-methyl-benzene
	B25	mobley_9478823	Cc1ccccc1C	o-xylene
	B26	mobley_3187514	Cc1ccccc1N	2-methylaniline
	B27	mobley_3169935	c1ccc(c(c1)N)Cl	2-chloroaniline
	B28	mobley_5518547	Cc1ccc(cc1)N	4-methylaniline

FreeSolv[155,173] database as a starting point. The atomic charges were generated using the AM1-BCC[306,307] approach. The input files for the single-topology TI calculations in GROMACS were prepared with FESetup.[263] The input files for the GROMOS RE-EDS simulations were prepared using *amber2gromos* as well as the GROMOS++[282] programs *pdb2g96*, *red_top*, and *prep_eds*. The molecules were aligned for the RE-EDS simulations using the RDKit[210] module *rdFMCS* and the *AllChem.AlignMol* function. The molecule pairs that were aligned to each other were chosen manually to optimize the overlap of the rings and substituents. For some molecule pairs, all atom types were matched (*rdFMCS.AtomCompare.CompareAny*) while for others, only heavy atom types were matched (*rdFMCS.AtomCompare.CompareAnyHeavyAtom*). Since the coordinates of all molecules are present separately in the system, the molecules are in principle able to drift away from each other during a simulation. To

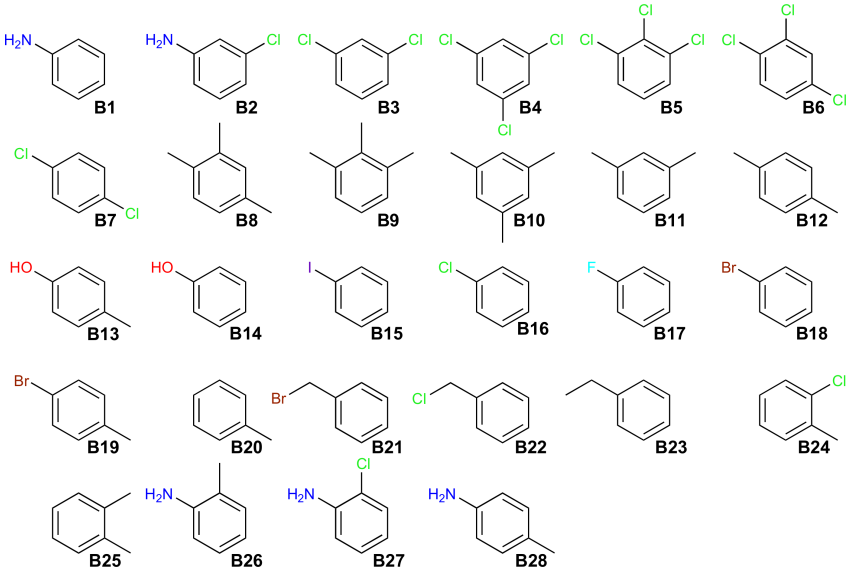


Figure 6.3: Set B consists of 28 benzene derivatives, selected from the FreeSolv[155,173] database. Set B was further subdivided into subset Ba (B1 - B14) and subset Bb (B1 along with B15 - B28). A table with the molecule indices, the FreeSolv identifiers, the SMILES strings, and the name of the molecules can be found in Table 6.1, bottom.

ensure that the molecules remain well-aligned during the whole simulation, atomic distance restraints were applied. The pairwise distance restraints were generated with *RestraintMaker* (see Chapter 5).[254] *RestraintMaker* chooses reference distances r^0 between restrained atoms according to the input alignment. For some molecule pairs, the ring atoms did not perfectly overlap in the initial alignment. The reference distances assigned by *RestraintMaker* were manually set to 0 for those pairs. Four atoms of each molecule were restrained to four atoms of two other molecules, forming a chain of pairwise distance restraints. For set B with 28 molecules, the chain arrangement allowed for relatively large deviations between the molecules furthest apart in the chain. Therefore, additional distance restraints were manually added for four molecule pairs on opposite sides of the chain. The workflow to generate the input files for (RE-)EDS simulations with GAFF[55] parameters is illustrated in Figure 6.4.

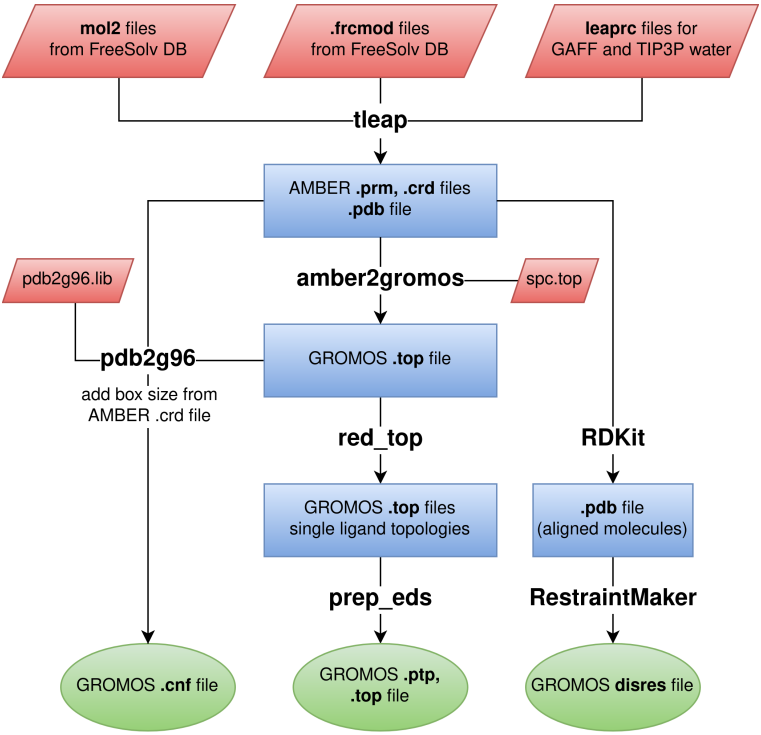


Figure 6.4: Schematic illustration of the RE-EDS input file preparation. The input files (topology, perturbed topology, coordinates and distance restraints) for the (RE-)EDS simulations in GROMOS were created from the frcmod and mol2 files of the FreeSolv[155,173] database. This workflow can easily be extended to also perform the molecule parameterization (*i.e.*, to generate mol2 and frcmod files) using *antechamber* and *parmchk*.[55,133,134]

The RE-EDS simulations were performed with a modified version of GROMOS[27,308] 1.5.0 and the open-source Python3[211] *reeds* module.[278] The TI simulations were performed in GROMACS[25,305] version 2016.6. For the simulations in water, the TIP3P water model[298] was used. A single cutoff radius of 1.2 nm was used for the calculation of the non-bonded interactions. The integration timestep was set to 2 fs and the pairlist was updated every five steps. Long-range nonbonded interactions were calculated in GROMOS using a RF correction[114] with $\varepsilon_{\mathrm{RF}} = 1$ for the simulations in vacuum and $\varepsilon_{\mathrm{RF}} = 78.5$ for the simulations in

water.[309,310] In GROMACS, the long-range nonbonded interactions were calculated using a plain cut-off in vacuum, and smooth Particle Mesh Ewald (SPME)[99,116,118] in water with a grid spacing of 0.1 nm and an interpolation order of 6. To maintain the temperature at 298.15 K and the pressure at 0.06102 kJ mol^{-1} nm^{-3} ($\approx$ 1 atm), a Berendsen thermostat and barostat[45] were used in GROMOS for the simulations in water. For the GROMACS TI calculations and the RE-EDS calculations in vacuum, the leap-frog stochastic dynamics integrator was used, so that no temperature scaling was necessary. In the TI calculations in water, the pressure was kept constant at 1.01325 bar ($\approx$ 1 atm) using a Parrinello-Rahman barostat.[44] All bonds were constrained with the LINCS algorithm[86] (for the TI calculations, 12^{th} order) or the SHAKE algorithm[84] (for the RE-EDS calculations, relative tolerance 10^{-4}), respectively, and harmonic bond-angle bending was employed. In the RE-EDS simulations in GROMOS, the force constant for the distance restraints was set to 5000 kJ/(mol·nm^2).[254]

The calculated hydration free energies reported in the FreeSolv[155,173] database were obtained from alchemical MBAR[158] simulations of 5 ns length at 20 λ-values in GROMACS. In the first five intermediate states, the electrostatic interactions were modified, and in the last 15 states, the Lennard-Jones terms were changed.[155] In the present chapter, the relative hydration free energies for set A were obtained from pairwise TI simulations and from RE-EDS simulations in vacuum/water. For set B, they were determined from RE-EDS simulations in vacuum/water. Here, it is important to note that the three methods used different pathways of the thermodynamic cycle to calculate the relative hydration free energies. Considering Figure 6.1, the MBAR calculations[155,173] correspond to the yellow pathways, the TI calculations to the pink pathways, and the multistate RE-EDS calculations to the blue pathways.

The input files for the RE-EDS simulations can be found at `https://github.com/rinikerlab/reeds/tree/main/examples/systems/benzenes_amber2gromos`.

Set A

To obtain the pairwise free-energy differences in vacuum/water with TI calculations in GROMACS, 21 λ-values were used in vacuum and 27 λ-values in water. In vacuum they were spread in steps of 0.05 between 0 and 1. In water, they were spread in steps of 0.05 between 0.1 and 0.9, and more densely in steps of 0.02 around the extreme values (*i.e.*, between 0 and 0.1, and between 0.9 and 1). The values were spread more densely around the extreme values for the simulations in water to smooth discontinuities in the $\langle \partial V/\partial \lambda \rangle$-curves. Such discontinuities were mainly observed between 0.0 and 0.1, but also occurred between 0.9 and 1.0 for molecule pairs A1 - A2, A1 - A3, and A2 - A3 (Figure 6.9). The masses of the end-state atoms were not perturbed, but kept to the masses of the first end-state. After a short steepest-descent energy minimization with maximally 5000 steps, the systems were equilibrated for 0.5 ns. The free-energy differences were calculated from five independent production runs in vacuum/water. The production runs were 5 ns long at each λ-point.

To generate relevant configurations of all end-states for the RE-EDS calculations, six EDS simulations in vacuum/water of 2 ns length were carried out at $s = 1$. The energy offsets were biased towards one of the end-states (molecules) in each of the simulations by setting one energy offset to 500 kJ mol^{-1} and the others to -500 kJ mol^{-1}. The optimized configurations were used for the starting state mixing (SSM) approach.[166] Simultaneously, 21 EDS simulations of 0.2 ns length with s-values logarithmically distributed between 1 and 10^{-5} were used to determine the lower bound of s for the RE-EDS simulations, which were set to 0.0178 in vacuum and 0.01 in water. RE-EDS simulations of 0.8 ns length were carried out to estimate the energy offsets with 11 replicas in vacuum and 12 replicas in water. Next, the s-distribution was optimized to achieve frequent round trips. Following the SSM approach,[166] the s-values were re-distributed to include a replica with $s = 1$ and optimized initial coordinates for each end-state (in total 12 replicas in vacuum, 13

in water). In vacuum, four replicas were added after one *s*-optimization step of 0.5 ns length. In water, two *s*-optimization steps of 0.5 ns and 1.0 ns lengths, respectively, were used, adding in total eight replicas. To achieve good sampling of all end-states, the initial energy offsets in water were rebalanced over two 0.5 ns RE-EDS simulations. Finally, the free-energy differences were calculated from five independent production runs of 0.5 ns in vacuum (11 replicas) and water (16 replicas). For the production runs, the additional replicas with $s = 1$, which were added during the optimization phase, were removed again.

Set B

For set B, 28 EDS simulations of 2 ns length were carried out to generate optimized configurations. The setup for the lower bound search was analogous to set A, and the values were set to 0.01 in vacuum and 0.0056 in water. For the energy offset estimation, 34 replicas were used in vacuum and 35 in water with 0.8 ns length. The *s*-optimization steps were analogous to set A, adding four replicas in vacuum and eight replicas in water. For set B, rebalancing was also required in vacuum. Both in vacuum and in water, four rebalancing steps were used, of 0.5 ns length each. The production run was 1 ns long in vacuum (34 replicas) and 2 ns long in water (39 replicas). As for set A, the obtained free-energy differences were averaged over five independent production runs.

To generate optimized coordinates, 14 EDS simulations were performed for subset Ba and 15 EDS simulations for subset Bb. The lower bounds for *s* were 0.01 for subset Ba in vacuum, 0.032 for subset Bb in vacuum, and 0.01 for both subsets in water. For the energy offset estimation, 20 (Ba, vacuum), 19 (Bb, vacuum), 20 (Ba, water), and 21 (Bb, water) replicas were used. For both subsets in vacuum, one *s*-optimization step was sufficient. For subset Ba in water, two *s*-optimization steps of 0.5 ns and 1.0 ns, respectively, were necessary, while for subset Bb in water, only one *s*-optimization step was needed. In vacuum, three energy offset rebalancing steps were carried out for both subsets. In water,

four rebalancing steps were used for subset Ba, and three for subset Bb. Finally, the production runs were 1 ns long in vacuum (20 replicas for Ba, 19 replicas for Bb) and 2 ns long in water (24 replicas for Ba, 21 replicas for Bb). Again, the obtained free-energy differences were averaged over five independent production runs.

6.3.5 Analysis

The analysis of the simulations was carried out using GROMOS++[282] and PyGromosTools.[279] The following Python packages were used for visualization and analysis: Matplotlib,[284] mpmath,[287] NumPy,[285] Pandas,[283] SciPy,[286] and Seaborn.[311] For all sets/subsets, the root-mean-square error (RMSE), the mean absolute error (MAE) and the Spearman[289] correlation coefficient was calculated between the different simulation methods and the experimental values.

6.4 RESULTS

6.4.1 Validation of amber2gromos

For the molecules of set A, the manual topology comparison showed that the topologies generated by *amber2gromos* were almost identical to the GROMACS topologies generated by ParmEd[293] (apart from differences in units, functional forms, and slight numerical differences). Apart from negligible numerical differences, the potential-energy terms of the 0^{th} integration step in vacuum and water were identical for the simulations in GROMOS and in GROMACS.

After this initial validation, simulations of 5 ns length were performed in vacuum. The distributions of different energy terms as well as the system temperature were compared. They were all qualitatively similar, with almost identical mean values (Figure 6.5, and Figure 6.15 in Appendix Sec. 6.A.2).

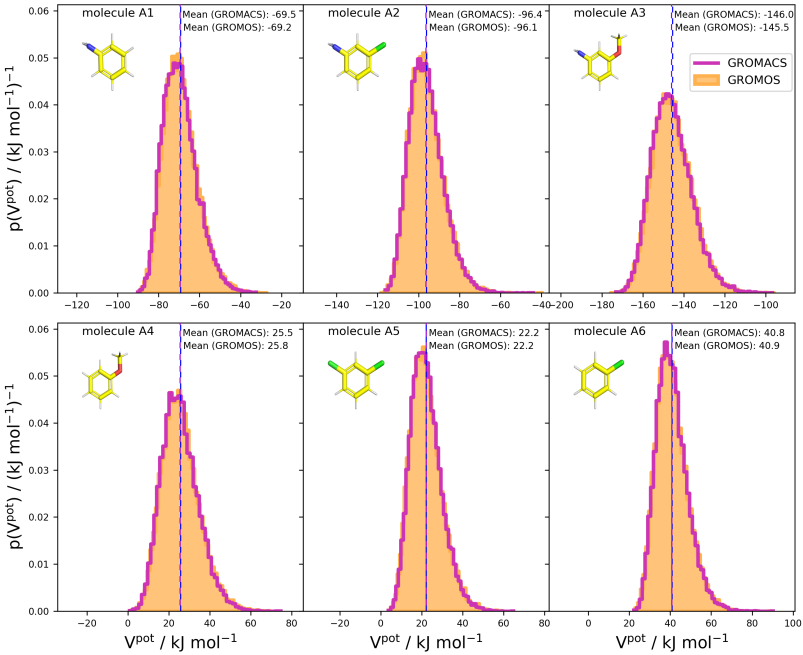


Figure 6.5: Potential-energy distributions of single-molecule simulations of set A based on 5 ns vacuum simulations in GROMOS[27] (orange bars) and GROMACS[25] (pink lines). The topologies for the simulations are based on the AMBER topologies taken from the FreeSolv[155,173] database. The GROMACS topologies were translated with ParmEd[293] and the GROMOS topologies were converted with *amber2gromos*. Energies were written every 100 timesteps (*i.e.*, every 200 fs), and the first 1.25 ns of the simulations were discarded as equilibration.

6.4.2 Calculation of Relative Hydration Free Energies for Set A

For set A, the 15 pairwise free-energy differences in vacuum/water obtained from the RE-EDS calculations were first compared to the ones from the TI calculations in GROMACS. With an RMSE of 1.9 kJ mol^{-1} (vacuum) and 2.0 kJ mol^{-1} (water) and a MAE of 1.6 kJ mol^{-1} (vacuum) and 1.7 kJ mol^{-1} (water), the results agreed well. The Spearman correlation coefficient was 1.00 for the vacuum and the water simulations (Figure 6.6).

Next, the relative hydration free energies from the different sources (*i.e.*,

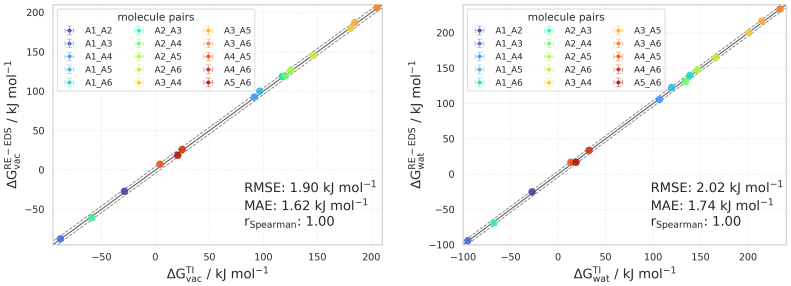


Figure 6.6: Comparison of the free-energy differences of set A. (Left): Comparison of the free-energy differences in vacuum calculated with RE-EDS in GROMOS ($\Delta G_{\text{vac}}^{\text{RE-EDS}}$) and the ones calculated with TI in GROMACS ($\Delta G_{\text{vac}}^{\text{TI}}$). (Right): Comparison of the free-energy differences in water calculated with RE-EDS in GROMOS ($\Delta G_{\text{wat}}^{\text{RE-EDS}}$) and the ones calculated with TI in GROMACS ($\Delta G_{\text{wat}}^{\text{TI}}$).

TI in GROMACS, RE-EDS in GROMOS, MBAR in GROMACS,[155] and experimental[155,173]) were compared. The results from all three simulation methods agreed well with the other calculated results, as well as with the experimental values (Figures 6.7 and 6.8, and Table 6.2). The RMSEs against the experimental results were 2.4 kJ mol^{-1} (TI), 2.4 kJ mol^{-1} (RE-EDS), and 3.1 kJ mol^{-1} (MBAR). The corresponding MAEs were 2.0 kJ mol^{-1} (TI), 2.1 kJ mol^{-1} (RE-EDS), and 2.7 kJ mol^{-1} (MBAR), respectively. The calculated relative hydration free energies also had a high correlation with the experimental results, with Spearman correlation coefficients ranging between 0.93 and 0.94. The full details are provided in Appendix Sec. 6.A.3 in Table 6.4.

While all three free-energy methods achieve comparable and accurate results, the RE-EDS calculations require by far the lowest accumulated simulation time. The total simulation time (pre-processing and production) for the RE-EDS calculations was about 115 ns. For the TI calculations, the total simulation time (equilibration and production) was 3960 ns. This could of course be reduced by calculating only the minimal number of required pairwise free-energy differences (*i.e.*, N-1, which is five for set A). The production time could likely also be reduced to a total of 2 - 3 ns without affecting the convergence significantly. Both measures

Table 6.2: Set A: Overview of statistical metrics (RMSE, MAE, and Spearman correlation coefficients) with respect to the experimental results, and total simulation time for the different free-energy methods. The RE-EDS and TI results were averaged over five independent production runs in vacuum/water and the errors of the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSE and MAE when a random selection of up to four molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time. The full table can be found in Table 6.4 in Appendix Sec. 6.A.3.

	$\Delta\Delta G_{\text{hyd}}^{\text{MBAR 155,173}}$	$\Delta\Delta G_{\text{hyd}}^{\text{TI}}$	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$
RMSE [kJ mol^{-1}]	3.1 ± 0.4	2.4 ± 0.4	2.4 ± 0.3
MAE [kJ mol^{-1}]	2.7 ± 0.3	2.0 ± 0.4	2.1 ± 0.3
r_{Spearman}	0.94	0.93	0.94
$t_{\text{preparation}}$ [ns]	18	360	101.3
$t_{\text{production}}$ [ns]	600	3600	13.5

would reduce the total simulation time required for the TI calculations to about 600 - 840 ns. However, even then, the total simulation time would still be more than five times longer than for the RE-EDS calculations. The calculated values reported in FreeSolv^{155,173} required 618 ns total simulation time (equilibration and production). Also here, the simulation time was chosen with convergence in mind, and could probably be reduced. However, even with 2 - 3 ns production runs, the total simulation time would still be about 2 - 3 times the simulation time required by RE-EDS. Plots of the convergence of the free-energy calculations with RE-EDS as well as the λ-curves for the TI simulations can be found in Figures 6.9 and 6.10.

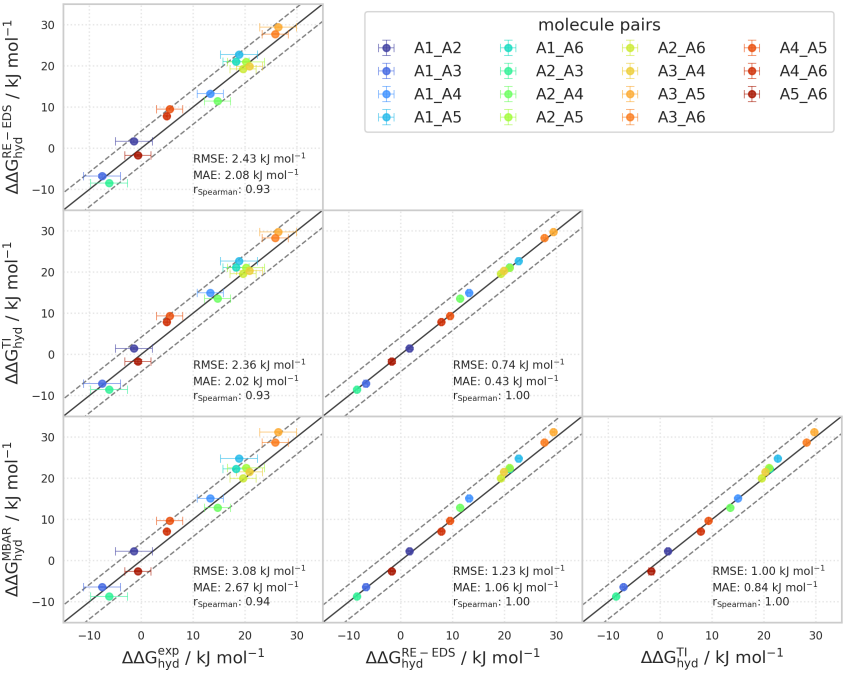


Figure 6.7: Comparison of the relative hydration free energies of set A. The pairwise comparisons of the relative hydration free energies calculated with RE-EDS ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$), TI ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$), MBAR ($\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$)[155,173] and the experimental results.[155,173] The RE-EDS and TI results were averaged over five independent production runs in vacuum/water and the errors of the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The numerical values are provided in Appendix Sec. 6.A.3 in Table 6.4.

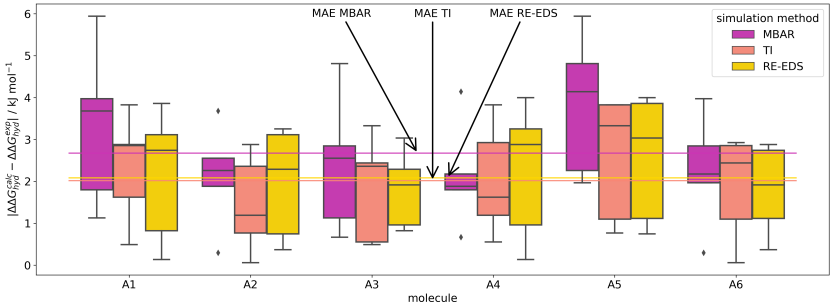


Figure 6.8: Absolute deviations of the calculated relative hydration free energies from experiment for set A. For the relative hydration free energies estimated with MBAR,[155,173] TI and RE-EDS calculations, the spread of the absolute deviation from the reported experimental values[155,173] is shown. The distribution for a molecule i was assembled from the 5 pairwise relative hydration free energies to which it contributed.

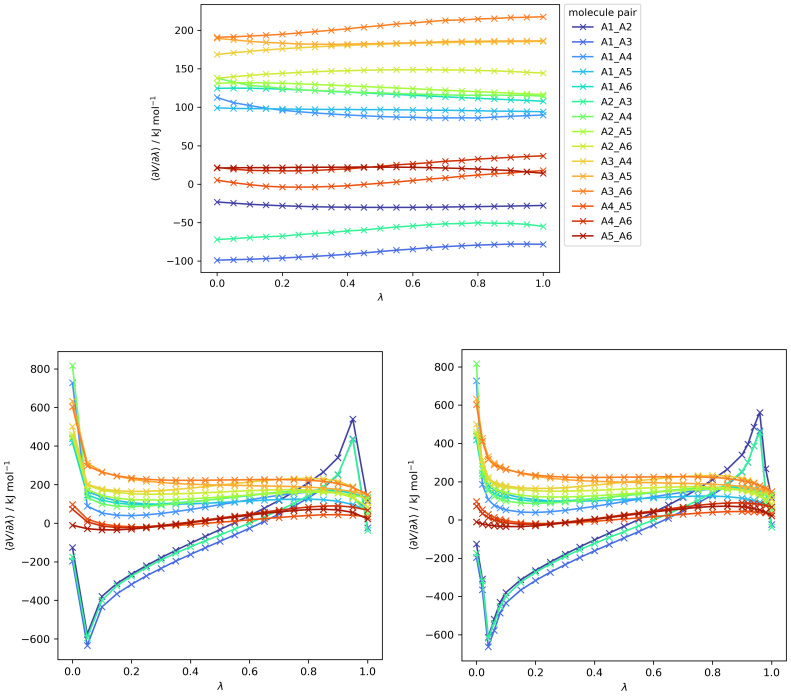


Figure 6.9: $\left\langle \frac{\partial V(\lambda)}{\partial \lambda} \right\rangle_\lambda$ as a function of λ for the 15 pairwise TI calculations in GROMACS of set A in vacuum (top), in water with 21 λ-values (bottom, left), and in water with 27 λ-values (bottom, right) from one of the five independent production runs used to calculate the relative hydration free energies reported in Table 6.4 in Appendix Sec. 6.A.3. Note that while there are still discontinuities in the curves of the simulations in water with 27 λ-values for molecule pairs A1 - A2, A1 - A3, and A2 - A3, the curves are generally smoother, especially in the range between 0.0 and 0.1. The production runs were 5 ns per λ-point.

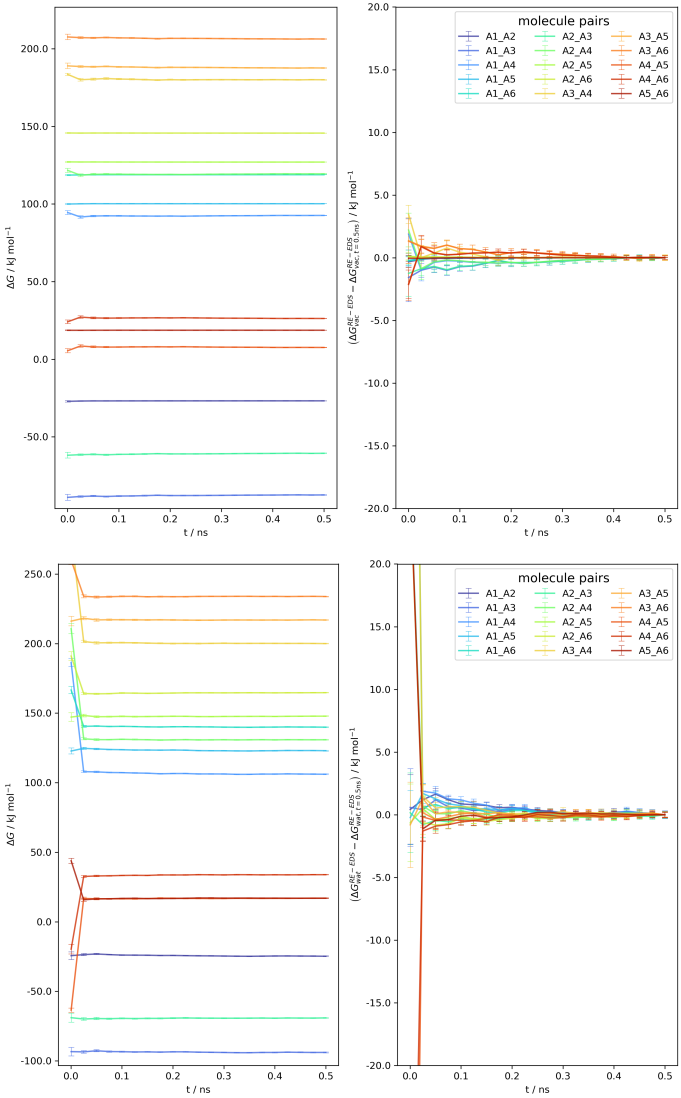


Figure 6.10: Convergence of ΔG_{vac} and ΔG_{wat} as a function of the simulation time for the RE-EDS simulation of set A ($s = 1.0$) in vacuum (top) and in water (bottom) from one of the five independent production runs used to calculate the relative hydration free energies reported in Table 6.4 in Appendix Sec. 6.A.3.

6.4.3 Calculation of Relative Hydration Free Energies for Set B

For set B, we found again an excellent agreement between the results obtained from the RE-EDS simulations and the calculated and experimental values reported in the FreeSolv[155,173] database (Figure 6.11, and Table 6.3). The RMSE against the experimental results was 2.6 kJ mol^{-1} for RE-EDS and 2.0 kJ mol^{-1} for MBAR.[155,173] The corresponding MAEs were 2.2 kJ mol^{-1} and 1.6 kJ mol^{-1}, respectively. The correlation with experiment was high for both methods, with $r^{\text{RE-EDS}}_{\text{Spearman}} = 0.89$, and $r^{\text{MBAR}}_{\text{Spearman}} = 0.92$. The agreement between the two simulation methods was also good, with an RMSE of 1.0 kJ mol^{-1}, a MAE of 0.7 kJ mol^{-1}, and a Spearman correlation coefficient of 0.99. For RE-EDS, molecule pairs B1 - B4, B4 - B8, B4 - B9, B4 - B23, B4 - B25, B6 - B8, and B6 - B9 showed deviations above 5.5 kJ mol^{-1}, the largest for molecule pair B4 - B9 with 6.7 kJ mol^{-1} (Figure 6.12). For MBAR, molecule pairs B1 - B4, B1 - B6, B4 - B8, B4 - B25, B4 - B27, and B4 - B28 deviated by more than 4.3 kJ mol^{-1} from experiment. Here, the highest absolute deviation was observed for molecule pair B1 - B4 with 4.9 kJ mol^{-1}. The Spearman correlation coefficient of the absolute deviations from experiment for the two simulation methods was relatively high at 0.88, indicating that some of the deviations might be related to shortcomings in the force field or in the experimental determination.

To investigate the efficiency and accuracy of RE-EDS free-energy calculations for a larger number of molecules, the RE-EDS results of set B were compared to the ones obtained from RE-EDS simulations of subsets Ba and Bb (Figures 6.11 and 6.12, and Table 6.3). With an RMSE against the experimental values[155,173] of 2.4 kJ mol^{-1}, a MAE of 2.0 kJ mol^{-1}, and a Spearman correlation coefficient of 0.90, the combined results from the two separate RE-EDS pipelines (*i.e.*, Ba and Bb) were marginally more accurate. The agreement with the MBAR results[155,173] was slightly higher with an RMSE of 0.8 kJ mol^{-1} and a MAE of 0.5 kJ mol^{-1}. The

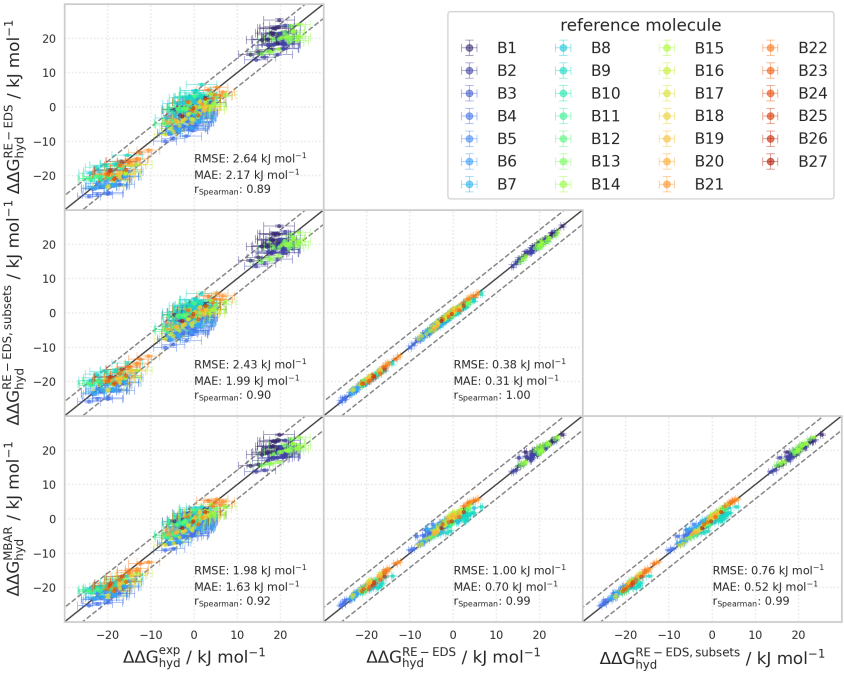


Figure 6.11: Comparison of the relative hydration free energies of set B. The pairwise comparison of the relative hydration free energies calculated with RE-EDS ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$), RE-EDS on the combined subsets Ba and Bb ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,subsets}}$), MBAR ($\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$) [155,173] and the experimental results. [155,173] The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The $\Delta\Delta G_{\text{hyd}}^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The RE-EDS results were averaged over five independent production runs in vacuum/water and the errors of the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The numerical values are provided in Appendix Sec. 6.A.4 in Table 6.5.

RMSE between the RE-EDS results for set B versus subsets Ba and Bb was 0.4 kJ mol^{-1} with perfect correlation ($r_{\text{Spearman}} = 1.00$). The slightly higher agreement of the results obtained from the two subsets with the experimental and MBAR results indicates that there is a small "diffusion effect" for this system when more molecules are added to the simulation. As more end-states are added to a system, the number of frames where an end-state contributes maximally to the reference state

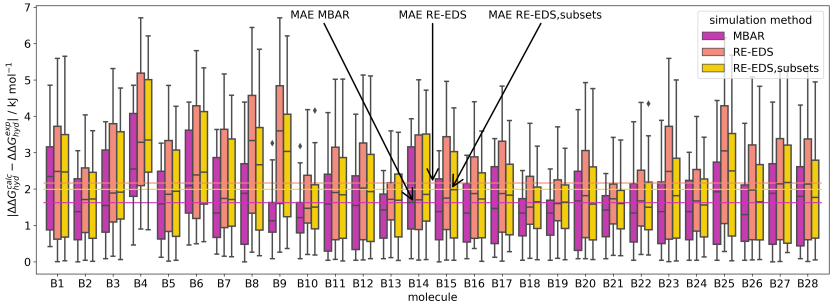


Figure 6.12: Absolute deviation of calculated relative hydration free energies from experiment for set B. For the relative hydration free energies estimated with MBAR,[155,173] the full RE-EDS simulations and the RE-EDS calculations on subsets Ba and Bb, the spread of the absolute deviation from the reported experimental values[155,173] is shown. The distribution for a molecule i was assembled from the 27 pairwise relative hydration free energies to which it contributed.

Table 6.3: Set B: Overview of statistical metrics (RMSE, MAE, and Spearman correlation coefficients) with respect to the experimental results, and total simulation time for the different free-energy methods. The RE-EDS and TI results were averaged over five independent production runs in vacuum/water and the errors of the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. For RE-EDS, both the results for the full set B and for the combined subsets Ba and Bb are reported. The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSE and MAE when a random selection of up to 26 molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time. The full table can be found in Appendix Sec. 6.A.4 in Table 6.5.

	$\Delta\Delta G_{hyd}^{MBAR\ 155,173}$	$\Delta\Delta G_{hyd}^{RE\text{-}EDS}$	$\Delta\Delta G_{hyd}^{RE\text{-}EDS,subsets}$
RMSE [kJ mol^{-1}]	2.0 ± 0.2	2.6 ± 0.3	2.4 ± 0.3
MAE [kJ mol^{-1}]	1.6 ± 0.2	2.2 ± 0.2	2.0 ± 0.2
$r_{Spearman}$	0.92	0.89	0.90
$t_{preparation}$ [ns]	84 ns	549 ns	514 ns
$t_{production}$ [ns]	2800 ns	112 ns	129 ns

decreases. Additionally, more s-values are required to obtain frequent round-trips and more energy offset rebalancing iterations are needed to obtain approximately equal sampling of all the end-states.

Also here, the relatively small simulation time required for the RE-EDS calculations can be highlighted. The total simulation time for the RE-EDS simulations of set B was about 661 ns, compared to 2884 ns for MBAR.[155,173] If pairwise TI simulations in vacuum/water had been used as for set A, the total simulation time would have been between

7128 ns (minimal 27 pairs = N-1) and 99627 ns (all pairs). The total simulation time for the RE-EDS pipelines of subsets Ba and Bb combined was about 643 ns. This is slightly shorter than the simulation length for the full set B, mainly due to the fact that for subset Bb in water, only four instead of eight replicas were added during the s-optimization, and one less rebalancing step was required for both subsets in vacuum and for Bb in water. Plots of the convergence of the free-energy calculations with RE-EDS can be found in Figures 6.13 and 6.14.

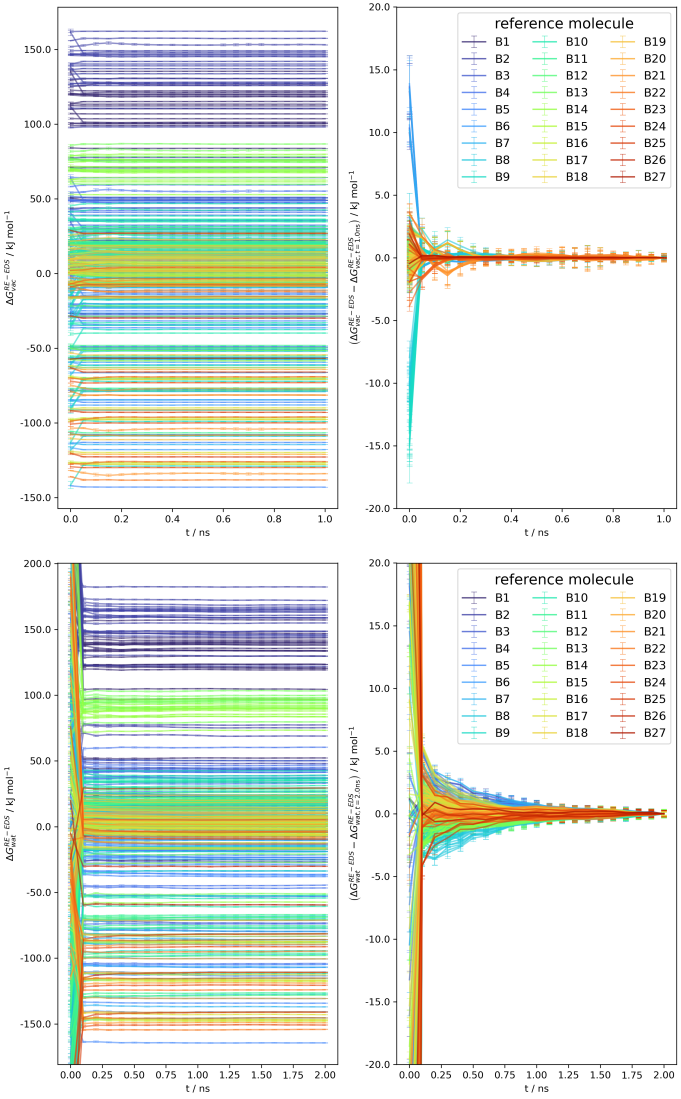


Figure 6.13: Convergence of $\Delta G_{\text{vac}}^{\text{RE-EDS}}$ and $\Delta G_{\text{wat}}^{\text{RE-EDS}}$ as a function of the simulation time for the RE-EDS simulation of set B ($s = 1.0$) in vacuum (top) and in water (bottom) from one of the five independent production runs used to calculate the relative hydration free energies reported in Table 6.5 in Appendix Sec. 6.A.3.

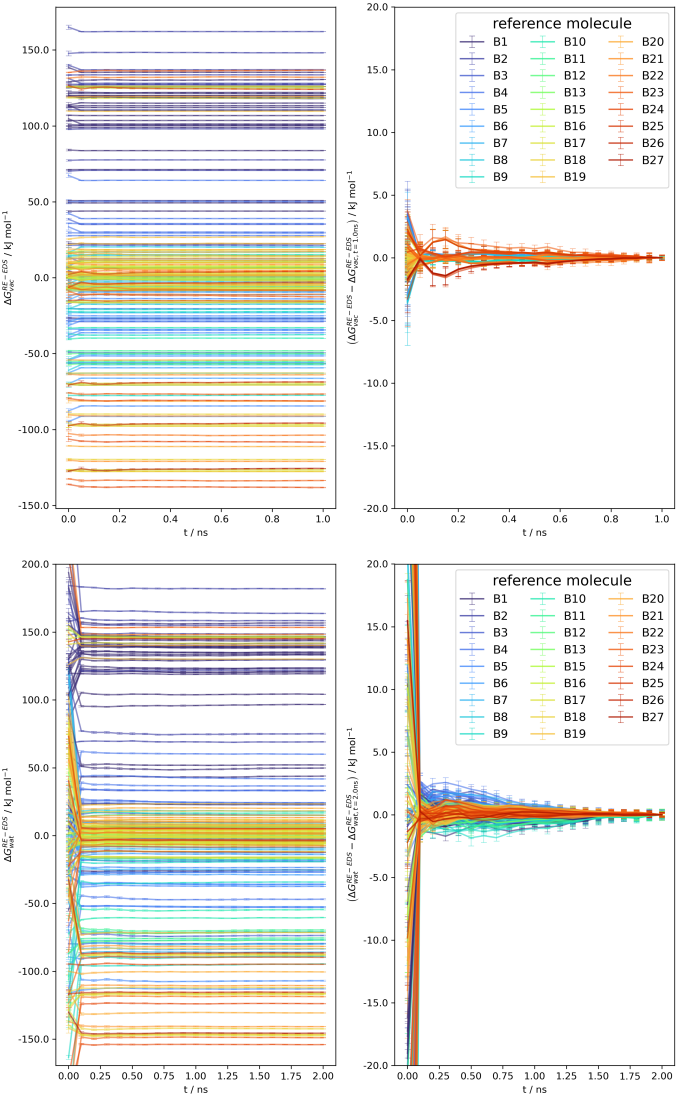


Figure 6.14: Convergence of $\Delta G_{\text{vac}}^{\text{RE-EDS,subsets}}$ and $\Delta G_{\text{wat}}^{\text{RE-EDS,subsets}}$ as a function of the simulation time for the RE-EDS simulation of subsets Ba and Bb ($s = 1.0$) in vacuum (top) and in water (bottom) from one of the five independent production runs used to calculate the relative hydration free energies reported in Table 6.5 in Appendix Sec. 6.A.3.

6.5 CONCLUSION

In the present chapter, the GROMOS++ program *amber2gromos* was introduced to convert topologies from the AMBER to the GROMOS file format. An overview of the differences between AMBER and GROMOS force fields was presented together with a description of the conversion of the AMBER topology parameters to GROMOS topology parameters, and the necessary slight modification to the GROMOS source code. A workflow was outlined to prepare topology, coordinate, and distance restraint input files for RE-EDS free-energy calculations with GAFF parameters in GROMOS. The extension of this workflow to the OpenFF family of force fields is straightforward.

Two sets of benzene derivatives were selected from the FreeSolv database with six (set A) and 28 molecules (set B). Set A was used to validate the implementation of *amber2gromos* and the related source-code changes to the GROMOS MD engine. The generated GROMOS topologies for the six benzene derivatives were compared to GROMACS topologies generated by ParmEd from the same AMBER topologies. Single-molecule simulations in vacuum performed in GROMOS and in GROMACS showed nearly identical energy and temperature distributions.

Finally, relative hydration free energies were calculated for both sets. For set A, both TI and RE-EDS simulations were carried out in vacuum/water to estimate the 15 pairwise free-energy differences. These results were compared to the relative hydration free energies obtained from experiment as well as the hydration free energies reported in the FreeSolv database (calculated with MBAR). Overall, an excellent agreement was observed between the different free-energy methods and with experiment. While all methods delivered highly accurate results, the RE-EDS calculations required the least amount of total simulation time. The system size was increased to 28 molecules in set B to challenge the RE-EDS pipeline. Again, the results agreed well with the ones from MBAR and with the experimental values.

To test if it is more efficient to use a large set of molecules or two subsets with a shared molecule, set B was divided into two subsets Ba (molecules B1 - B14) and Bb (molecule B1 and molecules B15 - B28). While both the results and simulation time of the two RE-EDS approaches were almost identical, smaller subsets may offer some advantages in practice. RE-EDS simulations are in principle highly parallelizable, as large parts of both the replicas and the interactions within the replicas can be carried out independently with relatively infrequent communication. Nevertheless, as more molecules/replicas are added to the system, the wall-clock time of the simulations increases (more interactions to calculate, larger communication overhead, more replicas). Using two subsets decreased the elapsed real-time of the RE-EDS pipeline. Further research will be needed to determine optimal splits of datasets into subsets as well as the choice of the common molecule(s). The aim will be to find a balance between avoiding a diffusion effect from too many end-states in one system, and error propagation due to too small subsets or sub-optimal common molecule(s).

Overall, it has been shown that hydration free-energy calculations with RE-EDS and GAFF parameters executed in GROMOS accurately reproduce both experimental values and results obtained with different free-energy estimators and MD engines. While the molecules of the chosen datasets were relatively small and contained a well-defined common benzene core, previous studies successfully used RE-EDS to calculate binding and hydration free energies for molecule sets involving larger structural changes such as R-group modifications, ring opening/closing and ring size changes.

6.A APPENDIX

6.A.1 Example of an amber2gromos Translation

Listing 6.1: AMBER topology for aniline. The topology was created with AmberTools, [55,133,134,292] using parameters from GAFF. [55]

```
%VERSION  VERSION_STAMP = V0001.000  DATE = 02/11/21  17:57:11
%FLAG TITLE
%FORMAT(20a4)
aniline
%FLAG POINTERS
%FORMAT(10I8)
      14       4       7       7      13       8      26       9       0       0
      63       1       7       8       9       4       5       3       4       0
       0       0       0       0       0       0       0       0      14       0
       0
%FLAG ATOM_NAME
%FORMAT(20a4)
C1  C2  C3  C4  C5  C6  N1  H1  H2  H3  H4  H5  H6  H7
%FLAG CHARGE
%FORMAT(5E16.8)
 -3.15063567E+00 -1.70378505E+00 -3.45859254E+00  2.47094388E+00 -3.45859254E+00
 -1.70378505E+00 -1.48985525E+01  2.36525454E+00  2.35614339E+00  2.37618792E+00
  2.37618792E+00  2.35614339E+00  7.03927449E+00  7.03927449E+00
%FLAG ATOMIC_NUMBER
%FORMAT(10I8)
       6       6       6       6       6       6       7       1       1       1
       1       1       1       1
%FLAG MASS
%FORMAT(5E16.8)
  1.20100000E+01  1.20100000E+01  1.20100000E+01  1.20100000E+01  1.20100000E+01
  1.20100000E+01  1.40100000E+01  1.00800000E+00  1.00800000E+00  1.00800000E+00
  1.00800000E+00  1.00800000E+00  1.00800000E+00  1.00800000E+00
%FLAG ATOM_TYPE_INDEX
%FORMAT(10I8)
       1       1       1       1       1       1       2       3       3       3
       3       3       4       4
%FLAG NUMBER_EXCLUDED_ATOMS
%FORMAT(10I8)
      10       9      10       9       8       5       4       2       1       1
       1       1       1       1
%FLAG NONBONDED_PARM_INDEX
%FORMAT(10I8)
       1       2       4       7       2       3       5       8       4       5
       6       9       7       8       9      10
%FLAG RESIDUE_LABEL
%FORMAT(20a4)
LI1
%FLAG RESIDUE_POINTER
%FORMAT(10I8)
       1
%FLAG BOND_FORCE_CONSTANT
%FORMAT(5E16.8)
  4.78400000E+02  3.44300000E+02  4.49000000E+02  4.01200000E+02
%FLAG BOND_EQUIL_VALUE
%FORMAT(5E16.8)
  1.38700000E+00  1.08700000E+00  1.36400000E+00  1.01400000E+00
```

```
%FLAG ANGLE_FORCE_CONSTANT
%FORMAT(5E16.8)
  6.71800000E+01  4.84600000E+01  6.93400000E+01  4.90800000E+01  4.00500000E+01
%FLAG ANGLE_EQUIL_VALUE
%FORMAT(5E16.8)
  2.09387240E+00  2.09457053E+00  2.09666493E+00  2.02685173E+00  2.00451150E+00
%FLAG DIHEDRAL_FORCE_CONSTANT
%FORMAT(5E16.8)
  3.62500000E+00  1.05000000E+00  1.10000000E+00
%FLAG DIHEDRAL_PERIODICITY
%FORMAT(5E16.8)
  2.00000000E+00  2.00000000E+00  2.00000000E+00
%FLAG DIHEDRAL_PHASE
%FORMAT(5E16.8)
  3.14159400E+00  3.14159400E+00  3.14159400E+00
%FLAG SCEE_SCALE_FACTOR
%FORMAT(5E16.8)
  1.20000000E+00  1.20000000E+00  0.00000000E+00
%FLAG SCNB_SCALE_FACTOR
%FORMAT(5E16.8)
  2.00000000E+00  2.00000000E+00  0.00000000E+00
%FLAG SOLTY
%FORMAT(5E16.8)
  0.00000000E+00  0.00000000E+00  0.00000000E+00  0.00000000E+00
%FLAG LENNARD_JONES_ACOEF
%FORMAT(5E16.8)
  8.19971662E+05  8.82619071E+05  9.44293233E+05  7.62451550E+04  7.91627154E+04
  5.71629601E+03  2.27577561E+03  2.12601181E+03  8.90987508E+01  1.39982777E-01
%FLAG LENNARD_JONES_BCOEF
%FORMAT(5E16.8)
  5.31102864E+02  6.53361429E+02  8.01323529E+02  1.04660679E+02  1.26451907E+02
  1.85196588E+01  1.82891803E+01  2.09604198E+01  2.33864085E+00  9.37598976E-02
%FLAG BONDS_INC_HYDROGEN
%FORMAT(10I8)
       0      21       2       3      24       2       6      27       2      12
      30       2      15      33       2      18      36       4      18      39
       4
%FLAG BONDS_WITHOUT_HYDROGEN
%FORMAT(10I8)
       0      15       1       0       3       1       3       6       1       6
       9       1       9      12       1       9      18       3      12      15
       1
%FLAG ANGLES_INC_HYDROGEN
%FORMAT(10I8)
       0      15      33       2       0       3      24       2       3       0
      21       2       3       6      27       2       6       3      24       2
       9       6      27       2       9      12      30       2       9      18
      36       4       9      18      39       4      12      15      33       2
      15       0      21       2      15      12      30       2      36      18
      39       5
%FLAG ANGLES_WITHOUT_HYDROGEN
%FORMAT(10I8)
       0      15      12       1       0       3       6       1       3       0
      15       1       3       6       9       1       6       9      12       1
       6       9      18       3       9      12      15       1      12       9
      18       3
%FLAG DIHEDRALS_INC_HYDROGEN
%FORMAT(10I8)
       0      15      12      30       1       0       3       6      27       1
       3       0      15      33       1      21       0       3       6       1
       6       9      12      30       1       6       9      18      36       2
       6       9      18      39       2       9       6       3      24       1
       9      12      15      33       1      12       9       6      27       1
```

```
      12       9      18      36       2      12       9      18      39       2
      21       0      15      12       1      15       0       3      24       1
      18       9       6      27       1      18       9      12      30       1
      21       0      15      33       1      21       0       3      24       1
      24       3       6      27       1      30      12      15      33       1
      21       0     -15      -3       3       0       6      -3     -24       3
       3       9      -6     -27       3       9      15     -12     -30       3
       0      12     -15     -33       3       9      36     -18     -39       3
%FLAG DIHEDRALS_WITHOUT_HYDROGEN
%FORMAT(10I8)
       0      15      12       9       1       0       3      -6       9       1
       3       0      15      12       1       3       6      -9      12       1
       3       6       9      18       1      15       0       3       6       1
       6       9     -12      15       1      15      12       9      18       1
       6      12      -9     -18       3
%FLAG EXCLUDED_ATOMS_LIST
%FORMAT(10I8)
       2       3       4       5       6       8       9      10      11      12
       3       4       5       6       7       8       9      10      12       4
       5       6       7       8       9      10      11      13      14       5
       6       7       9      10      11      12      13      14       6       7
       8      10      11      12      13      14       7       8       9      11
      12      10      11      13      14       9      12      10       0      12
       0      14       0
%FLAG HBOND_ACOEF
%FORMAT(5E16.8)

%FLAG HBOND_BCOEF
%FORMAT(5E16.8)

%FLAG HBCUT
%FORMAT(5E16.8)

%FLAG AMBER_ATOM_TYPE
%FORMAT(20a4)
ca  ca  ca  ca  ca  ca  nh  ha  ha  ha  ha  hn  hn
%FLAG TREE_CHAIN_CLASSIFICATION
%FORMAT(20a4)
BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA BLA
%FLAG JOIN_ARRAY
%FORMAT(10I8)
       0       0       0       0       0       0       0       0       0       0
       0       0       0       0
%FLAG IROTAT
%FORMAT(10I8)
       0       0       0       0       0       0       0       0       0       0
       0       0       0       0
%FLAG RADIUS_SET
%FORMAT(1a80)
modified Bondi radii (mbondi)
%FLAG RADII
%FORMAT(5E16.8)
  1.70000000E+00  1.70000000E+00  1.70000000E+00  1.70000000E+00  1.70000000E+00
  1.70000000E+00  1.55000000E+00  1.30000000E+00  1.30000000E+00  1.30000000E+00
  1.30000000E+00  1.30000000E+00  1.30000000E+00  1.30000000E+00
%FLAG SCREEN
%FORMAT(5E16.8)
  7.20000000E-01  7.20000000E-01  7.20000000E-01  7.20000000E-01  7.20000000E-01
  7.20000000E-01  7.90000000E-01  8.50000000E-01  8.50000000E-01  8.50000000E-01
  8.50000000E-01  8.50000000E-01  8.50000000E-01  8.50000000E-01
%FLAG IPOL
%FORMAT(1I8)
       0
```

Listing 6.2: GROMOS topology for aniline. The topology was converted with *amber2gromos* from the AMBER topology in Listing 6.1.

```
TITLE
 AMBER topology translated into GROMOS
END
PHYSICALCONSTANTS
# FPEPSI: 1.0/(4.0*PI*EPS0) (EPS0 is the permittivity of vacuum)
138.9354846
# HBAR: Planck's constant HBAR = H/(2* PI)
0.0635078077
# SPDL: Speed of light (nm/ps)
299792.458
# BOLTZ: Boltzmann's constant kB
0.008314511212
END
TOPVERSION
2.0
END
ATOMTYPENAME
# NRATT: number of van der Waals atom types
5
# TYPE: atom type names
ca
nh
ha
hn
DUM
END
RESNAME
# NRAA2: number of residues in a solute molecule
1
# AANM: residue names
LI1
END
SOLUTEATOM
#   NRP: number of solute atoms
   14
#  ATNM: atom number
#  MRES: residue number
#  PANM: atom name of solute atom
#   IAC: integer (van der Waals) atom type code
#  MASS: mass of solute atom
#    CG: charge of solute atom
#   CGC: charge group code (0 or 1)
#   INE: number of excluded atoms
# INE14: number of 1-4 interactions
# ATNM MRES PANM IAC     MASS       CG  CGC INE
#                                      INE14
     1    1   C1   1 12.01000 -0.17290  0     7     2     3     5     6     8     9
                                               12
                                        3     4    10    11
     2    1   C2   1 12.01000 -0.09350  0     6     3     4     6     8     9    10
                                        3     5     7    12
     3    1   C3   1 12.01000 -0.18980  0     5     4     5     7     9    10
                                        5     6     8    11    13    14
     4    1   C4   1 12.01000  0.13560  0     7     5     6     7    10    11    13
                                              14
                                        2     9    12
     5    1   C5   1 12.01000 -0.18980  0     4     6     7    11    12
                                        4     8    10    13    14
     6    1   C6   1 12.01000 -0.09350  0     3     8    11    12
                                        2     7     9
```

```
     7    1   N1    2 14.01000 -0.81760  0     2    13    14
                                             2    10    11
     8    1   H1    3  1.00800  0.12980  0     0
                                             2     9    12
     9    1   H2    3  1.00800  0.12930  0     0
                                             1    10
    10    1   H3    3  1.00800  0.13030  0     0
                                             0
    11    1   H4    3  1.00800  0.13030  0     0
                                             1    12
    12    1   H5    3  1.00800  0.12930  0     0
                                             0
    13    1   H6    4  1.00800  0.38630  0     1    14
                                             0
    14    1   H7    4  1.00800  0.38630  1     0
                                             0
END
BONDSTRETCHTYPE
#  NBTY: number of covalent bond types
4
#  CB:  quartic force constant
#  CHB: harmonic force constant
#  BO:  bond length at minimum energy
#         CB          CHB          BO
     1.04047e+07     4.00325e+05     1.38700e-01
     1.21918e+07     2.88110e+05     1.08700e-01
     1.00974e+07     3.75723e+05     1.36400e-01
     1.63259e+07     3.35724e+05     1.01400e-01
END
BONDH
#  NBONH: number of bonds involving H atoms in solute
7
#  IBH, JBH: atom sequence numbers of atoms forming a bond
#  ICBH: bond type code
#   IBH    JBH ICBH
      1      8    2
      2      9    2
      3     10    2
      5     11    2
      6     12    2
      7     13    4
      7     14    4
END
BOND
#  NBON: number of bonds NOT involving H atoms in solute
7
#  IB, JB: atom sequence numbers of atoms forming a bond
#  ICB: bond type code
#    IB     JB  ICB
      1      2    1
      1      6    1
      2      3    1
      3      4    1
      4      5    1
      4      7    3
      5      6    1
END
BONDANGLEBENDTYPE
#  NTTY: number of bond angle types
5
#  CT:  force constant (based on potential
#       harmonic in the angle cosine)
#  CHT: force constant (based on potential
```

```
#       harmonic in the angle)
#  T0:  bond angle at minimum energy in degrees
#         CT          CHT          T0
     7.49930e+02     1.71244e-01     1.19970e+02
     5.41626e+02     1.23526e-01     1.20010e+02
     7.76516e+02     1.76750e-01     1.20130e+02
     5.10377e+02     1.25107e-01     1.16130e+02
     4.07869e+02     1.02089e-01     1.14850e+02
END
BONDANGLEH
#  NTHEH: number of bond angles involving H atoms in solute
13
#  ITH, JTH, KTH: atom sequence numbers
#    of atoms forming a bond angle in solute
#  ICTH: bond angle type code
#   ITH    JTH    KTH ICTH
      2      1      8    2
      6      1      8    2
      1      2      9    2
      3      2      9    2
      2      3     10    2
      4      3     10    2
      4      5     11    2
      6      5     11    2
      1      6     12    2
      5      6     12    2
# 10
      4      7     13    4
      4      7     14    4
     13      7     14    5
END
BONDANGLE
#  NTHE: number of bond angles NOT
#        involving H atoms in solute
8
#  IT, JT, KT: atom sequence numbers of atoms
#     forming a bond angle
#  ICT: bond angle type code
#    IT     JT     KT  ICT
      2      1      6    1
      1      2      3    1
      2      3      4    1
      3      4      5    1
      3      4      7    3
      5      4      7    3
      4      5      6    1
      1      6      5    1
END
IMPDIHEDRALTYPE
#  NQTY: number of improper dihedrals
1
#  CQ: force constant of improper dihedral per degrees square
#  Q0: improper dihedral angle at minimum energy in degrees
#           CQ            Q0
    5.10000e-02    0.00000e+00
END
IMPDIHEDRALH
#  NQHIH: number of improper dihedrals
#         involving H atoms in the solute
0
#  IQH,JQH,KQH,LQH: atom sequence numbers
#     of atoms forming an improper dihedral
#  ICQH: improper dihedral type code
```

```
#   IQH    JQH    KQH    LQH ICQH
END
IMPDIHEDRAL
#  NQHI: number of improper dihedrals NOT
#    involving H atoms in solute
0
#  IQ,JQ,KQ,LQ: atom sequence numbers of atoms
#    forming an improper dihedral
#  ICQ: improper dihedral type code
#    IQ     JQ     KQ     LQ  ICQ
END
TORSDIHEDRALTYPE
#  NPTY: number of dihedral types
3
#  CP: force constant
#  PD: phase-shift angle
#  NP: multiplicity
#       CP          PD  NP
  15.16700  180.00008   2
   4.39320  180.00008   2
   4.60240  180.00008   2
END
DIHEDRALH
#  NPHIH: number of dihedrals involving H atoms in solute
26
#  IPH, JPH, KPH, LPH: atom sequence numbers
#    of atoms forming a dihedral
#  ICPH: dihedral type code
#   IPH    JPH    KPH    LPH ICPH
      6      1      2      9    1
      8      1      2      3    1
      8      1      2      9    1
      2      1      6     12    1
      8      1      6      2    3
      8      1      6      5    1
      8      1      6     12    1
      1      2      3     10    1
      9      2      3      1    3
      9      2      3      4    1
# 10
      9      2      3     10    1
     10      3      4      2    3
     10      3      4      5    1
     10      3      4      7    1
      3      4      5     11    1
      7      4      5     11    1
      3      4      7     13    2
      3      4      7     14    2
      5      4      7     13    2
      5      4      7     14    2
# 20
      1      5      6     12    3
      4      5      6     12    1
     11      5      6      1    1
     11      5      6      4    3
     11      5      6     12    1
     14      7     13      4    3
END
DIHEDRAL
#  NPHI: number of dihedrals NOT involving H atoms in solute
9
#  IP, JP, KP, LP: atom sequence numbers
#    of atoms forming a dihedral
```

```
#  ICP: dihedral type code
#    IP      JP      KP      LP  ICP
      6       1       2       3    1
      2       1       6       5    1
      1       2       3       4    1
      2       3       4       5    1
      2       3       4       7    1
      3       4       5       6    1
      7       4       5       3    3
      7       4       5       6    1
      4       5       6       1    1
END
CROSSDIHEDRALH
#  NPHIH: number of cross dihedrals involving H atoms in solute
0
#  APH, BPH, CPH, DPH, EPH, FPH, GPH, HPH: atom sequence numbers
#    of atoms forming a dihedral
#  ICCH: dihedral type code
#   APH     BPH     CPH     DPH     EPH     FPH     GPH     HPH ICCH
END
CROSSDIHEDRAL
#  NPPC: number of cross dihedrals NOT involving H atoms in solute
0
#  AP, BP, CP, DP, EP, FP, GP, HP: atom sequence numbers
#     of atoms forming a dihedral
#  ICC: dihedral type code
#    AP      BP      CP      DP      EP      FP      GP      HP  ICC
END
LJPARAMETERS
#  NRATT2: number of LJ interaction types = NRATT*(NRATT+1)/2
15
#  IAC,JAC: integer (van der Waals) atom type code
#  C12: r**(-12) term in nonbonded interactions
#   C6: r**(-6) term in nonbonded interactions
# CS12: r**(-12) term in 1-4 nonbonded interactions
#  CS6: r**(-6) term in 1-4 nonbonded interactions
# IAC  JAC           C12            C6          CS12           CS6
    1    1  3.430761e-06  2.222134e-03  1.715381e-06  1.111067e-03
#
    1    2  3.692878e-06  2.733664e-03  1.846439e-06  1.366832e-03
    2    2  3.950923e-06  3.352738e-03  1.975461e-06  1.676369e-03
#
    1    3  3.190097e-07  4.379003e-04  1.595049e-07  2.189501e-04
    2    3  3.312168e-07  5.290748e-04  1.656084e-07  2.645374e-04
    3    3  2.391698e-08  7.748625e-05  1.195849e-08  3.874313e-05
#
    1    4  9.521845e-09  7.652193e-05  4.760923e-09  3.826097e-05
    2    4  8.895233e-09  8.769840e-05  4.447617e-09  4.384920e-05
    3    4  3.727892e-10  9.784873e-06  1.863946e-10  4.892437e-06
    4    4  5.856879e-13  3.922914e-07  2.928440e-13  1.961457e-07
#
    1    5  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
    2    5  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
    3    5  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
    4    5  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
    5    5  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
#
END
SOLUTEMOLECULES
# NSPM: number of separate molecules in solute block
# NSP[1...NSPM]: atom sequence number of last atom
#                of the successive submolecules
#      NSPM  NSP[1...NSPM]
```

```
         1
    14
END
TEMPERATUREGROUPS
# NSTM: number of temperature atom groups
# NST[1...NSTM]: atom sequence number of last atom
#                of the successive temperature atom groups
#      NSTM  NST[1...NSTM]
         1
    14
END
PRESSUREGROUPS
# NSVM: number of pressure atom groups
# NSV[1...NSVM]: atom sequence number of last atom
#                of the successive pressure atom groups
#      NSVM  NSV[1...NSVM]
         1
    14
END
LJEXCEPTIONS
# This block defines special LJ-interactions based on atom numbers
# This overrules the normal LJ-parameters (including 1-4 interactions)
# NEX: number of exceptions
0
# AT1  AT2           C12            C6
END
SOLVENTATOM
#  NRAM: number of atoms per solvent molecule
3
#     I: solvent atom sequence number
#  IACS: integer (van der Waals) atom type code
#  ANMS: atom name of solvent atom
#  MASS: mass of solvent atom
#   CGS: charge of solvent atom
#  I  ANMS IACS      MASS        CGS
   1    OW   5   15.99940   -0.82000
   2   HW1  21    1.00800    0.41000
   3   HW2  21    1.00800    0.41000
END
SOLVENTCONSTR
#  NCONS: number of constraints
3
#  ICONS, JCONS: atom sequence numbers forming constraint
#   CONS constraint length
#ICONS JCONS         CONS
    1    2      0.1000000
    1    3      0.1000000
    2    3      0.1632990
END
# end of topology file
```

6.A.2 Single-Molecule Simulations in Vacuum

Additional energy and temperature distribution plots for the single-molecule simulations in vacuum with GROMOS and GROMACS are provided in Figure 6.15.

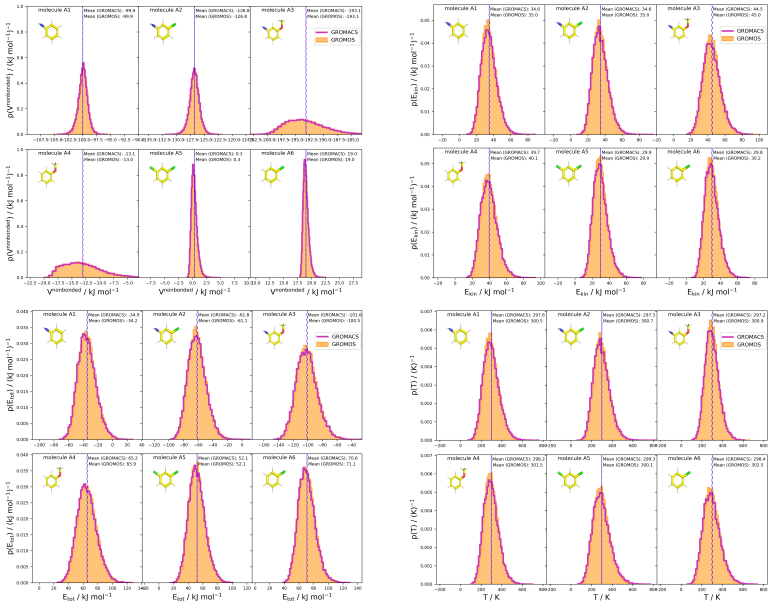


Figure 6.15: Energy and temperature distributions of single molecule simulations in vacuum. The nonbonded potential energy distributions (top left), nonbonded kinetic energy distributions (top right), total energy distributions (bottom left), as well as the temperature distributions (bottom right) for the molecules in set A are compared for 5 ns long vacuum simulations in GROMOS [27,231,312] and GROMACS. [25,305] The topologies for the simulations are based on the AMBER topologies taken from the FreeSolv database. [155,173] The GROMACS topologies were translated with ParmEd [293] and the GROMOS topologies were converted with *amber2gromos*. Energies were written every 100 timesteps (*i.e.*, every 200 fs) and the first 1.25 ns of the simulations were discarded as equilibration.

