DISS. ETH NO. 27332

SYSTEMATIC OPTIMIZATION OF EMPIRICAL FORCE FIELDS AGAINST EXPERIMENTAL DATA

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
MARINA PEREIRA OLIVEIRA

M.Sc. University of São Paulo in Chemistry

born on 16.06.1992

citizen of
Brazil

accepted on the recommendation of

Prof. Dr. Philippe Hünenberger, examiner
Prof. Dr. Sereina Riniker, co-examiner
Prof. Dr. Bruno A. C. Horta, co-examiner

2021

Summary

The main topic of this thesis is the design, development, and application of an integrated scheme, called CombiFF, to refine force-field parameters against condensed-phase data for small organic molecules. In this context, Chapter 1 gives a brief introduction to the area of force-field development.

Chapters 2-4 present the implementation and application of the CombiFF scheme to different families of compounds. The main steps of the CombiFF scheme are: (i) definition of a molecule family, (ii) combinatorial enumeration of all isomers, (ii) 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. A key component of the force-field design choice is that the atomic partial charges are not specified explicitly within topology building blocks for the various functional groups, but determined implicitly using an electronegativity-equalization scheme. Accordingly, atomic hardness and electronegativity values are specified instead of atomic partial charges. These parameters are expected to be less dependent on the covalent environment of an atom and to produce to charges that take induction effects into account. One significant advantage of the CombiFF scheme is that once the time-consuming task of target-data selection/curation has been performed, the optimization of a force field only takes a few days. As a result, this scheme enables a straightforward assessment of the consequences of functional-form and simulation-parameter decisions on the intrinsic accuracy of the classical force-field representation at an optimal level of parametrization.

Chapter 2 describes the workflow of the CombiFF scheme, along with an initial application of the approach to the saturated acyclic haloalkane family. The optimization of the non-bonded parameters is performed against liquid density ρ_{liq} and vaporization enthalpy ΔH_{vap}, and requires less than three days of wall-clock computing time given access to a few hundred processors. A remarkable level of agreement with experiment in terms of the two properties considered can be achieved using a simple united-atom representation without any specific description of sigma-holes and halogen-bonding. Of course, an improved representation of the anisotropy in the halogen electron density might become important when considering other systems and properties (*i.e.* directional interactions in protein-ligand binding). The trends in the optimized parameters along with the halogen series and across the compound family are in line with chemical intuition based on considerations related to size, polarizability, softness, electronegativity, induction, and hyperconjugation. This observation is particularly remarkable considering that the force-field calibration does not include the result of any quantum mechanical calculation.

In Chapter 3, the same scheme is applied to the construction of a force field for saturated acyclic compounds encompassing eight common chemical functional groups involving oxygen and/or nitrogen atoms. Again, ρ_{liq} and ΔH_{vap} are used to calibrate and validate the non-bonded interaction parameters of the force field. An excellent level of agreement with experiment is achieved, although the errors are not homogeneously distributed across the chemical functional groups.

In Chapter 4, a comparison between united-atom (UA) and all-atom (AA) representations is performed, in the context of the saturated acyclic (halo)alkanes. To this purpose, two force field variants are optimized using the CombiFF approach, one at the UA and one at AA level. In both cases, the calibration is performed against ρ_{liq} and ΔH_{vap} values for (halo)alkane compounds, and validated against the same properties. Further comparison between the UA and AA resolutions also involves nine other thermodynamic, dielectric, transport, and solvation properties, namely the surface-tension coefficient γ, the isothermal compressibility κ_T,

the isobaric thermal-expansion coefficient α_P, the isobaric heat capacity c_P, the static relative dielectric permittivity ϵ, the self-diffusion coefficient D, the shear viscosity η, the hydration free energy ΔG_{wat}, and the free energy of solvation ΔG_{che} in cyclohexane. In terms of the target properties ρ_{liq} and ΔH_{vap}, the UA and AA resolutions reach very similar levels of accuracy after optimization. Concerning the nine other properties, the AA representation leads to more accurate results in terms of D and η, comparably accurate results in terms of γ, κ_T, α_P, ϵ, and ΔG_{che}, and less accurate results in terms of c_P and ΔG_{wat}. This work also represents the first steps towards the calibration of a GROMOS-compatible force field at the AA resolution.

