Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations

Journal Article

Author(s):
Bleiziffer, Patrick; Schaller, Kay; Riniker, Sereina

Publication date:
2018-03-26

Permanent link:
https://doi.org/https://doi.org/10.3929/ethz-b-000255680

Originally published in:
Journal of Chemical Information and Modeling 58(3), https://doi.org/10.1021/acs.jcim.7b00663

Funding acknowledgement:
- Computer simulation of small organic molecules for computer-aided drug design ()
- A triple-resoluation QM/MM/coarse-grained model for the simulation of membrane-bound proteins ()

ACS Editors' Choice

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Cite This: J. Chem. Inf. Model. 2018, 58, 579−590

pubs.acs.org/jcim

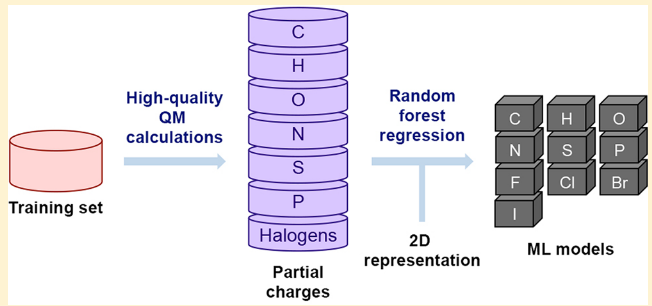
Machine Learning of Partial Charges Derived from High-Quality Quantum-Mechanical Calculations

Patrick Bleiziffer, Kay Schaller, and Sereina Riniker*

Laboratory of Physical Chemistry, ETH Zurich, Vladimir-Prelog-Weg 2, 8093 Zurich, Switzerland

Supporting Information

ABSTRACT: Parametrization of small organic molecules for classical molecular dynamics simulations is not trivial. The vastness of the chemical space makes approaches using building blocks challenging. The most common approach is therefore an individual parametrization of each compound by deriving partial charges from semiempirical or ab initio calculations and inheriting the bonded and van der Waals (Lennard-Jones) parameters from a (bio)molecular force field. The quality of the partial charges generated in this fashion depends on the level of the quantum-chemical calculation as well as on the extraction procedure used. Here, we present a machine learning (ML) based approach for predicting partial charges extracted from density functional theory (DFT) electron densities. The training set was chosen with the goal to provide a broad coverage of the known chemical space of druglike molecules. In addition to the speed of the approach, the partial charges predicted by ML are not dependent on the three-dimensional conformation in contrast to the ones obtained by fitting to the electrostatic potential (ESP). To assess the quality and compatibility with standard force fields, we performed benchmark calculations for the free energy of hydration and liquid properties such as density and heat of vaporization.



INTRODUCTION

A classical fixed-charge force field involves hundreds of parameters, which need to be chosen such as to represent the true Hamiltonian of the system. Parametrization is typically based on fitting to properties calculated using quantum-mechanical (QM) approaches and/or experimentally accessible properties. For the majority of biomolecules such as proteins, DNA, lipids, and carbohydrates, only a relatively small number of building blocks is required, which makes parametrization and validation straightforward. Standard biomolecular force fields such as AMBER,[1] OPLS,[2] CHARMM,[3] and GROMOS[4] have been continuously improved over the past decades and tested and applied in a myriad of studies.

For small organic molecules, the situation is different. The size and diversity of just the drug-like chemical space is vast, making it difficult to construct building blocks and thus requiring an individual parametrization of each ligand. The most common approach is thereby to inherit the parameters of the bonded and van der Waals (Lennard-Jones) terms from a biomolecular force field using a matching strategy. The partial charges, on the other hand, are generated from a QM calculation (semiempirical or ab initio) of the new molecule. In order to reduce the computational effort, often QM methods with limited accuracy are used. Examples of such force fields for organic molecules are the general Amber force field (GAFF),[5] the OPLS all-atom force field (OPLS-AA),[6,7] the CHARMM general force field (CGenFF),[8,9] or the GROMOS automatic topology builder (ATB).[10,11] Other force fields such as the Merck molecular force field (MMFF)[12−14] and the universal force field (UFF)[15] are intended to be more general and do not require an individual parametrization of each ligand. However, such force fields employ more complex potential energy functions than most biomolecular force fields and are thus not compatible.

The extraction of partial charges from ab initio data is a long-standing problem, which has no unique solution. This can be explained by the fact that an innumerable number of charge distributions can reproduce the electrostatic potential (ESP) outside a surface, which encapsulates all the charges. This underdetermined mathematical problem is especially severe for so-called "buried atoms" in a molecule. The most straightforward approach is to reproduce the ESP at a large number of grid points around the molecule.[16] In order to remove part of the conformational dependency as well as to symmetrize the charges, a two-step scheme with charge restraints in the fitting process (RESP) was developed.[17] It is used in the Amber biomolecular force field[1] and often with GAFF.[5] In the GROMOS ATB,[10] initial partial charges are derived by fitting the ESP using the Singh and Kollman scheme.[16] The initial charges are subsequently optimized by averaging symmetrical groups and assigning charge groups, for which the overall charge should be an integer. This reduces the conformational dependency of the resulting charge distribution.[10]

Received: November 17, 2017
Published: February 20, 2018

Besides the mentioned schemes, in which the integrated number of electrons is partitioned, it is also possible to partition the electron density operator $\hat{\varrho}(\mathbf{r})$ at each spatial position. Such approaches fall into the realm of atoms-in-molecule (AIM) methods and have the appealing property of not being explicitly basis-set dependent like the Mulliken[18] population analysis. In AIM methods, the electron density operator

$$\hat{\varrho}(\mathbf{r}) = \sum_{i=1}^{N} \delta(\mathbf{r} - \mathbf{r}_i) \tag{1}$$

is divided into atomic contributions $\hat{\varrho}_A(\mathbf{r})$ such that the total electron density $\varrho(\mathbf{r})$ can be expressed as a linear combination of atomic densities $\varrho_A(\mathbf{r})$,

$$\varrho(\mathbf{r}) = \sum_{A}^{N_{\text{atoms}}} \varrho_A(\mathbf{r}) \tag{2}$$

With the condition that the electron density is strictly positive at all points in space, one can readily define the weight of the atomic density with respect to the total electron density:

$$\varrho_A(\mathbf{r}) = \frac{w_A(\mathbf{r})}{\sum_i w_i(\mathbf{r})} \varrho(\mathbf{r}) \tag{3}$$

AIM schemes like that of Hirshfeld,[19] iterative Hirshfeld,[20] and iterative stockholder atom (ISA)[21] differ in the definition of the weighting factors $w_A(\mathbf{r})$. Often Bader's quantum chemical topology (QCT)[22−24] is mentioned in connection with AIM methods for historical reasons, although it is strictly speaking not partitioning the electron density into atomic contributions.[25] The density-derived electrostatic and chemical (DDEC)[25−27] method is designed to simultaneously resemble reference ion states and provide an approximately spherical atomic electron distribution, thus combining the advantages of the Hirshfeld methods and ISA method. A more detailed overview of the development and state of research of charge fitting procedures can be found in ref 28. Once the atomic contributions to the electron density are obtained, the partial charge q_A is determined by subtracting the integrated atomic density from the effective nuclear charge Z_A

$$q_A = Z_A - \int \mathrm{d}\mathbf{r}\, \varrho_A(\mathbf{r}) \tag{4}$$

It should be mentioned that the optimization toward spherical partial charges is not appropriate in some cases. For atoms with lone pairs or a σ-hole, the electron density is usually anisotropic.[29] This problem can be addressed by distributing the atomic charge over off-site charges in close proximity.[30−32] The amount and position of the spread charge has to be fitted to reproduce the molecular dipole moment. Note that all AIM methods are implicitly basis-set dependent, since the underlying electronic structure method usually requires an orbital basis for the generation of the electron density (this effect is less pronounced when plane waves are used as orbital basis).

The issue of basis-set dependence of the Mulliken population analysis is addressed by the natural population analysis (NPA) method,[33] which on the other hand shows a somewhat larger conformational dependence than the DDEC method. This behavior is desired in 3D quantitative structure−activity/property relationship (QSAR/QSPR) models where the partial charges are used as descriptor.[34,35] Recently, Cole et al.[32] extended the AIM approach to Lennard-Jones (LJ) parameters, where they relied on the Tkatchenko-Scheffler scheme[36] for the computation of C_6 coefficients.

