Title

Hydration Free Energies in the FreeSolv Database Calculated with Polarized Iterative Hirshfeld Charges.

Permalink
https://escholarship.org/uc/item/8jj5444k

Journal
Journal of chemical information and modeling, 58(9)

ISSN
1549-9596

Authors
Riquelme, Maximiliano
Lara, Alejandro
Mobley, David L
et al.

Publication Date
2018-09-01

DOI
10.1021/acs.jcim.8b00180

Peer reviewed

Published in final edited form as:
J Chem Inf Model. 2018 September 24; 58(9): 1779–1797. doi:10.1021/acs.jcim.8b00180.

Hydration Free Energies in the FreeSolv Database Calculated with Polarized Iterative Hirshfeld Charges

Maximiliano Riquelme[†], **Alejandro Lara**[†], **David L. Mobley**[‡], **Toon Verstraelen**[¶], **Adelio R. Matamala**[†], and **Esteban Vöhringer-Martinez**[†]

[†]Departamen to de Físico-Química, Facultad de Ciencias Químicas, Universidad de Concepción, 4070386 Concepción, Chile

[‡]Departments of Pharmaceutical Sciences and Chemistry, 147 Bison Modular, University of California, Irvine, Irvine, California, USA 92617

[¶]Center for Molecular Modeling (CMM), Ghent University, Technologiepark 903, B-9052, Ghent, Belgium

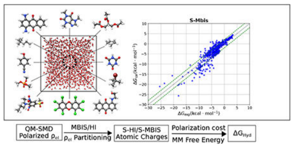
Abstract

Computer simulations of bio-molecular systems often use force fields, which are combinations of simple empirical atom-based functions to describe the molecular interactions. Even though polarizable force fields give a more detailed description of intermolecular interactions, nonpolarizable force fields, developed several decades ago, are often still preferred because of their reduced computation cost. Electrostatic interactions play a major role in bio-molecular systems and are therein described by atomic point charges. In this work, we address the performance of different atomic charges to reproduce experimental hydration free energies in the FreeSolv database in combination with the GAFF force field. Atomic charges were calculated by two atoms-in-molecules approaches, Hirshfeld-I and Minimal Basis Iterative Stockholder (MBIS). To account for polarization effects, the charges were derived from the solute's electron density computed with an implicit solvent model and the energy required to polarize the solute was added to the free energy cycle. The calculated hydration free energies were analyzed with an error model, revealing systematic errors associated with specific functional groups or chemical elements. The best agreement with the experimental data is observed for the AM1-BCC and the MBIS atomic charge methods. The latter includes the solvent polarization and present a root mean square error of 2.0 kcal mol^{-1} for the 613 organic molecules studied. The largest deviation was observed for phosphorus-containing molecules and the molecules with amide, ester and amine functional groups.

Graphical TOC Entry

evohringer@udec.cl, Phone: +56 41 2204986.

Supporting Information Available
The electronic structure method and the water model dependence of the hydration free energies for the S-HI atomic charges is shown in Figure S1 and S3. Solvation free energies obtained with the SMD model from the electronic structure calculations are shown in Figure S4 with their outliers in Figure S5. Figure S6–8 displays the hydration free energies per functional groups calculated with the S-HI, semipolarized and AM1-BCC atomic charges. Tables S1–3 show additional information on the error model and in Table S4 all hydration free energies are shown for each set of atomic charges.
This material is available free of charge via the Internet at http://pubs.acs.org/.



Introduction

The first step in modeling a molecular system at atomic resolution demands a correct description of the interactions between their atoms. In principle, to describe these interactions accurately the electrons of each atom must be accounted for by quantum mechanics. In biological systems, however, the large number of atoms makes a quantum mechanical treatment prohibitive, especially since most experimental observables require extensive sampling of the configurational space. In the last decades force fields, which are mathematical simple functions describing the atomic interactions, have been developed and applied to various biological problems such as ligand binding to protein or DNA, protein folding or enzyme catalysis.

However, the applicability of these force fields to predict biological function of complex systems is limited by their ability to reproduce atomic interactions correctly. The binding process of a ligand to a protein for example is associated with its partial desolvation, the stripping off of the surrounding water molecules. Considering only this first simple step of the binding process, various computational methods and force fields have been tested in their ability to reproduce the free energy associated with the transfer of a molecule or ligand from aqueous solution into the gas phase: the hydration free energy ΔG_{Hyd}. – The SAMPL4 challenge is one example where various methods have been applied to calculate the hydration free energies of 42 organic compounds. The number of compounds was recently extended to over 600 molecules in the FreeSolv database, which contains calculated hydration free energies with the Generalized Amber Force Field (GAFF) and the AM1 - Bond Charge Corrected (AM1-BCC) charges– reported by one of us, and experimental reference values. The extensive number of compounds, the variety of functional groups, and the consistent experimental data make this database ideally suited for the validation of force fields for small organic ligands and especially new methods to derive force-field parameters.

For polar molecules, the dominant factor in the hydration free energies are the electrostatic interactions represented by the atomic charges in force fields. Different methods were proposed to determine these atomic charges for bio-molecular force fields. For example, atomic charges are adjusted to reproduce hydration free energies (OPLS or GROMOS), or interaction energies with some discrete water molecules (CH ARMM) for small representative molecules. In the AMBER force field, atomic charges are obtained with the RESP method, which relies on the molecular electrostatic potential obtained from HF/6–31G(d) electronic structure calculation in vacuum. The GAFF force field is often combined with the AM1-BCC charge derivation method, which relies on an empirical AM1 electronic structure Hamiltonian and bond charge corrected atomic charges, which were parameterized to reproduce the charges from HF/6–31G(d) calculations and corrected to reproduce experimental hydration free energies.– However, both methods do not account for the varying polarization in different molecular environments; although both have shown success in reproducing hydration free energies due to a possible overpolarization of the molecular electron density by the employed electronic structure method or due to ad-hoc bond charge corrections.

Many phenomena, such as ligand binding, inherently involve the transfer of a guest molecule from one environment to another, resulting in a change of polarization of the guest. This affects the internal energy of the guest molecule (cost of polarization) and enhances the interaction with its environment. Such polarization effects were introduced recently in biomolecular force fields, e.g. through atomic inducible dipoles in the AMOEBA force field or with Drude oscillators in the CHARMM polarizable force field. The main drawbacks of these models are their increased complexity (more parameters) and increased computational cost. In addition, it was recently shown that polarization energy of atomic dipole polarizable force fields can deviate severely from the induction energy of Symmetry-Adapted Perturbation Theory (SAPT) calculations on the molecular dimers in the S66 set. Because a polarizable force field and the SAPT induction term try to describe the same physics, this discrepancy discourages the use more expensive polarizable force fields. For these reasons, it is still relevant to understand how one can account for polarization in conventional non-polarizable force fields.

Maintaining the simple character of non-polarizable force fields, we propose to derive atomic charges by partitioning the molecular electron density (Hirshfeld-I or Minimal Basis Iterative Stockholder) from electronic structure calculations of the solute. Hydration free energies were computed with these atomic charges, in combination with remaining GAFF van-der-Waals and bonded force-field parameters for the solute, and two water models. Our results were compared to the corresponding values in the FreeSolv database (v0.51) - particularly, calculated values with GAFF and AM1-BCC charges, as well as experimental values. We also address the effect of solvent polarization on the electronic structure calculation prior to the density partitioning with the implicit SMD model.

Hirshfeld-I and MBIS were selected because they rely only on the electron density without any additional parameters, reproduce the electrostatic potential in the gas phase, present only a minor conformational dependence,– and reproduce electrostatic interaction energies of molecular dimers better than most other charge definitions. Hirshfeld-I (HI) charges are derived from atom-in-molecule densities that maximize the similarity to so-called proatom densities. The method is iterative because it employs a self-consistent loop that enforces the same charge on every atom and corresponding pro-atom. Fractionally charged pro-atoms, needed in Hirshfeld-I, are constructed by taking linear combinations of densities of free atoms or ions with an integer charge. One drawback is that Hirshfeld-I becomes poorly defined when it requires densities of anions, to construct pro-atoms, which are not stable in gas phase (most elements can only bind one excess electron). Different ad hoc recipes exist to stabilize these anions by embedding them in a somewhat arbitrary confining potential,, resulting in different Hirsfeld-I charges. The Minimal Basis Iterative Stockholder (MBIS) method does not require isolated atom densities as input but uses a simple analytic proatom instead. A numerical assessment on molecular databases shows that MBIS charges are effective for modeling both electrostatic potentials and electrostatic interactions.

Methods

The FreeSolv Database

The FreeSolv database version 0.51 consists of 642 neutral molecules containing a variety of functional groups covering a significant range of chemistry. Many of these compounds come from earlier literature databases, but additional values/compounds have been added from the series of SAMPL blind challenges.,– Compounds are typically small and fragmentlike, with few rotatable bonds, though some compounds do have appreciable flexibility and multiple rotatable bonds. Likewise many compounds have relatively few functional groups in combination, but some have several., The database is freely available on GitHub, and includes experimental values from the literature, as well as calculated hydration free energies from TIP3P and GAFF/AM 1-BCC.

For this study we left out the carboxylic acids and iodine-containing molecules. For the carboxylic acids, the problem was related to the increased atomic charge of the hydroxyl hydrogen atom with no vdW parameters that led to unusually large forces with the surrounding solvent, resulting in free energy calculations which unfailingly crashed. This may indicate a need to introduce Lennard-Jones parameters on the hydroxyl hydrogen, which has been done for very related reasons in the new smirnoff99Frosst force field. The chemical element iodine was not present in the basis set employed for the electronic structure calculations and therefore the iodine-containing molecules were left out. Additionally, three molecules with ID 1189457, 5948990 and 7794077 were not included because no minimum in the potential energy during the geometry optimization cycle with the SMD solvation model and the ORCA package was found. In total 614 molecules were considered in the study from which one is a duplicate yielding a total of 613 molecules used for the analysis.

Force-field parameterization with Hirshfeld-I and MBIS atomic charges

Hydration free energies for all considered molecules of the FreeSolv database (version 0.51) were computed with several variants of the GAFF force field, which only differ in the method used to assign the atomic charges. In addition to the conventional AM1-BCC charges, we also derived atomic charges from the electron density of the solute, either computed in vacuum or with the SMD continuum solvent model., Geometry optimizations and electron densities were computed at the BLYP– and B3LYP level of theory with the def2-TZVP, basis, using ORCA 3.0.2. Most results were obtained with the BLYP method while the B3LYP results were only used to briefly assess the influence of exact exchange on the partial charges and the hydration free energies.

The atomic charges were derived from electron densities with the Hirshfeld-I and Minimal Basis Iterative Stockholder (MBIS) methods, as implemented in the HORTON package version 2.0.0. Both methods partition the molecular electron density, ρ_{mol}, into atom-in-molecule densities, ρ_{A}, with the stockholder formula originally proposed by Hirshfeld:

$$\rho_A(r) = \rho_{\mathrm{mol}}(r)\frac{\rho_A^0(r)}{\sum_B \rho_B^0(r)} \quad (1)$$

in which $\rho_A^0(r)$ is the spherical pro-atom density of atom A centered on the corresponding nucleus and the denominator contains the pro-molecule density. Originally, these pro-atoms were spherically averaged densities of neutral isolated atoms, but this choice is somewhat arbitrary and leads to charges which are nearly zero, such that electrostatic interactions are underestimated., In the Hirshfeld-I method, each pro-atom is a linear interpolation between the densities of spherically averaged isolated atoms and/or ions. Consistency between the pro-atom and atom-in-molecule charge is imposed by an iterative self-consistent algorithm. Hirshfeld-I charges are considerably larger in magnitude than Hirshfeld charges, they reproduce electrostatic potentials of organic molecules reasonably well and they have a much lower conformational sensitivity than RESP charges. However, in some corner cases, most notably metal-oxides, Hirshfeld-I charges become ill-defined. For example, the oxygen dianion is unstable in gas phase and there exists no ground state density of this isolated dianion. The second excess electron would rather drift away to infinity than being bound. In actual calculations, electrons are somewhat localized by basis set limitations and one then obtains an artificial and very diffuse dianion density instead. An oxygen pro-atom with a fractional charge between −1.0 e and −2.0 e is then also artificial because it requires an interpolation between the isolated anion and dianion density. Such pro-atoms are typically needed when computing Hirshfeld-I charges of a molecule in which oxygen is bound to an element with a high oxidation state. In these cases, Hirshfeld-I charges reach very large magnitudes and tend to overestimate the molecular dipole moment., Several refinements to Hirshfeld-I were proposed, including embedding methods to stabilize (di)anions, or methods that do not rely on unstable ions.,, To avoid difficulties with the unstable oxygen dianion, Hirshfeld-I charges were not computed for molecules containing phosphates and sulfonates. For those cases, Hirshfeld-E charges were used instead. The Hirshfeld-E method uses as pro-atom a linear combination of spherically averaged Fukui functions, such that unstable ions are no longer needed.

In the MBIS method, the pro-atom is approximated by a sum of exponentially decaying spherical functions, with one such function per shell. The amplitudes and widths of the exponential functions are optimized by minimizing the Kullback-Leibler divergence between the pro-molecule and DFT electron densities.

After the calculation of the atomic charge with each method described above, chemically equivalent atoms, which result from molecular symmetry or free rotations, e.g. hydrogen atoms in methyl group, were identified with the OpenEye Python Toolkit (version 2017.2.1) and their charges were averaged. To distinguish the different charge sets, the ones from vacuum calculations with the Hirshfeld-I partitioning were named HI, the ones with the same partitioning method but with an implicit solvation model S-HI, the ones obtained with the MBIS partitioning method and the implicit solvation model S-MBIS and the original ones from the FreeSolv database AM1-BCC.