In Chapter 5, the force-field parameters developed in Chapters 2 and 3 for haloalkanes and for compounds involving common oxygen and nitrogen functional groups are further tested against the same nine thermodynamic, dielectric, transport, and solvation properties. In addition, the transferability of the parameters is also tested for hetero-polyfunctional molecules in terms of ρ_{liq} and ΔH_{vap}. In terms of overall agreement between simulation and experiment for the nine additional properties, excellent agreement is generally observed in terms of γ, good results in terms of κ_T, α_P, ΔG_{wat} and ΔG_{che}, and more pronounced and systematic deviations in terms of c_P, ϵ, D and η. The results for hetero-polyfunctional molecules suggest that the force-field parameters are not fully transferable to molecules combining different types of functional groups simultaneously. A good agreement with experiment is achieved for ρ_{liq}, but significant deviations are observed for ΔH_{vap}.

Finally, Chapter 6 provides general conclusions as well as an outlook into possible future directions.

Zusammenfassung

Das Hauptthema dieser Arbeit ist der Entwurf, die Entwicklung und die Anwendung eines integrierenden Schemas zur Optimierung von Kraftfeldparametern anhand von Daten für kleine organische Moleküle in kondensierter Phase, genannt CombiFF. In diesem Zusammenhang gibt Kapitel 1 eine kurze Einführung in den Bereich der Kraftfeldentwicklung. In den Kapiteln 2-4 wird die Umsetzung und Anwendung des CombiFF-Schemas auf verschiedene Familien von chemischen Verbindungen vorgestellt.

Die grundlegenden Schritte des CombiFF-Schemas sind: (i) Definition einer Molekülfamilie, (ii) kombinatorische Aufzählung aller Isomere, (ii) Suche nach experimentellen Daten, (iv) automatische Konstruktion der molekularen Topologien durch Zusammensetzung von Fragmenten, und (v) iterative Verfeinerung der Kraftfeldparameter unter Berücksichtigung der gesamten Familie. Eine Schlüsselkomponente bei der Entscheidung für das Kraftfelddesigns ist, dass die atomaren Partialladungen nicht explizit innerhalb der Topologiebausteine für die verschiedenen Funktionsgruppen angegeben werden, sondern implizit unter Verwendung eines Elektronegativitäts-Ausgleichsschemas festgelegt sind. Dementsprechend werden Atomhärte- und Elektronegativitätswerte anstelle von atomaren Teilladungen angegeben. Diese Parameter sollten weniger abhängig von der kovalenten Umgebung eines Atoms sein und Ladungen produzieren, die Induktionseffekte berücksichtigen. Ein wesentlicher Vorteil der CombiFF-Regelung ist, dass sobald die zeitintensive Aufgabe der Zieldaten-Auswahl/Kuration durchgeführt wurde, die Optimierung eines Kraftfeldes nur wenige Tage dauert. Daher ermöglicht dieses Schema eine einfache Abschätzung der Folgen von Funktionsform- und Simulationsparameter-Entscheidungen über die intrinsische Genauigkeit der klassischen Kraftfelddarstellung auf einem optimalen Parametrisierungsniveau.

Kapitel 2 beschreibt den Arbeitsablauf des CombiFF-Schemas, zusammen mit einer ersten Anwendung des Ansatzes auf die gesättigte azyklische Haloalkan-Familie. Die Optimierung der nicht gebundenen Parameter wird gegen die Flüssigkeitsdichte ρ_{liq} und die Verdampfungsenthalpie ΔH_{vap} durchgeführt, und erfordert weniger als drei Tage Rechenzeit unter Verwendung von hundert Prozessoren. Ein bemerkenswertes Mass an Übereinstimmung mit dem Experiment in Bezug auf die beiden betrachteten Eigenschaften kann unter Verwendung der einfachen Darstellung des vereinigten Atome Prinzips ohne spezifische Beschreibung von Sigma-Löchern und Halogenbindungen erreicht werden. Natürlich könnte eine verbesserte Darstellung der Anisotropie in der Halogen-Elektronendichte wichtig werden, wenn andere Systeme und Eigenschaften (z.B. gerichtete Wechselwirkungen bei der Protein-Ligand-Bindung) betrachtet werden. Die Trends bei den optimierten Parametern zusammen mit der Halogenreihe und über die gesamte Verbindungsfamilie stehen im Einklang mit der chemischen Intuition, die auf Überlegungen zur Grösse, Polarisierbarkeit, Weichheit, Elektronegativität, Induktion und Hyperkonjugation beruht. Diese Beobachtung ist besonders bemerkenswert wenn man bedenkt, dass die Kraftfeldkalibrierung nicht das Ergebnis einer quantenmechanischen Berechnung enthält.