Performing a QM calculation for each compound individually to derive the partial charges has a number of drawbacks. First, as mentioned before the associated computational cost often results in the use of a QM method with limited accuracy, especially for high throughput. Second, ESP-based extraction methods have a relatively high conformational dependence, which is not desired in classical fixed-charge force fields, and third, the transferability of partial charges between the similar substructures in different molecules is not guaranteed. Reference 37 provides an excellent overview of successful applications of machine learning (ML) methods for the prediction of molecular properties. ML methods lend themselves to derive parameters for semiempirical[38,39] or force fields[40−43] from QM reference data. Recently, Li et al.[43] obtained parameters for the nonbonded interactions of methanol via a genetic algorithm based exclusively on QM data. Although the approach did not make use of any experimental data for training, the parameters yielded good condensed phase properties.

Only a few efforts have been made so far in the direction of learning and predicting force-field parameters. In 2013, Rai and Bakken have proposed a ML-based approach to predict partial charges using ESP-fit charges and random forest regression (RFR) models for the elements H, C, O, N, F, S, and Cl.[40] As descriptors the local environments in three dimensions were used. The training set consisted of 80 000 compounds randomly selected from the public database ZINC[44] and the Pfizer corporate library. Validation was conducted in terms of reproduction of the ESP and dipole moments. In addition, hydration free energies were calculated in combination with the OPLS-2005[45] force field. In 2015, a genetic algorithm (GA) optimization to obtain partial charges was proposed, which was, however, limited to rather small systems.[42] Another ML-based approach to predict multipole coefficients for a number of atom types, which are most common in organic molecules, was developed using kernel-ridge regression and a Coulomb matrix[46] with the inverse pairwise distances between atoms as descriptors.[41] The model was validated by computing intermolecular interaction energies in molecular dimers and organic crystals.

A computationally less demanding approach than the above-mentioned charge assignment schemes is offered by the electronegativity equalization method (EEM).[47] There, the key quantity for obtaining the partial charges is not directly the electron density itself but the equalization of the chemical potential for interacting atoms or molecules. Although it is also based on QM calculations, it is an empirical approach, where a model is fit to a set of reference electron densities. Compared to the schemes mentioned before, EEM is computationally less demanding, but on the other hand careful parametrization is required.[48]

In this work, we propose a ML-based approach to predict partial charges, building up on the work of Rai and Bakken.[40] The accuracy of QM calculations that go beyond a "HF/6-31G*" setup is combined with the conformationally robust charge-extraction scheme DDEC and descriptors that depend only on the 2D topological structure of the molecules. The latter removes the remaining conformational dependency of the partial charges and ensures a high transferability. By carefully selecting the molecules in the training set from the public databases ZINC[44] and ChEMBL,[49] a good coverage of the lead-like chemical space is obtained. The approach is evaluated by calculating hydration

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

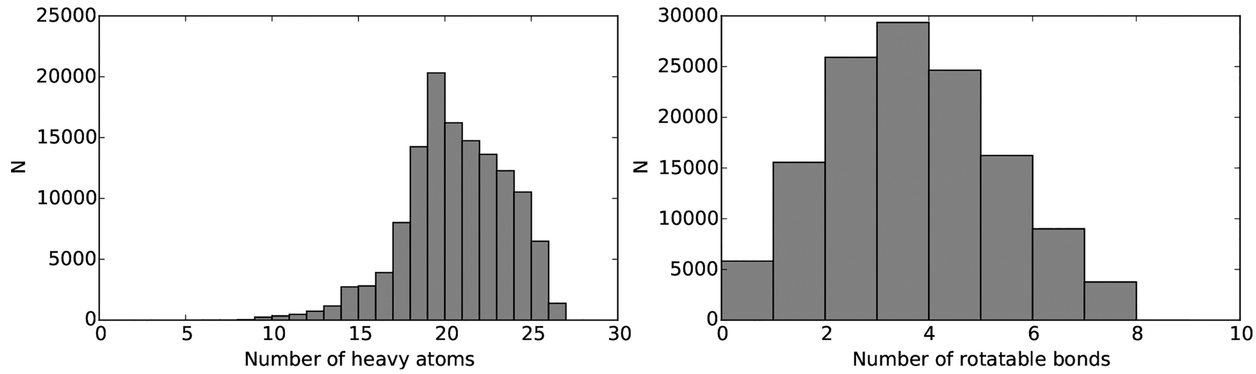


Figure 1. Distribution of the number of heavy atoms and the number of rotatable bonds in the combined training set, consisting of 130 267 lead-like molecules from the ZINC and ChEMBL databases.

free energies in combination with the GAFF force field, as well as densities and heat of vaporization in combination with the GAFF and OPLS-AA force field.

■ METHODS

Selection of the Training Set. We concentrated in the selection of compounds for the training set on lead-like molecules, i.e. molecules with a molecular weight of 250−350 g/mol, the number of rotatable bonds ≤7 and the octanol/water partition coefficient log $P \leq 3.5$ (as defined in the ZINC database[44]). Only molecules with the "organic" elements C, H, N, O, S, P, F, Cl, Br, and I were considered. Given this definition, the "clean leads" set from ZINC[44] version 2014 had roughly 4.6 million lead-like molecules (no molecules containing P) and ChEMBL,[49] version 21 consisted of 350 259 lead-like molecules. The local environment of an atom was described using an atom-centered atom pairs (AP) fingerprint[50] with a maximum path length of two. For this, the RDKit[51] function GetHashedAtomPairFingerprintAsBitVec() was used with the arguments maxLength = 2, nBits = 2048, and nBitsPerEntry = 4.

The protonation state at pH = 7 and the dominant tautomeric form were determined using the UNICON program.[52] This resulted in 117 767 unique fingerprints/atom environments in the ZINC set and 126 752 unique atom environments in ChEMBL. 64 785 atom environments found in ChEMBL were not present in ZINC (i.e., overlap of about 50%). To select a subset of the molecules such that all local environments were represented, we developed the following stepwise procedure for the ZINC molecules:

1. All molecules that contained atom environments, which occurred ≤4 times in the lead-like ZINC set and with no similar environment (Tanimoto similarity cutoff = 0.8) were chosen, i.e. 12 163 molecules.
2. For each of the remaining atom environments of the elements S, F, Cl, Br, and I, the four molecules with the highest number of not-yet-represented environments (of all elements) were picked, i.e., 2630 additional molecules.
3. For each of the remaining atom environments of the elements N and O, the four molecules with the highest number of not-yet-represented environments (of all elements) were picked, i.e. 19 815 additional molecules.
4. For each of the remaining atom environments of the element C, the four molecules with the highest number of not-yet-represented environments (of all elements) were picked, i.e., 65 497 additional molecules.

The final number of molecules selected from ZINC was 100 105. For the ChEMBL set, molecules were iteratively picked until all atom environments were represented, i.e. 40 325 molecules.

Conformer Generation. For 90 248 of the ZINC molecules and 40 019 of the ChEMBL molecules, conformers could be successfully generated using the RDKit distance geometry implementation. The final combined training set contained 130 267 molecules. The distribution of the number of heavy atoms and the number of rotatable bonds in the training set are shown in Figure 1.

Initial benchmarking on a small set of molecules showed that DDEC6 partial charges had a low conformational dependence (see Figures S3 and S4 in the Supporting Information), which is consistent with the results in refs 25 and 53. Therefore, only one conformer per molecule was used. In order to obtain a conformer that represents the molecule in water, the "Electrostatically Least-Interacting Functional Groups (ELF)" procedure[54] proposed by Bayly and McKay was applied. In short, 100 conformers were generated using the EmbedMultipleConfs() function in the RDKit with the options maxAttempts = 100 and pruneRMSthresh = 1.0, and subsequently optimized with the MMFF force field[12−14] using the MMFFOptimizeMoleculeConfs() function in the RDKit. For each conformer, all MMFF partial charges were set to their absolute value and the intramolecular Coulomb energy was calculated (excluding 1,2- and 1,3-neighbors). The conformer with the lowest energy, i.e. the one with the least intramolecular interactions, was selected for further processing.

The geometry (re)optimization of the selected conformers was performed with the semiempirical PM7 method[55] as implemented in the MOPAC package.[56] In addition to optimizations in vacuum, we also carried out calculations in implicit solvent, using the COSMO model[57] with dielectric constants of $\epsilon = 4$ (used to model the protein interior[58]) and $\epsilon = 78$ (water).

Electronic Structure Calculation. The quality of the extracted partial charges can only be as good as the underlying reference electron density. The best trade-off between accuracy and computational cost is provided by density functional theory (DFT).[59,60] To assess which density functional to use, electronic densities for a set of small organic molecules were benchmarked against coupled cluster singles doubles (CCSD) reference values for the following functionals: PBE0,[61] TPSS,[62] TPSSh,[63] B3LYP,[64−67] M06.[68] We also tested the popular choice of Hartree−Fock (HF) in a 6-31G* orbital basis. The results are shown in Figure S2 in the Supporting Information. The smallest deviation from the CCSD reference was found with the TPSSh

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

functional,[63] which was therefore employed for all subsequent electronic structure calculations in this work. The HF/6-31G* setup exhibited an error more than three times as high as that of TPSSh.