6.A.3 Relative Hydration Free Energies of Set A

Table 6.4: $\Delta\Delta G_{\mathrm{hyd}}$ for the six molecules in set A from experiment,[155,173] calculated from the hydration free energies obtained with MBAR,[155,173] from the pairwise relative calculations with TI, and from the multistate relative free-energy calculations with RE-EDS. The RE-EDs and TI results were averaged over five independent production runs in vacuum/water and the errors of the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to 4 molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		Experiment[155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$[155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{TI}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\text{RE-EDS}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
A1	A2	-1.4 ± 3.6	2.3 ± 0.1	1.5 ± 0.2	1.7 ± 0.3
A1	A3	-7.5 ± 3.6	-6.4 ± 0.2	-7.0 ± 0.2	-6.7 ± 0.4
A1	A4	13.3 ± 2.5	15.1 ± 0.1	15.0 ± 0.1	13.2 ± 0.3
A1	A5	18.9 ± 3.6	24.8 ± 0.1	22.7 ± 0.1	22.7 ± 0.3
A1	A6	18.3 ± 2.5	22.3 ± 0.1	21.1 ± 0.0	21.0 ± 0.3
A2	A3	-6.2 ± 3.6	-8.7 ± 0.2	-8.5 ± 0.2	-8.4 ± 0.2
A2	A4	14.7 ± 2.5	12.8 ± 0.1	13.5 ± 0.1	11.5 ± 0.2
A2	A5	20.3 ± 3.6	22.5 ± 0.1	21.0 ± 0.1	21.0 ± 0.2
A2	A6	19.7 ± 2.5	20.0 ± 0.1	19.6 ± 0.1	19.3 ± 0.1
A3	A4	20.9 ± 2.5	21.5 ± 0.2	20.3 ± 0.1	19.9 ± 0.2
A3	A5	26.4 ± 3.6	31.2 ± 0.2	29.7 ± 0.0	29.4 ± 0.2
A3	A6	25.8 ± 2.5	28.7 ± 0.2	28.3 ± 0.1	27.7 ± 0.2
A4	A5	5.5 ± 2.5	9.7 ± 0.1	9.3 ± 0.0	9.5 ± 0.2
A4	A6	4.9 ± 0.1	7.1 ± 0.1	7.9 ± 0.0	7.8 ± 0.2
A5	A6	-0.6 ± 2.5	-2.6 ± 0.1	-1.7 ± 0.5	-1.7 ± 0.2
RMSE			3.08 ± 0.4	2.36 ± 0.4	2.43 ± 0.3
MAE			2.67 ± 0.3	2.02 ± 0.4	2.08 ± 0.3
r_{Spearman}			0.94	0.93	0.93
$t_{\mathrm{preparation}}$			18 ns	360 ns	101.3 ns
$t_{\mathrm{production}}$			600 ns	3600 ns	13.5 ns

6.A.4 Relative Hydration Free Energies of Set B

Table 6.5: $\Delta\Delta G_{\mathrm{hyd}}$ for the 28 molecules in set B from experiment,[155,173] calculated from the hydration free energies obtained with MBAR[155,173] and from the multistate relative free-energy calculations with RE-EDS. The RE-EDS results were averaged over five independent production runs in vacuum/water and the errors of the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. For RE-EDS, both the results for the full set and for the combined subsets Ba and Bb are reported. The uncertainty of the RMSE and MAE were estimated from the distribution of RMSEs and MAEs when a random selection of up to 26 molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		Experiment [155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ [155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,subsets}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
B1	B2	−1.4 ± 3.6	1.7 ± 0.2	1.5 ± 0.2	1.7 ± 0.4
B1	B3	18.9 ± 3.6	22.7 ± 0.2	23.1 ± 0.1	23.1 ± 0.3
B1	B4	19.7 ± 3.6	24.6 ± 0.2	25.3 ± 0.4	25.4 ± 0.3
B1	B5	17.8 ± 3.6	21.0 ± 0.2	21.5 ± 0.2	21.5 ± 0.4
B1	B6	18.3 ± 3.6	22.7 ± 0.2	23.0 ± 0.2	23.1 ± 0.4
B1	B7	18.7 ± 3.6	22.4 ± 0.2	22.8 ± 0.1	22.8 ± 0.4
B1	B8	19.4 ± 3.6	19.8 ± 0.2	18.5 ± 0.2	19.2 ± 0.5
B1	B9	17.9 ± 3.6	19.5 ± 0.2	16.8 ± 0.2	17.3 ± 0.5
B1	B10	19.2 ± 3.6	20.9 ± 0.2	20.6 ± 0.4	20.7 ± 0.4
B1	B11	19.5 ± 3.6	20.3 ± 0.2	20.1 ± 0.1	20.1 ± 0.2
B1	B12	19.6 ± 3.6	20.4 ± 0.2	20.1 ± 0.1	20.2 ± 0.3
B1	B13	−2.7 ± 3.6	−0.2 ± 0.2	−0.2 ± 0.3	0.3 ± 0.5
B1	B14	−4.6 ± 2.6	−0.7 ± 0.2	−0.8 ± 0.1	−0.5 ± 0.4
B1	B15	15.7 ± 3.6	18.7 ± 0.2	19.5 ± 0.1	19.4 ± 0.2
B1	B16	18.3 ± 3.6	21.2 ± 0.2	21.6 ± 0.1	21.3 ± 0.2
B1	B17	19.6 ± 3.6	23.0 ± 0.2	23.3 ± 0.2	22.9 ± 0.2
B1	B18	16.9 ± 3.6	19.2 ± 0.2	19.6 ± 0.1	19.3 ± 0.2
B1	B19	17.2 ± 3.6	19.5 ± 0.2	19.8 ± 0.2	19.7 ± 0.3
B1	B20	19.2 ± 2.6	19.9 ± 0.2	19.9 ± 0.3	20.1 ± 0.2
B1	B21	13.0 ± 2.6	15.4 ± 0.2	15.3 ± 0.3	15.3 ± 0.3
B1	B22	14.9 ± 2.6	15.9 ± 0.2	16.1 ± 0.3	16.2 ± 0.3
B1	B23	19.7 ± 3.6	20.6 ± 0.2	19.7 ± 0.3	20.3 ± 0.2
B1	B24	18.2 ± 3.6	21.2 ± 0.2	21.1 ± 0.3	21.1 ± 0.3
B1	B25	19.2 ± 3.6	19.6 ± 0.2	18.6 ± 0.1	19.2 ± 0.3
B1	B26	−0.2 ± 3.6	0.9 ± 0.2	0.3 ± 0.3	0.7 ± 0.3
B1	B27	2.4 ± 3.6	2.9 ± 0.2	2.8 ± 0.3	2.7 ± 0.3
B1	B28	−0.3 ± 3.6	0.2 ± 0.2	0.0 ± 0.2	0.4 ± 0.1
B2	B3	20.3 ± 3.6	21.0 ± 0.2	21.5 ± 0.1	21.4 ± 0.3
B2	B4	21.1 ± 3.6	22.9 ± 0.2	23.8 ± 0.3	23.7 ± 0.3
B2	B5	19.2 ± 3.6	19.4 ± 0.2	20.0 ± 0.2	19.8 ± 0.4
B2	B6	19.7 ± 3.6	21.0 ± 0.2	21.4 ± 0.2	21.4 ± 0.4
B2	B7	20.1 ± 3.6	20.7 ± 0.2	21.2 ± 0.2	21.1 ± 0.4
B2	B8	20.8 ± 3.6	18.2 ± 0.2	17.0 ± 0.2	17.5 ± 0.5
B2	B9	19.3 ± 3.6	17.8 ± 0.2	15.2 ± 0.2	15.7 ± 0.5
B2	B10	20.6 ± 3.6	19.2 ± 0.2	19.1 ± 0.3	19.0 ± 0.4
B2	B11	20.9 ± 3.6	18.6 ± 0.2	18.5 ± 0.1	18.5 ± 0.2
B2	B12	21.0 ± 3.6	18.7 ± 0.2	18.5 ± 0.1	18.5 ± 0.3
B2	B13	−1.3 ± 3.6	−1.8 ± 0.2	−1.7 ± 0.4	−1.3 ± 0.5
B2	B14	−3.3 ± 2.6	−2.4 ± 0.2	−2.3 ± 0.2	−2.2 ± 0.4
B2	B15	17.1 ± 3.6	17.1 ± 0.2	18.0 ± 0.1	17.7 ± 0.4
B2	B16	19.7 ± 3.6	19.5 ± 0.2	20.0 ± 0.1	19.7 ± 0.4
B2	B17	21.0 ± 3.6	21.3 ± 0.2	21.8 ± 0.1	21.3 ± 0.4
B2	B18	18.2 ± 3.6	17.5 ± 0.2	18.0 ± 0.1	17.7 ± 0.4
B2	B19	18.5 ± 3.6	17.8 ± 0.2	18.2 ± 0.2	18.0 ± 0.5
B2	B20	20.6 ± 2.6	18.2 ± 0.2	18.3 ± 0.3	18.4 ± 0.5
B2	B21	14.4 ± 2.6	13.8 ± 0.2	13.8 ± 0.2	13.7 ± 0.5
B2	B22	16.3 ± 2.6	14.2 ± 0.2	14.6 ± 0.2	14.5 ± 0.5
B2	B23	21.0 ± 3.6	19.0 ± 0.2	18.1 ± 0.2	18.7 ± 0.5
B2	B24	19.6 ± 3.6	19.5 ± 0.2	19.5 ± 0.2	19.4 ± 0.5
B2	B25	20.6 ± 3.6	17.9 ± 0.2	17.1 ± 0.1	17.5 ± 0.5
B2	B26	1.2 ± 3.6	−0.8 ± 0.2	−1.2 ± 0.3	−1.0 ± 0.5
B2	B27	3.8 ± 3.6	1.2 ± 0.2	1.2 ± 0.3	1.1 ± 0.5
B2	B28	1.0 ± 3.6	−1.5 ± 0.2	−1.5 ± 0.3	−1.3 ± 0.4
B3	B4	0.8 ± 3.6	1.8 ± 0.2	2.2 ± 0.3	2.3 ± 0.3
B3	B5	−1.1 ± 3.6	−1.7 ± 0.2	−1.6 ± 0.1	−1.6 ± 0.4

B3	B6	−0.6 ± 3.6	−0.0 ± 0.2	−0.1 ± 0.2	−0.0 ± 0.4
B3	B7	−0.1 ± 3.6	−0.3 ± 0.2	−0.3 ± 0.1	−0.3 ± 0.4
B3	B8	0.5 ± 3.6	−2.9 ± 0.2	−4.5 ± 0.2	−3.9 ± 0.5
B3	B9	−1.0 ± 3.6	−3.2 ± 0.2	−6.3 ± 0.2	−5.7 ± 0.5
B3	B10	0.3 ± 3.6	−1.8 ± 0.2	−2.5 ± 0.3	−2.4 ± 0.4
B3	B11	0.6 ± 3.6	−2.5 ± 0.2	−3.0 ± 0.1	−3.0 ± 0.2
B3	B12	0.8 ± 3.6	−2.3 ± 0.2	−3.0 ± 0.1	−2.9 ± 0.3
B3	B13	−21.5 ± 3.6	−22.9 ± 0.2	−23.3 ± 0.3	−22.8 ± 0.5
B3	B14	−23.5 ± 2.6	−23.4 ± 0.2	−23.8 ± 0.1	−23.6 ± 0.4
B3	B15	−3.2 ± 3.6	−4.0 ± 0.2	−3.5 ± 0.1	−3.7 ± 0.4
B3	B16	−0.6 ± 3.6	−1.5 ± 0.2	−1.5 ± 0.1	−1.8 ± 0.3
B3	B17	0.8 ± 3.6	0.3 ± 0.2	0.3 ± 0.1	−0.1 ± 0.4
B3	B18	−2.0 ± 3.6	−3.5 ± 0.2	−3.5 ± 0.1	−3.8 ± 0.4
B3	B19	−1.7 ± 3.6	−3.3 ± 0.2	−3.3 ± 0.2	−3.4 ± 0.4
B3	B20	0.3 ± 2.6	−2.8 ± 0.2	−3.2 ± 0.2	−3.0 ± 0.4
B3	B21	−5.9 ± 2.6	−7.3 ± 0.2	−7.8 ± 0.2	−7.8 ± 0.4
B3	B22	−4.0 ± 2.6	−6.8 ± 0.2	−7.0 ± 0.2	−6.9 ± 0.4
B3	B23	0.8 ± 3.6	−2.1 ± 0.2	−3.4 ± 0.2	−2.8 ± 0.4
B3	B24	−0.7 ± 3.6	−1.5 ± 0.2	−2.0 ± 0.2	−2.0 ± 0.5
B3	B25	0.3 ± 3.6	−3.1 ± 0.2	−4.4 ± 0.1	−3.9 ± 0.4
B3	B26	−19.0 ± 3.6	−21.8 ± 0.2	−22.7 ± 0.3	−22.4 ± 0.4
B3	B27	−16.4 ± 3.6	−19.8 ± 0.2	−20.3 ± 0.3	−20.4 ± 0.4
B3	B28	−19.2 ± 3.6	−22.5 ± 0.2	−23.1 ± 0.2	−22.7 ± 0.3
B4	B5	−1.9 ± 3.6	−3.5 ± 0.2	−3.8 ± 0.3	−3.9 ± 0.4
B4	B6	−1.4 ± 3.6	−1.9 ± 0.2	−2.3 ± 0.3	−2.3 ± 0.4
B4	B7	−1.0 ± 3.6	−2.2 ± 0.2	−2.5 ± 0.3	−2.6 ± 0.4
B4	B8	−0.3 ± 3.6	−4.7 ± 0.2	−6.8 ± 0.3	−6.2 ± 0.5
B4	B9	−1.8 ± 3.6	−5.1 ± 0.2	−8.5 ± 0.2	−8.0 ± 0.5
B4	B10	−0.5 ± 3.6	−3.7 ± 0.2	−4.7 ± 0.2	−4.7 ± 0.4
B4	B11	−0.2 ± 3.6	−4.3 ± 0.2	−5.2 ± 0.3	−5.2 ± 0.2
B4	B12	−0.1 ± 3.6	−4.1 ± 0.2	−5.2 ± 0.3	−5.2 ± 0.3
B4	B13	−22.4 ± 3.6	−24.7 ± 0.2	−25.5 ± 0.6	−25.0 ± 0.5
B4	B14	−24.4 ± 2.6	−25.3 ± 0.2	−26.1 ± 0.4	−25.8 ± 0.4
B4	B15	−4.0 ± 3.6	−5.8 ± 0.2	−5.8 ± 0.3	−6.0 ± 0.3
B4	B16	−1.4 ± 3.6	−3.3 ± 0.2	−3.7 ± 0.3	−4.0 ± 0.3
B4	B17	−0.1 ± 3.6	−1.5 ± 0.2	−2.0 ± 0.3	−2.4 ± 0.4
B4	B18	−2.8 ± 3.6	−5.4 ± 0.2	−5.7 ± 0.3	−6.0 ± 0.4
B4	B19	−2.6 ± 3.6	−5.1 ± 0.2	−5.5 ± 0.3	−5.7 ± 0.4
B4	B20	−0.5 ± 2.6	−4.7 ± 0.2	−5.4 ± 0.4	−5.3 ± 0.4
B4	B21	−6.7 ± 2.6	−9.1 ± 0.2	−10.0 ± 0.3	−10.0 ± 0.4
B4	B22	−4.8 ± 2.6	−8.7 ± 0.2	−9.2 ± 0.3	−9.2 ± 0.4
B4	B23	−0.0 ± 3.6	−3.9 ± 0.2	−5.6 ± 0.3	−5.0 ± 0.4
B4	B24	−1.5 ± 3.6	−3.3 ± 0.2	−4.2 ± 0.4	−4.3 ± 0.4
B4	B25	−0.5 ± 3.6	−4.9 ± 0.2	−6.7 ± 0.4	−6.2 ± 0.4
B4	B26	−19.9 ± 3.6	−23.7 ± 0.2	−25.0 ± 0.2	−24.7 ± 0.4
B4	B27	−17.3 ± 3.6	−21.7 ± 0.2	−22.5 ± 0.2	−22.6 ± 0.4
B4	B28	−20.0 ± 3.6	−24.4 ± 0.2	−25.3 ± 0.3	−25.0 ± 0.3
B5	B6	0.5 ± 3.6	1.6 ± 0.2	1.5 ± 0.2	1.6 ± 0.4
B5	B7	1.0 ± 3.6	1.3 ± 0.2	1.3 ± 0.1	1.3 ± 0.4
B5	B8	1.6 ± 3.6	−1.2 ± 0.2	−3.0 ± 0.1	−2.3 ± 0.5
B5	B9	0.1 ± 3.6	−1.5 ± 0.2	−4.7 ± 0.2	−4.2 ± 0.5
B5	B10	1.4 ± 3.6	−0.2 ± 0.2	−0.9 ± 0.4	−0.8 ± 0.4
B5	B11	1.7 ± 3.6	−0.8 ± 0.2	−1.4 ± 0.1	−1.4 ± 0.2
B5	B12	1.8 ± 3.6	−0.6 ± 0.2	−1.4 ± 0.1	−1.3 ± 0.3
B5	B13	−20.5 ± 3.6	−21.2 ± 0.2	−21.7 ± 0.3	−21.2 ± 0.5
B5	B14	−22.4 ± 2.6	−21.8 ± 0.2	−22.3 ± 0.2	−22.0 ± 0.4
B5	B15	−2.1 ± 3.6	−2.3 ± 0.2	−2.0 ± 0.1	−2.1 ± 0.5
B5	B16	0.5 ± 3.6	0.2 ± 0.2	0.1 ± 0.1	−0.2 ± 0.5
B5	B17	1.8 ± 3.6	2.0 ± 0.2	1.8 ± 0.2	1.5 ± 0.5
B5	B18	−0.9 ± 3.6	−1.8 ± 0.2	−2.0 ± 0.1	−2.2 ± 0.5
B5	B19	−0.6 ± 3.6	−1.6 ± 0.2	−1.8 ± 0.2	−1.8 ± 0.5
B5	B20	1.4 ± 2.6	−1.2 ± 0.2	−1.6 ± 0.1	−1.4 ± 0.5
B5	B21	−4.8 ± 2.6	−5.6 ± 0.2	−6.2 ± 0.2	−6.2 ± 0.5
B5	B22	−2.9 ± 2.6	−5.1 ± 0.2	−5.4 ± 0.2	−5.3 ± 0.5
B5	B23	1.9 ± 3.6	−0.4 ± 0.2	−1.9 ± 0.2	−1.2 ± 0.5
B5	B24	0.4 ± 3.6	0.2 ± 0.2	−0.4 ± 0.2	−0.4 ± 0.5
B5	B25	1.4 ± 3.6	−1.4 ± 0.2	−2.9 ± 0.2	−2.3 ± 0.5
B5	B26	−17.9 ± 3.6	−20.2 ± 0.2	−21.2 ± 0.2	−20.8 ± 0.5
B5	B27	−15.4 ± 3.6	−18.2 ± 0.2	−18.8 ± 0.3	−18.8 ± 0.5
B5	B28	−18.1 ± 3.6	−20.8 ± 0.2	−21.5 ± 0.1	−21.1 ± 0.4
B6	B7	0.5 ± 3.6	−0.3 ± 0.2	−0.2 ± 0.1	−0.3 ± 0.4
B6	B8	1.1 ± 3.6	−2.8 ± 0.2	−4.5 ± 0.1	−3.9 ± 0.5
B6	B9	−0.4 ± 3.6	−3.2 ± 0.2	−6.2 ± 0.2	−5.7 ± 0.5
B6	B10	0.9 ± 3.6	−1.8 ± 0.2	−2.4 ± 0.3	−2.4 ± 0.4
B6	B11	1.2 ± 3.6	−2.4 ± 0.2	−2.9 ± 0.1	−2.9 ± 0.2
B6	B12	1.3 ± 3.6	−2.3 ± 0.2	−2.9 ± 0.1	−2.9 ± 0.3
B6	B13	−21.0 ± 3.6	−22.8 ± 0.2	−23.2 ± 0.3	−22.7 ± 0.5
B6	B14	−22.9 ± 2.6	−23.4 ± 0.2	−23.7 ± 0.2	−23.5 ± 0.4

B6	B15	-2.6 ± 3.6	-3.9 ± 0.2	-3.4 ± 0.1	-3.7 ± 0.5
B6	B16	0.0 ± 3.6	-1.5 ± 0.2	-1.4 ± 0.1	-1.7 ± 0.5
B6	B17	1.3 ± 3.6	0.3 ± 0.2	0.4 ± 0.1	-0.1 ± 0.5
B6	B18	-1.4 ± 3.6	-3.5 ± 0.2	-3.4 ± 0.1	-3.7 ± 0.5
B6	B19	-1.1 ± 3.6	-3.2 ± 0.2	-3.2 ± 0.1	-3.4 ± 0.5
B6	B20	0.9 ± 2.6	-2.8 ± 0.2	-3.1 ± 0.2	-3.0 ± 0.5
B6	B21	-5.3 ± 2.6	-7.2 ± 0.2	-7.7 ± 0.2	-7.7 ± 0.5
B6	B22	-3.4 ± 2.6	-6.8 ± 0.2	-6.9 ± 0.3	-6.9 ± 0.5
B6	B23	1.4 ± 3.6	-2.1 ± 0.2	-3.3 ± 0.2	-2.7 ± 0.5
B6	B24	-0.1 ± 3.6	-1.5 ± 0.2	-1.9 ± 0.3	-2.0 ± 0.5
B6	B25	0.9 ± 3.6	-3.1 ± 0.2	-4.3 ± 0.2	-3.9 ± 0.5
B6	B26	-18.5 ± 3.6	-21.8 ± 0.2	-22.6 ± 0.2	-22.4 ± 0.5
B6	B27	-15.9 ± 3.6	-19.8 ± 0.2	-20.2 ± 0.2	-20.3 ± 0.5
B6	B28	-18.6 ± 3.6	-22.5 ± 0.2	-23.0 ± 0.2	-22.7 ± 0.4
B7	B8	0.6 ± 3.6	-2.6 ± 0.2	-4.3 ± 0.2	-3.6 ± 0.5
B7	B9	-0.8 ± 3.6	-2.9 ± 0.2	-6.0 ± 0.1	-5.4 ± 0.5
B7	B10	0.5 ± 3.6	-1.5 ± 0.2	-2.2 ± 0.3	-2.1 ± 0.4
B7	B11	0.8 ± 3.6	-2.1 ± 0.2	-2.7 ± 0.0	-2.6 ± 0.2
B7	B12	0.9 ± 3.6	-2.0 ± 0.2	-2.7 ± 0.0	-2.6 ± 0.3
B7	B13	-21.4 ± 3.6	-22.6 ± 0.2	-23.0 ± 0.3	-22.4 ± 0.5
B7	B14	-23.4 ± 2.6	-23.1 ± 0.2	-23.5 ± 0.2	-23.2 ± 0.4
B7	B15	-3.1 ± 3.6	-3.6 ± 0.2	-3.3 ± 0.1	-3.4 ± 0.5
B7	B16	-0.5 ± 3.6	-1.2 ± 0.2	-1.2 ± 0.1	-1.4 ± 0.5
B7	B17	0.9 ± 3.6	0.6 ± 0.2	0.5 ± 0.2	0.2 ± 0.5
B7	B18	-1.9 ± 3.6	-3.2 ± 0.2	-3.2 ± 0.0	-3.4 ± 0.5
B7	B19	-1.6 ± 3.6	-2.9 ± 0.2	-3.0 ± 0.1	-3.1 ± 0.5
B7	B20	0.5 ± 2.6	-2.5 ± 0.2	-2.9 ± 0.2	-2.7 ± 0.5
B7	B21	-5.7 ± 2.6	-6.9 ± 0.2	-7.5 ± 0.2	-7.4 ± 0.5
B7	B22	-3.8 ± 2.6	-6.5 ± 0.2	-6.7 ± 0.2	-6.6 ± 0.5
B7	B23	0.9 ± 3.6	-1.8 ± 0.2	-3.1 ± 0.2	-2.4 ± 0.5
B7	B24	-0.5 ± 3.6	-1.2 ± 0.2	-1.7 ± 0.3	-1.7 ± 0.6
B7	B25	0.5 ± 3.6	-2.8 ± 0.2	-4.2 ± 0.2	-3.6 ± 0.5
B7	B26	-18.9 ± 3.6	-21.5 ± 0.2	-22.4 ± 0.2	-22.1 ± 0.5
B7	B27	-16.3 ± 3.6	-19.5 ± 0.2	-20.0 ± 0.2	-20.0 ± 0.5
B7	B28	-19.1 ± 3.6	-22.2 ± 0.2	-22.8 ± 0.1	-22.4 ± 0.4
B8	B9	-1.5 ± 3.6	-0.3 ± 0.2	-1.7 ± 0.2	-1.8 ± 0.5
B8	B10	-0.2 ± 3.6	1.0 ± 0.2	2.1 ± 0.4	1.5 ± 0.4
B8	B11	0.1 ± 3.6	0.4 ± 0.2	1.6 ± 0.2	0.9 ± 0.2
B8	B12	0.3 ± 3.6	0.6 ± 0.2	1.6 ± 0.2	1.0 ± 0.3
B8	B13	-22.0 ± 3.6	-20.0 ± 0.2	-18.7 ± 0.4	-18.9 ± 0.5
B8	B14	-24.0 ± 2.6	-20.5 ± 0.2	-19.3 ± 0.2	-19.7 ± 0.4
B8	B15	-3.7 ± 3.6	-1.1 ± 0.2	1.0 ± 0.2	0.2 ± 0.5
B8	B16	-1.1 ± 3.6	1.4 ± 0.2	3.1 ± 0.2	2.1 ± 0.5
B8	B17	0.3 ± 3.6	3.2 ± 0.2	4.8 ± 0.2	3.8 ± 0.5
B8	B18	-2.5 ± 3.6	-0.6 ± 0.2	1.0 ± 0.2	0.2 ± 0.5
B8	B19	-2.2 ± 3.6	-0.4 ± 0.2	1.2 ± 0.2	0.5 ± 0.5
B8	B20	-0.2 ± 2.6	0.0 ± 0.2	1.3 ± 0.1	0.9 ± 0.5
B8	B21	-6.4 ± 2.6	-4.4 ± 0.2	-3.2 ± 0.2	-3.9 ± 0.6
B8	B22	-4.5 ± 2.6	-3.9 ± 0.2	-2.4 ± 0.3	-3.0 ± 0.5
B8	B23	0.3 ± 3.6	0.8 ± 0.2	1.1 ± 0.2	1.1 ± 0.5
B8	B24	-1.2 ± 3.6	1.4 ± 0.2	2.6 ± 0.3	1.9 ± 0.6
B8	B25	-0.2 ± 3.6	-0.2 ± 0.2	0.1 ± 0.3	-0.0 ± 0.5
B8	B26	-19.5 ± 3.6	-19.0 ± 0.2	-18.2 ± 0.2	-18.5 ± 0.5
B8	B27	-16.9 ± 3.6	-16.9 ± 0.2	-15.8 ± 0.2	-16.5 ± 0.5
B8	B28	-19.7 ± 3.6	-19.6 ± 0.2	-18.5 ± 0.1	-18.8 ± 0.5
B9	B10	1.3 ± 3.6	1.4 ± 0.2	3.8 ± 0.3	3.4 ± 0.4
B9	B11	1.6 ± 3.6	0.8 ± 0.2	3.3 ± 0.2	2.8 ± 0.2
B9	B12	1.7 ± 3.6	0.9 ± 0.2	3.3 ± 0.2	2.8 ± 0.3
B9	B13	-20.6 ± 3.6	-19.7 ± 0.2	-17.0 ± 0.4	-17.0 ± 0.5
B9	B14	-22.6 ± 2.6	-20.2 ± 0.2	-17.5 ± 0.3	-17.8 ± 0.4
B9	B15	-2.2 ± 3.6	-0.8 ± 0.2	2.7 ± 0.2	2.0 ± 0.5
B9	B16	0.4 ± 3.6	1.7 ± 0.2	4.8 ± 0.2	4.0 ± 0.5
B9	B17	1.7 ± 3.6	3.5 ± 0.2	6.5 ± 0.2	5.6 ± 0.5
B9	B18	-1.0 ± 3.6	-0.3 ± 0.2	2.8 ± 0.2	2.0 ± 0.5
B9	B19	-0.8 ± 3.6	-0.0 ± 0.2	3.0 ± 0.2	2.3 ± 0.6
B9	B20	1.3 ± 2.6	0.4 ± 0.2	3.1 ± 0.2	2.7 ± 0.5
B9	B21	-4.9 ± 2.6	-4.1 ± 0.2	-1.5 ± 0.3	-2.0 ± 0.6
B9	B22	-3.0 ± 2.6	-3.6 ± 0.2	-0.7 ± 0.3	-1.1 ± 0.5
B9	B23	1.8 ± 3.6	1.1 ± 0.2	2.9 ± 0.2	3.0 ± 0.5
B9	B24	0.3 ± 3.6	1.7 ± 0.2	4.3 ± 0.4	3.7 ± 0.6
B9	B25	1.3 ± 3.6	0.1 ± 0.2	1.8 ± 0.3	1.8 ± 0.5
B9	B26	-18.1 ± 3.6	-18.6 ± 0.2	-16.4 ± 0.1	-16.7 ± 0.6
B9	B27	-15.5 ± 3.6	-16.6 ± 0.2	-14.0 ± 0.1	-14.6 ± 0.5
B9	B28	-18.2 ± 3.6	-19.3 ± 0.2	-16.8 ± 0.2	-17.0 ± 0.5
B10	B11	0.3 ± 3.6	-0.6 ± 0.2	-0.5 ± 0.3	-0.6 ± 0.2
B10	B12	0.4 ± 3.6	-0.5 ± 0.2	-0.5 ± 0.3	-0.5 ± 0.3
B10	B13	-21.9 ± 3.6	-21.0 ± 0.2	-20.8 ± 0.6	-20.4 ± 0.5
B10	B14	-23.8 ± 2.6	-21.6 ± 0.2	-21.4 ± 0.4	-21.2 ± 0.4

B10	B15	−3.5 ± 3.6	−2.1 ± 0.2	−1.1 ± 0.4	−1.3 ± 0.4
B10	B16	−0.9 ± 3.6	0.3 ± 0.2	1.0 ± 0.3	0.6 ± 0.4
B10	B17	0.4 ± 3.6	2.1 ± 0.2	2.7 ± 0.3	2.3 ± 0.4
B10	B18	−2.3 ± 3.6	−1.7 ± 0.2	−1.1 ± 0.3	−1.4 ± 0.4
B10	B19	−2.1 ± 3.6	−1.4 ± 0.2	−0.9 ± 0.3	−1.0 ± 0.5
B10	B20	0.0 ± 2.6	−1.0 ± 0.2	−0.7 ± 0.4	−0.6 ± 0.5
B10	B21	−6.2 ± 2.6	−5.4 ± 0.2	−5.3 ± 0.4	−5.4 ± 0.5
B10	B22	−4.3 ± 2.6	−5.0 ± 0.2	−4.5 ± 0.4	−4.5 ± 0.5
B10	B23	0.5 ± 3.6	−0.3 ± 0.2	−0.9 ± 0.3	−0.4 ± 0.4
B10	B24	−1.0 ± 3.6	0.3 ± 0.2	0.5 ± 0.4	0.4 ± 0.5
B10	B25	0.0 ± 3.6	−1.3 ± 0.2	−2.0 ± 0.4	−1.5 ± 0.5
B10	B26	−19.4 ± 3.6	−20.0 ± 0.2	−20.3 ± 0.3	−20.0 ± 0.5
B10	B27	−16.8 ± 3.6	−18.0 ± 0.2	−17.8 ± 0.3	−18.0 ± 0.5
B10	B28	−19.5 ± 3.6	−20.7 ± 0.2	−20.6 ± 0.4	−20.3 ± 0.4
B11	B12	0.1 ± 3.6	0.2 ± 0.2	0.0 ± 0.1	0.0 ± 0.3
B11	B13	−22.2 ± 3.6	−20.4 ± 0.2	−20.3 ± 0.3	−19.8 ± 0.5
B11	B14	−24.1 ± 2.6	−21.0 ± 0.2	−20.8 ± 0.2	−20.6 ± 0.4
B11	B15	−3.8 ± 3.6	−1.5 ± 0.2	−0.5 ± 0.1	−0.8 ± 0.3
B11	B16	−1.2 ± 3.6	1.0 ± 0.2	1.5 ± 0.1	1.2 ± 0.3
B11	B17	0.1 ± 3.6	2.8 ± 0.2	3.3 ± 0.1	2.8 ± 0.3
B11	B18	−2.6 ± 3.6	−1.0 ± 0.2	−0.5 ± 0.1	−0.8 ± 0.3
B11	B19	−2.3 ± 3.6	−0.8 ± 0.2	−0.3 ± 0.1	−0.4 ± 0.4
B11	B20	−0.3 ± 2.6	−0.4 ± 0.2	−0.2 ± 0.2	−0.0 ± 0.3
B11	B21	−6.5 ± 2.6	−4.8 ± 0.2	−4.8 ± 0.2	−4.8 ± 0.4
B11	B22	−4.6 ± 2.6	−4.4 ± 0.2	−4.0 ± 0.2	−3.9 ± 0.3
B11	B23	0.2 ± 3.6	0.4 ± 0.2	−0.4 ± 0.2	0.2 ± 0.3
B11	B24	−1.3 ± 3.6	1.0 ± 0.2	1.0 ± 0.3	0.9 ± 0.4
B11	B25	−0.3 ± 3.6	−0.6 ± 0.2	−1.4 ± 0.1	−1.0 ± 0.4
B11	B26	−19.7 ± 3.6	−19.4 ± 0.2	−19.7 ± 0.2	−19.5 ± 0.4
B11	B27	−17.1 ± 3.6	−17.4 ± 0.2	−17.3 ± 0.2	−17.4 ± 0.4
B11	B28	−19.8 ± 3.6	−20.0 ± 0.2	−20.1 ± 0.2	−19.8 ± 0.3
B12	B13	−22.3 ± 3.6	−20.6 ± 0.2	−20.3 ± 0.4	−19.8 ± 0.5
B12	B14	−24.3 ± 2.6	−21.1 ± 0.2	−20.8 ± 0.2	−20.7 ± 0.4
B12	B15	−3.9 ± 3.6	−1.7 ± 0.2	−0.5 ± 0.1	−0.8 ± 0.4
B12	B16	−1.3 ± 3.6	0.8 ± 0.2	1.5 ± 0.0	1.2 ± 0.4
B12	B17	0.0 ± 3.6	2.6 ± 0.2	3.3 ± 0.2	2.8 ± 0.4
B12	B18	−2.8 ± 3.6	−1.2 ± 0.2	−0.5 ± 0.1	−0.8 ± 0.4
B12	B19	−2.5 ± 3.6	−1.0 ± 0.2	−0.3 ± 0.1	−0.5 ± 0.4
B12	B20	−0.4 ± 2.6	−0.5 ± 0.2	−0.2 ± 0.2	−0.1 ± 0.4
B12	B21	−6.6 ± 2.6	−5.0 ± 0.2	−4.8 ± 0.2	−4.9 ± 0.5
B12	B22	−4.7 ± 2.6	−4.5 ± 0.2	−4.0 ± 0.2	−4.0 ± 0.4
B12	B23	0.0 ± 3.6	0.2 ± 0.2	−0.4 ± 0.2	0.2 ± 0.4
B12	B24	−1.4 ± 3.6	0.8 ± 0.2	1.0 ± 0.3	0.9 ± 0.5
B12	B25	−0.4 ± 3.6	−0.8 ± 0.2	−1.4 ± 0.2	−1.0 ± 0.4
B12	B26	−19.8 ± 3.6	−19.5 ± 0.2	−19.7 ± 0.2	−19.5 ± 0.4
B12	B27	−17.2 ± 3.6	−17.5 ± 0.2	−17.3 ± 0.2	−17.4 ± 0.4
B12	B28	−20.0 ± 3.6	−20.2 ± 0.2	−20.1 ± 0.2	−19.8 ± 0.4
B13	B14	−2.0 ± 2.6	−0.5 ± 0.2	−0.6 ± 0.2	−0.8 ± 0.4
B13	B15	18.4 ± 3.6	18.9 ± 0.2	19.7 ± 0.3	19.0 ± 0.5
B13	B16	21.0 ± 3.6	21.4 ± 0.2	21.8 ± 0.3	21.0 ± 0.5
B13	B17	22.3 ± 3.6	23.2 ± 0.2	23.5 ± 0.3	22.6 ± 0.5
B13	B18	19.5 ± 3.6	19.4 ± 0.2	19.7 ± 0.3	19.0 ± 0.5
B13	B19	19.8 ± 3.6	19.6 ± 0.2	19.9 ± 0.4	19.4 ± 0.6
B13	B20	21.9 ± 2.6	20.0 ± 0.2	20.1 ± 0.4	19.8 ± 0.6
B13	B21	15.7 ± 2.6	15.6 ± 0.2	15.5 ± 0.4	15.0 ± 0.6
B13	B22	17.6 ± 2.6	16.1 ± 0.2	16.3 ± 0.4	15.9 ± 0.6
B13	B23	22.3 ± 3.6	20.8 ± 0.2	19.9 ± 0.4	20.0 ± 0.6
B13	B24	20.9 ± 3.6	21.4 ± 0.2	21.3 ± 0.3	20.7 ± 0.6
B13	B25	21.9 ± 3.6	19.8 ± 0.2	18.8 ± 0.3	18.9 ± 0.6
B13	B26	2.5 ± 3.6	1.0 ± 0.2	0.5 ± 0.5	0.3 ± 0.6
B13	B27	5.1 ± 3.6	3.1 ± 0.2	3.0 ± 0.5	2.4 ± 0.6
B13	B28	2.3 ± 3.6	0.4 ± 0.2	0.2 ± 0.4	0.0 ± 0.5
B14	B15	20.3 ± 2.6	19.5 ± 0.2	20.3 ± 0.1	19.8 ± 0.4
B14	B16	22.9 ± 2.6	21.9 ± 0.2	22.3 ± 0.2	21.8 ± 0.4
B14	B17	24.3 ± 2.6	23.7 ± 0.2	24.1 ± 0.2	23.4 ± 0.4
B14	B18	21.5 ± 2.6	19.9 ± 0.2	20.3 ± 0.1	19.8 ± 0.4
B14	B19	21.8 ± 2.6	20.2 ± 0.2	20.5 ± 0.2	20.2 ± 0.5
B14	B20	23.8 ± 1.2	20.6 ± 0.2	20.6 ± 0.2	20.6 ± 0.4
B14	B21	17.7 ± 1.2	16.2 ± 0.2	16.1 ± 0.3	15.8 ± 0.5
B14	B22	19.5 ± 1.2	16.6 ± 0.2	16.9 ± 0.2	16.7 ± 0.4
B14	B23	24.3 ± 2.6	21.3 ± 0.2	20.4 ± 0.3	20.8 ± 0.4
B14	B24	22.8 ± 2.6	21.9 ± 0.2	21.8 ± 0.2	21.6 ± 0.5
B14	B25	23.8 ± 2.6	20.3 ± 0.2	19.4 ± 0.1	19.7 ± 0.5
B14	B26	4.5 ± 2.6	1.6 ± 0.2	1.1 ± 0.4	1.1 ± 0.5
B14	B27	7.1 ± 2.6	3.6 ± 0.2	3.5 ± 0.4	3.2 ± 0.5
B14	B28	4.3 ± 2.6	0.9 ± 0.2	0.8 ± 0.2	0.9 ± 0.4
B15	B16	2.6 ± 3.6	2.5 ± 0.2	2.0 ± 0.1	2.0 ± 0.1
B15	B17	3.9 ± 3.6	4.3 ± 0.2	3.8 ± 0.2	3.6 ± 0.1

B15	B18	1.2 ± 3.6	0.5 ± 0.2	0.0 ± 0.0	−0.0 ± 0.0
B15	B19	1.5 ± 3.6	0.7 ± 0.2	0.2 ± 0.1	0.3 ± 0.2
B15	B20	3.5 ± 2.6	1.1 ± 0.2	0.3 ± 0.2	0.7 ± 0.1
B15	B21	−2.7 ± 2.6	−3.3 ± 0.2	−4.2 ± 0.2	−4.0 ± 0.3
B15	B22	−0.8 ± 2.6	−2.8 ± 0.2	−3.4 ± 0.2	−3.2 ± 0.1
B15	B23	4.0 ± 3.6	1.9 ± 0.2	0.1 ± 0.2	1.0 ± 0.1
B15	B24	2.5 ± 3.6	2.5 ± 0.2	1.5 ± 0.2	1.7 ± 0.2
B15	B25	3.5 ± 3.6	0.9 ± 0.2	−0.9 ± 0.1	−0.2 ± 0.1
B15	B26	−15.9 ± 3.6	−17.9 ± 0.2	−19.2 ± 0.3	−18.7 ± 0.1
B15	B27	−13.3 ± 3.6	−15.9 ± 0.2	−16.8 ± 0.3	−16.6 ± 0.2
B15	B28	−16.0 ± 3.6	−18.5 ± 0.2	−19.5 ± 0.2	−19.0 ± 0.2
B16	B17	1.3 ± 3.6	1.8 ± 0.1	1.8 ± 0.1	1.6 ± 0.1
B16	B18	−1.4 ± 3.6	−2.0 ± 0.2	−2.0 ± 0.1	−2.0 ± 0.1
B16	B19	−1.1 ± 3.6	−1.8 ± 0.2	−1.8 ± 0.1	−1.6 ± 0.1
B16	B20	0.9 ± 2.6	−1.3 ± 0.2	−1.7 ± 0.2	−1.2 ± 0.1
B16	B21	−5.3 ± 2.6	−5.8 ± 0.2	−6.3 ± 0.2	−6.0 ± 0.2
B16	B22	−3.4 ± 2.6	−5.3 ± 0.2	−5.5 ± 0.2	−5.1 ± 0.1
B16	B23	1.4 ± 3.6	−0.6 ± 0.2	−1.9 ± 0.2	−1.0 ± 0.1
B16	B24	−0.1 ± 3.6	0.0 ± 0.2	−0.5 ± 0.2	−0.3 ± 0.2
B16	B25	0.9 ± 3.6	−1.6 ± 0.2	−2.9 ± 0.1	−2.1 ± 0.1
B16	B26	−18.5 ± 3.6	−20.3 ± 0.2	−21.2 ± 0.2	−20.7 ± 0.2
B16	B27	−15.9 ± 3.6	−18.3 ± 0.2	−18.8 ± 0.2	−18.6 ± 0.2
B16	B28	−18.6 ± 3.6	−21.0 ± 0.2	−21.6 ± 0.2	−20.9 ± 0.2
B17	B18	−2.8 ± 3.6	−3.8 ± 0.2	−3.8 ± 0.1	−3.6 ± 0.1
B17	B19	−2.5 ± 3.6	−3.6 ± 0.2	−3.6 ± 0.2	−3.3 ± 0.2
B17	B20	−0.4 ± 2.6	−3.1 ± 0.2	−3.5 ± 0.3	−2.9 ± 0.1
B17	B21	−6.6 ± 2.6	−7.6 ± 0.2	−8.0 ± 0.2	−7.6 ± 0.3
B17	B22	−4.7 ± 2.6	−7.1 ± 0.2	−7.2 ± 0.2	−6.7 ± 0.1
B17	B23	0.0 ± 3.6	−2.4 ± 0.2	−3.7 ± 0.2	−2.6 ± 0.1
B17	B24	−1.4 ± 3.6	−1.8 ± 0.2	−2.3 ± 0.2	−1.9 ± 0.2
B17	B25	−0.4 ± 3.6	−3.4 ± 0.2	−4.7 ± 0.2	−3.8 ± 0.1
B17	B26	−19.8 ± 3.6	−22.1 ± 0.2	−23.0 ± 0.3	−22.3 ± 0.1
B17	B27	−17.2 ± 3.6	−20.1 ± 0.2	−20.6 ± 0.3	−20.2 ± 0.2
B17	B28	−20.0 ± 3.6	−22.8 ± 0.2	−23.3 ± 0.3	−22.6 ± 0.2
B18	B19	0.3 ± 3.6	0.3 ± 0.2	0.2 ± 0.1	0.3 ± 0.1
B18	B20	2.3 ± 2.6	0.7 ± 0.2	0.3 ± 0.2	0.8 ± 0.1
B18	B21	−3.8 ± 2.6	−3.8 ± 0.2	−4.2 ± 0.2	−4.0 ± 0.3
B18	B22	−2.0 ± 2.6	−3.3 ± 0.2	−3.4 ± 0.2	−3.1 ± 0.1
B18	B23	2.8 ± 3.6	1.4 ± 0.2	0.1 ± 0.2	1.0 ± 0.1
B18	B24	1.3 ± 3.6	2.0 ± 0.2	1.5 ± 0.2	1.7 ± 0.2
B18	B25	2.3 ± 3.6	0.4 ± 0.2	−0.9 ± 0.1	−0.2 ± 0.1
B18	B26	−17.0 ± 3.6	−18.3 ± 0.2	−19.2 ± 0.2	−18.7 ± 0.1
B18	B27	−14.4 ± 3.6	−16.3 ± 0.2	−16.8 ± 0.3	−16.6 ± 0.1
B18	B28	−17.2 ± 3.6	−19.0 ± 0.2	−19.5 ± 0.2	−19.0 ± 0.2
B19	B20	2.1 ± 2.6	0.4 ± 0.2	0.1 ± 0.3	0.4 ± 0.1
B19	B21	−4.1 ± 2.6	−4.0 ± 0.2	−4.4 ± 0.3	−4.4 ± 0.2
B19	B22	−2.3 ± 2.6	−3.6 ± 0.2	−3.6 ± 0.3	−3.5 ± 0.1
B19	B23	2.5 ± 3.6	1.2 ± 0.2	−0.1 ± 0.3	0.6 ± 0.2
B19	B24	1.0 ± 3.6	1.8 ± 0.2	1.3 ± 0.3	1.4 ± 0.3
B19	B25	2.1 ± 3.6	0.2 ± 0.2	−1.1 ± 0.2	−0.5 ± 0.2
B19	B26	−17.3 ± 3.6	−18.6 ± 0.2	−19.4 ± 0.2	−19.0 ± 0.2
B19	B27	−14.7 ± 3.6	−16.6 ± 0.2	−17.0 ± 0.2	−17.0 ± 0.1
B19	B28	−17.5 ± 3.6	−19.2 ± 0.2	−19.7 ± 0.2	−19.3 ± 0.3
B20	B21	−6.2 ± 1.2	−4.4 ± 0.2	−4.5 ± 0.2	−4.8 ± 0.2
B20	B22	−4.3 ± 1.2	−4.0 ± 0.2	−3.8 ± 0.3	−3.9 ± 0.1
B20	B23	0.5 ± 2.6	0.8 ± 0.2	−0.2 ± 0.3	0.2 ± 0.2
B20	B24	−1.0 ± 2.6	1.3 ± 0.2	1.2 ± 0.3	1.0 ± 0.2
B20	B25	0.0 ± 2.6	−0.3 ± 0.2	−1.2 ± 0.3	−0.9 ± 0.1
B20	B26	−19.4 ± 2.6	−19.0 ± 0.2	−19.5 ± 0.2	−19.4 ± 0.1
B20	B27	−16.8 ± 2.6	−17.0 ± 0.2	−17.1 ± 0.3	−17.4 ± 0.2
B20	B28	−19.5 ± 2.6	−19.7 ± 0.2	−19.9 ± 0.1	−19.7 ± 0.2
B21	B22	1.9 ± 1.2	0.5 ± 0.2	0.8 ± 0.2	0.9 ± 0.2
B21	B23	6.7 ± 2.6	5.2 ± 0.2	4.3 ± 0.2	5.0 ± 0.3
B21	B24	5.2 ± 2.6	5.8 ± 0.2	5.8 ± 0.2	5.8 ± 0.4
B21	B25	6.2 ± 2.6	4.2 ± 0.2	3.3 ± 0.2	3.9 ± 0.2
B21	B26	−13.2 ± 2.6	−14.6 ± 0.2	−15.0 ± 0.3	−14.7 ± 0.3
B21	B27	−10.6 ± 2.6	−12.6 ± 0.2	−12.6 ± 0.3	−12.6 ± 0.3
B21	B28	−13.3 ± 2.6	−15.2 ± 0.2	−15.3 ± 0.3	−14.9 ± 0.3
B22	B23	4.8 ± 2.6	4.7 ± 0.2	3.6 ± 0.3	4.1 ± 0.2
B22	B24	3.3 ± 2.6	5.3 ± 0.2	5.0 ± 0.2	4.9 ± 0.2
B22	B25	4.3 ± 2.6	3.7 ± 0.2	2.5 ± 0.2	3.0 ± 0.1
B22	B26	−15.1 ± 2.6	−15.0 ± 0.2	−15.8 ± 0.4	−15.5 ± 0.2
B22	B27	−12.5 ± 2.6	−13.0 ± 0.2	−13.3 ± 0.4	−13.5 ± 0.2
B22	B28	−15.2 ± 2.6	−15.7 ± 0.2	−16.1 ± 0.3	−15.8 ± 0.2
B23	B24	−1.5 ± 3.6	0.6 ± 0.2	1.4 ± 0.3	0.7 ± 0.2
B23	B25	−0.5 ± 3.6	−1.0 ± 0.2	−1.0 ± 0.3	−1.1 ± 0.2
B23	B26	−19.8 ± 3.6	−19.7 ± 0.2	−19.3 ± 0.3	−19.7 ± 0.2
B23	B27	−17.2 ± 3.6	−17.7 ± 0.2	−16.9 ± 0.3	−17.6 ± 0.2

B23	B28	−20.0 ± 3.6	−20.4 ± 0.2	−19.7 ± 0.3	−19.9 ± 0.2
B24	B25	1.0 ± 3.6	−1.6 ± 0.2	−2.4 ± 0.2	−1.9 ± 0.2
B24	B26	−18.4 ± 3.6	−20.3 ± 0.2	−20.7 ± 0.4	−20.4 ± 0.2
B24	B27	−15.8 ± 3.6	−18.3 ± 0.2	−18.3 ± 0.4	−18.3 ± 0.3
B24	B28	−18.5 ± 3.6	−21.0 ± 0.2	−21.1 ± 0.3	−20.7 ± 0.3
B25	B26	−19.4 ± 3.6	−18.7 ± 0.2	−18.3 ± 0.4	−18.5 ± 0.2
B25	B27	−16.8 ± 3.6	−16.7 ± 0.2	−15.9 ± 0.4	−16.5 ± 0.2
B25	B28	−19.5 ± 3.6	−19.4 ± 0.2	−18.6 ± 0.3	−18.8 ± 0.2
B26	B27	2.6 ± 3.6	2.0 ± 0.2	2.4 ± 0.1	2.1 ± 0.1
B26	B28	−0.2 ± 3.6	−0.7 ± 0.2	−0.3 ± 0.2	−0.3 ± 0.3
B27	B28	−2.8 ± 3.6	−2.7 ± 0.2	−2.8 ± 0.2	−2.3 ± 0.3
RMSE			1.98 ± 0.2	2.64 ± 0.3	2.43 ± 0.3
MAE			1.63 ± 0.2	2.17 ± 0.2	1.99 ± 0.2
r_{Spearman}			0.92	0.89	0.90
$t_{\text{preparation}}$			84 ns	549.1 ns	514.3 ns
$t_{\text{production}}$			2800 ns	112 ns	129 ns

7

Leveraging the Sampling Efficiency of RE-EDS in OpenMM Using a Shifted Reaction-Field With an Atom-Based Cutoff*

"Nature, it seems, is the popular name for milliards and milliards and milliards of particles playing their infinite game of billiards and billiards and billiards."

Atomyriades by Piet Hein [313]

Replica-exchange enveloping distribution sampling (RE-EDS) is a pathway-independent multistate free-energy method, currently implemented in the GROMOS software package for molecular dynamics (MD) simulations. It has a high intrinsic sampling efficiency as the interactions between the unperturbed particles have to be calculated only once for multiple end-states. As a result, RE-EDS is an attractive method for the calculation of relative solvation and binding free energies. An essential requirement for reaching this high efficiency is the

* This chapter is reproduced in part from Rieder, S. R.; Ries, B.; Kubincová, A.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Leveraging the Sampling Efficiency of RE-EDS in OpenMM Using a Shifted Reaction-Field With an Atom-Based Cutoff, *J. Phys. Chem.* **2022**, *157*, 104117.

separability of the nonbonded interactions into solute-solute, solute-environment, and environment-environment contributions. Such a partitioning is trivial when using a Coulomb term with a reaction-field (RF) correction to model the electrostatic interactions, but not when using lattice-sum schemes. To avoid cutoff artifacts, the RF correction is typically used in combination with a charge-group based cutoff, which is not supported by most small-molecule force fields and other MD engines. To address this issue, we investigate the combination of RE-EDS simulations with a recently introduced RF scheme including a shifting function that enables the rigorous calculation of RF electrostatics with atom-based cutoffs. The resulting approach is validated by calculating solvation free energies with the generalized AMBER force field (GAFF) in water and chloroform using both the GROMOS software package and a proof-of-concept implementation in OpenMM.

7.1 INTRODUCTION

Classical molecular dynamics (MD) simulations are a powerful tool to investigate molecular systems *in silico*, providing complementary insights to experiments. Within the discipline of computational chemistry, free-energy calculations are an important (albeit challenging) task, which is nowadays a routine part of computer-aided drug design workflows.[143–151] Thermodynamic integration (TI),[156] free-energy perturbation (FEP),[153] Bennett's acceptance ratio (BAR),[157] and multistate BAR (MBAR)[158] are examples of well-established pathway-dependent pairwise free-energy methods. In recent years, multistate free-energy methods, such as multisite λ-dynamics[159–161] and enveloping distribution sampling (EDS),[162,163] have emerged, enabling the calculation of pairwise free-energy differences for multiple end-states from a single simulation. While λ-dynamics is also

a pathway-dependent method, no pathway (*i.e.*, coupling parameter or coupling-parameter space) is specified in EDS, which offers additional flexibility for sampling. Replica-exchange EDS (RE-EDS)[164–166,175] and accelerated EDS (A-EDS)[167,168] are extensions of EDS that aim at increasing the quality and robustness of the obtained free-energy differences. Currently, both RE-EDS and A-EDS are implemented in the GROMOS MD engine.[27]

In EDS,[162,163] N end-states are combined into a reference-state potential V_R as,[162]

$$V_R\left(\mathbf{r}; s, \mathbf{E}^R\right) = -\frac{1}{\beta s} \ln \left[\sum_{i=1}^{N} e^{-\beta s \left(V_i(\mathbf{r}) - E_i^R \right)} \right] , \tag{7.1}$$

where V_i is (typically) the nonbonded potential energy of end-state i (*i.e.*, the intramolecular nonbonded energy of the end-state plus its nonbonded energy with the environment), $s > 0$ is the smoothness parameter, $\mathbf{E}^R$ is a vector of energy offsets, and $\beta = 1/(k_B\, T)$, where k_B is the Boltzmann constant and T the absolute temperature. While the parameter s smooths the potential-energy landscape and consequently decreases energy barriers, the energy offsets govern the contributions of the individual end-states to the reference-state potential. If the nonbonded interactions are rigorously pairwise separable, the interactions between the unperturbed particles in the system have to be calculated only once for V_R in Eq. (7.1) (*i.e.*, not for each end-state), leading to the inherent sampling efficiency of EDS.

The force resulting from the reference-state potential on a particle k

can be calculated by applying the chain rule as[162,163]

$$\begin{aligned}\mathbf{f}_k(t) &= -\frac{\partial V_R(\mathbf{r}; s, \mathbf{E}^R)}{\partial \mathbf{r}_k} \\ &= \sum_{i=1}^{N} \left[\frac{e^{-\beta s(V_i(\mathbf{r})-E_i^R)}}{\sum_{j=1}^{N} e^{-\beta s(V_j(\mathbf{r})-E_j^R)}} \left(-\frac{\partial V_i(\mathbf{r})}{\partial \mathbf{r}_k} \right) \right] \\ &= \sum_{i=1}^{N} \left[f_i^{\text{scal}} \left(-\frac{\partial V_i(\mathbf{r})}{\partial \mathbf{r}_k} \right) \right] ,\end{aligned} \tag{7.2}$$

i.e., the force contribution of each end-state potential V_i is scaled by the scaling factor

$$f_i^{\text{scal}} = \frac{e^{-\beta s(V_i(\mathbf{r})-E_i^R)}}{\sum_{j=1}^{N} e^{-\beta s(V_j(\mathbf{r})-E_j^R)}} = \frac{e^{-\beta s(V_i(\mathbf{r})-E_i^R)}}{e^{-\beta s V_R(\mathbf{r};s,\mathbf{E}^R)}} . \tag{7.3}$$

Note that by definition, it holds that $\sum_{i=1}^{N} f_i^{\text{scal}} = 1$.

From a single EDS simulation, the free-energy difference between any end-state pair in the system can be calculated as,[162,163]

$$\Delta G^{ji} = -\frac{1}{\beta} \ln \frac{\langle e^{-\beta(V_j - V_R)} \rangle_R}{\langle e^{-\beta(V_i - V_R)} \rangle_R} . \tag{7.4}$$

In practice, the accuracy of free-energy differences obtained from EDS simulations relies critically on the choice of the s-value and of the energy offsets.[247] To mitigate the choice of optimal parameters, EDS was combined with Hamiltonian replica exchange (RE)[271,272] to enhance sampling by simulating multiple EDS replicas with decreasing s-values (but constant $\mathbf{E}^R$), and attempting replica exchanges at fixed intervals,[164] following an idea introduced by Brooks and co-workers for constant pH simulations.[314] To decide whether the s-values of two replicas k and l with s-values s_i and s_j should be exchanged, a Metropolis-Hastings[42] criterion is employed.

The probability of an exchange is determined as,[164,272]

$$p_{k,l} = \begin{cases} 1 & \Delta \leq 0 \\ e^{-\beta\Delta} & \Delta > 0 \end{cases}, \tag{7.5}$$

with $\Delta = (V_R(\mathbf{r}_k; s_j) + V_R(\mathbf{r}_l; s_i)) - (V_R(\mathbf{r}_k; s_i) + V_R(\mathbf{r}_l; s_j))$, where $\mathbf{r}_k$ and $\mathbf{r}_l$ are the current coordinates of replicas k and l, respectively. In recent studies, RE-EDS was applied to calculate relative binding and hydration free energies for (*i*) molecules containing relatively large structural changes, such as R-group modifications, ring opening/closing, and ring size changes, and (*ii*) systems containing a large number of end-states.[166,175,254]

The calculation of the (pairwise) nonbonded interactions is usually the most expensive part of an MD simulation due to the large number (in principle $\mathcal{O}(N^2)$ for N particles) of particle pairs (depending on the functional form, potentially also triplets, etc.). To improve the computational efficiency, in practice, the nonbonded interactions are only calculated explicitly within a given cutoff distance. The interactions beyond the cutoff are either neglected completely, resulting in a truncation of the nonbonded potential energy, or approximated using a (mean-field or periodic) long-range correction. While a straight truncation is less problematic for van der Waals interactions,[315] it can lead to serious cutoff artifacts for electrostatic interactions.[107–111] Typically, the long-range electrostatic interactions are therefore approximated by employing either a reaction-field (RF) correction[114] or a lattice-sum scheme such as Ewald summation,[116] particle-particle particle-mesh (P3M),[117] or particle-mesh Ewald (PME).[99,118] Recently, Kubincová *et al.* proposed a shifted RF correction that avoids the occurrence of artifacts at the cutoff.[304] In the context of RE-EDS simulations, using a RF correction is particularly convenient, as the nonbonded potential-energy contribution of the different end-states can easily be separated. Such a partitioning is required to calculate the reference-state potential V_R (Eq. (7.1)) efficiently. For lattice-sum schemes, on the other hand, additional fast Fourier transfor-

mations (FFTs)[316] would be required to achieve a partitioning of the end-state energies.[317]

In the present study, we investigate the use of the shifted RF scheme by Kubincová *et al.*[304] in RE-EDS simulations, such that the sampling efficiency of RE-EDS is retained while enabling a rigorously conservative treatment of force fields with QM-derived charges (without the need of charge re-distributions to achieve neutral charge groups) and facilitating the implementation of RE-EDS in additional MD software packages. For this, different choices for the treatment of the electrostatic energy (*i.e.*, functional form, cutoff distance, atom or charge-group based cutoff) are first compared in the context of solvation free-energy calculations with RE-EDS in the GROMOS MD engine.[27] Solvation free energies in water and in chloroform are used as a straightforward (and computationally relatively cheap) test system to compare methods,[170,172] and implementations. Second, a new implementation of RE-EDS in OpenMM[28,29] is presented and tested using the shifted RF scheme of Kubincová *et al.*[304] The results are compared to the experimental and calculated values reported in the FreeSolv[155,173] and Minnesota solvation[174] databases.

7.2 THEORY

7.2.1 Reaction-Field Correction for Long-Range Electrostatics

In the GROMOS MD engine,[26,27] long-range electrostatic interactions are usually handled by employing a reaction-field (RF) correction,[114]

$$V^{\mathrm{ele}} = \sum_i \sum_{j \in \mathrm{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_{\mathrm{cs}}} \left[\frac{1}{r_{ij}} - \frac{C_{\mathrm{RF}} r_{ij}^2}{2R_{\mathrm{RF}}^3} - \frac{1 - 0.5 C_{\mathrm{RF}}}{R_{\mathrm{RF}}} \right], \qquad (7.6)$$

where q_i and q_j are the charges of atoms i and j, ε_0 is the permittivity of vacuum, $\varepsilon_{\mathrm{cs}}$ is the background dielectric permittivity, r_{ij} is the (minimum-

image) distance between atoms i and j, and R_{RF} is the cutoff distance for the pairlist construction.[33] The notation $j \in \mathrm{PL}(i)$ indicates an atom j with $j > i$, where j is in the pairlist of i. The constant C_{RF} characterizes the effect of the RF continuum as,[33,114]

$$C_{\mathrm{RF}} = \frac{(2\varepsilon_{\mathrm{cs}} - 2\varepsilon_{\mathrm{RF}})(1 + \kappa_{\mathrm{RF}} R_{\mathrm{RF}}) - \varepsilon_{\mathrm{RF}}(\kappa_{\mathrm{RF}}\ R_{\mathrm{RF}})^2}{(\varepsilon_{\mathrm{cs}} + 2\varepsilon_{\mathrm{RF}})(1 + \kappa_{\mathrm{RF}} R_{\mathrm{RF}}) + \varepsilon_{\mathrm{RF}}(\kappa_{\mathrm{RF}}\ R_{\mathrm{RF}})^2}\,, \tag{7.7}$$

where $\varepsilon_{\mathrm{RF}}$ is the RF permittivity, and κ_{RF} is the inverse Debye screening length.[34] In simulations with explicit solvent, $\varepsilon_{\mathrm{cs}}$ is usually set to one and κ_{RF} is set to zero.[34] In this case, the RF contribution can be calculated as

$$V^{\mathrm{ele}} = \sum_i \sum_{j \in \mathrm{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} \left[\frac{1}{r_{ij}} + A_{\mathrm{RF}} r_{ij}^2 - B_{\mathrm{RF}} \right], \tag{7.8}$$

where the parameters A_{RF} and B_{RF} are calculated as

$$A_{\mathrm{RF}} = \frac{\varepsilon_{\mathrm{RF}} - 1}{1 + 2\varepsilon_{\mathrm{RF}}} \frac{1}{R_{\mathrm{RF}}^3}\,, \tag{7.9}$$

and

$$B_{\mathrm{RF}} = \frac{1}{R_{\mathrm{RF}}} + A_{\mathrm{RF}} R_{\mathrm{RF}}^2 = \frac{3\varepsilon_{\mathrm{RF}}}{1 + 2\varepsilon_{\mathrm{RF}}} \frac{1}{R_{\mathrm{RF}}}\,. \tag{7.10}$$

In the GROMOS MD engine, there is additionally an RF contribution for excluded neighbors as well as a self-interaction. The self-interaction may be interpreted as the reversible work needed to individually charge the atoms at infinite separation.[303] Thus, the total electrostatic potential

energy is calculated as[231,310,312]

$$V^{\mathrm{ele,orig}} = \sum_i \sum_{j \in \mathrm{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} \left[\frac{\alpha_{ij}}{r_{ij}} + A_{\mathrm{RF}} r_{ij}^2 - B_{\mathrm{RF}} \right] - \frac{1}{2} \frac{B_{\mathrm{RF}}}{4\pi\varepsilon_0} \left[\sum_i q_i^2 - \frac{1}{\varepsilon_{\mathrm{RF}}} \left(\sum_i q_i \right)^2 \right] , \tag{7.11}$$

where α_{ij} is set to zero if atoms i and j are excluded neighbors, and to one otherwise. Note that the sum over all charges, $\sum_i q_i$, is zero for neutral systems and the corresponding term $\varepsilon_{\mathrm{RF}}^{-1} \left(\sum_i q_i\right)^2$ is currently not implemented in GROMOS. When using AMBER/GAFF topologies[55] in GROMOS,[175] the scaling of the electrostatic 1,4-interactions by a factor 1/1.2[55,292] is accounted for by setting α_{ij} to 1/1.2 if atoms i and j are third neighbors. It should be noted that the choice of B_{RF} in Eq. (7.10) only leads to an interaction energy that is continuous at R_{RF} for normal pairs, *i.e.*, those for which $\alpha_{ij} = 1$. Although continuity could be enforced also in this case by adjusting B_{RF} for these pairs, this choice is not made here for two reasons: (*i*) excluded and third-neighbor pairs correspond to close covalent neighbors, so that they are and stay within the cutoff throughout a simulation; and (*ii*) close-neighbor modifications are meant to reduce the direct Coulombic interactions between the atoms without altering their effective interactions *via* the environment. Further, note that the B_{RF} term does not have an inherent physical meaning. It merely ensures that the electrostatic potential is continuous (and thus differentiable) at the cutoff R_{RF}. A "physical alternative" to $V^{\mathrm{ele,orig}}$ is therefore

$$V^{\mathrm{ele,phys}} = \sum_i \sum_{j \in \mathrm{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} \left[\frac{\alpha_{ij}}{r_{ij}} + A_{\mathrm{RF}} r_{ij}^2 \right] . \tag{7.12}$$

For computational efficiency and to avoid cutoff noise when employing a straight-cutoff scheme, multiple atoms of a molecule can be grouped into

"charge groups" (CGs) with integer total charges.[318] In the GROMOS MD engine, the position of a solute CG is calculated as its center of geometry, whereas the position of a solvent CG is set to the position of the first atom of the solvent molecule (*e.g.*, the oxygen atom for water).[34] When employing a CG based cutoff, the pairlist algorithm takes the CG positions into account instead of the individual atom coordinates to determine whether two atoms are currently within the cutoff (Figure 7.1, right). GROMOS (compatible) force fields typically make use of CGs,[33,65,68,72–74] whereas most other force-field families use an atom based (AT) cutoff (Figure 7.1, left).[33]

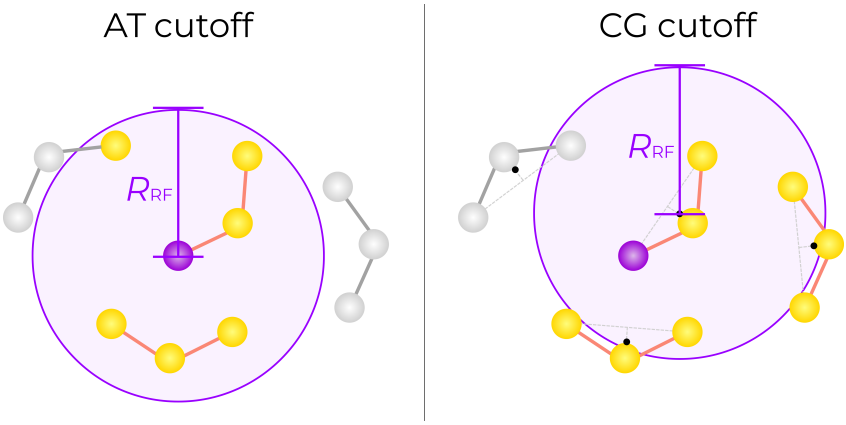


Figure 7.1: Schematic illustration of atom based (AT) and charge-group based (CG) cutoff for nonbonded interactions. The currently considered atom is colored in purple, and the atoms in its pairlist are colored in yellow, whereas the atoms outside the pairlist are colored in grey. Here, the CGs are defined such that they each contain all three atoms of the depicted molecules. Note that the configurations and currently considered atom are identical on the left and on the right. (Left): For AT cutoff, all atoms within R_{RF} of the current atom are considered for the nonbonded interactions. (Right): For CG cutoff, all atoms belonging to CGs whose center of geometry (COG, black dot) is within R_{RF} of the COG of the current atom's CG are considered for the nonbonded interactions. Note that instead of the COG, sometimes the first atom of the CG is used as the reference point (*e.g.*, for solvent molecules in GROMOS).[34]

CG cutoff is implemented *e.g.* in the GROMOS[27] and GROMACS[25] MD engines (though its use is deprecated in the latter since GROMACS 5.0[319]). In the AMBER[21] and OpenMM[29] MD engines, on the other

hand, only AT cutoff is implemented. As the AMBER force fields use QM-derived partial charges, no CGs are defined. In the previous Chapter 6 on converting AMBER topologies with the *amber2gromos* program for RE-EDS simulations using the GROMOS MD engine,[175] the solute molecules were small enough that a single CG per molecule could be justified. For larger molecules, however, this is no longer appropriate. To address this issue, we want to bypass CGs altogether by employing an AT cutoff in the RE-EDS simulations. This is achieved using the RF scheme with a shifting function developed by Kubincová *et al.*[304] The scheme significantly reduces cutoff artifacts, *i.e.*, cutoff noise in the radial distribution functions and dipole-dipole orientation correlation functions of several model liquids.

7.2.2 Reaction-Field Scheme for Atom Based Cutoff

In the RF scheme with a shifting function, the electrostatic potential energy is defined as,[304]

$$\begin{aligned} V^{\mathrm{ele,shift}} = & \sum_i \sum_{j \in \mathrm{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} \left[\frac{\alpha_{ij}}{r_{ij}} + A_{\mathrm{RF}} r_{ij}^2 + a_{\mathrm{RF},4} r_{ij}^4 + a_{\mathrm{RF},6} r_{ij}^6 - B_{\mathrm{RF}}^{\mathrm{shift}} \right] \\ & - \frac{1}{2} \frac{B_{\mathrm{RF}}^{\mathrm{shift}}}{4\pi\varepsilon_0} \left[\sum_i q_i^2 - \frac{1}{\varepsilon_{\mathrm{RF}}} \left(\sum_i q_i \right)^2 \right] , \end{aligned} \tag{7.13}$$

where $a_{\mathrm{RF},4}$ and $a_{\mathrm{RF},6}$ are shifting parameters chosen such that the potential-energy function is not modified at $r_{ij} \to 0$, is continuous at the cutoff, is constant at and beyond the cutoff, the exerted force (*i.e.*, first derivative) resulting from it is zero at the cutoff, and the exerted continuous force (*i.e.*, second derivative) is also zero at the cutoff.[304] Further, $B_{\mathrm{RF}}^{\mathrm{shift}}$ is defined as

$$B_{\mathrm{RF}}^{\mathrm{shift}} = B_{\mathrm{RF}} + a_{\mathrm{RF},4} R_{\mathrm{RF}}^4 + a_{\mathrm{RF},6} R_{\mathrm{RF}}^6 \, . \tag{7.14}$$

Note that in the modified GROMOS implementation used in the current chapter, $V^{\text{ele,shift}}$ is calculated *via* the equivalent term[304]

$$V^{\text{ele,shift}} = \sum_i \sum_{j \in \text{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} \left[\frac{\alpha_{ij}}{r_{ij}} + A_{\text{RF}} r_{ij}^2 + a_{\text{RF},4} r_{ij}^4 + a_{\text{RF},6} r_{ij}^6 \right] + \sum_i \sum_{j \notin \text{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} B_{\text{RF}}^{\text{shift}} . \tag{7.15}$$

The second term of Eq. (7.15) (*i.e.*, the sum over the atom pairs outside the pairlist) can be calculated as,[304]

$$\sum_i \sum_{j \notin \text{PL}(i)} \frac{q_i q_j}{4\pi\varepsilon_0} B_{\text{RF}}^{\text{shift}} = \frac{B_{\text{RF}}^{\text{shift}}}{4\pi\varepsilon_0} \left[\sum_i \sum_{j>i} q_i q_j - \sum_i \sum_{j \in \text{PL}(i)} q_i q_j \right] . \tag{7.16}$$

Note that in Eq. (7.16), the prefactor $B_{\text{RF}}^{\text{shift}}/(4\pi\varepsilon_0)$ and the first term in the brackets are configuration independent, *i.e.*, they need to be calculated only once at the start of the simulation and do not induce any atomic forces.[304] It can be shown that for neutral systems (*i.e.*, $\left(\sum_i q_i\right)^2 = 0$), Eq. (7.13) and Eq. (7.15) are equivalent. In the near future, the modified GROMOS implementation will be adapted to use Eq. (7.13) instead as this formulation is more similar to the implementation of the native GROMOS RF and will require fewer modifications to the outer and inner nonbonded loops. The corresponding code will be included in the next release of GROMOS. An illustration of the three presented functional forms of V^{ele} is provided in Figure 7.2.