In Kapitel 3 wird dasselbe Schema auf die Konstruktion eines Kraftfeldes für gesättigte azyklische Verbindungen angewendet, die acht gemeinsame chemische funktionelle Gruppen unter Beteiligung von Sauerstoff- und/oder Stickstoffatomen umfassen. Auch hier werden ρ_{liq} und ΔH_{vap}, die nicht gebundenen Wechselwirkungsparameter des Kraftfeldes, zur Kalibrierung und Validierung verwendet. Es wird eine sehr gute Übereinstimmung mit dem Experiment erreicht, allerdings sind die Fehler nicht homogen über die chemischen funktionellen Gruppen verteilt.

In Kapitel 4 wird ein Vergleich zwischen Molekül Darstellungen mit vereinigten Atomen (UA) und allen Atomen (AA) im Zusammenhang mit den gesättigten azyklischen (Ha-

lo)alkanen durchgeführt. Zu diesem Zweck werden zwei Kraftfeldvarianten mit dem CombiFF-Ansatz optimiert, eine auf UA- und eine auf AA-Ebene. In beiden Fällen wird die Kalibrierung gegen ρ_{liq}- und ΔH_{vap}-Werte für (Halo)alkanverbindungen durchgeführt und gegen dieselben Eigenschaften validiert. Ein weiterer Vergleich zwischen den UA- und AA-Resolutionen bezieht zusätzlich neun weitere thermodynamische, dielektrische, Transport- und Solvatationseigenschaften mit ein, nämlich der Oberflächenspannungskoeffizient γ, die isotherme Kompressibilität κ_T, der isobare Wärmeausdehnungskoeffizient α_P, die isobare Wärmekapazität c_P, die statische relative dielektrische Permittivität ϵ, der Selbstdiffusionskoeffizient D, die Scherviskosität η, die hydratationsfreie Energie ΔG_{wat}, und die freie Solvatationsenergie ΔG_{che} in Cyclohexan. In Bezug auf die Zieleigenschaften ρ_{liq} und ΔH_{vap} erreichen die UA- und AA-Auflösungen nach der Optimierung sehr ähnliche Genauigkeitsniveaus. Hinsichtlich der neun anderen Eigenschaften führt die Darstellung der AA zu genaueren Ergebnissen in Bezug auf D und η, vergleichbar genauen Ergebnissen in Bezug auf γ, κ_T, α_P, ϵ und ΔG_{che} und weniger genauen Ergebnissen in Bezug auf c_P und ΔG_{wat}. Diese Arbeit stellt auch die ersten Schritte zur Kalibrierung eines GROMOS-kompatiblen Kraftfeldes mit der AA-Auflösung dar.

In Kapitel 5 werden die Kraftfeldparameter, die in den Kapiteln 2 und 3 für Haloalkane und für Verbindungen mit gemeinsamen funktionellen Sauerstoff- und Stickstoffgruppen entwickelt wurden, weiter gegen dieselben neun thermodynamischen, dielektrischen, Transport- und Solvatationseigenschaften getestet. Darüber hinaus wird die Übertragbarkeit der Parameter auch für hetero-polyfunktionelle Moleküle in Bezug auf ρ_{liq} und ΔH_{vap} getestet. Die Ergebnisse für die neun zusätzlichen Eigenschaften zeigen eine sehr gute Übereinstimmung in Bezug auf γ, vernünftige Ergebnisse in Bezug auf κ_T, α_P, ΔG_{wat} und ΔG_{che}, sowie ausgeprägtere und systematischere Abweichungen in Bezug auf c_P, ϵ, D und η. Die Ergebnisse für hetero-polyfunktionelle Moleküle deuten darauf hin, dass die Kraftfeldparameter nicht vollständig auf Moleküle übertragbar sind, die verschiedene Arten von funktionellen Gruppen gleichzeitig kombinieren. Eine gute Übereinstimmung mit dem Experiment wird für ρ_{liq} erreicht, aber signifikante Abweichungen werden für ΔH_{vap} beobachtet.

Schliesslich enthält Kapitel 6 allgemeine Schlussfolgerungen sowie einen Ausblick auf mögliche zukünftige Forschungsrichtungen.