The DFT calculations were performed with the Turbomole package[69] (Version 7.0.2). As orbital basis, the def2-TZVP set[70,71] was used. The resolution of the identity was invoked in the computation of the two-center electron integrals.

To account for solvent effects in the DFT calculations, the implicit solvation model COSMO-RS[72,73] as implemented in Turbomole was employed wherever possible. For elements and atom types not available in COSMO-RS, the simpler COSMO was used instead. For a review of the COSMO and the COSMO-RS model, the reader is referred to ref 74. Three different dielectric permittivities of the continuum were considered: (i) $\epsilon = 1$ (vacuum), (ii) $\epsilon = 4$, and (iii) $\epsilon = 78$. For the calculation with $\epsilon = 1$ or 4, the molecules were neutralized when possible. The neutralization was performed by replacing the charged substructure with its neutral form using the ReplaceSubstructs() functionality in the RDKit.

Extraction of Partial Charges. Partial charges were extracted from the electron densities with the DDEC[25−27] method using the DDEC6 program.[25] This approach was chosen as it gives the least conformational dependence compared to ESP and RESP (Figure S4 in the Supporting Information). More details on the benchmarking are given in the Supporting Information.

Machine Learning. Atom-centered AP fingerprints[50] were used as descriptors to train the ML models. In the AP fingerprint, the atom type consists of the element, the number of π-electrons and the number of heavy-atom neighbors. Pairs of atoms together with their topological distance (path length) are hashed into bits. Note that this descriptor contains only 2D topological information, whereas the one in ref 40 is based on radial and angular symmetry functions, which describe the 3D environment. The descriptor was chosen to remove any remaining conformational dependence of the predicted partial charges as this is not desired in fixed-charge force fields. The RDKit function GetHashedAtomPairFingerprintAsBitVec() was used with the arguments nBits = 2048 and nBitsPerEntry = 4. The maximum path length (maxLength) was a parameter for optimization.

For each element (C, H, N, O, S, P, F, Cl, Br, and I), a separate random forest regression (RFR) model[75] was trained using the scikit-learn[76] library (version 0.18.1) and the following parameters: number of trees = 100, maximum depth = 6, minimum number of samples to split = 6, and minimum number of samples in leafs = 6. Random forests consist of an ensemble of decision trees, each tree trained with a random subset of features and samples.[75] The final prediction is obtained by taking a majority vote across all trees. This procedure reduces overfitting. In addition, random forests are fast and can handle a large amount of training data. They have also been used successfully in ref 40.

Each partial charge is predicted independently from the others in a molecule. Thus, the sum of the predicted charges q_i may differ from the formal charge or net charge Q^{net} of the molecule, as defined by the valency of the molecule. This excess charge

$$\Delta Q = \sum_{i=1}^{N_{\text{atoms}}} q_i - Q^{\text{net}} \tag{5}$$

can be viewed as a rough estimator of the accuracy of the prediction and has to be eliminated such that the predicted partial charges can be employed in MD simulations. This can be achieved by spreading the excess charge over the molecule. However, correctly predicted partial charges should not be perturbed. We employed the procedure described in ref 40, which uses the agreement among the trees of the RFR model as a measure. The standard deviation of the predicted charge for atom i by the trees T is defined as

$$\sigma_i = \sqrt{\frac{\sum_j^{N_{\text{trees}}} (q_i(T_j) - \bar{q}_i)^2}{N_{\text{trees}}}} \tag{6}$$

where N_{trees} is the number of trees in the RFR model, $q_i(T_j)$ is the partial charge predicted by tree j, and $\bar{q}_i$ is the overall charge predicted by the RFR model. σ_i can be utilized as weighting factor in the computation of the corrected partial charge

$$q_i^{\text{corr}} = q_i - \frac{\sigma_i |q_i| \Delta Q}{\sum_a^{N_{\text{atoms}}} \sigma_a |q_a|} \tag{7}$$

where the sum in the denominator runs over all atoms in the molecule (N_{atoms}).

We also investigated an alternative weighting scheme for eq 7 based on the occurrence of an atom environment (and/or environments with a high Tanimoto similarity) in the training set. For a fast calculation of the Tanimoto similarity, atom-centered AP fingerprints were calculated in the fps format using a modified version of the ChemFp package[77] for each molecule in the training set. A similarity threshold of 0.5 and 0.8 was tested. We found that large deviations between predicted and reference partial charge correlated with a low occurrence of the corresponding atom environment in the training set. However, the opposite is not true, i.e. the partial charges of many atom environments with low occurrence were correctly predicted. Therefore, this heuristic is not suited as weighting factor.

External Test Sets. As the training set contains only lead-like compounds, we constructed two external test sets to probe the transferability of the approach for real-world applications. Test set 1 contains 146 molecules, taken from ref 78, that are in the liquid phase at room temperature. These molecules have a molecular weight below 200 g/mol and less than seven rotatable bonds and are thus smaller than lead-like compounds.

Test set 2 contains standard drug-like molecules including some that are larger than the molecules in the training set. For this, the set of FDA approved drugs was taken from the ZINC database[79] version 2015, which consisted of 1385 molecules. Compounds already present in the training set and those with a molecular weight above 500 g/mol and/or with more than 14 rotatable bonds were removed. In addition, 11 molecules were too negatively charged, which led to convergence problems in the DFT calculations. The final set contained 1081 compounds. Of this, 380 had a molecular weight above 350 g/mol and 159 had more than seven rotatable bonds.

Molecular Dynamics Simulations. We tested the ML-predicted partial charges on two benchmark sets: (i) the set of 146 organic liquids with experimental thermodynamic properties reported in refs 78 and 80, and (ii) the set of 642 molecules with experimental hydration free energy data compiled in the FreeSolv[81−83] database.

The topologies, coordinates, and input files for the simulation of the set of organic liquids were obtained from http://virtualchemistry.org. The equilibrated boxes of the pure liquids

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

from ref 78 were taken as starting coordinates for the simulations. The topologies for GAFF (with AM1-BCC[84] partial charges) and OPLS-AA were used. The original partial charges were replaced with either DDEC reference partial charges or ML-predicted partial charges. Topologies were checked to have a vanishing net charge after replacement of the original partial charges with the ML-predicted partial charges. The DDEC reference partial charges were derived from electron densities obtained using the implicit solvation model with a dielectric permittivity ϵ = 1, 4, or 78. The ML-predicted partial charges were with ϵ = 4. The simulations were carried out under NPT conditions at 298 K and 1 atm with the GROMACS suite, version 4.6.7 using the input file from ref 78. The cutoff was 1.1 nm, which was also the switching distance for the particle mesh Ewald (PME) algorithm. For temperature coupling, the velocity rescaling with a stochastic term was used.[85] The time step was 2 fs.

The calculation of the hydration free energies were carried out using the same protocol for free energy perturbation (FEP)[86] as described in ref 83. Topologies, coordinates, and input files were taken from http://github.com/mobleylab/FreeSolv. The underlying force field is GAFF with AM1-BCC partial charges. The original partial charges were replaced with either DDEC reference partial charges or ML-predicted partial charges. Topologies were checked to have a vanishing net charge after replacement of the original partial charges with the ML-predicted partial charges. The DDEC reference partial charges were derived from electron densities obtained using the implicit solvation model with a dielectric permittivity ϵ = 1, 4, or 78. The ML-predicted partial charges were with ϵ = 4. The simulations were performed with the GROMACS suite, version 4.6.7 with 20 λ-points, where the first 5 λ-points correspond to changes in the electrostatic interactions, and the remaining points modify the LJ interactions. The simulation length was 5 ns per λ-point, under NPT conditions at 298 K and 1 atm. The cutoff for the electrostatic interactions was 1.2 and 1.0 nm for the van der Waals interactions. For temperature coupling, the Nosé–Hoover thermostat with a coupling time of 1 ps was used. The time step was 2 fs. Further details are given in ref 83.

■ RESULTS

Training Set. RFR models for the elements C, H, N, O, S, P, F, Cl, Br, and I were trained on the 130 267 lead-like molecules in the training set. The number of samples per element are shown in Figure 2.