Note that the electrostatic term in the GAFF force field uses simple point charges and does not account for short-range screening due to the finite extent of the atomic electron densities, also known as the penetration effect.,,– While the inclusion of such screening makes the electrostatic term more physically sound, it is not recommended in our case. The GAFF Lennard-Jones parameters account implicitly for this screening effect and adding it also explicitly, which drastically changes the interaction potential, would result in a double counting.

Because we can assume to good approximation that Hirshfeld-I and MBIS charges are not geometry-dependent, we treat them as constants during the MD simulations in this work. In other words, we do not include charge polarizability during the MD simulations. The influence of the solvent on the charges is estimated prior to the MD simulation and assumed to be constant.

Free Energy Calculations

The Gibbs free energy of hydration for each molecule was obtained by alchemical free energy calculations conducted with standard protocols that keep errors due to sampling and free energy estimators relatively small. Each molecule was solvated in approximately 1000 water molecules and energy minimized with the GROMACS simulation package. Water molecules were described by the SPC/E model, which reproduces the liquid properties of water reasonably well while keeping the computational cost low. In case of the S-HI atomic charges, we also used the TIP3P water model, which was originally used to compute hydration free energies in the FreeSolv database (with AM1-BCC atomic charges), to test the dependence of the results on the water model.

After energy minimization, the whole system was equilibrated in the NPT ensemble at 298.15 K and 1 bar. The simulations were performed with the GROMACS 5.0.4 software package using a time step of 2 fs in combination with stochastic dynamics (τ = 2ps) and the Parrinello-Rahman pressure coupling (τ_p = 1ps) algorithm using the compressibility of water. The electrostatic interactions were calculated with the Particle-Mesh-Ewald method, a cut-off radius of 1.2 nm, a PME-order of 6 and a spacing of 0.1 nm. The van der Waals interactions were scaled to zero via a switching function, which switches the potential to zero between 1.0 and 1.2 nm. The neighbor list was updated every 10 steps with the verlet cutoff-scheme implemented in GROMACS 5.0.4 and its cut-off was set to 1.2 nm. All bonds were constrained with the LINCS algorithm of order 4 and the isotropic correction to the energy pressure due to missing van-der-Waals interactions was applied.

The Gibbs free energy of hydration was then calculated from the equilibrated solvated systems with alchemical molecular dynamics simulations, as described above, where the intermolecular interactions between each solute and the solvent were switched off through a λ parameter using soft core potentials with values of σ = 0.3 and α = 0.5 and p = 1 originally proposed by Beutler *et al.* and implemented in Gromacs 5.0.4 package. First the electrostatic interactions were decoupled using the values of the λ parameter [0.00, 0.25, 0.50, 0.75, 1.00] followed by the van-der-Waals interactions at values of [0.00, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95, 1.00]. The total simulation time per value of λ was 1 ns. The analysis of the free energy simulations was performed with different free energy estimators implemented in alchemical-analysis tool and the MBAR method was used for final free energy estimates; its results were in general consistent with those of other estimators.

Additionally, the Gibbs free energy of hydration was also calculated using a purely quantum mechanical approach together with the SMD solvation model and the ORCA package 3.0.2: the free energy difference of each molecule calculated in vacuum was considered together with the free energies of the implicit solvent in the approximation of rigid rotor and harmonic oscillator including zero-point energy correction assuming the standard state of 1 mol L^{-1} as in the database.

Polarization correction to the hydration free energy

The hydration of a molecule includes a polarization of the solute's electron density. When atomic charges used in free energy calculations are calculated from the molecular electron density of the solute in vacuum, this polarization contribution is neglected and one assumes that the electronic densities in gas phase and in solution are the same. By using atomic charges derived from electronic structure calculations with an implicit solvation model, one can model the enhanced interaction with the solvent due to polarization of the solute. But, in free energy calculation based on non polarizable force fields the energy required to polarize the solute electron density, $E_{\mathrm{pol,s}}$, has to be accounted for:

$$E_{\mathrm{pol,s}} = \left\langle \Psi_{solv} \middle| \hat{H}_{vac} \middle| \Psi_{solv} \right\rangle - \left\langle \Psi_{vac} \middle| \hat{H}_{vac} \middle| \Psi_{vac} \right\rangle \quad (2)$$

The first term is the expectation value of the Hamiltonian of the vacuum calculation for the wave function obtained with the implicit solvation model and the second term is the ground state energy of the vacuum Hamiltonian. Note that this approach polarizes the solute through a reaction field resulting from the polarized solvent represented by a dielectric neglecting specific polarization by solvent molecules due e.g. the formation of hydrogen bonds.

Instead of correcting the hydration free energy with the polarization energy, another way to account for the polarization is to modify the atomic charges as proposed in the implicitly polarized charge method IPolQ by Cerutti *et al.* This method is based on linear response theory and was proposed by Karamertzanis *et al.* who showed that under the approximation of linear electronic response, the sum of electrostatic and polarization interactions of a molecule with an external electric field is equal to the electrostatic interaction of a half-polarized molecule with the same field., This means that atomic multipoles in a non-polarizable force field should be based on a (distributed) multipole expansion of the average of the vacuum and solvated electron densities. In IPolQ, this principle is applied to just atomic monopoles, i.e. atomic charges. This greatly facilitates the computation of hydration free energies with standard force fields, without the need for a post hoc polarization correction.

We followed a simplified version of the IPolQ approach and used the average of Hirshfeld-I charges from a vacuum and an implicit solvent electron density. We call these atomic charges *semipolarized* since they do not correspond to the atomic charges derived from the polarized molecular electron density. The hydration free energies obtained with these atomic charges were not corrected with the polarization energy, since this is accounted for in the atomic charge model.

The overall IPolQ approach also includes a variety of other unique aspects we do not utilize here, such as empirical refinements (derived from experimental hydration free energies of amino acid side chain analogues) to van-der-Waals parameters. Another difference from our work is that IPolQ uses RESP charges. Also the reaction field produced by the solvent differs. The reaction field in the IPolQ method is obtained iteratively from molecular dynamics simulations with an explicit solvent, whereas our model uses a computationally less demanding implicit solvation model. It is worth noting that IPolQ is now the default method used to obtain atomic charges in the new AMBER ff14ipq and ff15ipq protein force fields.,

Error Analysis and Correction Model

The hydration free energies computed with different sets of atomic charges were compared to the experimental values and the deviations were analyzed as function of the chemical elements and functional groups present in each molecule. While a classification of solutes by functional group can already provide insight into the errors, it is troubled by the presence of multiple, possibly different, functional groups in many solutes. To better explain the deviations between theory and experiment, we also developed an error model based on the number of occurrences of certain features in a solute. Two sets of features were used for this purpose, one counting occurrences of functional groups with the Checkmol program and a simpler alternative that just counts the chemical elements present (as in a chemical formula). We only retained a subset of the most essential and prevalent functional groups reported by Checkmol. The feature and their occurrences summed over all molecules in the FreeSolv database, are shown in Table 1. Even though counting functional groups facilitates the chemical interpretation of the error model, it also has some limitations. Rare functional groups cannot be included because there is too little data to make statistically sound statements on their contribution to the error, yet such infrequent groups might cause large deviations. By just counting chemical elements, this difficulty is avoided but also some chemical information is discarded.

The error model assumes that the deviation between a theoretical prediction and the experimental reference for solute *j* is a sum of independent normally distributed errors:

$$G_{\text{calc},j} - G_{\text{exp},j} = \Delta_{\text{tot},j} = \Delta_{\text{FE},j} + \Delta_{\text{exp},j} + \Delta_{\text{general}} + \sum_{i} N_{i,j}\Delta_{\text{feat},i} \quad (3)$$

The first term is an uncertainty with known variance estimated from the free-energy calculation. The second term is the experimental error, whose standard deviation is given in FreeSolv or set to a default value of 0.6 kcal/mol in case the experimental error is not available. While the first two terms have a fixed variance and zero mean, all remaining terms have an unknown mean and variance, which will be fitted to the observed errors. Keep in mind that all terms in the right-hand-side of Eq. (3) are normally distributed stochastic quantities, characterized by a mean and variance. The mean corresponds to a mean (as in transferable or systematic) error while the variance represents a random (as in non-transferable or unexplainable) error. Δ_{general} is an error contribution uncorrelated to the molecular size or contained features, e.g. due to simulation settings or the solvent model. $\Delta_{\text{feat,i}}$ is an error associated with the presence of feature *i*, e.g. due to an error in the force field parameters related to this feature. $N_{i,j}$ is the number of times feature *i* is present in solute *j*. With this model, the total error is also normally distributed; however, the mean and variance depend on the features present:

$$\mu_{\text{tot},j} = \mu_{\text{general}} + \sum_i N_{i,j}\mu_{\text{feat},i} \quad (4)$$

$$\sigma^2_{\text{tot},j} = \sigma^2_{\text{FE},j} + \sigma^2_{\text{exp},j} + \sigma^2_{\text{general}} + \sum_i N_{i,j}\sigma^2_{\text{feat},i} \quad (5)$$

The likelihood to observe a particular difference between theory and experiment, for a given value of the mean and variance parameters, is a product of probability densities, which include one factor for each solute:

$$\mathscr{L} = \prod_{j=1}^{N_{\text{mol}}} \frac{1}{\sqrt{2\pi\sigma^2_{\text{tot},j}}}\exp\left(-\frac{\left(\Delta_{\text{tot},j} - \mu_{\text{tot},j}\right)^2}{2\sigma^2_{\text{tot},j}}\right) \quad (6)$$

where $\Delta_{\text{tot,j}}$ is the difference between the theoretical and experimental hydration free energy for molecule *j*. The unknown parameters are found by maximizing the likelihood function. In practice, this is done by a minimization of the following objective function.

$$-\ln\mathscr{L} = \frac{1}{2}\sum_{j=1}^{N_{\text{mol}}} \frac{\left(\Delta_{\text{tot},j} - \mu_{\text{tot},j}\right)^2}{\sigma^2_{\text{tot},j}} + \frac{1}{2}\sum_{j=1}^{N_{\text{mol}}} \ln\left(2\pi\sigma^2_{\text{tot},j}\right) \quad (7)$$

The first term is a typical least-squares cost function. The second term is needed because we also want to optimize the variance parameters. This objective function was implemented in Python and its analytic derivatives with respect to the unknowns were computed with AutoGrad. The minimization is carried out with the L-BFGS-B minimizer in SciPy, constraining all variance parameters to be positive. Due to the large number of parameters, in case of functional groups and in case of chemical elements compared to 613 data points, the minimization is ill-conditioned and different solutions can be found with small changes in the training set. The ill-conditioned minimum introduces an uncertainty in the parameters, which we estimated by performing 100 bootstrapping iterations. In each of these 100 iterations, the training set was randomly resampled with replacement to obtain a new similar training set of the same size to which the parameters were fitted. No further perturbations were added to the resampled training sets to mimic experimental uncertainty because this error is already accounted for in the error model.

The error model can be used to identify sources of systematic and random errors in the free energy calculations, by analyzing the optimized mean and variance parameters, which is our primary interest in this work. It can also be used to predict systematic and random errors for any new solute that is similar to those in the training set, using the following expressions:

$$\bar{\mu}_{\mathrm{tot},\,j} = \bar{\mu}_{\mathrm{general}} + \sum_i N_{i,\,j}\bar{\mu}_{\mathrm{feat},\,i} \quad (8)$$

$$\bar{\sigma}^2_{\mathrm{tot},\,j} = \sigma^2_{\mathrm{FE},\,j} + \sigma^2_{\mathrm{exp},\,j} + \bar{\sigma}^2_{\mathrm{general}} + \sum_i N_{i,\,j}\bar{\sigma}^2_{\mathrm{feat},\,i} \quad (9)$$

with

$$\bar{\mu}_{\mathrm{general}} = \mathrm{E}\left[\mu_{\mathrm{general}}\right] \quad (10)$$

$$\bar{\mu}_{\mathrm{feat},\,i} = \mathrm{E}\left[\mu_{\mathrm{feat},\,i}\right] \quad (11)$$

$$\bar{\sigma}^2_{\mathrm{general}} = \mathrm{E}\left[\sigma^2_{\mathrm{general}}\right] + \mathrm{VAR}\left[\mu_{\mathrm{general}}\right] \quad (12)$$

$$\bar{\sigma}^2_{\mathrm{feat},\,i} = \mathrm{E}\left[\sigma^2_{\mathrm{feat},\,i}\right] + \mathrm{VAR}\left[\mu_{\mathrm{feat},\,i}\right] \quad (13)$$

where averages (E) and variances (VAR) are computed over all bootstrap iterations. The predictive variance parameters, $\bar{\sigma}^2_{\mathrm{general}}$ and $\bar{\sigma}^2_{\mathrm{feat},i}$, have two contributions: the uncertainty estimate by the likelihood maximization and the uncertainty on the most-likely average estimated with the bootstrapping method.

In addition to fitting the error model to the full FreeSolv database, we also partitioned the database into stratified training and test sets to test the predictive power of the error model and the robustness of the parameters.

Results

The capability of the different atomic charges in combination with the other parameters of the GAFF force field to reproduce the experimental hydration free energies of the FreeSolv database was addressed by state of the art free energy calculations and molecular dynamics simulations in explicit solvent. Several factors were examined, including the polarizing effect of the solvent on the electron density in the electronic structure calculations, the polarization correction to the free energy, the influence of the electronic structure method on the atomic charges and the derived hydration free energies, and finally also the influence of the water model on the free energy calculations. Hydration free energies obtained with charges from the Hirshfeld-I and Minimal Basis Iterative Stockholder (MBIS) partitioning methods (see Methods section) were compared to values obtained with the AM1-BCC atomic charges. For comparison hydration free energies were also obtained from the electronic structure calculations with the SMD solvation model.