The modified electrostatic energy term $V^{\text{ele,shift}}$ can easily be "corrected" to $V^{\text{ele,orig}}$ (Eq. (7.11)) or $V^{\text{ele,phys}}$ (Eq. (7.12)) for a given energy trajectory by keeping track of the extra energy (*i.e.*, the difference between $V^{\text{ele,shift}}$ and $V^{\text{ele,orig}}$ or $V^{\text{ele,phys}}$, respectively) during the simulation. For a RE-EDS simulation in which $V^{\text{ele,shift}}$ was used to propagate the system,

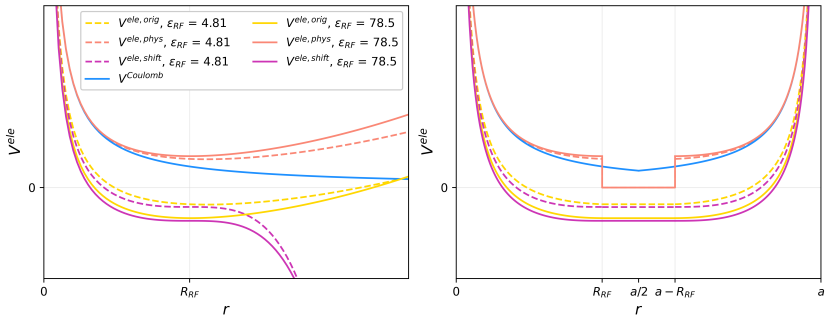


Figure 7.2: Schematic illustration of different functional forms of V^{ele}. Three different functional forms of V^{ele} are shown for the electrostatic interaction between two model particles ($q_0, q_1 > 0$), namely $V^{\mathrm{ele,orig}}$ (Eq. (7.11), yellow), $V^{\mathrm{ele,phys}}$ (Eq. (7.12), orange), and $V^{\mathrm{ele,shift}}$ (Eq. (7.13), purple), for two different RF permittivities, namely $\varepsilon_{\mathrm{RF}} = 4.81$ (chloroform, dashed lines) and $\varepsilon_{\mathrm{RF}} = 78.5$ (water, solid lines).[309,310] Additionally, $V_{0,1}^{\mathrm{Coulomb}} = q_0 q_1 / (4\pi\varepsilon_0 r_{0,1})$ is depicted in blue. (Left): The potential-energy functions are shown for a non-periodic system without cutoff truncation (note that the use of RF without a spherical cutoff represents a hypothetical scenario and is not used in practice). (Right): The potential-energy functions are shown with the cutoff R_{RF} in a periodic box with box length a.

the corrected end-state energy for end-state i can then be calculated as

$$V_i^{\mathrm{shift}\to\mathrm{orig}} = V_i + V_i^{\mathrm{extra,orig}} \tag{7.17}$$

$$V_i^{\mathrm{shift}\to\mathrm{phys}} = V_i + V_i^{\mathrm{extra,phys}} \,. \tag{7.18}$$

The corrected reference-state energy can then be calculated as

$$V_R^{\mathrm{shift}\to\mathrm{orig}} = -\frac{1}{\beta s} \ln \left[\sum_{i=1}^{N} e^{-\beta s \left(V_i + V_i^{\mathrm{extra,orig}} - E_i^R \right)} \right] \tag{7.19}$$

$$V_R^{\mathrm{shift}\to\mathrm{phys}} = -\frac{1}{\beta s} \ln \left[\sum_{i=1}^{N} e^{-\beta s \left(V_i + V_i^{\mathrm{extra,phys}} - E_i^R \right)} \right] . \tag{7.20}$$

This allows to propagate the system based on the modified electrostatic potential $V^{\mathrm{ele,shift}}$, avoiding artifacts at the cutoff, but calculate the Hamiltonian and free-energy differences based on the energies obtained with the "original" GROMOS electrostatic energies, or the "physical" electrostatic

energies. The former option potentially achieves more accurate results when using force fields that were parameterized with $V^{\text{ele,orig}}$ (*i.e.*, most GROMOS or GROMOS-compatible force fields), while the latter option might provide more accurate short-range interaction energies.

The modified functional form of V^{ele} (*i.e.*, $V^{\text{ele,shift}}$) allows for the use of a RF correction with an AT cutoff, while avoiding artifacts at the cutoff. As the partitioning of the end-state energies is trivial with such an RF correction, it can conveniently be combined with RE-EDS. This is a key advantage for porting the RE-EDS free-energy method to MD engines that do not support the use of CG cutoff.

7.3 METHODS

7.3.1 Comparison of Functional Forms of the Electrostatic Potential Energy

To assess whether RE-EDS calculations in GROMOS achieve accurate free-energy estimates with an AT cutoff, free-energy differences in vacuum ($\Delta G^{ji}_{\text{vac}}$), water ($\Delta G^{ji}_{\text{wat}}$), and chloroform ($\Delta G^{ji}_{\text{CHCl}_3}$) were calculated with RE-EDS for two sets of molecules (see Sec. 7.3.4). The corresponding relative hydration free energies ($\Delta\Delta G^{ji}_{\text{hyd}} = \Delta G^{ji}_{\text{wat}} - \Delta G^{ji}_{\text{vac}}$) and relative solvation free energies in chloroform ($\Delta\Delta G^{ji}_{\text{CHCl}_3} = \Delta G^{ji}_{\text{CHCl}_3} - \Delta G^{ji}_{\text{vac}}$) were calculated and subsequently compared to experimental and calculated reference values ($\Delta\Delta G^{ji}_{\text{hyd}} = \Delta G^{j}_{\text{hyd}} - \Delta G^{i}_{\text{hyd}}$ and $\Delta\Delta G^{ji}_{\text{CHCl}_3} = \Delta G^{j}_{\text{CHCl}_3} - \Delta G^{i}_{\text{CHCl}_3}$, respectively).

Three schemes to calculate V^{ele} were compared: (*i*) $V^{\text{ele,orig}}$ with CG cutoff; (*ii*) $V^{\text{ele,orig}}$ with AT cutoff; and (*iii*) $V^{\text{ele,shift}}$ with AT cutoff. In the following, the combination of $V^{\text{ele,orig}}$ with a CG cutoff will be referred to as CG^{orig}, $V^{\text{ele,orig}}$ with an AT cutoff will be referred to as AT^{orig}, and $V^{\text{ele,shift}}$ with an AT cutoff will be referred to as AT^{shift}. Additionally, for the simulations performed using AT^{shift}, we investigated whether using

the corrected end-state and reference-state potential energies $V^{\text{shift}\rightarrow\text{orig}}$ and $V^{\text{shift}\rightarrow\text{pyhs}}$ (Eqs. (7.17) – (7.20)) would improve the accuracy of free-energy calculations with RE-EDS.

7.3.2 RE-EDS Implementation in OpenMM

A proof-of-concept implementation of RE-EDS using the OpenMM MD engine was developed. It consists of a simple Python3 module that relies on the *openmm*,[29] *parmed*,[293] *numpy*,[285] and *pandas*[283] modules. The source code is available at `https://github.com/rinikerlab/reeds/blob/openmm/reeds/openmm/reeds_openmm.py` with example scripts provided at `https://github.com/rinikerlab/reeds/tree/main/examples/openmm`.

Calculation of the Electrostatic Potential Energy

For RE-EDS to be efficient, the separation of the nonbonded potential-energy contributions from the solute-solute, solute-environment, and environment-environment interactions is essential. While it is trivial for the van der Waals interactions, we decided to use $V^{\text{ele,shift}}$ (Eq. (7.13)) for the electrostatic interactions.

OpenMM provides the possibility to create so-called *custom forces*.[29] A custom force is defined by an algebraic expression representing an interaction between particles (or between a particle and an external force), *i.e.*, a potential-energy term, which is then analytically differentiated by OpenMM to obtain the resulting force.[29] In our module, the nonbonded interactions (including the shifted RF) are implemented *via* four *CustomNonbondedForce* and *CustomBondForce* terms for each end-state. The per particle and per bond parameters required to characterize the custom forces are taken from the default *NonbondedForce* of the system. Note that the parameters are loaded *via ParmEd*,[293] allowing the conversion of different topology formats to a format compatible with OpenMM. Upon creation of the custom forces, the default *NonbondedForce* is removed from

the system and replaced by the custom forces. For each EDS end-state, a separate force group is defined containing the four custom forces of that end-state, allowing for the separate evaluation of the nonbonded potential energies of the different end-states.

First, for each end-state k, a *CustomNonbondedForce* term is created as,

$$\begin{aligned} V_{k,ij}^{\mathrm{LJ,CRF}} =& 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \\ &+\frac{q_i q_j}{4\pi\varepsilon_0}\left(\frac{1}{r_{ij}}+A_{\mathrm{RF}}r_{ij}^2+a_{\mathrm{RF},4}r_{ij}^4+a_{\mathrm{RF},6}r_{ij}^6-B_{\mathrm{RF}}^{\mathrm{shift}}\right), \end{aligned} \tag{7.21}$$

to account for the nonbonded interactions of particles i and j that are neither third-neighbor nor excluded pairs. For each end-state, the interaction groups are defined such that the particles of the current end-state interact with the other particles of the current end-state, as well as with the environment particles. Note that the particles of an end-state do not include any dummy particles (*i.e.*, the number of particles in end-state i is not necessarily the same as the number of particles in end-state j), which is different from the GROMOS implementation. In addition to the force groups for the end-states, there is one instance of $V_{\mathrm{env},ij}^{\mathrm{LJ,CRF}}$ (analogous to Eq. (7.21)) that accounts for the intramolecular and intermolecular interactions of the environment particles.

Next, for each end-state k (plus the environment), a *CustomBondForce* term is created as,

$$\begin{aligned} V_{k,ij}^{\mathrm{LJ,CRF,1\text{-}4}} =& 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12}-\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]+\frac{q_{ij}^{\mathrm{scal}}}{4\pi\varepsilon_0}\frac{1}{r_{ij}} \\ &+\frac{q_{ij}}{4\pi\varepsilon_0}\left(A_{\mathrm{RF}}r_{ij}^2+a_{\mathrm{RF},4}r_{ij}^4+a_{\mathrm{RF},6}r_{ij}^6-B_{\mathrm{RF}}^{\mathrm{shift}}\right), \end{aligned} \tag{7.22}$$

to account for the third-neighbor interactions. Here, a "bond" is added for each third-neighbor pair of the current end-state molecule (or the environment molecules). The parameter q_{ij}^{scal} corresponds to the scaled charge product of the excluded atom pair according to the default *Non-*

bondedForce, whereas $q_{ij} = q_i q_j$ is the unscaled charge product calculated from the particle charges.

Next, another *CustomBondForce* term is defined as,

$$V_{k,ij}^{\mathrm{CRF,excl}} = \frac{q_{ij}}{4\pi\varepsilon_0} \left(A_{\mathrm{RF}} r_{ij}^2 + a_{\mathrm{RF},4} r_{ij}^4 + a_{\mathrm{RF},6} r_{ij}^6 - B_{\mathrm{RF}}^{\mathrm{shift}} \right), \quad (7.23)$$

for each end-state k (plus the environment) to account for the RF contribution of the excluded atom pairs. Here, a "bond" is added for each pair of excluded atoms of the current end-state molecule (or the environment molecules).

Finally, for each end-state k (plus the environment), the *CustomBondForce* term is defined as,

$$V_{k,ii}^{\mathrm{self}} = -\frac{1}{2} \frac{B_{\mathrm{RF}}^{\mathrm{shift}}}{4\pi\varepsilon_0} q_{ii}, \quad (7.24)$$

to account for the self-term. Here, a "bond" is added for each atom with itself, and the parameter q_{ii} is calculated as $q_{ii} = q_i q_i$. Note that, analogous to the GROMOS implementation, the term $\varepsilon_{\mathrm{RF}}^{-1} \left(\sum_i q_i \right)^2$ of Eq. (7.13) (which is zero for a neutral system) is currently not implemented for simplicity, but can easily be added.

EDS Integration

To perform an integration step of an EDS simulation, the nonbonded forces of the end-states have to be scaled according to Eq. (7.2). This is achieved in OpenMM *via* a four-step process: (*i*) calculate the nonbonded potential energies of all end-states (Eqs. (7.21) - (7.24)); (*ii*) calculate the scaling factor f_i^{scal} (Eq. (7.3)) for each end-state i based on the end-state potential energies calculated in step (*i*); (*iii*) scale the nonbonded force resulting from each end-state i by the corresponding scaling factor f_i^{scal}; and (*iv*) perform an OpenMM simulation step. For the scaling of the nonbonded forces, the nonbonded potential-energy terms (Eqs. (7.21) - (7.24)) of each end-state i are multiplied by the scaling factor

f_i^{scal} obtained in step (*ii*). In other words, the force resulting from the reference-state potential V_R on a particle k is calculated as

$$\mathbf{f}_k(t) = \sum_{i=1}^{N} \left[-\frac{\partial \left(f_i^{\text{scal}} V_i(\mathbf{r}) \right)}{\partial \mathbf{r}_k} \right] . \tag{7.25}$$

Note that this integration scheme results in the custom forces of each end-state being evaluated twice per integration step. First, to calculate the energies and scaling factors, and second, to calculate the forces. This introduces of course an inefficiency but is necessary to remain in the Python layer of OpenMM. It will be improved in a future implementation.

Atomic Distance Restraints

When using a dual topology approach,[258,261] the coordinates of all molecules are explicitly present during a (RE-)EDS simulation.[254] To prevent the molecules from drifting apart during the simulation, atomic distance restraints can be employed.[247,254] Here, the distance restraint for a restrained atom pair i and j was added as a *CustomBondForce* with the harmonic potential

$$V_{ij}^{\text{restr}} = \frac{1}{2} K r_{ij}^2 , \tag{7.26}$$

where K is the force constant of the distance restraint.

RE-EDS Simulation

The previous sections describe the three ingredients needed to implement an EDS simulation with OpenMM: separable nonbonded energy terms, atomic distance restraints, and an integration procedure. For a RE-EDS simulation, two more ingredients are needed: (*i*) EDS simulations of independent replicas at different s-values at the same time; and (*ii*) replica exchanges. In the current "proof-of-concept" implementation, the replicas are integrated serially on the GPU, *i.e.*, n EDS integration steps are performed for the first replica (where n is the number of time steps

between exchanges), then n steps are performed for the next replica, etc. This is of course inefficient and will be parallelized in a future implementation. After the n steps have been performed for each replica, replica exchanges are attempted. When comparing two neighboring replicas, the exchange probability is calculated according to Eq. (7.5).

7.3.3 RE-EDS Pipeline

To perform RE-EDS simulations in GROMOS, the pipeline recently proposed by Ries *et al.*[166] was used. The RE-EDS pipeline is carried out using the Python3[211] *reeds* module freely available on Github at `https://github.com/rinikerlab/reeds`.[278] It consists of three main steps: parameter exploration, parameter optimization, and production. During the parameter exploration, relevant configurations are generated for all end-states, a lower bound is determined for the s-values, and initial estimates for the energy offsets are generated. Next, during the parameter optimization, the distribution of the s-values and the values of the energy offsets $\mathbf{E}^R$ are optimized such that frequent round-trips are observed for all replicas, and all end-states are sampled approximately equally at $s = 1$. Finally, the free-energy differences between all end-state pairs are calculated from a RE-EDS production run with the optimized parameters. To perform the RE-EDS calculations in OpenMM, a slightly modified version of the *reeds* module was used.

7.3.4 Datasets

In order to (a) assess the performance of the different RF schemes and cut-off values; and (b) validate the implementation of RE-EDS in OpenMM, three different sets of molecules were considered. A set of six benzene derivatives (labeled set A) was taken from a previous study (see Chapter 6, Sec. 6.3.3).[175] Set A consists of six benzene derivatives selected

from the FreeSolv[155,173] database with available experimental and calculated reference data. In addition, two new sets of molecules (labeled set C and set D, respectively) were assembled. For both sets, the selected molecules were contained in both the FreeSolv and the Minnesota (chloroform) solvation[174] database, such that experimental reference data was available both for hydration free energies as well as solvation free energies in chloroform. This allowed the comparison of the different RF schemes and cutoff values both in a high permittivity (water) and in a low permittivity (chloroform) environment. Set C contains 14 benzene derivatives with small substituents (Figure 7.3 top) and is a subset of set B from Chapter 6 (*i.e.*, the molecules of set B for which solvation free energies in chloroform were available from the Minnesota solvation database). Set D consists of thirteen benzene, pyridine, and pyrazine derivatives with larger substituents (Figure 7.3 bottom).

Table 7.1: Sets C and D. Molecule ID, molecule identifier in the FreeSolv database, molecule identifier in the Minnesota solvation database, SMILES string, and name of the 14 benzene derivatives of set C (top) and of the 13 molecules of set D (bottom).[155,173,174]

Set	Molecule	ID FreeSolv	ID Minnesota	SMILES	Name
C	C1	mobley_1873346	0036tol	Cc1ccccc1	toluene
	C2	mobley_8127829	0037eth	CCc1ccccc1	ethylbenzene
	C3	mobley_9478823	0038oxy	Cc1ccccc1C	*o*-xylene
	C4	mobley_1424265	0039mxy	Cc1cccc(c1)C	*m*-xylene
	C5	mobley_20524	0053phe	c1ccc(cc1)O	phenol
	C6	mobley_2925352	0057pcr	Cc1ccc(cc1)O	*p*-cresol
	C7	mobley_4883284	0118ani	c1ccc(cc1)N	aniline
	C8	mobley_4483973	0157flu	c1ccc(cc1)F	fluorobenzene
	C9	mobley_7608462	0174chl	c1ccc(cc1)Cl	chlorobenzene
	C10	mobley_4553008	0176pdi	c1cc(ccc1Cl)Cl	1,4-dichlorobenzene
	C11	mobley_7599023	0186bro	c1ccc(cc1)Br	bromobenzene
	C12	mobley_3187514	n009	Cc1ccccc1N	2-methylaniline
	C13	mobley_5518547	n011	Cc1ccc(cc1)N	4-methylaniline
	C14	mobley_3398536	test4001	c1ccc(cc1)I	iodobenzene
D	D1	mobley_4035953	0055ocr	Cc1ccccc1O	*o*-cresol
	D2	mobley_7295828	0068ani	COc1ccccc1	anisole
	D3	mobley_3969312	0074ben	c1ccc(cc1)C=O	benzaldehyde
	D4	mobley_7497999	0084met	CC(=O)c1ccccc1	1-phenylethanone
	D5	mobley_4287564	0119met	Cc1ccccn1	2-methylpyridine
	D6	mobley_5977084	0120met	Cc1cccnc1	3-methylpyridine
	D7	mobley_1520842	0121met	Cc1ccncc1	4-methylpyridine
	D8	mobley_4584540	0125dim	Cc1cccc(n1)C	2,6-dimethylpyridine
	D9	mobley_7988076	0151phy	c1cc(ccc1C=O)O	4-hydroxybenzaldehyde
	D10	mobley_1733799	0215pbr	c1cc(ccc1O)Br	4-bromophenol
	D11	mobley_5220185	0230eth	CCc1cnccn1	2-ethylpyrazine
	D12	mobley_3968739	0240met	COC(=O)c1ccccc1	methyl benzoate
	D13	mobley_2763835	0246eth	CCOc1ccccc1	ethoxybenzene

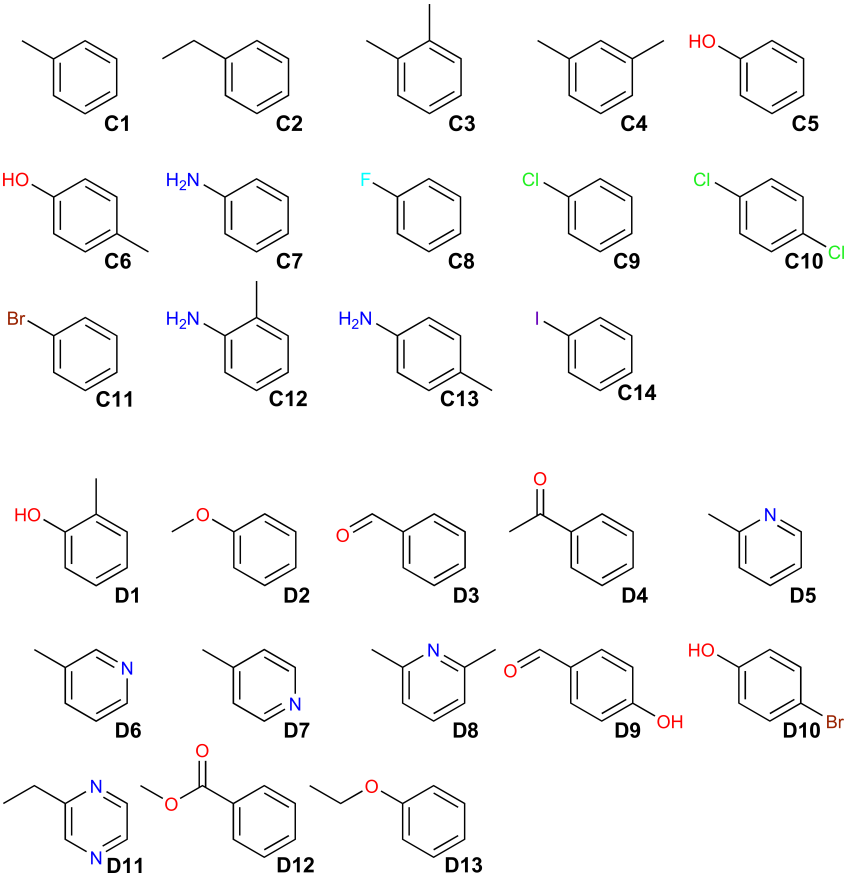


Figure 7.3: (Top): Set C consists of 14 benzene derivatives, selected from the FreeSolv [155,173] and the Minnesota (chloroform) solvation [174] database. A list of the corresponding molecule indices, the FreeSolv identifiers, the Minnesota solvation database identifiers, the SMILES strings, and the names of the molecules can be found in Table 7.1, top. (Bottom): Set D consists of 13 benzene, pyridine, and pyrazine derivatives, selected from the FreeSolv [155,173] and the Minnesota (chloroform) solvation [174] database. A list of the corresponding molecule indices, the FreeSolv identifiers, the Minnesota solvation database identifiers, the SMILES strings, and the names of the molecules can be found in Table 7.1, bottom.

7.3.5 Simulation Details

All topologies were generated with *tleap* (AmberTools16) [292] based on the mol2 and frcmod files provided in the FreeSolv database. [155,173] For the

simulations in GROMOS, the topologies were converted with *amber2gromos* and for the simulations in OpenMM, the topologies were converted using *ParmEd*.[293]

The partial charges were generated with antechamber[55,133,134] using the AM1-BCC[306,307] method. The input files for the RE-EDS simulations in GROMOS were prepared using *amber2gromos* as well as the GROMOS++[282] programs *pdb2g96*, *red_top*, and *prep_eds*.[175] The alignment of the molecules was generated with the RDKit[210] module *rdFMCS* and the *AllChem.AlignMol* function. The aligned molecule pairs were selected manually to maximize the overlap of the respective aromatic rings and substituents. For some of the molecule pairs, all atom types were matched (*rdFMCS.AtomCompare.CompareAny*) whereas for others, only heavy atom types were matched (*rdFMCS.AtomCompare.CompareAnyHeavyAtom*).[175] To prevent the molecules from drifting apart from each other during the RE-EDS simulations, atomic distance restraints were employed. They were generated with *RestraintMaker* (see Chapter 5).[254] The same atom pairs were restrained for the simulations in GROMOS and in OpenMM.

The RE-EDS simulations in GROMOS were performed with a modified version of GROMOS[27,308] 1.5.0. The RE-EDS pipeline was carried out using the open-source *reeds* module.[278] The RE-EDS simulations in OpenMM were performed with OpenMM version 7.7.0[29] and the RE-EDS pipeline was carried out using a slightly modified version of the *reeds* module. The TIP3P[298] water model was used for the simulations in water. The integration time step was set to 2 fs. The RF permittivity $\varepsilon_{\mathrm{RF}}$ was set to 1 in vacuum, 4.81 in chloroform,[115] and 78.5 in water.[309,310,320] For the simulations in GROMOS, three different choices of R_{RF} (*i.e.*, 0.8, 1.0, and 1.2 nm) were compared. For the simulations in OpenMM, R_{RF} was set to 1.2 nm. The temperature was maintained at 298.15 K in all environments and the pressure at 0.06102 kJ mol^{-1} nm^{-3} $\approx$ 1.01325 bar $\approx$ 1 atm in chloroform/water. In GROMOS, Berendsen thermostat and barostat[45] were employed for the simulations in water and chloroform. In vacuum,

the leap-frog stochastic dynamics integrator was used such that no temperature scaling was necessary. In OpenMM, a Monte Carlo barostat[321] was employed for the simulations in chloroform/water. For the OpenMM simulations in all environments, the *LangevinMiddleIntegrator* was used, thus no additional temperature scaling was necessary. All bonds were constrained with the SHAKE algorithm[84] in GROMOS (relative tolerance 10^{-4}), or with a mix of SETTLE,[85] SHAKE,[84] and CCMA[87] in OpenMM (*i.e.*, default). The force constant for the atomic distance restraints was set to[254] 5000 kJ mol^{-1} nm^{-2}. All simulations were executed on the Euler cluster of ETH Zürich. The input files for the RE-EDS simulations can be found at `https://github.com/rinikerlab/reeds/tree/main/examples/systems/shifted_reaction_field`.

The calculated hydration free energies reported in the FreeSolv[155,173] database were obtained from alchemical MBAR[157,158,322] simulations performed with the GROMACS MD engine.[24,25] 20 λ-values were used and the simulations at each λ-value were 5 ns long. The electrostatic interactions were modified along the first five intermediate states, and the Lennard-Jones interactions were modified along the last 15 intermediate states.[155]

Set A

For the RE-EDS simulations in GROMOS/OpenMM, six independent EDS simulations of 1 ns length with $s = 1$ were conducted to generate optimized coordinates for each end-state in vacuum/water. For each EDS simulation there was one "favored" end-state. The energy offset of the favored end-state was set to 500 kJ mol^{-1} whereas the energy offset of the other end-states were set to -500 kJ mol^{-1}. 21 independent EDS simulations of 0.2 ns length with logarithmically distributed s-values between 1 and 10^{-5} were used to determine a lower bound for the s-values. The determined lower bounds were identical for the GROMOS and OpenMM simulations (0.0178 in vacuum and 0.01 in water). Next, the energy offsets were estimated from a RE-EDS simulation of 0.8 ns

length with 11 replicas (vacuum) and 12 replicas (water), respectively, where the initial energy offsets were all set to zero. In vacuum, one s-optimization run of 0.5 ns length with 12 replicas was required to obtain frequent round-trips in vacuum, adding four s-values. In water, two s-optimization iterations, one of 0.5 ns length with 13 replicas and one of 1.0 ns length with 17 replicas were required, adding eight s-values in total. In vacuum, no energy offset rebalancing was needed. In water, two energy offset rebalancing runs of 0.5 ns length each with 21 replicas were required to sample all end-states approximately equally. Finally, the production runs were 0.5 ns in vacuum (11 replicas) and in water (16 replicas). The free-energy differences in vacuum/water were calculated from five independent production runs.

Set C

For set C, nine independent RE-EDS pipelines were executed in GROMOS for each environment (vacuum/water/chloroform). Three different schemes to calculate the electrostatic potential energy were used: $\mathrm{CG}^{\mathrm{orig}}$, $\mathrm{AT}^{\mathrm{orig}}$, and $\mathrm{AT}^{\mathrm{shift}}$. Further, three different cutoffs were used for each scheme: 0.8 nm, 1.0 nm, and 1.2 nm. For each RE-EDS pipeline, 14 independent EDS simulations were conducted analogously to set A to generate optimized coordinates for the RE-EDS simulations. The lower bound search was analogous to set A for all systems. The determined lower bounds were between 0.0178 and 0.01. For each RE-EDS pipeline, the energy offsets were estimated from RE-EDS simulations of 1.2 ns. The number of replicas was either 19 (lower bound 0.0178) or 20 (lower bound 0.01). In vacuum/chloroform, one s-optimization run of 1 ns length was conducted, adding four s-values. In water, two s-optimization runs of 1 ns and 1.5 ns, respectively, were performed, adding eight s-values in total. Four energy offset rebalancing runs of 0.5 ns length each were conducted in vacuum, two in chloroform, and three in water. The production runs were 1 ns in vacuum and 2 ns in chloroform/water. The free-energy differences in vacuum/chloroform/water were calculated from five independent

production runs each.

For the RE-EDS calculations in OpenMM, only the system with the scheme AT^{shift} and a 1.2 nm cutoff was investigated. The coordinate optimization was analogous to the RE-EDS simulations in GROMOS, as was the lower bound search. The determined lower bounds were identical to the ones obtained from the simulations in GROMOS. For the energy offset estimation, simulations of 0.8 ns length were conducted in each environment with 19 replicas (vacuum) and 20 replicas (chloroform/water), respectively. In vacuum/chloroform, one s-optimization run of 0.5 ns, adding four s-values, was sufficient, whereas in water, two s-optimization runs of 0.5 ns were used, adding eight s-values in total. In vacuum/water, three energy offset rebalancing runs of 0.5 ns length each were conducted, and in chloroform, two energy offset rebalancing runs of 0.5 ns length each were performed. The production runs were 1 ns in vacuum and 2 ns in chloroform/water. Again, the free-energy differences in vacuum/chloroform/water were calculated from five independent production runs.

Set D

For set D, nine independent RE-EDS pipelines were executed in GROMOS in each of the three environments, analogously to set C. The coordinate optimization (13 independent EDS simulations) and lower bound search were analogous to set C. The determined lower bounds were between 0.0178 and 0.00316. The number of replicas for the energy offset estimation was between 18 (lower bound 0.0178) and 22 (lower bound 0.00316). The s-optimization runs were analogous to set C. For the simulations in vacuum, three energy offset rebalancing runs of 0.5 ns length each were required, while the energy offset rebalancing runs were identical to set C in chloroform/water. The production runs were analogous to set C in all three environments and the free-energy differences were again calculated from five independent production runs.

As for set C, only the system with the scheme AT^{shift} and a 1.2 ns

cutoff was investigated with OpenMM. Also here, the coordinate optimization and lower bound search was analogous to the RE-EDS simulations in GROMOS. The lower bounds were 0.0178 (vacuum) and 0.0056 (chloroform/water), identical to the ones obtained in GROMOS. 18 replicas were used for the energy offset estimation in vacuum and 20 replicas for the energy offset estimation in water/chloroform. In vacuum/chloroform one s-optimization run of 0.5 ns length was sufficient, adding four s-values, whereas in water two s-optimization runs of 0.5 ns length each were used, adding eight s-values in total. In vacuum/chloroform, two energy offset rebalancing runs (0.5 ns each) were required, and in water, three rebalancing runs (0.5 ns each) were required. The production runs were analogous to the simulations in GROMOS and the free-energy differences were calculated from five independent production runs.

7.3.6 Analysis

The analysis of the simulations was conducted with GROMOS++[282] and PyGromosTools.[279] Further, the following Python packages were used for visualization and analysis: Matplotlib,[284] mpmath,[287] NumPy,[285] Pandas,[283] SciPy,[286] and Seaborn.[311] For all systems, the root-mean-square error (RMSE), the mean absolute error (MAE), and the Spearman[289] correlation coefficient between the different simulation methods and the experimental values are reported. The free-energy differences obtained from the RE-EDS simulations in GROMOS were calculated with the GROMOS++ program *dfmult*, whereas the free-energy differences obtained from the RE-EDS simulations in OpenMM were calculated directly with Python using the NumPy functions *log*, *exp*, and *mean*.

7.4 RESULTS

7.4.1 Comparison of Different RF Schemes and Cutoff Values for RE-EDS

The relative hydration free energies, $\Delta\Delta G_{\text{hyd}}$, and the relative solvation free energies in chloroform, $\Delta\Delta G_{\text{CHCl}_3}$, were calculated from the free-energy differences obtained with RE-EDS in GROMOS for sets C and D. The resulting $\Delta\Delta G_{\text{hyd}}$ values were compared to the values obtained from the calculated and experimental ΔG_{hyd} values reported in the FreeSolv database.[155,173] The $\Delta\Delta G_{\text{CHCl}_3}$ values were compared to the ones obtained from the experimental ΔG_{CHCl_3} values reported in the Minnesota solvation database.[174]

RF Schemes and Cutoff Values

First, the $\Delta\Delta G_{\text{hyd}}$ and $\Delta\Delta G_{\text{CHCl}_3}$ values obtained from the RE-EDS calculations in GROMOS for the different RF schemes (*i.e.*, CG^{orig}, AT^{orig}, AT^{shift}) with different cutoff values (*i.e.*, 0.8 nm, 1.0 nm, 1.2 nm) were evaluated.

Set C For set C, the agreement between the $\Delta\Delta G_{\text{hyd}}$ values obtained from RE-EDS in GROMOS and the calculated/experimental results reported by FreeSolv was high for all schemes. Relative to the experimental values, the RMSE values were between 2.1 and 2.8 kJ mol^{-1}, and the Spearman correlation coefficients between 0.86 and 0.92 (bottom left panel in Figure 7.4, Table 7.2). While the reported metrics were slightly better for CG^{orig} than for the two schemes with AT cutoff, the differences were essentially negligible at $R_{\text{RF}} = 1.2$ nm. The RMSE values compared to MBAR were between 0.4 and 1.1 kJ mol^{-1}, and the Spearman correlation coefficients between 0.97 and 0.99 (bottom middle panel in Figure 7.4, Table 7.5 in Appendix Sec. 7.A.1).

The agreement between the $\Delta\Delta G_{\mathrm{CHCl_3}}$ values obtained from RE-EDS in GROMOS and the experimental results reported in the Minnesota solvation database[174] was lower than for the hydration free energies. Apart from one outlier ($\mathrm{CG^{orig}}$ with $R_{\mathrm{RF}} = 0.8$), the RMSE values were between 3.3 and 3.7 kJ mol^{-1}, and the Spearman correlation coefficients between 0.78 and 0.82 (bottom right panel in Figure 7.4, Table 7.2). The reported metrics improved with an increase in R_{RF} for all three RF schemes.

Table 7.2: Overview of statistical metrics (RMSE, MAE, Spearman correlation coefficient), as well as simulation time (t_{prep} and t_{prod}), for different RF schemes and cutoff distances for sets C and D. The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve (set C) or eleven (set D) molecules was removed from the calculations (5000 repetitions). (Left): $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$ versus $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$ [155,173] (top) and $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ [155,173] versus $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$ [155,173] (bottom). For the MBAR calculations, the cutoff was 1.2 nm for the electrostatic interactions and 1.0 nm for the vdW interactions (with a switch at 0.9 nm and a long-range dispersion correction) [155] (Right): $\Delta\Delta G_{\mathrm{CHCl_3}}^{\mathrm{RE\text{-}EDS}}$ versus $\Delta\Delta G_{\mathrm{CHCl_3}}^{\mathrm{exp}}$. [174]

$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$ vs $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$

Set	Scheme	Cutoff [nm]	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}	t_{prep} [ns]	t_{prod} [ns]
C	$\mathrm{CG^{orig}}$	0.8	2.2 ± 0.2	1.8 ± 0.2	0.92	309.2	67
		1.0	2.1 ± 0.2	1.7 ± 0.2	0.92	309.2	67
		1.2	2.2 ± 0.2	1.8 ± 0.2	0.92	309.2	67
	$\mathrm{AT^{orig}}$	0.8	2.3 ± 0.2	1.9 ± 0.2	0.91	309.2	67
		1.0	2.3 ± 0.2	1.9 ± 0.2	0.91	309.2	67
		1.2	2.2 ± 0.2	1.8 ± 0.2	0.92	309.2	67
	$\mathrm{AT^{shift}}$	0.8	2.8 ± 0.2	2.3 ± 0.2	0.86	309.2	67
		1.0	2.4 ± 0.2	2.0 ± 0.2	0.90	309.2	67
		1.2	2.3 ± 0.2	1.9 ± 0.2	0.91	309.2	67
D	$\mathrm{CG^{orig}}$	0.8	6.4 ± 0.6	5.4 ± 0.5	0.78	281	66
		1.0	6.3 ± 0.6	5.3 ± 0.5	0.79	286.2	68
		1.2	6.2 ± 0.7	5.2 ± 0.6	0.80	281	66
	$\mathrm{AT^{orig}}$	0.8	5.8 ± 0.6	4.9 ± 0.5	0.80	281	66
		1.0	6.0 ± 0.6	5.0 ± 0.6	0.80	281	66
		1.2	6.2 ± 0.6	5.2 ± 0.6	0.80	281	66
	$\mathrm{AT^{shift}}$	0.8	5.6 ± 0.6	4.7 ± 0.6	0.81	281	66
		1.0	6.0 ± 0.6	5.0 ± 0.6	0.80	281	66
		1.2	6.1 ± 0.6	5.1 ± 0.6	0.81	281	66

$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ vs $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$

Set	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}	t_{prep} [ns]	t_{prod} [ns]
C	1.9 ± 0.1	1.6 ± 0.1	0.94	42	1400
D	5.9 ± 0.6	4.9 ± 0.6	0.80	39	1300

$\Delta\Delta G_{\mathrm{CHCl_3}}^{\mathrm{RE\text{-}EDS}}$ vs $\Delta\Delta G_{\mathrm{CHCl_3}}^{\mathrm{exp}}$

Set	Scheme	Cutoff [nm]	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}	t_{prep} [ns]	t_{prod} [ns]
C	$\mathrm{CG^{orig}}$	0.8	5.6 ± 0.5	4.6 ± 0.5	0.66	234	57
		1.0	3.5 ± 0.3	2.9 ± 0.2	0.80	277.2	59
		1.2	3.4 ± 0.3	2.7 ± 0.3	0.80	277.2	59
	$\mathrm{AT^{orig}}$	0.8	3.7 ± 0.3	3.1 ± 0.3	0.78	277.2	59
		1.0	3.4 ± 0.3	2.7 ± 0.2	0.80	277.2	59
		1.2	3.3 ± 0.3	2.7 ± 0.3	0.82	277.2	59
	$\mathrm{AT^{shift}}$	0.8	3.6 ± 0.3	3.0 ± 0.3	0.80	277.2	59
		1.0	3.4 ± 0.3	2.8 ± 0.3	0.80	277.2	59
		1.2	3.4 ± 0.3	2.7 ± 0.3	0.80	277.2	59
D	$\mathrm{CG^{orig}}$	0.8	5.6 ± 0.7	4.4 ± 0.7	0.31	207.8	56
		1.0	4.8 ± 0.5	4.0 ± 0.5	0.58	211	58
		1.2	4.9 ± 0.5	4.1 ± 0.4	0.58	211	58
	$\mathrm{AT^{orig}}$	0.8	5.1 ± 0.5	4.2 ± 0.5	0.52	211	58
		1.0	5.0 ± 0.5	4.1 ± 0.4	0.57	211	58
		1.2	4.9 ± 0.4	4.1 ± 0.4	0.58	211	58
	$\mathrm{AT^{shift}}$	0.8	5.0 ± 0.5	4.2 ± 0.5	0.52	211	58
		1.0	5.1 ± 0.5	4.2 ± 0.5	0.55	211	58
		1.2	5.0 ± 0.5	4.1 ± 0.4	0.59	211	58

Set D While the agreement between RE-EDS and MBAR was again excellent, the deviations from the experimental reference data were considerably higher for set D than for set C. This indicates that the deviations

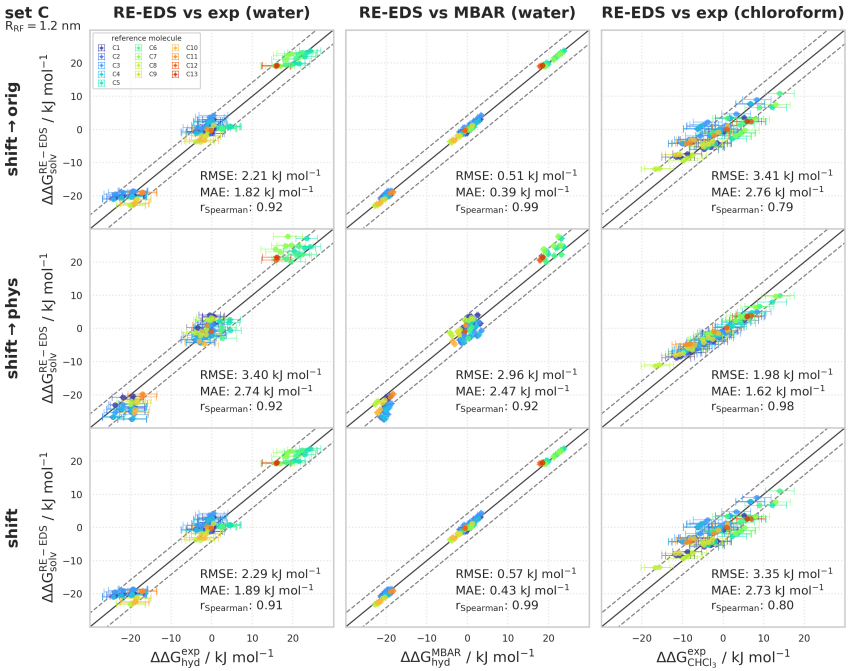


Figure 7.4: Comparison of the relative solvation free energies of set C: $\Delta\Delta G_{\text{solv}}^{\text{RE-EDS}}$ obtained from RE-EDS calculations in GROMOS with $R_{\text{RF}} = 1.2$ nm, propagated with the AT^{shift} scheme versus the experimental and calculated reference values. The three columns correspond to the comparison against $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ [155,173] (left), $\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$ [155,173] (middle), and $\Delta\Delta G_{\text{CHCl}_3}^{\text{exp}}$ [174] (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\text{shift}\rightarrow\text{orig}}$ (top), $V^{\text{shift}\rightarrow\text{pyhs}}$ (middle), and the shifted electrostatic potential energy $V^{\text{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in each environment and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The complete numerical values for all combinations of the three RF schemes and of the three cutoff distances are provided in Tables 7.6 - 7.9 in Appendix Sec. 7.A.1, and Tables 7.15 and 7.16 in Appendix Sec. 7.A.2. The corresponding plots are provided in Figures 7.7 - 7.9 in Appendix Sec. 7.A.1 and Figures 7.13 - 7.15 in Appendix Sec. 7.A.2.

may be related to shortcomings in the force field or experimental determination. The RMSE values compared to experiment were between 5.6 and 6.4 kJ mol^{-1}, and the Spearman correlation coefficients between 0.78 and 0.81 (bottom left panel in Figure 7.5, Table 7.2). Overall, the

obtained $\Delta\Delta G_{\text{hyd}}$ values were again very similar for all schemes, and small differences should not be over interpreted. Relative to MBAR, the RMSE values are between 0.4 and 1.1 kJ mol^{-1}, and the Spearman correlation coefficients between 0.99 and 1.00 (bottom middle panel in Figure 7.5, Table 7.5 in Appendix Sec. 7.A.1).

The agreement between the $\Delta\Delta G_{\text{CHCl}_3}$ values obtained from RE-EDS in GROMOS and the experimental values was similar than for the hydration free energies. The RMSE values were between 4.8 and 5.6 kJ mol^{-1}, and the Spearman correlation coefficients between 0.31 and 0.59 (bottom right panel in Figure 7.5, Table 7.2).

Timings The comparison of the total simulation time to obtain converged relative hydration free energies with RE-EDS (*i.e.*, 376 ns for set C and 347 ns for set D with AT$^{\text{shift}}$) compared to the total simulation time for the MBAR results (*i.e.*, 1442 ns for set C and 1339 ns for set D) highlights the high intrinsic sampling efficiency of RE-EDS (Table 7.2). Note that the small variations in the required simulation time for the RE-EDS simulations in Table 7.2 stem from additional/fewer replicas due to a lower/higher lower bound for the s-values. A more detailed discussion on timings of RE-EDS calculations compared to other simulation methods (TI and MBAR) is provided in Chapter 6.

Based on these results, using a cutoff of 1.0 or 1.2 nm is appropriate for both considered sets of molecules. With the exception of the $\Delta\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ values using CG$^{\text{orig}}$, even a low cutoff of 0.8 nm leads to comparable accuracy. As the ΔG_{hyd} values calculated with MBAR and reported in the FreeSolv database were obtained with a cutoff of 1.2 nm for the electrostatic interactions,[155] $R_{\text{RF}} = 1.2$ nm is the most appropriate choice to compare to $\Delta\Delta G^{\text{MBAR}}_{\text{hyd}}$. The results obtained with the different RF schemes achieve comparable accuracy within the uncertainty of the calculations. However, as the AT$^{\text{shift}}$ scheme was shown to decrease cutoff artifacts for simulations with an atom-based cutoff significantly,[304] this is the most appropriate scheme to employ when using force fields that do

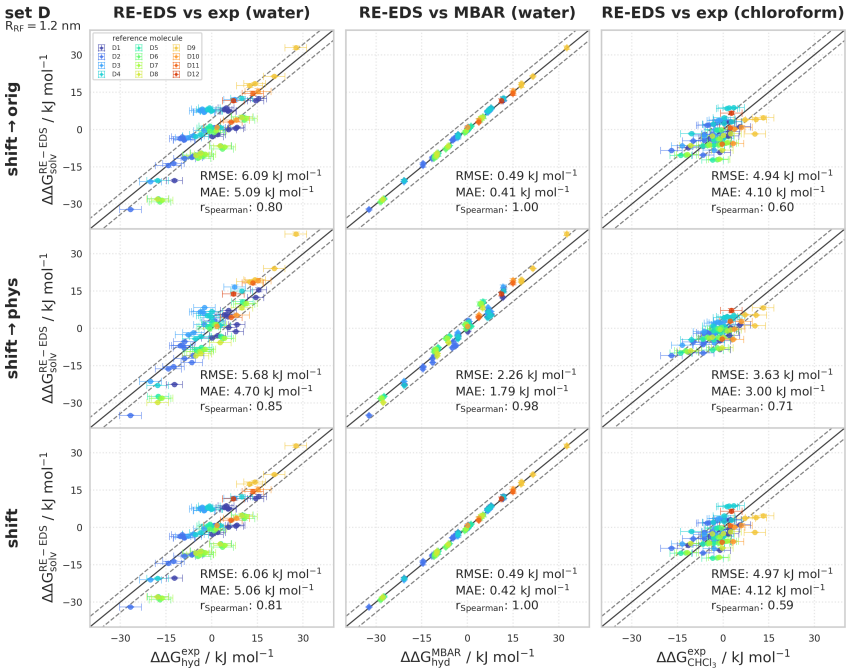


Figure 7.5: Comparison of the relative solvation free energies of set D: $\Delta\Delta G_{\text{solv}}^{\text{RE-EDS}}$ obtained from RE-EDS calculations in GROMOS with $R_{\text{RF}} = 1.2$ nm, propagated with the AT$^{\text{shift}}$ scheme versus the experimental and calculated reference values. The three columns correspond to the comparison against $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ [155,173] (left), $\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$ [155,173] (middle), and $\Delta\Delta G_{\text{CHCl}_3}^{\text{exp}}$ [174] (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\text{shift}\rightarrow\text{orig}}$ (top), $V^{\text{shift}\rightarrow\text{pyhs}}$ (middle), and the shifted electrostatic potential energy $V^{\text{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in each environment and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The complete numerical values for all combinations of the three RF schemes and of the three cutoff distances are provided in Tables 7.10 - 7.13 in Appendix Sec. 7.A.1, and Tables 7.17 and 7.18 in Appendix Sec. 7.A.2. The corresponding plots are provided in Figures 7.10 - 7.12 in Appendix Sec. 7.A.1 and Figures 7.16 - 7.18 in Appendix Sec. 7.A.2.

not rely on charge groups.

Free-Energy Differences from Corrected Energies

The $\Delta\Delta G$ values obtained from the RE-EDS calculations in GROMOS with AT$^{\text{shift}}$ were compared to the values obtained with the "corrected"

energies $V^{\text{shift}\to\text{orig}}$ and $V^{\text{shift}\to\text{phys}}$ (Eqs. (7.19) and (7.20)). The systems were propagated with the AT^{shift} scheme, and the extra energy terms $V_i^{\text{extra,orig}}$ and $V_i^{\text{extra,phys}}$ were stored in the energy trajectory for all the end-states i.

Set C For set C, the use of corrected end-state and reference-state electrostatic potential energies $V^{\text{shift}\to\text{orig}}$ to calculate the relative hydration free energies resulted in negligible changes (left panel in Figure 7.4, Table 7.3). Using $V^{\text{shift}\to\text{phys}}$, on the other hand, significantly diminished the agreement with the experimental values. The same trends were observed for the agreement with the values obtained from MBAR (middle panel in Figure 7.4, Table 7.14 in Appendix Sec. 7.A.2). In the low permittivity environment of chloroform, on the other hand, using $V^{\text{shift}\to\text{phys}}$ to calculate the relative solvation free energies significantly increased the agreement with the experimental reference values. At $R_{\text{RF}} = 1.2$ nm, the RMSE was 2.0 kJ mol^{-1} compared to a RMSE of 3.4 kJ mol^{-1} when $V^{\text{ele,shift}}$ was used directly. Using $V^{\text{shift}\to\text{orig}}$ had a negligible impact on the agreement with experiment (right panel in Figure 7.4, Table 7.3).

Set D For set D, using $V^{\text{shift}\to\text{orig}}$ to calculate the $\Delta\Delta G_{\text{hyd}}$ values had again a negligible impact on the agreement with experiment, while using $V^{\text{shift}\to\text{phys}}$ slightly improved the reported metrics (*e.g.*, at $R_{\text{RF}} = 1.2$ nm, the RMSE was 5.7 kJ mol^{-1} compared to 6.1 kJ mol^{-1}, see the left panel in Figure 7.5, Table 7.3). The agreement with the MBAR results, on the other hand, significantly decreased when using $V^{\text{shift}\to\text{phys}}$, analogous to set C. Here, the RMSE was 2.3 kJ mol^{-1} at $R_{\text{RF}} = 1.2$ nm compared to 0.5 kJ mol^{-1} when using $V^{\text{ele,shift}}$ directly (middle panel in Figure 7.5, Table 7.14 in Appendix Sec. 7.A.2). As for set C, the agreement with experiment for the relative solvation free energies in chloroform significantly improved upon using the corrected energies $V^{\text{shift}\to\text{phys}}$. For example, at $R_{\text{RF}} = 1.2$ nm, the RMSE was 3.6 kJ mol^{-1} compared to 5.0 kJ mol^{-1} (right panel in Figure 7.5, Table 7.3).

Table 7.3: Overview of statistical metrics (RMSE, MAE, Spearman correlation coefficient) for RE-EDS simulations using the AT$^{\text{shift}}$ scheme with and without corrected energy terms for sets C and D. The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve (set C) or eleven (set D) molecules was removed from the calculations (5000 repetitions). (Left): $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ versus $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$.[155,173] (Right): $\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS}}$ versus $\Delta\Delta G_{\text{CHCl}_3}^{\text{exp}}$.[174]

$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ vs $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$

Set	V^{ele}	Cutoff [nm]	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}
C	shift→orig	0.8	2.4 ± 0.2	1.9 ± 0.2	0.90
		1.0	2.2 ± 0.2	1.8 ± 0.2	0.92
		1.2	2.2 ± 0.2	1.8 ± 0.2	0.92
	shift→phys	0.8	3.5 ± 0.5	2.8 ± 0.4	0.91
		1.0	3.3 ± 0.4	2.7 ± 0.4	0.93
		1.2	3.4 ± 0.5	2.7 ± 0.4	0.92
	shift	0.8	2.8 ± 0.2	2.3 ± 0.2	0.86
		1.0	2.4 ± 0.2	2.0 ± 0.2	0.90
		1.2	2.3 ± 0.2	1.9 ± 0.2	0.91
D	shift→orig	0.8	5.8 ± 0.7	4.8 ± 0.5	0.80
		1.0	6.1 ± 0.6	5.1 ± 0.6	0.80
		1.2	6.1 ± 0.6	5.1 ± 0.6	0.80
	shift→phys	0.8	5.4 ± 0.6	4.4 ± 0.6	0.86
		1.0	5.5 ± 0.6	4.6 ± 0.6	0.85
		1.2	5.7 ± 0.5	4.7 ± 0.5	0.85
	shift	0.8	5.6 ± 0.6	4.7 ± 0.6	0.81
		1.0	6.0 ± 0.6	5.0 ± 0.6	0.80
		1.2	6.1 ± 0.6	5.1 ± 0.6	0.81

$\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS}}$ vs $\Delta\Delta G_{\text{CHCl}_3}^{\text{exp}}$

Set	V^{ele}	Cutoff [nm]	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}
C	shift→orig	0.8	3.8 ± 0.3	3.1 ± 0.2	0.77
		1.0	3.5 ± 0.3	2.9 ± 0.3	0.78
		1.2	3.4 ± 0.3	2.8 ± 0.2	0.79
	shift→phys	0.8	2.4 ± 0.3	1.9 ± 0.2	0.99
		1.0	2.0 ± 0.2	1.6 ± 0.2	0.98
		1.2	2.0 ± 0.2	1.6 ± 0.2	0.98
	shift	0.8	3.6 ± 0.3	3.0 ± 0.3	0.80
		1.0	3.4 ± 0.3	2.8 ± 0.3	0.80
		1.2	3.4 ± 0.3	2.7 ± 0.3	0.80
D	shift→orig	0.8	5.0 ± 0.5	4.1 ± 0.5	0.53
		1.0	5.0 ± 0.5	4.2 ± 0.4	0.56
		1.2	4.9 ± 0.5	4.1 ± 0.4	0.60
	shift→phys	0.8	3.6 ± 0.4	3.0 ± 0.3	0.53
		1.0	3.7 ± 0.3	3.1 ± 0.3	0.70
		1.2	3.6 ± 0.3	3.0 ± 0.3	0.71
	shift	0.8	5.0 ± 0.5	4.2 ± 0.5	0.52
		1.0	5.1 ± 0.5	4.2 ± 0.5	0.55
		1.2	5.0 ± 0.5	4.1 ± 0.4	0.59

Based on these results, propagating a system with the electrostatic potential-energy term $V^{\text{ele,shift}}$ and using the corrected energy $V^{\text{shift}\to\text{phys}} = V^{\text{shift}} + V^{\text{extra,phys}}$ for analysis could potentially have a significantly positive effect on the accuracy of MD (free-energy) calculations. However, as the trend was not consistent across all systems, more investigations are needed before drawing a firm conclusion. In contrast, it is evident that using $V^{\text{shift}\to\text{orig}}$ does not improve the free-energy differences obtained based on GAFF topologies.

7.4.2 RE-EDS Simulations in OpenMM

To validate the implementation of RE-EDS in OpenMM, the relative hydration free energies of sets A, C, and D were compared to the experimental reference values,[155,173] the calculated values obtained with MBAR,[155,173] as well as the values obtained with RE-EDS in GROMOS. Further, the relative free energies in chloroform for sets C and D were compared to the experimental reference values,[174] as well as the results

obtained with RE-EDS in GROMOS. For the RE-EDS simulations in OpenMM, the AT$^{\text{shift}}$ scheme with a cutoff of 1.2 nm was chosen for the electrostatic interactions.

For set A, there was excellent agreement between the $\Delta\Delta G_{\text{hyd}}$ values calculated with RE-EDS in OpenMM and in GROMOS. The RMSE and MAE were 0.2 kJ mol^{-1} and the Spearman correlation coefficient was 1.00. According to the reported metric, the agreement with the experimental reference values was slightly better for the values obtained with RE-EDS in OpenMM than with RE-EDS in GROMOS and with MBAR[155,173] in GROMACS (left panel in Figure 7.6, Table 7.4).

Table 7.4: Overview of statistical metrics (RMSE, MAE, Spearman correlation coefficient) against experiment, as well as simulation time (t_{prep} and t_{prod}) for RE-EDS simulations in GROMOS and OpenMM (using the AT$^{\text{shift}}$ scheme and $R_{\text{RF}} = 1.2$ nm), as well as MBAR simulations in GROMACS as reported by FreeSolv.[155,173] The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSEs and MAEs when a random selection of up to four (set A), twelve (set C), or eleven (set D) molecules was removed from the calculations (5000 repetitions each).

Set	$\Delta\Delta G$	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}	t_{prep} [ns]	t_{prod} [ns]
A	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$	2.6 ± 0.3	2.2 ± 0.3	0.93	89.3	13.5
	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$	2.5 ± 0.3	2.2 ± 0.3	0.93	89.3	13.5
	$\Delta\Delta G_{\text{hyd}}^{\text{MBAR 155,173}}$	3.1 ± 0.4	2.7 ± 0.3	0.94	18	600
C	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$	2.3 ± 0.2	1.9 ± 0.2	0.91	309.2	67
	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$	2.0 ± 0.1	1.6 ± 0.1	0.93	216.1	67
	$\Delta\Delta G_{\text{hyd}}^{\text{MBAR 155,173}}$	1.9 ± 0.1	1.6 ± 0.1	0.94	42	1400
	$\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS,GROMOS}}$	3.4 ± 0.3	2.7 ± 0.3	0.80	277.2	59
	$\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS,OpenMM}}$	2.9 ± 0.2	2.3 ± 0.2	0.87	177.1	59
D	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$	6.1 ± 0.6	5.1 ± 0.5	0.81	282.6	66
	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$	6.0 ± 0.6	5.0 ± 0.5	0.80	191.8	66
	$\Delta\Delta G_{\text{hyd}}^{\text{MBAR 155,173}}$	5.9 ± 0.6	4.9 ± 0.5	0.80	39	1300
	$\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS,GROMOS}}$	5.0 ± 0.5	4.1 ± 0.5	0.59	211	58
	$\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS,OpenMM}}$	5.0 ± 0.4	4.0 ± 0.4	0.59	153.8	58

Analogous to set A, the agreement between the $\Delta\Delta G_{\text{hyd}}$ values obtained from RE-EDS in GROMOS and in OpenMM was excellent for set C (top middle panel in Figure 7.6). The RMSE and MAE were 0.4 kJ mol^{-1}

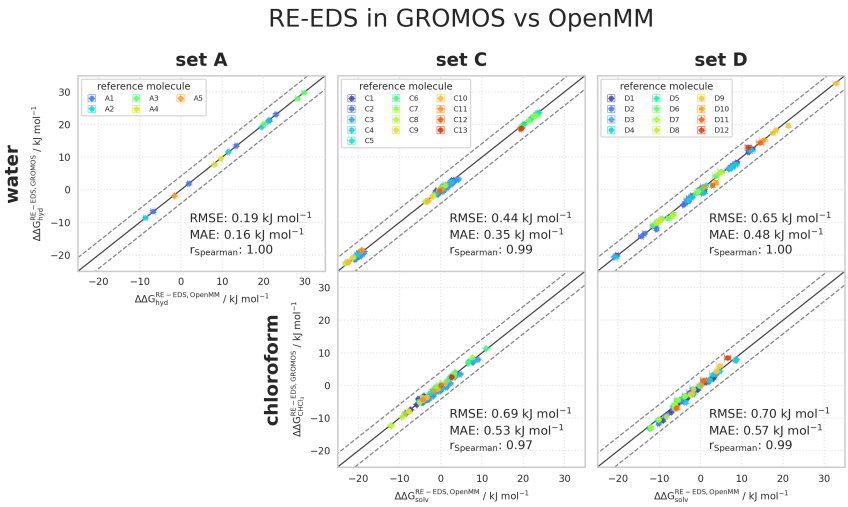


Figure 7.6: Comparison of the relative solvation free energies in water of set A (left), and in water and chloroform of set C (middle) and D (right). (Top): $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$ obtained from RE-EDS calculations in GROMOS versus $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$ obtained from RE-EDS calculations in OpenMM (both with the AT$^{\text{shift}}$ scheme and $R_{\text{RF}} = 1.2$ nm). (Bottom): $\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS,GROMOS}}$ obtained from RE-EDS calculations in GROMOS versus $\Delta\Delta G_{\text{CHCl}_3}^{\text{RE-EDS,OpenMM}}$ obtained from RE-EDS calculations in OpenMM (both with the AT$^{\text{shift}}$ scheme and $R_{\text{RF}} = 1.2$ nm). The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results were averaged over five repeats in each environment and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.19 - 7.23 in Appendix Sec. 7.A.3. All pairwise comparisons between the different simulation methods and the experimental results are shown in Figures 7.19, 7.22, 7.25, 7.27, and 7.30 in Appendix Sec. 7.A.3. Plots of the convergence of the free-energy calculations in each environment with RE-EDS in OpenMM and GROMOS are provided in Figures 7.20, 7.21, 7.23, 7.24, 7.26, 7.28, 7.29, and 7.31 in Appendix Sec. 7.A.3.

and the Spearman correlation coefficient was 0.99. Also here, the agreement with the experimental reference values was slightly better for the OpenMM results than the GROMOS ones (Table 7.4). In chloroform, the agreement between the relative solvation free energies obtained with RE-EDS in GROMOS and OpenMM was also excellent with an RMSE of 0.7 kJ mol^{-1}, and a Spearman correlation coefficient of 0.97 (bottom middle panel in Figure 7.6). The agreement with the experimental values

was slightly higher for the results obtained with OpenMM than with GROMOS (Table 7.4).

Finally, similar observations were made for set D in water and chloroform, showing excellent agreement between the RE-EDS implementations in GROMOS and OpenMM (right panel in Figure 7.6) and comparable results against experiment (Table 7.4).

Taken together, the results for the three sets A, C, and D demonstrate that the RE-EDS implementation in OpenMM can be used to calculate free-energy differences with a rigorously conservative scheme and a high sampling efficiency.

7.5 CONCLUSION

To exploit the inherent sampling efficiency of RE-EDS with force fields using QM-derived charges and in MD engines without support for CG based cutoff, we wanted to combine RE-EDS with the RF scheme including a shifting function that enables the use of AT based cutoff in a rigorously conservative manner. For this, we first compared different RF schemes (CG^{orig}, AT^{orig}, and AT^{shift}) and cutoff values in the GROMOS MD engine using solvation free energies in water and chloroform as test system. The results indicated that the AT^{shift} scheme can be used with RE-EDS by propagating the system with the electrostatic potential-energy term $V^{ele,shift}$, and using either directly $V^{ele,shift}$ or the corrected energy $V^{shift \rightarrow phys} = V^{shift} + V^{extra,phys}$ for the calculation of the free-energy difference.

Next, we implemented RE-EDS with the optimal RF scheme in the OpenMM MD engine. There were four main ingredients required for the implementation: (*i*) the definition of the nonbonded interactions based on V^{LJ} and $V^{ele,shift}$ *via* OpenMM's *CustomNonbondedForce* and *CustomBondForce*; (*ii*) the definition of atomic distance restraints *via* a *CustomBondForce* to prevent the molecules from drifting apart during the

simulation; (*iii*) the EDS integration, *i.e.*, the scaling of the nonbonded forces of the end-states based on the reference potential; and (*iv*) the simulation of independent replicas together with a replica-exchange criterion. The current proof-of-concept implementation is a simple Python3 module where the replicas are simulated serially. An implementation with improved parallelization is part of future work. The implementation in OpenMM was validated using solvation free energies in water and chloroform for three sets of molecules at $R_{\mathrm{RF}} = 1.2$ nm. The obtained $\Delta\Delta G_{\mathrm{hyd}}$ and $\Delta\Delta G_{\mathrm{CHCl_3}}$ values were compared to the analogous results obtained from RE-EDS in GROMOS, from MBAR in GROMACS (as reported by FreeSolv), and the experimental reference values. The agreement of the RE-EDS results obtained with the two MD engines was excellent, both for $\Delta\Delta G_{\mathrm{hyd}}$ as well as $\Delta\Delta G_{\mathrm{CHCl_3}}$, with RMSEs ≤ 0.7 kJ mol^{-1}. Similar high agreement was found for the results obtained with the state-of-the art method MBAR. Compared to experiment, the RE-EDS calculations in OpenMM provided a small but consistent improvement.

The developments presented in this chapter enable free-energy calculations with RE-EDS in GROMOS and OpenMM using force fields with QM-derived charges.