The parameters of the RFR models were determined using k-fold cross validation with k = 10. Nine fractions were used for training and one fraction for testing, thus each atom was predicted once. From the predictions, the root-mean-square error (RMSE) was calculated for each fold and then averaged. Note that for the cross validation, the "raw" predicted partial charges were used, i.e., without a correction using eq 7. First, optimal parameters of the RFR models (e.g., number of trees, maximum depth, etc.) were determined by minimizing the RMSE (data not shown). Second, the maximum path length in the atom-centered AP fingerprint used as descriptor was optimized. In Figure 3, the RMSE as a function of the maximum path length in the descriptor is shown for the different elements. The baseline performance is given when only the reference atoms are considered. For this, a Morgan[87] fingerprint with radius zero was used, as AP fingerprints with a maximum path length of zero are not possible. The best performance was obtained for a maximum path length of four, and this setting was

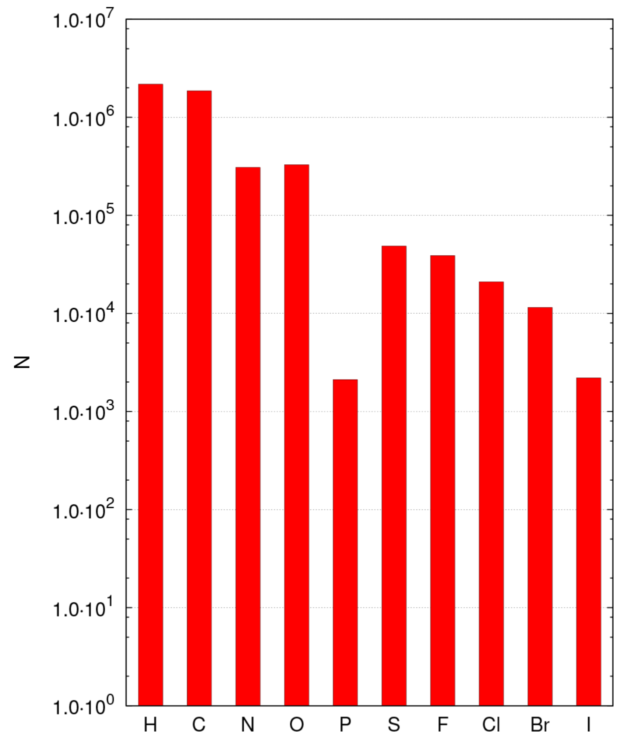


Figure 2. Number of samples (i.e., atom with environment and reference partial charge) per element in the combined training set of 130 267 molecules.

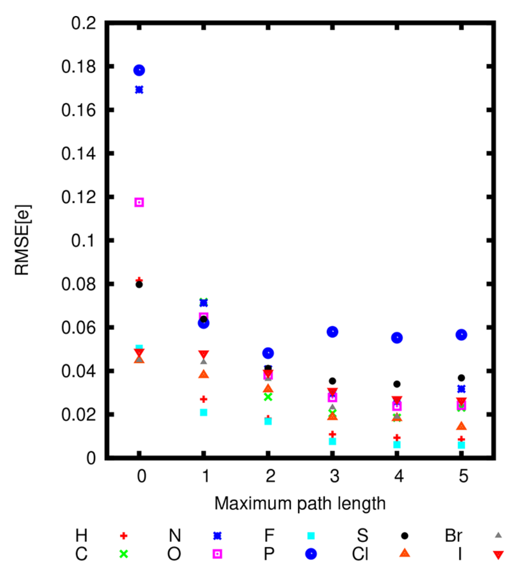


Figure 3. Average RMSE of the 10-fold cross validation of the training set as a function of the maximum path length in the atom-centered AP fingerprints used as descriptors. For each element, a separate RFR model was trained. Note that the predicted partial charges were not corrected using eq 7.

therefore used in the following. Increasing the path length further likely adds noise.

In the top panel in Figure 4, the predicted (not corrected) versus the reference partial charges from the 10-fold cross validation are shown for carbon as example. The same plots for the other elements are shown in the Figure S6 in the Supporting Information. The vast majority of the 3 100 446 predictions for carbon lies in an accuracy window of ±0.05 e (shown as black

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

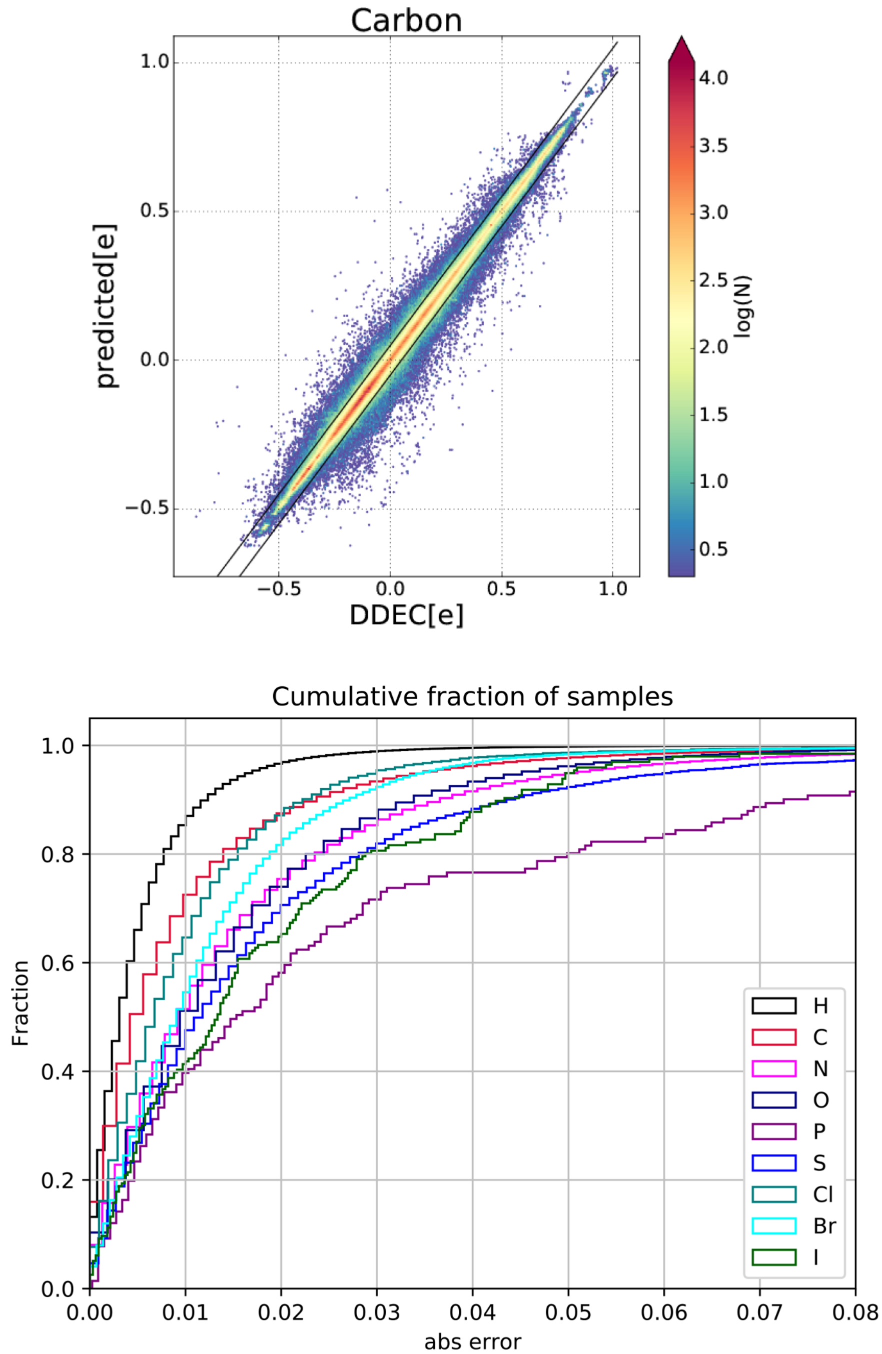


Figure 4. (top) Predicted versus reference partial charges for carbon in the 10-fold cross validation of the training set. The black lines indicate the accuracy window of ±0.05 e. (bottom) Cumulative fraction of samples (i.e., atom with environment and reference partial charge) as a function of the prediction error for all elements. A maximum path length of four was used for the atom-centered AP fingerprints. Note that the predicted partial charges were not corrected using eq 7.

solid lines). To quantify the prediction accuracy, the cumulative fraction of samples as a function of prediction error was calculated for all elements (bottom panel in Figure 4 and Table 1).

The highest prediction accuracy is obtained for H and F, and the lowest for P. The latter is likely due to the low number of samples (Figure 2) combined with a large range of possible partial charges (Figure S6 in the Supporting Information). For I on the other hand, there are also relatively few examples, but the range of partial charges is much smaller.