The deviations from the experimental values for each charge set were rationalized through statistical analysis for several chemical elements and functional groups in the molecules. The obtained parameters of the statistical model may serve in the future to anticipate errors in calculated hydration free energies with the S-MBIS atomic charges if the new molecules share the same functional groups or chemical elements. In the context of force field development the statistical model will identify systematic errors associated with the description of the intermolecular interactions of specific functional groups or atoms.

Hydration free energies using different atomic charge derivation methods

The atomic charges used to calculate hydration free energies stem from different partitioning methods in which the polarization of the electron density was accounted for by the SMD solvation model. Final free energies were corrected with the energy required to polarize the electron density. Only in the case of the semipolarized atomic charges based on the linear response approximation and the AM1-BCC method no polarization correction was included because this is neglected in the latter or accounted for in the approximation in the former.

We first wanted to see the effect of the polarization on the hydration free energies and calculated Hirshfeld-I atomic charges from the molecular electron density in vacuum. Figure 1 shows the parity plot between the calculated hydration free energies and the experimental reference values in the FreeSolv database. The comparison shows a poor correlation and a large value for the root mean square error (RMSE) of 5.0 kcal mol^{-1}.

Figure 2 shows parity plots of all hydration free energies calculated with atomic charges which account for the polarization of the solute and the AM1-BCC charges against the reported experimental value in the FreeSolv database. In Figure 2A the same Hirshfeld-I paritioning method was applied to polarized molecular electron densities from electronic structure calculations with the SMD solvation model (BLYP/def2-TZVP) (S-HI atomic charges) and corrected by the polarization energy (for sulfates and phosphates the Hirshfeld-E method was used and they are shown as circles). By comparison to Figure 1 on concludes that the effect of the electron density polarization through the solvent prior to the partitioning into atomic charges is essential to improve the calculated hydration free energies. Adding the implicit solvation model in the electronic structure calculation polarizes the electron density and leads to larger absolute values of the atomic charges and more negative free energies, which after the polarization correction reveal better agreement with the experimental values as shown in Figure 2A (RMSE=2.9 kcal mol^{-1}). The polarization of the electron density and the polarization correction energy become more important for polar molecules, which also have the most negative hydration free energies and the strongest electrostatic interactions with the solvent. Despite the success of using charges derived from polarized densities and the polarization correction, most of the computed hydration free energies (Figure 2A) still overestimate the experimental value (average error = 2.2 kcal mol^{-1}). This systematic error will be analyzed in detail in the following paragraphs.

To test the sensitivity of Hirshfeld-I atomic charges to the employed electronic structure method, we also computed electron densities with the computationally more demanding B3LYP/def2-TZVP method and the SMD implicit solvent model. With the resulting atomic charges, hydration free energies were recalculated for the whole database applying the respective polarization corrections. The change in the electronic structure method has only a small effect for molecules with hydration free energies above −10 kcal mol^{-1} and for more polar molecules it provides more negative values (see Figure S1 in Supporting Information). This is related to the self-interaction error of semi-local exchange-correlation functionals like BLYP, which results in more homogeneous electron densities. This error has been shown to provide smaller dipole moments, which is also the case in this work. Furthermore, we observe that atomic charges derived from BLYP densities are smaller compared to their B3LYP counterparts. However, the absolute unsigned deviation of the calculated hydration free energies (B3LYP method) with respect to ones obtained with the BLYP atomic charges is 0.7 kcal mol^{-1} with the largest deviations in the range of 2–3 kcal mol^{-1} (RMSE = 3.0 kcal mol^{-1} with respect to experimental reference values). For the molecules containing sulfates and phosphates, which atomic charges were calculated with the alternative Hirshfeld-E method, a larger dependence on the electronic structure methods was observed. Considering the computational cost of the B3LYP functional (3 times larger than BLYP) and the increased RMSE value with respect to the experimental reference data, BLYP in combination with the employed basis set presents an attractive alternative for the calculation of a large number of molecules as performed in this study.

For all calculations the SPC/E water model was considered due to its better reproduction of liquid water properties and low computational cost related to only three interaction sites. Because hydration free energies also depend on the water model used in the molecular dynamics simulations, we tested the effect of changing the SPC/E water model to the TIP3P model, which was used previously in combination with the AM1-BCC charges in the FreeSolv database. In previous work, the TIP3P model was selected because it is often recommended for the AMBER and GAFF force fields. Our comparison of the hydration free energies with the two water models reveals a root mean square deviation of 0.56 kcal mol^{-1} (RMSE(TIP3P model) = 2.6 kcal mol^{-1} with respect to experimental reference values). Therefore, the SPC/E model was considered for the rest of the study due to its better reproduction of the liquid water properties at the same computational cost.

Figure 2B shows the results for semipolarized Hirshfeld-I charges, which is an alternative method based on linear response theory to account for polarization of the solute and similar to the implicitly polarized charge method (IPolQ) by Cerutti *et al.* (see Methods section). Instead of correcting the hydration free energy with a polarization energy, here the atomic charges are altered to account for polarization. Semipolarized atomic charges lead to relatively large errors and in most cases the hydration free energy is overestimated, which is reflected in the average error and other statistical descriptors shown in the inset. This is not surprising because the semipolarized charges rely on two main approximations: (i) only linear response effects are included and (ii) the distributed multipole expansion of induced density fluctuations is truncated after the atomic monopoles. In contrast, the polarization correction in Eq. (2) makes fewer approximations: non-linear effects are included and there is no truncation of the distributed multipole expansion.

Finally, Figure 2C shows hydration free energies calculated with atomic charges obtained from the MBIS partitioning method employing the BLYP/def2-TZVP electronic structure method and SMD solvation model and the post-hoc polarization correction of the free energies. When the hydration free energies of this charge set are compared to the Hirshfeld-I partitioning method in Figure 2A, an overall better correlation with the experimental values is observed, as evidenced by the statistical descriptors in the inset. A detailed error analysis for each charge set will be provided in the last section. For comparison, we added the results obtained with the AM1-BCC charges present in the FreeSolv database in Figure 2D.

In summary, of the three new charge sets analyzed (S-HI, semipolarized and S-MBIS) the S-MBIS charges clearly reproduce the hydration free energies best with a performance comparable to the AM1-BCC charges. As shown in Figure 1, atomic charges computed in vacuum without polarization by the solvent lead to a very poor prediction of the hydration free energy. Semipolarized charges yield better results, but the improvement is limited. The use of Hirshfeld-I or MBIS charges from a polarized electron density and adding the polarization correction improves the correlation with experimental values considerably. Comparing the statistical descriptors of the S-MBIS charges to the AM1-BCC atomic charges we conclude that S-MBIS atomic charges present smaller average errors (0.29 vs 0.79 kcal mol^{-1}) with slightly larger values for the root mean square error (RMSE) and the absolute unsigned error (AUE). The difference in both Kendall τ and Pearson R correlation coefficients are small and lie almost in the error margin.

These statistical descriptors in Figure 2 are often used to validate the performance of a certain model; however, it is also interesting to analyze the number of molecules whose calculated hydration free energies are within the range of the experimental error. For the S-MBIS atomic charges we found that this amounts to 23 % which equals the fraction of molecules for the AM1-BCC charges. Taking the other extreme, we then identified the molecules that had an error of more than five times the experimental error and displayed their structure in Figure 3. A comparison of the different molecules shows that most of them present phosphor or sulfur atoms and more than one functional group. Assuming that the functional groups contribute independently to the total absolute error, it is to be expected that solutes with many functional groups have a larger error. A detailed analysis of the error per chemical element and functional group to rationalize these large deviations will be provided in the last section.

To analyze systematic deviations of the calculated hydration free energies we show in Figure 4 the distribution of absolute errors ($\Delta G_{calc} — \Delta G_{exp}$) for each atomic charge set. Both atomic charge sets employing the Hirshfeld-I charges (S-HI and semipolarized) present a broad distribution with systematic positive errors resulting from too positive calculated hydration free energies. This points to a deficiency of the Hirshfeld-I method and its derived atomic charges when used with force fields for free energy calculations. For the S-MBIS atomic charges a fast decaying normal distribution (D ‘Agostino and Pearson’s normality test p value = 3.76×10^{-5}) is observed with a maximum slightly shifted from zero. For the AM1-BCC atomic charges also a normal distribution is observed ($p = 2.15 \times 10^{-7}$), which might originate from the parametrization of bond charge corrections added to AM1 atomic charges to match experimental hydration free energies mentioned in the original AM1-BCC reference (pages 1637–1638). The slightly larger shift of the maximum is mostly related to the SPC/E water model, as comparison with the calculated hydration free energies in the FreeSolv database using the TIP3P model reveals.

Finally, to see how well the new atomic charges perform, we compare our results with the MBIS partitioning including polarization correction directly to the previously published results (Figure 5) which used the AM1-BCC charges together with the TIP3P water model. The comparison shows that the S-MBIS atomic charges provide some improvements for the molecules with the most negative hydration free energies, which possess more than one hydroxyl group. For some of the more polar molecules the S-MBIS results underestimate compared to experiment, while AM1-BCC results were rather good. The S-MBIS hydration free energies of phosphor and sulfur containing molecules, with experimental values between −10 and −5 kcal mol^{-1}, are considerably underestimated and will be discussed in the last section. If this group of molecules is removed from the data, the RMSE is 1.8 for the SMBIS atomic charges compared to 1.3 kcal mol^{-1} for the AM1-BCC ones in the FreeSolv database. When we calculate the hydration free energy with the same AM1-BCC atomic charges but SPC/E water model the RMSE increases to 1.7 kcal mol^{-1} (see Figure 2). One should put the good performance of the AM1-BCC charges in perspective: some of the bond-charge corrections (BCCs) in AM1-BCC were tuned to improve the reproduction of hydration free energies, while such empirical corrections are not present in the S-MBIS approach.

We also calculated the hydration free energies using the SMD solvation model that is used to obtain the polarized molecular electron density directly from the electronic structure calculations under the rigid rotor harmonic oscillator approximation including the zero-point energy correction using geometries optimized both in vacuum and with the solvation model (see Methods section). The obtained hydration free energies are shown in a parity plot in Figure S4 of the Supporting Information. The calculated hydration free energies present statistical descriptors which lie in between the values of the S-MBIS and the AM1-BCC atomic charges (Average error = 0.12, RMSE = 1.84, AUE 1.34 kcal mol^{-1}) but a considerably worse value of Pearson R (R=0.88) and Kendall τ (τ=0.69) in addition to some outliers which structure is provided in Figure S5 of the Supporting Information. The good performance of the SMD model may be biased: most of the parameters for the non-electrostatic contribution to the hydration free energy in the SMD solvation model were parameterized using experimental hydration free energies for similar or the same compounds. The molecules for which we observe a significant deviation present nitro functional groups, various halogen atoms or phosphate or thiophosphate esters (see Figure S6 Supporting Information). A possible explanation could be that the identified outliers were not present or underrepresented in the SMD training set.

Molecular Dipole Moments in Solution Obtained from the Different Atomic Charge Sets

Our primary validation of our partial atomic charges relies on hydration free energies, but we also examined how well our charge sets reproduce molecular dipole moments. To a first approximation, the atomic charges should reproduce the magnitude of the static molecular dipole moment of the solute in the minimum structure in their molecular environment. Since the atomic charges will be used in condensed phase simulations and experimental data on the molecular dipole moment in aqueous solution is not available, we used the magnitude of the molecular dipole moment of the solute in aqueous solution calculated directly from the molecular electron density at the BLYP/def2-TZVP level with the implicit SMD model as our reference. In principle, the S-HI and S-MBIS charges are derived from the same electron density and therefore should reproduce the molecular dipole moment. As shown in Figure 6A, S-HI charges show a good correlation for small values but a systematic underestimation is observed for larger values of the molecular dipole moment as evidenced by the negative value of the average error. The S-MBIS atomic charges present a much better correlation as shown in Figure 6C with some outliers whose Lewis structure is shown in Figure 7. The outliers are large molecular structures with various functional groups containing chlorine, sulfur and phosphor atoms. In the Hirshfeld-I and MBIS partitioning methods, point charges only describe the leading monopole term in the atomic multipole expansion and atomic dipole contributions to the molecular dipole moment are neglected. One possible explanation for the deviation of larger molecules is the increased number of atoms whose dipole moment is neglected and may therefore exhibit larger error in Figures 6A and 6C. This already explains why the outliers are large molecules. The outliers also contain functional groups for which large atomic dipoles should be expected, e.g. due to the σ–hole in heavier halogens or due to polar groups such as sulphates and phosphates. In principle, off-center charges or atom-centered multipoles can improve the representation of electron density around the atom but they result in a larger computational cost of the force field.

Semipolarized Hirshfeld-I charges are not intended to reproduce dipole moments of solvated molecules. Instead they contain deviations from the equilibrium charge distribution to account for the polarization energy. This is clearly visible in Figure 6B and from the statistical descriptors in the inset. These atomic charges systematically underestimate the dipole moment, which is expected since only half of the difference to the fully polarized atomic charges is added. Interestingly, in some cases very large deviations in the dipole moments are observed (molecules 1, 2, and 3 in Figure 7B). For molecule 1 and 2 the partitioning method and the limitations of atomic charges discussed above may contribute considerably to the observed deviation since these two outliers are also present in Figure 6A of the S-HI atomic charges.