7.A APPENDIX

7.A.1 Comparison of RF Schemes and Cutoff Values in GROMOS

Overview of Statistical Metrics Against MBAR

Table 7.5: Overview of statistical metrics (RMSE, MAE, Spearman correlation coefficient) for $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ for different RF schemes and cutoff distances for sets C and D versus $\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$.[155,173] The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve (set C) or eleven (set D) molecules was removed from the calculations (5000 repetitions). For the MBAR calculations, the cutoff was 1.2 nm for the electrostatic interactions and 1.0 nm for the vdW interactions (with a switch at 0.9 nm and a long-range dispersion correction).[155]

$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ vs $\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$

Set	Scheme	Cutoff [nm]	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	r_{Spearman}
C	CG^{orig}	0.8	0.90 ± 0.1	0.72 ± 0.1	0.98
		1.0	0.47 ± 0.1	0.37 ± 0.1	0.99
		1.2	0.43 ± 0.1	0.35 ± 0.1	0.99
	AT^{orig}	0.8	0.67 ± 0.1	0.54 ± 0.1	0.99
		1.0	0.58 ± 0.1	0.46 ± 0.1	0.99
		1.2	0.44 ± 0.1	0.35 ± 0.1	0.99
	AT^{shift}	0.8	1.10 ± 0.1	0.88 ± 0.1	0.97
		1.0	0.65 ± 0.1	0.52 ± 0.1	0.99
		1.2	0.57 ± 0.1	0.43 ± 0.1	0.99
D	CG^{orig}	0.8	1.13 ± 0.1	0.94 ± 0.1	0.99
		1.0	0.66 ± 0.1	0.55 ± 0.1	1.00
		1.2	0.54 ± 0.0	0.45 ± 0.0	1.00
	AT^{orig}	0.8	0.41 ± 0.1	0.34 ± 0.0	1.00
		1.0	0.43 ± 0.1	0.36 ± 0.0	1.00
		1.2	0.53 ± 0.1	0.43 ± 0.1	1.00
	AT^{shift}	0.8	0.85 ± 0.2	0.64 ± 0.2	1.00
		1.0	0.48 ± 0.1	0.41 ± 0.1	1.00
		1.2	0.49 ± 0.1	0.42 ± 0.1	1.00

Relative Hydration Free Energies of Set C

Table 7.6: $\Delta\Delta G_{\mathrm{hyd}}$ for the 14 molecules of set C calculated from RE-EDS calculations using the RF schemes $\mathrm{CG^{orig}}$, $\mathrm{AT^{orig}}$, and $\mathrm{AT^{shift}}$ with $R_{\mathrm{RF}} = 0.8$ nm and $R_{\mathrm{RF}} = 1.0$ nm. The results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

		$R_{\mathrm{RF}} = 0.8$ nm			$R_{\mathrm{RF}} = 1.0$ nm		
Molecules		$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,CG^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{shift}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,CG^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{shift}}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	1.0 ± 0.3	0.5 ± 0.2	0.6 ± 0.2	0.5 ± 0.3	0.3 ± 0.2	0.3 ± 0.2
C1	C3	−0.3 ± 0.2	−0.9 ± 0.1	−1.0 ± 0.1	−1.0 ± 0.1	−1.1 ± 0.1	−1.0 ± 0.1
C1	C4	1.2 ± 0.2	0.6 ± 0.2	0.4 ± 0.1	0.8 ± 0.5	0.4 ± 0.2	0.5 ± 0.1
C1	C5	−21.0 ± 0.1	−20.9 ± 0.3	−20.3 ± 0.3	−20.8 ± 0.2	−20.9 ± 0.4	−20.5 ± 0.2
C1	C6	−19.5 ± 0.3	−20.0 ± 0.4	−19.6 ± 0.3	−20.3 ± 0.1	−19.9 ± 0.2	−19.5 ± 0.5
C1	C7	−20.2 ± 0.2	−19.9 ± 0.2	−20.1 ± 0.2	−20.1 ± 0.2	−20.1 ± 0.1	−19.7 ± 0.2
C1	C8	2.2 ± 0.2	2.8 ± 0.0	3.6 ± 0.2	2.7 ± 0.2	3.0 ± 0.1	3.2 ± 0.1
C1	C9	1.5 ± 0.1	1.4 ± 0.1	2.2 ± 0.1	1.2 ± 0.3	1.4 ± 0.1	1.7 ± 0.1
C1	C10	3.2 ± 0.1	3.2 ± 0.2	3.6 ± 0.1	2.5 ± 0.2	2.8 ± 0.1	3.0 ± 0.1
C1	C11	0.1 ± 0.1	−0.3 ± 0.1	0.4 ± 0.1	−0.6 ± 0.2	−0.5 ± 0.1	−0.3 ± 0.1
C1	C12	−18.8 ± 0.1	−19.6 ± 0.2	−19.6 ± 0.2	−19.4 ± 0.3	−19.6 ± 0.1	−19.5 ± 0.2
C1	C13	−19.4 ± 0.2	−19.5 ± 0.3	−19.7 ± 0.2	−20.0 ± 0.4	−19.7 ± 0.2	−19.9 ± 0.3
C1	C14	0.5 ± 0.2	−0.0 ± 0.2	1.1 ± 0.2	−0.5 ± 0.2	−0.3 ± 0.1	−0.1 ± 0.0
C2	C3	−1.3 ± 0.2	−1.4 ± 0.2	−1.6 ± 0.3	−1.5 ± 0.2	−1.4 ± 0.2	−1.3 ± 0.3
C2	C4	0.2 ± 0.3	0.1 ± 0.2	−0.2 ± 0.2	0.3 ± 0.3	0.1 ± 0.3	0.2 ± 0.2
C2	C5	−22.0 ± 0.3	−21.4 ± 0.4	−20.9 ± 0.3	−21.3 ± 0.2	−21.2 ± 0.5	−20.8 ± 0.2
C2	C6	−20.5 ± 0.3	−20.5 ± 0.3	−20.2 ± 0.5	−20.8 ± 0.3	−20.2 ± 0.4	−20.2 ± 0.5
C2	C7	−21.1 ± 0.3	−20.4 ± 0.2	−20.6 ± 0.4	−20.6 ± 0.3	−20.4 ± 0.2	−20.0 ± 0.2
C2	C8	1.2 ± 0.3	2.3 ± 0.2	3.0 ± 0.2	2.2 ± 0.2	2.7 ± 0.2	2.9 ± 0.2
C2	C9	0.5 ± 0.3	0.9 ± 0.2	1.6 ± 0.2	0.7 ± 0.2	1.1 ± 0.2	1.4 ± 0.2
C2	C10	2.2 ± 0.3	2.7 ± 0.2	3.0 ± 0.2	2.0 ± 0.1	2.5 ± 0.2	2.8 ± 0.2
C2	C11	−0.9 ± 0.3	−0.8 ± 0.2	−0.2 ± 0.2	−1.2 ± 0.2	−0.9 ± 0.1	−0.6 ± 0.2
C2	C12	−19.8 ± 0.3	−20.1 ± 0.2	−20.2 ± 0.3	−19.9 ± 0.4	−19.9 ± 0.2	−19.8 ± 0.3
C2	C13	−20.4 ± 0.3	−20.0 ± 0.3	−20.3 ± 0.2	−20.5 ± 0.3	−20.0 ± 0.2	−20.2 ± 0.3
C2	C14	−0.4 ± 0.3	−0.5 ± 0.2	0.5 ± 0.2	−1.0 ± 0.1	−0.6 ± 0.2	−0.4 ± 0.1
C3	C4	1.5 ± 0.2	1.5 ± 0.2	1.4 ± 0.1	1.7 ± 0.4	1.5 ± 0.3	1.5 ± 0.1
C3	C5	−20.7 ± 0.2	−20.0 ± 0.4	−19.3 ± 0.3	−19.8 ± 0.2	−19.8 ± 0.5	−19.5 ± 0.3
C3	C6	−19.2 ± 0.3	−19.1 ± 0.4	−18.6 ± 0.3	−19.4 ± 0.2	−18.8 ± 0.2	−18.9 ± 0.5
C3	C7	−19.9 ± 0.1	−19.0 ± 0.2	−19.1 ± 0.2	−19.1 ± 0.2	−19.0 ± 0.2	−18.7 ± 0.3
C3	C8	2.5 ± 0.1	3.7 ± 0.1	4.6 ± 0.2	3.6 ± 0.2	4.1 ± 0.2	4.3 ± 0.2
C3	C9	1.8 ± 0.1	2.3 ± 0.1	3.2 ± 0.1	2.2 ± 0.3	2.5 ± 0.2	2.7 ± 0.2
C3	C10	3.5 ± 0.1	4.1 ± 0.2	4.6 ± 0.1	3.4 ± 0.2	3.9 ± 0.1	4.1 ± 0.2
C3	C11	0.4 ± 0.1	0.6 ± 0.1	1.4 ± 0.2	0.3 ± 0.2	0.5 ± 0.1	0.7 ± 0.2
C3	C12	−18.5 ± 0.2	−18.7 ± 0.1	−18.6 ± 0.2	−18.4 ± 0.3	−18.5 ± 0.2	−18.5 ± 0.2
C3	C13	−19.1 ± 0.1	−18.7 ± 0.3	−18.7 ± 0.3	−19.0 ± 0.4	−18.6 ± 0.2	−18.9 ± 0.3
C3	C14	0.9 ± 0.2	0.8 ± 0.1	2.1 ± 0.2	0.4 ± 0.2	0.8 ± 0.2	0.9 ± 0.1
C4	C5	−22.2 ± 0.2	−21.5 ± 0.4	−20.7 ± 0.3	−21.6 ± 0.3	−21.3 ± 0.5	−21.0 ± 0.2
C4	C6	−20.7 ± 0.2	−20.6 ± 0.4	−20.0 ± 0.4	−21.1 ± 0.4	−20.2 ± 0.2	−20.4 ± 0.4
C4	C7	−21.4 ± 0.2	−20.5 ± 0.3	−20.4 ± 0.3	−20.9 ± 0.4	−20.5 ± 0.2	−20.2 ± 0.2
C4	C8	1.0 ± 0.2	2.3 ± 0.2	3.2 ± 0.2	1.9 ± 0.3	2.6 ± 0.2	2.7 ± 0.1
C4	C9	0.3 ± 0.1	0.9 ± 0.2	1.8 ± 0.1	0.5 ± 0.2	1.1 ± 0.2	1.2 ± 0.1
C4	C10	2.0 ± 0.2	2.7 ± 0.2	3.2 ± 0.1	1.7 ± 0.2	2.4 ± 0.2	2.5 ± 0.1
C4	C11	−1.1 ± 0.1	−0.9 ± 0.2	0.0 ± 0.2	−1.4 ± 0.3	−0.9 ± 0.2	−0.9 ± 0.1
C4	C12	−20.0 ± 0.1	−20.2 ± 0.2	−20.0 ± 0.2	−20.2 ± 0.6	−20.0 ± 0.3	−20.0 ± 0.2
C4	C13	−20.6 ± 0.2	−20.1 ± 0.3	−20.1 ± 0.2	−20.7 ± 0.5	−20.0 ± 0.4	−20.4 ± 0.2
C4	C14	−0.7 ± 0.2	−0.6 ± 0.1	0.7 ± 0.1	−1.3 ± 0.2	−0.7 ± 0.2	−0.6 ± 0.1
C5	C6	1.5 ± 0.3	0.9 ± 0.4	0.7 ± 0.6	0.5 ± 0.2	1.1 ± 0.5	0.6 ± 0.3
C5	C7	0.8 ± 0.2	1.0 ± 0.3	0.2 ± 0.4	0.7 ± 0.2	0.8 ± 0.4	0.8 ± 0.2
C5	C8	23.2 ± 0.2	23.8 ± 0.3	23.9 ± 0.3	23.5 ± 0.1	23.9 ± 0.4	23.7 ± 0.2
C5	C9	22.5 ± 0.1	22.3 ± 0.3	22.5 ± 0.2	22.0 ± 0.2	22.4 ± 0.4	22.2 ± 0.2
C5	C10	24.2 ± 0.1	24.1 ± 0.3	23.9 ± 0.3	23.2 ± 0.1	23.7 ± 0.4	23.5 ± 0.3
C5	C11	21.1 ± 0.2	20.6 ± 0.3	20.7 ± 0.2	20.1 ± 0.1	20.4 ± 0.4	20.2 ± 0.2
C5	C12	2.2 ± 0.2	1.3 ± 0.5	0.7 ± 0.2	1.4 ± 0.5	1.3 ± 0.5	1.0 ± 0.4
C5	C13	1.6 ± 0.2	1.4 ± 0.3	0.6 ± 0.1	0.8 ± 0.4	1.3 ± 0.4	0.6 ± 0.4
C5	C14	21.5 ± 0.2	20.9 ± 0.4	21.3 ± 0.3	20.3 ± 0.1	20.6 ± 0.5	20.4 ± 0.2
C6	C7	−0.7 ± 0.3	0.1 ± 0.4	−0.5 ± 0.4	0.2 ± 0.1	−0.2 ± 0.3	0.2 ± 0.5
C6	C8	21.7 ± 0.2	22.8 ± 0.4	23.2 ± 0.4	23.0 ± 0.2	22.9 ± 0.2	23.1 ± 0.4
C6	C9	21.0 ± 0.3	21.4 ± 0.4	21.8 ± 0.4	21.6 ± 0.3	21.3 ± 0.2	21.6 ± 0.4
C6	C10	22.7 ± 0.3	23.2 ± 0.4	23.2 ± 0.4	22.8 ± 0.2	22.7 ± 0.2	23.0 ± 0.5
C6	C11	19.6 ± 0.2	19.7 ± 0.3	20.0 ± 0.4	19.7 ± 0.1	19.3 ± 0.3	19.6 ± 0.4
C6	C12	0.7 ± 0.2	0.4 ± 0.3	−0.0 ± 0.5	0.9 ± 0.4	0.3 ± 0.3	0.4 ± 0.6
C6	C13	0.1 ± 0.3	0.5 ± 0.3	−0.1 ± 0.5	0.4 ± 0.5	0.2 ± 0.3	−0.0 ± 0.6
C6	C14	20.1 ± 0.2	20.0 ± 0.4	20.7 ± 0.5	19.8 ± 0.2	19.6 ± 0.2	19.8 ± 0.4
C7	C8	22.3 ± 0.1	22.8 ± 0.2	23.7 ± 0.2	22.8 ± 0.2	23.1 ± 0.1	22.9 ± 0.2
C7	C9	21.6 ± 0.2	21.3 ± 0.1	22.3 ± 0.2	21.3 ± 0.3	21.5 ± 0.1	21.4 ± 0.2
C7	C10	23.4 ± 0.1	23.2 ± 0.1	23.7 ± 0.2	22.5 ± 0.2	22.9 ± 0.1	22.8 ± 0.2
C7	C11	20.2 ± 0.1	19.6 ± 0.1	20.5 ± 0.3	19.4 ± 0.1	19.6 ± 0.1	19.4 ± 0.2
C7	C12	1.3 ± 0.2	0.3 ± 0.2	0.4 ± 0.2	0.7 ± 0.5	0.5 ± 0.2	0.2 ± 0.2
C7	C13	0.8 ± 0.1	0.4 ± 0.3	0.4 ± 0.3	0.1 ± 0.5	0.4 ± 0.2	−0.2 ± 0.2
C7	C14	20.7 ± 0.2	19.9 ± 0.2	21.1 ± 0.2	19.6 ± 0.2	19.8 ± 0.1	19.6 ± 0.2

C8	C9	−0.7	±	0.1	−1.4	±	0.1	−1.4	±	0.1	−1.4	±	0.1	−1.6	±	0.1	−1.5	±	0.1
C8	C10	1.0	±	0.2	0.4	±	0.1	−0.0	±	0.2	−0.2	±	0.1	−0.2	±	0.1	−0.2	±	0.1
C8	C11	−2.1	±	0.1	−3.2	±	0.1	−3.2	±	0.1	−3.3	±	0.1	−3.5	±	0.1	−3.6	±	0.1
C8	C12	−21.0	±	0.1	−22.5	±	0.2	−23.2	±	0.1	−22.1	±	0.5	−22.6	±	0.2	−22.7	±	0.3
C8	C13	−21.6	±	0.1	−22.4	±	0.3	−23.3	±	0.2	−22.6	±	0.4	−22.7	±	0.2	−23.2	±	0.3
C8	C14	−1.6	±	0.1	−2.9	±	0.1	−2.5	±	0.1	−3.2	±	0.1	−3.3	±	0.1	−3.3	±	0.1
C9	C10	1.7	±	0.1	1.8	±	0.1	1.4	±	0.1	1.2	±	0.1	1.4	±	0.0	1.4	±	0.1
C9	C11	−1.4	±	0.1	−1.8	±	0.1	−1.8	±	0.1	−1.9	±	0.2	−2.0	±	0.1	−2.0	±	0.1
C9	C12	−20.3	±	0.1	−21.0	±	0.2	−21.8	±	0.1	−20.6	±	0.6	−21.0	±	0.2	−21.2	±	0.3
C9	C13	−20.9	±	0.2	−21.0	±	0.3	−21.9	±	0.2	−21.2	±	0.3	−21.1	±	0.2	−21.6	±	0.3
C9	C14	−0.9	±	0.1	−1.5	±	0.1	−1.1	±	0.1	−1.8	±	0.1	−1.7	±	0.1	−1.8	±	0.1
C10	C11	−3.1	±	0.1	−3.6	±	0.1	−3.2	±	0.1	−3.1	±	0.1	−3.3	±	0.1	−3.4	±	0.1
C10	C12	−22.1	±	0.2	−22.8	±	0.2	−23.2	±	0.2	−21.8	±	0.5	−22.4	±	0.2	−22.6	±	0.3
C10	C13	−22.6	±	0.2	−22.8	±	0.3	−23.3	±	0.2	−22.4	±	0.3	−22.5	±	0.2	−23.0	±	0.3
C10	C14	−2.7	±	0.2	−3.3	±	0.1	−2.5	±	0.2	−3.0	±	0.1	−3.1	±	0.1	−3.1	±	0.1
C11	C12	−18.9	±	0.1	−19.3	±	0.2	−20.0	±	0.1	−18.7	±	0.5	−19.1	±	0.2	−19.2	±	0.3
C11	C13	−19.5	±	0.1	−19.2	±	0.2	−20.1	±	0.1	−19.3	±	0.4	−19.1	±	0.2	−19.6	±	0.3
C11	C14	0.5	±	0.1	0.3	±	0.1	0.7	±	0.2	0.1	±	0.1	0.2	±	0.1	0.2	±	0.1
C12	C13	−0.6	±	0.2	0.1	±	0.3	−0.1	±	0.2	−0.6	±	0.5	−0.1	±	0.2	−0.4	±	0.1
C12	C14	19.4	±	0.1	19.6	±	0.1	20.7	±	0.2	18.9	±	0.5	19.3	±	0.2	19.4	±	0.2
C13	C14	19.9	±	0.1	19.5	±	0.2	20.8	±	0.2	19.4	±	0.4	19.4	±	0.3	19.8	±	0.3
RMSE		2.19	±	0.2	2.27	±	0.2	2.82	±	0.2	2.08	±	0.2	2.27	±	0.2	2.39	±	0.2
MAE		1.83	±	0.2	1.87	±	0.2	2.31	±	0.2	1.73	±	0.2	1.86	±	0.2	1.96	±	0.2
r_{Spearman}		0.92			0.91			0.86			0.92			0.91			0.90		
$t_{\text{preparation}}$		309.2 ns			309.2 ns			309.2 ns			309.2 ns			309.2 ns			309.2 ns		
$t_{\text{production}}$		67 ns			67 ns			67 ns			67 ns			67 ns			67 ns		

Table 7.7: $\Delta\Delta G_{\text{hyd}}$ for the 14 molecules of set C from experiment,[155,173] calculated from the hydration free energies obtained with MBAR,[155,173] and from RE-EDS calculations using the RF schemes CG^{orig}, AT^{orig}, and AT^{shift} with $R_{\text{RF}} = 1.2$ nm. The RE-EDS results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ [155,173]	$\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$ [155,173]	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,CG}^{\text{orig}}}$	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,AT}^{\text{orig}}}$	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,AT}^{\text{shift}}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	0.5 ± 2.6	0.8 ± 0.2	0.4 ± 0.2	0.4 ± 0.2	0.1 ± 0.2
C1	C3	0.0 ± 2.6	−0.3 ± 0.2	−1.0 ± 0.2	−0.9 ± 0.1	−1.2 ± 0.1
C1	C4	0.3 ± 2.6	0.4 ± 0.2	0.5 ± 0.2	0.6 ± 0.1	0.4 ± 0.1
C1	C5	−23.8 ± 1.2	−20.6 ± 0.2	−20.7 ± 0.3	−20.7 ± 0.2	−20.6 ± 0.2
C1	C6	−21.9 ± 2.6	−20.0 ± 0.2	−20.5 ± 0.3	−19.9 ± 0.6	−20.1 ± 0.2
C1	C7	−19.2 ± 2.6	−19.9 ± 0.2	−19.9 ± 0.1	−19.9 ± 0.2	−19.9 ± 0.1
C1	C8	0.4 ± 2.6	3.1 ± 0.2	3.0 ± 0.2	3.1 ± 0.2	3.1 ± 0.1
C1	C9	−0.9 ± 2.6	1.3 ± 0.2	1.3 ± 0.1	1.4 ± 0.1	1.4 ± 0.2
C1	C10	−0.5 ± 2.6	2.5 ± 0.2	2.7 ± 0.2	2.8 ± 0.2	2.6 ± 0.1
C1	C11	−2.3 ± 2.6	−0.7 ± 0.2	−0.7 ± 0.1	−0.6 ± 0.1	−0.6 ± 0.1
C1	C12	−19.4 ± 2.6	−19.0 ± 0.2	−19.5 ± 0.1	−19.3 ± 0.2	−19.7 ± 0.2
C1	C13	−19.5 ± 2.6	−19.7 ± 0.2	−20.0 ± 0.1	−19.8 ± 0.2	−20.0 ± 0.4
C1	C14	−3.5 ± 2.6	−1.1 ± 0.2	−0.7 ± 0.1	−0.5 ± 0.1	−0.5 ± 0.2
C2	C3	−0.5 ± 3.6	−1.0 ± 0.2	−1.4 ± 0.2	−1.3 ± 0.2	−1.3 ± 0.2
C2	C4	−0.2 ± 3.6	−0.4 ± 0.2	0.0 ± 0.2	0.2 ± 0.2	0.3 ± 0.2
C2	C5	−24.3 ± 2.6	−21.3 ± 0.2	−21.1 ± 0.3	−21.1 ± 0.3	−20.8 ± 0.2
C2	C6	−22.3 ± 3.6	−20.8 ± 0.2	−20.9 ± 0.4	−20.3 ± 0.6	−20.3 ± 0.2
C2	C7	−19.7 ± 3.6	−20.6 ± 0.2	−20.4 ± 0.2	−20.3 ± 0.3	−20.0 ± 0.2
C2	C8	−0.0 ± 3.6	2.4 ± 0.2	2.5 ± 0.2	2.7 ± 0.2	3.0 ± 0.1
C2	C9	−1.4 ± 3.6	0.6 ± 0.2	0.9 ± 0.2	1.0 ± 0.2	1.3 ± 0.1
C2	C10	−0.9 ± 3.6	1.8 ± 0.2	2.2 ± 0.2	2.4 ± 0.2	2.5 ± 0.1
C2	C11	−2.8 ± 3.6	−1.4 ± 0.2	−1.2 ± 0.2	−1.0 ± 0.2	−0.8 ± 0.1
C2	C12	−19.8 ± 3.6	−19.7 ± 0.2	−19.9 ± 0.2	−19.7 ± 0.3	−19.9 ± 0.3
C2	C13	−20.0 ± 3.6	−20.4 ± 0.2	−20.5 ± 0.3	−20.2 ± 0.2	−20.2 ± 0.4
C2	C14	−4.0 ± 3.6	−1.9 ± 0.2	−1.1 ± 0.2	−0.9 ± 0.2	−0.6 ± 0.1
C3	C4	0.3 ± 3.6	0.6 ± 0.2	1.5 ± 0.2	1.6 ± 0.2	1.6 ± 0.2
C3	C5	−23.8 ± 2.6	−20.3 ± 0.2	−19.7 ± 0.4	−19.8 ± 0.2	−19.5 ± 0.2
C3	C6	−21.9 ± 3.6	−19.8 ± 0.2	−19.5 ± 0.5	−19.0 ± 0.4	−19.0 ± 0.2
C3	C7	−19.2 ± 3.6	−19.6 ± 0.2	−18.9 ± 0.3	−18.9 ± 0.2	−18.7 ± 0.2
C3	C8	0.4 ± 3.6	3.4 ± 0.2	4.0 ± 0.3	4.1 ± 0.2	4.3 ± 0.1
C3	C9	−0.9 ± 3.6	1.6 ± 0.2	2.3 ± 0.2	2.4 ± 0.1	2.6 ± 0.2
C3	C10	−0.5 ± 3.6	2.8 ± 0.2	3.6 ± 0.1	3.7 ± 0.2	3.8 ± 0.1
C3	C11	−2.3 ± 3.6	−0.4 ± 0.2	0.3 ± 0.3	0.3 ± 0.1	0.5 ± 0.1
C3	C12	−19.4 ± 3.6	−18.7 ± 0.2	−18.5 ± 0.2	−18.4 ± 0.2	−18.6 ± 0.2
C3	C13	−19.5 ± 3.6	−19.4 ± 0.2	−19.0 ± 0.4	−18.9 ± 0.2	−18.9 ± 0.3
C3	C14	−3.5 ± 3.6	−0.9 ± 0.2	0.3 ± 0.2	0.4 ± 0.1	0.7 ± 0.1
C4	C5	−24.1 ± 2.6	−21.0 ± 0.2	−21.2 ± 0.4	−21.3 ± 0.3	−21.0 ± 0.1
C4	C6	−22.2 ± 3.6	−20.4 ± 0.2	−21.0 ± 0.4	−20.5 ± 0.5	−20.6 ± 0.2
C4	C7	−19.5 ± 3.6	−20.3 ± 0.2	−20.4 ± 0.3	−20.5 ± 0.2	−20.3 ± 0.1
C4	C8	0.1 ± 3.6	2.8 ± 0.2	2.5 ± 0.2	2.5 ± 0.2	2.7 ± 0.2
C4	C9	−1.2 ± 3.6	1.0 ± 0.2	0.8 ± 0.2	0.8 ± 0.1	1.0 ± 0.2
C4	C10	−0.8 ± 3.6	2.1 ± 0.2	2.2 ± 0.2	2.2 ± 0.1	2.2 ± 0.1
C4	C11	−2.6 ± 3.6	−1.0 ± 0.2	−1.2 ± 0.3	−1.3 ± 0.1	−1.1 ± 0.2
C4	C12	−19.7 ± 3.6	−19.4 ± 0.2	−20.0 ± 0.2	−20.0 ± 0.3	−20.2 ± 0.3
C4	C13	−19.8 ± 3.6	−20.0 ± 0.2	−20.5 ± 0.4	−20.4 ± 0.1	−20.4 ± 0.4

C4	C14	−3.8	±	3.6	−1.5	±	0.2	−1.2	±	0.2	−1.1	±	0.1	−0.9	±	0.1			
C5	C6	2.0	±	2.6	0.5	±	0.2	0.2	±	0.2	0.8	±	0.6	0.5	±	0.2			
C5	C7	4.6	±	2.6	0.7	±	0.2	0.8	±	0.1	0.9	±	0.3	0.7	±	0.2			
C5	C8	24.3	±	2.6	23.7	±	0.2	23.7	±	0.3	23.9	±	0.2	23.8	±	0.2			
C5	C9	22.9	±	2.6	21.9	±	0.2	22.0	±	0.2	22.2	±	0.2	22.0	±	0.2			
C5	C10	23.4	±	2.6	23.1	±	0.2	23.3	±	0.4	23.5	±	0.3	23.2	±	0.1			
C5	C11	21.5	±	2.6	19.9	±	0.2	20.0	±	0.3	20.1	±	0.2	20.0	±	0.2			
C5	C12	4.5	±	2.6	1.6	±	0.2	1.2	±	0.3	1.4	±	0.2	0.9	±	0.4			
C5	C13	4.3	±	2.6	0.9	±	0.2	0.6	±	0.3	0.9	±	0.3	0.6	±	0.5			
C5	C14	20.3	±	2.6	19.5	±	0.2	20.0	±	0.2	20.2	±	0.2	20.2	±	0.2			
C6	C7	2.7	±	3.6	0.2	±	0.2	0.6	±	0.3	0.1	±	0.4	0.3	±	0.3			
C6	C8	22.3	±	3.6	23.2	±	0.2	23.5	±	0.4	23.1	±	0.5	23.3	±	0.2			
C6	C9	21.0	±	3.6	21.4	±	0.2	21.8	±	0.4	21.4	±	0.5	21.5	±	0.3			
C6	C10	21.4	±	3.6	22.6	±	0.2	23.1	±	0.5	22.7	±	0.5	22.8	±	0.2			
C6	C11	19.5	±	3.6	19.4	±	0.2	19.8	±	0.3	19.3	±	0.5	19.5	±	0.2			
C6	C12	2.5	±	3.6	1.0	±	0.2	1.0	±	0.3	0.6	±	0.6	0.4	±	0.3			
C6	C13	2.3	±	3.6	0.4	±	0.2	0.5	±	0.3	0.1	±	0.6	0.1	±	0.4			
C6	C14	18.4	±	3.6	18.9	±	0.2	19.8	±	0.4	19.4	±	0.5	19.7	±	0.2			
C7	C8	19.6	±	3.6	23.0	±	0.2	22.9	±	0.2	23.0	±	0.1	23.0	±	0.1			
C7	C9	18.3	±	3.6	21.2	±	0.2	21.2	±	0.1	21.3	±	0.2	21.3	±	0.2			
C7	C10	18.7	±	3.6	22.4	±	0.2	22.6	±	0.2	22.7	±	0.2	22.5	±	0.1			
C7	C11	16.9	±	3.6	19.2	±	0.2	19.2	±	0.1	19.2	±	0.2	19.2	±	0.2			
C7	C12	−0.2	±	3.6	0.9	±	0.2	0.4	±	0.2	0.5	±	0.3	0.1	±	0.3			
C7	C13	−0.3	±	3.6	0.2	±	0.2	−0.1	±	0.2	0.1	±	0.3	−0.1	±	0.4			
C7	C14	15.7	±	3.6	18.7	±	0.2	19.2	±	0.1	19.3	±	0.1	19.4	±	0.2			
C8	C9	−1.3	±	3.6	−1.8	±	0.1	−1.7	±	0.1	−1.7	±	0.2	−1.8	±	0.1			
C8	C10	−0.9	±	3.6	−0.6	±	0.2	−0.3	±	0.2	−0.3	±	0.1	−0.5	±	0.1			
C8	C11	−2.8	±	3.6	−3.8	±	0.2	−3.7	±	0.2	−3.8	±	0.1	−3.8	±	0.0			
C8	C12	−19.8	±	3.6	−22.1	±	0.2	−22.5	±	0.3	−22.5	±	0.3	−22.9	±	0.2			
C8	C13	−20.0	±	3.6	−22.8	±	0.2	−23.0	±	0.3	−23.0	±	0.3	−23.2	±	0.3			
C8	C14	−3.9	±	3.6	−4.3	±	0.2	−3.7	±	0.1	−3.7	±	0.1	−3.6	±	0.1			
C9	C10	0.5	±	3.6	1.2	±	0.2	1.3	±	0.1	1.4	±	0.1	1.2	±	0.1			
C9	C11	−1.4	±	3.6	−2.0	±	0.2	−2.0	±	0.1	−2.1	±	0.1	−2.0	±	0.1			
C9	C12	−18.5	±	3.6	−20.3	±	0.2	−20.8	±	0.1	−20.8	±	0.2	−21.1	±	0.3			
C9	C13	−18.6	±	3.6	−21.0	±	0.2	−21.3	±	0.3	−21.3	±	0.2	−21.4	±	0.4			
C9	C14	−2.6	±	3.6	−2.5	±	0.2	−2.0	±	0.0	−2.0	±	0.1	−1.9	±	0.1			
C10	C11	−1.9	±	3.6	−3.2	±	0.2	−3.4	±	0.2	−3.5	±	0.1	−3.3	±	0.1			
C10	C12	−18.9	±	3.6	−21.5	±	0.2	−22.1	±	0.2	−22.2	±	0.3	−22.4	±	0.3			
C10	C13	−19.1	±	3.6	−22.2	±	0.2	−22.7	±	0.3	−22.6	±	0.2	−22.6	±	0.4			
C10	C14	−3.1	±	3.6	−3.6	±	0.2	−3.3	±	0.1	−3.3	±	0.1	−3.1	±	0.1			
C11	C12	−17.0	±	3.6	−18.3	±	0.2	−18.8	±	0.2	−18.7	±	0.2	−19.1	±	0.2			
C11	C13	−17.2	±	3.6	−19.0	±	0.2	−19.3	±	0.1	−19.2	±	0.2	−19.4	±	0.3			
C11	C14	−1.2	±	3.6	−0.5	±	0.2	0.1	±	0.1	0.1	±	0.1	0.2	±	0.1			
C12	C13	−0.2	±	3.6	−0.7	±	0.2	−0.5	±	0.2	−0.5	±	0.3	−0.3	±	0.2			
C12	C14	15.9	±	3.6	17.9	±	0.2	18.8	±	0.1	18.8	±	0.3	19.3	±	0.3			
C13	C14	16.0	±	3.6	18.5	±	0.2	19.4	±	0.2	19.3	±	0.2	19.6	±	0.4			
RMSE					1.94	±	0.1	2.15	±	0.2	2.19	±	0.2	2.29	±	0.2			
MAE					1.60	±	0.1	1.78	±	0.2	1.80	±	0.2	1.89	±	0.2			
r_{Spearman}							0.94			0.92			0.92			0.91			
$t_{\text{preparation}}$							42 ns			309.2 ns			309.2 ns			309.2 ns			
$t_{\text{production}}$							1400 ns			67 ns			67 ns			67 ns			

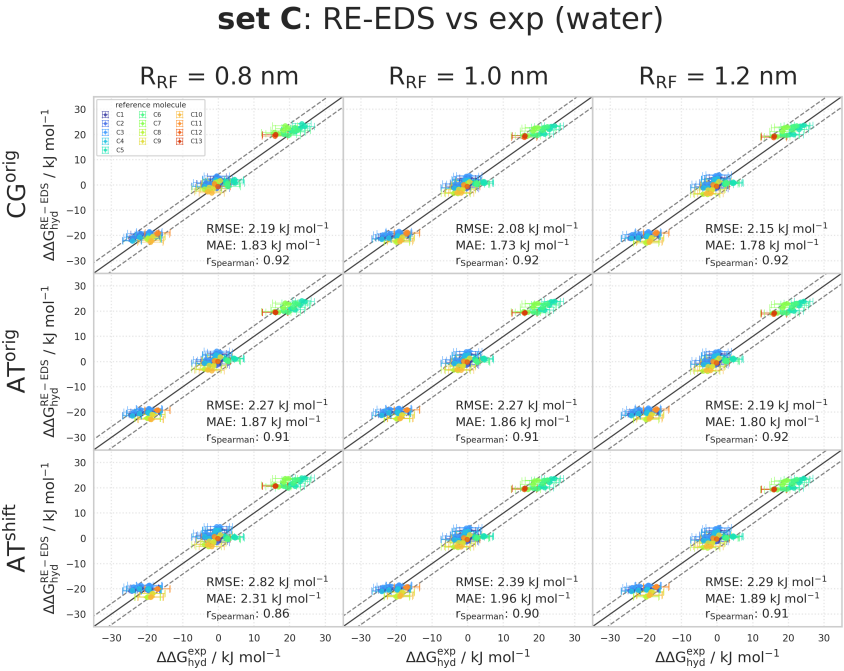


Figure 7.7: Comparison of the relative hydration free energies of set C: $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$ obtained from RE-EDS calculations in GROMOS versus $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\mathrm{RF}} = 0.8$ nm (left), $R_{\mathrm{RF}} = 1.0$ nm (middle), and $R_{\mathrm{RF}} = 1.2$ nm (right), respectively. The three rows correspond to the RF schemes $\mathrm{CG}^{\mathrm{orig}}$ (top), $\mathrm{AT}^{\mathrm{orig}}$ (middle), and $\mathrm{AT}^{\mathrm{shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.6 and 7.7.

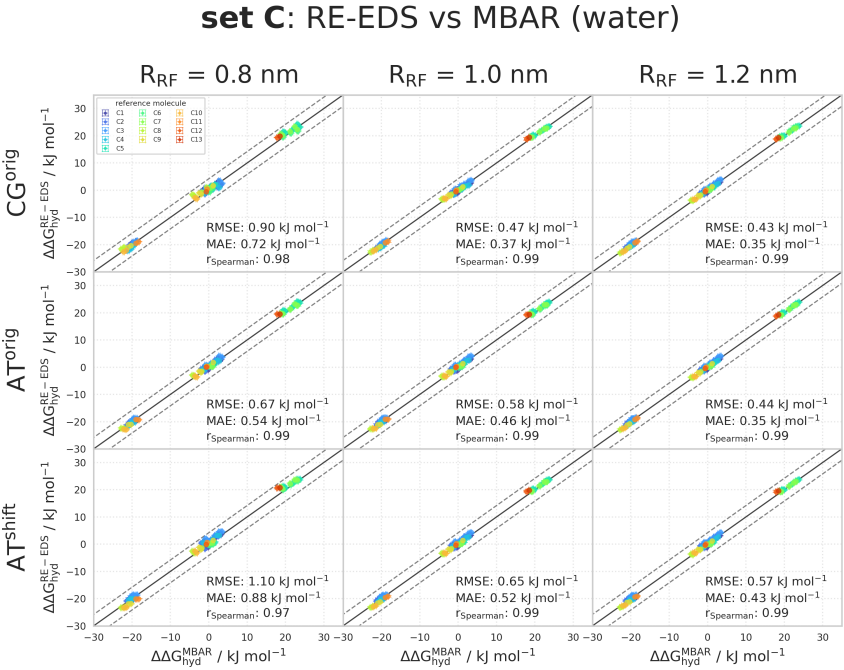


Figure 7.8: Comparison of the relative hydration free energies of set C: $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$ obtained from RE-EDS calculations in GROMOS[27] versus $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\mathrm{RF}} = 0.8$ nm (left), $R_{\mathrm{RF}} = 1.0$ nm (middle), and $R_{\mathrm{RF}} = 1.2$ nm (right), respectively. The three rows correspond to the RF schemes $\mathrm{CG}^{\mathrm{orig}}$ (top), $\mathrm{AT}^{\mathrm{orig}}$ (middle), and $\mathrm{AT}^{\mathrm{shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.6 and 7.7.

Relative Solvation Free Energies in Chloroform of set C

Table 7.8: $\Delta\Delta G_{\mathrm{CHCl_3}}$ for the 14 molecules of set C calculated from RE-EDS calculations using the RF schemes $\mathrm{CG^{orig}}$, $\mathrm{AT^{orig}}$, and $\mathrm{AT^{shift}}$ with $R_{\mathrm{RF}} = 0.8$ nm and $R_{\mathrm{RF}} = 1.0$ nm. The results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

		$R_{\mathrm{RF}} = 0.8$ nm			$R_{\mathrm{RF}} = 1.0$ nm		
Molecules		$\Delta\Delta G^{\mathrm{RE\text{-}EDS,CG^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{shift}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,CG^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{shift}}}_{\mathrm{CHCl_3}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	−1.4 ± 0.2	−4.4 ± 0.3	−3.9 ± 0.2	−3.7 ± 0.1	−4.0 ± 0.3	−4.1 ± 0.2
C1	C3	−1.9 ± 0.1	−5.8 ± 0.3	−5.2 ± 0.1	−5.2 ± 0.3	−5.3 ± 0.3	−5.5 ± 0.1
C1	C4	−1.3 ± 0.3	−4.8 ± 0.4	−3.9 ± 0.3	−3.9 ± 0.3	−3.9 ± 0.5	−4.2 ± 0.3
C1	C5	0.4 ± 0.1	−2.7 ± 0.4	−2.2 ± 0.2	−2.6 ± 0.1	−3.1 ± 0.1	−2.8 ± 0.2
C1	C6	−1.0 ± 0.1	−7.1 ± 0.4	−6.4 ± 0.4	−6.7 ± 0.1	−7.5 ± 0.2	−7.2 ± 0.2
C1	C7	−0.4 ± 0.1	−3.6 ± 0.4	−3.4 ± 0.2	−3.5 ± 0.1	−3.9 ± 0.2	−3.8 ± 0.2
C1	C8	1.3 ± 0.1	3.0 ± 0.3	3.7 ± 0.2	3.4 ± 0.1	3.4 ± 0.1	3.7 ± 0.2
C1	C9	0.8 ± 0.1	−0.7 ± 0.3	−0.0 ± 0.2	−0.5 ± 0.1	−0.8 ± 0.1	−0.5 ± 0.2
C1	C10	0.4 ± 0.1	−4.7 ± 0.3	−4.1 ± 0.2	−4.4 ± 0.1	−5.1 ± 0.2	−4.8 ± 0.2
C1	C11	0.3 ± 0.1	−3.4 ± 0.3	−2.8 ± 0.2	−3.5 ± 0.1	−4.0 ± 0.1	−3.6 ± 0.3
C1	C12	−2.8 ± 0.3	−8.3 ± 0.3	−7.6 ± 0.1	−7.8 ± 0.3	−8.1 ± 0.1	−8.4 ± 0.1
C1	C13	−2.2 ± 0.2	−8.0 ± 0.2	−7.6 ± 0.2	−7.6 ± 0.1	−8.5 ± 0.2	−8.4 ± 0.2
C1	C14	0.2 ± 0.1	−5.0 ± 0.3	−4.2 ± 0.2	−5.0 ± 0.1	−5.9 ± 0.1	−5.4 ± 0.2
C2	C3	−0.5 ± 0.2	−1.4 ± 0.2	−1.3 ± 0.2	−1.5 ± 0.3	−1.3 ± 0.2	−1.5 ± 0.2
C2	C4	0.1 ± 0.3	−0.4 ± 0.3	−0.0 ± 0.3	−0.2 ± 0.3	0.1 ± 0.3	−0.1 ± 0.3
C2	C5	1.9 ± 0.2	1.7 ± 0.3	1.7 ± 0.2	1.1 ± 0.2	1.0 ± 0.3	1.3 ± 0.2
C2	C6	0.4 ± 0.2	−2.7 ± 0.3	−2.5 ± 0.3	−3.0 ± 0.2	−3.5 ± 0.2	−3.1 ± 0.2
C2	C7	1.0 ± 0.2	0.8 ± 0.3	0.6 ± 0.2	0.2 ± 0.1	0.2 ± 0.2	0.2 ± 0.2
C2	C8	2.7 ± 0.2	7.4 ± 0.2	7.7 ± 0.2	7.1 ± 0.2	7.5 ± 0.3	7.7 ± 0.2
C2	C9	2.2 ± 0.2	3.7 ± 0.2	3.9 ± 0.2	3.2 ± 0.2	3.3 ± 0.2	3.6 ± 0.2
C2	C10	1.8 ± 0.2	−0.3 ± 0.2	−0.1 ± 0.2	−0.7 ± 0.2	−1.1 ± 0.3	−0.7 ± 0.2
C2	C11	1.7 ± 0.2	1.0 ± 0.2	1.1 ± 0.2	0.2 ± 0.2	0.1 ± 0.3	0.5 ± 0.2
C2	C12	−1.4 ± 0.3	−3.9 ± 0.2	−3.7 ± 0.2	−4.1 ± 0.3	−4.1 ± 0.2	−4.3 ± 0.1
C2	C13	−0.7 ± 0.2	−3.6 ± 0.2	−3.7 ± 0.3	−3.9 ± 0.2	−4.4 ± 0.2	−4.3 ± 0.2
C2	C14	1.6 ± 0.2	−0.6 ± 0.2	−0.2 ± 0.2	−1.3 ± 0.1	−1.8 ± 0.3	−1.3 ± 0.2
C3	C4	0.6 ± 0.2	1.0 ± 0.2	1.3 ± 0.2	1.3 ± 0.3	1.4 ± 0.3	1.4 ± 0.3
C3	C5	2.4 ± 0.1	3.1 ± 0.2	3.0 ± 0.1	2.6 ± 0.2	2.2 ± 0.2	2.7 ± 0.2
C3	C6	0.9 ± 0.2	−1.3 ± 0.2	−1.2 ± 0.3	−1.5 ± 0.3	−2.2 ± 0.1	−1.7 ± 0.2
C3	C7	1.5 ± 0.1	2.1 ± 0.1	1.9 ± 0.1	1.7 ± 0.2	1.5 ± 0.1	1.7 ± 0.2
C3	C8	3.2 ± 0.1	8.8 ± 0.1	9.0 ± 0.1	8.6 ± 0.3	8.7 ± 0.2	9.2 ± 0.2
C3	C9	2.7 ± 0.1	5.1 ± 0.1	5.2 ± 0.1	4.7 ± 0.3	4.5 ± 0.2	5.1 ± 0.2
C3	C10	2.3 ± 0.1	1.1 ± 0.1	1.2 ± 0.1	0.8 ± 0.3	0.2 ± 0.2	0.8 ± 0.2
C3	C11	2.2 ± 0.1	2.4 ± 0.1	2.4 ± 0.1	1.7 ± 0.2	1.3 ± 0.2	1.9 ± 0.3
C3	C12	−0.9 ± 0.2	−2.5 ± 0.1	−2.4 ± 0.1	−2.6 ± 0.1	−2.8 ± 0.2	−2.8 ± 0.1
C3	C13	−0.2 ± 0.2	−2.2 ± 0.1	−2.4 ± 0.2	−2.4 ± 0.3	−3.1 ± 0.2	−2.9 ± 0.2
C3	C14	2.1 ± 0.1	0.8 ± 0.1	1.1 ± 0.1	0.3 ± 0.2	−0.6 ± 0.2	0.2 ± 0.2
C4	C5	1.7 ± 0.3	2.1 ± 0.2	1.7 ± 0.2	1.3 ± 0.2	0.9 ± 0.4	1.4 ± 0.2
C4	C6	0.3 ± 0.4	−2.3 ± 0.2	−2.5 ± 0.3	−2.8 ± 0.4	−3.6 ± 0.3	−3.0 ± 0.3
C4	C7	0.9 ± 0.2	1.1 ± 0.1	0.6 ± 0.2	0.4 ± 0.3	0.1 ± 0.3	0.3 ± 0.2
C4	C8	2.6 ± 0.3	7.8 ± 0.2	7.7 ± 0.1	7.3 ± 0.2	7.4 ± 0.4	7.8 ± 0.2
C4	C9	2.1 ± 0.3	4.1 ± 0.2	3.9 ± 0.2	3.4 ± 0.3	3.2 ± 0.4	3.7 ± 0.1
C4	C10	1.7 ± 0.3	0.1 ± 0.2	−0.1 ± 0.2	−0.5 ± 0.2	−1.1 ± 0.4	−0.6 ± 0.2
C4	C11	1.6 ± 0.3	1.4 ± 0.2	1.1 ± 0.2	0.4 ± 0.2	−0.0 ± 0.4	0.6 ± 0.1
C4	C12	−1.5 ± 0.2	−3.5 ± 0.2	−3.7 ± 0.3	−3.9 ± 0.2	−4.2 ± 0.4	−4.2 ± 0.2
C4	C13	−0.9 ± 0.2	−3.2 ± 0.2	−3.6 ± 0.3	−3.7 ± 0.3	−4.5 ± 0.4	−4.2 ± 0.1
C4	C14	1.5 ± 0.3	−0.3 ± 0.2	−0.2 ± 0.1	−1.0 ± 0.3	−1.9 ± 0.5	−1.2 ± 0.2
C5	C6	−1.4 ± 0.1	−4.4 ± 0.1	−4.2 ± 0.2	−4.1 ± 0.1	−4.5 ± 0.1	−4.4 ± 0.1
C5	C7	−0.8 ± 0.1	−1.0 ± 0.1	−1.1 ± 0.1	−0.9 ± 0.1	−0.8 ± 0.1	−1.0 ± 0.1
C5	C8	0.8 ± 0.1	5.7 ± 0.1	6.0 ± 0.1	6.0 ± 0.1	6.5 ± 0.1	6.5 ± 0.0
C5	C9	0.4 ± 0.1	2.0 ± 0.1	2.2 ± 0.1	2.1 ± 0.1	2.3 ± 0.1	2.4 ± 0.1
C5	C10	−0.0 ± 0.0	−2.0 ± 0.2	−1.8 ± 0.0	−1.8 ± 0.1	−2.0 ± 0.1	−2.0 ± 0.1
C5	C11	−0.2 ± 0.1	−0.7 ± 0.1	−0.5 ± 0.1	−0.9 ± 0.1	−0.9 ± 0.1	−0.8 ± 0.1
C5	C12	−3.2 ± 0.3	−5.6 ± 0.1	−5.4 ± 0.1	−5.2 ± 0.2	−5.1 ± 0.1	−5.6 ± 0.1
C5	C13	−2.6 ± 0.1	−5.3 ± 0.1	−5.3 ± 0.1	−5.0 ± 0.1	−5.4 ± 0.1	−5.6 ± 0.1
C5	C14	−0.2 ± 0.1	−2.3 ± 0.1	−1.9 ± 0.1	−2.3 ± 0.1	−2.8 ± 0.1	−2.6 ± 0.1
C6	C7	0.6 ± 0.1	3.4 ± 0.2	3.1 ± 0.2	3.2 ± 0.1	3.7 ± 0.1	3.4 ± 0.1
C6	C8	2.3 ± 0.1	10.1 ± 0.2	10.2 ± 0.2	10.1 ± 0.2	11.0 ± 0.1	10.9 ± 0.1
C6	C9	1.8 ± 0.1	6.4 ± 0.2	6.4 ± 0.2	6.2 ± 0.1	6.8 ± 0.1	6.8 ± 0.2
C6	C10	1.4 ± 0.1	2.4 ± 0.2	2.4 ± 0.2	2.3 ± 0.1	2.4 ± 0.1	2.4 ± 0.2
C6	C11	1.3 ± 0.1	3.7 ± 0.2	3.6 ± 0.2	3.2 ± 0.2	3.6 ± 0.1	3.6 ± 0.2
C6	C12	−1.8 ± 0.3	−1.2 ± 0.2	−1.2 ± 0.3	−1.1 ± 0.3	−0.6 ± 0.2	−1.2 ± 0.2
C6	C13	−1.2 ± 0.3	−0.9 ± 0.2	−1.2 ± 0.3	−0.9 ± 0.2	−0.9 ± 0.1	−1.2 ± 0.2
C6	C14	1.2 ± 0.1	2.1 ± 0.1	2.3 ± 0.2	1.7 ± 0.1	1.7 ± 0.1	1.8 ± 0.1
C7	C8	1.7 ± 0.1	6.7 ± 0.1	7.1 ± 0.1	6.9 ± 0.2	7.3 ± 0.1	7.5 ± 0.1
C7	C9	1.2 ± 0.1	3.0 ± 0.1	3.3 ± 0.0	3.0 ± 0.1	3.1 ± 0.1	3.4 ± 0.1
C7	C10	0.8 ± 0.1	−1.0 ± 0.1	−0.7 ± 0.1	−0.9 ± 0.1	−1.2 ± 0.2	−1.0 ± 0.2
C7	C11	0.7 ± 0.0	0.2 ± 0.1	0.6 ± 0.0	0.0 ± 0.1	−0.1 ± 0.1	0.2 ± 0.1
C7	C12	−2.4 ± 0.2	−4.6 ± 0.1	−4.3 ± 0.1	−4.3 ± 0.2	−4.3 ± 0.1	−4.5 ± 0.1
C7	C13	−1.8 ± 0.1	−4.3 ± 0.2	−4.2 ± 0.2	−4.1 ± 0.1	−4.6 ± 0.1	−4.6 ± 0.1
C7	C14	0.6 ± 0.1	−1.4 ± 0.1	−0.8 ± 0.1	−1.5 ± 0.1	−2.0 ± 0.1	−1.5 ± 0.1

C8	C9	−0.5 ± 0.0	−3.7 ± 0.1	−3.8 ± 0.1	−3.9 ± 0.1	−4.2 ± 0.0	−4.1 ± 0.0
C8	C10	−0.9 ± 0.0	−7.7 ± 0.0	−7.8 ± 0.1	−7.9 ± 0.1	−8.5 ± 0.1	−8.4 ± 0.1
C8	C11	−1.0 ± 0.0	−6.4 ± 0.1	−6.5 ± 0.1	−6.9 ± 0.1	−7.4 ± 0.0	−7.3 ± 0.1
C8	C12	−4.1 ± 0.3	−11.3 ± 0.1	−11.4 ± 0.2	−11.3 ± 0.3	−11.5 ± 0.1	−12.0 ± 0.1
C8	C13	−3.4 ± 0.2	−11.0 ± 0.1	−11.3 ± 0.2	−11.0 ± 0.1	−11.9 ± 0.1	−12.0 ± 0.0
C8	C14	−1.1 ± 0.0	−8.1 ± 0.1	−7.9 ± 0.0	−8.4 ± 0.1	−9.3 ± 0.1	−9.0 ± 0.0
C9	C10	−0.4 ± 0.0	−4.0 ± 0.1	−4.0 ± 0.1	−4.0 ± 0.1	−4.3 ± 0.1	−4.3 ± 0.1
C9	C11	−0.5 ± 0.0	−2.7 ± 0.1	−2.8 ± 0.0	−3.0 ± 0.1	−3.2 ± 0.0	−3.2 ± 0.1
C9	C12	−3.6 ± 0.2	−7.6 ± 0.1	−7.6 ± 0.1	−7.4 ± 0.2	−7.3 ± 0.0	−7.9 ± 0.1
C9	C13	−3.0 ± 0.2	−7.3 ± 0.1	−7.6 ± 0.2	−7.1 ± 0.1	−7.7 ± 0.1	−7.9 ± 0.0
C9	C14	−0.6 ± 0.0	−4.3 ± 0.1	−4.1 ± 0.1	−4.5 ± 0.1	−5.1 ± 0.1	−4.9 ± 0.1
C10	C11	−0.1 ± 0.0	1.3 ± 0.1	1.3 ± 0.1	0.9 ± 0.1	1.1 ± 0.1	1.2 ± 0.2
C10	C12	−3.2 ± 0.2	−3.6 ± 0.1	−3.6 ± 0.1	−3.4 ± 0.2	−3.0 ± 0.2	−3.6 ± 0.2
C10	C13	−2.6 ± 0.2	−3.3 ± 0.1	−3.5 ± 0.2	−3.2 ± 0.1	−3.4 ± 0.2	−3.6 ± 0.1
C10	C14	−0.2 ± 0.0	−0.4 ± 0.1	−0.1 ± 0.1	−0.5 ± 0.1	−0.8 ± 0.2	−0.6 ± 0.1
C11	C12	−3.1 ± 0.2	−4.9 ± 0.1	−4.8 ± 0.1	−4.3 ± 0.2	−4.1 ± 0.1	−4.8 ± 0.2
C11	C13	−2.4 ± 0.1	−4.6 ± 0.1	−4.8 ± 0.2	−4.1 ± 0.1	−4.5 ± 0.1	−4.8 ± 0.1
C11	C14	−0.1 ± 0.0	−1.6 ± 0.1	−1.4 ± 0.1	−1.5 ± 0.1	−1.9 ± 0.1	−1.8 ± 0.1
C12	C13	0.6 ± 0.3	0.3 ± 0.1	0.0 ± 0.2	0.2 ± 0.3	−0.3 ± 0.1	−0.0 ± 0.1
C12	C14	3.0 ± 0.2	3.3 ± 0.1	3.5 ± 0.2	2.9 ± 0.2	2.3 ± 0.1	3.0 ± 0.1
C13	C14	2.4 ± 0.1	3.0 ± 0.1	3.4 ± 0.2	2.7 ± 0.1	2.6 ± 0.1	3.0 ± 0.1
RMSE		5.61 ± 0.5	3.68 ± 0.3	3.58 ± 0.3	3.53 ± 0.3	3.39 ± 0.3	3.39 ± 0.3
MAE		4.63 ± 0.5	3.05 ± 0.3	2.96 ± 0.3	2.87 ± 0.2	2.73 ± 0.2	2.78 ± 0.3
r_{Spearman}		0.66	0.78	0.80	0.80	0.80	80
$t_{\text{preparation}}$		234 ns	277.2 ns	277.2 ns	277.2 ns	277.2 ns	277.2 ns
$t_{\text{production}}$		57 ns	59 ns	59 ns	59 ns	59 ns	59 ns

Table 7.9: $\Delta\Delta G_{\text{CHCl}_3}$ for the 14 molecules of set C from experiment,[174] and from RE-EDS calculations using the RF schemes CG^{orig}, AT^{orig}, and AT^{shift} with $R_{\text{RF}} = 1.2$ nm. The RE-EDS results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G^{\text{exp}}_{\text{CHCl}_3}$ [174]	$\Delta\Delta G^{\text{RE-EDS,CG}^{\text{orig}}}_{\text{CHCl}_3}$	$\Delta\Delta G^{\text{RE-EDS,AT}^{\text{orig}}}_{\text{CHCl}_3}$	$\Delta\Delta G^{\text{RE-EDS,AT}^{\text{shift}}}_{\text{CHCl}_3}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	−1.5 ± 3.6	−4.0 ± 0.4	−3.9 ± 0.2	−4.1 ± 0.2
C1	C3	−3.1 ± 3.6	−5.4 ± 0.2	−5.4 ± 0.2	−5.4 ± 0.2
C1	C4	−1.6 ± 3.6	−4.2 ± 0.2	−4.0 ± 0.3	−4.3 ± 0.2
C1	C5	−6.9 ± 3.6	−3.1 ± 0.1	−3.1 ± 0.1	−3.1 ± 0.1
C1	C6	−8.8 ± 3.6	−7.5 ± 0.3	−7.5 ± 0.2	−7.3 ± 0.1
C1	C7	−7.8 ± 3.6	−4.0 ± 0.2	−4.0 ± 0.1	−4.0 ± 0.1
C1	C8	5.1 ± 3.6	3.5 ± 0.2	3.7 ± 0.1	3.6 ± 0.1
C1	C9	0.1 ± 3.6	−0.8 ± 0.2	−0.7 ± 0.1	−0.7 ± 0.1
C1	C10	−3.5 ± 3.6	−5.2 ± 0.2	−5.2 ± 0.1	−5.1 ± 0.2
C1	C11	−2.5 ± 3.6	−4.1 ± 0.2	−4.1 ± 0.1	−4.1 ± 0.1
C1	C12	−11.5 ± 3.6	−8.3 ± 0.2	−8.5 ± 0.2	−8.5 ± 0.2
C1	C13	−10.6 ± 3.6	−8.4 ± 0.2	−8.5 ± 0.2	−8.4 ± 0.2
C1	C14	−4.7 ± 3.6	−5.8 ± 0.1	−5.9 ± 0.1	−5.8 ± 0.1
C2	C3	−1.6 ± 3.6	−1.4 ± 0.2	−1.4 ± 0.2	−1.3 ± 0.3
C2	C4	−0.1 ± 3.6	−0.2 ± 0.2	−0.1 ± 0.2	−0.2 ± 0.2
C2	C5	−5.4 ± 3.6	0.9 ± 0.3	0.8 ± 0.2	1.0 ± 0.2
C2	C6	−7.3 ± 3.6	−3.5 ± 0.3	−3.6 ± 0.2	−3.3 ± 0.2
C2	C7	−6.3 ± 3.6	−0.0 ± 0.3	−0.1 ± 0.2	0.0 ± 0.1
C2	C8	6.7 ± 3.6	7.5 ± 0.3	7.6 ± 0.2	7.7 ± 0.1
C2	C9	1.6 ± 3.6	3.2 ± 0.3	3.2 ± 0.2	3.4 ± 0.2
C2	C10	−2.0 ± 3.6	−1.2 ± 0.3	−1.3 ± 0.2	−1.0 ± 0.1
C2	C11	−1.0 ± 3.6	−0.1 ± 0.3	−0.2 ± 0.2	0.0 ± 0.2
C2	C12	−10.0 ± 3.6	−4.3 ± 0.3	−4.6 ± 0.2	−4.4 ± 0.2
C2	C13	−9.1 ± 3.6	−4.4 ± 0.3	−4.6 ± 0.2	−4.3 ± 0.1
C2	C14	−3.2 ± 3.6	−1.8 ± 0.3	−2.0 ± 0.2	−1.7 ± 0.1
C3	C4	1.5 ± 3.6	1.2 ± 0.2	1.4 ± 0.2	1.1 ± 0.1
C3	C5	−3.8 ± 3.6	2.3 ± 0.2	2.3 ± 0.2	2.3 ± 0.2
C3	C6	−5.7 ± 3.6	−2.0 ± 0.3	−2.1 ± 0.2	−1.9 ± 0.2
C3	C7	−4.6 ± 3.6	1.4 ± 0.3	1.3 ± 0.2	1.4 ± 0.3
C3	C8	8.3 ± 3.6	8.9 ± 0.3	9.0 ± 0.1	9.0 ± 0.3
C3	C9	3.3 ± 3.6	4.6 ± 0.2	4.7 ± 0.1	4.7 ± 0.2
C3	C10	−0.4 ± 3.6	0.2 ± 0.2	0.2 ± 0.1	0.3 ± 0.2
C3	C11	0.7 ± 3.6	1.3 ± 0.2	1.3 ± 0.2	1.3 ± 0.2
C3	C12	−8.4 ± 3.6	−2.9 ± 0.2	−3.2 ± 0.1	−3.1 ± 0.2
C3	C13	−7.4 ± 3.6	−3.0 ± 0.2	−3.1 ± 0.2	−3.0 ± 0.2
C3	C14	−1.5 ± 3.6	−0.4 ± 0.2	−0.6 ± 0.2	−0.4 ± 0.2
C4	C5	−5.4 ± 3.6	1.1 ± 0.1	0.9 ± 0.2	1.2 ± 0.1
C4	C6	−7.2 ± 3.6	−3.3 ± 0.2	−3.5 ± 0.3	−3.1 ± 0.2
C4	C7	−6.2 ± 3.6	0.2 ± 0.1	−0.0 ± 0.2	0.2 ± 0.3
C4	C8	6.7 ± 3.6	7.6 ± 0.2	7.6 ± 0.2	7.9 ± 0.2
C4	C9	1.7 ± 3.6	3.4 ± 0.1	3.3 ± 0.2	3.6 ± 0.2
C4	C10	−1.9 ± 3.6	−1.0 ± 0.1	−1.2 ± 0.2	−0.8 ± 0.2
C4	C11	−0.9 ± 3.6	0.1 ± 0.1	−0.1 ± 0.2	0.2 ± 0.2

C4	C12	−9.9 ± 3.6	−4.2 ± 0.1	−4.6 ± 0.2	−4.2 ± 0.2
C4	C13	−9.0 ± 3.6	−4.2 ± 0.2	−4.5 ± 0.2	−4.2 ± 0.2
C4	C14	−3.1 ± 3.6	−1.7 ± 0.1	−2.0 ± 0.2	−1.6 ± 0.2
C5	C6	−1.9 ± 3.6	−4.4 ± 0.2	−4.4 ± 0.2	−4.2 ± 0.1
C5	C7	−0.8 ± 3.6	−0.9 ± 0.2	−0.9 ± 0.1	−0.9 ± 0.1
C5	C8	12.1 ± 3.6	6.5 ± 0.1	6.7 ± 0.1	6.7 ± 0.1
C5	C9	7.1 ± 3.6	2.3 ± 0.1	2.4 ± 0.1	2.4 ± 0.1
C5	C10	3.4 ± 3.6	−2.1 ± 0.1	−2.1 ± 0.1	−2.0 ± 0.1
C5	C11	4.5 ± 3.6	−1.0 ± 0.1	−1.0 ± 0.1	−1.0 ± 0.1
C5	C12	−4.6 ± 3.6	−5.3 ± 0.1	−5.5 ± 0.1	−5.4 ± 0.2
C5	C13	−3.6 ± 3.6	−5.3 ± 0.1	−5.4 ± 0.1	−5.3 ± 0.1
C5	C14	2.3 ± 3.6	−2.8 ± 0.1	−2.9 ± 0.1	−2.7 ± 0.1
C6	C7	1.0 ± 3.6	3.4 ± 0.1	3.5 ± 0.2	3.3 ± 0.1
C6	C8	14.0 ± 3.6	10.9 ± 0.2	11.1 ± 0.2	11.0 ± 0.1
C6	C9	9.0 ± 3.6	6.7 ± 0.1	6.8 ± 0.2	6.7 ± 0.1
C6	C10	5.3 ± 3.6	2.3 ± 0.2	2.3 ± 0.1	2.3 ± 0.1
C6	C11	6.4 ± 3.6	3.4 ± 0.2	3.4 ± 0.2	3.3 ± 0.1
C6	C12	−2.7 ± 3.6	−0.9 ± 0.2	−1.1 ± 0.1	−1.1 ± 0.2
C6	C13	−1.8 ± 3.6	−0.9 ± 0.2	−1.0 ± 0.3	−1.1 ± 0.2
C6	C14	4.1 ± 3.6	1.6 ± 0.2	1.5 ± 0.2	1.5 ± 0.1
C7	C8	12.9 ± 3.6	7.5 ± 0.1	7.7 ± 0.1	7.7 ± 0.1
C7	C9	7.9 ± 3.6	3.2 ± 0.1	3.3 ± 0.1	3.4 ± 0.1
C7	C10	4.3 ± 3.6	−1.2 ± 0.1	−1.2 ± 0.1	−1.0 ± 0.1
C7	C11	5.3 ± 3.6	−0.1 ± 0.1	−0.1 ± 0.1	−0.0 ± 0.1
C7	C12	−3.7 ± 3.6	−4.3 ± 0.1	−4.5 ± 0.1	−4.4 ± 0.3
C7	C13	−2.8 ± 3.6	−4.4 ± 0.2	−4.5 ± 0.1	−4.4 ± 0.1
C7	C14	3.1 ± 3.6	−1.8 ± 0.1	−1.9 ± 0.1	−1.8 ± 0.1
C8	C9	−5.0 ± 3.6	−4.3 ± 0.1	−4.3 ± 0.1	−4.3 ± 0.1
C8	C10	−8.7 ± 3.6	−8.7 ± 0.1	−8.8 ± 0.1	−8.7 ± 0.1
C8	C11	−7.6 ± 3.6	−7.5 ± 0.1	−7.7 ± 0.1	−7.7 ± 0.1
C8	C12	−16.7 ± 3.6	−11.8 ± 0.1	−12.2 ± 0.2	−12.1 ± 0.3
C8	C13	−15.7 ± 3.6	−11.9 ± 0.1	−12.1 ± 0.1	−12.0 ± 0.1
C8	C14	−9.8 ± 3.6	−9.3 ± 0.1	−9.6 ± 0.1	−9.4 ± 0.1
C9	C10	−3.6 ± 3.6	−4.4 ± 0.1	−4.5 ± 0.1	−4.4 ± 0.1
C9	C11	−2.6 ± 3.6	−3.3 ± 0.0	−3.4 ± 0.1	−3.4 ± 0.1
C9	C12	−11.6 ± 3.6	−7.6 ± 0.1	−7.9 ± 0.1	−7.8 ± 0.2
C9	C13	−10.7 ± 3.6	−7.6 ± 0.1	−7.8 ± 0.1	−7.8 ± 0.1
C9	C14	−4.8 ± 3.6	−5.0 ± 0.1	−5.2 ± 0.1	−5.2 ± 0.1
C10	C11	1.0 ± 3.6	1.1 ± 0.1	1.1 ± 0.1	1.0 ± 0.1
C10	C12	−8.0 ± 3.6	−3.2 ± 0.1	−3.4 ± 0.1	−3.4 ± 0.2
C10	C13	−7.1 ± 3.6	−3.2 ± 0.1	−3.3 ± 0.2	−3.3 ± 0.1
C10	C14	−1.2 ± 3.6	−0.7 ± 0.1	−0.7 ± 0.1	−0.7 ± 0.1
C11	C12	−9.0 ± 3.6	−4.3 ± 0.1	−4.5 ± 0.2	−4.4 ± 0.2
C11	C13	−8.1 ± 3.6	−4.3 ± 0.1	−4.4 ± 0.1	−4.4 ± 0.2
C11	C14	−2.2 ± 3.6	−1.8 ± 0.0	−1.8 ± 0.0	−1.7 ± 0.1
C12	C13	0.9 ± 3.6	−0.0 ± 0.1	0.1 ± 0.2	0.0 ± 0.3
C12	C14	6.8 ± 3.6	2.5 ± 0.1	2.6 ± 0.2	2.6 ± 0.2
C13	C14	5.9 ± 3.6	2.6 ± 0.2	2.5 ± 0.1	2.6 ± 0.1
RMSE			3.38 ± 0.3	3.29 ± 0.3	3.35 ± 0.3
MAE			2.74 ± 0.3	2.66 ± 0.3	2.73 ± 0.3
r_{Spearman}			0.80	0.82	80
$t_{\text{preparation}}$			277.2 ns	277.2 ns	277.2 ns
$t_{\text{production}}$			59 ns	59 ns	59 ns

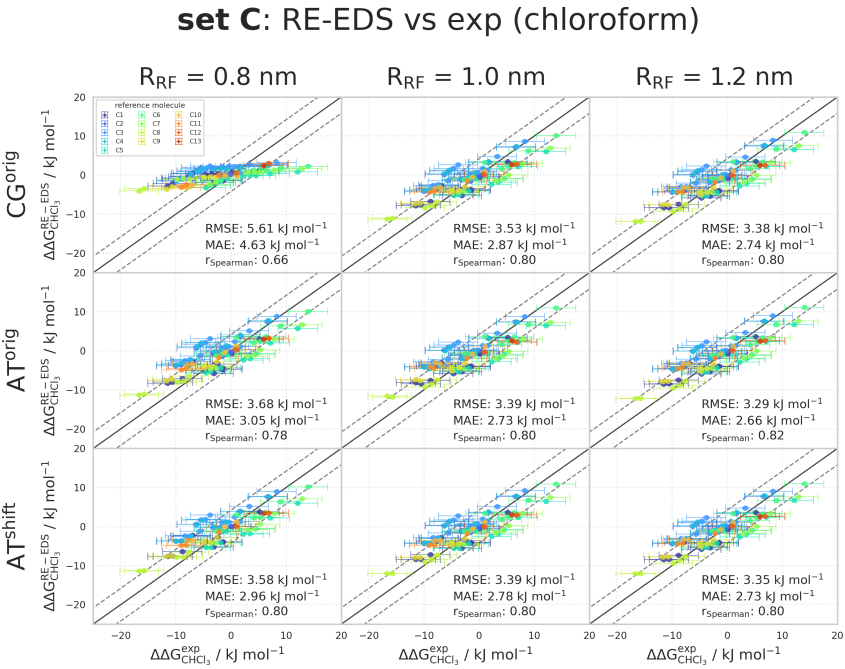


Figure 7.9: Comparison of the relative solvation free energies in chloroform of set C: $\Delta\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ obtained from RE-EDS calculations in GROMOS versus $\Delta\Delta G^{\text{exp}}_{\text{CHCl}_3}$ as reported in the Minnesota solvation[174] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to the RF schemes CG^{orig} (top), AT^{orig} (middle), and AT^{shift} (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.8 and 7.9.