Performance on External Test Sets. The ML models were trained on lead-like compounds. The extrapolation potential to smaller and larger molecules was tested using two external test sets. Test set 1 contains 146 organic liquids with a molecular weight below 200 g/mol and less than seven rotatable bonds. The predicted versus reference partial charges are shown in the top panel of Figure 5. A low RMSE of 0.030 e was found, which does not change significantly when the excess charge is removed. The outliers stem mainly from very small molecules such as formaldehyde, 1,1-dichloroethene, dichlorofluoromethane, and 1,1,2,2-tetrachloroethane (Figure S7 in the Supporting Informa-

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

Table 1. Fraction of Samples (i.e., Atom with Environment and Reference Partial Charge) within an Accuracy Window of ±0.01, ± 0.05, and ±0.1 e in the 10-fold Cross Validation

	±0.01 e	±0.05 e	±0.1 e
H	0.85	1.00	1.00
C	0.73	0.98	1.00
N	0.51	0.95	0.99
O	0.51	0.97	1.00
F	0.92	1.00	1.00
P	0.38	0.79	0.94
S	0.44	0.92	0.98
Cl	0.65	0.99	1.00
Br	0.57	0.98	1.00
I	0.40	0.93	0.99

tion), which are not well represented in the training set. A straightforward way of reducing the numbers of outliers is to extend the training set with molecules containing substructures that are not yet well represented. For test set 2, the RMSE of 0.016 e is even lower than for test set 1, which indicates that it is easier for the ML models to extrapolate to larger molecules than to smaller ones. This is highly beneficial, as the computational cost of QM calculations increases with the size of the molecule and thus a fast ML-based approach is especially desirable for larger molecules.

A critical aspect for the application of a ML-based method is the computational efficiency. The memory requirements of the ML model depends on the number of samples and the range of partial charges. The scikit-learn RFR models are known to be rather large. The compressed files of the RFR models are up to

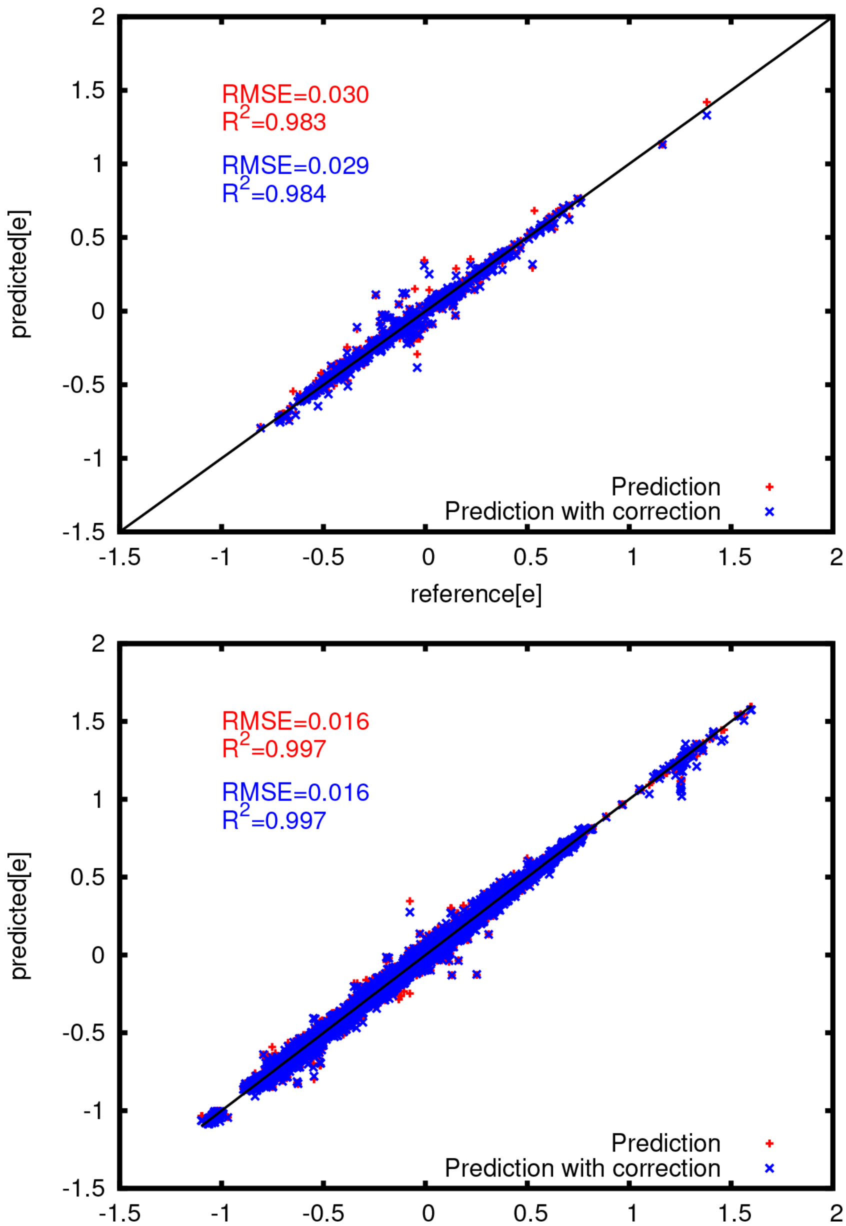


Figure 5. Predicted versus reference partial charges for the external test sets 1 (organic liquids, top) and 2 (FDA approved drugs, bottom). In contrast to the cross validation of the training set, the partial charges of the whole molecules are predicted here. The "raw" predicted charges are shown in red, and the ones corrected using eq 7 are in blue. The DDEC reference partial charges were derived from electron densities obtained using the implicit solvation model with $\epsilon = 4$. The ML models were trained also with $\epsilon = 4$.

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

660 MB for carbon. For predicting the partial charges of a whole molecule, all ML models need to be kept in memory, which can require up to 6 GB. This can, however, be provided by a normal workstation nowadays. Loading the ML models into memory is by far the slowest step in a prediction but needs to be done only once. In Table 2, the runtime and memory requirements of our

Table 2. Computational Requirements for the Charge Assignment of 1081 Drug-like Molecules from the External Test Set 2[a]

	ML	AM1-BCC (1SCF)	AM1-BCC
walltime [min]	3.0	11.3	1514
RAM [GB]	6.04	0.093	0.180

[a]The timings were estimated by running the corresponding method five times on a single core and taking the average of the runtime.

ML-based approach are compared to that of the antechamber program (version 17.3)[88] with AM1-BCC charges (semi-empirical method). Partial charges were derived for the external test set 2. The procedure was repeated five times and the runtime averaged. The timings were carried out on a XeonE3 1585Lv5 workstation using only a single core, although both approaches could be run in parallel. AM1-BCC was performed with and without geometry optimization (1SCF). The geometry optimization increases the computational cost dramatically, but is often necessary in a practical setting. In the ML-based approach, this work is done once for the training set and afterward the prediction requires only the 2D topological information on the molecule. The computational speed-up compared to a DFT-calculation with subsequent charge assignment using the DDEC scheme is several orders of magnitude.

Combination with Existing General Force Fields. Computing condensed-phase properties such as density, heat of vaporization or free energy of hydration (ΔG^{hyd}) is a common strategy for force-field validation. These properties can be accurately measured in experiment and reference data are available for a wide range of compounds.

Since compatibility of the ML-predicted partial charges with GAFF or OPLS-AA cannot be better than that of the DDEC reference partial charges, we first calculated the density and heat of vaporization for the 146 organic liquids with DDEC reference partial charges as baseline. The parametrization of GAFF and OPLS-AA were done with partial charges obtained in vacuum and fitted to experimental data of organic liquids.[5,6] Here, we investigated how the extracted partial charges are influenced by the dielectric permittivity of an implicit solvent used in the QM calculations. Three different permittivities were used: $\epsilon = 1$, 4, or 78. In Table 3, the partial charges of water for the TIP3P[89] model are compared with DDEC partial charges obtained using different permittivities for the implicit solvation model. The DDEC partial charges obtained with $\epsilon = 4$ present a good compromise between the underpolarization with $\epsilon = 1$ and overpolarization with $\epsilon = 78$. In addition, $\epsilon = 4$ mimics the environment of a molecule in a protein.[58,90] We therefore expect that DDEC partial charges obtained using $\epsilon = 4$ for the continuum model match best with the other parameters in the force field for the calculation of hydration free energies.