Finally, in Figure 6D, we also compare dipole moments of the AM1-BCC charges with those obtained from electronic structure calculations with implicit solvation. This may seem to be an unfair comparison because the AM1-BCC charges are derived from the AM1 Hamiltonian in vacuum. However, the bond charge corrections are added to account for the polarization of the molecules in water and to reproduce the values obtained with the RESP method at the HF/6–31G(d) level of theory. Some additional corrections were then introduced to reproduce the hydration free energies. As evidenced from the statistical descriptors in Figure 6D, these atomic charges show a systematic underestimation of the dipole moments with some pronounced outliers and larger scatter. The poor correlation might be related to bond-charge corrections (BCCs), which were directly fitted to reproduce hydration free energies without ensuring a reasonable dipole moment.

One has to note at this point that the reference molecular dipole moment obtained with the BLYP functional is known to produce small dipole moments in vacuum. This deficiency of the BLYP functional is partially corrected for by the computationally more demanding B3LYP functional. Calculating the dipole moments with the B3LYP functional would increase the magnitude of the reference values resulting in even larger deviations for the AM1-BCC charges. However, for the other three atomic charge sets (S-HI, semipolarized and S-MBIS) the conclusions will be similar since their atomic charges would be derived from the more polarized B3LYP molecular electron density leading to larger dipole moments.

In summary, one can conclude that the magnitude of the static molecular dipole moments of the minimum solute structure calculated with electronic structure calculations and the SMD solvation model are best reproduced by the S-MBIS charges followed by the S-HI charges with some systematic deviations. Semipolarized Hirshfeld-I charges in some cases and the AM1-BCC charges have particularly poor agreement with ab-initio molecular dipole moments. The obtained conclusions for the MBIS and Hirshfeld-I partitioning method agree well with earlier results using the MBIS charges in vacuum by Verstraelen et al. In fairness, neither AM1-BCC nor semipolarized Hirshfeld-I charges should necessarily reproduce dipole moments particularly well, since they are not designed to do so and might predict hydration free energies reasonably well without yielding accurate dipole moments.

Hydration free energies for varying functional groups

The analysis so far took the whole database into account without examining trends with respect to specific functional groups which are key in driving solvation behavior, so we divided the dataset into functional group categories and analyzed subsets. Some molecules present more than one kind of functional group and the assignment is not unambiguous. For this analysis, only the S-MBIS atomic charge set was considered since this performs best; for the other atomic charge sets the results are summarized in Figure S6-S8 in the Supporting Information.

Figure 8 shows the calculated hydration free energies for molecules containing at least one carbonyl, amine, ether or one hydroxyl group and the aliphatics, aromatics, and hetero-cycles in the top three parity plots. At the bottom the nitro and carbonitrile groups are shown followed by halogen-containing compounds and molecules with sulfur or phosphor atoms.

For ketones and aldehydes, most of the hydration free energies have errors with magnitudes less than 2 kcal mol^{-1}. The identified outliers 1,2, and 3 (see Figure 9) present more than one functional group, or amine groups as in molecule 1. For molecule 3, a competition between an intramolecular hydrogen bond and a hydrogen bond to the surrounding water molecules is expected. The implicit solvation model used in this study to derive the atomic charges does not account for this competition. Therefore, an approach which includes the explicit description of the water molecules in the derivation of the atomic charges and a proper sampling of the relevant conformations would be needed. Recently one of us proposed a method, which replaces the atomic charges of the solute during molecular dynamics simulations by charges derived from QM/MM calculations in explicit solvent after a fixed number of steps until a constant value of the atomic charges is reached., We plan to use this method in a future study on molecules with intramolecular hydrogen bonds and larger conformational flexibility to test, if better agreement with the experimental hydration free energies can be reached.

As shown in Figure 8 for carboxylic acid esters and amides systematically more negative hydration free energies are obtained, which explains outlier 2 where the error of the three ester functional groups might sum up. For the alcohols, the largest deviations are observed when more than three functional groups are present (molecules 4, 5 and 6), which suggests that each functional group contributes to the error, resulting in a large total deviation. Also for these molecules, the competition between intra- and intermolecular hydrogen bonds, plays a major role and may alter the stability of the various different conformations in aqueous solution. In the amines, molecules 7 and 8, which are tertiary amines, present the largest deviation together with molecule 9, which also contains a nitro group bound to a benzene ring. This nitro group together with the amine group in para position forms a conjugated π electron system, which might not be well described with atom-centered fixed atomic charges. The same molecule was also identified to be problematic using the AM1-BCC atomic charges. In the amine group, molecule 10 also possesses a large deviation from the experimental value probably due to the addition of the errors associated with the tertiary amine, the amide group and the heterocycle with two nitrogen atoms, which provides an additional error to the free energy as will be shown below. The hydration free energies of compounds containing ether functional groups are all well reproduced.

Calculated hydration free energies of aliphatics and aromatics lie mostly within the 2 kcal mol^{-1} error band with outlier 13 being a molecule with an urea group, which may contribute to the deviation due to the two nitrogen atoms. In the heterocycles, pronounced deviations are observed in the molecules 11, 12 and 14, which contain more than one nitrogen atom in the heterocyle. For molecules containing a nitro or carbonitrile functional group, S-MBIS atomic charges provide accurate free energies — more accurate than the values reported with AM1-BCC charges. Halogen-containing molecules are closer to the experimental values for fluorine, with outlier 16 being a compound with two amine groups, which might contribute to the absolute error. Chlorine containing molecules in general present larger deviations, which might also stem from other functional groups and the presence of multiple chlorine atoms in the molecule. Among the bromine-containing compounds, the most pronounced outlier is 15, which also has a heterocycle with two nitrogens, known to cause large errors. In the last parity plot, we show the calculated hydration free energies for the sulfur and phosphor containing compounds. The correlation with experimental hydration free energies is better for sulfur containing compounds if the hetero-atom is bonded to carbon atoms. The largest deviation corresponds to molecules with sulfonyl or sulfonate and phosphonate groups. In these groups, atomic partial charges are relatively large in absolute value due to the high oxidation state of the sulfur or phosphor atom. With such pronounced charge distributions, electrostatic interactions between atomic charges and atomic dipoles also become important, which are neglected in any force-field using a point-charge model. It is therefore not surprising that these pronounced charge distribution systematically lead to larger errors on the hydration free energy. One may bias charges in these groups (or the van der Waals parameters) to compensate for the absence of atomic dipoles in the force field but it is not clear how this can be done systematically.

This general analysis already qualitatively identifies some functional groups with larger deviations, such as sulfonates, phosphonates, amines and heterocycles with two nitrogen atoms or amines. However, a quantitative analysis is not possible with the classification by functional group, since one molecule might also present multiple groups contributing to the absolute error. To quantify the error due to each functional group or even due to each element in a solute, we developed a statistical model. This model counts features (functional groups or elements) in each solute and it assumes the total error between computation and experiment is a sum of independent normal errors due to (i) features present in the solute, (ii) an uncertainty due to general model errors (e.g. in the solvent), (iii) the fixed experimental measurement error and (iv) the statistical error from the free energy calculation. The mean and variance of the uncertainty associated with each feature and the general error were fitted to the differences between the S-MBIS and experimental hydration free energies.

The optimal parameters (averaged over 100 bootstrapping iterations) are given in Tables 2 and 3, using functional groups and chemical elements as feature sets, respectively. Table 4 shows some key statistics of the error between S-MBIS predictions and experimental hydration free energies at three stages: (i) prior to applying any correction, (ii) after correcting for systematic errors with Eq. (8) and (iii) after dividing by the predicted random error with Eq. (9). After correcting for systematic errors, the mean error almost vanishes and the standard deviation is reduced, which means the error model captures a part of the systematic error made by the free energy simulations. Even though the error model is primarily meant to estimate the risk that a free energy simulation is unreliable, it can also be useful as a simple correction. After dividing deviations between corrected predictions and experiment by the predicted error for each molecule, one obtains normalized residuals. The normalized residuals closely resemble samples from a standard normal distribution (mean zero and standard deviation one). This is further confirmed by QQ plots in Figure 10, which show the deviation between real errors and the errors one would get if they were distributed normally with the same variance and mean. The uncorrected error shows visible deviations from normality, while the normalized residuals are closer to a normal distribution, especially when chemical elements are used as features. This confirms a key assumption of the error model, namely that the errors between theory and experiment are normally distributed, yet with a different mean and variance for each molecule. There are only a few outliers that cannot be explained well by the error model, which are visible as deviations from the parity line in Figures 10b and 10c (representative structures are shown in Figure S10 of the Supporting Information). However, for most molecules, the error estimate, both $\bar{\mu}_{\text{tot},j}$ and $\bar{\sigma}_{\text{tot},j}$, are consistent with the data. For any new molecule, similar to the ones in FreeSolv, these estimates will give *a priori* a useful judgement of the reliability of the hydration free energy calculation.

Some systematic errors in Tables 2 and 3 are relatively big but can compensate each other to large extent when multiple functional groups are present in one molecule. This is clear in Figure 11, which contains plots of the predicted systematic and random error for each molecule, see Eqs. (8) and (9), against the actual residual. The systematic error underestimates the residual, especially when the residual is large. Due to this compensation of systematic error contributions, we should be careful with their interpretation. Nevertheless, it is clear that the hydration free energy of phosphorus-containing molecules is systematically underestimated, hinting at too strong electrostatic or van der Waals interactions with the solvent in the force-field model. A similar problem is also observed for esters and amides and to lesser extent also for ketones, aldehydes, nitro groups and nitriles. Secondary and tertiary amines cause large positive systematic errors. The random errors cannot cancel each other out and are therefore easier to interpret. It is clear that predictions for solutes with sulfur, heterocycles or phosphor are the most problematic, followed by tertiary amines, amides and secondary amines. The error model based on elemental features sketches a similar picture, the presence of nitrogen, sulfur, oxygen and phosphor may lead to large errors. This analysis is consistent with the analysis of outliers in Figure 8 but it provides more detailed insight into functional groups whose force-field parameters may need to be refined.

To verify that all results for the error model (tables 2, 3 and 4) are not affected by overfitting, we have repeated the parameter estimation (for both feature sets) with a training set containing only 75% of the data and then tested the performance of the model on the remaining 25%. This 75% training set was constructed such that all features are sufficiently present. For example, if no sulfurs would be present in the 75% training set, one should not expect the error model to work at all for sulfur-containing molecules. The parameters are shown in Tables S1 and S2 in the Supporting information and they exhibit only minor deviations from the parameters fitted to all the data: in total, only 11 out of 60 parameters differ by more than 0.2 kcal/mol. The performance statistics for the training and test sets, comparable to Table 4, are shown in Table S3 in the Supporting information. These results show that the error model performs almost equally on the training and the test sets, especially when using chemical elements as features. One can therefore conclude that the parameters fitted to the complete data in tables 2 and 3 are capable of predicting systematic and random errors for new molecules that are similar to the ones in FreeSolv.

Conclusions and outlook

Hydration free energies were calculated for different sets of atomic charges in combination with the GAFF force field parameters. Accurate hydration free energies were obtained (i) if charges were derived from BLYP electron densities polarized with the SMD implicit solvent model and (ii) if the energy needed to polarize the BLYP density is included in the free energy cycle. Predictions obtained with semipolarized charges, a method to account for polarization effects in a non-polarizable force field, were not satisfactory. Of the two new charge sets considered, both based on atoms-in-molecules methods, the MBIS set is found to result in hydration free energies closest to the experimental reference values with a comparable performance as the AM1-BCC atomic charges. A statistical error model identified phosphorus-containing compounds with the largest deviation together with molecules containing functional groups as ester, amides and amines.

The error model can also be used in the future to anticipate errors in calculated hydration free energies of new molecules, if they contain only functional groups or chemical elements that are prevalent in the FreeSolv database. Due to its simple structure, the error model does not account for correlations between two features in a molecule, but is nevertheless surprisingly successful. This means simple improvements by refining parameters of individual atom types should be feasible. Polar groups, such as sulfonates and phosphates, are obvious candidates for future improvement of the electrostatic term in the force field. Some of the observed systematic errors may also stem from the GAFF Lennard-Jones parameters.

For some molecules the competition of intra-molecular and inter-molecular solute solvent hydrogen bonds becomes important, which is poorly described in implicit solvent models, used in this work. We will address this in a future study by including an explicit description of the solvent in the charge derivation method. Finally, in this work we only tested new methods to assign atomic charges. Obviously, refined Lennard-Jones parameters could further improve the reproduction of experimental hydration free energies. The GAFF Lennard-Jones parameters were originally taken from the Amber99 force field. In Amber99, these parameters were either adjusted to reproduce liquid properties of alkanes in combination with RESP charges or were taken from the OPLS force field with some adjustments. Given their historical origin, we expect the Lennard-Jones parameters, and then also torsional parameters, can still be refined in combination with MBIS charges in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors acknowledge financial support provided by Fondecyt No. 1160197. MR thanks CONICYT for the graduate scholarship No. 21140918. DLM appreciates financial support from the National Institutes of Health (1R01GM108889–01), the National Science Foundation (CHE 1352608). T.V. acknowledges the Foundation of Scientific Research—Flanders (FWO) and the Research Board of Ghent University (BOF) for their financial support.

References

(1). Shirts MR; Pande VS Solvation Free Energies of Amino Acid Side Chain Analogs for Common Molecular Mechanics Water Models. J. Chem. Phys 2005, 122, 134508. [PubMed: 15847482]

(2). Hess B; van der Vegt NFA Hydration Thermodynamic Properties of Amino Acid Analogues: A Systematic Comparison of Biomolecular Force Fields and Water Models. J. Phys. Chem. B 2006, 110, 17616–17626. [PubMed: 16942107]

(3). Kollman P Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. Chem. Rev 1993, 93, 2395–2417.