Relative Hydration Free Energies of Set D

Table 7.10: $\Delta\Delta G_{\mathrm{hyd}}$ for the 13 molecules of set D calculated from RE-EDS calculations using the RF schemes $\mathrm{CG^{orig}}$, $\mathrm{AT^{orig}}$, and $\mathrm{AT^{shift}}$ with $R_{\mathrm{RF}} = 0.8$ nm and $R_{\mathrm{RF}} = 1.0$ nm. The results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

		$R_{\mathrm{RF}} = 0.8$ nm			$R_{\mathrm{RF}} = 1.0$ nm		
Molecules		$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,CG^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{shift}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,CG^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{orig}}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,AT^{shift}}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
D1	D2	11.6 ± 0.5	11.5 ± 0.4	11.7 ± 0.4	11.6 ± 0.3	11.7 ± 0.3	11.4 ± 0.5
D1	D3	−0.9 ± 0.5	0.3 ± 0.2	1.8 ± 0.3	−0.3 ± 0.3	0.3 ± 0.3	0.7 ± 0.4
D1	D4	−0.5 ± 0.4	0.1 ± 0.1	1.4 ± 0.3	−0.4 ± 0.3	−0.2 ± 0.3	0.2 ± 0.4
D1	D5	7.0 ± 0.4	7.0 ± 0.1	7.8 ± 0.3	7.0 ± 0.3	7.0 ± 0.4	7.2 ± 0.4
D1	D6	7.9 ± 0.6	8.3 ± 0.1	9.3 ± 0.2	8.3 ± 0.2	8.4 ± 0.4	8.6 ± 0.7
D1	D7	7.5 ± 0.7	7.5 ± 0.3	8.9 ± 0.3	7.7 ± 0.4	7.8 ± 0.1	8.3 ± 0.6
D1	D8	8.3 ± 0.5	7.3 ± 0.2	7.9 ± 0.3	7.6 ± 0.4	7.2 ± 0.3	7.3 ± 0.4
D1	D9	−21.6 ± 0.4	−20.3 ± 0.3	−19.0 ± 0.2	−21.2 ± 0.4	−20.5 ± 0.4	−20.1 ± 0.5
D1	D10	−2.0 ± 0.8	−2.6 ± 0.3	−1.6 ± 0.6	−2.9 ± 0.6	−3.0 ± 0.4	−2.8 ± 0.5
D1	D11	−2.0 ± 0.5	−2.3 ± 0.3	−1.3 ± 0.5	−2.2 ± 0.3	−2.2 ± 0.5	−2.2 ± 0.5
D1	D12	0.8 ± 0.4	1.3 ± 0.3	2.4 ± 0.3	0.7 ± 0.3	0.7 ± 0.5	0.8 ± 0.4
D1	D13	13.4 ± 0.6	12.8 ± 0.4	12.8 ± 0.3	12.8 ± 0.5	12.6 ± 0.5	12.3 ± 0.7
D2	D3	−12.5 ± 0.1	−11.2 ± 0.3	−9.9 ± 0.2	−11.8 ± 0.1	−11.4 ± 0.1	−10.7 ± 0.2
D2	D4	−12.1 ± 0.2	−11.5 ± 0.4	−10.3 ± 0.2	−12.0 ± 0.2	−11.9 ± 0.2	−11.2 ± 0.3
D2	D5	−4.6 ± 0.1	−4.6 ± 0.3	−3.9 ± 0.3	−4.6 ± 0.2	−4.6 ± 0.1	−4.2 ± 0.2
D2	D6	−3.7 ± 0.3	−3.2 ± 0.3	−2.4 ± 0.2	−3.2 ± 0.3	−3.3 ± 0.2	−2.8 ± 0.3
D2	D7	−4.1 ± 0.3	−4.0 ± 0.2	−2.8 ± 0.2	−3.9 ± 0.3	−3.9 ± 0.3	−3.2 ± 0.3
D2	D8	−3.3 ± 0.1	−4.3 ± 0.3	−3.8 ± 0.3	−4.0 ± 0.2	−4.4 ± 0.1	−4.2 ± 0.2
D2	D9	−33.1 ± 0.2	−31.8 ± 0.3	−30.7 ± 0.4	−32.8 ± 0.3	−32.2 ± 0.2	−31.6 ± 0.2
D2	D10	−13.6 ± 0.6	−14.1 ± 0.3	−13.3 ± 0.3	−14.4 ± 0.5	−14.7 ± 0.5	−14.3 ± 0.3
D2	D11	−13.6 ± 0.2	−13.9 ± 0.4	−13.0 ± 0.4	−13.8 ± 0.3	−13.8 ± 0.2	−13.6 ± 0.4
D2	D12	−10.7 ± 0.2	−10.3 ± 0.4	−9.3 ± 0.2	−10.9 ± 0.2	−10.9 ± 0.3	−10.6 ± 0.3
D2	D13	1.8 ± 0.4	1.3 ± 0.5	1.1 ± 0.4	1.2 ± 0.5	0.9 ± 0.4	0.9 ± 0.4
D3	D4	0.4 ± 0.2	−0.3 ± 0.2	−0.4 ± 0.1	−0.1 ± 0.2	−0.4 ± 0.2	−0.5 ± 0.2
D3	D5	7.9 ± 0.1	6.6 ± 0.2	6.0 ± 0.2	7.3 ± 0.2	6.8 ± 0.1	6.5 ± 0.1
D3	D6	8.8 ± 0.2	8.0 ± 0.1	7.5 ± 0.1	8.6 ± 0.3	8.1 ± 0.2	7.9 ± 0.3
D3	D7	8.4 ± 0.3	7.2 ± 0.2	7.1 ± 0.2	8.0 ± 0.2	7.5 ± 0.3	7.6 ± 0.3
D3	D8	9.2 ± 0.2	6.9 ± 0.2	6.1 ± 0.2	7.9 ± 0.1	7.0 ± 0.1	6.6 ± 0.1
D3	D9	−20.7 ± 0.2	−20.7 ± 0.3	−20.8 ± 0.3	−20.9 ± 0.3	−20.8 ± 0.1	−20.8 ± 0.2
D3	D10	−1.1 ± 0.6	−2.9 ± 0.3	−3.4 ± 0.4	−2.6 ± 0.4	−3.3 ± 0.5	−3.5 ± 0.2
D3	D11	−1.1 ± 0.2	−2.7 ± 0.2	−3.1 ± 0.3	−1.9 ± 0.3	−2.4 ± 0.2	−2.9 ± 0.2
D3	D12	1.8 ± 0.2	0.9 ± 0.2	0.6 ± 0.2	1.0 ± 0.2	0.5 ± 0.3	0.1 ± 0.2
D3	D13	14.3 ± 0.5	12.5 ± 0.4	11.0 ± 0.3	13.0 ± 0.6	12.3 ± 0.4	11.6 ± 0.4
D4	D5	7.5 ± 0.2	6.9 ± 0.2	6.4 ± 0.2	7.4 ± 0.2	7.2 ± 0.1	7.0 ± 0.2
D4	D6	8.4 ± 0.3	8.3 ± 0.1	7.9 ± 0.1	8.7 ± 0.2	8.6 ± 0.2	8.6 ± 0.4
D4	D7	8.0 ± 0.3	7.5 ± 0.3	7.5 ± 0.2	8.1 ± 0.2	7.9 ± 0.3	8.0 ± 0.4
D4	D8	8.8 ± 0.2	7.2 ± 0.2	6.5 ± 0.2	8.0 ± 0.3	7.4 ± 0.2	7.1 ± 0.1
D4	D9	−21.1 ± 0.1	−20.4 ± 0.3	−20.4 ± 0.4	−20.8 ± 0.3	−20.3 ± 0.2	−20.3 ± 0.2
D4	D10	−1.5 ± 0.5	−2.6 ± 0.3	−3.1 ± 0.4	−2.5 ± 0.6	−2.8 ± 0.6	−3.0 ± 0.2
D4	D11	−1.5 ± 0.3	−2.4 ± 0.2	−2.7 ± 0.3	−1.8 ± 0.3	−2.0 ± 0.3	−2.4 ± 0.3
D4	D12	1.3 ± 0.1	1.2 ± 0.2	1.0 ± 0.2	1.1 ± 0.1	0.9 ± 0.3	0.6 ± 0.2
D4	D13	13.9 ± 0.5	12.8 ± 0.3	11.4 ± 0.3	13.2 ± 0.5	12.8 ± 0.4	12.1 ± 0.5
D5	D6	0.9 ± 0.2	1.4 ± 0.1	1.5 ± 0.2	1.3 ± 0.1	1.4 ± 0.1	1.4 ± 0.3
D5	D7	0.5 ± 0.3	0.6 ± 0.2	1.1 ± 0.2	0.7 ± 0.2	0.7 ± 0.3	1.0 ± 0.4
D5	D8	1.3 ± 0.2	0.3 ± 0.1	0.1 ± 0.2	0.6 ± 0.2	0.2 ± 0.1	0.1 ± 0.1
D5	D9	−28.6 ± 0.2	−27.3 ± 0.3	−26.8 ± 0.2	−28.2 ± 0.2	−27.6 ± 0.1	−27.3 ± 0.2
D5	D10	−9.0 ± 0.6	−9.5 ± 0.3	−9.4 ± 0.5	−9.9 ± 0.6	−10.0 ± 0.6	−10.0 ± 0.1
D5	D11	−9.0 ± 0.1	−9.3 ± 0.3	−9.1 ± 0.3	−9.2 ± 0.3	−9.2 ± 0.2	−9.4 ± 0.3
D5	D12	−6.1 ± 0.2	−5.7 ± 0.3	−5.4 ± 0.1	−6.3 ± 0.1	−6.3 ± 0.3	−6.4 ± 0.1
D5	D13	6.4 ± 0.5	5.9 ± 0.4	5.0 ± 0.3	5.8 ± 0.5	5.6 ± 0.4	5.1 ± 0.4
D6	D7	−0.4 ± 0.4	−0.8 ± 0.2	−0.4 ± 0.2	−0.6 ± 0.2	−0.6 ± 0.4	−0.4 ± 0.4
D6	D8	0.4 ± 0.3	−1.1 ± 0.2	−1.5 ± 0.2	−0.7 ± 0.4	−1.1 ± 0.1	−1.3 ± 0.3
D6	D9	−29.4 ± 0.3	−28.6 ± 0.3	−28.3 ± 0.3	−29.5 ± 0.3	−28.9 ± 0.1	−28.7 ± 0.3
D6	D10	−9.9 ± 0.6	−10.9 ± 0.2	−11.0 ± 0.4	−11.2 ± 0.6	−11.4 ± 0.6	−11.4 ± 0.3
D6	D11	−9.9 ± 0.2	−10.7 ± 0.2	−10.6 ± 0.3	−10.5 ± 0.2	−10.6 ± 0.2	−10.8 ± 0.4
D6	D12	−7.0 ± 0.3	−7.1 ± 0.2	−6.9 ± 0.2	−7.6 ± 0.2	−7.6 ± 0.2	−7.8 ± 0.4
D6	D13	5.5 ± 0.6	4.5 ± 0.4	3.5 ± 0.3	4.4 ± 0.5	4.2 ± 0.4	3.7 ± 0.3
D7	D8	0.8 ± 0.2	−0.3 ± 0.2	−1.1 ± 0.3	−0.1 ± 0.3	−0.5 ± 0.3	−1.0 ± 0.3
D7	D9	−29.1 ± 0.3	−27.9 ± 0.3	−27.9 ± 0.3	−28.9 ± 0.2	−28.3 ± 0.4	−28.4 ± 0.3
D7	D10	−9.5 ± 0.6	−10.1 ± 0.3	−10.6 ± 0.3	−10.6 ± 0.7	−10.7 ± 0.4	−11.1 ± 0.5
D7	D11	−9.5 ± 0.3	−9.9 ± 0.3	−10.2 ± 0.4	−9.9 ± 0.3	−9.9 ± 0.4	−10.4 ± 0.5
D7	D12	−6.6 ± 0.3	−6.3 ± 0.4	−6.5 ± 0.1	−7.0 ± 0.1	−7.0 ± 0.5	−7.4 ± 0.5
D7	D13	5.9 ± 0.5	5.3 ± 0.5	3.9 ± 0.3	5.0 ± 0.5	4.8 ± 0.5	4.1 ± 0.6
D8	D9	−29.8 ± 0.1	−27.6 ± 0.2	−26.9 ± 0.3	−28.8 ± 0.4	−27.8 ± 0.2	−27.4 ± 0.2
D8	D10	−10.2 ± 0.6	−9.8 ± 0.1	−9.5 ± 0.5	−10.5 ± 0.4	−10.2 ± 0.6	−10.1 ± 0.2
D8	D11	−10.3 ± 0.2	−9.6 ± 0.2	−9.2 ± 0.2	−9.8 ± 0.3	−9.4 ± 0.2	−9.5 ± 0.2
D8	D12	−7.4 ± 0.2	−6.0 ± 0.3	−5.4 ± 0.2	−6.9 ± 0.2	−6.5 ± 0.2	−6.5 ± 0.2
D8	D13	5.1 ± 0.4	5.6 ± 0.4	5.0 ± 0.3	5.2 ± 0.6	5.3 ± 0.4	5.0 ± 0.4
D9	D10	19.6 ± 0.6	17.7 ± 0.1	17.4 ± 0.6	18.3 ± 0.7	17.5 ± 0.5	17.3 ± 0.2
D9	D11	19.5 ± 0.2	18.0 ± 0.2	17.7 ± 0.4	19.0 ± 0.4	18.3 ± 0.2	18.0 ± 0.4
D9	D12	22.4 ± 0.1	21.6 ± 0.2	21.4 ± 0.3	21.9 ± 0.3	21.3 ± 0.3	20.9 ± 0.2
D9	D13	34.9 ± 0.4	33.1 ± 0.5	31.8 ± 0.4	34.0 ± 0.6	33.1 ± 0.4	32.4 ± 0.4

D10	D11	−0.0	±	0.6	0.2	±	0.2	0.3	±	0.5	0.7	±	0.6	0.8	±	0.6	0.6	±	0.3
D10	D12	2.8	±	0.6	3.8	±	0.3	4.1	±	0.3	3.6	±	0.6	3.7	±	0.7	3.6	±	0.1
D10	D13	15.4	±	0.7	15.4	±	0.5	14.5	±	0.5	15.6	±	0.8	15.6	±	0.7	15.1	±	0.4
D11	D12	2.9	±	0.3	3.6	±	0.2	3.7	±	0.3	2.9	±	0.2	2.9	±	0.2	3.0	±	0.3
D11	D13	15.4	±	0.5	15.2	±	0.5	14.1	±	0.4	15.0	±	0.5	14.8	±	0.4	14.5	±	0.5
D12	D13	12.5	±	0.4	11.6	±	0.4	10.4	±	0.4	12.1	±	0.5	11.8	±	0.5	11.5	±	0.4
RMSE		6.37	±	0.6	5.82	±	0.6	5.64	±	0.7	6.25	±	0.6	6.01	±	0.6	5.99	±	0.6
MAE		5.35	±	0.5	4.88	±	0.5	4.70	±	0.6	5.25	±	0.5	5.03	±	0.5	5.00	±	0.6
r_{Spearman}		0.78			0.80			0.81			0.79			0.80			0.80		
$t_{\text{preparation}}$		281 ns			281 ns			281 ns			286.2 ns			281 ns			281 ns		
$t_{\text{production}}$		66 ns			66 ns			66 ns			68 ns			66 ns			66 ns		

Table 7.11: $\Delta\Delta G_{\text{hyd}}$ for the 13 molecules of set D from experiment,[155,173] calculated from the hydration free energies obtained with MBAR,[155,173] and from RE-EDS calculations using the RF schemes CG^{orig}, AT^{orig}, and AT^{shift} with $R_{\text{RF}} = 1.2$ nm. The RE-EDS results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ [155,173]	$\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$ [155,173]	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,CG}^{\text{orig}}}$	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,AT}^{\text{orig}}}$	$\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,AT}^{\text{shift}}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
D1	D2	14.4 ± 2.6	11.5 ± 0.2	11.8 ± 0.2	11.7 ± 0.4	11.6 ± 0.3
D1	D3	7.9 ± 2.6	0.1 ± 0.2	0.3 ± 0.1	0.2 ± 0.4	0.7 ± 0.2
D1	D4	5.5 ± 2.6	0.0 ± 0.2	−0.1 ± 0.2	−0.2 ± 0.4	0.1 ± 0.3
D1	D5	5.3 ± 2.6	6.6 ± 0.2	7.4 ± 0.2	7.2 ± 0.3	7.5 ± 0.2
D1	D6	4.7 ± 2.6	7.8 ± 0.2	8.6 ± 0.2	8.8 ± 0.3	8.8 ± 0.4
D1	D7	4.1 ± 2.6	7.3 ± 0.2	8.3 ± 0.2	7.9 ± 0.5	8.1 ± 0.3
D1	D8	5.5 ± 2.6	6.9 ± 0.2	7.6 ± 0.2	7.3 ± 0.4	7.5 ± 0.3
D1	D9	−12.3 ± 2.6	−20.8 ± 0.2	−20.6 ± 0.3	−20.6 ± 0.5	−20.3 ± 0.4
D1	D10	0.2 ± 3.6	−3.1 ± 0.2	−3.1 ± 0.4	−2.9 ± 0.4	−2.8 ± 0.6
D1	D11	1.9 ± 2.6	−3.1 ± 0.2	−2.1 ± 0.4	−2.1 ± 0.5	−2.0 ± 0.4
D1	D12	8.3 ± 2.6	0.7 ± 0.2	0.9 ± 0.1	0.7 ± 0.5	0.9 ± 0.2
D1	D13	15.4 ± 2.6	11.8 ± 0.2	12.3 ± 0.5	12.3 ± 0.6	12.5 ± 0.8
D2	D3	−6.6 ± 3.6	−11.5 ± 0.2	−11.5 ± 0.2	−11.4 ± 0.1	−11.0 ± 0.2
D2	D4	−8.9 ± 3.6	−11.5 ± 0.2	−11.9 ± 0.2	−11.9 ± 0.2	−11.5 ± 0.3
D2	D5	−9.1 ± 3.6	−4.9 ± 0.2	−4.4 ± 0.2	−4.5 ± 0.2	−4.2 ± 0.2
D2	D6	−9.7 ± 3.6	−3.8 ± 0.2	−3.2 ± 0.2	−2.9 ± 0.4	−2.9 ± 0.2
D2	D7	−10.4 ± 3.6	−4.3 ± 0.2	−3.5 ± 0.2	−3.7 ± 0.2	−3.6 ± 0.3
D2	D8	−9.0 ± 3.6	−4.7 ± 0.2	−4.2 ± 0.2	−4.3 ± 0.1	−4.1 ± 0.2
D2	D9	−26.7 ± 3.6	−32.3 ± 0.2	−32.4 ± 0.3	−32.3 ± 0.2	−32.0 ± 0.2
D2	D10	−14.2 ± 4.3	−14.7 ± 0.2	−14.9 ± 0.5	−14.6 ± 0.4	−14.4 ± 0.6
D2	D11	−12.6 ± 3.6	−14.6 ± 0.2	−13.9 ± 0.3	−13.7 ± 0.2	−13.6 ± 0.4
D2	D12	−6.2 ± 3.6	−10.9 ± 0.2	−10.9 ± 0.1	−10.9 ± 0.1	−10.7 ± 0.2
D2	D13	1.0 ± 3.6	0.3 ± 0.2	0.5 ± 0.4	0.7 ± 0.4	0.8 ± 0.8
D3	D4	−2.3 ± 3.6	−0.1 ± 0.2	−0.3 ± 0.2	−0.4 ± 0.2	−0.5 ± 0.1
D3	D5	−2.6 ± 3.6	6.5 ± 0.2	7.1 ± 0.2	7.0 ± 0.1	6.8 ± 0.1
D3	D6	−3.1 ± 3.6	7.7 ± 0.2	8.4 ± 0.2	8.6 ± 0.4	8.1 ± 0.2
D3	D7	−3.8 ± 3.6	7.2 ± 0.2	8.0 ± 0.2	7.7 ± 0.2	7.4 ± 0.3
D3	D8	−2.4 ± 3.6	6.8 ± 0.2	7.3 ± 0.2	7.1 ± 0.2	6.9 ± 0.1
D3	D9	−20.1 ± 3.6	−20.9 ± 0.2	−20.9 ± 0.2	−20.8 ± 0.2	−21.0 ± 0.2
D3	D10	−7.7 ± 4.3	−3.2 ± 0.2	−3.1 ± 0.4	−3.1 ± 0.4	−3.6 ± 0.5
D3	D11	−6.0 ± 3.6	−3.1 ± 0.2	−2.4 ± 0.4	−2.3 ± 0.2	−2.6 ± 0.3
D3	D12	0.4 ± 3.6	0.6 ± 0.2	0.6 ± 0.1	0.5 ± 0.2	0.3 ± 0.0
D3	D13	7.5 ± 3.6	11.8 ± 0.2	12.0 ± 0.4	12.1 ± 0.5	11.8 ± 0.8
D4	D5	−0.2 ± 3.6	6.6 ± 0.2	7.4 ± 0.1	7.4 ± 0.2	7.3 ± 0.2
D4	D6	−0.8 ± 3.6	7.8 ± 0.2	8.7 ± 0.3	9.0 ± 0.3	8.7 ± 0.3
D4	D7	−1.5 ± 3.6	7.3 ± 0.2	8.3 ± 0.1	8.2 ± 0.2	7.9 ± 0.3
D4	D8	−0.0 ± 3.6	6.9 ± 0.2	7.7 ± 0.2	7.5 ± 0.2	7.4 ± 0.2
D4	D9	−17.8 ± 3.6	−20.8 ± 0.2	−20.5 ± 0.3	−20.4 ± 0.3	−20.5 ± 0.2
D4	D10	−5.3 ± 4.3	−3.1 ± 0.2	−3.0 ± 0.5	−2.7 ± 0.4	−2.9 ± 0.5
D4	D11	−3.6 ± 3.6	−3.1 ± 0.2	−2.0 ± 0.3	−1.8 ± 0.2	−2.1 ± 0.4
D4	D12	2.8 ± 3.6	0.7 ± 0.2	0.9 ± 0.2	1.0 ± 0.2	0.8 ± 0.2
D4	D13	9.9 ± 3.6	11.8 ± 0.2	12.3 ± 0.5	12.5 ± 0.4	12.3 ± 0.8
D5	D6	−0.6 ± 3.6	1.2 ± 0.2	1.3 ± 0.2	1.6 ± 0.3	1.3 ± 0.2
D5	D7	−1.3 ± 3.6	0.7 ± 0.2	0.9 ± 0.0	0.8 ± 0.2	0.6 ± 0.2
D5	D8	0.2 ± 3.6	0.3 ± 0.2	0.3 ± 0.3	0.1 ± 0.2	0.1 ± 0.1
D5	D9	−17.6 ± 3.6	−27.4 ± 0.2	−27.9 ± 0.4	−27.8 ± 0.2	−27.8 ± 0.2
D5	D10	−5.1 ± 4.3	−9.7 ± 0.2	−10.4 ± 0.5	−10.1 ± 0.3	−10.3 ± 0.5
D5	D11	−3.4 ± 3.6	−9.7 ± 0.2	−9.5 ± 0.3	−9.2 ± 0.2	−9.5 ± 0.3
D5	D12	3.0 ± 3.6	−5.9 ± 0.2	−6.5 ± 0.2	−6.4 ± 0.2	−6.5 ± 0.1
D5	D13	10.1 ± 3.6	5.2 ± 0.2	4.9 ± 0.5	5.1 ± 0.5	5.0 ± 0.8
D6	D7	−0.7 ± 3.6	−0.5 ± 0.2	−0.4 ± 0.2	−0.8 ± 0.3	−0.7 ± 0.3
D6	D8	0.8 ± 3.6	−0.9 ± 0.2	−1.0 ± 0.3	−1.4 ± 0.4	−1.3 ± 0.1
D6	D9	−17.0 ± 3.6	−28.6 ± 0.2	−29.2 ± 0.3	−29.4 ± 0.5	−29.1 ± 0.2
D6	D10	−4.5 ± 4.3	−10.9 ± 0.2	−11.7 ± 0.4	−11.7 ± 0.5	−11.6 ± 0.6
D6	D11	−2.8 ± 3.6	−10.8 ± 0.2	−10.7 ± 0.5	−10.8 ± 0.4	−10.8 ± 0.3
D6	D12	3.6 ± 3.6	−7.1 ± 0.2	−7.8 ± 0.2	−8.0 ± 0.4	−7.9 ± 0.3
D6	D13	10.7 ± 3.6	4.1 ± 0.2	3.6 ± 0.4	3.5 ± 0.5	3.7 ± 0.8
D7	D8	1.4 ± 3.6	−0.4 ± 0.2	−0.7 ± 0.2	−0.6 ± 0.2	−0.5 ± 0.2
D7	D9	−16.3 ± 3.6	−28.1 ± 0.2	−28.9 ± 0.3	−28.6 ± 0.3	−28.4 ± 0.4
D7	D10	−3.8 ± 4.3	−10.4 ± 0.2	−11.3 ± 0.5	−10.9 ± 0.5	−10.9 ± 0.3
D7	D11	−2.2 ± 3.6	−10.3 ± 0.2	−10.4 ± 0.3	−10.0 ± 0.3	−10.1 ± 0.2

D7	D12	4.2	±	3.6	−6.6	±	0.2	−7.4	±	0.1	−7.2	±	0.2	−7.1	±	0.3
D7	D13	11.3	±	3.6	4.6	±	0.2	4.0	±	0.5	4.4	±	0.5	4.4	±	0.9
D8	D9	−17.7	±	3.6	−27.7	±	0.2	−28.2	±	0.3	−27.9	±	0.1	−27.9	±	0.2
D8	D10	−5.3	±	4.3	−10.0	±	0.2	−10.7	±	0.6	−10.2	±	0.3	−10.3	±	0.5
D8	D11	−3.6	±	3.6	−9.9	±	0.2	−9.7	±	0.3	−9.4	±	0.3	−9.5	±	0.3
D8	D12	2.8	±	3.6	−6.2	±	0.2	−6.8	±	0.1	−6.6	±	0.1	−6.6	±	0.1
D8	D13	9.9	±	3.6	5.0	±	0.2	4.6	±	0.5	5.0	±	0.5	4.9	±	0.7
D9	D10	12.5	±	4.3	17.7	±	0.2	17.5	±	0.4	17.7	±	0.4	17.5	±	0.6
D9	D11	14.1	±	3.6	17.7	±	0.2	18.5	±	0.5	18.6	±	0.4	18.3	±	0.4
D9	D12	20.5	±	3.6	21.5	±	0.2	21.4	±	0.3	21.4	±	0.2	21.3	±	0.2
D9	D13	27.7	±	3.6	32.6	±	0.2	32.8	±	0.3	32.9	±	0.6	32.8	±	0.7
D10	D11	1.7	±	4.3	0.1	±	0.2	1.0	±	0.8	0.9	±	0.4	0.8	±	0.3
D10	D12	8.1	±	4.3	3.8	±	0.2	3.9	±	0.4	3.7	±	0.4	3.7	±	0.5
D10	D13	15.2	±	4.3	15.0	±	0.2	15.3	±	0.6	15.2	±	0.6	15.3	±	1.1
D11	D12	6.4	±	3.6	3.7	±	0.2	2.9	±	0.3	2.8	±	0.3	2.9	±	0.3
D11	D13	13.5	±	3.6	14.9	±	0.2	14.4	±	0.6	14.4	±	0.4	14.5	±	0.9
D12	D13	7.1	±	3.6	11.2	±	0.2	11.4	±	0.4	11.6	±	0.5	11.5	±	0.7
RMSE					5.86	±	0.6	6.20	±	0.7	6.15	±	0.6	6.06	±	0.7
MAE					4.92	±	0.5	5.19	±	0.6	5.15	±	0.6	5.06	±	0.6
^{r}Spearman							0.80			0.80			0.80			0.81
^{t}preparation							39 ns			286.2 ns			281 ns			281 ns
^{t}production							1300 ns			68 ns			66 ns			66 ns

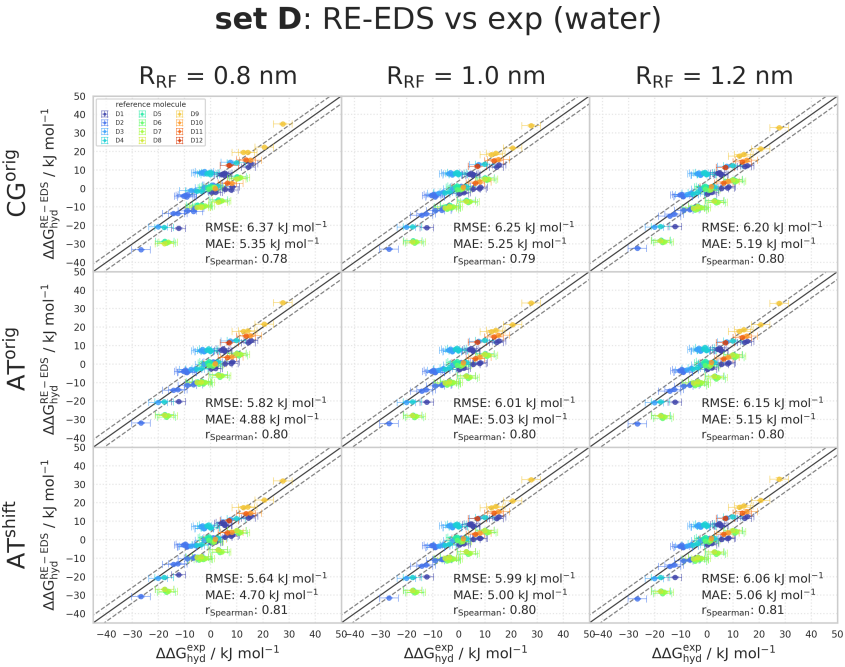


Figure 7.10: Comparison of the relative hydration free energies of set D: $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ obtained from RE-EDS calculations in GROMOS versus $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ as reported in the FreeSolv [155,173] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to the RF schemes CG^{orig} (top), AT^{orig} (middle), and AT^{shift} (bottom), respectively. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.10 and 7.11.

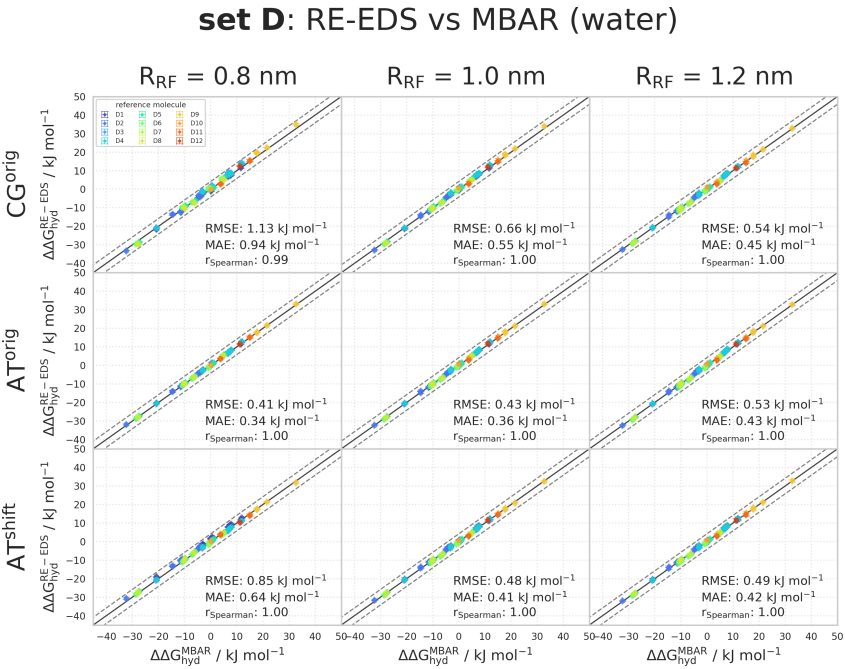


Figure 7.11: Comparison of the relative hydration free energies of set D: $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ obtained from RE-EDS calculations in GROMOS[27] versus $\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to the RF schemes CG^{orig} (top), AT^{orig} (middle), and AT^{shift} (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.10 and 7.11.

Relative Solvation Free Energies in Chloroform of set D

Table 7.12: $\Delta\Delta G_{\mathrm{CHCl_3}}$ for the 13 molecules of set D calculated from RE-EDS calculations using the RF schemes $\mathrm{CG^{orig}}$, $\mathrm{AT^{orig}}$, and $\mathrm{AT^{shift}}$ with $R_{\mathrm{RF}} = 0.8$ nm and $R_{\mathrm{RF}} = 1.0$ nm. The results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

		$R_{\mathrm{RF}} = 0.8$ nm			$R_{\mathrm{RF}} = 1.0$ nm		
Molecules		$\Delta\Delta G^{\mathrm{RE\text{-}EDS,CG^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{shift}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,CG^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{orig}}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,AT^{shift}}}_{\mathrm{CHCl_3}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
D1	D2	1.9 ± 0.4	1.3 ± 0.1	1.0 ± 0.3	1.2 ± 0.2	1.1 ± 0.2	1.1 ± 0.2
D1	D3	−0.4 ± 0.3	−0.6 ± 0.2	0.2 ± 0.1	−0.4 ± 0.1	−0.6 ± 0.1	−0.2 ± 0.1
D1	D4	−5.1 ± 0.5	−5.7 ± 0.2	−4.5 ± 0.2	−5.1 ± 0.2	−5.6 ± 0.2	−5.1 ± 0.2
D1	D5	1.1 ± 0.4	2.9 ± 0.1	3.4 ± 0.1	3.0 ± 0.1	3.1 ± 0.1	3.1 ± 0.1
D1	D6	1.1 ± 0.3	2.7 ± 0.1	3.3 ± 0.1	3.0 ± 0.3	3.0 ± 0.2	3.3 ± 0.3
D1	D7	1.6 ± 0.7	2.9 ± 0.2	3.1 ± 0.2	2.7 ± 0.3	2.7 ± 0.3	3.2 ± 0.2
D1	D8	−2.9 ± 0.6	−1.1 ± 0.2	−0.5 ± 0.2	−0.6 ± 0.1	−0.6 ± 0.3	−0.6 ± 0.2
D1	D9	−2.9 ± 0.4	−6.3 ± 0.2	−5.8 ± 0.2	−6.3 ± 0.1	−6.9 ± 0.1	−6.6 ± 0.1
D1	D10	0.9 ± 0.3	−2.3 ± 0.2	−2.3 ± 0.2	−2.7 ± 0.2	−3.2 ± 0.2	−2.8 ± 0.1
D1	D11	−2.0 ± 0.4	−3.4 ± 0.3	−2.6 ± 0.2	−2.9 ± 0.2	−3.3 ± 0.2	−3.3 ± 0.2
D1	D12	−6.8 ± 0.5	−8.6 ± 0.3	−7.9 ± 0.2	−8.2 ± 0.3	−9.1 ± 0.1	−8.8 ± 0.2
D1	D13	−2.6 ± 0.6	−2.6 ± 0.3	−2.6 ± 0.3	−2.0 ± 0.5	−2.5 ± 0.4	−2.6 ± 0.3
D2	D3	−2.3 ± 0.2	−1.9 ± 0.1	−0.8 ± 0.3	−1.6 ± 0.1	−1.7 ± 0.1	−1.3 ± 0.2
D2	D4	−6.9 ± 0.4	−7.0 ± 0.2	−5.5 ± 0.2	−6.3 ± 0.1	−6.7 ± 0.3	−6.2 ± 0.2
D2	D5	−0.8 ± 0.2	1.6 ± 0.1	2.4 ± 0.2	1.8 ± 0.2	1.9 ± 0.1	2.0 ± 0.2
D2	D6	−0.8 ± 0.1	1.5 ± 0.2	2.3 ± 0.4	1.8 ± 0.3	1.9 ± 0.1	2.2 ± 0.4
D2	D7	−0.3 ± 0.4	1.6 ± 0.1	2.1 ± 0.3	1.5 ± 0.3	1.6 ± 0.2	2.1 ± 0.2
D2	D8	−4.8 ± 0.5	−2.4 ± 0.3	−1.5 ± 0.2	−1.7 ± 0.2	−1.7 ± 0.2	−1.7 ± 0.2
D2	D9	−4.8 ± 0.3	−7.6 ± 0.2	−6.8 ± 0.3	−7.5 ± 0.2	−8.0 ± 0.2	−7.7 ± 0.1
D2	D10	−1.0 ± 0.2	−3.6 ± 0.2	−3.3 ± 0.3	−3.9 ± 0.2	−4.4 ± 0.2	−3.9 ± 0.2
D2	D11	−3.9 ± 0.2	−4.7 ± 0.2	−3.5 ± 0.4	−4.1 ± 0.2	−4.4 ± 0.2	−4.4 ± 0.3
D2	D12	−8.6 ± 0.4	−9.9 ± 0.2	−8.9 ± 0.4	−9.4 ± 0.2	−10.2 ± 0.2	−9.8 ± 0.2
D2	D13	−4.4 ± 0.5	−3.9 ± 0.3	−3.6 ± 0.5	−3.2 ± 0.4	−3.6 ± 0.4	−3.7 ± 0.3
D3	D4	−4.6 ± 0.3	−5.1 ± 0.1	−4.7 ± 0.1	−4.7 ± 0.2	−5.0 ± 0.3	−4.9 ± 0.2
D3	D5	1.5 ± 0.2	3.5 ± 0.1	3.1 ± 0.1	3.4 ± 0.1	3.7 ± 0.1	3.3 ± 0.1
D3	D6	1.5 ± 0.2	3.3 ± 0.2	3.1 ± 0.2	3.4 ± 0.2	3.6 ± 0.1	3.5 ± 0.3
D3	D7	2.0 ± 0.4	3.5 ± 0.2	2.9 ± 0.1	3.1 ± 0.2	3.3 ± 0.2	3.4 ± 0.2
D3	D8	−2.5 ± 0.5	−0.5 ± 0.2	−0.7 ± 0.2	−0.1 ± 0.1	0.0 ± 0.2	−0.4 ± 0.2
D3	D9	−2.5 ± 0.2	−5.7 ± 0.1	−6.0 ± 0.1	−5.9 ± 0.1	−6.3 ± 0.2	−6.4 ± 0.1
D3	D10	1.3 ± 0.2	−1.7 ± 0.2	−2.5 ± 0.1	−2.3 ± 0.1	−2.6 ± 0.2	−2.6 ± 0.1
D3	D11	−1.6 ± 0.2	−2.8 ± 0.2	−2.8 ± 0.2	−2.5 ± 0.2	−2.7 ± 0.2	−3.1 ± 0.3
D3	D12	−6.3 ± 0.3	−8.0 ± 0.1	−8.1 ± 0.2	−7.8 ± 0.2	−8.5 ± 0.1	−8.6 ± 0.1
D3	D13	−2.1 ± 0.6	−2.0 ± 0.3	−2.8 ± 0.3	−1.6 ± 0.5	−1.9 ± 0.4	−2.4 ± 0.3
D4	D5	6.1 ± 0.4	8.6 ± 0.2	7.8 ± 0.1	8.1 ± 0.2	8.7 ± 0.3	8.2 ± 0.2
D4	D6	6.2 ± 0.4	8.4 ± 0.2	7.8 ± 0.3	8.1 ± 0.2	8.6 ± 0.3	8.4 ± 0.3
D4	D7	6.7 ± 0.5	8.6 ± 0.2	7.6 ± 0.2	7.8 ± 0.3	8.3 ± 0.3	8.3 ± 0.2
D4	D8	2.1 ± 0.8	4.3 ± 0.2	3.9 ± 0.2	4.6 ± 0.2	5.0 ± 0.4	4.9 ± 0.3
D4	D9	2.1 ± 0.4	−0.6 ± 0.2	−1.3 ± 0.2	−1.2 ± 0.3	−1.3 ± 0.2	−1.5 ± 0.1
D4	D10	5.9 ± 0.5	3.4 ± 0.2	2.2 ± 0.2	2.4 ± 0.2	2.4 ± 0.4	2.3 ± 0.2
D4	D11	3.0 ± 0.3	2.2 ± 0.2	1.9 ± 0.3	2.2 ± 0.2	2.3 ± 0.3	1.8 ± 0.3
D4	D12	−1.7 ± 0.2	−2.9 ± 0.2	−3.4 ± 0.3	−3.1 ± 0.1	−3.5 ± 0.3	−3.7 ± 0.1
D4	D13	2.5 ± 0.9	3.1 ± 0.4	1.9 ± 0.3	3.1 ± 0.5	3.1 ± 0.5	2.5 ± 0.4
D5	D6	0.0 ± 0.3	−0.2 ± 0.2	−0.1 ± 0.2	−0.0 ± 0.2	−0.1 ± 0.2	0.1 ± 0.3
D5	D7	0.5 ± 0.4	−0.0 ± 0.1	−0.2 ± 0.2	−0.3 ± 0.2	−0.4 ± 0.2	0.0 ± 0.1
D5	D8	−4.0 ± 0.4	−4.0 ± 0.3	−3.9 ± 0.2	−3.6 ± 0.1	−3.7 ± 0.3	−3.7 ± 0.1
D5	D9	−4.0 ± 0.4	−9.2 ± 0.1	−9.2 ± 0.1	−9.3 ± 0.1	−10.0 ± 0.1	−9.8 ± 0.1
D5	D10	−0.2 ± 0.2	−5.2 ± 0.2	−5.6 ± 0.2	−5.7 ± 0.2	−6.3 ± 0.2	−6.0 ± 0.1
D5	D11	−3.1 ± 0.2	−6.4 ± 0.2	−5.9 ± 0.2	−5.9 ± 0.2	−6.4 ± 0.2	−6.5 ± 0.2
D5	D12	−7.8 ± 0.3	−11.5 ± 0.2	−11.2 ± 0.2	−11.2 ± 0.2	−12.1 ± 0.1	−11.9 ± 0.2
D5	D13	−3.6 ± 0.6	−5.5 ± 0.3	−5.9 ± 0.3	−5.0 ± 0.5	−5.6 ± 0.4	−5.7 ± 0.4
D6	D7	0.5 ± 0.4	0.2 ± 0.2	−0.2 ± 0.1	−0.3 ± 0.3	−0.3 ± 0.2	−0.1 ± 0.4
D6	D8	−4.0 ± 0.5	−3.9 ± 0.2	−3.8 ± 0.2	−3.6 ± 0.3	−3.6 ± 0.2	−3.9 ± 0.5
D6	D9	−4.0 ± 0.2	−9.1 ± 0.2	−9.1 ± 0.1	−9.3 ± 0.2	−9.9 ± 0.2	−9.9 ± 0.3
D6	D10	−0.2 ± 0.1	−5.0 ± 0.2	−5.6 ± 0.2	−5.7 ± 0.3	−6.2 ± 0.2	−6.1 ± 0.3
D6	D11	−3.1 ± 0.2	−6.2 ± 0.2	−5.8 ± 0.1	−5.9 ± 0.3	−6.3 ± 0.3	−6.6 ± 0.4
D6	D12	−7.9 ± 0.4	−11.4 ± 0.3	−11.1 ± 0.1	−11.2 ± 0.2	−12.0 ± 0.2	−12.0 ± 0.3
D6	D13	−3.7 ± 0.5	−5.4 ± 0.3	−5.8 ± 0.4	−5.0 ± 0.5	−5.5 ± 0.4	−5.9 ± 0.4
D7	D8	−4.5 ± 0.6	−4.0 ± 0.4	−3.6 ± 0.2	−3.3 ± 0.2	−3.3 ± 0.2	−3.8 ± 0.1
D7	D9	−4.5 ± 0.4	−9.2 ± 0.2	−8.9 ± 0.1	−9.0 ± 0.3	−9.6 ± 0.2	−9.8 ± 0.2
D7	D10	−0.7 ± 0.5	−5.2 ± 0.3	−5.4 ± 0.2	−5.4 ± 0.3	−5.9 ± 0.3	−6.0 ± 0.2
D7	D11	−3.6 ± 0.4	−6.3 ± 0.2	−5.7 ± 0.1	−5.6 ± 0.3	−6.0 ± 0.3	−6.5 ± 0.2
D7	D12	−8.4 ± 0.4	−11.5 ± 0.2	−11.0 ± 0.2	−10.9 ± 0.2	−11.8 ± 0.2	−11.9 ± 0.2
D7	D13	−4.2 ± 0.8	−5.5 ± 0.3	−5.7 ± 0.4	−4.7 ± 0.5	−5.2 ± 0.4	−5.7 ± 0.4
D8	D9	0.0 ± 0.6	−5.2 ± 0.3	−5.3 ± 0.2	−5.8 ± 0.2	−6.3 ± 0.3	−6.0 ± 0.2
D8	D10	3.8 ± 0.5	−1.2 ± 0.1	−1.8 ± 0.3	−2.1 ± 0.1	−2.6 ± 0.3	−2.2 ± 0.2
D8	D11	0.9 ± 0.6	−2.3 ± 0.3	−2.0 ± 0.2	−2.4 ± 0.2	−2.7 ± 0.3	−2.7 ± 0.2
D8	D12	−3.8 ± 0.7	−7.5 ± 0.3	−7.3 ± 0.3	−7.6 ± 0.2	−8.5 ± 0.3	−8.2 ± 0.2
D8	D13	0.4 ± 0.6	−1.5 ± 0.4	−2.0 ± 0.3	−1.5 ± 0.5	−1.9 ± 0.5	−2.0 ± 0.4
D9	D10	3.8 ± 0.4	4.0 ± 0.2	3.5 ± 0.1	3.6 ± 0.2	3.6 ± 0.2	3.8 ± 0.1
D9	D11	0.9 ± 0.3	2.9 ± 0.2	3.3 ± 0.2	3.4 ± 0.2	3.6 ± 0.2	3.3 ± 0.2
D9	D12	−3.8 ± 0.4	−2.3 ± 0.1	−2.0 ± 0.2	−1.9 ± 0.3	−2.2 ± 0.2	−2.1 ± 0.2
D9	D13	0.4 ± 0.6	3.7 ± 0.4	3.3 ± 0.4	4.3 ± 0.5	4.4 ± 0.4	4.0 ± 0.3

D10	D11	−2.9 ± 0.2	−1.1 ± 0.3	−0.3 ± 0.3	−0.2 ± 0.2	−0.0 ± 0.3	−0.5 ± 0.2
D10	D12	−7.6 ± 0.4	−6.3 ± 0.2	−5.6 ± 0.3	−5.5 ± 0.3	−5.8 ± 0.3	−5.9 ± 0.2
D10	D13	−3.4 ± 0.5	−0.3 ± 0.4	−0.3 ± 0.4	0.7 ± 0.5	0.7 ± 0.4	0.3 ± 0.3
D11	D12	−4.7 ± 0.3	−5.2 ± 0.2	−5.3 ± 0.2	−5.3 ± 0.2	−5.8 ± 0.2	−5.4 ± 0.3
D11	D13	−0.5 ± 0.6	0.8 ± 0.4	−0.0 ± 0.3	0.9 ± 0.5	0.8 ± 0.4	0.8 ± 0.5
D12	D13	4.2 ± 0.9	6.0 ± 0.4	5.3 ± 0.4	6.2 ± 0.5	6.5 ± 0.5	6.2 ± 0.4
RMSE		5.61 ± 0.7	5.08 ± 0.5	4.97 ± 0.6	4.83 ± 0.5	5.00 ± 0.5	5.05 ± 0.5
MAE		4.39 ± 0.7	4.24 ± 0.5	4.16 ± 0.5	4.02 ± 0.5	4.14 ± 0.5	4.21 ± 0.5
r_{Spearman}		0.31	0.52	0.52	0.58	0.57	0.55
$t_{\text{preparation}}$		207.8 ns	211 ns	211 ns	211 ns	211 ns	211 ns
$t_{\text{production}}$		56 ns	58 ns	58 ns	58 ns	58 ns	58 ns

Table 7.13: $\Delta\Delta G_{\text{CHCl}_3}$ for the 13 molecules of set D from experiment,[174] and from RE-EDS calculations using the RF schemes CG^{orig}, AT^{orig}, and AT^{shift} with $R_{\text{RF}} = 1.2$ nm. The RE-EDS results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G^{\text{exp}}_{\text{CHCl}_3}$[174]	$\Delta\Delta G^{\text{RE-EDS,CG}^{\text{orig}}}_{\text{CHCl}_3}$	$\Delta\Delta G^{\text{RE-EDS,AT}^{\text{orig}}}_{\text{CHCl}_3}$	$\Delta\Delta G^{\text{RE-EDS,AT}^{\text{shift}}}_{\text{CHCl}_3}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
D1	D2	5.5 ± 3.6	1.1 ± 0.2	1.1 ± 0.2	1.1 ± 0.2
D1	D3	1.9 ± 3.6	−0.5 ± 0.1	−0.8 ± 0.3	−0.6 ± 0.1
D1	D4	−1.1 ± 3.6	−5.3 ± 0.2	−5.4 ± 0.2	−5.5 ± 0.3
D1	D5	2.4 ± 3.6	3.1 ± 0.2	3.1 ± 0.2	3.1 ± 0.2
D1	D6	0.8 ± 3.6	2.9 ± 0.1	3.0 ± 0.1	3.0 ± 0.2
D1	D7	0.2 ± 3.6	2.9 ± 0.2	2.9 ± 0.4	2.8 ± 0.2
D1	D8	−0.8 ± 3.6	−0.6 ± 0.2	−0.8 ± 0.3	−0.9 ± 0.2
D1	D9	−11.5 ± 3.6	−6.8 ± 0.2	−7.2 ± 0.2	−7.2 ± 0.2
D1	D10	−4.4 ± 3.6	−3.3 ± 0.2	−3.4 ± 0.2	−3.4 ± 0.3
D1	D11	−0.7 ± 3.6	−3.3 ± 0.3	−3.5 ± 0.3	−3.3 ± 0.3
D1	D12	−1.1 ± 3.6	−8.6 ± 0.1	−8.9 ± 0.3	−9.1 ± 0.2
D1	D13	1.6 ± 3.6	−2.5 ± 0.3	−2.8 ± 0.5	−2.6 ± 0.7
D2	D3	−3.6 ± 3.6	−1.6 ± 0.2	−1.9 ± 0.2	−1.7 ± 0.2
D2	D4	−6.6 ± 3.6	−6.4 ± 0.3	−6.5 ± 0.2	−6.6 ± 0.3
D2	D5	−3.1 ± 3.6	2.1 ± 0.2	2.0 ± 0.1	2.0 ± 0.2
D2	D6	−4.6 ± 3.6	1.8 ± 0.3	1.9 ± 0.2	1.9 ± 0.2
D2	D7	−5.3 ± 3.6	1.8 ± 0.2	1.8 ± 0.2	1.8 ± 0.2
D2	D8	−6.3 ± 3.6	−1.7 ± 0.4	−1.9 ± 0.4	−2.0 ± 0.2
D2	D9	−17.0 ± 3.6	−7.9 ± 0.3	−8.3 ± 0.3	−8.3 ± 0.2
D2	D10	−9.8 ± 3.6	−4.4 ± 0.3	−4.5 ± 0.2	−4.5 ± 0.3
D2	D11	−6.2 ± 3.6	−4.4 ± 0.3	−4.6 ± 0.2	−4.3 ± 0.3
D2	D12	−6.6 ± 3.6	−9.7 ± 0.3	−10.0 ± 0.2	−10.2 ± 0.2
D2	D13	−3.8 ± 3.6	−3.6 ± 0.4	−3.8 ± 0.6	−3.7 ± 0.7
D3	D4	−3.0 ± 3.6	−4.7 ± 0.2	−4.6 ± 0.2	−4.9 ± 0.3
D3	D5	0.5 ± 3.6	3.7 ± 0.1	3.9 ± 0.2	3.8 ± 0.1
D3	D6	−1.1 ± 3.6	3.5 ± 0.1	3.8 ± 0.3	3.7 ± 0.1
D3	D7	−1.7 ± 3.6	3.4 ± 0.2	3.7 ± 0.3	3.5 ± 0.1
D3	D8	−2.7 ± 3.6	−0.1 ± 0.3	−0.0 ± 0.3	−0.2 ± 0.2
D3	D9	−13.4 ± 3.6	−6.3 ± 0.2	−6.5 ± 0.2	−6.6 ± 0.1
D3	D10	−6.3 ± 3.6	−2.7 ± 0.2	−2.6 ± 0.1	−2.8 ± 0.3
D3	D11	−2.6 ± 3.6	−2.8 ± 0.2	−2.7 ± 0.1	−2.6 ± 0.2
D3	D12	−3.0 ± 3.6	−8.1 ± 0.2	−8.2 ± 0.1	−8.5 ± 0.1
D3	D13	−0.3 ± 3.6	−1.9 ± 0.4	−2.0 ± 0.5	−1.9 ± 0.7
D4	D5	3.5 ± 3.6	8.4 ± 0.2	8.5 ± 0.2	8.7 ± 0.2
D4	D6	1.9 ± 3.6	8.2 ± 0.1	8.4 ± 0.3	8.6 ± 0.2
D4	D7	1.3 ± 3.6	8.2 ± 0.2	8.3 ± 0.4	8.4 ± 0.3
D4	D8	0.3 ± 3.6	4.7 ± 0.3	4.6 ± 0.4	4.7 ± 0.3
D4	D9	−10.4 ± 3.6	−1.6 ± 0.2	−1.8 ± 0.3	−1.7 ± 0.3
D4	D10	−3.3 ± 3.6	2.0 ± 0.3	2.0 ± 0.2	2.1 ± 0.3
D4	D11	0.4 ± 3.6	1.9 ± 0.3	1.9 ± 0.2	2.3 ± 0.3
D4	D12	0.0 ± 3.6	−3.4 ± 0.2	−3.5 ± 0.2	−3.6 ± 0.2
D4	D13	2.7 ± 3.6	2.8 ± 0.3	2.7 ± 0.5	3.0 ± 0.8
D5	D6	−1.5 ± 3.6	−0.2 ± 0.1	−0.1 ± 0.2	−0.1 ± 0.1
D5	D7	−2.2 ± 3.6	−0.3 ± 0.1	−0.2 ± 0.2	−0.3 ± 0.1
D5	D8	−3.2 ± 3.6	−3.8 ± 0.3	−3.9 ± 0.3	−4.0 ± 0.1
D5	D9	−13.9 ± 3.6	−10.0 ± 0.2	−10.4 ± 0.2	−10.3 ± 0.1
D5	D10	−6.7 ± 3.6	−6.4 ± 0.2	−6.5 ± 0.2	−6.6 ± 0.3
D5	D11	−3.1 ± 3.6	−6.5 ± 0.3	−6.6 ± 0.2	−6.4 ± 0.2
D5	D12	−3.5 ± 3.6	−11.8 ± 0.2	−12.1 ± 0.2	−12.3 ± 0.1
D5	D13	−0.8 ± 3.6	−5.6 ± 0.4	−5.9 ± 0.6	−5.7 ± 0.7
D6	D7	−0.6 ± 3.6	−0.0 ± 0.1	−0.1 ± 0.4	−0.2 ± 0.2
D6	D8	−1.6 ± 3.6	−3.5 ± 0.2	−3.8 ± 0.2	−3.9 ± 0.2
D6	D9	−12.3 ± 3.6	−9.8 ± 0.2	−10.2 ± 0.2	−10.2 ± 0.1
D6	D10	−5.2 ± 3.6	−6.2 ± 0.2	−6.4 ± 0.2	−6.5 ± 0.3
D6	D11	−1.5 ± 3.6	−6.3 ± 0.3	−6.5 ± 0.3	−6.3 ± 0.2
D6	D12	−1.9 ± 3.6	−11.6 ± 0.2	−12.0 ± 0.3	−12.2 ± 0.1
D6	D13	0.8 ± 3.6	−5.4 ± 0.3	−5.8 ± 0.6	−5.6 ± 0.7
D7	D8	−1.0 ± 3.6	−3.5 ± 0.3	−3.7 ± 0.4	−3.7 ± 0.1
D7	D9	−11.7 ± 3.6	−9.7 ± 0.2	−10.1 ± 0.4	−10.1 ± 0.1

D7	D10	−4.6 ± 3.6	−6.2 ± 0.3	−6.3 ± 0.3	−6.3 ± 0.3
D7	D11	−0.9 ± 3.6	−6.2 ± 0.3	−6.4 ± 0.3	−6.1 ± 0.2
D7	D12	−1.3 ± 3.6	−11.5 ± 0.2	−11.8 ± 0.3	−12.0 ± 0.2
D7	D13	1.4 ± 3.6	−5.4 ± 0.4	−5.6 ± 0.7	−5.4 ± 0.7
D8	D9	−10.7 ± 3.6	−6.2 ± 0.3	−6.4 ± 0.2	−6.3 ± 0.2
D8	D10	−3.6 ± 3.6	−2.7 ± 0.2	−2.6 ± 0.3	−2.6 ± 0.2
D8	D11	0.1 ± 3.6	−2.7 ± 0.3	−2.7 ± 0.3	−2.4 ± 0.2
D8	D12	−0.3 ± 3.6	−8.0 ± 0.2	−8.1 ± 0.3	−8.3 ± 0.2
D8	D13	2.4 ± 3.6	−1.9 ± 0.4	−1.9 ± 0.6	−1.7 ± 0.7
D9	D10	7.2 ± 3.6	3.6 ± 0.3	3.8 ± 0.2	3.8 ± 0.3
D9	D11	10.8 ± 3.6	3.5 ± 0.3	3.7 ± 0.2	4.0 ± 0.2
D9	D12	10.4 ± 3.6	−1.8 ± 0.3	−1.7 ± 0.2	−1.9 ± 0.1
D9	D13	13.1 ± 3.6	4.3 ± 0.4	4.5 ± 0.6	4.6 ± 0.7
D10	D11	3.6 ± 3.6	−0.1 ± 0.1	−0.1 ± 0.2	0.2 ± 0.3
D10	D12	3.3 ± 3.6	−5.3 ± 0.2	−5.5 ± 0.2	−5.7 ± 0.3
D10	D13	6.0 ± 3.6	0.8 ± 0.4	0.7 ± 0.5	0.9 ± 0.7
D11	D12	−0.4 ± 3.6	−5.3 ± 0.3	−5.4 ± 0.1	−5.9 ± 0.2
D11	D13	2.3 ± 3.6	0.8 ± 0.4	0.8 ± 0.5	0.7 ± 0.7
D12	D13	2.7 ± 3.6	6.1 ± 0.3	6.2 ± 0.5	6.6 ± 0.7
RMSE			4.90 ± 0.5	4.94 ± 0.5	4.97 ± 0.5
MAE			4.07 ± 0.4	4.10 ± 0.4	4.12 ± 0.4
τ_{Spearman}			0.58	0.58	0.59
$t_{\text{preparation}}$			211 ns	211 ns	211 ns
$t_{\text{production}}$			58 ns	58 ns	58 ns

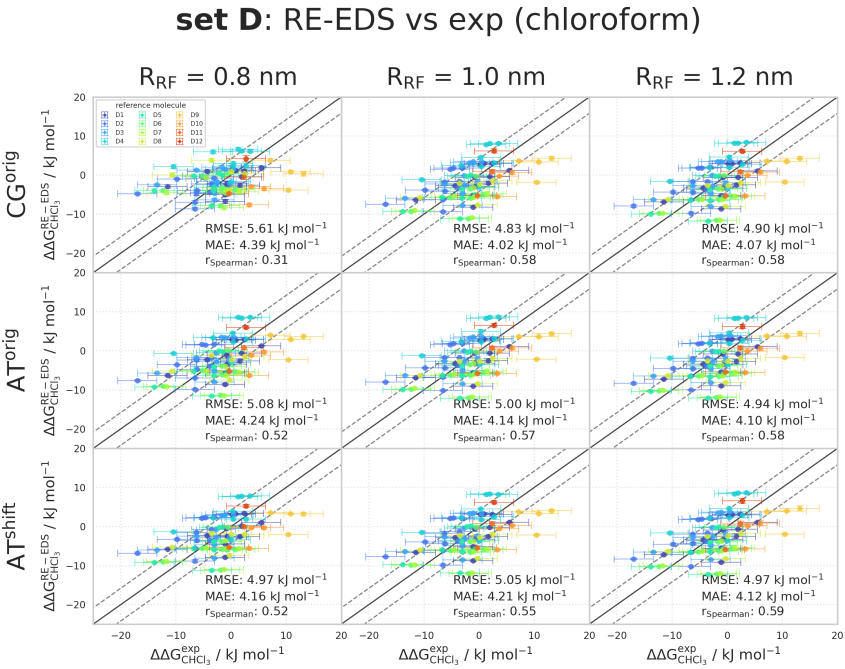


Figure 7.12: Comparison of the relative solvation free energies in chloroform of set D: $\Delta\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ obtained from RE-EDS calculations in GROMOS versus $\Delta\Delta G^{\text{exp}}_{\text{CHCl}_3}$ as reported in the Minnesota solvation[174] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to the RF schemes CG^{orig} (top), AT^{orig} (middle), and AT^{shift} (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Tables 7.12 and 7.13.

7.A.2 Comparison of Corrected Electrostatic Potential Energies in GROMOS

Overview of Statistical Metrics Against MBAR

Table 7.14: Overview of statistical metrics (RMSE, MAE, Spearman correlation coefficient) for RE-EDS simulations using the AT^{shift} scheme with and without corrected energy terms for sets C and D versus $\Delta\Delta G_{hyd}^{MBAR}$.[155,173] The uncertainties of the RMSE and MAE values were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve (set C) or eleven (set D) molecules was removed from the calculations (5000 repetitions). For the MBAR calculations, the cutoff was 1.2 nm for the electrostatic interactions and 1.0 nm for the vdW interactions (with a switch at 0.9 nm and a long-range dispersion correction)[155]

$\Delta\Delta G_{hyd}^{RE\text{-}EDS}$ vs $\Delta\Delta G_{hyd}^{exp}$

Set	V^{ele}	Cutoff [nm]	RMSE [kJ mol^{-1}]	MAE [kJ mol^{-1}]	$r_{Spearman}$
C	shift→orig	0.8	0.69 ± 0.1	0.54 ± 0.1	0.99
		1.0	0.52 ± 0.1	0.41 ± 0.1	0.99
		1.2	0.51 ± 0.1	0.39 ± 0.1	0.99
	shift→phys	0.8	2.96 ± 0.3	2.47 ± 0.3	0.90
		1.0	2.94 ± 0.4	2.46 ± 0.3	0.92
		1.2	2.96 ± 0.4	2.47 ± 0.4	0.92
	shift	0.8	1.10 ± 0.1	0.88 ± 0.1	0.97
		1.0	0.65 ± 0.1	0.52 ± 0.1	0.99
		1.2	0.57 ± 0.1	0.43 ± 0.1	0.99
D	shift→orig	0.8	0.62 ± 0.1	0.49 ± 0.1	1.00
		1.0	0.44 ± 0.0	0.37 ± 0.0	1.00
		1.2	0.49 ± 0.0	0.41 ± 0.0	1.00
	shift→phys	0.8	2.61 ± 0.2	2.12 ± 0.2	0.97
		1.0	2.31 ± 0.2	1.88 ± 0.2	0.97
		1.2	2.26 ± 0.2	1.79 ± 0.2	0.98
	shift	0.8	0.85 ± 0.2	0.64 ± 0.2	1.00
		1.0	0.48 ± 0.1	0.41 ± 0.1	1.00
		1.2	0.49 ± 0.1	0.42 ± 0.1	1.00

Relative Hydration Free Energies of Set C

Table 7.15: $\Delta\Delta G_{\mathrm{hyd}}$ for the 14 molecules of set C calculated from RE-EDS simulations using AT$^{\mathrm{shift}}$, corrected to $V^{\mathrm{ele,orig}}$ and $V^{\mathrm{ele,phys}}$, with $R_{\mathrm{RF}} = 0.8$ nm, $R_{\mathrm{RF}} = 1.0$ nm, and $R_{\mathrm{RF}} = 1.2$ nm. The results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions).

$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$

		$R_{\mathrm{RF}} = 0.8$ nm		$R_{\mathrm{RF}} = 1.0$ nm		$R_{\mathrm{RF}} = 1.2$ nm	
Molecules		shift→orig	shift→phys	shift→orig	shift→phys	shift→orig	shift→phys
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	0.6 ± 0.2	3.7 ± 0.2	0.3 ± 0.2	3.3 ± 0.2	0.1 ± 0.2	3.7 ± 0.1
C1	C3	−1.0 ± 0.1	2.3 ± 0.1	−1.0 ± 0.1	2.0 ± 0.1	−1.1 ± 0.1	2.2 ± 0.1
C1	C4	0.4 ± 0.1	3.3 ± 0.2	0.5 ± 0.1	3.2 ± 0.1	0.5 ± 0.1	3.5 ± 0.1
C1	C5	−20.3 ± 0.3	−23.1 ± 0.1	−20.5 ± 0.2	−23.6 ± 0.1	−20.6 ± 0.2	−23.1 ± 0.0
C1	C6	−19.6 ± 0.3	−20.2 ± 0.2	−19.9 ± 0.5	−20.7 ± 0.1	−20.1 ± 0.2	−20.8 ± 0.1
C1	C7	−19.7 ± 0.2	−23.4 ± 0.1	−19.5 ± 0.2	−23.5 ± 0.0	−19.8 ± 0.1	−23.7 ± 0.0
C1	C8	3.1 ± 0.2	1.5 ± 0.1	3.0 ± 0.1	1.2 ± 0.1	3.1 ± 0.1	1.5 ± 0.0
C1	C9	1.7 ± 0.1	1.3 ± 0.1	1.5 ± 0.1	0.9 ± 0.1	1.3 ± 0.2	1.2 ± 0.0
C1	C10	3.2 ± 0.1	4.4 ± 0.1	2.9 ± 0.1	3.8 ± 0.1	2.6 ± 0.1	4.1 ± 0.0
C1	C11	−0.1 ± 0.1	−0.2 ± 0.1	−0.6 ± 0.1	−0.8 ± 0.1	−0.7 ± 0.1	−0.6 ± 0.0
C1	C12	−19.2 ± 0.2	−20.5 ± 0.1	−19.3 ± 0.2	−21.1 ± 0.1	−19.6 ± 0.2	−20.4 ± 0.1
C1	C13	−19.3 ± 0.2	−20.7 ± 0.1	−19.7 ± 0.3	−21.5 ± 0.0	−19.9 ± 0.4	−21.3 ± 0.1
C1	C14	0.2 ± 0.2	1.2 ± 0.0	−0.4 ± 0.0	0.0 ± 0.1	−0.6 ± 0.2	0.2 ± 0.1
C2	C3	−1.6 ± 0.3	−1.3 ± 0.2	−1.3 ± 0.3	−1.2 ± 0.3	−1.3 ± 0.2	−1.5 ± 0.1
C2	C4	−0.1 ± 0.2	−0.4 ± 0.3	0.2 ± 0.2	−0.1 ± 0.2	0.3 ± 0.2	−0.2 ± 0.2
C2	C5	−20.9 ± 0.3	−26.8 ± 0.2	−20.8 ± 0.2	−26.9 ± 0.2	−20.8 ± 0.2	−26.8 ± 0.1
C2	C6	−20.2 ± 0.5	−23.9 ± 0.2	−20.2 ± 0.5	−24.0 ± 0.2	−20.3 ± 0.2	−24.5 ± 0.1
C2	C7	−20.3 ± 0.4	−27.1 ± 0.2	−19.8 ± 0.2	−26.8 ± 0.2	−19.9 ± 0.2	−27.4 ± 0.1
C2	C8	2.5 ± 0.2	−2.1 ± 0.2	2.7 ± 0.2	−2.0 ± 0.2	2.9 ± 0.1	−2.2 ± 0.1
C2	C9	1.1 ± 0.2	−2.3 ± 0.2	1.2 ± 0.2	−2.3 ± 0.2	1.2 ± 0.1	−2.5 ± 0.1
C2	C10	2.6 ± 0.2	0.8 ± 0.2	2.6 ± 0.2	0.6 ± 0.1	2.4 ± 0.1	0.4 ± 0.1
C2	C11	−0.7 ± 0.2	−3.9 ± 0.2	−0.9 ± 0.2	−4.1 ± 0.2	−0.9 ± 0.1	−4.3 ± 0.1
C2	C12	−19.8 ± 0.3	−24.2 ± 0.2	−19.6 ± 0.3	−24.4 ± 0.2	−19.8 ± 0.3	−24.1 ± 0.1
C2	C13	−19.9 ± 0.2	−24.4 ± 0.2	−20.0 ± 0.3	−24.7 ± 0.2	−20.1 ± 0.4	−25.0 ± 0.1
C2	C14	−0.3 ± 0.2	−2.5 ± 0.2	−0.7 ± 0.1	−3.2 ± 0.2	−0.7 ± 0.1	−3.5 ± 0.1
C3	C4	1.4 ± 0.1	0.9 ± 0.2	1.6 ± 0.1	1.2 ± 0.2	1.6 ± 0.2	1.2 ± 0.2
C3	C5	−19.4 ± 0.3	−25.4 ± 0.1	−19.5 ± 0.3	−25.7 ± 0.1	−19.5 ± 0.2	−25.3 ± 0.1
C3	C6	−18.7 ± 0.3	−22.5 ± 0.1	−18.9 ± 0.5	−22.7 ± 0.2	−19.0 ± 0.2	−23.0 ± 0.1
C3	C7	−18.7 ± 0.2	−25.7 ± 0.1	−18.5 ± 0.3	−25.5 ± 0.1	−18.7 ± 0.2	−25.9 ± 0.1
C3	C8	4.0 ± 0.2	−0.8 ± 0.1	4.0 ± 0.2	−0.8 ± 0.1	4.2 ± 0.1	−0.7 ± 0.1
C3	C9	2.7 ± 0.1	−1.0 ± 0.1	2.5 ± 0.2	−1.1 ± 0.2	2.5 ± 0.2	−1.0 ± 0.1
C3	C10	4.2 ± 0.1	2.1 ± 0.1	3.9 ± 0.2	1.8 ± 0.2	3.7 ± 0.1	1.8 ± 0.1
C3	C11	0.8 ± 0.2	−2.5 ± 0.2	0.5 ± 0.2	−2.8 ± 0.1	0.4 ± 0.1	−2.9 ± 0.1
C3	C12	−18.2 ± 0.2	−22.9 ± 0.1	−18.3 ± 0.2	−23.2 ± 0.1	−18.5 ± 0.2	−22.6 ± 0.1
C3	C13	−18.3 ± 0.3	−23.0 ± 0.1	−18.7 ± 0.3	−23.5 ± 0.1	−18.8 ± 0.3	−23.5 ± 0.1
C3	C14	1.2 ± 0.2	−1.1 ± 0.2	0.9 ± 0.2	−1.4 ± 0.1	0.5 ± 0.1	−1.6 ± 0.1
C4	C5	−20.8 ± 0.3	−26.4 ± 0.2	−21.1 ± 0.2	−26.8 ± 0.2	−21.1 ± 0.1	−26.6 ± 0.1
C4	C6	−20.1 ± 0.4	−23.5 ± 0.2	−20.5 ± 0.4	−23.9 ± 0.2	−20.6 ± 0.2	−24.2 ± 0.1
C4	C7	−20.1 ± 0.3	−26.6 ± 0.1	−20.1 ± 0.2	−26.7 ± 0.1	−20.3 ± 0.1	−27.1 ± 0.1
C4	C8	2.6 ± 0.2	−1.7 ± 0.1	2.5 ± 0.1	−2.0 ± 0.2	2.6 ± 0.2	−1.9 ± 0.1
C4	C9	1.3 ± 0.1	−1.9 ± 0.2	0.9 ± 0.1	−2.3 ± 0.1	0.9 ± 0.2	−2.3 ± 0.2
C4	C10	2.8 ± 0.1	1.2 ± 0.1	2.4 ± 0.1	0.6 ± 0.2	2.1 ± 0.1	0.6 ± 0.1
C4	C11	−0.6 ± 0.2	−3.4 ± 0.2	−1.1 ± 0.1	−4.0 ± 0.1	−1.2 ± 0.2	−4.1 ± 0.1
C4	C12	−19.6 ± 0.2	−23.8 ± 0.2	−19.8 ± 0.2	−24.3 ± 0.1	−20.1 ± 0.3	−23.9 ± 0.1
C4	C13	−19.7 ± 0.2	−24.0 ± 0.2	−20.3 ± 0.2	−24.7 ± 0.1	−20.4 ± 0.4	−24.8 ± 0.2
C4	C14	−0.2 ± 0.1	−2.1 ± 0.2	−1.0 ± 0.1	−3.2 ± 0.1	−1.1 ± 0.1	−3.2 ± 0.1
C5	C6	0.7 ± 0.6	2.9 ± 0.1	0.6 ± 0.3	2.9 ± 0.1	0.5 ± 0.2	2.3 ± 0.1
C5	C7	0.7 ± 0.4	−0.3 ± 0.0	1.0 ± 0.2	0.1 ± 0.1	0.8 ± 0.2	−0.6 ± 0.0
C5	C8	23.4 ± 0.3	24.6 ± 0.0	23.5 ± 0.2	24.9 ± 0.0	23.7 ± 0.2	24.6 ± 0.0
C5	C9	22.0 ± 0.2	24.5 ± 0.0	22.0 ± 0.2	24.6 ± 0.1	21.9 ± 0.2	24.3 ± 0.1
C5	C10	23.6 ± 0.3	27.5 ± 0.1	23.4 ± 0.3	27.5 ± 0.1	23.2 ± 0.1	27.2 ± 0.1
C5	C11	20.2 ± 0.2	22.9 ± 0.1	20.0 ± 0.2	22.8 ± 0.1	19.9 ± 0.2	22.4 ± 0.0
C5	C12	1.2 ± 0.2	2.6 ± 0.0	1.2 ± 0.4	2.5 ± 0.1	1.0 ± 0.4	2.7 ± 0.1
C5	C13	1.1 ± 0.1	2.4 ± 0.0	0.8 ± 0.4	2.2 ± 0.0	0.7 ± 0.5	1.8 ± 0.1
C5	C14	20.6 ± 0.3	24.3 ± 0.1	20.1 ± 0.2	23.7 ± 0.1	20.0 ± 0.2	23.3 ± 0.1
C6	C7	−0.0 ± 0.4	−3.2 ± 0.1	0.4 ± 0.5	−2.8 ± 0.1	0.3 ± 0.3	−2.9 ± 0.1
C6	C8	22.7 ± 0.4	21.7 ± 0.1	22.9 ± 0.4	21.9 ± 0.1	23.2 ± 0.2	22.3 ± 0.0
C6	C9	21.3 ± 0.4	21.5 ± 0.1	21.4 ± 0.4	21.6 ± 0.1	21.4 ± 0.3	22.0 ± 0.1
C6	C10	22.9 ± 0.4	24.6 ± 0.1	22.8 ± 0.5	24.5 ± 0.1	22.7 ± 0.2	24.8 ± 0.1
C6	C11	19.5 ± 0.4	20.0 ± 0.2	19.4 ± 0.4	19.9 ± 0.1	19.4 ± 0.2	20.1 ± 0.1
C6	C12	0.5 ± 0.5	−0.3 ± 0.1	0.6 ± 0.6	−0.4 ± 0.1	0.5 ± 0.3	0.4 ± 0.1
C6	C13	0.4 ± 0.5	−0.5 ± 0.1	0.2 ± 0.6	−0.8 ± 0.1	0.2 ± 0.4	−0.5 ± 0.1
C6	C14	19.9 ± 0.5	21.4 ± 0.1	19.5 ± 0.4	20.7 ± 0.1	19.5 ± 0.2	21.0 ± 0.1
C7	C8	22.7 ± 0.2	24.9 ± 0.0	22.5 ± 0.2	24.7 ± 0.1	22.9 ± 0.1	25.2 ± 0.0
C7	C9	21.4 ± 0.2	24.7 ± 0.0	21.0 ± 0.2	24.4 ± 0.1	21.1 ± 0.2	24.9 ± 0.0
C7	C10	22.9 ± 0.2	27.8 ± 0.1	22.4 ± 0.2	27.3 ± 0.1	22.4 ± 0.1	27.7 ± 0.1
C7	C11	19.5 ± 0.3	23.2 ± 0.1	19.0 ± 0.2	22.7 ± 0.0	19.1 ± 0.2	23.0 ± 0.0
C7	C12	0.5 ± 0.2	2.8 ± 0.0	0.2 ± 0.2	2.4 ± 0.1	0.2 ± 0.3	3.3 ± 0.0
C7	C13	0.4 ± 0.3	2.7 ± 0.0	−0.2 ± 0.2	2.0 ± 0.0	−0.1 ± 0.4	2.4 ± 0.1
C7	C14	19.9 ± 0.2	24.6 ± 0.1	19.1 ± 0.2	23.5 ± 0.0	19.2 ± 0.2	23.9 ± 0.1

C8	C9	−1.4	±	0.1	−0.2	±	0.0	−1.5	±	0.1	−0.3	±	0.1	−1.8	±	0.1	−0.3	±	0.1
C8	C10	0.2	±	0.2	2.9	±	0.1	−0.1	±	0.1	2.6	±	0.1	−0.5	±	0.1	2.5	±	0.1
C8	C11	−3.2	±	0.1	−1.7	±	0.1	−3.6	±	0.1	−2.0	±	0.1	−3.8	±	0.0	−2.2	±	0.0
C8	C12	−22.2	±	0.1	−22.1	±	0.0	−22.3	±	0.3	−22.4	±	0.1	−22.7	±	0.2	−21.9	±	0.0
C8	C13	−22.3	±	0.2	−22.2	±	0.0	−22.7	±	0.3	−22.7	±	0.1	−23.0	±	0.3	−22.8	±	0.1
C8	C14	−2.8	±	0.1	−0.3	±	0.0	−3.4	±	0.1	−1.2	±	0.1	−3.7	±	0.1	−1.3	±	0.1
C9	C10	1.5	±	0.1	3.1	±	0.0	1.4	±	0.1	2.9	±	0.1	1.2	±	0.1	2.9	±	0.0
C9	C11	−1.8	±	0.1	−1.5	±	0.1	−2.0	±	0.1	−1.8	±	0.1	−2.0	±	0.1	−1.8	±	0.0
C9	C12	−20.9	±	0.1	−21.9	±	0.0	−20.8	±	0.3	−22.1	±	0.1	−21.0	±	0.3	−21.6	±	0.1
C9	C13	−21.0	±	0.2	−22.0	±	0.0	−21.2	±	0.3	−22.4	±	0.1	−21.2	±	0.4	−22.5	±	0.1
C9	C14	−1.5	±	0.1	−0.1	±	0.1	−1.9	±	0.1	−0.9	±	0.1	−1.9	±	0.1	−1.0	±	0.1
C10	C11	−3.4	±	0.1	−4.6	±	0.1	−3.5	±	0.1	−4.6	±	0.1	−3.3	±	0.1	−4.7	±	0.0
C10	C12	−22.4	±	0.2	−25.0	±	0.1	−22.2	±	0.3	−25.0	±	0.1	−22.2	±	0.3	−24.5	±	0.1
C10	C13	−22.5	±	0.2	−25.1	±	0.1	−22.6	±	0.3	−25.3	±	0.1	−22.5	±	0.4	−25.4	±	0.1
C10	C14	−3.0	±	0.2	−3.2	±	0.1	−3.3	±	0.1	−3.8	±	0.1	−3.2	±	0.1	−3.8	±	0.0
C11	C12	−19.0	±	0.1	−20.4	±	0.1	−18.7	±	0.3	−20.3	±	0.1	−18.9	±	0.2	−19.8	±	0.1
C11	C13	−19.1	±	0.1	−20.5	±	0.1	−19.2	±	0.3	−20.7	±	0.1	−19.2	±	0.3	−20.7	±	0.1
C11	C14	0.4	±	0.2	1.4	±	0.1	0.1	±	0.1	0.9	±	0.0	0.1	±	0.1	0.9	±	0.1
C12	C13	−0.1	±	0.2	−0.2	±	0.0	−0.4	±	0.1	−0.3	±	0.1	−0.3	±	0.2	−0.9	±	0.1
C12	C14	19.4	±	0.2	21.7	±	0.1	18.9	±	0.2	21.2	±	0.1	19.0	±	0.3	20.6	±	0.1
C13	C14	19.5	±	0.2	21.9	±	0.1	19.3	±	0.3	21.5	±	0.0	19.3	±	0.4	21.5	±	0.1
RMSE		2.37	±	0.2	3.46	±	0.5	2.21	±	0.2	3.33	±	0.4	2.21	±	0.2	3.40	±	0.5
MAE		1.94	±	0.2	2.80	±	0.4	1.82	±	0.2	2.70	±	0.4	1.82	±	0.2	2.74	±	0.4
r_{Spearman}				0.90			0.91			0.92			0.93			0.92			0.92

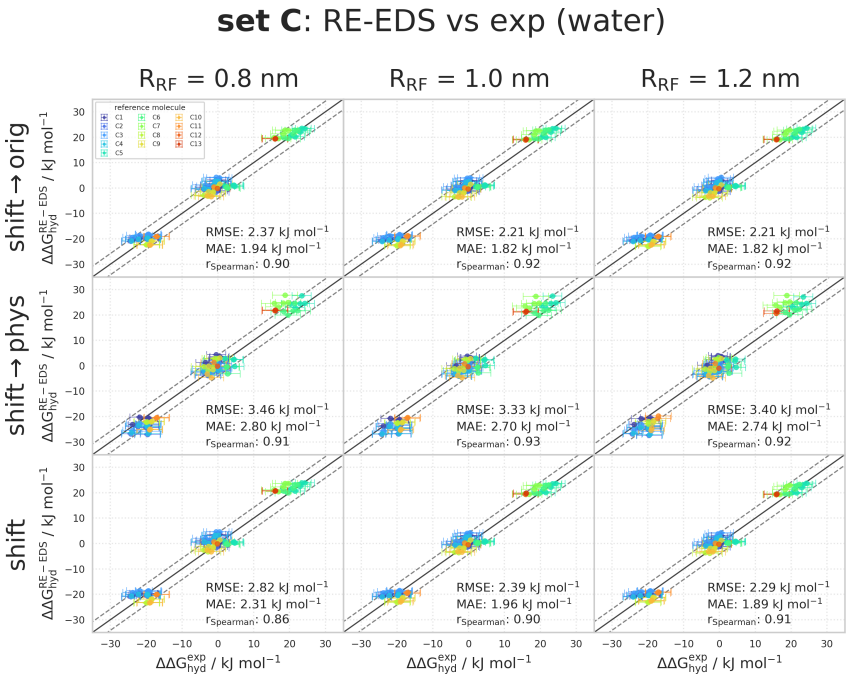


Figure 7.13: Comparison of the relative hydration free energies of set C: $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ obtained from RE-EDS calculations in GROMOS propagated with the AT$^{\text{shift}}$ scheme versus $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\text{shift}\rightarrow\text{orig}}$ (top), $V^{\text{shift}\rightarrow\text{pyhs}}$ (middle), and the uncorrected electrostatic potential energy $V^{\text{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.15.