Table 3. Partial Charges for O and H in a Water Molecule for the TIP3P[89] Model and Using DDEC from Electron Densities Obtained Using the Implicit Solvation Model with ϵ = 1, 4, and 78

	TIP3P	$\epsilon = 1$	$\epsilon = 4$	$\epsilon = 78$
q(O)	−0.834	−0.790	−0.880	−0.990
q(H)	+0.417	+0.395	+0.440	+0.495

Density and Heat of Vaporization. In Table 4, the RMSE to the experimental values is given for the different partial-charge sets. The correlation plot for OPLS-AA and $\epsilon = 4$ is shown in Figure 6, the same plots for GAFF and the other permittivity values are shown in Figures S8−S13 in the Supporting Information. The LJ parameters for bromine in GAFF are known to be problematic, which results in underestimated densities and overestimated heat of vaporization.[78] Therefore, the RMSE for the whole data set and for all molecules, which do not contain Br, is given. In general, the DDEC partial charges exhibit a good compatibility with both force fields, with overall the best performance for the partial charges obtained with $\epsilon = 4$. To further improve the performance, the LJ parameters would have to be refitted to the DDEC partial charges. Partial charges obtained with $\epsilon = 78$ are overpolarized and do not match with the LJ parameters. The ML-predicted partial charges result in similar densities and heats of vaporization as the DDEC reference partial charges.

Table 4. RMSE of the Calculated Densities and Heats of Vaporization for the Set of 146 Organic Liquids[78,a]

force field	charge set	RMSE ϱ [kg/m^3]	RMSE ϱ^b [kg/m^3]	RMSE ΔH_{vap} [kJ/mol]	RMSE $\Delta H_{\mathrm{vap}}{}^b$ [kJ/mol]
GAFF	AM1-BCC	82.5	51.6	10.5	10.7
	DDEC ($\epsilon = 1$)	83.5	46.3	10.7	10.5
	DDEC ($\epsilon = 4$)	90.0	44.4	9.1	9.1
	DDEC ($\epsilon = 78$)	79.2	51.5	14.6	14.3
	ML ($\epsilon = 4$)	83.1	44.1	8.9	9.0
OPLS-AA	CM1	40.4	34.0	7.8	7.9
	DDEC ($\epsilon = 1$)	48.4	45.2	9.0	9.2
	DDEC ($\epsilon = 4$)	45.5	43.3	9.4	9.4
	DDEC ($\epsilon = 78$)	56.2	45.5	14.4	14.6
	ML ($\epsilon = 4$)	42.5	38.5	8.9	9.1

[a]Reference values for the original GAFF (AM1-BCC) and OPLS-AA (CM1[91]) force fields were taken from ref 78. The DDEC reference partial charges were obtained using the implicit solvation model with ϵ = 1, 4, and 78. The ML-predicted partial charges were with $\epsilon = 4$. The partial charges were combined with the GAFF and the OPLS-AA force field. [b]Without compounds containing Br.

Hydration Free Energy. The 642 molecules for which experimental hydration free energy data is available range from small organic liquids to larger polyfunctional molecules.[81−83] The partial charge prediction accuracy of the ML models for this set of molecules is shown in Figure S14 in the Supporting Information. With a RMSE of 0.021 e, it is comparable to that observed for the external test sets. Again, the outliers are molecules like ammonia, hydrogen sulfide, and ethylene, which are much smaller compared to the ones in the training set. Note that the following six molecules were also present in the training

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

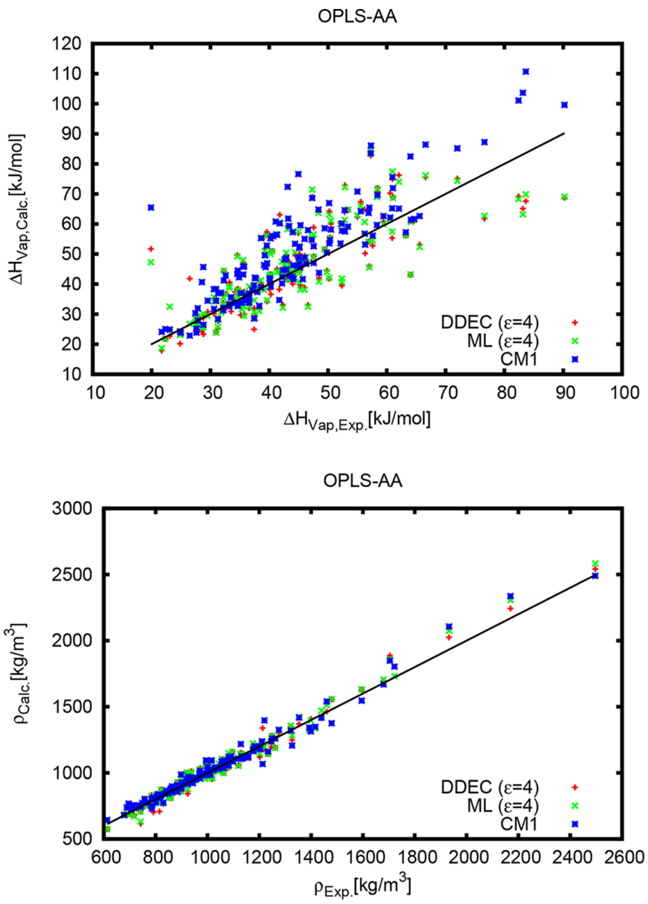


Figure 6. Experimental versus calculated heats of vaporization (top) or density (bottom) using the OPLS-AA force field in combination with DDEC reference partial charges (implicit solvent with ϵ = 4) and the ML-predicted partial charges (also ϵ = 4). For comparison, the results with CM1 partial charges taken from ref 78 are shown.

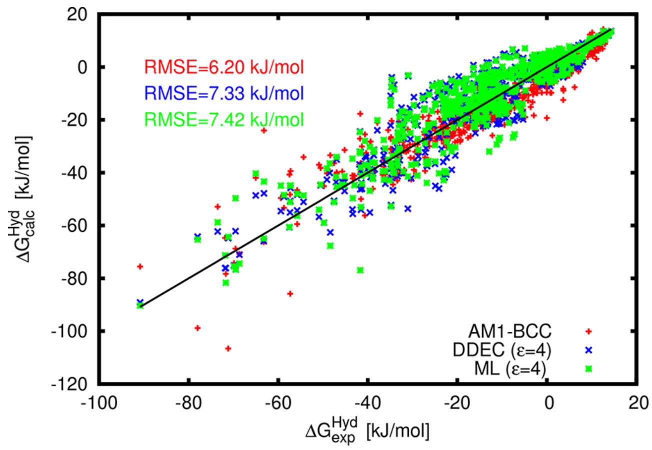


Figure 7. Correlation between calculated and experimental hydration free energies ΔG^{hyd} for the set of 642 compounds taken from the FreeSolv database.[81−83] Reference values for the original GAFF (AM1-BCC) force field were taken from ref 83. The DDEC reference partial charges were obtained using the implicit solvation model with ϵ = 4. The ML-predicted partial charges were with ϵ = 4. The partial charges were combined with the GAFF force field. The results for ϵ = 1 are shown separately in Figure S15 in the Supporting Information.

Table 5. RMSE, Maximal Error, Excess Kurtosis, and R^2 Values of the Calculated versus Experimental Hydration Free Energies ΔG^{hyd} for the Set of 642 Compounds Taken from the FreeSolv database[81−83,a]

force field	charge set	RMSE [kJ/mol]	MAX [kJ/mol]	excess kurtosis	R^2
GAFF	AM1-BCC	6.2	39.2	3.62	0.87
	DDEC (ϵ = 1)	9.4	35.1	2.62	0.71
	DDEC (ϵ = 4)	7.3	30.8	1.66	0.82
	ML (ϵ = 4)	7.4	35.2	2.02	0.82

[a]Reference values for the original GAFF (AM1-BCC) force field were taken from ref 83. The DDEC reference partial charges were obtained using the implicit solvation model with ϵ = 1 and 4. The ML-predicted partial charges were with ϵ = 4. The partial charges were combined with the GAFF force field.

set (ID in the FreeSolv database): 1417007, 1770205, 194273, 5076071, 6334915, and 9281946.

The correlation between the calculated and experimental ΔG^{hyd} values are shown in Figure 7 and Figure S15 in the Supporting Information. As the implicit solvation model with a dielectric permittivity ϵ = 78 leads to overpolarized partial charges (with respect to the LJ parameters), only partial charges obtained with ϵ = 1 or 4 were used. The RMSE with the DDEC partial charges (with ϵ = 4) is slightly higher than with AM1-BCC, but the total number of outliers is reduced as can be seen by the lower excess kurtosis (Table 5).

For ϵ = 1, the RMSE of 9.4 kJ/mol is considerably larger than the RMSE of 7.3 kJ/mol for ϵ = 4. The systematic underestimation of ΔG^{hyd} with the ϵ = 1 partial charges (Figure S15 in the Supporting Information) is a direct consequence of the too little polarized vacuum-derived charges. Although the AM1-BCC charges are generated in vacuum as well, the corresponding values for ΔG^{hyd} reproduce much better the experimental ones. This can be explained by the influence of the bond charge corrections (BCC), which compensate for the lack of polarization in the bare AM1 charges. With the DDEC partial charges, the ΔG^{hyd} values of polar and polyfunctional molecules are better reproduced. The results can certainly be improved when the LJ parameters are reparametrized with the DDEC partial charges.