(4). Mobley DL; Bayly CI; Cooper MD; Shirts MR; Dill KA Small Molecule Hydration Free Energies in Explicit Solvent: An Extensive Test of Fixed-Charge Atomistic Simulations. J. Chem. Theory Comput 2009, 5, 350–358. [PubMed: 20150953]

(5). Mobley DL; Liu S; Cerutti DS; Swope WC; Rice JE Alchemical Prediction of Hydration Free Energies for SAMPL. J. Comput. Aided Mol. Des 2012, 26, 551–562. [PubMed: 22198475]

(6). Mobley DL; Guthrie JP FreeSolv: A Database of Experimental and Calculated Hydration Free Energies, With Input Files. J. Comput. Aided Mol. Des 2014, 28, 711–720. [PubMed: 24928188]

(7). Konig G; Boresch S Hydration Free Energies of Amino Acids: Why Side Chain Analog Data Are Not Enough. J. Phys. Chem. B 2009, 113, 8967–8974. [PubMed: 19507836]

(8). Shirts MR; Bair E; Hooker G; Pande VS Equilibrium Free Energies from Nonequilibrium Measurements Using Maximum-Likelihood Methods. Phys. Rev. Lett 2003, 91, 140601. [PubMed: 14611511]

(9). Ben-Naim AY Solvation Thermodynamics; Springer Science & Business Media, 2013.

(10). Wolf MG; Groenhof G Evaluating Nonpolarizable Nucleic Acid Force Fields: a Systematic Comparison of the Nucleobases Hydration Free Energies And Chloroform-to-Water Partition Coefficients. J. Comput. Chem 2012, 33, 2225–2232. [PubMed: 22782700]

(11). Takahashi H; Omi A; Morita A; Matubayasi N Simple and Exact Approach to the Electronic Polarization Effect on the Solvation Free Energy: Formulation for Quantum-Mechanical/ Molecular-Mechanical System and its Applications to Aqueous Solutions. J. Chem. Phys 2012, 136, 214503–214503. [PubMed: 22697554]

(12). Shivakumar D; Deng Y; Roux B Computations of Absolute Solvation Free Energies of Small Molecules Using Explicit and Implicit Solvent Model. J. Chem. Theory Comput 2009, 5, 919–930. [PubMed: 26609601]

(13). Mobley DL; Wymer KL; Lim NM; Guthrie JP Blind Prediction of Solvation Free Energies from the SAMPL4 Challenge. J. Comput. Aided Mol. Des 2014, 28, 135–150. [PubMed: 24615156]

(14). Wang J; Wolf RM; Caldwell JW; Kollman PA; Case DA Development and Testing of a General Amber Force Field. J. Comput. Chem 2004, 25, 1157–1174. [PubMed: 15116359]

(15). Jakalian A; Jack DB; Bayly CI Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. J. Comput. Chem 2002, 23, 1623–1641. [PubMed: 12395429]

(16). Jakalian A; Bush BL; Jack DB; Bayly CI Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: I. Method. J. Comput. Chem 2000, 21, 132–146.

(17). Duarte Ramos Matos G; Kyu DY; Loeffler HH; Chodera JD; Shirts MR; Mobley DL Approaches for Calculating Solvation Free Energies and Enthalpies Demonstrated with an Update of the FreeSolv Database. J. Chem. Eng. Data 2017, 62, 1559–1569. [PubMed: 29056756]

(18). Jorgensen W; Maxwell D; TiradoRives 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.

(19). Oostenbrink C; Villa A; Mark AE; 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. Comp. Chem 2004, 25, 1656–1676. [PubMed: 15264259]

(20). Mackerell AD; Bashford D; Bellott M; Dunbrack RL; Evanseck JD; Field MJ; Fischer S; Gao J; Guo H; Ha S; Joseph-McCarthy D; Kuchnir L; Kuczera K; Lau FT; Mattos C; Michnick S; Ngo T; Nguyen DT; Prod- hom B; Reiher WE; Roux B; Schlenkrich M; Smith JC; Stote R; Straub J; Watanabe M; Wiorkiewicz-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. [PubMed: 24889800]

(21). Wang J; Cieplak P; Kollman PA How well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules? J. Comput. Chem 2000, 21, 1049–1074.

(22). Bayly CI; Cieplak P; Cornell W; Kollman PA A well-Behaved Electrostatic Potential Based Method Using Charge Restraints for Deriving Atomic Charges: the RESP Model. J. Phys. Chem 1993, 97, 10269–10280.

(23). Ponder JW; Wu C; Ren P; Pande VS; Chodera JD; Schnieders MJ; Haque I; Mobley DL; Lambrecht DS; DiStasio RA; Head-Gordon M; Clark GNI; Johnson ME; Head-Gordon T Current Status of the AMOEBA Polarizable Force Field. J. Phys. Chem. B 2010, 114, 2549–2564. [PubMed: 20136072]

(24). Savelyev A; MacKerell AD All-Atom Polarizable Force Field for DNA Based on the Classical Drude Oscillator Model. J. Comput. Chem 2014, 35, 1219–1239. [PubMed: 24752978]

(25). Vandenbrande S; Waroquier M; Van Speybroeck V; Verstraelen T The Monomer Electron Density Force Field (MEDFF): A Physically Inspired Model for Noncovalent Interactions. J. Chem. Theory Comput 2017, 13, 161–179. [PubMed: 27935712]

(26). Bultinck P; Van Alsenoy C; Ayers PW.; Carbo-Dorca, R. Critical Analysis and Extension of the Hirshfeld Atoms in Molecules. J. Chem. Phys 2007, 126, 144111. [PubMed: 17444705]

(27). Verstraelen T; Vandenbrande S; Heidar-Zadeh F; Vanduyfhuys L; Van Speybroeck V; Waroquier M; Ayers PW Minimal Basis Iterative Stockholder: Atoms in Molecules for Force-Field Development. J. Chem. Theory Comput 2016, 12, 3894–3912. [PubMed: 27385073]

(28). Marenich AV; Cramer CJ; Truhlar DG Universal Solvation Model Based on the Generalized Born Approximation with Asymmetric Descreening. J. Chem. Theory Comput 2009, 5, 2447–2464. [PubMed: 26616625]

(29). Van Damme S; Bultinck P; Fias S Electrostatic Potentials from Self-Consistent Hirshfeld Atomic Charges. J. Chem. Theory Comput 2009, 5, 334–340. [PubMed: 26610109]

(30). Verstraelen T; Pauwels E; De Proft F; Van Speybroeck V; Geerlings P; Waroquier M Assessment of Atomic Charge Models for Gas-Phase Computations on Polypeptides. J. Chem. Theory Comput 2012, 8, 661–676. [PubMed: 26596614]

(31). Heidar-Zadeh F; Ayers PW.; Verstraelen T; Vinogradov I; Vhringer-Martinez E; Bultinck P. Information-Theoretic Approaches to Atoms-in-Molecules: Hirshfeld Family of Partitioning Schemes. J. Phys. Chem. A 2018, 122, 4219–4245. [PubMed: 29148815]

(32). Manz TA; Sholl DS Chemically Meaningful Atomic Charges That Reproduce the Electrostatic Potential in Periodic and Nonperiodic Materials. J. Chem. Theory Comput 2010, 6, 2455–2468. [PubMed: 26613499]

(33). Gould T; Bucko TC 6 Coefficients and Dipole Polarizabilities for All Atoms and Many Ions in Rows 16 of the Periodic Table. J. Chem. Theory Comput 2016, 12, 3603–3613. [PubMed: 27304856]

(34). Guthrie JP A Blind Challenge for Computational Solvation Free Energies: Introduction and Overview. J. Phys. Chem. B 2009, 113, 4501–4507. [PubMed: 19338360]

(35). Geballe M; Skillman A; Nicholls A; Guthrie J; Taylor P The SAMPL2 Blind Prediction Challenge: Introduction and Overview. J. Comput. Aided Mol. Des 2010, 24, 259–279. [PubMed: 20455007]

(36). Geballe M; Guthrie J The SAMPL3 Blind Prediction Challenge: Transfer Energy Overview. J. Comput. Aided Mol. Des 2012, 26, 489–496. [PubMed: 22476552]

(37). Bannan CC; Burley KH; Chiu M; Shirts MR; Gilson MK; Mobley DL Blind Prediction of Cyclohexane-Water Distribution Coefficients from the SAMPL5 Challenge. J. Comput. Aided Mol. Des 2016, 30, 927–944. [PubMed: 27677750]

(38). Mobley D; Bannan CC; Rizzi A; Bayly CI; Chodera JD; Lim VT; Lim NM; Beauchamp KA; Shirts MR; Gilson MK; Eastman PK Open Force Field Consortium: Escaping atom types using direct chemical perception with SMIRNOFF v0.1. bioRxiv 2018,

(39). Neese F The ORCA Program System. Wiley Interdiscip. Rev. Comput. Mol. Sci 2011, 2, 73–78.

(40). Curutchet C; Cramer CJ; Truhlar DG; Ruiz-Lopez MF; Orozco M; Luque FJ Electrostatic Component of Solvation: Comparison of SCRF Continuum Models. J. Comput. Chem 2003, 24, 284–297. [PubMed: 12548720]

(41). Becke AD Density-Functional Exchange-Energy Approximation with Correct Asymptotic Behavior. Phys. Rev. A 1988, 38, 3098–3100.

(42). Lee C; Yang W.; Parr RG Development of the Colle-Salvetti Correlation-Energy Formula into a Functional of the Electron Density. Phys. Rev. B 1988, 37, 785–789.

(43). Miehlich B; Savin A; Stoll H; Preuss H Results Obtained W ith the Correlation Energy Density Functionals of Becke and Lee, Yang and Parr. Chem. Phys. Lett 1989, 157, 200–206.

(44). Becke AD Density-Functional Thermochemistry. III. The Role of Exact Exchange. J. Chem. Phys 1993, 98, 5648–5652.

(45). 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. [PubMed: 16240044]

(46). Weigend F Accurate Coulomb-Fitting Basis Sets for H to Rn. Phys. Chem. Chem. Phys 2006, 8, 1057. [PubMed: 16633586]

(47). Verstraelen T; Tecmer P; Heidar-Zadeh F; Boguslawski K; Chan M; Zhao Y; Kim TD; Vandenbrande S; Yang D; Gonzlez-Espinoza CE; Fias S; Limacher PA; Berrocal D; Malek A; Ayers PW HORTON 2.0.0 2015; http://theochem.github.com/horton/, Accessed: January 25, 2018.

(48). Hirshfeld FL Bonded-Atom Fragments for Describing Molecular Charge Densities. Theoret. Chim. Acta 1977, 44, 129–138.

(49). Davidson ER; Chakravorty S A Test of the Hirshfeld Definition of Atomic Charges and Moments. Theor. Chem. Acc 1992, 83, 319–330.

(50). Ayers P Atoms in Molecules, an Axiomatic Approach. I. Maximum Transferability. The J. Chem. Phys 2000, 113, 10886–10898.

(51). Verstraelen T; Ayers PW; Van Speybroeck V; Waroquier M The Conformational Sensitivity of Iterative Stockholder Partitioning Schemes. Chem. Phys. Lett 2012, 545, 138–143.

(52). Elking DM; Perera L; Pedersen LG HPAM: Hirshfeld partitioned atomic multipoles. Comput. Phys. Commun 2012, 183, 390–397. [PubMed: 22140274]

(53). Lillestolen TC; Wheatley RJ Redefining the Atom: Atomic Charge Densities Produced by an Iterative Stockholder Approach. Chem. Commun 2008, 0, 5909–5911.

(54). Verstraelen T; Ayers PW; Van Speybroeck V; Waroquier M Hirshfeld-E Partitioning: AIM Charges with an Improved Trade-off between Robustness and Accurate Electrostatics. J. Chem. Theory Comput 2013, 9, 2221–2225. [PubMed: 26583716]

(55). Spackman M The use of the promolecular charge density to approximate the penetration contribution to intermolecular electrostatic energies. Chem. Phys. Lett 2006, 418, 158–162.

(56). Wang B; Truhlar DG Including Charge Penetration Effects in Molecular Modeling. J. Chem. Theory Comput 2010, 6, 3330–3342. [PubMed: 26617087]

(57). Wang B; Truhlar DG Screened Electrostatic Interactions in Molecular Mechanics. J. Chem. Theory Comput 2014, 10, 4480–4487. [PubMed: 26588144]

(58). Wang Q; Rackers JA; He C; Qi R; Narth C; Lagardere L; Gresh N; Ponder JW; Piquemal JP; Ren P General Model for Treating Short-Range Electrostatic Penetration in a Molecular Mechanics Force Field. J. Chem. Theory Comput 2015, 11, 2609–2618. [PubMed: 26413036]

(59). Xie W; Orozco M; Truhlar DG; Gao J X-Pol Potential: An Electronic Structure-Based Force Field for Molecular Dynamics Simulation of a Solvated Protein in Water. J. Chem. Theory Comp 2009, 5, 459–467.

(60). Abraham MJ; Murtola T; Schulz R; Pall S; Smith JC; Hess B; Lindahl E GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism from Laptops to Supercomputers. SoftwareX 2015, 1–2, 19–25.

(61). Berendsen H; Grigera JR; Straatsma TP The Missing Term in Effective Pair Potentials. J. Phys. Chem 1987, 91, 6269–6271.