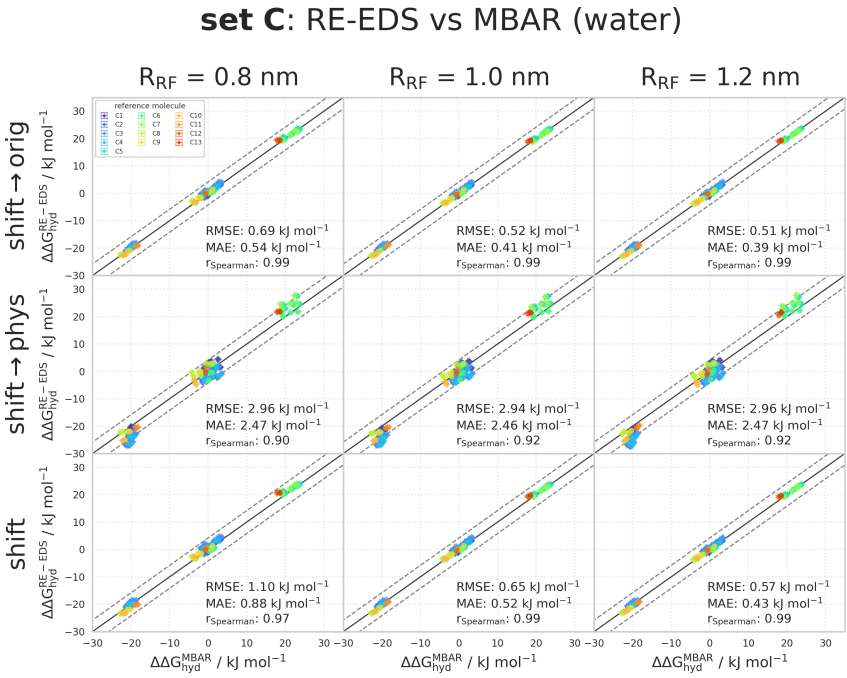


Figure 7.14: Comparison of the relative hydration free energies of set C: $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$ obtained from RE-EDS calculations in GROMOS[27] propagated with the AT$^{\mathrm{shift}}$ scheme versus $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\mathrm{RF}} = 0.8$ nm (left), $R_{\mathrm{RF}} = 1.0$ nm (middle), and $R_{\mathrm{RF}} = 1.2$ nm (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\mathrm{shift}\to\mathrm{orig}}$ (top), $V^{\mathrm{shift}\to\mathrm{pyhs}}$ (middle), and the uncorrected electrostatic potential energy $V^{\mathrm{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.15.

Relative Solvation Free Energies in Chloroform of Set C

Table 7.16: $\Delta\Delta G_{\mathrm{CHCl_3}}$ for the 14 molecules of set C calculated from RE-EDS simulations using AT$^{\mathrm{shift}}$, corrected to Vele,orig and $V^{\mathrm{ele,phys}}$, with $R_{\mathrm{RF}} = 0.8$ nm, $R_{\mathrm{RF}} = 1.0$ nm, and $R_{\mathrm{RF}} = 1.2$ nm. The results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions).

		$\Delta\Delta G^{\text{RE-EDS}}_{\mathrm{CHCl_3}}$											
		$R_{\mathrm{RF}} = 0.8$ nm				$R_{\mathrm{RF}} = 1.0$ nm				$R_{\mathrm{RF}} = 1.2$ nm			
Molecules		shift→orig		shift→phys		shift→orig		shift→phys		shift→orig		shift→phys	
i	j	[kJ mol^{-1}]		[kJ mol^{-1}]		[kJ mol^{-1}]		[kJ mol^{-1}]		[kJ mol^{-1}]		[kJ mol^{-1}]	
C1	C2	−3.9	0.2	−1.0	0.2	−4.1	0.2	−0.9	0.1	−4.1	0.2	−1.1	0.2
C1	C3	−5.2	0.1	−2.2	0.1	−5.5	0.1	−2.6	0.1	−5.4	0.2	−2.6	0.1
C1	C4	−3.9	0.3	−1.8	0.1	−4.1	0.3	−1.9	0.2	−4.2	0.2	−1.9	0.1
C1	C5	−2.3	0.2	−4.7	0.1	−2.8	0.2	−5.0	0.1	−3.1	0.1	−5.5	0.0
C1	C6	−6.4	0.4	−6.5	0.1	−7.2	0.2	−6.9	0.1	−7.3	0.1	−7.4	0.1
C1	C7	−3.0	0.2	−6.2	0.1	−3.7	0.2	−6.8	0.1	−4.0	0.1	−7.2	0.1
C1	C8	3.2	0.2	2.2	0.1	3.5	0.2	2.6	0.1	3.5	0.1	2.5	0.0
C1	C9	−0.5	0.2	−0.7	0.1	−0.7	0.2	−0.3	0.1	−0.7	0.1	−0.7	0.0
C1	C10	−4.4	0.2	−3.3	0.1	−4.9	0.2	−3.3	0.1	−5.1	0.2	−3.4	0.1
C1	C11	−3.3	0.2	−3.2	0.1	−3.8	0.3	−3.3	0.1	−4.2	0.1	−3.9	0.1
C1	C12	−7.2	0.1	−8.1	0.1	−8.2	0.1	−8.8	0.1	−8.4	0.2	−8.8	0.1
C1	C13	−7.1	0.2	−8.0	0.1	−8.2	0.2	−8.3	0.1	−8.3	0.2	−8.5	0.1
C1	C14	−4.9	0.2	−4.4	0.1	−5.7	0.2	−4.0	0.1	−5.9	0.1	−4.9	0.1
C2	C3	−1.3	0.2	−1.3	0.2	−1.5	0.2	−1.7	0.2	−1.3	0.3	−1.4	0.2
C2	C4	0.0	0.3	−0.8	0.2	−0.1	0.3	−1.1	0.2	−0.2	0.2	−0.8	0.2
C2	C5	1.6	0.2	−3.7	0.2	1.2	0.2	−4.1	0.1	1.0	0.2	−4.4	0.2
C2	C6	−2.5	0.3	−5.5	0.2	−3.1	0.2	−6.0	0.2	−3.3	0.2	−6.3	0.2
C2	C7	0.9	0.2	−5.2	0.2	0.4	0.2	−6.0	0.1	0.1	0.1	−6.1	0.2
C2	C8	7.1	0.2	3.2	0.2	7.5	0.2	3.5	0.1	7.6	0.1	3.6	0.1
C2	C9	3.4	0.2	0.3	0.2	3.4	0.2	0.6	0.2	3.3	0.2	0.4	0.1
C2	C10	−0.5	0.2	−2.4	0.2	−0.9	0.2	−2.4	0.1	−1.1	0.1	−2.3	0.1
C2	C11	0.6	0.2	−2.3	0.2	0.2	0.2	−2.5	0.2	−0.1	0.2	−2.8	0.1
C2	C12	−3.3	0.2	−7.1	0.2	−4.1	0.1	−8.0	0.1	−4.3	0.2	−7.7	0.2
C2	C13	−3.2	0.3	−7.0	0.2	−4.1	0.2	−7.4	0.2	−4.2	0.1	−7.3	0.2
C2	C14	−1.0	0.2	−3.4	0.2	−1.6	0.2	−3.1	0.2	−1.9	0.1	−3.7	0.1
C3	C4	1.3	0.2	0.4	0.1	1.4	0.3	0.6	0.3	1.1	0.1	0.6	0.0
C3	C5	2.8	0.1	−2.4	0.1	2.7	0.2	−2.4	0.2	2.3	0.2	−2.9	0.1
C3	C6	−1.3	0.3	−4.2	0.1	−1.7	0.2	−4.3	0.1	−1.9	0.2	−4.8	0.1
C3	C7	2.2	0.1	−3.9	0.1	1.9	0.2	−4.3	0.1	1.4	0.3	−4.6	0.1
C3	C8	8.4	0.1	4.4	0.1	9.0	0.2	5.2	0.1	8.9	0.3	5.1	0.1
C3	C9	4.7	0.1	1.6	0.1	4.9	0.2	2.3	0.1	4.6	0.2	1.8	0.1
C3	C10	0.8	0.1	−1.1	0.1	0.6	0.2	−0.7	0.2	0.3	0.2	−0.9	0.1
C3	C11	1.9	0.1	−1.0	0.1	1.7	0.3	−0.8	0.2	1.2	0.2	−1.4	0.1
C3	C12	−2.0	0.1	−5.9	0.1	−2.6	0.1	−6.2	0.1	−3.0	0.2	−6.3	0.1
C3	C13	−2.0	0.2	−5.7	0.1	−2.7	0.2	−5.7	0.2	−2.9	0.2	−5.9	0.1
C3	C14	0.3	0.1	−2.1	0.1	−0.2	0.2	−1.4	0.2	−0.6	0.2	−2.3	0.2
C4	C5	1.5	0.2	−2.9	0.1	1.3	0.2	−3.1	0.1	1.1	0.1	−3.6	0.1
C4	C6	−2.6	0.2	−4.7	0.2	−3.1	0.3	−5.0	0.2	−3.1	0.2	−5.5	0.1
C4	C7	0.8	0.2	−4.4	0.1	0.5	0.2	−4.9	0.1	0.3	0.3	−5.3	0.1
C4	C8	7.1	0.1	4.0	0.1	7.6	0.2	4.6	0.2	7.8	0.2	4.4	0.1
C4	C9	3.4	0.2	1.1	0.1	3.5	0.1	1.7	0.2	3.5	0.2	1.2	0.1
C4	C10	−0.5	0.2	−1.6	0.2	−0.8	0.2	−1.4	0.2	−0.9	0.2	−1.5	0.1
C4	C11	0.6	0.2	−1.4	0.2	0.3	0.1	−1.4	0.1	0.1	0.2	−2.0	0.1
C4	C12	−3.3	0.3	−6.3	0.1	−4.0	0.2	−6.9	0.2	−4.1	0.2	−6.9	0.1
C4	C13	−3.3	0.3	−6.2	0.2	−4.0	0.1	−6.4	0.1	−4.1	0.2	−6.6	0.1
C4	C14	−1.0	0.1	−2.6	0.2	−1.6	0.2	−2.1	0.1	−1.7	0.2	−2.9	0.2
C5	C6	−4.1	0.2	−1.8	0.1	−4.4	0.1	−1.9	0.1	−4.2	0.1	−1.9	0.1
C5	C7	−0.7	0.1	−1.5	0.0	−0.8	0.1	−1.8	0.0	−0.8	0.1	−1.7	0.1
C5	C8	5.6	0.1	6.8	0.1	6.3	0.0	7.6	0.0	6.7	0.1	8.0	0.0
C5	C9	1.8	0.1	4.0	0.0	2.2	0.0	4.7	0.1	2.4	0.1	4.8	0.0
C5	C10	−2.1	0.1	1.3	0.1	−2.1	0.1	1.7	0.1	−2.0	0.1	2.1	0.1
C5	C11	−0.9	0.1	1.4	0.0	−1.0	0.1	1.7	0.1	−1.0	0.1	1.6	0.0
C5	C12	−4.9	0.1	−3.4	0.1	−5.3	0.1	−3.8	0.1	−5.3	0.2	−3.3	0.1
C5	C13	−4.8	0.2	−3.3	0.1	−5.3	0.0	−3.3	0.1	−5.2	0.1	−3.0	0.1
C5	C14	−2.6	0.1	0.3	0.1	−2.9	0.1	1.0	0.1	−2.8	0.1	0.7	0.1
C6	C7	3.4	0.2	0.3	0.1	3.5	0.1	0.1	0.1	3.4	0.1	0.2	0.0
C6	C8	9.7	0.2	8.7	0.1	10.7	0.1	9.5	0.1	10.9	0.1	9.9	0.1
C6	C9	5.9	0.2	5.8	0.1	6.5	0.2	6.6	0.1	6.6	0.1	6.7	0.1
C6	C10	2.1	0.2	3.1	0.1	2.3	0.2	3.6	0.1	2.2	0.1	4.0	0.0
C6	C11	3.2	0.2	3.3	0.1	3.4	0.2	3.6	0.1	3.2	0.1	3.5	0.0
C6	C12	−0.7	0.3	−1.6	0.1	−1.0	0.2	−1.9	0.1	−1.0	0.2	−1.4	0.1
C6	C13	−0.7	0.3	−1.5	0.1	−1.0	0.2	−1.4	0.1	−1.0	0.2	−1.1	0.1
C6	C14	1.5	0.2	2.1	0.1	1.5	0.1	2.9	0.1	1.4	0.1	2.6	0.1
C7	C8	6.3	0.1	8.4	0.1	7.1	0.1	9.5	0.0	7.5	0.1	9.7	0.0
C7	C9	2.5	0.0	5.5	0.1	3.0	0.1	6.6	0.1	3.2	0.1	6.5	0.1
C7	C10	−1.4	0.1	2.8	0.1	−1.3	0.2	3.5	0.1	−1.2	0.1	3.8	0.0
C7	C11	−0.3	0.0	3.0	0.1	−0.2	0.1	3.5	0.1	−0.2	0.1	3.3	0.0
C7	C12	−4.2	0.1	−1.9	0.0	−4.5	0.1	−2.0	0.1	−4.4	0.3	−1.6	0.1
C7	C13	−4.1	0.2	−1.8	0.1	−4.5	0.1	−1.5	0.1	−4.4	0.1	−1.3	0.0
C7	C14	−1.9	0.1	1.8	0.1	−2.0	0.1	2.8	0.1	−2.0	0.1	2.4	0.1

C8	C9	−3.7	±	0.1	−2.9	±	0.1	−4.1	±	0.0	−2.9	±	0.1	−4.3	±	0.1	−3.2	±	0.0
C8	C10	−7.6	±	0.1	−5.5	±	0.1	−8.4	±	0.1	−5.9	±	0.1	−8.7	±	0.1	−5.9	±	0.0
C8	C11	−6.5	±	0.1	−5.4	±	0.1	−7.3	±	0.1	−6.0	±	0.1	−7.7	±	0.1	−6.4	±	0.0
C8	C12	−10.4	±	0.1	−10.3	±	0.0	−11.6	±	0.1	−11.4	±	0.1	−11.9	±	0.3	−11.3	±	0.1
C8	C13	−10.4	±	0.2	−10.2	±	0.1	−11.6	±	0.0	−10.9	±	0.1	−11.9	±	0.1	−11.0	±	0.1
C8	C14	−8.1	±	0.0	−6.6	±	0.1	−9.1	±	0.0	−6.6	±	0.1	−9.5	±	0.1	−7.4	±	0.1
C9	C10	−3.9	±	0.1	−2.7	±	0.1	−4.3	±	0.1	−3.0	±	0.1	−4.4	±	0.1	−2.7	±	0.1
C9	C11	−2.8	±	0.0	−2.6	±	0.0	−3.2	±	0.1	−3.0	±	0.1	−3.4	±	0.1	−3.2	±	0.0
C9	C12	−6.7	±	0.1	−7.4	±	0.1	−7.5	±	0.1	−8.5	±	0.1	−7.6	±	0.2	−8.1	±	0.1
C9	C13	−6.6	±	0.2	−7.3	±	0.1	−7.5	±	0.0	−8.0	±	0.1	−7.6	±	0.1	−7.8	±	0.1
C9	C14	−4.4	±	0.1	−3.7	±	0.1	−5.0	±	0.1	−3.7	±	0.1	−5.2	±	0.1	−4.1	±	0.1
C10	C11	1.1	±	0.1	0.1	±	0.1	1.1	±	0.2	−0.0	±	0.1	1.0	±	0.1	−0.5	±	0.0
C10	C12	−2.8	±	0.1	−4.7	±	0.1	−3.2	±	0.2	−5.5	±	0.1	−3.2	±	0.2	−5.4	±	0.1
C10	C13	−2.8	±	0.2	−4.6	±	0.1	−3.3	±	0.1	−5.0	±	0.1	−3.2	±	0.1	−5.1	±	0.1
C10	C14	−0.5	±	0.1	−1.0	±	0.1	−0.8	±	0.1	−0.7	±	0.1	−0.8	±	0.1	−1.4	±	0.1
C11	C12	−3.9	±	0.1	−4.9	±	0.1	−4.3	±	0.2	−5.5	±	0.1	−4.2	±	0.2	−4.9	±	0.1
C11	C13	−3.9	±	0.2	−4.7	±	0.1	−4.4	±	0.1	−5.0	±	0.1	−4.2	±	0.2	−4.6	±	0.1
C11	C14	−1.6	±	0.1	−1.2	±	0.1	−1.9	±	0.1	−0.7	±	0.1	−1.8	±	0.1	−0.9	±	0.1
C12	C13	0.0	±	0.2	0.1	±	0.1	−0.0	±	0.1	0.5	±	0.1	0.1	±	0.3	0.3	±	0.1
C12	C14	2.3	±	0.2	3.7	±	0.1	2.5	±	0.1	4.8	±	0.1	2.4	±	0.2	4.0	±	0.2
C13	C14	2.2	±	0.2	3.6	±	0.1	2.5	±	0.1	4.3	±	0.1	2.4	±	0.1	3.6	±	0.1
RMSE		3.84	±	0.3	2.37	±	0.3	3.52	±	0.3	1.97	±	0.2	3.41	±	0.3	1.98	±	0.2
MAE		3.11	±	0.2	1.94	±	0.2	2.86	±	0.3	1.60	±	0.2	2.76	±	0.3	1.62	±	0.2
r_{Spearman}				0.77			0.99			0.78			0.98			0.79			0.98

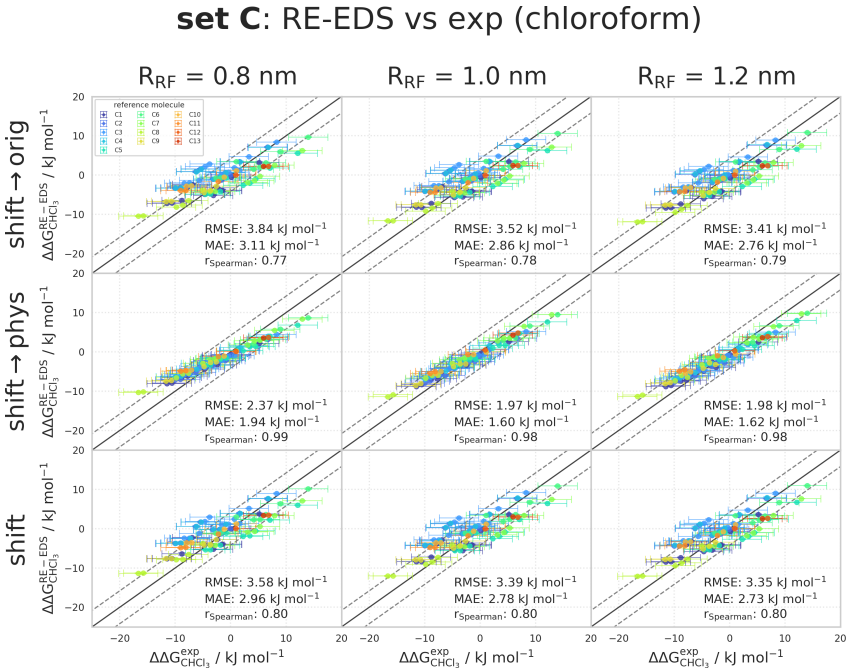


Figure 7.15: Comparison of the relative solvation free energies in chloroform of set C: $\Delta\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ obtained from RE-EDS calculations in GROMOS propagated with the AT$^{\text{shift}}$ scheme versus $\Delta\Delta G^{\text{exp}}_{\text{CHCl}_3}$ as reported in the Minnesota solvation [174] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\text{shift}\rightarrow\text{orig}}$ (top), $V^{\text{shift}\rightarrow\text{pyhs}}$ (middle), and the uncorrected electrostatic potential energy $V^{\text{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv. [155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.16.

Relative Hydration Free Energies of Set D

Table 7.17: $\Delta\Delta G_{\mathrm{hyd}}$ for the 13 molecules of set D calculated from RE-EDS simulations using $\mathrm{AT}^{\mathrm{shift}}$, corrected to $V^{\mathrm{ele,orig}}$ and $V^{\mathrm{ele,phys}}$, with $R_{\mathrm{RF}} = 0.8$ nm, $R_{\mathrm{RF}} = 1.0$ nm, and $R_{\mathrm{RF}} = 1.2$ nm. The results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions).

$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$

		$R_{\mathrm{RF}} = 0.8$ nm		$R_{\mathrm{RF}} = 1.0$ nm		$R_{\mathrm{RF}} = 1.2$ nm	
Molecules		shift→orig	shift→phys	shift→orig	shift→phys	shift→orig	shift→phys
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
D1	D2	11.6 ± 0.4	12.6 ± 0.2	11.4 ± 0.5	13.0 ± 0.3	11.6 ± 0.3	12.5 ± 0.2
D1	D3	0.7 ± 0.3	−1.3 ± 0.1	0.2 ± 0.4	−1.0 ± 0.2	0.4 ± 0.2	−1.2 ± 0.2
D1	D4	0.6 ± 0.3	1.0 ± 0.2	−0.1 ± 0.4	1.2 ± 0.2	−0.0 ± 0.3	0.4 ± 0.1
D1	D5	7.4 ± 0.3	4.7 ± 0.1	7.0 ± 0.4	5.5 ± 0.2	7.4 ± 0.2	4.8 ± 0.1
D1	D6	8.7 ± 0.2	5.4 ± 0.2	8.4 ± 0.7	5.9 ± 0.2	8.7 ± 0.4	5.6 ± 0.2
D1	D7	8.3 ± 0.3	6.0 ± 0.2	8.0 ± 0.6	5.8 ± 0.2	8.0 ± 0.3	5.5 ± 0.2
D1	D8	7.6 ± 0.3	7.6 ± 0.2	7.2 ± 0.4	8.1 ± 0.2	7.5 ± 0.3	7.2 ± 0.1
D1	D9	−19.7 ± 0.2	−21.3 ± 0.5	−20.4 ± 0.5	−21.5 ± 0.2	−20.5 ± 0.4	−22.5 ± 0.2
D1	D10	−2.0 ± 0.6	−3.2 ± 0.3	−3.0 ± 0.5	−2.6 ± 0.2	−2.9 ± 0.6	−3.7 ± 0.2
D1	D11	−1.5 ± 0.5	−2.9 ± 0.2	−2.3 ± 0.5	−2.4 ± 0.3	−2.0 ± 0.4	−2.9 ± 0.3
D1	D12	1.8 ± 0.3	3.5 ± 0.2	0.7 ± 0.4	2.4 ± 0.2	0.9 ± 0.2	1.6 ± 0.1
D1	D13	12.6 ± 0.3	16.0 ± 0.3	12.2 ± 0.7	15.9 ± 0.5	12.4 ± 0.8	15.5 ± 0.8
D2	D3	−10.8 ± 0.2	−14.0 ± 0.2	−11.2 ± 0.2	−14.0 ± 0.2	−11.2 ± 0.2	−13.7 ± 0.1
D2	D4	−11.0 ± 0.2	−11.7 ± 0.2	−11.5 ± 0.3	−11.8 ± 0.2	−11.7 ± 0.3	−12.1 ± 0.2
D2	D5	−4.2 ± 0.3	−8.0 ± 0.2	−4.4 ± 0.2	−7.5 ± 0.2	−4.2 ± 0.2	−7.7 ± 0.2
D2	D6	−2.8 ± 0.2	−7.3 ± 0.2	−3.0 ± 0.3	−7.1 ± 0.2	−2.9 ± 0.2	−6.8 ± 0.1
D2	D7	−3.3 ± 0.2	−6.6 ± 0.2	−3.4 ± 0.3	−7.3 ± 0.2	−3.7 ± 0.3	−7.0 ± 0.2
D2	D8	−4.0 ± 0.3	−5.0 ± 0.2	−4.2 ± 0.2	−5.0 ± 0.2	−4.1 ± 0.2	−5.3 ± 0.1
D2	D9	−31.3 ± 0.4	−33.9 ± 0.1	−31.8 ± 0.2	−34.5 ± 0.1	−32.1 ± 0.2	−35.0 ± 0.2
D2	D10	−13.6 ± 0.3	−15.9 ± 0.2	−14.4 ± 0.3	−15.7 ± 0.2	−14.5 ± 0.6	−16.2 ± 0.3
D2	D11	−13.0 ± 0.4	−15.5 ± 0.2	−13.7 ± 0.4	−15.4 ± 0.2	−13.7 ± 0.4	−15.3 ± 0.2
D2	D12	−9.8 ± 0.2	−9.1 ± 0.1	−10.7 ± 0.3	−10.6 ± 0.2	−10.7 ± 0.2	−10.9 ± 0.2
D2	D13	1.0 ± 0.4	3.3 ± 0.4	0.8 ± 0.4	2.8 ± 0.4	0.8 ± 0.7	3.0 ± 0.8
D3	D4	−0.2 ± 0.1	2.3 ± 0.1	−0.4 ± 0.2	2.2 ± 0.1	−0.5 ± 0.1	1.6 ± 0.1
D3	D5	6.6 ± 0.2	6.0 ± 0.0	6.8 ± 0.1	6.5 ± 0.1	7.0 ± 0.1	5.9 ± 0.1
D3	D6	8.0 ± 0.1	6.7 ± 0.0	8.2 ± 0.3	6.9 ± 0.0	8.2 ± 0.2	6.8 ± 0.1
D3	D7	7.6 ± 0.2	7.4 ± 0.1	7.8 ± 0.3	6.8 ± 0.0	7.5 ± 0.3	6.7 ± 0.1
D3	D8	6.8 ± 0.2	9.0 ± 0.0	6.4 ± 0.1	9.0 ± 0.1	7.0 ± 0.1	8.4 ± 0.1
D3	D9	−20.5 ± 0.3	−19.9 ± 0.1	−20.7 ± 0.2	−20.5 ± 0.1	−20.9 ± 0.2	−21.3 ± 0.0
D3	D10	−2.8 ± 0.4	−1.9 ± 0.2	−3.2 ± 0.2	−1.6 ± 0.1	−3.3 ± 0.5	−2.5 ± 0.3
D3	D11	−2.2 ± 0.3	−1.5 ± 0.1	−2.5 ± 0.2	−1.4 ± 0.2	−2.5 ± 0.3	−1.7 ± 0.2
D3	D12	1.1 ± 0.2	4.8 ± 0.1	0.4 ± 0.2	3.4 ± 0.1	0.4 ± 0.0	2.8 ± 0.1
D3	D13	11.8 ± 0.3	17.3 ± 0.3	12.0 ± 0.4	16.9 ± 0.4	12.0 ± 0.8	16.7 ± 0.8
D4	D5	6.8 ± 0.2	3.7 ± 0.1	7.2 ± 0.2	4.3 ± 0.1	7.4 ± 0.2	4.4 ± 0.1
D4	D6	8.2 ± 0.1	4.4 ± 0.1	8.5 ± 0.4	4.7 ± 0.1	8.7 ± 0.3	5.3 ± 0.1
D4	D7	7.7 ± 0.2	5.0 ± 0.1	8.1 ± 0.4	4.6 ± 0.1	8.0 ± 0.3	5.1 ± 0.1
D4	D8	7.0 ± 0.2	6.7 ± 0.1	7.3 ± 0.1	6.9 ± 0.1	7.5 ± 0.2	6.8 ± 0.1
D4	D9	−20.3 ± 0.4	−22.3 ± 0.1	−20.3 ± 0.2	−22.7 ± 0.1	−20.5 ± 0.2	−22.9 ± 0.1
D4	D10	−2.6 ± 0.4	−1.2 ± 0.2	−2.8 ± 0.2	−1.0 ± 0.1	−2.9 ± 0.5	−2.1 ± 0.3
D4	D11	−2.0 ± 0.3	−3.9 ± 0.2	−2.1 ± 0.3	−3.6 ± 0.2	−2.0 ± 0.4	−3.3 ± 0.2
D4	D12	1.2 ± 0.2	2.5 ± 0.1	0.8 ± 0.2	1.2 ± 0.1	0.9 ± 0.2	1.2 ± 0.1
D4	D13	12.0 ± 0.3	15.0 ± 0.3	12.4 ± 0.5	14.7 ± 0.4	12.5 ± 0.8	15.1 ± 0.8
D5	D6	1.4 ± 0.2	0.7 ± 0.1	1.4 ± 0.3	0.3 ± 0.0	1.3 ± 0.2	0.9 ± 0.1
D5	D7	0.9 ± 0.2	1.4 ± 0.1	1.0 ± 0.4	0.2 ± 0.0	0.6 ± 0.2	0.7 ± 0.0
D5	D8	0.2 ± 0.2	3.0 ± 0.1	0.1 ± 0.1	2.5 ± 0.1	0.1 ± 0.1	2.5 ± 0.0
D5	D9	−27.1 ± 0.2	−26.0 ± 0.1	−27.5 ± 0.2	−27.1 ± 0.1	−27.9 ± 0.2	−27.2 ± 0.1
D5	D10	−9.4 ± 0.5	−7.9 ± 0.2	−10.0 ± 0.1	−8.2 ± 0.1	−10.3 ± 0.5	−8.5 ± 0.2
D5	D11	−8.8 ± 0.3	−7.6 ± 0.1	−9.3 ± 0.3	−7.9 ± 0.2	−9.4 ± 0.3	−7.6 ± 0.1
D5	D12	−5.6 ± 0.1	−1.2 ± 0.1	−6.4 ± 0.1	−0.2 ± 0.1	−6.5 ± 0.1	−1.1 ± 0.1
D5	D13	5.2 ± 0.3	11.3 ± 0.4	5.2 ± 0.4	10.3 ± 0.4	5.0 ± 0.8	10.7 ± 0.8
D6	D7	−0.5 ± 0.2	0.7 ± 0.0	−0.4 ± 0.4	−0.1 ± 0.0	−0.7 ± 0.3	−0.2 ± 0.1
D6	D8	−1.2 ± 0.2	2.3 ± 0.0	−1.2 ± 0.3	2.2 ± 0.1	−1.2 ± 0.1	1.6 ± 0.0
D6	D9	−28.5 ± 0.3	−26.7 ± 0.1	−28.8 ± 0.3	−27.4 ± 0.1	−29.2 ± 0.2	−28.1 ± 0.1
D6	D10	−10.8 ± 0.4	−8.6 ± 0.2	−11.4 ± 0.3	−8.5 ± 0.1	−11.6 ± 0.6	−9.4 ± 0.2
D6	D11	−10.2 ± 0.3	−8.3 ± 0.1	−10.6 ± 0.4	−8.3 ± 0.1	−10.7 ± 0.3	−8.5 ± 0.1
D6	D12	−7.0 ± 0.2	−1.9 ± 0.1	−7.7 ± 0.4	−3.5 ± 0.1	−7.8 ± 0.3	−4.0 ± 0.1
D6	D13	3.8 ± 0.3	10.6 ± 0.3	3.8 ± 0.4	10.0 ± 0.4	3.7 ± 0.8	9.8 ± 0.8
D7	D8	−0.7 ± 0.3	1.6 ± 0.1	−0.8 ± 0.3	2.3 ± 0.1	−0.5 ± 0.2	1.7 ± 0.1
D7	D9	−28.0 ± 0.3	−27.3 ± 0.1	−28.4 ± 0.3	−27.3 ± 0.1	−28.4 ± 0.4	−28.0 ± 0.1
D7	D10	−10.3 ± 0.3	−9.3 ± 0.2	−11.0 ± 0.5	−8.4 ± 0.1	−10.8 ± 0.3	−9.2 ± 0.2
D7	D11	−9.8 ± 0.4	−8.9 ± 0.1	−10.2 ± 0.5	−8.1 ± 0.1	−10.0 ± 0.2	−8.4 ± 0.1
D7	D12	−6.5 ± 0.1	−2.5 ± 0.1	−7.3 ± 0.5	−3.4 ± 0.1	−7.1 ± 0.3	−3.9 ± 0.1
D7	D13	4.3 ± 0.3	9.9 ± 0.3	4.2 ± 0.6	10.1 ± 0.4	4.5 ± 0.9	10.0 ± 0.8
D8	D9	−27.3 ± 0.3	−28.9 ± 0.1	−27.6 ± 0.2	−29.6 ± 0.2	−28.0 ± 0.2	−29.7 ± 0.1
D8	D10	−9.6 ± 0.5	−10.9 ± 0.2	−10.2 ± 0.2	−10.7 ± 0.1	−10.4 ± 0.5	−10.9 ± 0.2
D8	D11	−9.0 ± 0.2	−10.5 ± 0.1	−9.4 ± 0.2	−10.4 ± 0.2	−9.5 ± 0.3	−10.1 ± 0.2
D8	D12	−5.8 ± 0.2	−4.1 ± 0.1	−6.5 ± 0.2	−5.7 ± 0.1	−6.6 ± 0.1	−5.6 ± 0.1
D8	D13	5.0 ± 0.3	8.3 ± 0.3	5.1 ± 0.4	7.8 ± 0.4	5.0 ± 0.7	8.3 ± 0.8
D9	D10	17.7 ± 0.6	18.1 ± 0.2	17.5 ± 0.3	18.9 ± 0.1	17.6 ± 0.6	18.8 ± 0.3
D9	D11	18.3 ± 0.4	18.4 ± 0.2	18.2 ± 0.4	19.1 ± 0.2	18.5 ± 0.4	19.6 ± 0.1
D9	D12	21.5 ± 0.3	24.8 ± 0.1	21.1 ± 0.3	23.9 ± 0.1	21.4 ± 0.2	24.1 ± 0.1
D9	D13	32.3 ± 0.4	37.3 ± 0.4	32.7 ± 0.4	37.4 ± 0.3	32.9 ± 0.7	38.0 ± 0.7

D10	D11	0.6	±	0.5	0.3	±	0.2	0.7	±	0.3	0.2	±	0.2	0.8	±	0.3	0.9	±	0.2
D10	D12	3.8	±	0.3	6.7	±	0.1	3.7	±	0.1	5.0	±	0.1	3.8	±	0.5	5.3	±	0.3
D10	D13	14.6	±	0.5	19.2	±	0.4	15.2	±	0.4	18.5	±	0.4	15.3	±	1.1	19.2	±	0.9
D11	D12	3.3	±	0.3	6.4	±	0.1	2.9	±	0.3	4.8	±	0.2	2.9	±	0.3	4.5	±	0.2
D11	D13	14.0	±	0.4	18.9	±	0.3	14.5	±	0.5	18.3	±	0.5	14.5	±	0.9	18.4	±	0.8
D12	D13	10.8	±	0.4	12.5	±	0.4	11.6	±	0.4	13.5	±	0.4	11.6	±	0.7	13.9	±	0.8
RMSE		5.78	±	0.6	5.39	±	0.6	6.05	±	0.6	5.54	±	0.6	6.09	±	0.6	5.68	±	0.5
MAE		4.84	±	0.6	4.43	±	0.6	5.05	±	0.5	4.57	±	0.5	5.09	±	0.6	4.70	±	0.5
r_{Spearman}				0.81			0.86			0.80			0.85			0.80			0.85

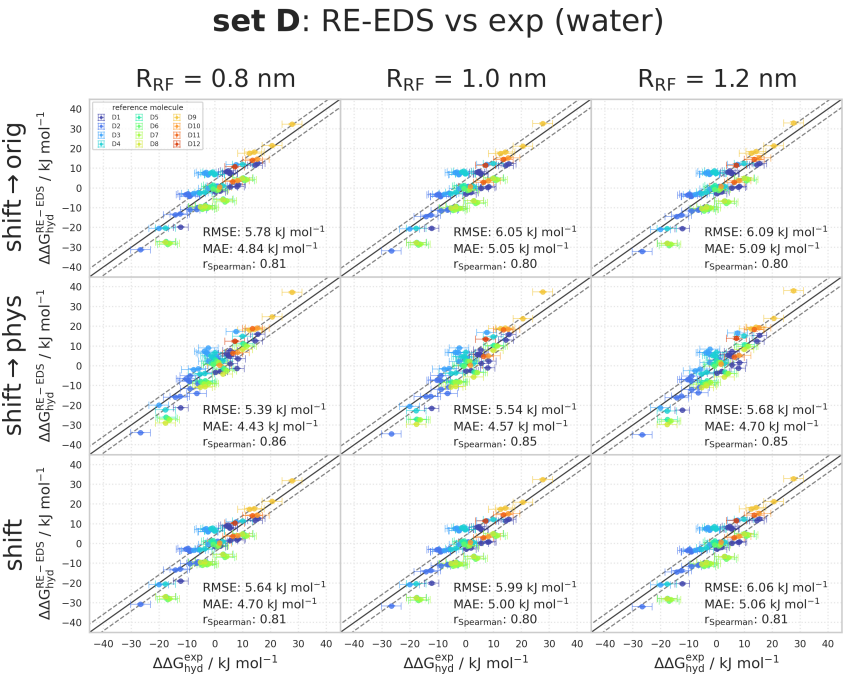


Figure 7.16: Comparison of the relative hydration free energies of set D: $\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS}}$ obtained from RE-EDS calculations in GROMOS propagated with the AT^{shift} scheme versus $\Delta\Delta G_{\text{hyd}}^{\text{exp}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\text{shift}\rightarrow\text{orig}}$ (top), $V^{\text{shift}\rightarrow\text{pyhs}}$ (middle), and the uncorrected electrostatic potential energy $V^{\text{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.17.

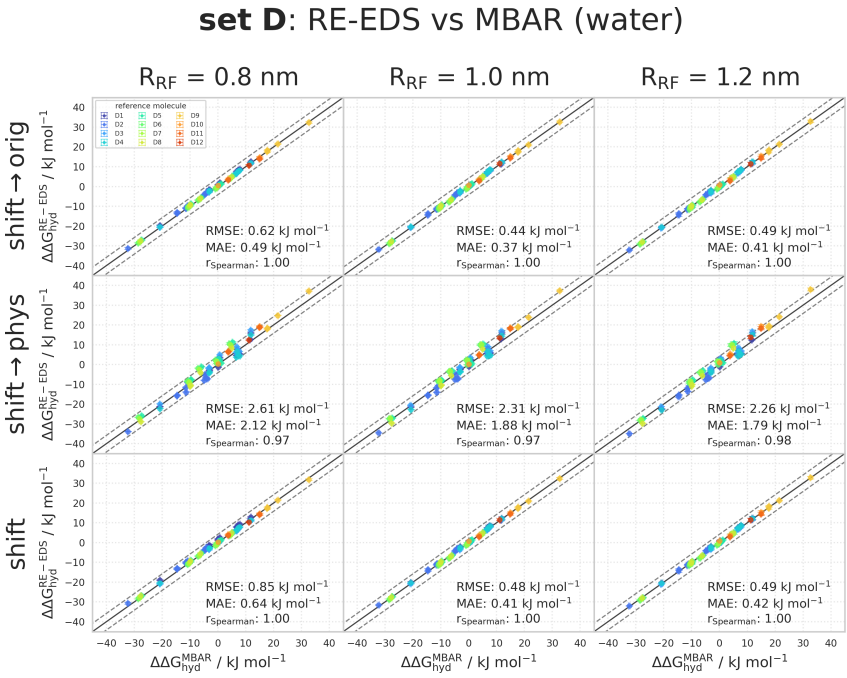


Figure 7.17: Comparison of the relative hydration free energies of set D: $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS}}$ obtained from RE-EDS calculations in GROMOS[27] propagated with the AT$^{\mathrm{shift}}$ scheme versus $\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ as reported in the FreeSolv[155,173] database. The three columns correspond to $R_{\mathrm{RF}} = 0.8$ nm (left), $R_{\mathrm{RF}} = 1.0$ nm (middle), and $R_{\mathrm{RF}} = 1.2$ nm (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\mathrm{shift}\rightarrow\mathrm{orig}}$ (top), $V^{\mathrm{shift}\rightarrow\mathrm{phys}}$ (middle), and the uncorrected electrostatic potential energy $V^{\mathrm{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within $\pm$ 4.184 kJ mol^{-1} ($\pm$ 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.17.

Relative Solvation Free Energies in Chloroform of Set D

Table 7.18: $\Delta\Delta G_{\mathrm{CHCl_3}}$ for the 13 molecules of set D calculated from RE-EDS simulations using AT$^{\mathrm{shift}}$, corrected to $V^{\mathrm{ele,orig}}$ and $V^{\mathrm{ele,phys}}$, with $R_{\mathrm{RF}} = 0.8$ nm, $R_{\mathrm{RF}} = 1.0$ nm, and $R_{\mathrm{RF}} = 1.2$ nm. The results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions).

$\Delta\Delta G^{\text{RE-EDS}}_{\mathrm{CHCl_3}}$

		$R_{\mathrm{RF}} = 0.8$ nm		$R_{\mathrm{RF}} = 1.0$ nm		$R_{\mathrm{RF}} = 1.2$ nm	
Molecules		shift→orig	shift→phys	shift→orig	shift→phys	shift→orig	shift→phys
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
D1	D2	0.9 ± 0.3	1.2 ± 0.2	1.0 ± 0.2	1.3 ± 0.2	1.1 ± 0.2	1.4 ± 0.2
D1	D3	−0.8 ± 0.1	−2.9 ± 0.1	−0.7 ± 0.1	−2.7 ± 0.1	−0.9 ± 0.1	−2.8 ± 0.2
D1	D4	−5.3 ± 0.2	−4.1 ± 0.1	−5.4 ± 0.2	−4.6 ± 0.2	−5.7 ± 0.3	−5.0 ± 0.2
D1	D5	2.9 ± 0.1	0.5 ± 0.1	3.0 ± 0.1	0.3 ± 0.1	3.1 ± 0.2	0.3 ± 0.1
D1	D6	2.7 ± 0.2	−0.6 ± 0.1	3.0 ± 0.3	−0.3 ± 0.1	2.9 ± 0.2	−0.6 ± 0.2
D1	D7	2.5 ± 0.2	−0.3 ± 0.1	2.9 ± 0.2	0.1 ± 0.1	2.7 ± 0.2	−0.2 ± 0.1
D1	D8	−0.8 ± 0.2	−0.6 ± 0.1	−0.7 ± 0.2	−0.6 ± 0.2	−0.9 ± 0.2	−0.6 ± 0.2
D1	D9	−6.5 ± 0.2	−8.4 ± 0.1	−6.9 ± 0.1	−8.7 ± 0.1	−7.4 ± 0.2	−9.5 ± 0.2
D1	D10	−2.7 ± 0.2	−3.2 ± 0.1	−3.0 ± 0.1	−3.4 ± 0.1	−3.5 ± 0.3	−4.0 ± 0.2
D1	D11	−2.7 ± 0.2	−3.4 ± 0.2	−3.4 ± 0.2	−3.9 ± 0.2	−3.3 ± 0.3	−4.2 ± 0.2
D1	D12	−8.5 ± 0.2	−7.0 ± 0.1	−8.9 ± 0.2	−7.8 ± 0.2	−9.2 ± 0.2	−8.4 ± 0.2
D1	D13	−2.8 ± 0.3	−0.6 ± 0.3	−2.7 ± 0.3	−1.0 ± 0.3	−2.6 ± 0.7	−1.2 ± 0.7
D2	D3	−1.7 ± 0.3	−4.1 ± 0.2	−1.7 ± 0.2	−3.9 ± 0.2	−1.9 ± 0.2	−4.2 ± 0.2
D2	D4	−6.2 ± 0.2	−5.4 ± 0.2	−6.5 ± 0.2	−5.8 ± 0.2	−6.8 ± 0.3	−6.4 ± 0.3
D2	D5	2.0 ± 0.2	−0.8 ± 0.2	1.9 ± 0.2	−1.0 ± 0.2	2.0 ± 0.2	−1.1 ± 0.2
D2	D6	1.8 ± 0.4	−1.8 ± 0.2	2.0 ± 0.4	−1.6 ± 0.2	1.9 ± 0.2	−2.0 ± 0.2
D2	D7	1.6 ± 0.3	−1.6 ± 0.2	1.8 ± 0.2	−1.1 ± 0.2	1.7 ± 0.2	−1.6 ± 0.2
D2	D8	−1.7 ± 0.2	−1.8 ± 0.1	−1.8 ± 0.2	−1.8 ± 0.2	−2.0 ± 0.2	−2.1 ± 0.2
D2	D9	−7.4 ± 0.3	−9.8 ± 0.1	−8.0 ± 0.1	−9.9 ± 0.1	−8.4 ± 0.2	−10.9 ± 0.2
D2	D10	−3.5 ± 0.3	−4.4 ± 0.2	−4.0 ± 0.2	−4.7 ± 0.2	−4.6 ± 0.3	−5.5 ± 0.2
D2	D11	−3.6 ± 0.4	−4.6 ± 0.2	−4.5 ± 0.3	−5.2 ± 0.2	−4.4 ± 0.3	−5.6 ± 0.2
D2	D12	−9.3 ± 0.4	−8.2 ± 0.2	−9.9 ± 0.2	−9.1 ± 0.2	−10.3 ± 0.2	−9.8 ± 0.2
D2	D13	−3.7 ± 0.4	−1.9 ± 0.4	−3.7 ± 0.3	−2.3 ± 0.3	−3.7 ± 0.7	−2.6 ± 0.7
D3	D4	−4.5 ± 0.1	−1.3 ± 0.1	−4.8 ± 0.2	−1.9 ± 0.1	−4.8 ± 0.3	−2.2 ± 0.2
D3	D5	3.7 ± 0.1	3.3 ± 0.0	3.6 ± 0.1	2.9 ± 0.1	3.9 ± 0.1	3.1 ± 0.0
D3	D6	3.5 ± 0.1	2.3 ± 0.1	3.7 ± 0.3	2.3 ± 0.0	3.8 ± 0.1	2.2 ± 0.1
D3	D7	3.3 ± 0.1	2.5 ± 0.1	3.5 ± 0.2	2.8 ± 0.1	3.6 ± 0.1	2.6 ± 0.1
D3	D8	0.0 ± 0.2	2.3 ± 0.1	−0.0 ± 0.2	2.1 ± 0.2	−0.1 ± 0.2	2.1 ± 0.1
D3	D9	−5.7 ± 0.1	−5.5 ± 0.1	−6.3 ± 0.2	−6.0 ± 0.1	−6.5 ± 0.1	−6.7 ± 0.1
D3	D10	−1.8 ± 0.1	−0.3 ± 0.1	−2.3 ± 0.1	−0.7 ± 0.0	−2.7 ± 0.3	−1.2 ± 0.2
D3	D11	−1.9 ± 0.2	−0.5 ± 0.2	−2.8 ± 0.3	−1.2 ± 0.2	−2.4 ± 0.2	−1.4 ± 0.1
D3	D12	−7.6 ± 0.2	−4.1 ± 0.1	−8.2 ± 0.1	−5.2 ± 0.1	−8.3 ± 0.1	−5.6 ± 0.1
D3	D13	−2.0 ± 0.3	2.2 ± 0.3	−2.0 ± 0.3	1.7 ± 0.3	−1.7 ± 0.7	1.6 ± 0.7
D4	D5	8.2 ± 0.1	4.6 ± 0.1	8.4 ± 0.2	4.8 ± 0.1	8.8 ± 0.2	5.3 ± 0.2
D4	D6	8.0 ± 0.3	3.6 ± 0.1	8.5 ± 0.3	4.2 ± 0.1	8.6 ± 0.2	4.4 ± 0.2
D4	D7	7.8 ± 0.2	3.8 ± 0.1	8.3 ± 0.2	4.7 ± 0.2	8.4 ± 0.3	4.8 ± 0.2
D4	D8	4.5 ± 0.2	3.5 ± 0.1	4.7 ± 0.3	4.0 ± 0.2	4.8 ± 0.3	4.4 ± 0.2
D4	D9	−1.2 ± 0.2	−4.2 ± 0.1	−1.5 ± 0.2	−4.1 ± 0.2	−1.7 ± 0.3	−4.5 ± 0.2
D4	D10	2.6 ± 0.2	1.0 ± 0.2	2.4 ± 0.2	1.1 ± 0.2	2.2 ± 0.3	1.0 ± 0.3
D4	D11	2.6 ± 0.3	1.0 ± 0.2	2.3 ± 0.3	0.7 ± 0.2	2.3 ± 0.8	0.8 ± 0.2
D4	D12	−3.2 ± 0.3	−2.9 ± 0.2	−3.5 ± 0.1	−3.3 ± 0.1	−3.5 ± 0.2	−3.4 ± 0.2
D4	D13	2.5 ± 0.3	3.5 ± 0.3	2.8 ± 0.4	3.6 ± 0.3	3.1 ± 0.8	3.8 ± 0.8
D5	D6	−0.2 ± 0.2	−1.0 ± 0.0	0.1 ± 0.3	−0.6 ± 0.0	−0.1 ± 0.1	−0.9 ± 0.1
D5	D7	−0.4 ± 0.2	−0.8 ± 0.1	−0.1 ± 0.1	−0.2 ± 0.1	−0.3 ± 0.1	−0.5 ± 0.1
D5	D8	−3.7 ± 0.2	−1.1 ± 0.1	−3.7 ± 0.1	−0.8 ± 0.1	−4.0 ± 0.1	−0.9 ± 0.1
D5	D9	−9.4 ± 0.1	−8.8 ± 0.1	−9.9 ± 0.1	−8.9 ± 0.1	−10.4 ± 0.1	−9.8 ± 0.1
D5	D10	−5.6 ± 0.1	−3.7 ± 0.1	−6.0 ± 0.1	−3.7 ± 0.1	−6.6 ± 0.3	−4.3 ± 0.2
D5	D11	−5.6 ± 0.2	−3.8 ± 0.1	−6.4 ± 0.2	−4.2 ± 0.2	−6.4 ± 0.2	−4.5 ± 0.1
D5	D12	−11.4 ± 0.2	−7.5 ± 0.1	−11.9 ± 0.2	−8.1 ± 0.1	−12.2 ± 0.1	−8.7 ± 0.1
D5	D13	−5.7 ± 0.3	−1.1 ± 0.4	−5.6 ± 0.4	−1.3 ± 0.4	−5.7 ± 0.7	−1.5 ± 0.7
D6	D7	−0.2 ± 0.1	0.2 ± 0.0	−0.2 ± 0.4	0.5 ± 0.1	−0.2 ± 0.2	0.4 ± 0.1
D6	D8	−3.5 ± 0.2	−0.1 ± 0.1	−3.8 ± 0.5	−0.2 ± 0.1	−3.9 ± 0.2	−0.0 ± 0.1
D6	D9	−9.2 ± 0.2	−7.8 ± 0.1	−10.0 ± 0.3	−8.3 ± 0.1	−10.3 ± 0.1	−8.9 ± 0.1
D6	D10	−5.4 ± 0.2	−2.6 ± 0.1	−6.1 ± 0.3	−3.1 ± 0.0	−6.4 ± 0.3	−3.4 ± 0.3
D6	D11	−5.4 ± 0.1	−2.8 ± 0.1	−6.5 ± 0.4	−3.6 ± 0.2	−6.2 ± 0.2	−3.6 ± 0.1
D6	D12	−11.2 ± 0.1	−6.4 ± 0.1	−11.9 ± 0.3	−7.5 ± 0.1	−12.1 ± 0.1	−7.8 ± 0.1
D6	D13	−5.5 ± 0.4	−0.1 ± 0.3	−5.7 ± 0.4	−0.7 ± 0.4	−5.5 ± 0.7	−0.6 ± 0.7
D7	D8	−3.3 ± 0.2	−0.3 ± 0.1	−3.6 ± 0.1	−0.7 ± 0.1	−3.7 ± 0.1	−0.5 ± 0.1
D7	D9	−9.0 ± 0.1	−8.0 ± 0.1	−9.8 ± 0.2	−8.8 ± 0.1	−10.1 ± 0.1	−9.3 ± 0.1
D7	D10	−5.2 ± 0.2	−2.8 ± 0.2	−5.9 ± 0.2	−3.5 ± 0.1	−6.2 ± 0.3	−3.9 ± 0.2
D7	D11	−5.2 ± 0.1	−3.0 ± 0.1	−6.3 ± 0.2	−4.0 ± 0.2	−6.0 ± 0.2	−4.0 ± 0.2
D7	D12	−11.0 ± 0.2	−6.7 ± 0.1	−11.8 ± 0.2	−8.0 ± 0.2	−11.9 ± 0.2	−8.2 ± 0.2
D7	D13	−5.3 ± 0.4	−0.3 ± 0.4	−5.5 ± 0.4	−1.1 ± 0.4	−5.3 ± 0.7	−1.0 ± 0.7
D8	D9	−5.7 ± 0.2	−7.7 ± 0.1	−6.2 ± 0.2	−8.1 ± 0.2	−6.4 ± 0.2	−8.9 ± 0.1
D8	D10	−1.8 ± 0.3	−2.6 ± 0.2	−2.3 ± 0.2	−2.9 ± 0.2	−2.6 ± 0.2	−3.4 ± 0.2
D8	D11	−1.9 ± 0.2	−2.8 ± 0.2	−2.7 ± 0.2	−3.3 ± 0.2	−2.4 ± 0.2	−3.5 ± 0.2
D8	D12	−7.6 ± 0.3	−6.4 ± 0.1	−8.2 ± 0.2	−7.3 ± 0.1	−8.3 ± 0.2	−7.8 ± 0.1
D8	D13	−2.0 ± 0.3	−0.0 ± 0.3	−1.9 ± 0.4	−0.4 ± 0.4	−1.7 ± 0.7	−0.5 ± 0.7
D9	D10	3.9 ± 0.1	5.2 ± 0.1	3.9 ± 0.1	5.3 ± 0.1	3.8 ± 0.3	5.5 ± 0.2
D9	D11	3.8 ± 0.2	5.0 ± 0.2	3.5 ± 0.2	4.8 ± 0.2	4.1 ± 0.2	5.3 ± 0.1
D9	D12	−2.0 ± 0.2	1.4 ± 0.1	−2.0 ± 0.2	0.8 ± 0.2	−1.8 ± 0.1	1.1 ± 0.1

D9	D13	3.7	±	0.4	7.7	±	0.4	4.3	±	0.3	7.7	±	0.3	4.8	±	0.7	8.3	±	0.7
D10	D11	−0.1	±	0.3	−0.2	±	0.2	−0.4	±	0.2	−0.5	±	0.2	0.2	±	0.3	−0.1	±	0.3
D10	D12	−5.8	±	0.3	−3.8	±	0.2	−5.9	±	0.2	−4.4	±	0.2	−5.7	±	0.3	−4.4	±	0.3
D10	D13	−0.2	±	0.4	2.6	±	0.4	0.4	±	0.3	2.4	±	0.3	0.9	±	0.7	2.8	±	0.7
D11	D12	−5.7	±	0.2	−3.6	±	0.1	−5.5	±	0.3	−3.9	±	0.2	−5.9	±	0.2	−4.2	±	0.1
D11	D13	−0.1	±	0.3	2.7	±	0.3	0.8	±	0.5	2.9	±	0.5	0.7	±	0.7	3.0	±	0.7
D12	D13	5.6	±	0.4	6.4	±	0.4	6.2	±	0.4	6.8	±	0.4	6.6	±	0.7	7.2	±	0.7
RMSE		4.94	±	0.5	3.57	±	0.4	5.01	±	0.5	3.73	±	0.3	4.94	±	0.4	3.63	±	0.3
MAE		4.11	±	0.4	2.98	±	0.3	4.16	±	0.5	3.10	±	0.3	4.10	±	0.4	3.00	±	0.3
r_{Spearman}		0.53			0.72			0.56			0.70			0.60			0.71		

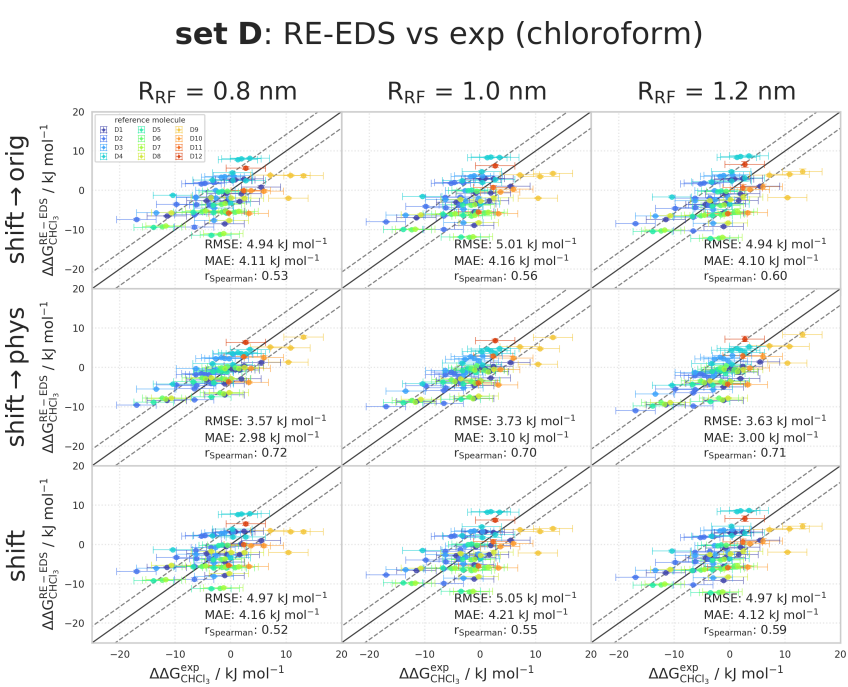


Figure 7.18: Comparison of the relative solvation free energies in chloroform of set D: $\Delta\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ obtained from RE-EDS calculations in GROMOS propagated with the AT$^{\text{shift}}$ scheme versus $\Delta\Delta G^{\text{exp}}_{\text{CHCl}_3}$ as reported in the Minnesota solvation [174] database. The three columns correspond to $R_{\text{RF}} = 0.8$ nm (left), $R_{\text{RF}} = 1.0$ nm (middle), and $R_{\text{RF}} = 1.2$ nm (right), respectively. The three rows correspond to using the corrected electrostatic potential energy $V^{\text{shift}\to\text{orig}}$ (top), $V^{\text{shift}\to\text{pyhs}}$ (middle), and the uncorrected electrostatic potential energy $V^{\text{ele,shift}}$ (bottom), respectively. The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The results obtained from RE-EDS were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv. [155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.18.

7.A.3 RE-EDS with OpenMM

Relative Hydration Free Energies of Set A

Table 7.19: $\Delta\Delta G_{\mathrm{hyd}}$ for the six molecules of set A from experiment,[155,173] calculated from the hydration free energies obtained with MBAR,[155,173] and from the RE-EDS calculations in GROMOS[27] and OpenMM[29] (with the $\mathrm{AT}^{\mathrm{shift}}$ scheme and $R_{\mathrm{RF}} = 1.2$ nm). The RE-EDS results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to four molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules i	Molecules j	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$ [155,173] [kJ mol^{-1}]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ [155,173] [kJ mol^{-1}]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,GROMOS}}$ [kJ mol^{-1}]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,OpenMM}}$ [kJ mol^{-1}]
A1	A2	-1.4 ± 3.6	2.3 ± 0.1	1.9 ± 0.3	1.9 ± 0.3
A1	A3	-7.5 ± 3.6	-6.4 ± 0.2	-6.8 ± 0.5	-6.7 ± 0.3
A1	A4	13.3 ± 2.5	15.1 ± 0.1	13.3 ± 0.3	13.5 ± 0.3
A1	A5	18.9 ± 3.6	24.8 ± 0.1	23.0 ± 0.2	23.1 ± 0.3
A1	A6	18.3 ± 2.5	22.3 ± 0.1	21.4 ± 0.2	21.2 ± 0.2
A2	A3	-6.2 ± 3.6	-8.7 ± 0.2	-8.7 ± 0.5	-8.6 ± 0.2
A2	A4	14.7 ± 2.5	12.8 ± 0.1	11.5 ± 0.3	11.6 ± 0.1
A2	A5	20.3 ± 3.6	22.5 ± 0.1	21.2 ± 0.1	21.2 ± 0.1
A2	A6	19.7 ± 2.5	20.0 ± 0.1	19.6 ± 0.1	19.3 ± 0.2
A3	A4	20.9 ± 2.5	21.5 ± 0.2	20.1 ± 0.4	20.2 ± 0.2
A3	A5	26.4 ± 3.6	31.2 ± 0.2	29.8 ± 0.5	29.8 ± 0.2
A3	A6	25.8 ± 2.5	28.7 ± 0.2	28.2 ± 0.5	27.9 ± 0.3
A4	A5	5.5 ± 2.5	9.7 ± 0.1	9.7 ± 0.3	9.6 ± 0.2
A4	A6	4.9 ± 0.1	7.1 ± 0.1	8.1 ± 0.2	7.7 ± 0.3
A5	A6	-0.6 ± 2.5	-2.6 ± 0.1	-1.6 ± 0.1	-1.9 ± 0.1
RMSE			3.08 ± 0.4	2.61 ± 0.3	2.53 ± 0.3
MAE			2.67 ± 0.3	2.20 ± 0.3	2.16 ± 0.3
r_{Spearman}			0.94	0.93	0.93
$t_{\mathrm{preparation}}$			18 ns	89.3 ns	89.3 ns
$t_{\mathrm{production}}$			600 ns	13.5 ns	13.5 ns

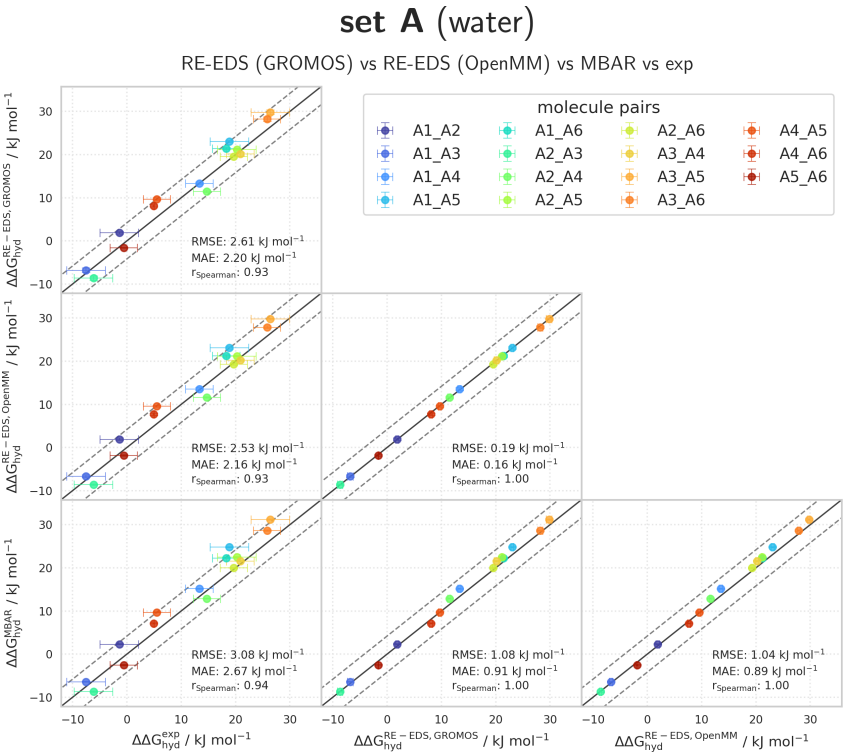


Figure 7.19: Comparison of the relative hydration free energies of set A. The pairwise comparisons of the relative hydration free energies calculated with RE-EDS in GROMOS[27] ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$), RE-EDS in OpenMM[29] ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$), MBAR ($\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$)[155,173] and the experimental results.[155,173] The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The numerical values are provided in Table 7.19.

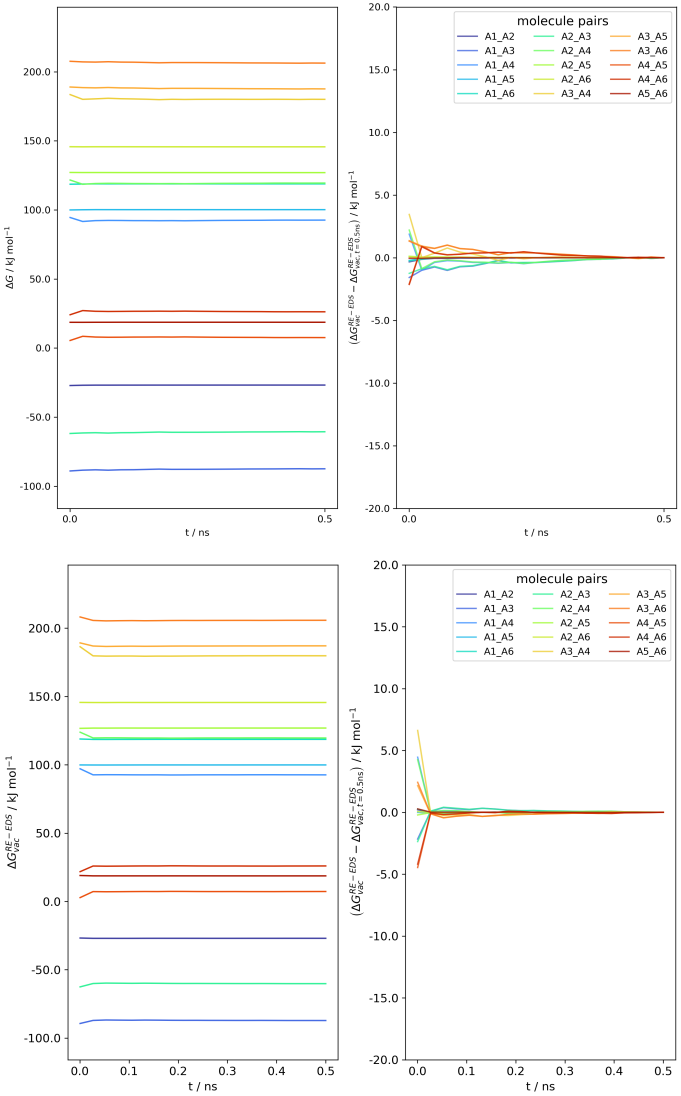


Figure 7.20: Convergence of $\Delta G_{\text{vac}}^{\text{RE-EDS}}$ as a function of the simulation time for the RE-EDS simulation of set A ($s = 1.0$) in GROMOS[27] (top) and in OpenMM[29] (bottom) from one of the five repeats used to calculate the relative hydration free energies reported in Table 7.19.

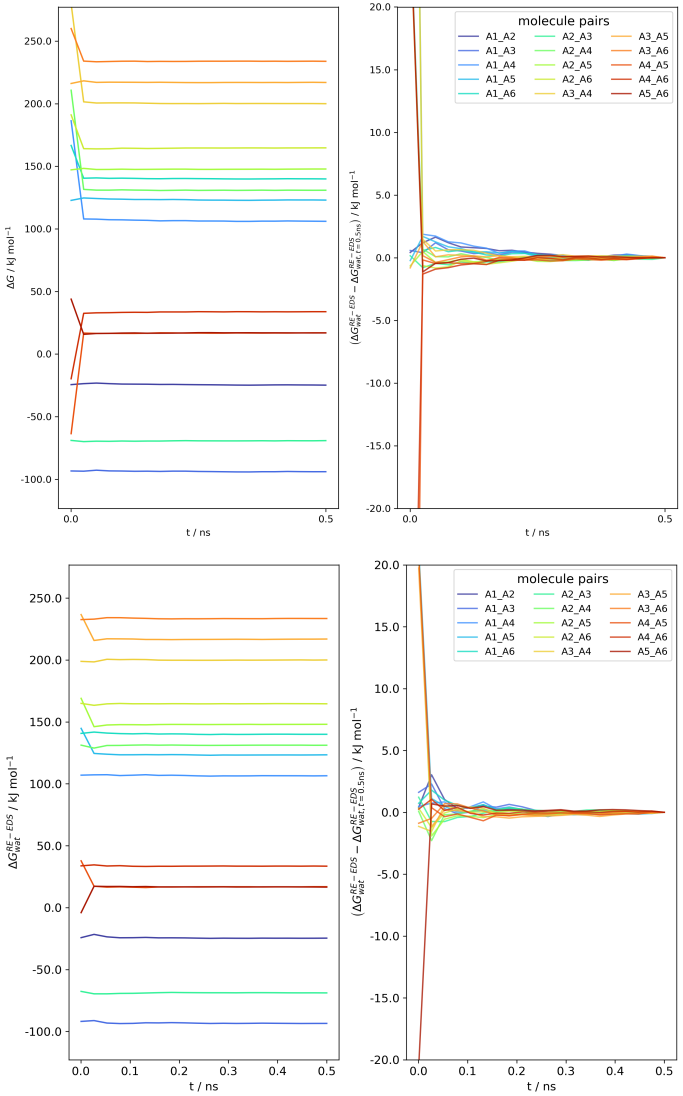


Figure 7.21: Convergence of $\Delta G_{\mathrm{wat}}^{\mathrm{RE\text{-}EDS}}$ as a function of the simulation time for the RE-EDS simulation of set A ($s = 1.0$) in GROMOS[27] (top) and in OpenMM[29] (bottom) from one of the five repeats used to calculate the relative hydration free energies reported in Table 7.19.