As expected from the good prediction of the DDEC partial charges by the ML models, the RMSE using the ML-predicted partial charges is only 0.1 kJ/mol higher than that using the DDEC reference partial charges (Table 5). The correlation between the ML-based values for ΔG^{hyd} and the DDEC reference values is shown in Figure S16 in the Supporting Information. The largest outliers, i.e. formaldehyde and α-endosulfan, can be explained by low occurrence of some atom environments in the training set.

■ CONCLUSIONS

In this work, an ML-based approach for the prediction of partial charges for classical fixed-charge force fields has been presented. The ML models were trained on a large and diverse data set of approximately 130 000 lead-like compounds, which allowed the reliable prediction of the partial charges of unseen molecules, both smaller and larger than the ones in the training set. The ML model can further be improved by extending the training set with substructures that are not yet well represented. The predicted partial charges are based on DFT calculations with the TPSSh functional and a reasonably large basis set, supplemented by an implicit solvent model. The accuracy of the electron densities was confirmed by benchmarking several exchange-correlation func-

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

tionals against CCSD reference densities. The chosen charge-extraction scheme DDEC is known to yield chemically meaningful and conformationally robust partial charges. The latter property is crucial for fixed-charge force fields. The atom-centered AP fingerprints, which were used as descriptors for the ML models, rely only on 2D topological information, such that any remaining conformational dependence of the partial charges is removed. This allows for an easy application of the ML models as only a SMILES string is necessary to predict the partial charges of a new molecule. In addition, the ML-based approach allows the assignment of high-quality partial charges within the fraction of a second for a typical drug-like molecule. This results in a speed-up of several orders of magnitude compared to the original derivation of DDEC (or RESP) partial charges.

As the molecules in the training set were chosen to represent all substructures found in the lead-like portion of the ZINC and ChEMBL database, also rather exotic substructures which do not occur often are part of the training set. It is challenging for the ML models to predict the partial charges of such substructures correctly. This is most evident for phosphorus (only present in ChEMBL molecules), which has a large range of possible partial charges but only relatively few examples. The excess-charge correction scheme can smooth some of the prediction error. Furthermore, by simply adding molecules with underrepresented substructures to our database in the future, the prediction accuracy of the partial charges can be increased. For typical drug-like molecules (as shown by the external test set 2 containing FDA approved drugs), the partial charges can already be predicted with high accuracy (RMSE = 0.016 e) with the current training set.

The compatibility of the DDEC partial charges with the OPLS-AA and GAFF force fields was successfully tested by computing basic thermodynamic quantities such as density, heat of vaporization and hydration free energy. The DDEC partial charges gave comparable results as the AM1-BCC or CM1 charges, if an appropriate dielectric permittivity (ϵ = 4) was chosen for the implicit solvation model used in the QM calculations. The ML-predicted partial charges performed similarly to the DDEC reference partial charges. For further improvement, LJ parameters would need to be reparametrized together with DDEC partial charges.

Besides the application in MD simulations, partial charges are widely used as descriptors in QSAR/QSPR models. Using the ML-based approach presented here, accurate partial charges can be obtained efficiently for application in high-throughput ligand-based virtual screening.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00663.

Results of the benchmarking for the electron density and additional Figures S1−S16 (PDF)

Text files with the results from the calculation of the densities, heats of vaporization, and hydration free energies as well as the SMILES for the 146 organic liquids (external test set 1) and the 642 molecules from the FreeSolv database (ZIP)

Python scripts to train the RFR models and use them for predicting the partial charges of an example compound (ZIP)

Text file with the SMILES for the 1081 molecules in external test set 2 (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: sriniker@ethz.ch.

ORCID

Sereina Riniker: 0000-0003-1893-4031

Notes

The authors declare no competing financial interest.

SDFs with the optimized geometries of the molecules in the training set together with the DDEC reference partial charges (with ϵ = 4 and ϵ = 78) can be downloaded from https://www.research-collection.ethz.ch/handle/20.500.11850/230799 free of charge.

■ ACKNOWLEDGMENTS

The authors thank Andrew Dalke for a modified version of the ChemFp package and Gregory Landrum for useful discussions. The authors gratefully acknowledge financial support by the Swiss National Science Foundation (Grant Number 200021-159713) and by ETH Zürich (ETH-08 15-1).

■ REFERENCES

(1) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.

(2) Jorgensen, W. L.; Tirado-Rives, J. The OPLS Potential Functions for Proteins. Energy Minimisations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.* **1988**, *110*, 1657−1666.

(3) Mackerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.

(4) 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.

(5) 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.

(6) 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.

(7) Jorgensen, W. L.; Tirado-Rives, J. Potential Energy Functions for Atomic-level Simulations of Water and Organic and Biomolecular Systems. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6665−6670.

(8) Vanommeslaeghe, K.; MacKerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. *J. Chem. Inf. Model.* **2012**, *52*, 3144−3154.

(9) Vanommeslaeghe, K.; MacKerell, A. D.; Raman, E. P. 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.

(10) 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.