(62). Jorgensen WL; Chandrasekhar J; Madura JD; Impey RW; Klein M, Comparison of Simple Potential Functions for Simulating Liquid Water. J. Chem. Phys 1983, 79, 926–935.

(63). Van Gunsteren WF; Berendsen HJC A Leap-frog Algorithm for Stochastic Dynamics. Mol. Simul 1988, 1, 173–185.

(64). Parrinello M; Rahman A Polymorphic Transitions in Single Crystals: A New Molecular Dynamics Method. J. Appl. Phys 1981, 52, 7182–7190.

(65). Essmann U; Perera L; Berkowitz ML; Darden T; Lee H; Pedersen LG A Smooth Particle Mesh Ewald Method. J. Chem. Phys 1995, 103, 8577–8593.

(66). Hess B; Bekker H; Berendsen H; Fraaije J LINCS: A Linear Constraint Solver for Molecular Simulations. J. Comput. Chem 1997, 18, 1463–1472.

(67). Shirts MR; Mobley DL; Chodera JD; Pande VS Accurate and Efficient Corrections for Missing Dispersion Interactions in Molecular Simulations. J. Phys. Chem. B 2007, 111, 13052–13063. [PubMed: 17949030]

(68). Beutler TC; Mark AE; van Schaik RC; Gerber PR; van Gunsteren WF Avoiding Singularities and Numerical Instabilities in Free Energy Calculations Based on Molecular Simulations. Chem. Phys. Lett 1994, 222, 529–539.

(69). Klimovich PV; Shirts MR; Mobley DL Guidelines for the Analysis of Free Energy Calculations. J. Comput. Aided Mol. Des 2015, 29, 397–411. [PubMed: 25808134]

(70). Shirts MR; Chodera JD Statistically Optimal Analysis of Samples from Multiple Equilibrium States. J. Chem. Phys 2008, 129, 124105–124105. [PubMed: 19045004]

(71). da Cunha AR; Duarte EL; Lamy MT; Coutinho K Protonation/Deprotonation Process of Emodin in Aqueous Solution and pKa Determination: UV/Visible Spec-trophotometric Titration and Quantum/Molecular Mechanics Calculations. Chem. Phys 2014, 440, 69–79.

(72). Cerutti DS; Rice JE; Swope WC; Case DA Derivation of Fixed Partial Charges for Amino Acids Accommodating a Specific Water Model and Implicit Polarization. J. Phys. Chem. B 2013, 117, 2328–2338. [PubMed: 23379664]

(73). Karamertzanis PG; Raiteri P; Galindo A The Use of Anisotropic Potentials in Modeling Water and Free Energies of Hydration. J. Chem. Theory Comput 2010, 6, 1590–1607. [PubMed: 26615693]

(74). Yu H; Karplus M A Thermodynamic Analysis of Solvation. J. Chem. Phys 1988, 89, 2366–2379.

(75). Vosmeer CR; Rustenburg AS; Rice JE; Horn HW.; Swope, W. C.; Geerke, D. P. QM/M M -Based Fitting of Atomic Polarizabilities for Use in Condensed-Phase Biomolecular Simulation. J. Chem. Theory Comput 2012, 8, 3839–3853. [PubMed: 26593025]

(76). Cerutti DS; Swope WC; Rice JE; Case DA ff14ipq: A Self-Consistent Force Field for Condensed-Phase Simulations of Proteins. J. Chem. Theory Comput 2014, 10, 4515–4534. [PubMed: 25328495]

(77). Debiec KT; Cerutti DS; Baker LR; Gronenborn AM; Case DA; Chong LT Further Along the Road Less Traveled: AM BER ff15ipq, an Original Protein Force Field Built on a Self-Consistent Physical Model. J. Chem. Theory Comput 2016, 12, 3926–3947. [PubMed: 27399642]

(78). Haider N Checkmol v0.5a 2016; http://merian.pch.univie.ac.at/~nhaider/cheminf/cmmm.html, Accessed: January 25, 2018.

(79). Maclaurin D; Duvenaud D; Johnson M AutoGrad 1.2.1 2017; Accessed: January 25, 2018.

(80). Byrd RH; Nocedal J; Schnabel RB Representations of Quasi-Newton Matrices and Their Use in Limited Memory Methods. Math. Program 1994, 63, 129–156.

(81). Pluta T; Kolaski M; Medved M; Budzak S Dipole Moment and Polarizability of the Low-Lying Excited States of Uracil. Chem. Phys. Lett 2012, 546, 24–29.

(82). Roseman MA Hydrophilicity of Polar Amino Acid Side-Chains is Markedly Reduced by Flanking Peptide Bonds. J. Mol. Biol 1988, 200, 513– 522. [PubMed: 3398047]

(83). Vöhringer-Martinez E; Verstraelen T; Ayers PW The Influence of Ser-154, Cys-113, and the Phosphorylated Threonine Residue on the Catalytic Reaction Mechanism of Pin1. J. Phys. Chem. B 2014, 118, 9871–9880. [PubMed: 25059768]

(84). Saez DA; Vöhringer-Martinez E A Consistent S-Adenosylmethionine Force Field Improved by Dynamic Hirshfeld-I Atomic Charges for Biomolecular Simulation. J. Comput. Aided Mol. Des 2015, 29, 951–961. [PubMed: 26276557]

(85). Konig G; Bruckner S; Boresch S Absolute Hydration Free Energies of Blocked Amino Acids: Implications for Protein Solvation and Stability. Biophys. J 2013, 104, 453–462. [PubMed: 23442867]

(86). Cornell WD; Cieplak P; Bayly CI; Gould IR; Merz KM; Ferguson DM; Spellmeyer DC; Fox T; Caldwell JW.; 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.

(87). Jorgensen W.; Tirado-Rives J The OPLS [Optimized Potentials for Liquid Simulations] Potential Functions for Proteins, Energy Minimizations for Crystals of Cyclic Peptides and Crambin. J. Am. Chem. Soc 1988, 110, 1657. [PubMed: 27557051]

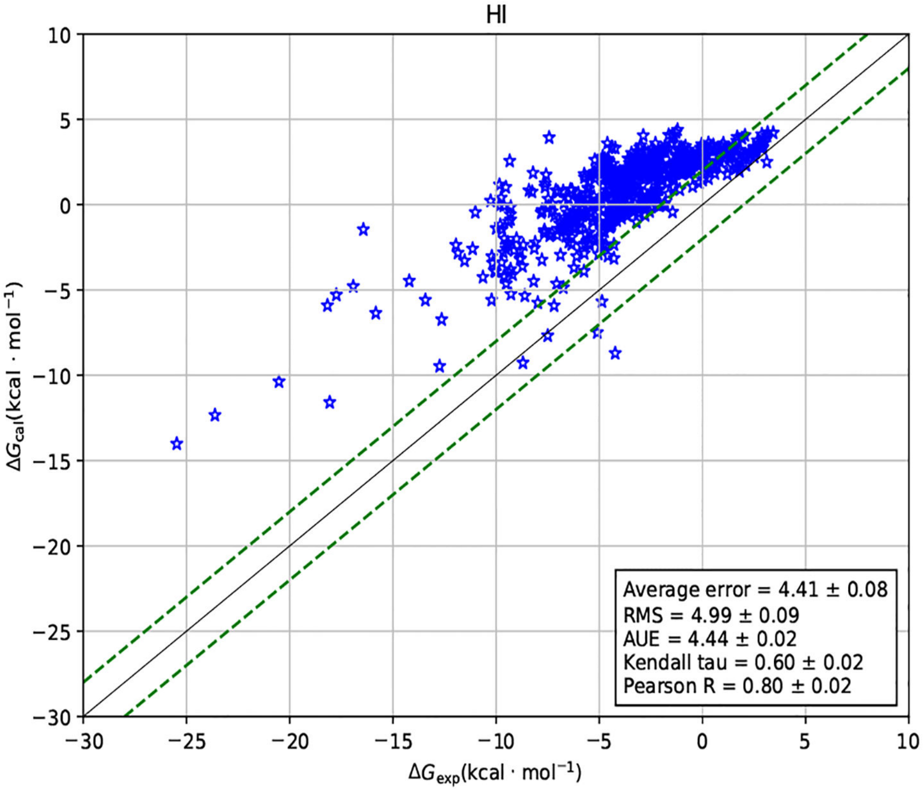


Figure 1:

Parity plot of calculated vs experimental hydration free energies for all FreeSolv molecules considered, using the GAFF force field and the atomic charges obtained from the Hirshfeld-I partitioning method of the BLYP/def2-TZVP molecular electron density in vacuum (The inset shows the result of error analysis: root mean square error (RMS), absolute unsigned error (AUE), Kendall tau and the Pearson R correlation coefficient for each charge set. Uncertainties were computed via bootstrapping as described elsewhere.”)

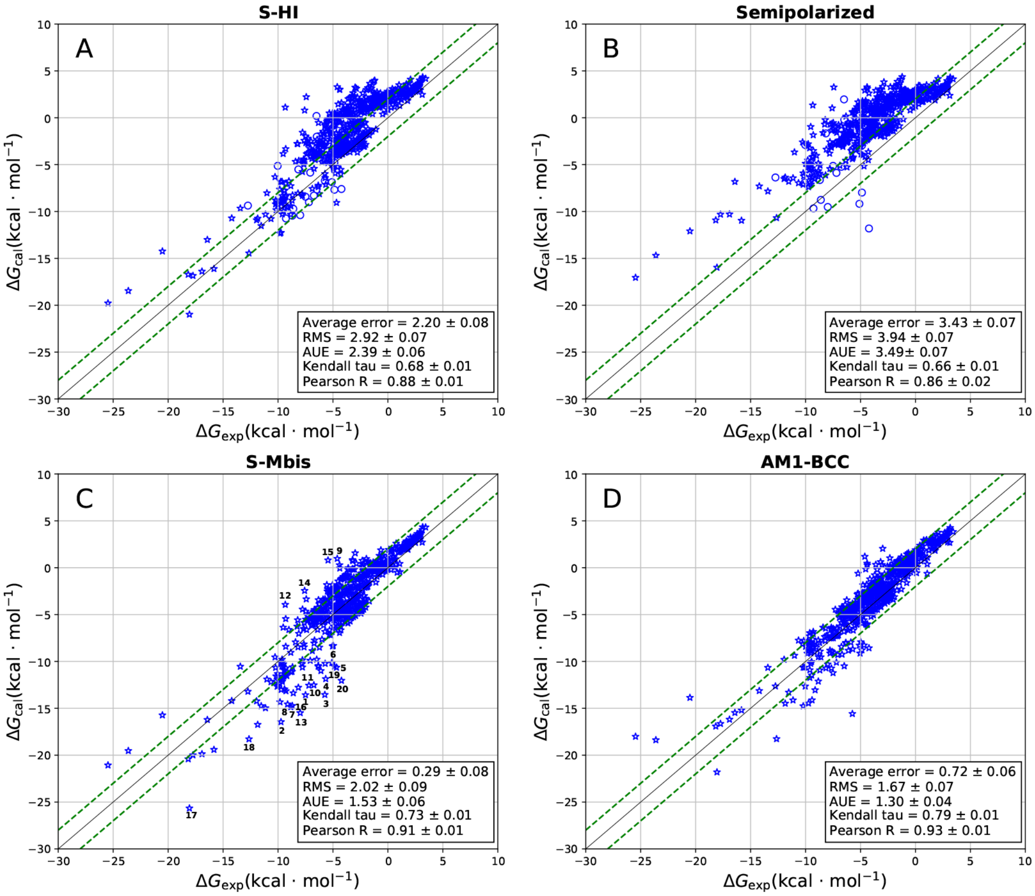


Figure 2:

Parity plot of calculated vs experimental hydration free energies for all FreeSolv molecules considered, using the GAFF force field and different methods to obtain atomic charges. (A) BLYP/def2-TZVP electronic structure method with SMD solvation model and Hirshfeld-I partitioning method adding the polarization correction to the free energy; (B) same electronic structure method with the semipolarized charges without applying any polarization correction; (C) same electronic structure method and SMD solvation model but using the Minimal Basis Iterative Stockholder (MBIS) partitioning correcting the transfer free energies for polarization and (D) using AM1-BCC atom charges in FreeSolv. Uncertainties were computed via bootstrapping as described elsewhere”