Relative Hydration Free Energies of Set C

Table 7.20: $\Delta\Delta G_{\mathrm{hyd}}$ for the 14 molecules of set C from experiment,[155,173] calculated from the hydration free energies obtained with MBAR,[155,173] and from the RE-EDS calculations in GROMOS[27] and OpenMM[29] (with the AT$^{\mathrm{shift}}$ scheme and $R_{\mathrm{RF}} = 1.2$ nm). The RE-EDS results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$ [155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ [155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,GROMOS}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,OpenMM}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	0.5 ± 2.6	0.8 ± 0.2	0.1 ± 0.2	0.6 ± 0.1
C1	C3	0.0 ± 2.6	−0.3 ± 0.2	−1.2 ± 0.1	−0.3 ± 0.1
C1	C4	0.3 ± 2.6	0.4 ± 0.2	0.4 ± 0.1	0.4 ± 0.1
C1	C5	−23.8 ± 1.2	−20.6 ± 0.2	−20.6 ± 0.2	−20.7 ± 0.2
C1	C6	−21.9 ± 2.6	−20.0 ± 0.2	−20.1 ± 0.2	−19.9 ± 0.3
C1	C7	−19.2 ± 2.6	−19.9 ± 0.2	−19.9 ± 0.1	−19.8 ± 0.1
C1	C8	0.4 ± 2.6	3.1 ± 0.2	3.1 ± 0.1	2.9 ± 0.1
C1	C9	−0.9 ± 2.6	1.3 ± 0.2	1.4 ± 0.2	1.4 ± 0.1
C1	C10	−0.5 ± 2.6	2.5 ± 0.2	2.6 ± 0.1	2.8 ± 0.0
C1	C11	−2.3 ± 2.6	−0.7 ± 0.2	−0.6 ± 0.1	−0.7 ± 0.0
C1	C12	−19.4 ± 2.6	−19.0 ± 0.2	−19.7 ± 0.2	−19.3 ± 0.1
C1	C13	−19.5 ± 2.6	−19.7 ± 0.2	−20.0 ± 0.4	−19.6 ± 0.1
C1	C14	−3.5 ± 2.6	−1.1 ± 0.2	−0.5 ± 0.2	−0.6 ± 0.1
C2	C3	−0.5 ± 3.6	−1.0 ± 0.2	−1.3 ± 0.2	−0.9 ± 0.1
C2	C4	−0.2 ± 3.6	−0.4 ± 0.2	0.3 ± 0.2	−0.2 ± 0.1
C2	C5	−24.3 ± 2.6	−21.3 ± 0.2	−20.8 ± 0.2	−21.3 ± 0.1
C2	C6	−22.3 ± 3.6	−20.8 ± 0.2	−20.3 ± 0.2	−20.5 ± 0.3
C2	C7	−19.7 ± 3.6	−20.6 ± 0.2	−20.0 ± 0.2	−20.4 ± 0.1
C2	C8	−0.0 ± 3.6	2.4 ± 0.2	3.0 ± 0.1	2.3 ± 0.1
C2	C9	−1.4 ± 3.6	0.6 ± 0.2	1.3 ± 0.1	0.8 ± 0.1
C2	C10	−0.9 ± 3.6	1.8 ± 0.2	2.5 ± 0.1	2.1 ± 0.1
C2	C11	−2.8 ± 3.6	−1.4 ± 0.2	−0.8 ± 0.1	−1.3 ± 0.1
C2	C12	−19.8 ± 3.6	−19.7 ± 0.2	−19.9 ± 0.3	−19.9 ± 0.1
C2	C13	−20.0 ± 3.6	−20.4 ± 0.2	−20.2 ± 0.4	−20.2 ± 0.2
C2	C14	−4.0 ± 3.6	−1.9 ± 0.2	−0.6 ± 0.1	−1.2 ± 0.1
C3	C4	0.3 ± 3.6	0.6 ± 0.2	1.6 ± 0.2	0.7 ± 0.2
C3	C5	−23.8 ± 2.6	−20.3 ± 0.2	−19.5 ± 0.2	−20.4 ± 0.2
C3	C6	−21.9 ± 3.6	−19.8 ± 0.2	−19.0 ± 0.2	−19.6 ± 0.3
C3	C7	−19.2 ± 3.6	−19.6 ± 0.2	−18.7 ± 0.2	−19.5 ± 0.2
C3	C8	0.4 ± 3.6	3.4 ± 0.2	4.3 ± 0.1	3.2 ± 0.2
C3	C9	−0.9 ± 3.6	1.6 ± 0.2	2.6 ± 0.2	1.7 ± 0.2
C3	C10	−0.5 ± 3.6	2.8 ± 0.2	3.8 ± 0.1	3.1 ± 0.1
C3	C11	−2.3 ± 3.6	−0.4 ± 0.2	0.5 ± 0.1	−0.4 ± 0.1
C3	C12	−19.4 ± 3.6	−18.7 ± 0.2	−18.6 ± 0.2	−18.9 ± 0.1
C3	C13	−19.5 ± 3.6	−19.4 ± 0.2	−18.9 ± 0.3	−19.3 ± 0.1
C3	C14	−3.5 ± 3.6	−0.9 ± 0.2	0.7 ± 0.1	−0.3 ± 0.2
C4	C5	−24.1 ± 2.6	−21.0 ± 0.2	−21.0 ± 0.1	−21.1 ± 0.2
C4	C6	−22.2 ± 3.6	−20.4 ± 0.2	−20.6 ± 0.2	−20.3 ± 0.2
C4	C7	−19.5 ± 3.6	−20.3 ± 0.2	−20.3 ± 0.1	−20.2 ± 0.1
C4	C8	0.1 ± 3.6	2.8 ± 0.2	2.7 ± 0.2	2.5 ± 0.1
C4	C9	−1.2 ± 3.6	1.0 ± 0.2	1.0 ± 0.2	1.0 ± 0.1
C4	C10	−0.8 ± 3.6	2.1 ± 0.2	2.2 ± 0.1	2.4 ± 0.0
C4	C11	−2.6 ± 3.6	−1.0 ± 0.2	−1.1 ± 0.2	−1.1 ± 0.1
C4	C12	−19.7 ± 3.6	−19.4 ± 0.2	−20.2 ± 0.3	−19.6 ± 0.1
C4	C13	−19.8 ± 3.6	−20.0 ± 0.2	−20.4 ± 0.4	−20.0 ± 0.2
C4	C14	−3.8 ± 3.6	−1.5 ± 0.2	−0.9 ± 0.1	−1.0 ± 0.1
C5	C6	2.0 ± 2.6	0.5 ± 0.2	0.5 ± 0.2	0.8 ± 0.3
C5	C7	4.6 ± 2.6	0.7 ± 0.2	0.7 ± 0.2	0.9 ± 0.1
C5	C8	24.3 ± 2.6	23.7 ± 0.2	23.8 ± 0.2	23.6 ± 0.1
C5	C9	22.9 ± 2.6	21.9 ± 0.2	22.0 ± 0.2	22.1 ± 0.1
C5	C10	23.4 ± 2.6	23.1 ± 0.2	23.2 ± 0.1	23.5 ± 0.2
C5	C11	21.5 ± 2.6	19.9 ± 0.2	20.0 ± 0.2	20.0 ± 0.1
C5	C12	4.5 ± 2.6	1.6 ± 0.2	0.9 ± 0.4	1.5 ± 0.2
C5	C13	4.3 ± 2.6	0.9 ± 0.2	0.6 ± 0.5	1.1 ± 0.2
C5	C14	20.3 ± 2.6	19.5 ± 0.2	20.2 ± 0.2	20.1 ± 0.1
C6	C7	2.7 ± 3.6	0.2 ± 0.2	0.3 ± 0.3	0.1 ± 0.3
C6	C8	22.3 ± 3.6	23.2 ± 0.2	23.3 ± 0.2	22.8 ± 0.3
C6	C9	21.0 ± 3.6	21.4 ± 0.2	21.5 ± 0.3	21.3 ± 0.3
C6	C10	21.4 ± 3.6	22.6 ± 0.2	22.8 ± 0.2	22.7 ± 0.3
C6	C11	19.5 ± 3.6	19.4 ± 0.2	19.5 ± 0.2	19.2 ± 0.3
C6	C12	2.5 ± 3.6	1.0 ± 0.2	0.4 ± 0.3	0.6 ± 0.2
C6	C13	2.3 ± 3.6	0.4 ± 0.2	0.1 ± 0.4	0.3 ± 0.4
C6	C14	18.4 ± 3.6	18.9 ± 0.2	19.7 ± 0.2	19.3 ± 0.3
C7	C8	19.6 ± 3.6	23.0 ± 0.2	23.0 ± 0.1	22.7 ± 0.1
C7	C9	18.3 ± 3.6	21.2 ± 0.2	21.3 ± 0.2	21.2 ± 0.1
C7	C10	18.7 ± 3.6	22.4 ± 0.2	22.5 ± 0.1	22.6 ± 0.1
C7	C11	16.9 ± 3.6	19.2 ± 0.2	19.2 ± 0.2	19.1 ± 0.1
C7	C12	−0.2 ± 3.6	0.9 ± 0.2	0.1 ± 0.3	0.6 ± 0.2
C7	C13	−0.3 ± 3.6	0.2 ± 0.2	−0.1 ± 0.4	0.3 ± 0.2
C7	C14	15.7 ± 3.6	18.7 ± 0.2	19.4 ± 0.2	19.2 ± 0.1

C8	C9	−1.3 ± 3.6	−1.8 ± 0.1	−1.8 ± 0.1	−1.6 ± 0.1
C8	C10	−0.9 ± 3.6	−0.6 ± 0.2	−0.5 ± 0.1	−0.2 ± 0.2
C8	C11	−2.8 ± 3.6	−3.8 ± 0.2	−3.8 ± 0.0	−3.6 ± 0.1
C8	C12	−19.8 ± 3.6	−22.1 ± 0.2	−22.9 ± 0.2	−22.2 ± 0.2
C8	C13	−20.0 ± 3.6	−22.8 ± 0.2	−23.2 ± 0.3	−22.5 ± 0.2
C8	C14	−3.9 ± 3.6	−4.3 ± 0.2	−3.6 ± 0.1	−3.5 ± 0.1
C9	C10	0.5 ± 3.6	1.2 ± 0.2	1.2 ± 0.1	1.4 ± 0.1
C9	C11	−1.4 ± 3.6	−2.0 ± 0.2	−2.0 ± 0.1	−2.1 ± 0.1
C9	C12	−18.5 ± 3.6	−20.3 ± 0.2	−21.1 ± 0.3	−20.6 ± 0.1
C9	C13	−18.6 ± 3.6	−21.0 ± 0.2	−21.4 ± 0.4	−20.9 ± 0.2
C9	C14	−2.6 ± 3.6	−2.5 ± 0.2	−1.9 ± 0.1	−2.0 ± 0.1
C10	C11	−1.9 ± 3.6	−3.2 ± 0.2	−3.3 ± 0.1	−3.5 ± 0.1
C10	C12	−18.9 ± 3.6	−21.5 ± 0.2	−22.4 ± 0.3	−22.0 ± 0.1
C10	C13	−19.1 ± 3.6	−22.2 ± 0.2	−22.6 ± 0.4	−22.3 ± 0.1
C10	C14	−3.1 ± 3.6	−3.6 ± 0.2	−3.1 ± 0.1	−3.4 ± 0.1
C11	C12	−17.0 ± 3.6	−18.3 ± 0.2	−19.1 ± 0.2	−18.5 ± 0.1
C11	C13	−17.2 ± 3.6	−19.0 ± 0.2	−19.4 ± 0.3	−18.8 ± 0.1
C11	C14	−1.2 ± 3.6	−0.5 ± 0.2	0.2 ± 0.1	0.1 ± 0.0
C12	C13	−0.2 ± 3.6	−0.7 ± 0.2	−0.3 ± 0.2	−0.3 ± 0.2
C12	C14	15.9 ± 3.6	17.9 ± 0.2	19.3 ± 0.3	18.6 ± 0.1
C13	C14	16.0 ± 3.6	18.5 ± 0.2	19.6 ± 0.4	18.9 ± 0.1
RMSE			1.94 ± 0.1	2.29 ± 0.2	2.01 ± 0.1
MAE			1.60 ± 0.1	1.89 ± 0.2	1.64 ± 0.1
r_{Spearman}			0.94	0.91	0.93
$t_{\text{preparation}}$			42 ns	309.2 ns	216.1 ns
$t_{\text{production}}$			1400 ns	67 ns	67 ns

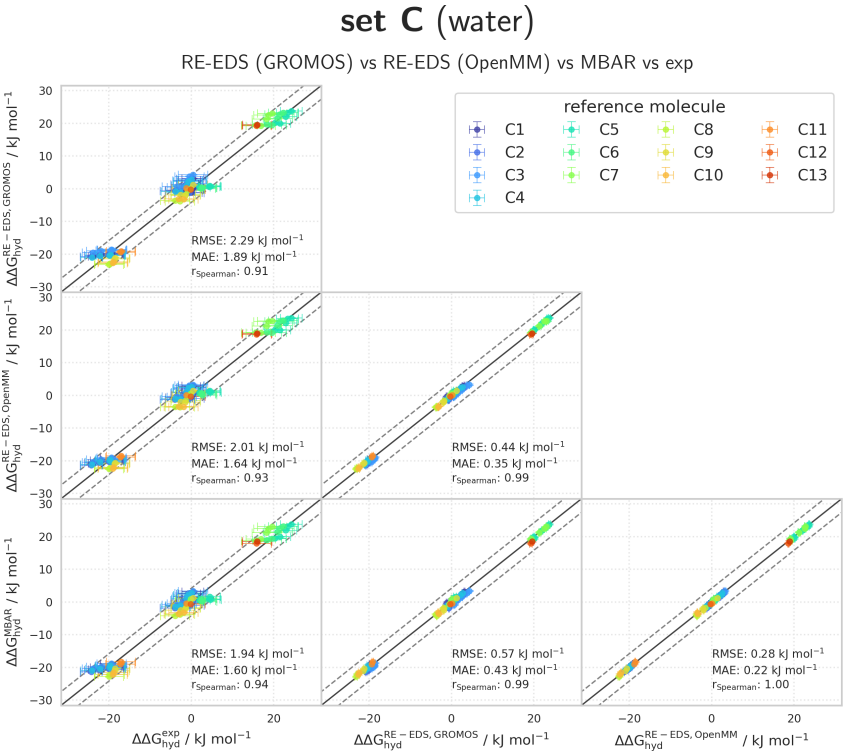


Figure 7.22: Comparison of the relative hydration free energies of set C. The pairwise comparisons of the relative hydration free energies calculated with RE-EDS in GROMOS[27] ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$), RE-EDS in OpenMM[29] ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$), MBAR ($\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$)[155,173] and the experimental results.[155,173] The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.20.

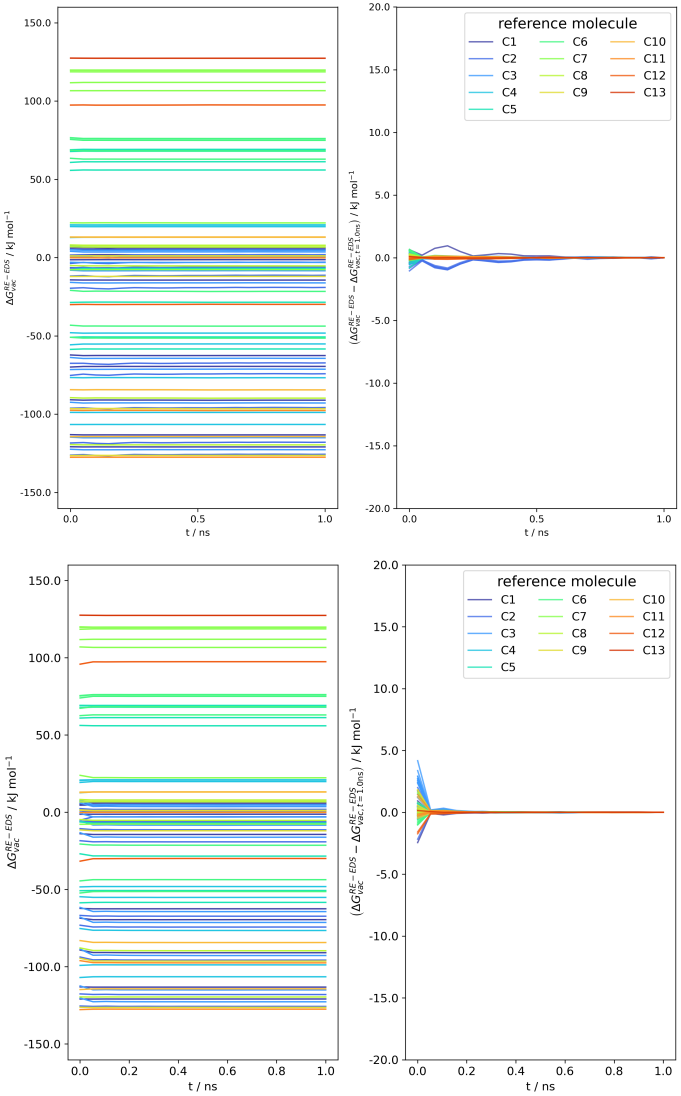


Figure 7.23: Convergence of $\Delta G_{\text{vac}}^{\text{RE-EDS}}$ as a function of the simulation time for the RE-EDS simulation of set C ($s = 1.0$) in GROMOS[27] (top) and in OpenMM[29] (bottom) from one of the five repeats used to calculate the relative hydration free energies reported in Table 7.20.

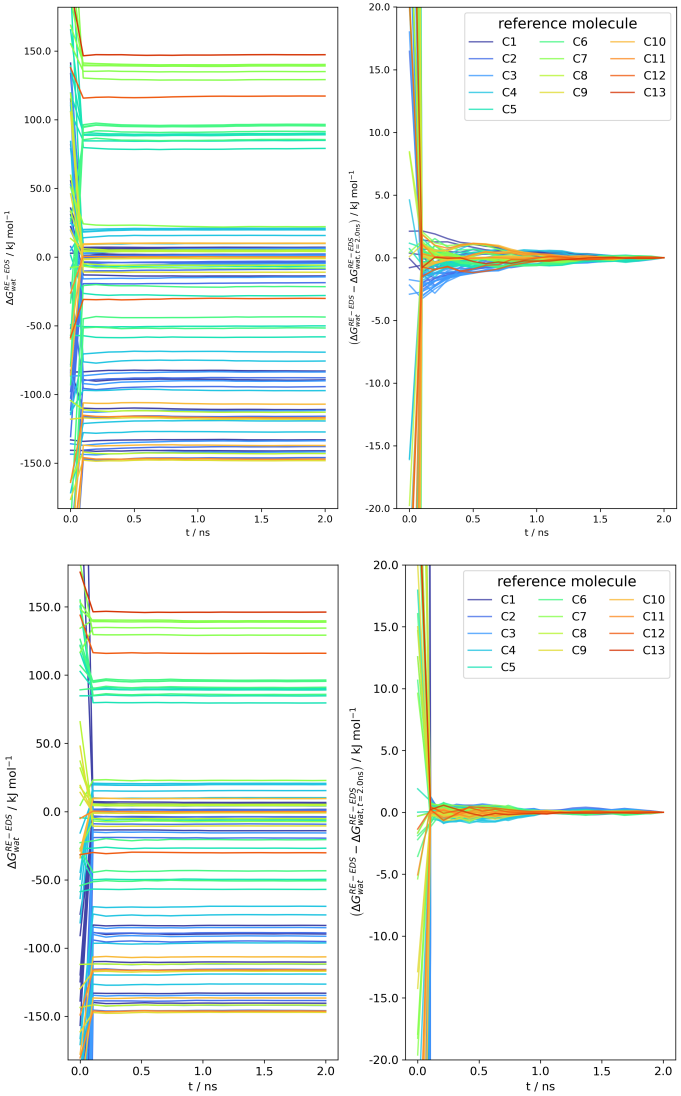


Figure 7.24: Convergence of $\Delta G_{\mathrm{wat}}^{\mathrm{RE\text{-}EDS}}$ as a function of the simulation time for the RE-EDS simulation of set C ($s = 1.0$) in GROMOS[27] (top) and in OpenMM[29] (bottom) from one of the five repeats used to calculate the relative hydration free energies reported in Table 7.20.

Relative Solvation Free Energies in Chloroform of Set C

Table 7.21: $\Delta\Delta G_{CHCl_3}$ for the 14 molecules of set C from experiment,[174] and calculated from the RE-EDS calculations in GROMOS[27] and OpenMM[29] (with the AT^{shift} scheme and $R_{RF} = 1.2$ nm). The RE-EDS results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to twelve molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G^{exp}_{CHCl_3}$ [174]	$\Delta\Delta G^{RE\text{-}EDS,GROMOS}_{CHCl_3}$	$\Delta\Delta G^{RE\text{-}EDS,OpenMM}_{CHCl_3}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	−1.5 ± 3.6	−4.1 ± 0.2	−3.2 ± 0.2
C1	C3	−3.1 ± 3.6	−5.4 ± 0.2	−4.1 ± 0.2
C1	C4	−1.6 ± 3.6	−4.3 ± 0.2	−3.8 ± 0.2
C1	C5	−6.9 ± 3.6	−3.1 ± 0.1	−3.5 ± 0.1
C1	C6	−8.8 ± 3.6	−7.3 ± 0.1	−7.4 ± 0.1
C1	C7	−7.8 ± 3.6	−4.0 ± 0.1	−4.7 ± 0.1
C1	C8	5.1 ± 3.6	3.6 ± 0.1	3.8 ± 0.1
C1	C9	0.1 ± 3.6	−0.7 ± 0.1	−0.6 ± 0.1
C1	C10	−3.5 ± 3.6	−5.1 ± 0.2	−4.8 ± 0.1
C1	C11	−2.5 ± 3.6	−4.1 ± 0.1	−4.2 ± 0.1
C1	C12	−11.5 ± 3.6	−8.5 ± 0.2	−8.4 ± 0.3
C1	C13	−10.6 ± 3.6	−8.4 ± 0.2	−8.5 ± 0.2
C1	C14	−4.7 ± 3.6	−5.8 ± 0.1	−5.9 ± 0.1
C2	C3	−1.6 ± 3.6	−1.3 ± 0.3	−0.9 ± 0.1
C2	C4	−0.1 ± 3.6	−0.2 ± 0.2	−0.5 ± 0.1
C2	C5	−5.4 ± 3.6	1.0 ± 0.2	−0.2 ± 0.2
C2	C6	−7.3 ± 3.6	−3.3 ± 0.2	−4.2 ± 0.1
C2	C7	−6.3 ± 3.6	0.0 ± 0.1	−1.4 ± 0.2
C2	C8	6.7 ± 3.6	7.7 ± 0.1	7.1 ± 0.2
C2	C9	1.6 ± 3.6	3.4 ± 0.2	2.6 ± 0.1
C2	C10	−2.0 ± 3.6	−1.0 ± 0.1	−1.6 ± 0.2
C2	C11	−1.0 ± 3.6	0.0 ± 0.2	−0.9 ± 0.1
C2	C12	−10.0 ± 3.6	−4.4 ± 0.2	−5.2 ± 0.2
C2	C13	−9.1 ± 3.6	−4.3 ± 0.1	−5.2 ± 0.1
C2	C14	−3.2 ± 3.6	−1.7 ± 0.1	−2.7 ± 0.1
C3	C4	1.5 ± 3.6	1.1 ± 0.1	0.3 ± 0.2
C3	C5	−3.8 ± 3.6	2.3 ± 0.2	0.6 ± 0.2
C3	C6	−5.7 ± 3.6	−1.9 ± 0.2	−3.3 ± 0.1
C3	C7	−4.6 ± 3.6	1.4 ± 0.3	−0.6 ± 0.2
C3	C8	8.3 ± 3.6	9.0 ± 0.3	8.0 ± 0.1
C3	C9	3.3 ± 3.6	4.7 ± 0.2	3.5 ± 0.1
C3	C10	−0.4 ± 3.6	0.3 ± 0.2	−0.7 ± 0.2
C3	C11	0.7 ± 3.6	1.3 ± 0.2	−0.0 ± 0.1
C3	C12	−8.4 ± 3.6	−3.1 ± 0.2	−4.3 ± 0.3
C3	C13	−7.4 ± 3.6	−3.0 ± 0.2	−4.4 ± 0.1
C3	C14	−1.5 ± 3.6	−0.4 ± 0.2	−1.8 ± 0.1
C4	C5	−5.4 ± 3.6	1.2 ± 0.1	0.3 ± 0.2
C4	C6	−7.2 ± 3.6	−3.1 ± 0.2	−3.6 ± 0.2
C4	C7	−6.2 ± 3.6	0.2 ± 0.3	−0.9 ± 0.3
C4	C8	6.7 ± 3.6	7.9 ± 0.2	7.6 ± 0.2
C4	C9	1.7 ± 3.6	3.6 ± 0.2	3.2 ± 0.2
C4	C10	−1.9 ± 3.6	−0.8 ± 0.2	−1.0 ± 0.2
C4	C11	−0.9 ± 3.6	0.2 ± 0.2	−0.4 ± 0.1
C4	C12	−9.9 ± 3.6	−4.2 ± 0.2	−4.7 ± 0.2
C4	C13	−9.0 ± 3.6	−4.2 ± 0.2	−4.7 ± 0.1
C4	C14	−3.1 ± 3.6	−1.6 ± 0.2	−2.1 ± 0.2
C5	C6	−1.9 ± 3.6	−4.2 ± 0.1	−3.9 ± 0.2
C5	C7	−0.8 ± 3.6	−0.9 ± 0.1	−1.2 ± 0.1
C5	C8	12.1 ± 3.6	6.7 ± 0.1	7.3 ± 0.1
C5	C9	7.1 ± 3.6	2.4 ± 0.1	2.9 ± 0.1
C5	C10	3.4 ± 3.6	−2.0 ± 0.1	−1.3 ± 0.0
C5	C11	4.5 ± 3.6	−1.0 ± 0.1	−0.7 ± 0.1
C5	C12	−4.6 ± 3.6	−5.4 ± 0.2	−4.9 ± 0.3
C5	C13	−3.6 ± 3.6	−5.3 ± 0.1	−5.0 ± 0.2
C5	C14	2.3 ± 3.6	−2.7 ± 0.1	−2.4 ± 0.1
C6	C7	1.0 ± 3.6	3.3 ± 0.1	2.7 ± 0.2
C6	C8	14.0 ± 3.6	11.0 ± 0.1	11.3 ± 0.1
C6	C9	9.0 ± 3.6	6.7 ± 0.1	6.8 ± 0.1
C6	C10	5.3 ± 3.6	2.3 ± 0.1	2.6 ± 0.1
C6	C11	6.4 ± 3.6	3.3 ± 0.1	3.3 ± 0.1
C6	C12	−2.7 ± 3.6	−1.1 ± 0.2	−1.0 ± 0.3
C6	C13	−1.8 ± 3.6	−1.1 ± 0.2	−1.1 ± 0.1
C6	C14	4.1 ± 3.6	1.5 ± 0.1	1.5 ± 0.0
C7	C8	12.9 ± 3.6	7.7 ± 0.1	8.5 ± 0.2
C7	C9	7.9 ± 3.6	3.4 ± 0.1	4.1 ± 0.1
C7	C10	4.3 ± 3.6	−1.0 ± 0.1	−0.1 ± 0.2
C7	C11	5.3 ± 3.6	−0.0 ± 0.1	0.5 ± 0.2
C7	C12	−3.7 ± 3.6	−4.4 ± 0.3	−3.7 ± 0.4
C7	C13	−2.8 ± 3.6	−4.4 ± 0.1	−3.8 ± 0.2

C7	C14	3.1 ± 3.6	−1.8 ± 0.1	−1.2 ± 0.1
C8	C9	−5.0 ± 3.6	−4.3 ± 0.1	−4.4 ± 0.1
C8	C10	−8.7 ± 3.6	−8.7 ± 0.1	−8.7 ± 0.1
C8	C11	−7.6 ± 3.6	−7.7 ± 0.1	−8.0 ± 0.1
C8	C12	−16.7 ± 3.6	−12.1 ± 0.3	−12.3 ± 0.3
C8	C13	−15.7 ± 3.6	−12.0 ± 0.1	−12.3 ± 0.1
C8	C14	−9.8 ± 3.6	−9.4 ± 0.1	−9.7 ± 0.1
C9	C10	−3.6 ± 3.6	−4.4 ± 0.1	−4.2 ± 0.1
C9	C11	−2.6 ± 3.6	−3.4 ± 0.1	−3.6 ± 0.1
C9	C12	−11.6 ± 3.6	−7.8 ± 0.2	−7.8 ± 0.3
C9	C13	−10.7 ± 3.6	−7.8 ± 0.1	−7.9 ± 0.1
C9	C14	−4.8 ± 3.6	−5.2 ± 0.1	−5.3 ± 0.1
C10	C11	1.0 ± 3.6	1.0 ± 0.1	0.7 ± 0.1
C10	C12	−8.0 ± 3.6	−3.4 ± 0.2	−3.6 ± 0.3
C10	C13	−7.1 ± 3.6	−3.3 ± 0.1	−3.7 ± 0.1
C10	C14	−1.2 ± 3.6	−0.7 ± 0.1	−1.1 ± 0.1
C11	C12	−9.0 ± 3.6	−4.4 ± 0.2	−4.3 ± 0.3
C11	C13	−8.1 ± 3.6	−4.4 ± 0.2	−4.3 ± 0.1
C11	C14	−2.2 ± 3.6	−1.7 ± 0.1	−1.7 ± 0.1
C12	C13	0.9 ± 3.6	0.0 ± 0.3	−0.0 ± 0.3
C12	C14	6.8 ± 3.6	2.6 ± 0.2	2.5 ± 0.3
C13	C14	5.9 ± 3.6	2.6 ± 0.1	2.6 ± 0.1
RMSE			3.35 ± 0.3	2.88 ± 0.2
MAE			2.73 ± 0.3	2.31 ± 0.2
r_{Spearman}			0.80	0.87
$t_{\text{preparation}}$			277.2 ns	177.1 ns
$t_{\text{production}}$			59 ns	59 ns

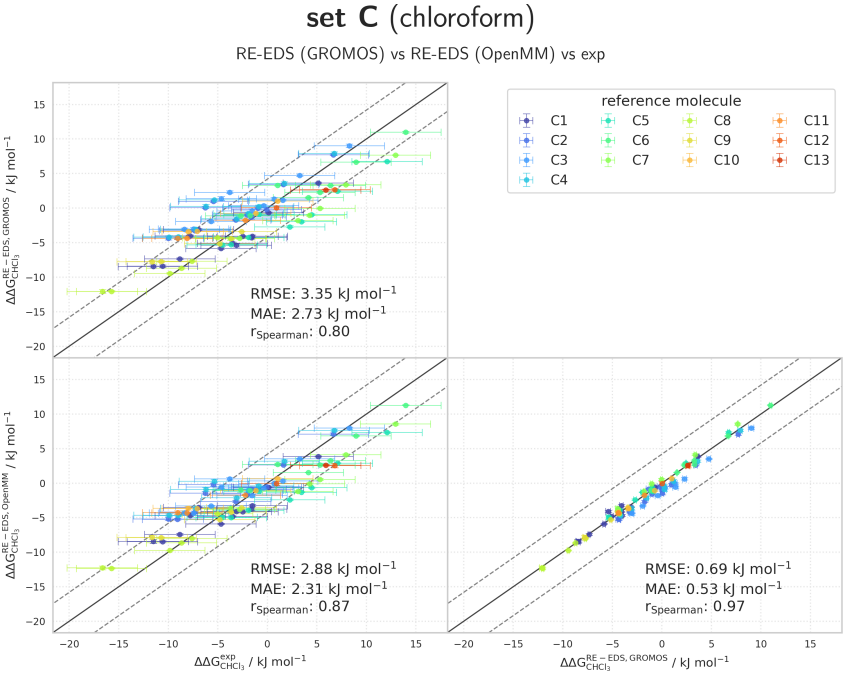


Figure 7.25: Comparison of the relative solvation free energies in chloroform of set C. The pairwise comparisons of the relative solvation free energies in chloroform calculated with RE-EDS in GROMOS[27] ($\Delta\Delta G^{\text{RE-EDS,GROMOS}}_{\text{CHCl}_3}$), RE-EDS in OpenMM[29] ($\Delta\Delta G^{\text{RE-EDS,OpenMM}}_{\text{CHCl}_3}$), and the experimental results.[174] The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.21.

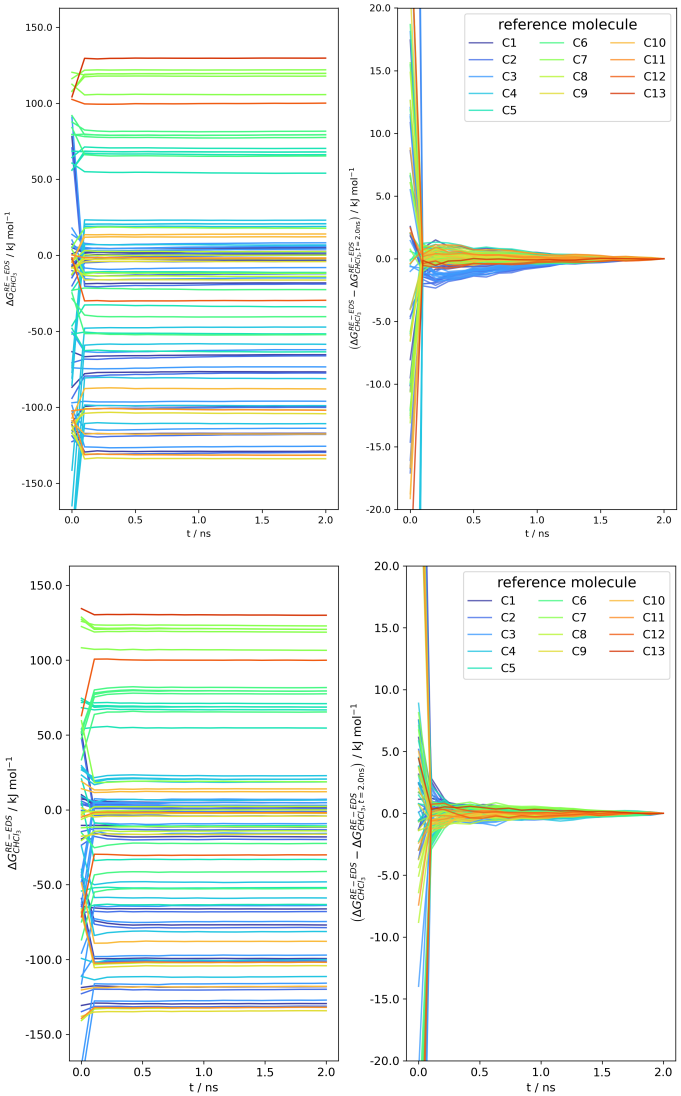


Figure 7.26: Convergence of $\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ as a function of the simulation time for the RE-EDS simulation of set C ($s = 1.0$) in GROMOS [27] (top) and in OpenMM [29] (bottom) from one of the five repeats used to calculate the relative solvation free energies in chloroform reported in Table 7.21.

Relative Hydration Free Energies of Set D

Table 7.22: $\Delta\Delta G_{\mathrm{hyd}}$ for the 13 molecules of set D from experiment,[155,173] calculated from the hydration free energies obtained with MBAR,[155,173] and from the RE-EDS calculations in GROMOS[27] and OpenMM[29] (with the AT$^{\mathrm{shift}}$ scheme and $R_{\mathrm{RF}} = 1.2$ nm). The RE-EDS results were averaged over five repeats in vacuum/water and the errors on the ΔG values correspond to the standard deviation over the five repeats. The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{exp}}$ [155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{MBAR}}$ [155,173]	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,GROMOS}}$	$\Delta\Delta G_{\mathrm{hyd}}^{\mathrm{RE\text{-}EDS,OpenMM}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	14.4 ± 2.6	11.5 ± 0.2	11.6 ± 0.3	11.6 ± 0.4
C1	C3	7.9 ± 2.6	0.1 ± 0.2	0.7 ± 0.2	0.5 ± 0.4
C1	C4	5.5 ± 2.6	0.0 ± 0.2	0.1 ± 0.3	0.4 ± 0.3
C1	C5	5.3 ± 2.6	6.6 ± 0.2	7.5 ± 0.2	7.0 ± 0.3
C1	C6	4.7 ± 2.6	7.8 ± 0.2	8.8 ± 0.4	8.3 ± 0.4
C1	C7	4.1 ± 2.6	7.3 ± 0.2	8.1 ± 0.3	7.6 ± 0.3
C1	C8	5.5 ± 2.6	6.9 ± 0.2	7.5 ± 0.3	7.1 ± 0.3
C1	C9	−12.3 ± 2.6	−20.8 ± 0.2	−20.3 ± 0.4	−20.1 ± 0.4
C1	C10	0.2 ± 3.6	−3.1 ± 0.2	−2.8 ± 0.6	−2.7 ± 0.5
C1	C11	1.9 ± 2.6	−3.1 ± 0.2	−2.0 ± 0.4	−1.8 ± 0.5
C1	C12	8.3 ± 2.6	0.7 ± 0.2	0.9 ± 0.2	−0.5 ± 0.6
C1	C13	15.4 ± 2.6	11.8 ± 0.2	12.5 ± 0.8	12.5 ± 0.3
C2	C3	−6.6 ± 3.6	−11.5 ± 0.2	−11.0 ± 0.2	−11.1 ± 0.2
C2	C4	−8.9 ± 3.6	−11.5 ± 0.2	−11.5 ± 0.3	−11.2 ± 0.3
C2	C5	−9.1 ± 3.6	−4.9 ± 0.2	−4.2 ± 0.2	−4.6 ± 0.1
C2	C6	−9.7 ± 3.6	−3.8 ± 0.2	−2.9 ± 0.2	−3.3 ± 0.3
C2	C7	−10.4 ± 3.6	−4.3 ± 0.2	−3.6 ± 0.3	−4.0 ± 0.4
C2	C8	−9.0 ± 3.6	−4.7 ± 0.2	−4.1 ± 0.2	−4.5 ± 0.1
C2	C9	−26.7 ± 3.6	−32.3 ± 0.2	−32.0 ± 0.2	−31.7 ± 0.1
C2	C10	−14.2 ± 4.3	−14.7 ± 0.2	−14.4 ± 0.6	−14.3 ± 0.3
C2	C11	−12.6 ± 3.6	−14.6 ± 0.2	−13.6 ± 0.4	−13.5 ± 0.2
C2	C12	−6.2 ± 3.6	−10.9 ± 0.2	−10.7 ± 0.2	−12.1 ± 0.5
C2	C13	1.0 ± 3.6	0.3 ± 0.2	0.8 ± 0.8	0.9 ± 0.3
C3	C4	−2.3 ± 3.6	−0.1 ± 0.2	−0.5 ± 0.1	−0.1 ± 0.4
C3	C5	−2.6 ± 3.6	6.5 ± 0.2	6.8 ± 0.1	6.5 ± 0.1
C3	C6	−3.1 ± 3.6	7.7 ± 0.2	8.1 ± 0.2	7.8 ± 0.2
C3	C7	−3.8 ± 3.6	7.2 ± 0.2	7.4 ± 0.3	7.1 ± 0.4
C3	C8	−2.4 ± 3.6	6.8 ± 0.2	6.9 ± 0.1	6.6 ± 0.1
C3	C9	−20.1 ± 3.6	−20.9 ± 0.2	−21.0 ± 0.2	−20.7 ± 0.2
C3	C10	−7.7 ± 4.3	−3.2 ± 0.2	−3.5 ± 0.5	−3.2 ± 0.4
C3	C11	−6.0 ± 3.6	−3.1 ± 0.2	−2.6 ± 0.3	−2.4 ± 0.2
C3	C12	0.4 ± 3.6	0.6 ± 0.2	0.3 ± 0.0	−1.0 ± 0.6
C3	C13	7.5 ± 3.6	11.8 ± 0.2	11.8 ± 0.8	12.0 ± 0.4
C4	C5	−0.2 ± 3.6	6.6 ± 0.2	7.3 ± 0.2	6.6 ± 0.3
C4	C6	−0.8 ± 3.6	7.8 ± 0.2	8.7 ± 0.3	7.9 ± 0.4
C4	C7	−1.5 ± 3.6	7.3 ± 0.2	7.9 ± 0.3	7.2 ± 0.5
C4	C8	−0.0 ± 3.6	6.9 ± 0.2	7.4 ± 0.2	6.7 ± 0.3
C4	C9	−17.8 ± 3.6	−20.8 ± 0.2	−20.5 ± 0.2	−20.5 ± 0.4
C4	C10	−5.3 ± 4.3	−3.1 ± 0.2	−2.9 ± 0.5	−3.1 ± 0.4
C4	C11	−3.6 ± 3.6	−3.1 ± 0.2	−2.1 ± 0.4	−2.3 ± 0.5
C4	C12	2.8 ± 3.6	0.7 ± 0.2	0.8 ± 0.2	−0.9 ± 0.4
C4	C13	9.9 ± 3.6	11.8 ± 0.2	12.3 ± 0.8	12.1 ± 0.3
C5	C6	−0.6 ± 3.6	1.2 ± 0.2	1.3 ± 0.2	1.3 ± 0.2
C5	C7	−1.3 ± 3.6	0.7 ± 0.2	0.6 ± 0.2	0.6 ± 0.3
C5	C8	0.2 ± 3.6	0.3 ± 0.2	0.1 ± 0.1	0.1 ± 0.1
C5	C9	−17.6 ± 3.6	−27.4 ± 0.2	−27.8 ± 0.2	−27.1 ± 0.2
C5	C10	−5.1 ± 4.3	−9.7 ± 0.2	−10.3 ± 0.5	−9.7 ± 0.3
C5	C11	−3.4 ± 3.6	−9.7 ± 0.2	−9.5 ± 0.3	−8.8 ± 0.3
C5	C12	3.0 ± 3.6	−5.9 ± 0.2	−6.5 ± 0.1	−7.5 ± 0.5
C5	C13	10.1 ± 3.6	5.2 ± 0.2	5.0 ± 0.8	5.5 ± 0.3
C6	C7	−0.7 ± 3.6	−0.5 ± 0.2	−0.7 ± 0.3	−0.7 ± 0.3
C6	C8	0.8 ± 3.6	−0.9 ± 0.2	−1.3 ± 0.1	−1.2 ± 0.2
C6	C9	−17.0 ± 3.6	−28.6 ± 0.2	−29.1 ± 0.2	−28.4 ± 0.3
C6	C10	−4.5 ± 4.3	−10.9 ± 0.2	−11.6 ± 0.6	−11.0 ± 0.2
C6	C11	−2.8 ± 3.6	−10.8 ± 0.2	−10.8 ± 0.3	−10.1 ± 0.4
C6	C12	3.6 ± 3.6	−7.1 ± 0.2	−7.9 ± 0.3	−8.8 ± 0.6
C6	C13	10.7 ± 3.6	4.1 ± 0.2	3.7 ± 0.8	4.2 ± 0.3
C7	C8	1.4 ± 3.6	−0.4 ± 0.2	−0.5 ± 0.2	−0.5 ± 0.3
C7	C9	−16.3 ± 3.6	−28.1 ± 0.2	−28.4 ± 0.4	−27.8 ± 0.3
C7	C10	−3.8 ± 4.3	−10.4 ± 0.2	−10.9 ± 0.3	−10.3 ± 0.3
C7	C11	−2.2 ± 3.6	−10.3 ± 0.2	−10.1 ± 0.2	−9.5 ± 0.5
C7	C12	4.2 ± 3.6	−6.6 ± 0.2	−7.1 ± 0.3	−8.1 ± 0.7
C7	C13	11.3 ± 3.6	4.6 ± 0.2	4.4 ± 0.9	4.9 ± 0.4
C8	C9	−17.7 ± 3.6	−27.7 ± 0.2	−27.9 ± 0.2	−27.3 ± 0.1
C8	C10	−5.3 ± 4.3	−10.0 ± 0.2	−10.3 ± 0.5	−9.8 ± 0.3
C8	C11	−3.6 ± 3.6	−9.9 ± 0.2	−9.5 ± 0.3	−9.0 ± 0.2
C8	C12	2.8 ± 3.6	−6.2 ± 0.2	−6.6 ± 0.1	−7.6 ± 0.5
C8	C13	9.9 ± 3.6	5.0 ± 0.2	4.9 ± 0.7	5.4 ± 0.3
C9	C10	12.5 ± 4.3	17.7 ± 0.2	17.5 ± 0.6	17.4 ± 0.3
C9	C11	14.1 ± 3.6	17.7 ± 0.2	18.3 ± 0.4	18.3 ± 0.2

C9	C12	20.5 ± 3.6	21.5 ± 0.2	21.3 ± 0.2	19.6 ± 0.6
C9	C13	27.7 ± 3.6	32.6 ± 0.2	32.8 ± 0.7	32.7 ± 0.4
C10	C11	1.7 ± 4.3	0.1 ± 0.2	0.8 ± 0.3	0.9 ± 0.4
C10	C12	8.1 ± 4.3	3.8 ± 0.2	3.7 ± 0.5	2.2 ± 0.6
C10	C13	15.2 ± 4.3	15.0 ± 0.2	15.3 ± 1.1	15.2 ± 0.3
C11	C12	6.4 ± 3.6	3.7 ± 0.2	2.9 ± 0.3	1.3 ± 0.6
C11	C13	13.5 ± 3.6	14.9 ± 0.2	14.5 ± 0.9	14.4 ± 0.5
C12	C13	7.1 ± 3.6	11.2 ± 0.2	11.5 ± 0.7	13.0 ± 0.5
RMSE			5.86 ± 0.6	6.06 ± 0.6	5.98 ± 0.6
MAE			4.92 ± 0.5	5.06 ± 0.5	5.01 ± 0.5
^{r}Spearman			0.80	0.81	0.80
$t_{\text{preparation}}$			39 ns	282.6 ns	191.8 ns
$t_{\text{production}}$			1300 ns	66 ns	66 ns

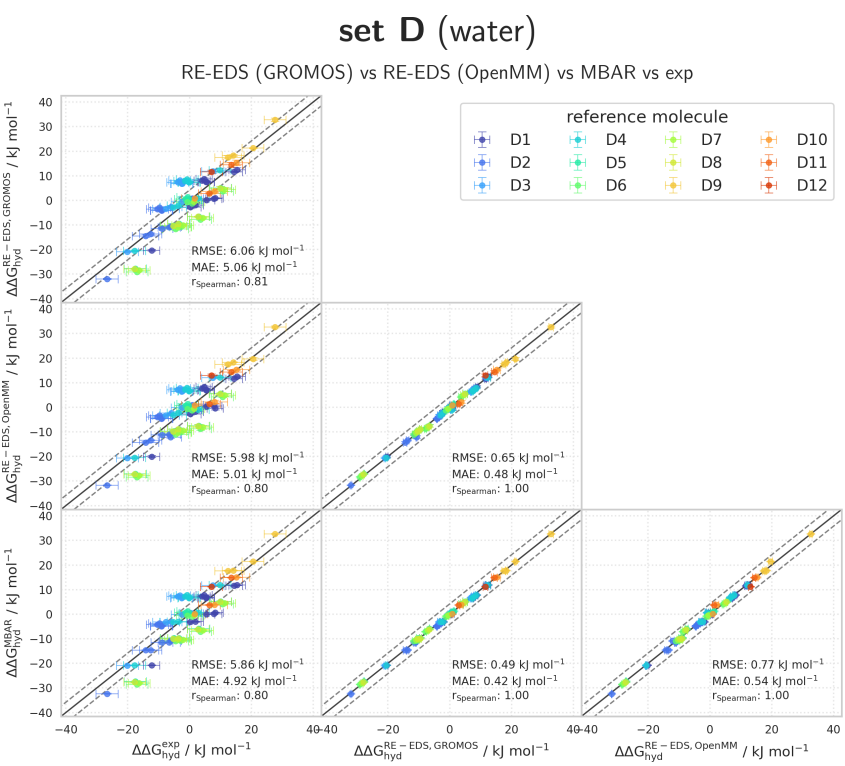


Figure 7.27: Comparison of the relative hydration free energies of set D. The pairwise comparisons of the relative hydration free energies calculated with RE-EDS in GROMOS[27] ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,GROMOS}}$), RE-EDS in OpenMM[29] ($\Delta\Delta G_{\text{hyd}}^{\text{RE-EDS,OpenMM}}$), MBAR ($\Delta\Delta G_{\text{hyd}}^{\text{MBAR}}$)[155,173] and the experimental results.[155,173] The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.22.

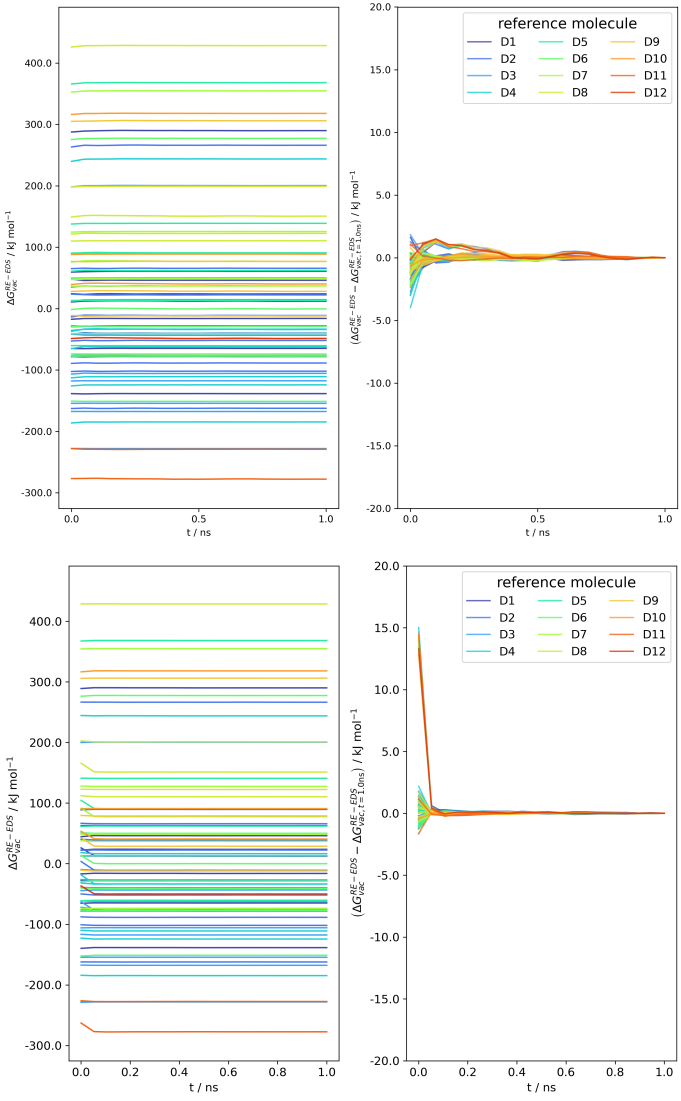


Figure 7.28: Convergence of $\Delta G_{\mathrm{vac}}^{\mathrm{RE\text{-}EDS}}$ as a function of the simulation time for the RE-EDS simulation of set D ($s = 1.0$) in GROMOS[27] (top) and in OpenMM[29] (bottom) from one of the five repeats used to calculate the relative hydration free energies reported in Table 7.22.

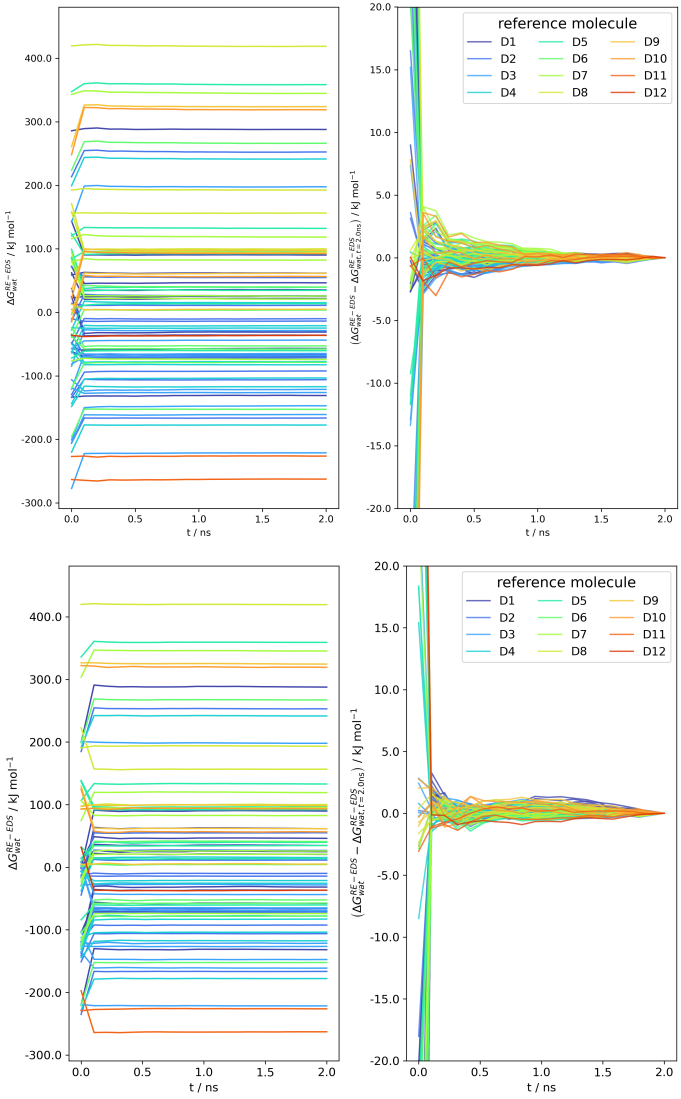


Figure 7.29: Convergence of $\Delta G_{\text{wat}}^{\text{RE-EDS}}$ as a function of the simulation time for the RE-EDS simulation of set D ($s = 1.0$) in GROMOS [27] (top) and in OpenMM [29] (bottom) from one of the five repeats used to calculate the relative hydration free energies reported in Table 7.22.

Relative Solvation Free Energies in Chloroform of Set D

Table 7.23: $\Delta\Delta G_{\mathrm{CHCl_3}}$ for the 13 molecules of set D from experiment,[174] and calculated from the RE-EDS calculations in GROMOS[27] and OpenMM[29] (with the AT$^{\mathrm{shift}}$ scheme and $R_{\mathrm{RF}} = 1.2$ nm). The RE-EDS results were averaged over five repeats in vacuum/chloroform and the errors on the ΔG values correspond to the standard deviation over the five repeats. The experimental errors of the free energies in chloroform were set to 0.6 kcal/mol (2.5104 kJ mol^{-1}), analogous to FreeSolv.[155,173] The error estimate of the $\Delta\Delta G$ values was calculated *via* Gaussian error propagation. The uncertainties of the RMSEs and MAEs were estimated from the distribution of RMSEs and MAEs when a random selection of up to eleven molecules was removed from the calculations (5000 repetitions). The accumulated simulation time is split into preparation (pre-processing, equilibration) and production time.

Molecules		$\Delta\Delta G^{\mathrm{exp}}_{\mathrm{CHCl_3}}$ [174]	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,GROMOS}}_{\mathrm{CHCl_3}}$	$\Delta\Delta G^{\mathrm{RE\text{-}EDS,OpenMM}}_{\mathrm{CHCl_3}}$
i	j	[kJ mol^{-1}]	[kJ mol^{-1}]	[kJ mol^{-1}]
C1	C2	5.5 ± 3.6	1.1 ± 0.2	0.9 ± 0.2
C1	C3	1.9 ± 3.6	−0.6 ± 0.1	−1.2 ± 0.2
C1	C4	−1.1 ± 3.6	−5.5 ± 0.3	−5.5 ± 0.1
C1	C5	2.4 ± 3.6	3.1 ± 0.2	2.5 ± 0.1
C1	C6	0.8 ± 3.6	3.0 ± 0.2	2.2 ± 0.2
C1	C7	0.2 ± 3.6	2.8 ± 0.2	2.2 ± 0.2
C1	C8	−0.8 ± 3.6	−0.9 ± 0.2	−1.1 ± 0.2
C1	C9	−11.5 ± 3.6	−7.2 ± 0.2	−8.2 ± 0.2
C1	C10	−4.4 ± 3.6	−3.4 ± 0.3	−4.0 ± 0.2
C1	C11	−0.7 ± 3.6	−3.3 ± 0.3	−3.8 ± 0.1
C1	C12	−1.1 ± 3.6	−9.1 ± 0.2	−10.7 ± 0.4
C1	C13	1.6 ± 3.6	−2.6 ± 0.7	−2.3 ± 0.3
C2	C3	−3.6 ± 3.6	−1.7 ± 0.2	−2.1 ± 0.2
C2	C4	−6.6 ± 3.6	−6.6 ± 0.3	−6.5 ± 0.2
C2	C5	−3.1 ± 3.6	2.0 ± 0.2	1.5 ± 0.1
C2	C6	−4.6 ± 3.6	1.9 ± 0.2	1.3 ± 0.2
C2	C7	−5.3 ± 3.6	1.8 ± 0.2	1.3 ± 0.1
C2	C8	−6.3 ± 3.6	−2.0 ± 0.2	−2.1 ± 0.3
C2	C9	−17.0 ± 3.6	−8.3 ± 0.2	−9.2 ± 0.1
C2	C10	−9.8 ± 3.6	−4.5 ± 0.3	−4.9 ± 0.1
C2	C11	−6.2 ± 3.6	−4.3 ± 0.3	−4.7 ± 0.2
C2	C12	−6.6 ± 3.6	−10.2 ± 0.2	−11.7 ± 0.5
C2	C13	−3.8 ± 3.6	−3.7 ± 0.7	−3.2 ± 0.4
C3	C4	−3.0 ± 3.6	−4.9 ± 0.3	−4.4 ± 0.2
C3	C5	0.5 ± 3.6	3.8 ± 0.1	3.6 ± 0.1
C3	C6	−1.1 ± 3.6	3.7 ± 0.1	3.4 ± 0.2
C3	C7	−1.7 ± 3.6	3.5 ± 0.1	3.4 ± 0.3
C3	C8	−2.7 ± 3.6	−0.2 ± 0.2	0.0 ± 0.4
C3	C9	−13.4 ± 3.6	−6.6 ± 0.1	−7.1 ± 0.1
C3	C10	−6.3 ± 3.6	−2.8 ± 0.3	−2.8 ± 0.1
C3	C11	−2.6 ± 3.6	−2.6 ± 0.2	−2.6 ± 0.2
C3	C12	−3.0 ± 3.6	−8.5 ± 0.1	−9.6 ± 0.5
C3	C13	−0.3 ± 3.6	−1.9 ± 0.7	−1.1 ± 0.3
C4	C5	3.5 ± 3.6	8.7 ± 0.2	8.0 ± 0.1
C4	C6	1.9 ± 3.6	8.6 ± 0.2	7.7 ± 0.2
C4	C7	1.3 ± 3.6	8.4 ± 0.3	7.7 ± 0.2
C4	C8	0.3 ± 3.6	4.7 ± 0.3	4.4 ± 0.2
C4	C9	−10.4 ± 3.6	−1.7 ± 0.3	−2.7 ± 0.2
C4	C10	−3.3 ± 3.6	2.1 ± 0.3	1.5 ± 0.2
C4	C11	0.4 ± 3.6	2.3 ± 0.3	1.7 ± 0.1
C4	C12	0.0 ± 3.6	−3.6 ± 0.2	−5.2 ± 0.4
C4	C13	2.7 ± 3.6	3.0 ± 0.8	3.2 ± 0.4
C5	C6	−1.5 ± 3.6	−0.1 ± 0.1	−0.3 ± 0.2
C5	C7	−2.2 ± 3.6	−0.3 ± 0.1	−0.3 ± 0.2
C5	C8	−3.2 ± 3.6	−4.0 ± 0.1	−3.6 ± 0.3
C5	C9	−13.9 ± 3.6	−10.3 ± 0.1	−10.7 ± 0.1
C5	C10	−6.7 ± 3.6	−6.6 ± 0.3	−6.4 ± 0.1
C5	C11	−3.1 ± 3.6	−6.4 ± 0.2	−6.3 ± 0.2
C5	C12	−3.5 ± 3.6	−12.3 ± 0.1	−13.2 ± 0.4
C5	C13	−0.8 ± 3.6	−5.7 ± 0.7	−4.7 ± 0.3
C6	C7	−0.6 ± 3.6	−0.2 ± 0.2	−0.0 ± 0.3
C6	C8	−1.6 ± 3.6	−3.9 ± 0.2	−3.4 ± 0.4
C6	C9	−12.3 ± 3.6	−10.2 ± 0.1	−10.4 ± 0.2
C6	C10	−5.2 ± 3.6	−6.5 ± 0.3	−6.2 ± 0.1
C6	C11	−1.5 ± 3.6	−6.3 ± 0.2	−6.0 ± 0.3
C6	C12	−1.9 ± 3.6	−12.2 ± 0.1	−12.9 ± 0.5
C6	C13	0.8 ± 3.6	−5.6 ± 0.7	−4.5 ± 0.2
C7	C8	−1.0 ± 3.6	−3.7 ± 0.1	−3.3 ± 0.2
C7	C9	−11.7 ± 3.6	−10.1 ± 0.1	−10.4 ± 0.1
C7	C10	−4.6 ± 3.6	−6.3 ± 0.3	−6.2 ± 0.2
C7	C11	−0.9 ± 3.6	−6.1 ± 0.2	−6.0 ± 0.3
C7	C12	−1.3 ± 3.6	−12.0 ± 0.2	−12.9 ± 0.4
C7	C13	1.4 ± 3.6	−5.4 ± 0.7	−4.5 ± 0.5
C8	C9	−10.7 ± 3.6	−6.3 ± 0.2	−7.1 ± 0.3
C8	C10	−3.6 ± 3.6	−2.6 ± 0.2	−2.8 ± 0.3
C8	C11	0.1 ± 3.6	−2.4 ± 0.2	−2.7 ± 0.3
C8	C12	−0.3 ± 3.6	−8.3 ± 0.2	−9.6 ± 0.4
C8	C13	2.4 ± 3.6	−1.7 ± 0.7	−1.1 ± 0.5
C9	C10	7.2 ± 3.6	3.8 ± 0.3	4.2 ± 0.1

C9	C11	10.8 ± 3.6	4.0 ± 0.2	4.4 ± 0.2
C9	C12	10.4 ± 3.6	−1.9 ± 0.1	−2.5 ± 0.4
C9	C13	13.1 ± 3.6	4.6 ± 0.7	6.0 ± 0.4
C10	C11	3.6 ± 3.6	0.2 ± 0.3	0.2 ± 0.3
C10	C12	3.3 ± 3.6	−5.7 ± 0.3	−6.7 ± 0.5
C10	C13	6.0 ± 3.6	0.9 ± 0.7	1.7 ± 0.3
C11	C12	−0.4 ± 3.6	−5.9 ± 0.2	−6.9 ± 0.4
C11	C13	2.3 ± 3.6	0.7 ± 0.7	1.5 ± 0.4
C12	C13	2.7 ± 3.6	6.6 ± 0.7	8.4 ± 0.5
RMSE			4.97 ± 0.5	4.95 ± 0.4
MAE			4.12 ± 0.5	3.99 ± 0.4
r^{Spearman}			0.59	0.59
$t^{\text{preparation}}$			211 ns	153.8 ns
$t^{\text{production}}$			58 ns	58 ns

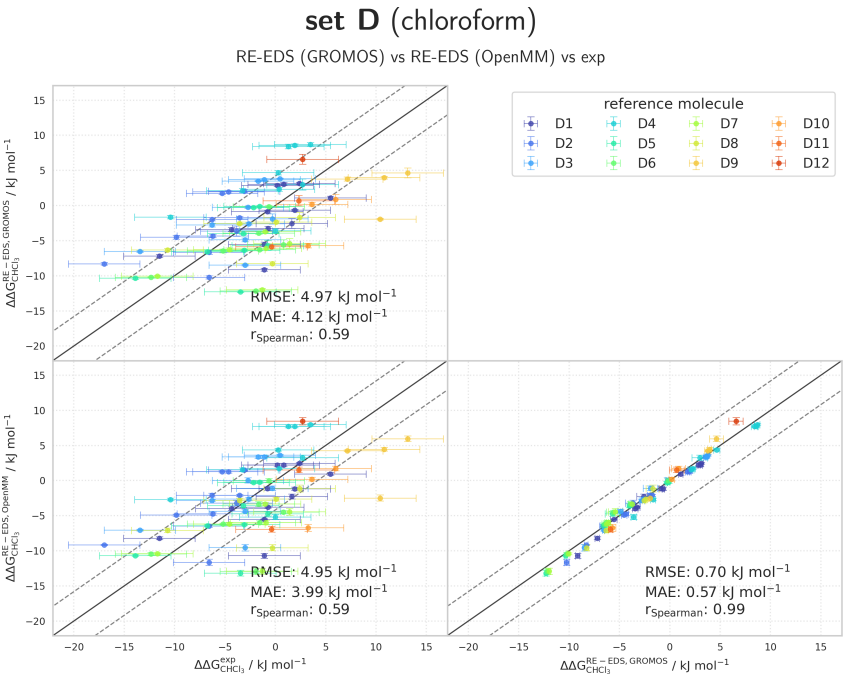


Figure 7.30: Comparison of the relative solvation free energies in chloroform of set D. The pairwise comparisons of the relative solvation free energies in chloroform calculated with RE-EDS in GROMOS[27] ($\Delta\Delta G^{\text{RE-EDS,GROMOS}}_{\text{CHCl}_3}$), RE-EDS in OpenMM[29] ($\Delta\Delta G^{\text{RE-EDS,OpenMM}}_{\text{CHCl}_3}$), and the experimental results.[174] The gray diagonal lines correspond to perfect alignment within ± 4.184 kJ mol^{-1} (± 1 kcal mol^{-1}). The pairwise RMSD, MAE, and Spearman correlation coefficient are reported. The $\Delta\Delta G^{ji}$ values are colored according to end-state i (*i.e.*, the "reference molecule" for the calculation). The numerical values are provided in Table 7.23.

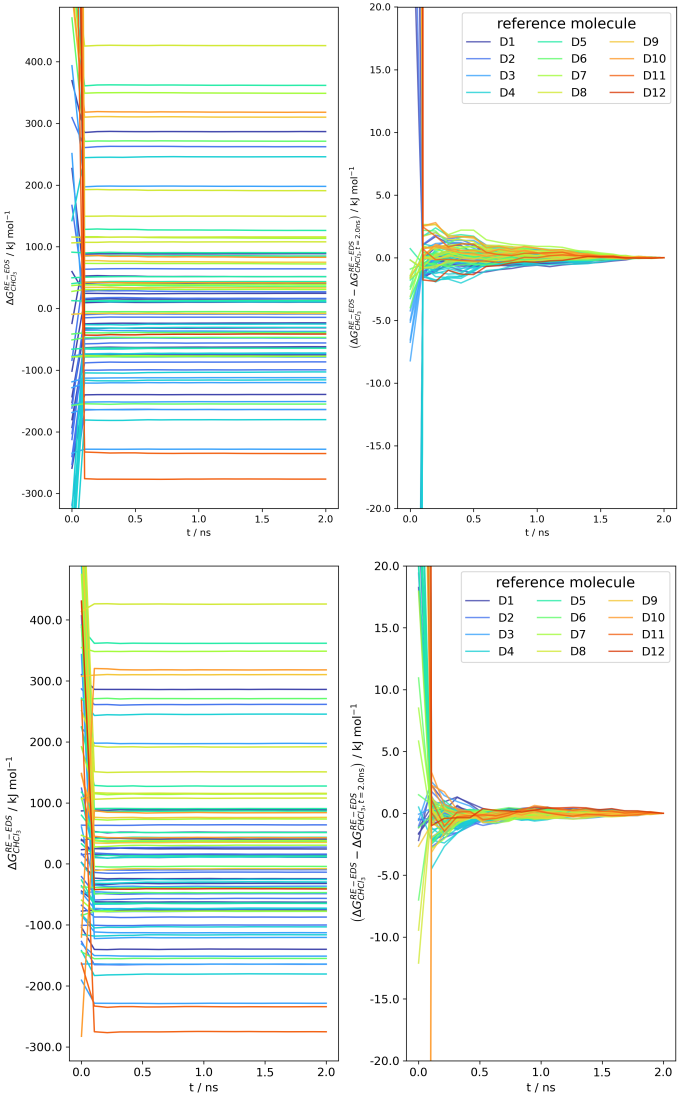


Figure 7.31: Convergence of $\Delta G^{\text{RE-EDS}}_{\text{CHCl}_3}$ as a function of the simulation time for the RE-EDS simulation of set D ($s = 1.0$) in GROMOS[27] (top) and in OpenMM[29] (bottom) from one of the five repeats used to calculate the relative solvation free energies in chloroform reported in Table 7.23.

8 Summary and Outlook

"Dulcius ex asperis"

"Sweeter after difficulties"

Clan Fergusson Motto[323]

8.1 FORCE-FIELD PARAMETERIZATION WITH COMBIFF

CombiFF[73,74] is an automated scheme to parameterize force-fields for condensed-phase MD simulations. It is a fragment-based approach rooted in the transferability principle, namely that force-field parameters for small molecules can be used as well for larger compounds. Two essential ingredients of the CombiFF scheme are: (*i*) the enumeration of all isomers for a given target family; and (*ii*) the automated generation of molecular topologies based on a fragment library. These two tasks are carried out by the C++ programs *enu* (Chapter 2) and *tbl* (Chapter 3), respectively.

The *enu* program enumerates the constitutional and spatial isomers of a given molecular formula based on the orderly enumeration of adjacency matrices and their respective automorphism group. The enumerated isomers are reported as canonical SMILES strings in a convenient XML format. The program offers the user a lot of flexibility to specify the molecules of interest, *e.g.*, by allowing ranges of atom counts in the

molecular formulas, or by filtering for substructures. However, there are still several shortcomings to be addressed, as well as possibilities to further improve the performance. These include the handling of aromaticity and variable valences (*e.g.*, for sulfur or phosphorus), the recognition of para stereocenters that do not depend on true stereocenters in cycles, and the pre-distribution of all atoms (instead of only hydrogen atoms) with valence one among the atoms with valence larger than one. The performance of the program could be further improved by parallelizing, *e.g.*, the enumeration of different canonical hydrogen vectors.

The *tbl* program employs the Ullmann[213] subgraph isomorphism algorithm to decompose a given molecule into smaller (pre-defined) fragments. In the context of CombiFF, fragments are defined as molecular building blocks, consisting of core atoms and link atoms, that can be connected to form larger molecules using so-called bond-linking. For a given list of molecules, the program creates three intermediate files, specifying: (*i*) the decomposition of the molecules into fragments; (*ii*) a molecular topology in XML format; and (*iii*) a GROMOS mtb file. The force-field parameters for step (*ii*) and (*iii*) are string macros, assembled by combining the parameters of the fragments. These string macros are subsequently replaced by concrete force-field parameters with a search-and-replace procedure using RegEx notation. In the future, the in-house MD engine SAMOS will be modified to read the macro parameters directly, and assign (and update) the concrete parameters internally during the optimization process. While it is unlikely that the automated topology generation using *tbl* becomes a bottleneck of the CombiFF approach, if the performance of *tbl* ever becomes an issue, the Ullmann algorithm could be replaced with a more efficient one, such as VF2.[214,215]

In Chapter 4, the conversion and canonicalization program *cnv* was outlined. It is able to canonicalize molecular properties, such as the chemical formula, the SMILES string, or the adjacency matrix, according to the conventions of CombiFF. There are, in principle, almost limitless possibilities to refine and extend this program. Possible extensions include,

for example, the recognition of aromaticity or the implementation of (or integration with existing) visualization tools.

8.2 FREE-ENERGY CALCULATIONS WITH RE-EDS

Replica-exchange enveloping distribution sampling (RE-EDS)[164–166] is a pathway-independent multistate free-energy method, combining Hamiltonian replica-exchange (RE)[271,272] with enveloping distribution sampling (EDS).[162,163] Although RE-EDS can be combined with a single or a hybrid-topology approach, in previous studies involving RE-EDS, a linked dual-topology approach was usually employed.[164–166,254] As all coordinates are explicitly present with the dual-topology approach, the molecules are in principle able to drift away from each other during the simulation. One strategy to prevent this is the assignment of atom-atom distance restraints. In Chapter 5, the program *RestraintMaker*, developed by Ries, Rieder *et al.*,[254] was used to generate distance restraints for relative hydration free-energy calculations with RE-EDS for two sets of molecules. The obtained results were compared to experimental and calculated values from the ATB server.[135] It was shown that the distance restraints are successful in keeping the molecules well-aligned, and that the conformational sampling of the end-states is only negligibly influenced by the distance restraints. Further research could involve the comparison of the accuracy of free-energy calculations with RE-EDS using single, hybrid, and dual topologies. In the future, distance restraints generated by *RestraintMaker* will also be tested for RE-EDS binding free-energy calculations.