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

(11) 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.
(12) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parametrisation, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(13) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *J. Comput. Chem.* **1996**, *17*, 553−586.
(14) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 Van der Waals and Electrostatic Parameters for Intermolecular Interactions. *J. Comput. Chem.* **1996**, *17*, 520−552.
(15) Casewit, C. J.; Colwell, K. S.; Rappé, A. K. Application of a Universal Force Field to Organic Molecules. *J. Am. Chem. Soc.* **1992**, *114*, 10035−10046.
(16) Singh, U. C.; Kollman, P. A. An Approach to Computing Electrostatic Charges for Molecules. *J. Comput. Chem.* **1984**, *5*, 129−145.
(17) Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. A Well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: The RESP model. *J. Phys. Chem.* **1993**, *97*, 10269−10280.
(18) Mulliken, R. S. Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833−1840.
(19) Hirshfeld, F. L. Bonded-atom Fragments for Describing Molecular Charge Densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.
(20) Bultinck, P.; van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical Analysis and Extension of the Hirshfeld Atoms in Molecules. *J. Chem. Phys.* **2007**, *126*, 144111.
(21) Lillestolen, T. C.; Wheatley, R. J. Redefining the Atom: Atomic Charge Densities Produced by an Iterative Stockholder Approach. *Chem. Commun.* **2008**, 5909−5911.
(22) Bader, R. F. W. Molecular fragments or chemical bonds. *Acc. Chem. Res.* **1975**, *8*, 34−40.
(23) Bader, R. F. W.; MacDougall, P. J.; Lau, C. D. H. Bonded and nonbonded charge concentrations and their relation to molecular geometry and reactivity. *J. Am. Chem. Soc.* **1984**, *106*, 1594−1605.
(24) Bader, R. F. W. Everyman's Derivation of the Theory of Atoms in Molecules. *J. Phys. Chem. A* **2007**, *111*, 7966−7972. PMID: 17629258.
(25) Manz, T. A.; Limas, N. G. Introducing DDEC6 Atomic Population Analysis: Part 1. Charge Partitioning Theory and Methodology. *RSC Adv.* **2016**, *6*, 47771−47801.
(26) Manz, T. A.; Sholl, D. S. Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.* **2010**, *6*, 2455−2468.
(27) Manz, T. A.; Sholl, D. S. Improved Atoms-In-Molecule Charge Partitioning Functional for Simultaneously Reproducing the Electrostatic Potential and Chemical States in Periodic and Nonperiodic Materials. *J. Chem. Theory Comput.* **2012**, *8*, 2844−2867.
(28) Cisneros, G. A.; Karttunen, M.; Ren, P.; Sagui, C. Classical Electrostatics for Biomolecular Simulations. *Chem. Rev.* **2014**, *114*, 779−814.
(29) Kramer, C.; Spinn, A.; Liedl, K. R. Charge Anisotropy: Where Atomic Multipoles Matter Most. *J. Chem. Theory Comput.* **2014**, *10*, 4488−4496.
(30) Dixon, R. W.; Kollman, P. A. Advancing Beyond the Atom-centered Model in Additive and Nonadditive Molecular Mechanics. *J. Comput. Chem.* **1997**, *18*, 1632−1646.
(31) Mahoney, M. W.; Jorgensen, W. L. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential Functions. *J. Chem. Phys.* **2000**, *112*, 8910−8922.
(32) Cole, D. J.; Vilseck, J. Z.; Tirado-Rives, J.; Payne, M. C.; Jorgensen, W. L. Biomolecular Force Field Parameterization Via Atoms-In-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2016**, *12*, 2312−2323.
(33) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural Population Analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.
(34) Wan, J.; Zhang, L.; Yang, G.; Zhan, C.-G. Quantitative Structure-Activity Relationship for Cyclic Imide Derivatives of Protoporphyrinogen Oxidase Inhibitors: A Study of Quantum Chemical Descriptors from Density Functional Theory. *J. Chem. Inf. Comp. Sci.* **2004**, *44*, 2099−2105.
(35) Finkelmann, A. R.; Goller, A. H.; Schneider, G. Robust Molecular Representations for Modelling and Design Derived from Atomic Partial Charges. *Chem. Commun.* **2016**, *52*, 681−684.
(36) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van der Waals Interactions from Ground-state Electron Density and Free-atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, 073005.
(37) Ramakrishnan, R.; von Lilienfeld, O. A. Machine Learning, Quantum Mechanics, and Chemical Compound Space. In *Reviews in Computational Chemistry*; Parrill, A. L., Lipkowitz, K. B., Eds.; Wiley-VCH, 2017; pp 225−256.
(38) Dral, P. O.; von Lilienfeld, O. A.; Thiel, W. Machine Learning of Parameters for Accurate Semiempirical Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2015**, *11*, 2120−2125.
(39) Shen, L.; Wu, J.; Yang, W. Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J. Chem. Theory Comput.* **2016**, *12*, 4934−4946.
(40) Rai, B. K.; Bakken, G. A. Fast and Accurate Generation of ab initio Quality Atomic Charges Using Nonparametric Statistical Regression. *J. Comput. Chem.* **2013**, *34*, 1661−1671.
(41) Bereau, T.; Andrienko, D.; von Lilienfeld, O. A. Transferable Atomic Multipole Machine Learning Models for Small Organic Molecules. *J. Chem. Theory Comput.* **2015**, *11*, 3225−3233.
(42) Ivanov, M. V.; Talipov, M. R.; Timerghazin, Q. K. Genetic Algorithm Optimization of Point Charges in Force Field Development: Challenges and Insights. *J. Phys. Chem. A* **2015**, *119*, 1422−1434.
(43) Li, Y.; Li, H.; Pickard, F. C.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S. K. R. S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *J. Chem. Theory Comput.* **2017**, *13*, 4492−4503.
(44) Irwin, J. J.; Shoichet, B. K. ZINC - a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(45) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J. Comput. Chem.* **2005**, *26*, 1752−1780.
(46) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies With Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(47) Oliferenko, A. A.; Pisarev, S. A.; Palyulin, V. A.; Zefirov, N. S. Atomic Charges Via Electronegativity Equalization: Generalizations and Perspectives. *Adv. Quantum Chem.* **2006**, *51*, 139−156.
(48) Raček, T.; Pazúriková, J.; Svobodová Vařeková, R.; Geidl, S.; Křenek, A.; Falginella, F. L.; Horský, V.; Hejret, V.; Koča, J. NEEMP: Software for Validation, Accurate Calculation and Fast Parameterization of EEM Charges. *J. Cheminf.* **2016**, *8*, 57.
(49) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.
(50) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-activity Studies: Definition and Applications. *J. Chem. Inf. Model.* **1985**, *25*, 64−73.
(51) RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org (accessed July 24, 2017).
(52) Sommer, K.; Friedrich, N.-O.; Bietz, S.; Hilbig, M.; Inhester, T.; Rarey, M. UNICON: A Powerful and Easy-to-use Compound Library Converter. *J. Chem. Inf. Model.* **2016**, *56*, 1105−1111.
(53) Limas, N. G.; Manz, T. A. Introducing DDEC6 atomic population analysis: part 2. Computed results for a wide range of periodic and nonperiodic materials. *RSC Adv.* **2016**, *6*, 45727−45747.

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590

(54) Bayly, C. I.; McKay, D. J. Atomic Partial Charges for Fixed-charge Force Fields: Dealing With Conformational Dependence. Presented at the *250th ACS National Meeting*, Boston, MA, August 16−20, 2015.
(55) Stewart, J. J. P. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-Optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1−32.
(56) Stewart, J. P. *MOPAC2016*; 2016.
(57) Klamt, A.; Schuurmann, G. COSMO: a New Approach to Dielectric Screening in Solvents With Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799−805.
(58) Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J. W.; Wang, J.; Kollman, P. A. 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.
(59) Hohenberg, P.; Kohn, W. Inhomogeneous Electron Gas. *Phys. Rev.* **1964**, *136*, B864−B871.
(60) Kohn, W.; Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **1965**, *140*, A1133−A1138.
(61) Adamo, C.; Barone, V. Toward Reliable Density Functional Methods Without Adjustable Parameters: The PBE0Model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(62) Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the Density Functional Ladder: Nonempirical Meta-generalized Gradient Approximation Designed for Molecules and Solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.
(63) Staroverov, V. N.; Scuseria, G. E.; Tao, J.; Perdew, J. P. Comparative Assessment of a New Nonempirical Density Functional: Molecules and Hydrogen-bonded Complexes. *J. Chem. Phys.* **2003**, *119*, 12129−12137.
(64) Becke, A. D. Becke's Three Parameter Hybrid Method Using the LYP Correlation Functional. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(65) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-energy Formula Into a Functional of the Electron Density. *Phys. Rev. B: Condens. Matter Mater. Phys.* **1988**, *37*, 785.
(66) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-Dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: a Critical Analysis. *Can. J. Phys.* **1980**, *58*, 1200−1211.
(67) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(68) Zhao, Y.; Truhlar, D. G. The M06 Suite of Density Functionals for Main Group Thermochemistry, Thermochemical Kinetics, Non-covalent Interactions, Excited States, and Transition Elements: Two New Functionals and Systematic Testing of Four M06-class Functionals and 12 Other Functionals. *Theor. Chem. Acc.* **2008**, *120*, 215−241.
(69) Ahlrichs, R.; Bär, M.; Häser, M.; Horn, H.; Kölmel, C. Electronic Structure Calculations on Workstation Computers: The Program System TURBOMOLE. *Chem. Phys. Lett.* **1989**, *162*, 165−169.
(70) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(71) Weigend, F. Accurate Coulomb-fitting Basis Sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057−1065.
(72) Eckert, F.; Klamt, A. Fast Solvent Screening Via Quantum Chemistry: COSMO-RS Approach. *AIChE J.* **2002**, *48*, 369−385.
(73) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. W. Refinement and Parametrization of COSMO-RS. *J. Phys. Chem. A* **1998**, *102*, 5074−5085.
(74) Klamt, A. The COSMO and COSMO-RS Solvation Models. *WIREs Comput. Mol. Sci.* **2011**, *1*, 699−709.
(75) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(76) Scikit-Learn−Machine Learning in Python. http://scikit-learn.org (accessed July 24, 2017).
(77) Dalke, A. The FPS Fingerprint Format and Chemfp Toolkit. *J. Cheminf.* **2013**, *5*, 36.
(78) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8*, 61−74.
(79) Sterling, T.; Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324−2337.
(80) van der Spoel, D.; van Maaren, P. J.; Caleman, C. GROMACS Molecule and Liquid Database. *Bioinformatics* **2012**, *28*, 752−753.
(81) Mobley, D. L.; Bayly, C. I.; Cooper, M. D.; Dill, K. A. Predictions of Hydration Free Energies from All-Atom Molecular Dynamics Simulations. *J. Phys. Chem. B* **2009**, *113*, 4533−4537.
(82) 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.
(83) 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.
(84) Jakalian, A.; Jack, D. B.; Bayly, C. Fast, Efficient Generation of High-quality Atomic Charges. AM1-BCC: II. Parametrization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(85) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling Through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.
(86) Zwanzig, R. W. High Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases. *J. Chem. Phys.* **1954**, *22*, 1420−1426.
(87) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(88) AmberTools17. http://ambermd.org/#AmberTools (accessed July 24, 2017).
(89) 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.
(90) Lee, L. P.; Cole, D. J.; Skylaris, C.-K.; Jorgensen, W. L.; Payne, M. C. Polarized Protein-Specific Charges from Atoms-in-Molecule Electron Density Partitioning. *J. Chem. Theory Comput.* **2013**, *9*, 2981−2991.
(91) Storer, J. W.; Giesen, D. J.; Cramer, C. J.; Truhlar, D. G. Class IV Charge Models: A New Semiempirical Approach in Quantum Chemistry. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 87−110.

DOI: 10.1021/acs.jcim.7b00663
J. Chem. Inf. Model. 2018, 58, 579−590