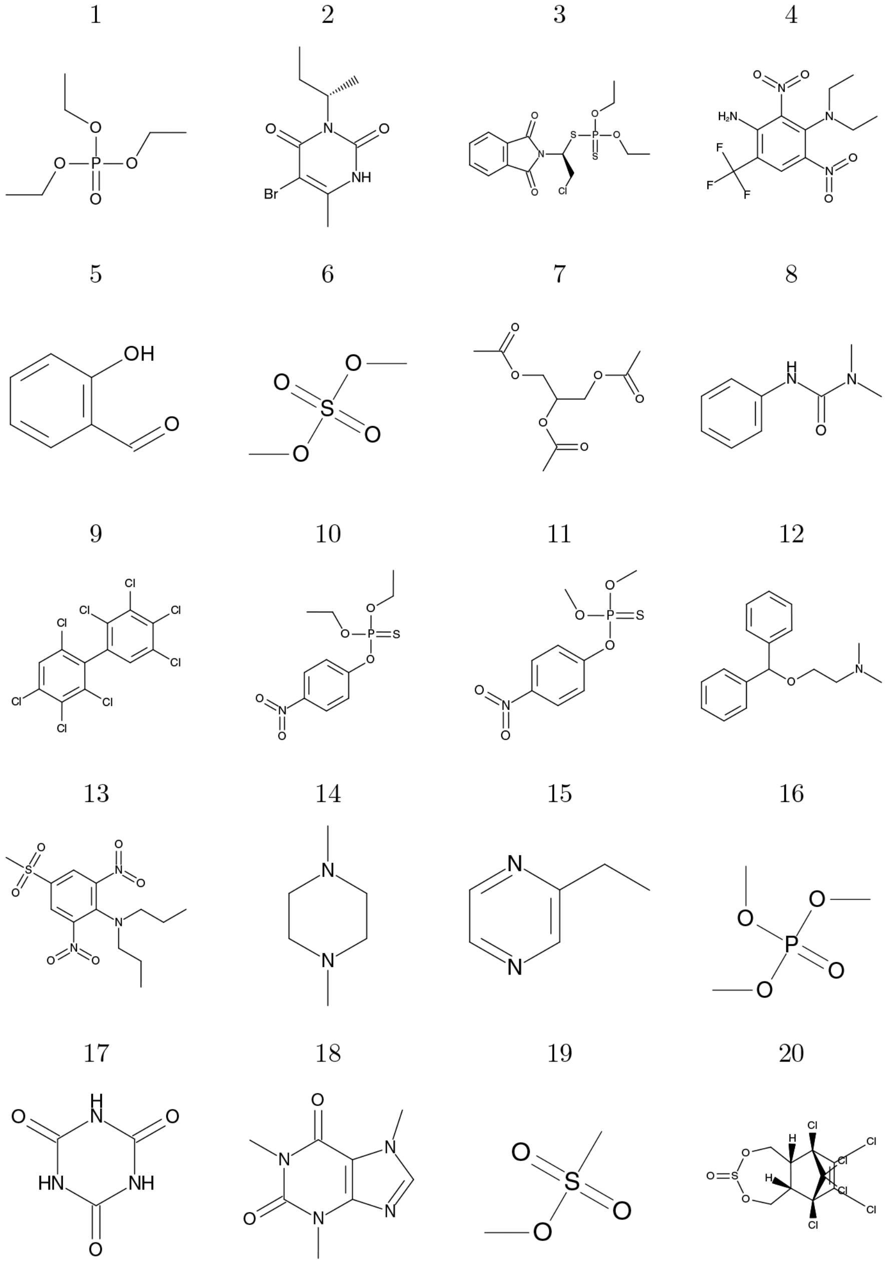


Figure 3:

Molecules which present a deviation of the calculated hydration free energy with the S-MBIS charges from the experimental value which is larger than five times the experimental error.

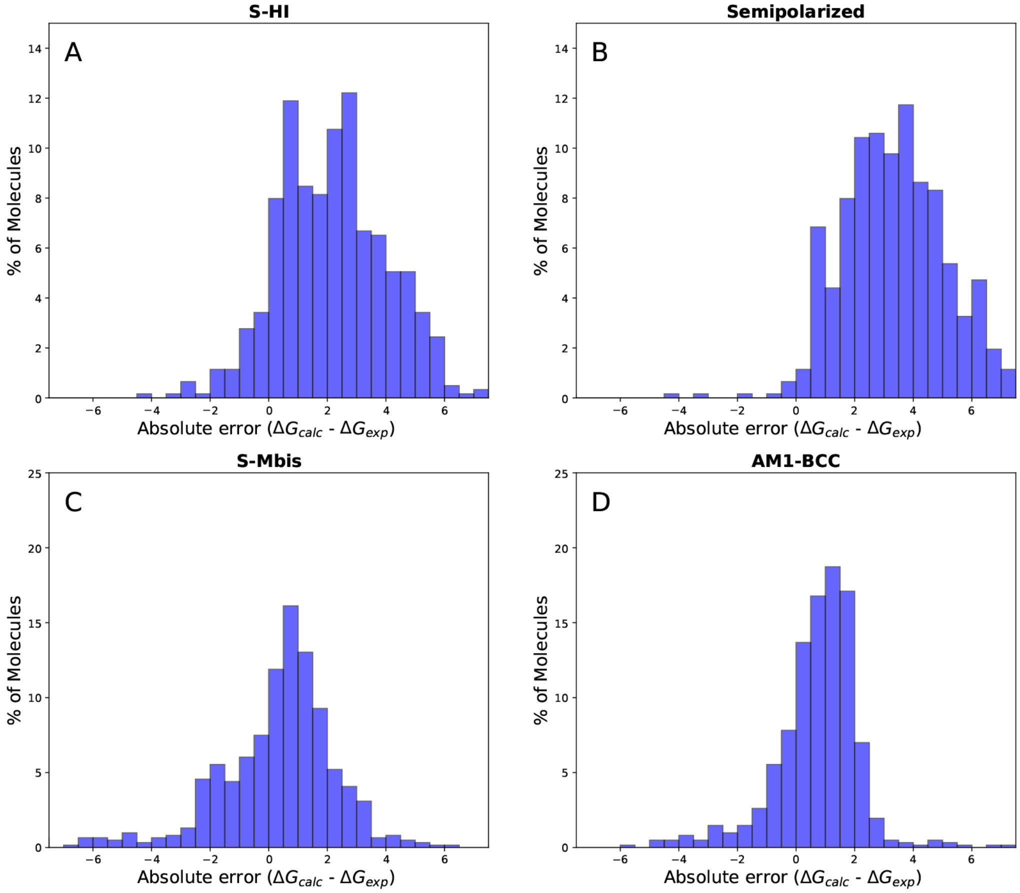


Figure 4:

Distribution of the absolute errors to the experimental value in the hydration free energies obtained with the S-HI charges (A), the semipolarized charges HI charges (B), the S-MBIS charges (C), and the AM1-BCC charges (D).

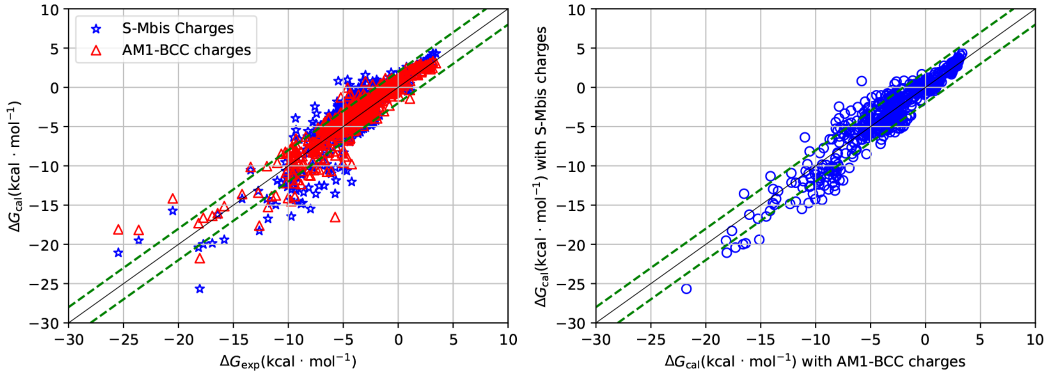


Figure 5:

Hydration free energies for all molecules of the FreeSolv Database obtained with the GAFF force field and the MBIS charges including the SMD solvent model in the electronic structure calculation (BLYP/def2-TZVP level) with the polarization correction (blue stars), and the hydration free energies with the same force field using the AM1-BCC atomic charges vs the experimental reference.

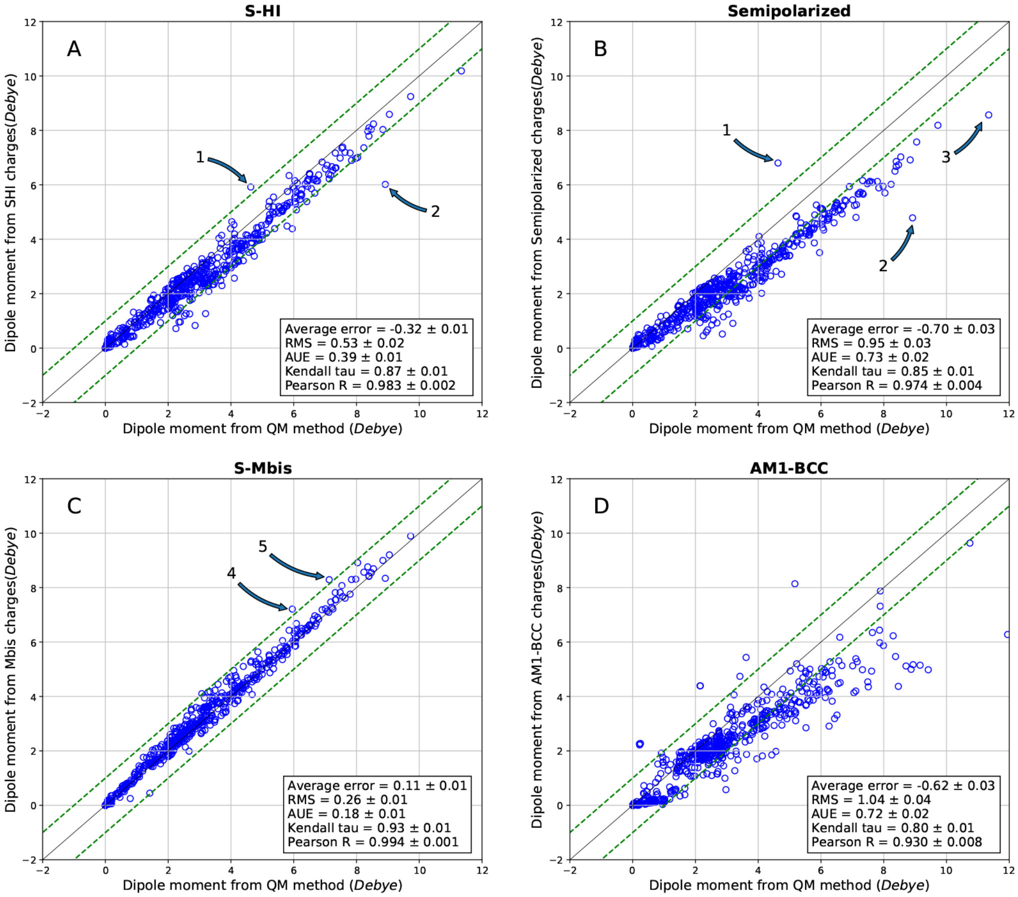


Figure 6:

Magnitude of the dipole moments of the molecules in the FreeSolv database calculated at the optimized geometry (BLYP/TZVP level with SMD solvent model) directly with the QM method vs the dipole moments resulting from the atomic charges obtained with the Hirshfeld-I (A) or the MBIS partitioning method (C) from the same electron density, the semipolarized charges with the Hirshfeld-I method (B) and the AM1-BCC atomic charges (D). The arrows identify outliers with structures shown in Figure 7.

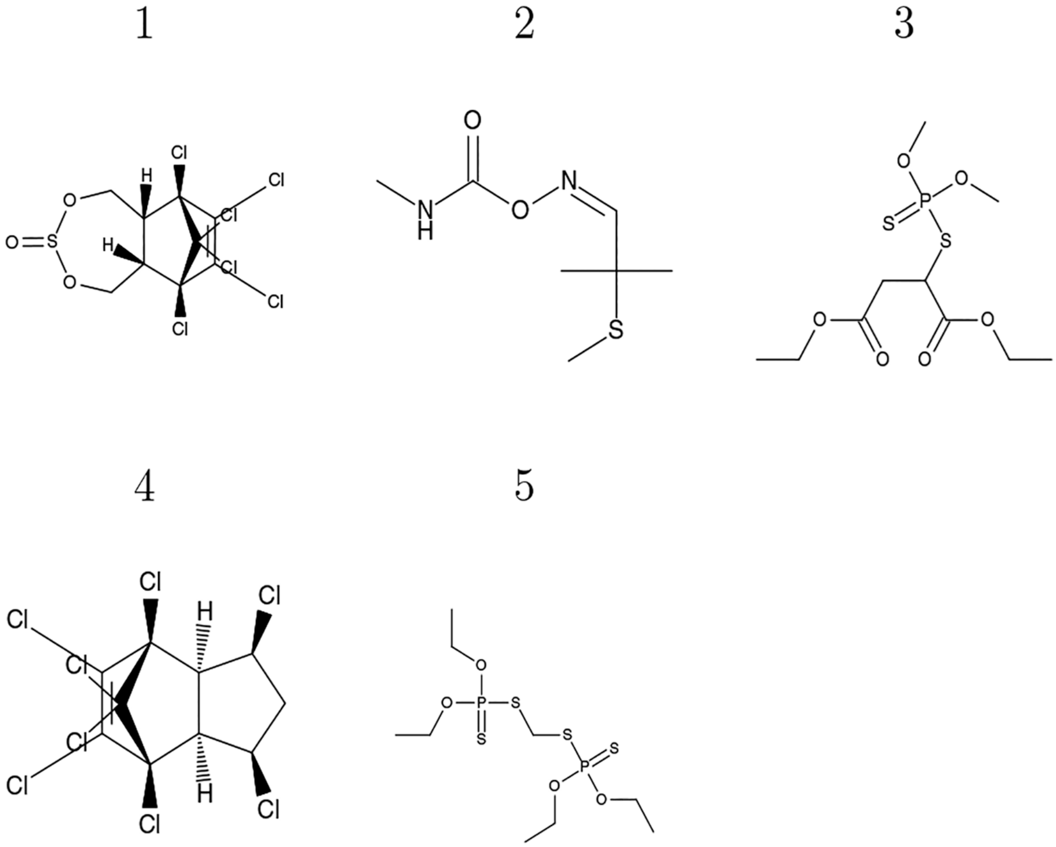


Figure 7:

Molecules labeled in Figure 6 which present a large deviation of their molecular dipole moment calculated with different atomic charges.