In Chapter 6, the topology conversion tool *amber2gromos* was presented. It converts AMBER topologies to GROMOS-compatible topologies. A particular advantage of generalized AMBER force-field (GAFF) topologies is the existence of AmberTools, enabling the automated generation and parameterization of molecular topologies.[55,133,134] The use

of GAFF topologies, converted with *amber2gromos*, for simulations in GROMOS was validated with single-molecule simulations in vacuum, as well as relative hydration free-energy calculations with RE-EDS for two sets of benzene derivatives. The obtained relative hydration free energies were compared to values obtained with thermodynamic integration (TI)[156] in GROMACS,[25] as well as with multistate Bennett's acceptance ratio (MBAR)[157,158] and experimental values as reported in the FreeSolv database.[155,173] While the first dataset only contained six molecules to enable the efficient validation of *amber2gromos*, the second dataset was much larger, containing 28 molecules. To investigate whether the performance of RE-EDS would be negatively affected by having many end-states in a system, the set of 28 molecules was divided into two subsets of 14 and 15 molecules, respectively, with one molecule in common. For both sets of molecules, as well as the two combined subsets, there was good agreement with the results obtained using other free-energy methods, as well as with the experimental values. In the future, RE-EDS free-energy calculations with AMBER/GAFF or OpenFF topologies in GROMOS will be used to calculate relative binding free-energies.

The calculation of (pairwise) nonbonded interactions is generally the most expensive part of an MD simulation. To mitigate this, a cutoff is typically employed and the interactions of particle pairs beyond the cutoff distance are neglected. Two strategies to account for the long-range electrostatic interactions are mean-field schemes such as a reaction-field (RF)[114] correction, or lattice-sum schemes such as Ewald summation,[116] particle-particle particle-mesh (P3M),[117] or particle-mesh Ewald (PME).[99,118] To decrease the time required for the pairlist update and to avoid artifacts at the cutoff when using a straight truncation, the GROMOS MD engine makes use of charge groups. Many modern force-fields – such as GAFF – and MD engines – such as OpenMM[29] – do not support charge groups. Recently, Kubincová *et al.*[304] published a modified RF correction, employing a shifting function to avoid artifacts at the cutoff when using an atom-based cutoff.[304] The GROMOS implementation of the shifted

RF correction will be updated to use Eq. (7.13) instead of Eq. (7.15), and will be added to the next GROMOS release. In Chapter 7, three different RF schemes, combined with three different cutoff distances, were compared for relative solvation free-energy calculations with RE-EDS in water (high permittivity) and chloroform (low permittivity), using GAFF topologies in GROMOS. The obtained results validated the use of the shifted RF scheme for RE-EDS calculations with GAFF topologies and an atom-based cutoff. Further research is needed to conclude whether using the corrected potential-energy function $V^{\text{shift}\rightarrow\text{phys}}$ (Eq. 7.18) generally improves the accuracy of free-energy calculations using the shifted RF correction. A proof-of-concept implementation of RE-EDS with OpenMM was written and successfully validated by repeating the relative solvation free-energy calculations in water and in chloroform, and comparing the obtained results to the results obtained with RE-EDS in GROMOS. In this implementation, the replicas are simulated serially. The high-performance GPU implementation of OpenMM leads to a considerable speed-up for EDS calculations. However, as more replicas are added for RE-EDS simulations, this speed-up decreases. In the future, the OpenMM implementation of RE-EDS will be modified to use parallel replicas. It will also be investigated whether the double evaluation of the end-state energies (once for the calculation of the scaling factors, and once for the calculation of the forces) can be avoided by using, *e.g.*, the *CustomIntegrator* or *CustomCVForce* class provided by OpenMM. Further speed-up could be achieved by implementing the shifted RF correction in the C++ layer of OpenMM. In the future, the OpenMM implementation of RE-EDS will also be used for binding free-energy calculations.

The three chapters on RE-EDS highlight the high sampling efficiency of the method compared to other free-energy methods, namely TI[156] and MBAR.[157,158] A particular advantage is the fact that environment-environment interactions only need to be calculated once (per replica), and not for all pairs of end-states separately. While RE-EDS is highly parallelizable, at the level of both the replicas and the atomic interactions

within the replicas, the wall-clock time still increases as more end-states are added to a system. In the future, this can be mitigated by: (*i*) dividing large datasets into smaller subsets with one (or several) end-states in common; and/or (*ii*) implementing a parallelized version of RE-EDS in a high-performance MD engine such as OpenMM.[29]

An interesting future research project could involve the combination of CombiFF and RE-EDS. For example, the sets of molecules used for the relative solvation free-energy calculations of Chapters 5 - 7 could be re-parameterized employing the shifted RF correction. The free-energy calculations could then be repeated with the optimized parameters and the results compared to the ones obtained with the ATB/GAFF parameters.

References

Cover: Plots adapted from Ref. 324, visualization of molecules generated with the RDKit [210] and MarvinSketch. [325]

[1] King, S. *11/22/63*; Simon and Schuster, 2011.

[2] Asimov, I. *I. Asimov: A Memoir*; Doubleday, 1994.

[3] Alder, B. J.; Wainwright, T. E. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **1957**, *27*, 1208–1209.

[4] Alder, B. J.; Wainwright, T. E. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.* **1959**, *31*, 459–466.

[5] Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press, New York, USA, 1987.

[6] Berendsen, H. J. C. *Simulating the Physical World*; Cambridge University Press, Cambridge, UK, 2007.

[7] Hirst, J. D.; Glowacki, D. R.; Baaden, M. Molecular Simulations and Visualization: Introduction and Overview. *Faraday Discuss.* **2014**, *169*, 9–22.

[8] Hoover, W. G.; Ladd, A. J. C.; Hoover, V. N. In *Molecular-Based Study of Fluids*; Haile, J. M., Mansoori, G. A., Eds.; ACS Publications, 1983; Chapter 2, pp 29–46.

[9] Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **1964**, *136*, A405–A411.

[10] Verlet, L. Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Phys. Rev.* **1967**, *159*, 98–103.

[11] Rahman, A.; Stillinger, F. H. Molecular Dynamics Study of Liquid Water. *J. Chem. Phys.* **1971**, *55*, 3336–3359.

[12] Berendsen, H. Models for Protein Dynamics. Report of the CECAM Workshop (Orsay). 1976.

[13] McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267*, 585–590.

[14] van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for Macromolecular Dynamics and Constraint Dynamics. *Mol. Phys.* **1977**, *34*, 1311–1327.

[15] van Gunsteren, W. F.; Berendsen, H. J. C. Computer Simulation of Molecular Dynamics: Methodology, Applications and Perspectives in Chemistry. *Angew. Chem. Int. Ed.* **1990**, *29*, 992–1023.

[16] Hansson, T.; Oostenbrink, C.; van Gunsteren, W. F. Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **2002**, *12*, 190–196.

[17] Karplus, M.; McCammon, J. A. Molecular Dynamics Simulations of Biomolecules. *Nature Struct. Biol.* **2002**, *9*, 646–652.

[18] van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. Biomolecular Modelling: Goals, Problems, Perspectives. *Angew. Chem. Int. Ed.* **2006**, *45*, 4064–4092.

[19] van Gunsteren, W. F.; Dolenc, J. Biomolecular Simulation: Historical Picture and Future Perspectives. *Biochem. Soc. Trans.* **2008**, *36*, 11–15.

[20] Weiner, P. K.; Kollman, P. A. AMBER: Assisted Model Building With Energy Refinement. A General Program for Modeling Molecules and Their Interactions. *J. Comput. Chem.* **1981**, *2*, 287–303.

[21] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 198–210.

[22] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

[23] Brooks, B. R.; Brooks III, C. L.; Mackerell Jr., A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545–1614.

[24] Bekker, H.; Berendsen, H. J. C.; Dijkstra, E. J.; Achterop, S.; van Drunen, R.; van der Spoel, D.; Sijbers, A.; Keegstra, H. GROMACS: A Parallel Computer for Molecular Dynamics Simulations. Physics Computing 92. 1993; pp 252–256.

[25] Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism From Laptops to Supercomputers. *SoftwareX* **2015**, *1*, 19–25.

[26] van Gunsteren, W. F.; Billeter, S. R.; Eising, A. A.; Hünenberger, P. H.; Krüger, P.; Mark, A. E.; Scott, W. R. P.; Tironi, I. G. *Biomolecular Simulation: The GROMOS96 Manual and User Guide*; Vdf Hochschulverlag AG an der ETH Zürich, Zürich, Switzerland, 1996.

[27] Schmid, N.; Christ, C. D.; Christen, M.; Eichenberger, A. P.; van Gunsteren, W. F. Architecture, Implementation and Parallelization of the GROMOS Software for Biomolecular Simulation. *Comp. Phys. Comm.* **2012**, *183*, 890–903.

[28] Eastman, P.; Pande, V. OpenMM: A Hardware-Independent Framework for Molecular Simulations. *Comput. Sci. Eng.* **2010**, *12*, 34–39.

[29] Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput. Biol.* **2017**, *13*, e1005659.

[30] Meier, K.; Choutko, A.; Dolenc, J.; Eichenberger, A. P.; Riniker, S.; van Gunsteren, W. F. Multi-Resolution Simulation of Biomolecular Systems: A Review of Methodological Issues. *Angew. Chem. Int. Ed. Engl.* **2013**, *52*, 2820–2834.

[31] Christen, M.; van Gunsteren, W. F. On Searching In, Sampling Of, and Dynamically Moving Through Conformational Space of Biomolecular Systems: A Review. *J. Comput. Chem.* **2008**, *29*, 157–166.

[32] Hünenberger, P. H. Thermostat Algorithms for Molecular Dynamics Simulations. *Adv. Polym. Sci.* **2005**, *173*, 105–149.

[33] Riniker, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. *J. Chem. Inf. Model.* **2018**, *58*, 565–578.

[34] The GROMOS Software for (Bio)Molecular Simulation, Volume 2. 2021.

[35] van Gunsteren, W. F.; Berendsen, H. J. C.; Rullmann, J. A. C. Stochastic Dynamics for Molecules With Constraints. Brownian Dynamics of N-Alkanes. *Mol. Phys.* **1981**, *44*, 69–95.

[36] van Gunsteren, W. F.; Berendsen, H. J. C. Algorithms for Brownian Dynamics. *Mol. Phys.* **1982**, *45*, 637–647.

[37] van Gunsteren, W. F.; Berendsen, H. J. C. A Leap-Frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1*, 173–185.

[38] Leimkuhler, B.; Matthews, C. Robust and Efficient Configurational Molecular Sampling via Langevin Dynamics. *J. Chem. Phys.* **2013**, *138*, 05B601_1.

[39] Leimkuhler, B.; Matthews, C. Rational Construction of Atochastic Numerical Methods for Molecular Sampling. *Appl. Math. Res. Express* **2013**, *2013*, 34–56.

[40] Zhang, Z.; Liu, X.; Yan, K.; Tuckerman, M. E.; Liu, J. Unified Efficient Thermostat Scheme for the Canonical Ensemble With Holonomic or Isokinetic Constraints via Molecular Dynamics. *J. Phys. Chem. A* **2019**, *123*, 6056–6079.

[41] Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

[42] Hastings, W. K. Monte Carlo Sampling Methods Using Markov Chains and Their Applications. *Biometrika* **1970**, *57*, 97–109.

[43] Andersen, H. C. Molecular Dynamics Simulations at Constant Pressure and/or Temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

[44] Parrinello, M.; Rahman, A. Crystal Structure and Pair Potentials: A Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196.

[45] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular Dynamics With Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[46] Nosé, S. A Molecular Dynamics Method for Simulations in the Canonical Ensemble. *Mol. Phys.* **1984**, *52*, 255–268.

[47] Nosé, S. A Unified Formulation of the Constant Temperature Molecular Dynamics Methods. *J. Chem. Phys.* **1984**, *81*, 511–519.

[48] Hoover, W. G. Canonical Dynamics: Equilibrium Phase-Space Distributions. *Phys. Rev. A* **1985**, *31*, 1695–1697.

[49] Martyna, G. J.; Klein, M. L.; Tuckerman, M. Nosé-Hoover Chains: The Canonical Ensemble via Continuous Dynamics. *J. Chem. Phys.* **1992**, *97*, 2635–2643.

[50] Monticelli, L.; Tieleman, D. P. Force Fields for Classical Molecular Dynamics. *Methods Mol. Biol.* **2013**, 197–213.

[51] Nerenberg, P. S.; Head-Gordon, T. New Developments in Force Fields for Biomolecular Simulations. *Curr. Opin. Struct. Biol.* **2018**, *49*, 129–138.

[52] Lifson, S.; Warshel, A. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *J. Chem. Phys.* **1968**, *49*, 5116–5129.

[53] Levitt, M.; Lifson, S. Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. *J. Mol. Biol.* **1969**, *46*, 269–279.

[54] Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Jr., S. P.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids & Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.

[55] Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

[56] Tian, C.; Kasavajhala, K.; Belfon, K. A.; Raguette, L.; Huang, H.; Migues, A. N.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; Simmerling, C. ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained Against Quantum Mechanics Energy Surfaces in Solution. *J. Chem. Theory Comput.* **2019**, *16*, 528–552.

[57] Klauda, J. B.; Venable, R. M.; Freites, J. A.; O'Connor, J. W.; Tobias, D. J.; Mondragon-Ramirez, C.; Vorobyov, I.; Jr., A. D. M.; Pastor, R. W. Update of the CHARMM All-Atom Additive Force Field for Lipids: Validation on Six Lipid Types. *J. Phys. Chem. B* **2010**, *114*, 7830–7843.

[58] Mallajosyula, S. S.; Guvench, O.; Hatcher, E.; Jr., A. D. M. CHARMM Additive All-Atom Force Field for Phosphate and Sulfate Linked to Carbohydrates. *J. Chem. Theory Comput.* **2012**, *8*, 759–776.

[59] Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr, A. D. CHARMM General Force Field: A Force Field for Drug-Like Molecules Compatible With the CHARMM All-Atom Additive Biological Force Fields. *J. Comput. Chem.* **2010**, *31*, 671–690.

[60] Vanommeslaeghe, K.; MacKerell Jr, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144–3154.

[61] Vanommeslaeghe, K.; Raman, E. P.; MacKerell Jr, A. D. Automation of the CHARMM General Force Field (CGenFF) II: Assignment of Bonded Parameters and Partial Atomic Charges. *J. Chem. Inf. Model.* **2012**, *52*, 3155–3168.

[62] Hermans, J.; Berendsen, H. J. C.; van Gunsteren, W. F.; Postma, J. P. M. A Consistent Empirical Potential for Water-Protein Interactions. *Biopolymers* **1984**, *23*, 1513–1518.

[63] Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An Improved GROMOS96 Force Field for Aliphatic Hydrocarbons in the Condensed Phase. *J. Comput. Chem.* **2001**, *22*, 1205–1218.

[64] Chandrasekhar, I.; Kastenholz, M. A.; Lins, R. D.; Oostenbrink, C.; Schuler, L. D.; Tieleman, D. P.; van Gunsteren, W. F. A Consistent Potential Energy Parameter Set for Lipids: Dipalmitoylphosphatidylcholine as a Benchmark of the GROMOS96 45A3 Force Field. *Eur. Biophys. J.* **2003**, *32*, 67–77.

[65] Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: the GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656–1676.

[66] Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions: 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843–856.

[67] Reif, M. M.; Hünenberger, P. H.; Oostenbrink, C. New Interaction Parameters for Charged Amino Acid Side Chains in the GROMOS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 3705–3723.

[68] Malde, A. K.; Zuo, L.; Breeze, M.; Stroet, M.; Poger, D.; Nair, P. C.; Oostenbrink, C.; Mark, A. E. An Automated Force Field Topology Builder (ATB) and Repository: Version 1.0. *J. Chem. Theory Comput.* **2011**, *7*, 4026–4037.

[69] Koziara, K. B.; Stroet, M.; Malde, A. K.; Mark, A. E. Testing and Validation of the Automated Topology Builder (ATB) Version 2.0: Prediction of Hydration Free Enthalpies. *J. Comput.-Aided Mol. Des.* **2014**, *28*, 221–233.

[70] Hansen, H. S.; Hünenberger, P. H. A Reoptimized GROMOS Force Field for Hexopyranose-Based Carbohydrates Accounting for the Relative Free Energies of Ring Conformers, Anomers, Epimers, Hydroxymethyl Rotamers and Glycosidic Linkage Conformers. *J. Comput. Chem.* **2011**, *32*, 998–1032.

[71] Plazinski, W.; Lonardi, A.; Hünenberger, P. H. Revision of the GROMOS 56A6$_{CARBO}$ Force Field: Improving the Description of Ring-Conformational Equilibria in Hexopyranose-Based Carbohydrates Chains. *J. Comput. Chem.* **2016**, *37*, 354–365.

[72] Horta, B. A. C.; Merz, P. T.; Fuchs, P.; Dolenc, J.; Riniker, S.; Hünenberger, P. H. A GROMOS-Compatible Force Field for Small Organic Molecules in the Condensed Phase: The 2016H66 Parameter Set. *J. Chem. Theory Comput* **2016**, *12*, 3825–3850.

[73] Oliveira, M. P.; Andrey, M.; Rieder, S. R.; Kern, L.; Hahn, D. F.; Riniker, S.; Horta, B. A. C.; Hūnenberger, P. H. Systematic Optimization of a Fragment-Based Force Field Against Experimental Pure-Liquid Properties Considering Large Compound Families: Application to Saturated Haloalkanes. *J. Chem. Theory Comput.* **2020**, *16*, 7525–7555.

[74] Oliveira, M. P.; Hünenberger, P. H. Systematic Optimization of a Fragment-Based Force Field Against Experimental Pure-Liquid Properties Considering Large Compound Families: Application to Oxygen and Nitrogen Compounds. *Phys. Chem. Chem. Phys.* **2021**, *23*, 17774–17793.

[75] Mobley, D. L.; Bannan, C. C.; Rizzi, A.; Bayly, C. I.; Chodera, J. D.; Lim, V. T.; Lim, N. M.; Beauchamp, K. A.; Slochower, D. R.; Shirts, M. R.; Gilson, M. K.; Eastman, P. K. Escaping Atom Types in Force Fields Using Direct Chemical Perception. *J. Chem. Theory Comput.* **2018**, *14*, 6076–6092.

[76] Qiu, Y.; Smith, D. G. A.; Boothroyd, S.; Jang, H.; Hahn, D. F.; Wagner, J.; Bannan, C. C.; Gokey, T.; Lim, V. T.; Stern, C. D.; Rizzi, A.; Tjanaka, B.; Tresadern, G.; Lucas, X.; Shirts, M. R.; Gilson, M. K.; Chodera, J. D.; Bayly, C. I.; Mobley, D. L.; Wang, L.-P. Development and Benchmarking of Open Force Field v1.0.0 – the Parsley Small-Molecule Force Field. *J. Chem. Theory Comput.* **2021**, *17*, 6262–6280.

[77] Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657–1666.

[78] Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

[79] Kaminski, G. A.; Friesner, R. A.; Tirado-Rives, J.; Jorgensen, W. L. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison With Accurate Quantum Chemical Calculations on Peptides. *J. Phys. Chem. B* **2001**, *105*, 6474–6487.

[80] Shivakumar, D.; Harder, E.; Damm, W.; Friesner, R. A.; Sherman, W. Improving the Prediction of Absolute Solvation Free Energies Using the Next Generation OPLS Force Field. *J. Chem. Theory Comput.* **2012**, *8*, 2553–2558.

[81] Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jorgensen, W. L.; Abel, R.; Friesner, R. A. OPLS3: A Force Field Providing Broad

Coverage of Drug-Like Small Molecules and Proteins. *J. Chem. Theory Comput.* **2016**, *12*, 281–296.

[82] Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. OPLS All-Atom Force Field for Carbohydrates. *J. Comput. Chem.* **1997**, *18*, 1955–1970.

[83] Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. A Fast SHAKE Algorithm to Solve Distance Constraint Equations for Small Molecules in Molecular Dynamics Simulations. *J. Comput. Chem.* **2001**, *22*, 501–508.

[84] Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System With Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

[85] Miyamoto, S.; Kollman, P. A. Settle: An Analytical Version of the SHAKE and RATTLE Algorithm for Rigid Water Models. *J. Comput. Chem.* **1992**, *13*, 952–962.

[86] Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A Linear Constraint Solver for Molecular Simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

[87] Eastman, P.; Pande, V. S. Constant Constraint Matrix Approximation: A Robust, Parallelizable Constraint Method for Molecular Simulations. *J. Chem. Theory Comput.* **2010**, *6*, 434–437.

[88] Tao, P.; Wu, X.; Brooks, B. R. Maintain Rigid Structures in Verlet Based Cartesian Molecular Dynamics Simulations. *J. Chem. Phys.* **2012**, *137*, 134110.

[89] van Gunsteren, W. F. Constrained Dynamics of Flexible Molecules. *Mol. Phys.* **1980**, *40*, 1015–1019.

[90] van Gunsteren, W. F.; Karplus, M. Effect of Constraints on the Dynamics of Macromolecules. *Macromolecules* **1982**, *15*, 1528–1544.

[91] Canzar, S.; El-Kebir, M.; Rool, R.; Elbassioni, K.; Malde, A. K.; Mark, A. E.; Geerke, D. P.; Stougie, L.; Klau, G. W. Charge Group Partitioning in Biomolecular Simulation. *RECOMB* **2012**, *2012*, 29–43.

[92] Gonçalves, Y. M. H.; Senac, C.; Fuchs, P. F. J.; Hünenberger, P. H.; Horta, B. A. C. Influence of the Treatment of Non-Bonded Interactions on the Thermodynamic and Transport Properties of Pure Liquids Calculated Using the 2016H66 Force Field. *J. Chem. Theory Comput.* **2019**, *15*, 1806–1826.

[93] Wu, X.; Brooks, B. R. Isotropic Periodic Sum: A Method for the Calculation of Long-Range Interactions. *J. Chem. Phys.* **2005**, *122*, 044107.

[94] Klauda, J. B.; Wu, X.; Pastor, R. W.; Brooks, B. R. Long-Range Lennard-Jones and Electrostatic Interactions in Interfaces: Application of the Isotropic Periodic Sum Method. *J. Phys. Chem. B* **2007**, *111*, 4393–4400.

[95] Wu, X.; Brooks, B. R. Using the Isotropic Periodic Sum Method to Calculate Long-Range Interactions of Heterogeneous Systems. *J. Chem. Phys.* **2008**, *129*, 154115.

[96] Wu, X.; Brooks, B. R. The Homogeneity Condition: A Simple Way to Derive Isotropic Periodic Sum Potentials for Efficient Calculation of Long-Range Interactions in Molecular Simulation. *J. Chem. Phys.* **2019**, *150*, 214109.

[97] Takahashi, K.; Yasuoka, K.; Narumi, T. Cutoff Radius Effect of Isotropic Periodic Sum Method for Transport Coefficients of Lennard-Jones Liquid. *J. Chem. Phys.* **2007**, *127*, 114511.

[98] Williams, D. E. Accelerated Convergence of Crystal-Lattice Potential Sums. *Acta Crystallogr. A* **1971**, *27*, 452–455.

[99] Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

[100] Shi, B.; Sinha, S.; Dir, V. K. Molecular Dynamics Simulation of the Density and Surface Tension by Particle-Particle Particle-Mesh Method. *J. Chem. Phys.* **2006**, *124*, 204715.

[101] in 't Veld, P. J.; Ismail, A. E.; Grest, G. S. Application of Ewald Summation to Long-Range Dispersion Forces. *J. Chem. Phys.* **2007**, *127*, 144711.

[102] Isele-Holder, R. E.; Mitchell, W.; Ismail, A. E. Development and Application of a Particle-Particle Particle-Mesh Ewald Method for Dispersion Interactions. *J. Chem. Phys.* **2012**, *137*, 174107.

[103] Wennberg, C. L.; Murtola, T.; Hess, B.; Lindahl, E. Lennard-Jones Lattice Summation in Bilayer Simulations Has Critical Effects on Surface Tension and Lipid Properties. *J. Chem. Theory Comput.* **2013**, *9*, 3527–3537.

[104] Wells, B. A.; Chaffee, A. L. Ewald Summation for Molecular Simulations. *J. Chem. Theory Comput.* **2015**, *11*, 3684–3695.

[105] Wennberg, C. L.; Murtola, T.; Páll, S.; Abraham, M. J.; Hess, B.; Lindahl, E. Direct-Space Corrections Enable Fast and Accurate Lorentz-Berthelot Combination Rule Lennard-Jones Lattice Summation. *J. Chem. Theory Comput.* **2015**, *11*, 5737–5746.

[106] Leonard, A. N.; Simmonett, A. C.; Pickard, F. C.; Huang, J.; Venable, R. M.; Klauda, J. B.; Brooks, B. R.; Pastor, R. W. Comparison of Additive and Polarizable Models With Explicit Treatment of Long-Range Lennard-Jones Interactions Using Alkane Simulations. *J. Chem. Theory Comput.* **2018**, *14*, 948–958.

[107] Brooks III, C. L. The Influence of Long-Range Force Truncation on the Thermodynamics of Aqueous Ionic Solutions. *J. Chem. Phys.* **1987**, *86*, 5156–5162.

[108] Schreiber, H.; Steinhauser, O. Cutoff Size Does Strongly Influence Molecular Dynamics Results on Solvated Polypeptides. *Biochemistry* **1992**, *31*, 5856–5860.

[109] Spohr, E. Effect of Electrostatic Boundary Conditions and System Size on the Interfacial Properties of Water and Aqueous Solutions. *J. Chem. Phys.* **1997**, *107*, 6342–6348.

[110] Hünenberger, P. H.; van Gunsteren, W. F. Alternative Schemes for the Inclusion of a Reaction-Field Correction Into Molecular Dynamics Simulations: Influence on the Simulated Energetic, Structural, and Dielectric Properties of Liquid Water. *J. Chem. Phys.* **1998**, *108*, 6117–6134.

[111] Reif, M. M.; Kräutler, V.; Kastenholz, M. A.; Daura, X.; Hünenberger, P. H. Molecular Dynamics Simulations of a Reversibly Folding β-Heptapeptide in Methanol: Influence of the Treatment of Long-Range Electrostatic Interactions. *J. Phys. Chem. B* **2009**, *113*, 3112–3128.

[112] Barker, J. A.; Watts, R. O. Monte Carlo Studies of the Dielectric Properties of Water-Like Models. *Mol. Phys.* **1973**, *26*, 789–792.

[113] Onsager, L. Electric Moments of Molecules in Liquids. *J. Am. Chem. Soc.* **1936**, *58*, 1486–1493.

[114] Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **1995**, *102*, 5451–5459.

[115] Tironi, I. G.; van Gunsteren, W. F. A Molecular Dynamics Simulation Study of Chloroform. *Mol. Phys.* **1994**, *83*, 381–403.

[116] Ewald, P. P. Ewald Summation. *Ann. Phys.* **1921**, *369*, 1–2.

[117] Hockney, R.; Eastwood, J. *Computer Simulation Using Particles*; CRC Press, 2021.

[118] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An Nlog(N) Method for Ewald Sums in Large Systems. *SoftwareX* **1993**, *98*, 10089–10092.

[119] Jorgensen, W. L. Transferable Intermolecular Potential Functions for Water, Alcohols, and Ethers. Application to Liquid Water. *J. Am. Chem. Soc.* **1981**, *103*, 335–340.

[120] van Gunsteren, W. F.; Berendsen, H. J. C. *Groningen Molecular Simulation (GROMOS) Library Manual*; BIOMOS, Groningen, The Netherlands, 1987.

[121] Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraint for Deriving Atomic Charges: The RESP Model. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

[122] Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W. Y.; R., C.; P., L.; R., L.; T., C.; J., W.; J., K.; P., A Point-Charge Force Field for Molecular Mechanics Simulations of Proteins Based on Condensed-Phase Quantum Mechanical Calculations. *J. Comput. Chem.* **2003**, *24*, 1999–2012.

[123] Reith, D.; Kirschner, K. N. A Modern Workflow for Force-Field Development – Bridging Quantum Mechanics and Atomistic Computational Models. *Comput. Phys. Commun.* **2011**, *182*, 2184–2191.

[124] Grimme, S. A General Quantum Mechanically Derived Force Field (QMDFF) for Molecules and Condensed Phase Simulations. *J. Chem. Theory Comput.* **2014**, *10*, 4497–4514.

[125] Prampolini, G.; Campetella, M.; Mitri, N. D.; Livotto, P. R.; Cacelli, I. Systematic and Automated Development of Quantum Mechanically Derived Force Fields: The Challenging Case of Halogenated Hydrocarbons. *Chem. Theory Comput.* **2016**, *12*, 5525–5540.

[126] Piquemal, J. P.; Jordan, K. D. Preface: Special Topic – From Quantum Mechanics to Force Fields. *J. Chem. Phys.* **2017**, *147*, 161401.

[127] Xu, P.; Guidez, E. B.; Bertoni, C.; Gordon, M. S. Perspective: *Ab initio* Force Field Methods Derived From Quantum Mechanics. *J. Chem. Phys.* **2018**, *148*, 090901.

[128] Horton, J. T.; Allen, A. E. A.; Dodda, L. S.; Cole, D. J. QUBEKit: Automating the Derivation of Force Field Parameters From Quantum Mechanics. *J. Chem. Inf. Model.* **2019**, *59*, 1366–1381.

[129] Allen, A. E. A.; Robertson, M. J.; Payne, M. C.; Cole, D. J. Development and Validation of the Quantum Mechanical Bespoke Protein Force Field. *ACS Omega* **2019**, *4*, 14537–14550.

[130] Galvelis, R.; Doerr, S.; Damas, J. M.; Harvey, M. J.; Fabritiis, G. D. A Scalable Molecular Force Field Parameterization Method Based on Density Functional Theory and Quantum-Level Machine Learning. *J. Chem. Inf. Model.* **2019**, *59*, 3485–3493.

[131] Kantonen, S. M.; Muddana, H. S.; Schauperl, M.; Henriksen, N. M.; Wang, L. Data-Driven Mapping of Gas-Phase Quantum Calculations to General Force Field Lennard-Jones Parameters. *J. Chem. Theory Comput.* **2020**, *16*, 1115–1127.

[132] Sami, S.; Menger, M. F.; Faraji, S.; Broer, R.; Havenith, R. W. Q-Force: Quantum mechanically augmented molecular force fields. *J. Chem. Theory Comput.* **2021**, *17*, 4946–4960.

[133] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Antechamber: An Accessory Software Package for Molecular Mechanical Calculations. *J. Am. Chem. Soc.* **2001**, *222*, U403.

[134] Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. Automatic Atom Type and Bond Type Perception in Molecular Mechanical Calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260.

[135] Stroet, M.; Caron, B.; Visscher, K. M.; Geerke, D. P.; Malde, A. K.; Mark, A. E. Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane. *J. Chem. Theory Comput.* **2018**, *14*, 5834–5845.

[136] Huang, L.; Roux, B. Automated Force Field Parameterization for Nonpolarizable and Polarizable Atomic Models Based on ab initio Target Data. *J. Chem. Theory Comput.* **2013**, *9*, 3543–3556.

[137] Boulanger, E.; Huang, L.; Rupakheti, C.; MacKerell Jr, A. D.; Roux, B. Optimized Lennard-Jones Parameters for Druglike Small Molecules. *J. Chem. Theory Comput.* **2018**, *14*, 3121–3131.

[138] Dodda, L. S.; Cabeza de Vaca, I.; Tirado-Rives, J.; Jorgensen, W. L. LigParGen Web Server: An Automatic OPLS-AA Parameter Generator for Organic Ligands. *Nucleic Acids Res.* **2017**, *45*, W331–W336.

[139] Schüttelkopf, A. W.; van Aalten, D. M. F. PRODRG: A Tool for High-Throughput Crystallography of Protein–Ligand Complexes. *Acta Crystallogr. D* **2004**, *60*, 1355–1363.

[140] Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F.-Y. RED Server: A Web service for Deriving RESP and ESP Charges and Building Force Field Libraries for New Molecules and Molecular Fragments. *Nucleic Acids Res.* **2011**, *39*, W511–W517.

[141] Zoete, V.; Cuendet, M. A.; Grosdidier, A.; Michielin, O. SwissParam: A Fast Force Field Generation Tool for Small Organic Molecules. *J. Comput. Chem.* **2011**, *32*, 2359–2368.

[142] Oliveira, M. P.; Gonçalves, Y. M. H.; Kashefolgheta, S.; Rieder, S. R.; Horta, B. A. C.; Hünenberger, P. H. Comparison of the United- and All-Atom Representation of (Halo)Alkanes Based on Two Condensed-Phase Force Fields Optimized Against the Same Experimental Data Set. *J. Chem. Theory Comput.* **2022**, submitted.

[143] Michel, J.; Foloppe, N.; Essex, J. W. Rigorous Free Energy Calculations in Structure-Based Drug Design. *Mol. Inform.* **2010**, *29*, 570–578.

[144] Borhani, D. W.; Shaw, D. E. The Future of Molecular Dynamics Simulations in Drug Discovery. *J. Comput. Aided* **2012**, *26*, 15–26.

[145] De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59*, 4035–4061.

[146] Cournia, Z.; Allen, B.; Sherman, W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.* **2017**, *57*, 2911–2937.

[147] Williams-Noonan, B. J.; Yuriev, E.; Chalmers, D. K. Free Energy Methods in Drug Design: Prospects of "Alchemical Perturbation" in Medicinal Chemistry. *J. Med. Chem.* **2018**, *61*, 638–649.

[148] Armacost, K. A.; Riniker, S.; Cournia, Z. Novel Directions in Free Energy Methods and Applications. *J. Chem. Inf. Model.* **2020**, *60*, 1–5.

[149] Cournia, Z.; Allen, B. K.; Beuming, T.; Pearlman, D. A.; Radak, B. K.; Sherman, W. Rigorous Free Energy Simulations in Virtual Screening. *J. Chem. Inf. Model.* **2020**, *60*, 4153–4169.

[150] Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee Jr, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5595–5623.

[151] Song, L. F.; Merz Jr, K. M. Evolution of Alchemical Free Energy Methods in Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60*, 5308–5318.

[152] Christ, C. D.; Mark, A. E.; van Gunsteren, W. F. Basic Ingredients of Free Energy Calculations: A Review. *J. Comput. Chem.* **2009**, *31*, 1569–1582.

[153] Zwanzig, R. W. High-Temperature Equation of State by a Pertubation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420–1426.

[154] Shirts, M. R.; Mobley, D. L.; Chodera, J. D. Chapter 4 Alchemical Free Energy Calculations: Ready for Prime Time? *Annu. Rep. Comput. Chem.* **2007**, *3*, 41–59.

[155] Duarte Ramos Matos, G.; Kyu, D. Y.; Loeffler, H. H.; Chodera, J. D.; Shirts, M. R.; Mobley, D. L. Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated With an Update of the FreeSolv Database. *J. Chem. Eng. Data* **2017**, *62*, 1559–1569.

[156] Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300–313.

[157] Bennett, C. H. Efficient Estimation of Free Energy Differences From Monte Carlo Data. *J. Comput. Phys.* **1976**, *22*, 245–268.

[158] Shirts, M. R.; Chodera, J. D. Statistically Optimal Analysis of Samples From Multiple Equilibrium States. *J. Chem. Phys.* **2008**, *129*, 124105.

[159] Knight, J. L.; Brooks III, C. L. Multisite λ Dynamics for Simulated Structure–Activity Relationship Studies. *J. Chem. Theory Comput.* **2011**, *7*, 2728–2739.

[160] Raman, E. P.; Paul, T. J.; Hayes, R. L.; Brooks III, C. L. Automated, Accurate, and Scalable Relative Protein–Ligand Binding Free-Energy Calculations Using λ Dynamics. *J. Chem. Theory Comput.* **2020**, *16*, 7895–7914.

[161] Hayes, R. L.; Buckner, J.; Brooks III, C. L. BLaDE: A Basic Lambda Dynamics Engine for GPU-Accelerated Molecular Dynamics Free Energy Calculations. *J. Chem. Theory Comput.* **2021**, *17*, 6799–6807.

[162] Christ, C. D.; van Gunsteren, W. F. Enveloping Distribution Sampling: A Method to Calculate Free Energy Differences From a Single Simulation. *J. Chem. Phys.* **2007**, *126*, 184110.

[163] Christ, C. D.; van Gunsteren, W. F. Multiple Free Energies From a Single Simulation: Extending Enveloping Distribution Sampling to Nonoverlapping Phase-space Distributions. *J. Chem. Phys.* **2008**, *128*, 174112.

[164] Sidler, D.; Schwaninger, A.; Riniker, S. Replica Exchange Enveloping Distribution Sampling (RE-EDS): A Robust Method to Estimate Multiple Free-Energy Differences From a Single Simulation. *J. Chem. Phys.* **2016**, *145*, 154114.

[165] Sidler, D.; Cristòfol-Clough, M.; Riniker, S. Efficient Round-Trip Time Optimization for Replica-Exchange Enveloping Distribution Sampling (RE-EDS). *J. Chem. Theory Comput.* **2017**, *13*, 3020–3030.

[166] Ries, B.; Normak, K.; Weiss, R. G.; Rieder, S.; Barros, E. P.; Champion, C.; König, G.; Riniker, S. Relative Free-Energy Calculations for Scaffold Hopping-Type Transformations With an Automated RE-EDS Sampling Procedure. *J. Comput. Aided Mol. Des.* **2022**, *36*, 117–130.

[167] Perthold, J. W.; Oostenbrink, C. Accelerated Enveloping Distribution Sampling: Enabling Sampling of Multiple End States While Preserving Local Energy Minima. *J. Phys. Chem. B* **2018**, *122*, 5030–5037.

[168] Perthold, J. W.; Petrov, D.; Oostenbrink, C. Toward Automated Free Energy Calculation With Accelerated Enveloping Distribution Sampling (A-EDS). *J. Chem. Inf. Model.* **2020**, *60*, 5395–5406.

[169] König, G.; Glaser, N.; Schroeder, B.; Kubincová, A.; Hünenberger, P. H.; Riniker, S. An Alternative to Conventional λ-Intermediate States in Alchemical Free Energy Calculations: λ-Enveloping Distribution Sampling. *J. Chem. Inf. Model.* **2020**, *60*, 5407–5423.

[170] Guthrie, J. P. A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. *J. Phys. Chem. B* **2009**, *113*, 4501–4507.

[171] Klimkovich, P. V.; Mobley, D. L. Predicting Hydration Free Energies Using All-Atom Molecular Dynamics Simulations and Multiple Starting Conformations. *J. Comput. Aided Mol. Des.* **2010**, *24*, 307–316.

[172] Kashef Ol Gheta, S.; Oliveira, M. P.; Rieder, S. R.; Horta, B. A. C.; Acree, W. E.; Hünenberger, P. H. Evaluating Classical Force Fields Against Experimental Cross-Solvation Free Energies. *J. Chem. Theory Comput.* **2020**, *16*, 7556–7580.

[173] Mobley, D. L.; Guthrie, J. P. FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, With Input Files. *J. Comput. Aided Mol. Des.* **2014**, *28*, 711–720.

[174] Marenich, A. V.; Kelly, C. P.; Thompson, J. D.; Hawkins, G. D.; Chambers, C. C.; Giesen, D. J.; Winget, P.; Cramer, C. J.; Truhlar, D. G. Minnesota Solvation Database, version 20. 2012.

[175] Rieder, S. R.; Ries, B.; Schaller, K.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Replica-Exchange Enveloping Distribution Sampling Using Generalized AMBER Force-Field Topologies: Application to Relative Hydration Free-Energy Calculations for Large Sets of Molecules. *J. Chem. Inf. Model.* **2022**, *62*, 3043–3056.

[176] Johnson, K. *Reaching for the Moon: The Autobiography of NASA Mathematician Katherine Johnson*; Simon and Schuster, 2019.

[177] Meringer, M. In *Handbook of Chemoinformatics Algorithms*; Faulon, J.-L., Bender, A., Eds.; Chapman & Hall/CRC: London, 2010; Chapter 8, pp 233–267.

[178] Klein, D. J.; Babić, D.; Trinajstić, N. In *Chemical Modelling: Applications and Theory, Volume 2*; Hinchliffe, A., Ed.; The Royal Society of Chemistry, 2002; Chapter 2, pp 56–95.

[179] Pólya, G. Kombinatorische Anzahlbestimmungen für Gruppen, Graphen und chemische Verbindungen. *Acta Mathematica* **1937**, *68*, 145–254.

[180] Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. *Applications of Artificial Intelligence for Organic Chemistry. The DENDRAL Project*; McGraw-Hill Companies, Inc.: New York, 1980.

[181] Faulon, J.-L.; Visco, D. P.; Roe, D. Enumerating Molecules. *Reviews in Computational Chemistry* **2005**, *21*, 209–286.

[182] Brown, H.; Masinter, L. *An Algorithm for the Construction of the Graphs of Organic Molecules*; Stanford University: Stanford, 1973.

[183] Melnikov, A. A.; Palyulin, V. A.; Zefirov, N. S. Generation of Molecular Graphs for QSAR Studies: An Approach Based on Supergraphs. *J. Chem. Inf. Model* **2007**, *47*, 2077–2088.

[184] Molchanova, M. S.; Shcherbukhin, V. V.; Zefirov, N. S. Computer Generation of Molecular Structures by the SMOG Program. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 888–899.

[185] Grund, R. Konstruktion Molekularer Graphen mit Gegebenen Hybridisierungen und Überlappungsfreien Fragmenten. Ph.D. thesis, Lehrstuhl II für Mathematik der Universität Bayreuth, 1994.

[186] Gugisch, R.; Kerber, A.; Kohnert, A.; Laue, R.; Meringer, M.; Rücker, C.; Wassermann, A. In *Advances in Mathematical Chemistry and Applications: Revised Edition*; Basak, S. C., Restrepo, G., Villaveces, J. L., Eds.; Bentham Science Publishers: Sharjah, 2015; Vol. 1; Chapter 6, pp 113–138.

[187] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J. L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.

[188] Yirik, M. A.; Sorokina, M.; Steinbeck, C. MAYGEN: An Open-Source Chemical Structure Generator for Constitutional Isomers Based on the Orderly Generation Principle. *J. Cheminform.* **2021**, *13*, 1–14.

[189] Stroustrup, B. *The C++ Programming Language*, 4th ed.; Addison-Wesley, 2013.

[190] MOLGEN. `https://www.molgen.de/`, Accessed: 2022/06/15.

[191] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

[192] Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.

[193] Schneider, N.; Sayle, R. A.; Landrum, G. A. Get Your Atoms in Order – An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm. *J. Chem. Inf. Model.* **2015**, *55*, 2111–2120.

[194] Walthard, S. R. Design and Implementation of an Algorithm for the Systematic Isomer Enumeration of Small Organic Molecules. M.Sc. thesis, ETH Zürich, 2018.

[195] Biggs, N. L.; Lloyd, E. K.; Wilson, R. J. *Graph Theory, 1736–1936*; Oxford University Press: Oxford, 1998.

[196] Foulds, L. R. *Graph Theory Applications*; Springer Science + Business Media: New York, 1992.

[197] McNaught, A. D.; Wilkinson, A. *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book")*; Blackwell Scientific Publications: Oxford, 1997; p 951.

[198] Read, R. C. In *Annals of Discrete Mathematics*; Alspach, B., Hell, P., Miller, D. J., Eds.; 1978; Vol. 2; pp 107–120.

[199] Faradzhev, I. A. Generation of Nonisomorphic Graphs With a Given Degree Sequence. *Algorithmic Studies in Combinatorics* **1978**, 11–19.

[200] Faradzhev, I. A. Constructive Enumeration of Combinatorial Objects. Problèmes Combinatoires et Théorie des Graphes. Colloq. Internat. CNRS, University of Orsay, Orsay. 1978; pp 131–135.

[201] Ottmann, T.; Widmayer, P. *Algorithmen und Datenstrukturen*; Spektrum Akademischer Verlag: Heidelberg, 2012.

[202] Daylight Chemical Information Systems - SMILES. `https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html`, Accessed: 2022/04/20.

[203] Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures. A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.

[204] Faulon, J.-L.; Collins, M. J.; Carr, R. D. The Signature Molecular Descriptor. 4. Canonizing Molecules Using Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 427–436.

[205] O'Boyle, N. M. Towards a Universal SMILES Representation – A Standard Method to Generate Canonical SMILES Based on the InChI. *J. Cheminform.* **2012**, *4*, 1–14.

[206] Krotko, D. G. Atomic Ring Invariant and Modified CANON Extended Connectivity Algorithm for Symmetry Perception in Molecular Graphs and Rigorous Canonicalization of SMILES. *J. Cheminform.* **2020**, *12*, 1–11.

[207] Razinger, M.; Balasubramanian, K.; Perdih, M.; Munk, M. E. Stereoisomer Generation in Computer-Enhanced Structure Elucidation. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 812–825.

[208] Dijkstra, E. W. A Note on Two Problems in Connexion With Graphs. *Numer. Math.* **1959**, *1*, 269–271.

[209] CMake. `https://cmake.org/`, Accessed: 2022/04/20.

[210] Landrum, G.; Tosco, P.; Kelley, B.; Riniker, S.; Ric,; gedeck,; Vianello, R.; Schneider, N.; Dalke, A.; N, D.; Eisuke, K.; Cole, B.; Turk, S.; Swain, M.; Alexander, S.; Cosgrove, D.; Vaucher, A.; Wójcikowski, M.; Jones, G.; Probst, D.; Godin, G.; Scalfani, V. F.; Pahl, A.; Francois, B.; JLVarjo,; strets123,; JP,; DoliathGavin,; Sforna, G.; Jensen, J. H. rdkit/rdkit: 2021_03_2 (Q1 2021) Release. `https://doi.org/10.5281/zenodo.4750957`, 2021.

[211] van Rossum, G.; Drake, F. L. Python 3 Reference Manual. 2009.

[212] Maupin, P. pdfrw. 2017; `https://github.com/pmaupin/pdfrw`.

[213] Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. ACM* **1976**, *23*, 31–42.

[214] Ehrlich, H. C.; Rarey, M. Systematic Benchmark of Substructure Search in Molecular Graphs – From Ullmann to VF2. *J. Cheminform.* **2012**, *4*, 1–17.

[215] Cordella, L. P.; Foggia, P.; Sansone, C.; Vento, M. A (Sub)Graph Isomorphism Algorithm for Matching Large Graphs. *IEEE Trans. Pattern Anal. Mach. Intell* **2004**, *26*, 1367–1372.

[216] Bóna, M. *Combinatorics of Permutations*, 2nd ed.; Chapman and Hall/CRC: New York, 2012.

[217] Faulon, J.-L.; Bender, A. *Handbook of Chemoinformatics Algorithms*; Chapman & Hall/CRC: London, 2010.

[218] Plewinsky, B.; Hennecke, M.; Oppermann, W. *Das Ingenieurwissen: Chemie*; Springer Berlin Heidelberg: Berlin, 2014; Chapter 11, p 86.

[219] Badertscher, M.; Bischofberger, K.; Munk, M. E.; Pretsch, E. A Novel Formalism to Characterize the Degree of Unsaturation of Organic Molecules. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 889–893.

[220] Bhattacharya, P. The Representation of Permutations by Trees. *Computers Math. Applic.* **1994**, *28*, 67–71.

[221] Truesdell, C.; Noll, W. In *The Non-Linear Field Theories of Mechanics*; Antman, S. S., Ed.; Springer, Berlin, Heidelberg, 2004; pp 1–579.

[222] Brooks III, C. L. Methodological Advances in Molecular Dynamics Simulations of Biological Systems. *Curr. Opin. Struct. Biol.* **1995**, *5*, 211–215.

[223] Elber, R. Novel Methods for Molecular Dynamics Simulations. *Curr. Opin. Struct. Biol.* **1996**, *6*, 232–235.

[224] Norberg, J.; Nilsson, L. Advances in Biomolecular Simulations: Methodology and Recent Applications. *Quart. Rev. Biophys.* **2003**, *36*, 257–306.

[225] van Gunsteren, W. F.; Dolenc, J. Thirty-Five Years of Biomolecular Simulation: Development of Methodology, Force Fields, and Software. *Mol. Simul.* **2012**, *38*, 1271–1281.

[226] Field, M. J. Technical Advances in Molecular Simulation Since the 1980s. *Arc. Biochem. Biophys.* **2015**, *582*, 3–9.

[227] The GROMOS Software for (Bio)Molecular Simulation, Volume 4. 2021.

[228] Kleene, S. C. Representation of Events in Nerve Nets and Finite Automata. *Ann. Math. Stud.* **1956**, *34*, 3–41.

[229] Thompson, K. Programming Techniques: Regular Expression Search Algorithm. *Commun. ACM* **1968**, *11*, 419–422.

[230] Josuttis, N. M. *The C++ Standard Library: A Tutorial and Reference*; Addison-Wesley, 2012; Chapter 14.

[231] Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS Software for Biomolecular Simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719–1751.

[232] Daura, X.; Mark, A. E.; van Gunsteren, W. F. Parametrization of Aliphatic CH_n United Atoms of GROMOS96 Force Field. *J. Comput. Chem.* **1998**, *19*, 535–547.

[233] Quadbeck-Seeger, H.-J. *"Der Wechsel allein ist das Beständige": Zitate und Gedanken für innovative Führungskräfte*; Wiley-VCH, 2007.

[234] Ramey, C. Bash, the Bourne-Again Shell. Proceedings of The Romanian Open Systems Conference & Exhibition (ROSE 1994), The Romanian UNIX User's Group (GURU). 1994; pp 3–5.

[235] Joy, W. *An Introduction to the C Shell*; University of California, Berkeley, 1980.

[236] Hill, E. A. On a System of Indexing Chemical Literature; Adopted by the Classification Division of the US Patent Office. *J. Am. Chem. Soc.* **1900**, *22*, 478–494.

[237] Becker, G. S. *The Economic Way of Looking at Behavior: The Nobel Lecture*; Hoover Press, 1996.

[238] Meier, K.; Bluck, J. P.; Christ, C. D. *Use of Free Energy Methods in the Drug Discovery Industry*; ACS Symposium Series; J. Am. Chem. Soc., 2021; Vol. 1397; pp 39–66.

[239] P. Barros, E.; Ries, B.; Böselt, L.; Champion, C.; Riniker, S. Recent Developments in Multiscale Free Energy Simulations. *Curr. Opin. Struct. Biol.* **2022**, *72*, 55–62.

[240] Heinzelmann, G.; Gilson, M. Automation of Absolute Protein-Ligand Binding Free Energy Calculations for Docking Refinement and Compound Evaluation. *Sci. Rep.* **2021**, *11*, 1116.

[241] Gapsys, V.; Pérez-Benito, L.; Aldeghi, M.; Seeliger, D.; van Vlijmen, H.; Tresadern, G.; de Groot, B. L. Large Scale Relative Protein Ligand Binding Affinities Using Non-Equilibrium Alchemy. *Chem. Sci.* **2020**, *11*, 1140–1152.

[242] Jespers, W.; Esguerra, M.; Åqvist, J.; Gutiérrez-De-Terán, H. QligFEP: An Automated Workflow for Small Molecule Free Energy Calculations in Q. *J. Cheminf.* **2019**, *11*, 26.

[243] Christ, C. D.; Fox, T. Accuracy Assessment and Automation of Free Energy Calculations for Drug Design. *J. Chem. Inf. Model.* **2014**, *54*, 108–120.

[244] Gao, Y.-D.; Hu, Y.; Crespo, A.; Wang, D.; Armacost, K. A.; Fells, J. I.; Fradera, X.; Wang, H.; Wang, H.; Sherborne, B.; Verras, A.; Peng, Z. Workflows and Performances in the Ranking Prediction of 2016 D3R Grand Challenge 2: Lessons Learned From a Collaborative Effort. *J. Comput. Aided Mol. Des.* **2018**, *32*, 129–142.

[245] Tielker, N.; Eberlein, L.; Beckstein, O.; Güssregen, S.; Iorga, B. I.; Kast, S. M.; Liu, S. *Free Energy Methods in Drug Discovery: Current State and Future Directions*; 2021; Chapter 3, pp 67–107.

[246] Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14*, 5567–5582.

[247] Riniker, S.; Christ, C. D.; Hansen, N.; Mark, A. E.; Nair, P. C.; van Gunsteren, W. F. Comparison of Enveloping Distribution Sampling and Thermodynamic Integration to Calculate Binding Free Energies of Phenylethanolamine N-Methyltransferase Inhibitors. *J. Chem. Phys.* **2011**, *135*, 24105.

[248] Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695–2703.

[249] Wang, L.; Deng, Y.; Wu, Y.; Kim, B.; LeBard, D. N.; Wandschneider, D.; Beachy, M.; Friesner, R. A.; Abel, R. Accurate Modeling of Scaffold Hopping Transformations in Drug Discovery. *J. Chem. Theory Comput.* **2017**, *13*, 42–54.

[250] Aldeghi, M.; Heifetz, A.; Bodkin, M. J.; Knapp, S.; Biggin, P. C. Accurate Calculation of the Absolute Free Energy of Binding for Drug Molecules. *Chem. Sci.* **2016**, *7*, 207–218.

[251] Yu, H. S.; Deng, Y.; Wu, Y.; Sindhikara, D.; Rask, A. R.; Kimura, T.; Abel, R.; Wang, L. Accurate and Reliable Prediction of the Binding Affinities of Macrocycles to Their Protein Targets. *J. Chem. Theory Comput.* **2017**, *13*, 6290–6300.

[252] Wei, J.; Chipot, C.; Roux, B. Computing Relative Binding Affinity of Ligands to Receptor: An Effective Hybrid Single-Dual-Topology Free-Energy Perturbation Approach in NAMD. *J. Chem. Inf. Model.* **2019**, *59*, 3794–3802.

[253] Paulsen, J. L.; Yu, H. S.; Sindhikara, D.; Wang, L.; Appleby, T.; Villasenor, A. G.; Schmitz, U.; Shivakumar, D. Evaluation of Free Energy Calculations for the Prioritization of Macrocycle Synthesis. *J. Chem. Inf. Model.* **2020**, *60*, 3489–3498.

[254] Ries, B.; Rieder, S. R.; Rhiner, C.; H., H. P.; Riniker, S. RestraintMaker: A Graph-Based Approach to Determine Distance Restraints for Free-Energy Calculations With Dual Topology. *J. Comput. Aided Mol. Des.* **2022**, *36*, 175–192.

[255] Boresch, S.; Karplus, M. The Role of Bonded Terms in Free Energy Simulations. 2. Calculation of Their Influence on Free Energy Differences of Solvation. *J. Phys. Chem. A* **1999**, *103*, 119–136.

[256] Rocklin, G. J.; Mobley, D. L.; Dill, K. A. Separated Topologies – A Method for Relative Binding Free Energy Calculations Using Orientational Restraints. *J. Chem. Phys.* **2013**, *138*, 085104.

[257] Fleck, M.; Wieder, M.; Boresch, S. Dummy Atoms in Alchemical Free Energy Calculations. *J. Chem. Theory and Comp.* **2021**, *17*, 4403–4419.

[258] Pearlman, D. A.; Kollman, P. A. The Overlooked Bond-Stretching Contribution in Free Energy Perturbation Calculations. *J. Chem. Phys.* **1991**, *94*, 4532–4545.

[259] Pearlman, D. A. A Comparison of Alternative Approaches to Free Energy Calculations. *J. Phys. Chem.* **1994**, *98*, 1487–1493.

[260] Shobana, S.; Roux, B.; Andersen, O. S. Free Energy Simulations: Thermodynamic Reversibility and Variability. *J. Phys. Chem. B* **2000**, *104*, 5179–5190.

[261] Gao, J.; Kuczera, K.; Tidor, B.; Karplus, M. Hidden Thermodynamics of Mutant Proteins: A Molecular Dynamics Analysis. *Science* **1989**, *244*, 1069–1072.

[262] Hedges, L. O.; Mey, A. S.; Laughton, C. A.; Gervasio, F. L.; Mulholland, A. J.; Woods, C. J.; Michel, J. BioSimSpace: An Interoperable Python Framework for Biomolecular Simulation. *J. Open Source Softw.* **2019**, *4*, 1831.

[263] Loeffler, H. H.; Michel, J.; Woods, C. FESetup: Automating Setup for Alchemical Free Energy Simulations. *J. Chem. Inf. Model.* **2015**, *55*, 2485–2490.

[264] Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. Lead Optimization Mapper: Automating Free Energy Calculations for Lead Optimization. *J. Comput. Aided Mol. Des.* **2013**, *27*, 755–770.

[265] Suruzhon, M.; Senapathi, T.; Bodnarchuk, M. S.; Viner, R.; Wall, I. D.; Barnett, C. B.; Naidoo, K. J.; Essex, J. W. ProtoCaller: Robust Automation of Binding Free Energy Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 1917–1921.

[266] Petrov, D. Perturbation Free-Energy Toolkit: An Automated Alchemical Topology Builder. *J. Chem. Inf. Model.* **2021**, *61*, 4382–4390.

[267] Zavitsanou, S.; Tsengenes, A.; Papadourakis, M.; Amendola, G.; Chatzigoulas, A.; Dellis, D.; Cosconati, S.; Cournia, Z. FEPrepare: A Web-Based Tool for Automating the Setup of Relative Binding Free Energy Calculations. *J. Chem. Inf. Model.* **2021**, *61*, 4131–4138.

[268] Homeyer, N.; Gohlke, H. FEW: A Workflow Tool for Free Energy Calculations of Ligand Binding. *J. Comput. Chem.* **2013**, *34*, 965–973.

[269] Carvalho Martins, L.; Cino, E. A.; Ferreira, R. S. PyAutoFEP: An Automated Free Energy Perturbation Workflow for GROMACS Integrating Enhanced Sampling Methods. *J. Chem. Theory and Comp.* **2021**, *17*, 4262–4273.

[270] Schrodinger, L.; DeLano, W. PyMOL. `http://www.pymol.org/pymol`.

[271] Hansmann, U. H. E. Parallel Tempering Algorithm for Conformational Studies of Biological Molecules. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

[272] Sugita, Y.; Kitao, A.; Okamoto, Y. Multidimensional Replica-Exchange Method for Free-Energy Calculations. *J. Chem. Phys.* **2000**, *113*, 6042–6051.

[273] Kruskal, J. B. On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proc. Amer. Math. Soc.* **1956**, *7*, 48–48.

[274] Wolfenden, R.; Liang, Y. L.; Matthews, M.; Williams, R. Cooperativity and Anticooperativity in Solvation by Water: Imidazoles, Quinones, Nitrophenols, Nitrophenolate, and Nitrothiophenolate Ions. *J. Am. Chem. Soc.* **1987**, *109*, 463–466.

[275] Rizzo, R. C.; Aynechi, T.; Case, D. A.; Kuntz, I. D. Estimation of Absolute Free Energies of Hydration Using Continuum Methods: Accuracy of Partial Charge Models and Optimization of Nonpolar Contributions. *J. Chem. Theory Comput.* **2006**, *2*, 128–139.

[276] Nicholls, A.; Mobley, D. L.; Guthrie, J. P.; Chodera, J. D.; Bayly, C. I.; Cooper, M. D.; Pande, V. S. Predicting Small-Molecule Solvation Free Energies: An Informal Blind Test for Computational Chemistry. *J. Med. Chem.* **2008**, *51*, 769–779.

[277] Guthrie, J. P. SAMPL4, a Blind Challenge for Computational Solvation Free Energies: The Compounds Considered. *J. Comput. Aided* **2014**, *28*, 151–168.

[278] Ries, B.; Rieder, S. R.; Champion, C.; Barros, E. P.; Riniker, S. rinikerlab/reeds: An Automatized RE-EDS Sampling Procedure (v1.0). `https://github.com/rinikerlab/reeds`, 2021.

[279] Lehner, M. T.; Ries, B.; Rieder, S. R.; Riniker, S. rinikerlab/PyGromosTools: PyGromosTools_V2 (v2.0). 2021; `https://doi.org/10.5281/zenodo.4621710`, Accessed: 2022/05/05.

[280] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Hermans, J. J. In *Intermolecular Forces*; Pullman, B., Ed.; Reidel: Dordrecht, The Netherlands, 1981; pp 331–342.

[281] Heinz, T. N.; van Gunsteren, W. F.; Hünenberger, P. H. Comparison of Four Methods to Compute the Dielectric Permittivity of Liquids From Molecular Dynamics Simulations. *J. Chem. Phys.* **2001**, *115*, 1125–1136.

[282] Eichenberger, A. P.; Allison, J. R.; Dolenc, J.; Geerke, D. P.; Horta, B. A. C.; Meier, K.; Oostenbrink, C.; Schmid, N.; Steiner, D.; Wang, D.; van Gunsteren, W. F. The GROMOS++ Software for the Analysis of Biomolecular Simulation Trajectories. *J. Chem. Theory Comput.* **2011**, *7*, 3379–3390.

[283] McKinney, W. Data Structures for Statistical Computing in Python. *Proc. of the 9th Python in Science Conf.* **2010**, *445*, 51–56.

[284] Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 99–104.

[285] van Der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30.

[286] Virtanen, P.; Gommers, R.; Oliphant, T. E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; van der Walt, S. J.; Brett, M.; Wilson, J.; Millman, K. J.; Mayorov, N.; Nelson, A. R.; Jones, E.; Kern, R.; Larson, E.; Carey, C. J.; Polat, I.; Feng, Y.; Moore, E. W.; VanderPlas, J.; Laxalde, D.; Perktold, J.; Cimrman, R.; Henriksen, I.; Quintero, E. A.; Harris, C. R.; Archibald, A. M.; Ribeiro, A. H.; Pedregosa, F.; van Mulbregt, P.; Vijaykumar, A.; Bardelli, A. P.; Rothberg, A.; Hilboll, A.; Kloeckner, A.; Scopatz, A.; Lee, A.; Rokem, A.; Woods, C. N.; Fulton, C.; Masson, C.; Häggström, C.; Fitzgerald, C.; Nicholson, D. A.; Hagen, D. R.; Pasechnik, D. V.; Olivetti, E.; Martin, E.; Wieser, E.; Silva, F.; Lenders, F.; Wilhelm, F.; Young, G.; Price, G. A.; Ingold, G. L.; Allen, G. E.; Lee, G. R.; Audren, H.; Probst, I.; Dietrich, J. P.; Silterra, J.; Webber, J. T.; Slavič, J.; Nothman, J.; Buchner, J.; Kulick, J.; Schönberger, J. L.; de Miranda Cardoso, J. V.; Reimer, J.; Harrington, J.; Rodríguez, J. L. C.; Nunez-Iglesias, J.; Kuczynski, J.; Tritz, K.; Thoma, M.; Newville, M.; Kümmerer, M.; Bolingbroke, M.; Tartre, M.; Pak, M.; Smith, N. J.; Nowaczyk, N.; Shebanov, N.; Pavlyk, O.; Brodtkorb, P. A.; Lee, P.; McGibbon, R. T.; Feldbauer, R.; Lewis, S.; Tygier, S.; Sievert, S.; Vigna, S.; Peterson, S.; More, S.; Pudlik, T.; Oshima, T.; Pingel, T. J.; Robitaille, T. P.; Spura, T.; Jones, T. R.; Cera, T.; Leslie, T.; Zito, T.; Krauss, T.; Upadhyay, U.; Halchenko, Y. O.; Vázquez-Baeza, Y. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* **2020**, *17*, 261–272.

[287] Johansson, F. mpmath: A Python Library for Arbitrary-Precision Floating-Point Arithmetic (Version 0.18). 2013; http://mpmath.org/.

[288] Kluyver, T.; Ragan-kelley, B.; Pérez, F.; Granger, B.; Bussonnier, M.; Frederic, J.; Kelley, K.; Hamrick, J.; Grout, J.; Corlay, S.; Ivanov, P.; Avila, D.; Abdalla, S.;

Willing, C.; Development Team, J. Jupyter Notebooks – A Publishing Format for Reproducible Computational Workflows. *ELPUB* **2016**, 87–90.

[289] Spearman, C. The Proof and Measurement of Association Between Two Things. *Am. J. Psychol.* **1904**, *15*, 72–101.

[290] Feynman, R. P. What Is Science? *Phys. Teach.* **1969**, *7*, 313–320.

[291] Hünenberger, P. H.; van Gunsteren, W. F. *Potential Energy Surfaces*; Springer, 1999; pp 177–214.

[292] Case, D. A.; Walker, R. C.; Cheatham, C., Thomas E. Simmerling; Roitberg, A.; Merz, K. M.; Luo, R.; Darden, T.; Wang, J.; Duke, R. E.; Roe, D. R.; LeGrand, S.; Swails, J.; Cerutti, D.; Monard, G.; Sagui, C.; Kaus, J.; Betz, R.; Madej, B.; Lin, C.; Mermelstein, D.; Li, P.; Onufriev, A.; Izadi, S.; Wolf, R. M.; Wu, X.; Götz, A. W.; Gohlke, H.; Homeyer, N.; Botello-Smith, W. M.; Xiao, L.; Luchko, T.; Giese, T.; Lee, T.; Nguyen, H. T.; Nguyen, H.; Janowski, P.; Omelyan, I.; Kovelnko, A.; Kollman, P. A. AMBER Reference Manual; University of California: San Francisco. 2016.

[293] Swails, J.; Hernandez, C.; Mobley, D. L.; Nguyen, H.; Wang, L.-P.; Janowski, P. ParmEd. `https://github.com/ParmEd/ParmEd`, 2010.

[294] da Silva, A. W. S.; Vranken, W. F. ACPYPE-Antechamber Python Parser Interface. *BMC Res. Notes* **2012**, *5*, 1–8.

[295] Lee, T. gromos2amber. `https://github.com/ATB-UQ/gromos2amber`, Accessed: 2022/06/15.

[296] Vermaas, J. V.; Hardy, D. J.; Stone, J. E.; Tajkhorshid, E.; Kohlmeyer, A. TopoGromacs: Automated Topology Conversion From CHARMM to GROMACS Within VMD. *J. Chem. Inf. Model.* **2016**, *56*, 1112–1116.

[297] Swails, J. M. Free Energy Simulations of Complex Biological Systems at Constant pH. Ph.D. thesis, University of Florida, 2013.

[298] Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935.

[299] The GROMOS Software for (Bio)Molecular Simulation, Volume 6. 2021.

[300] Jorgensen, W. L.; Buckner, J. K.; Boudon, S.; Tirado-Rives, J. Efficient Computation of Absolute Free Energies of Binding by Computer Simulations. Application to the Methane Dimer in Water. *J. Chem. Phys.* **1988**, *89*, 3742–3746.

[301] Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *J. Chem. Theory Comput.* **2014**, *10*, 2632–2647.

[302] prmtop.pdb. `https://ambermd.org/prmtop.pdf`, Accessed: 2021/11/30.

[303] Bergdorf, M.; Peter, C.; Hünenberger, P. H. Influence of Cut-Off Truncation and Artificial Periodicity of Electrostatic Interactions in Molecular Simulations of Solvated Ions: A Continuum Electrostatics Study. *J. Chem. Phys.* **2003**, *119*, 9129–9144.

[304] Kubincová, A.; Riniker, S.; Hünenberger, P. H. Reaction-Field Electrostatics in Molecular Dynamics Simulations: Development of a Conservative Scheme Compatible With an Atomic Cutoff. *Phys. Chem. Chem. Phys.* **2020**, *22*, 26419–26437.

[305] Berendsen, H. J. C.; van der Spoel, D.; van Drunen, R. GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

[306] Jakalian, A.; Bush, B. L.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC model: I. Method. *J. Comput. Chem.* **2000**, *21*, 132–146.

[307] Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623–1641.

[308] van Gunsteren, W. F. gromos.net. `http://www.gromos.net/`, 2021; Accessed: 2022/01/06.

[309] Glättli, A.; Daura, X.; van Gunsteren, W. F. Derivation of an Improved Simple Point Charge Model for Liquid Water: SPC/A and SPC/L. *J. Chem. Phys.* **2002**, *116*, 9811–9828.

[310] Riniker, S.; Kunz, A.-P. E.; van Gunsteren, W. F. On the Calculation of the Dielectric Permittivity and Relaxation of Molecular Models in the Liquid Phase. *J. Chem. Theory Comput.* **2011**, *7*, 1469–1475.

[311] Waskom, M. L. seaborn: Statistical Data Visualization. *J. Open Source Softw.* **2021**, *6*, 3021.

[312] Kunz, A.-P. E.; Allison, J. R.; Geerke, D. P.; Horta, B. A. C.; Hünenberger, P. H.; Riniker, S.; Schmid, N.; van Gunsteren, W. F. New Functionalities in the GROMOS Biomolecular Simulation Software. *J. Comput. Chem.* **2012**, *33*, 340–353.

[313] Hein, P. *Grooks II*; Doubleday & Company, Garden City, New York, USA, 1969.

[314] Lee, J.; Miller, B. T.; Damjanovic, A.; Brooks, B. R. Constant pH Molecular Dynamics in Explicit Solvent With Enveloping Distribution Sampling and Hamiltonian Exchange. *J. Chem. Theory Comput.* **2014**, *10*, 2738–2750.

[315] Chandler, D.; Weeks, J. D.; Andersen, H. C. van der Waals Picture of Liquids, Solids, and Phase Transformations. *Science* **1983**, *220*, 787–794.

[316] Cooley, J. W.; Tukey, J. W. An Algorithm for the Machine Calculation of Complex Fourier Series. *Math. Comput.* **1965**, *19*, 297–301.

[317] Hünenberger, P. H. In *Simulation and Theory of Electrostatic Interactions in Solution: Computational Chemistry, Biophysics, and Aqueous Solution.*; Hummer, G., Pratt, L. R., Eds.; American Institute of Physics, New York, USA, 1999; Vol. 492.

[318] van Gunsteren, W. F.; Berendsen, H. J. C.; Colonna, F.; Perahia, D.; Hollenberg, J. P.; Lellouch, D. On Searching Neighbors in Computer Simulations of Macromolecular Systems. *J. Comput. Chem.* **1984**, *5*, 272–279.

[319] Lindahl, E.; Abraham, M. J.; Hess, B.; van der Spoel, D. GROMACS 2019.6 Manual. 2020; `https://doi.org/10.5281/zenodo.3685925`, Accessed: 2020/05/19.

[320] Lide, D. R. *Handbook of Chemistry and Physics*, 88th ed.; CRC Press/Taylor and Francis: Boca Raton, FL, 2007-2008.

[321] Åqvist, J.; Wennerström, P.; Nervall, M.; Bjelic, S.; Brandsdal, B. O. Molecular Dynamics Simulations of Water and Biomolecules With a Monte Carlo Constant Pressure Algorithm. *Chem. Phys. Lett.* **2004**, *384*, 288–294.

[322] Shirts, M. R. Reweighting From the Mixture Distribution as a Better Way to Describe the Multistate Bennett Acceptance Ratio. *arXiv e-prints* **2017**, arXiv–1704.

[323] Scots Connection - Ferguson Clan. `https://www.scotsconnection.com/clan_crests/ferguson.htm`, Accessed: 2022/06/23.

[324] Rieder, S. R.; Ries, B.; Champion, C.; Barros, E. P.; Hünenberger, P. H.; Riniker, S. Replica-Exchange Enveloping Distribution Sampling: Calculation of Relative Free Energies in GROMOS. *CHIMIA* **2022**, *76*, 327–330.

[325] MarvinSketch version 20.11, ChemAxon. https://chemaxon.com, Accessed: 2022/06/30.

Curriculum Vitæ

Personal Data

Name	Salomé Ronja Rieder-Walthard
Date of Birth	07.04.1993
Place of Origin	Berne, BE, Switzerland
Citizen of	Switzerland

Education

2018 – 2022	Doctor of Sciences ETH Zurich (D-CHAB), Switzerland
2016 – 2018	MSc. ETH in Computational Science and Engineering ETH Zurich (D-MATH), Switzerland
2011 – 2016	Bsc. ETH in Computational Science and Engineering ETH Zurich (D-MATH), Switzerland
2005 – 2011	Bilingual Matura Kantonsschule Alpenquai, Lucerne, Switzerland

Experience

2020 –	Web Design Pneumatikhaus AG Luzern, Lucerne, Switzerland
2018 – 2022	Scientific Assistant ETH Zurich (D-CHAB), Switzerland
2017 – 2020	Member of the Board JF Stadt Luzern, Lucerne, Switzerland
2014 – 2017	Maths Tutor Various Cantonal School Students, Zurich, Switzerland
2010 – 2011	Cashier Cafeteria Kantonsschule Alpenquai, Lucerne, Switzerland

Honors

2021	Best Poster Presentation Award in Computational Chemistry Swiss Chemical Society (SCS) Fall Meeting 2021
2019	ETH Medal for Outstanding Master's Theses ETH Zurich (D-MATH), Switzerland