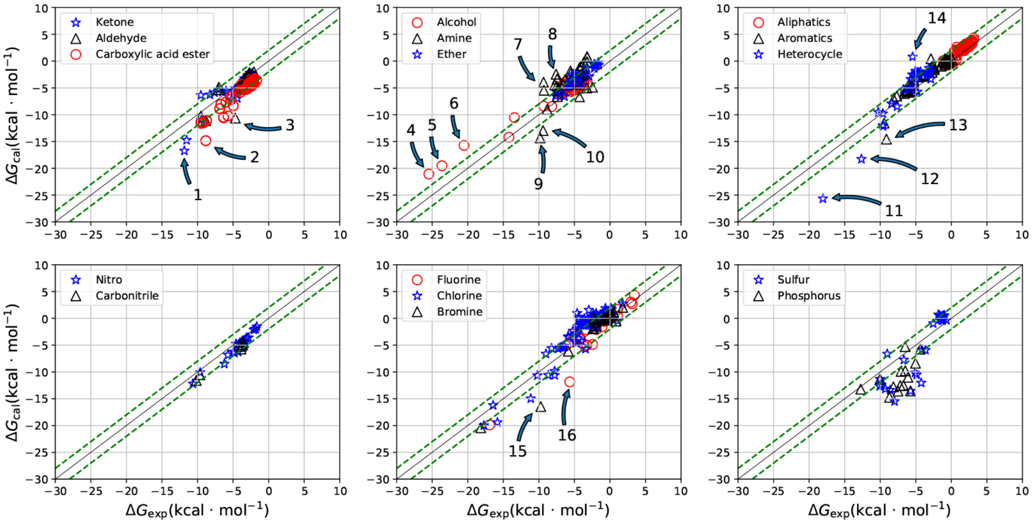


Figure 8:

Hydration free energies separated by functional groups in the FreeSolv database obtained with S-MBIS atomic charges. Outliers (numbers) are shown in Figure 9.

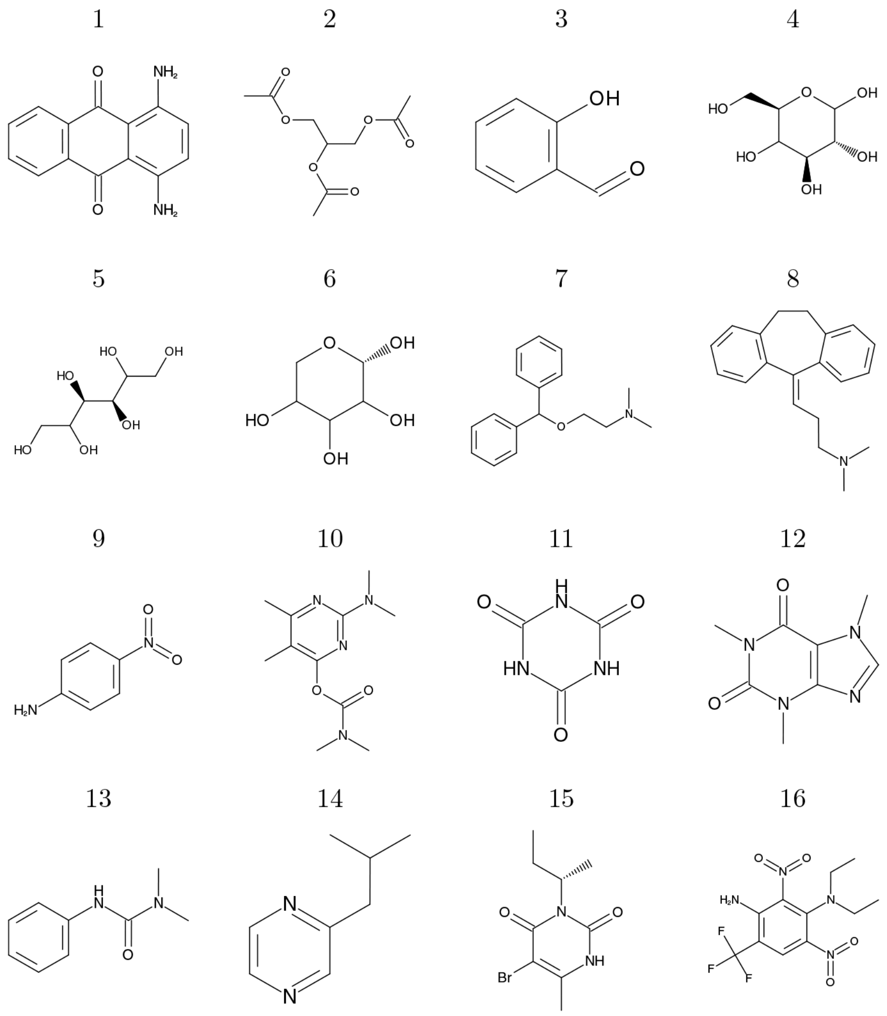


Figure 9:

Molecules identified as outliers in Figure 8.

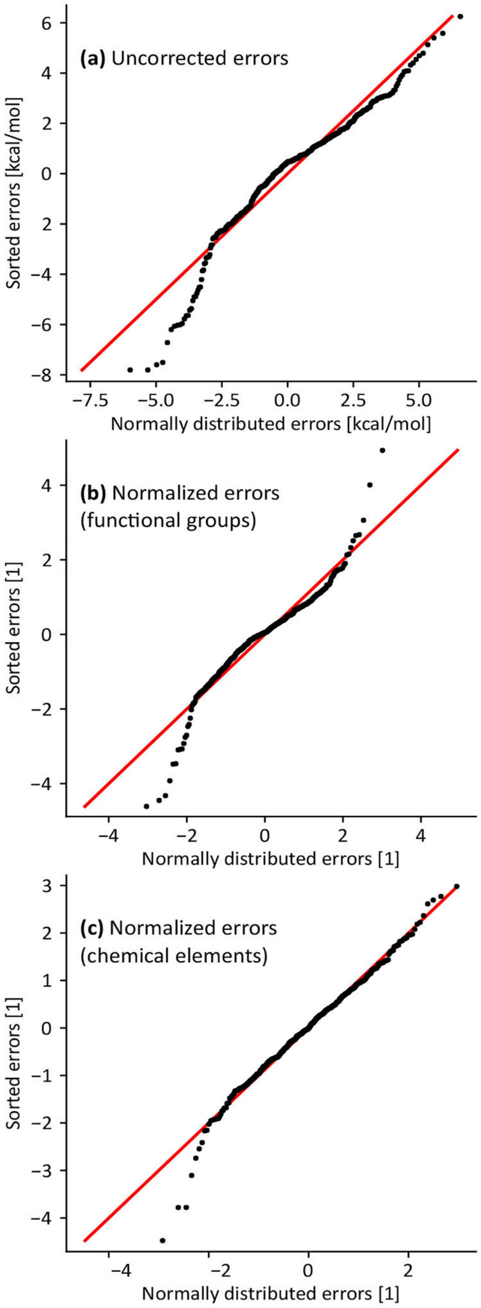


Figure 10:

Quantile-quantile plots of the error between S-MBIS and experimental hydration free energies: (a) uncorrected errors, (b) normalized errors using the error model based on functional groups and (c) normalized errors using the error model based on chemical elements. The sorted errors in each case are plotted against the corresponding quantile of normally distributed errors with the same mean and variance.

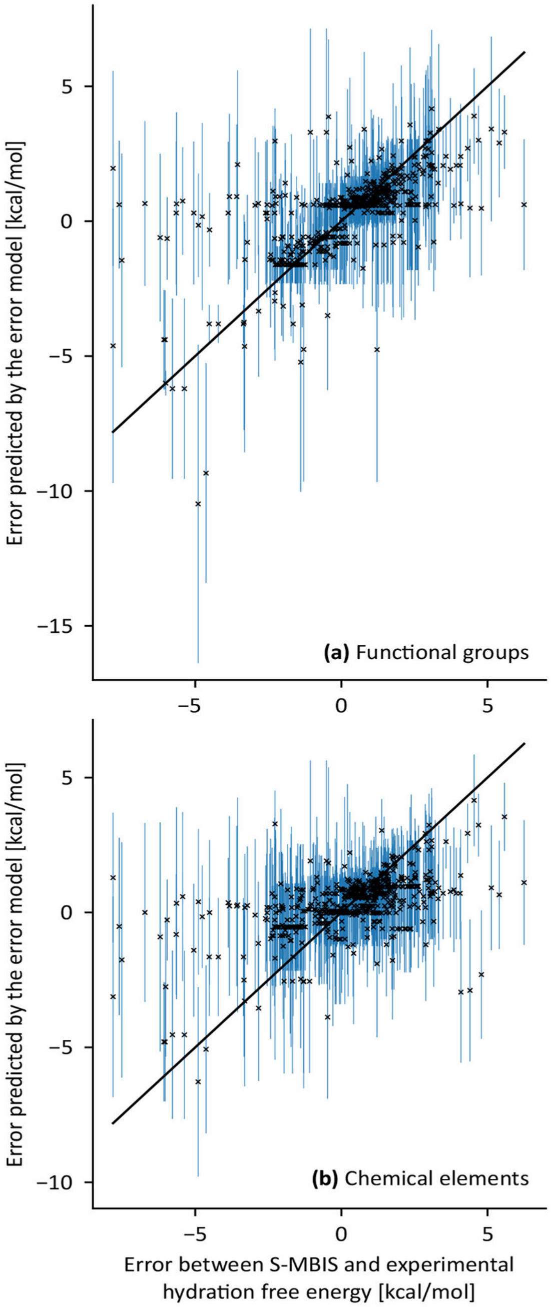


Figure 11:

The error predicted by the error model versus the actual error between the S-MBIS and experimental hydration free energy. Two variants of the error model were used, one based on (a) functional groups and (b) one on chemical elements. The systematic error is plotted with a black cross and the random error is represented by a blue vertical error bar.

Table 1:

Features used by the error model and their total occurrences (over the entire FreeSolv database). The first column shows in which feature set each feature is used: G=functional groups, E=chemical elements.

Set	Feature	Total
G	ketone	41
G	aldehyde	24
G	hydroxy	109
G	ester	58
G	ether	81
G	prim, amine	31
G	sec. amine	20
G	tert. amine	17
G	amide	12
G	nitro	31
G	nitrate	14
G	nitrile	12
G	aromatic	320
G	heterocycle	87
E	hydrogen	5772
E	carbon	3991
E	nitrogen	230
E	oxygen	624
E G	fluorine	99
E G	chlorine	304
E G	bromine	30
E G	phosphorus	15
E G	sulfur	51

Table 2:

Predictive error model parameters for the errors between S-MBIS calculations and experimental hydration free energies, using functional groups as features. All values are in kcal/mol. $\bar{\mu}_i$ represent systematic errors, where negative values indicate that theoretical predications underestimate the experimental hydration free energy. $\bar{\sigma}_i$ is the spread on random deviations, not transferable between different molecules.

Feature	$\bar{\mu}_i$	$\bar{\sigma}_i$
general	0.58	0.28
ketone	−1.17	1.02
aldehyde	−1.40	1.36
hydroxy	0.01	0.48
ester	−2.19	0.16
ether	0.50	0.50
prim. amine	0.63	0.73
sec. amine	1.48	1.12
tert. amine	1.49	1.73
amide	−2.00	1.24
nitro	−1.70	1.03
nitrate	−0.86	0.18
nitrile	−1.49	0.47
aromatic	0.17	0.41
heterocycle	−0.14	2.30
fluorine	−0.02	0.07
chlorine	0.30	0.41
bromine	0.04	0.27
sulfur	−0.28	2.56
phosphorus	−4.97	1.81

Table 3:

Predictive error model parameters for the errors between S-MBIS calculations and experimental hydration free energies, using chemical elements as features. All values are in kcal/mol. $\bar{\mu}_i$ represent systematic errors, where negative values indicate that theoretical predications underestimate the experimental hydration free energy. $\bar{\sigma}_i$ is the spread on random deviation, not transferable between different molecules.

Feature	$\bar{\mu}_i$	$\bar{\sigma}_i$
general	0.63	0.21
hydrogen	−0.03	0.02
carbon	0.07	0.04
nitrogen	0.17	1.57
oxygen	−0.59	1.05
fluorine	−0.06	0.09
chlorine	0.27	0.43
bromine	0.03	0.33
sulfur	0.38	1.14
phosphor	−2.98	0.65

Table 4:

Average and standard deviation on the error between S-MBIS predictions and ex perimental hydration free energies (i) without corrections, (ii) after correcting for systematic errors and (iii) after dividing by the predicted uncertainty.

		Functional groups	Chemical elements
(i)	$\mathrm{E}[\Delta_{\mathrm{tot},j}]$ [kcal/mol]	0.29	0.29
	$\mathrm{STD}[\Delta_{\mathrm{tot},j}]$ [kcal/mol]	1.99	1.99
(ii)	$\mathrm{E}[\Delta_{\mathrm{tot},j} - \overline{\mu}_j]$ [kcal/mol]	−0.03	−0.01
	$\mathrm{STD}[\Delta_{\mathrm{tot},j} - \overline{\mu}_j]$ [kcal/mol]	1.56	1.67
(iii)	$E[(\Delta_{\mathrm{tot},j} - \overline{\mu}_j)/\overline{\sigma}_j]$ [1]	−0.01	0.02
	$\mathrm{STD}[(\Delta_{\mathrm{tot},j} - \overline{\mu}_j)/\overline{\sigma}_j]$ [1]	0.96	0.93