A Supervised Fitting Approach to Force Field Parametrization with Application to the SIBFA Polarizable Force Field

Mike Devereux,*[a] Nohad Gresh,[b] Jean-Philip Piquemal,[c] and Markus Meuwly[a]

A supervised, semiautomated approach to force field parameter fitting is described and applied to the SIBFA polarizable force field. The I-NoLLS interactive, nonlinear least squares fitting program is used as an engine for parameter refinement while keeping parameter values within a physical range. Interactive fitting is shown to avoid many of the stability problems that frequently afflict highly correlated, nonlinear fitting problems occurring in force field parametrizations. The method is used to obtain parameters for the H_2O, formamide, and imidazole molecular fragments and their complexes with the Mg^{2+} cation. Reference data obtained from *ab initio* calculations using an auc-cc-pVTZ basis set exploit advances in modern computer hardware to provide a more accurate parametrization of SIBFA than has previously been available. © 2014 Wiley Periodicals, Inc.

DOI: 10.1002/jcc.23661

Introduction

Obtaining robust and transferable force field parameters is a prerequisite for all classical molecular mechanics methodologies. A description of the energy of a chemical system as a function of its nuclear coordinates is required to obtain the structural, energetic, and thermodynamic information that is the ultimate goal of all such computational studies.

The vast available chemical space means that force fields, and especially those with limited transferability of parameters between similar chemical environments, require whole libraries of fitted values to describe complex, heterogeneous systems such as solvated biomolecules. As more detailed potentials emerge[1–5] the number of fitted parameters required for each system also tends to increase. Manual fitting of parameters can be laborious and often demands a significant degree of user expertise to keep parameter values within a physical range to ensure that they will be compatible with other parameters in the force field and to maximize transferability between different chemical environments. In the following, we address the issue of automating the process of parameter fitting for the SIBFA polarizable potential.

Sum of Interactions Between Fragments *Ab Initio* (SIBFA)

SIBFA[6–10] is a highly detailed force field, including terms that explicitly account for electrostatic, exchange repulsion, electronic polarization, and charge-transfer contributions to the interaction energy:

$$E_{int} = E_{el} + E_{rep} + E_{pol} + E_{ct} \quad (1)$$

The additional terms incur a computational cost when compared with simpler models but excellent agreement with *ab initio* data can be achieved at a fraction of the computational cost of a full self consistent field (SCF) calculation. A particular strength of SIBFA has been application to metal containing systems such as metalloenzymes and ions in solution,[11–14] where the polarization and charge-transfer terms contribute significantly to the total energy and the neglect of these terms in simpler, faster force field approaches precludes their use for many applications. While other molecular mechanics approaches have been developed to describe transition-metal complexes,[1,15] SIBFA does not require a bonded term to describe short-range ligand–metal interactions and so a single parametrization can be used to describe essentially any arrangement of ligands around the central metal (important in catalytic protein binding sites), and even ligand–metal bond breaking and formation in dissociation and substitution reactions. The original parametrization of SIBFA, using lightweight SBK[16] (Stevens, Basch, Krauss) core potentials and basis sets for reference data, has been successfully used in many different studies.[6,13,14,17]

[a] *M. Devereux, M. Meuwly*
Department of Chemistry, University of Basel, Klingelbergstr 80, CH 4056, Switzerland
E-mail: Michael.Devereux@unibas.ch
[b] *N. Gresh*
Chemistry & Biology, Nucleo(s)tides & Immunology for Therapy (CBNIT), UMR 8601, CNRS, UFR Biomedicale, Paris, France
[c] *Jean-Philip Piquemal*
Laboratoire de Chimie Théorique, UPMC, Sorbonne Université, Campus de Jussieu, 4 place Jussieu, Paris, France
Contract grant sponsor: Schweizerischer Nationalfonds; Contract grant number: 200021-117810; Contract grant sponsor: NCCR MUST; Contract grant sponsor: Grand Equipement National de Calcul Intensif (GENCI): Institut du Développement et des Ressources en Informatique Scientifique (IDRIS), Centre Informatique de l'Enseignement Supérieur (CINES), France (project No. x2009–075009), and the Centre de Ressources Informatiques de Haute Normandie (CRIHAN, Rouen, France), project 1998053

The basic premise of SIBFA is to build a force field that can account for each of the terms of an energy decomposition of the *ab initio* intermolecular interaction energy, obtained using, for example, the reduced variational space (RVS) approximation of Stevens and Fink.[18] In RVS, the total interaction energy is broken down into first-order Coulomb (E_c) and exchange-repulsion (E_e) energies and second-order polarization (E_{pol}) and charge-transfer (E_{ct}) terms. Essentially any energy decomposition approach could be used, such as SAPT,[19] CSOV,[20] or PIEDA,[21] so long as the experimentally observable total binding energy is recovered by summing the individual contributions. Agreement with correlated, post-Hartree–Fock (HF) methodologies is achieved by inclusion of an additional dispersion energy term.[9] Accurate representation of these energy contributions within the force field allows SIBFA to reconstruct the total *ab initio* interaction energy for a given system. An accurate description of the potential energy surface provides the basis to study formation, interaction, and dynamics of complex chemical systems. Multipolar force fields of this nature have been shown to successfully describe bulk properties in condensed phase molecular dynamics simulations.[22,23]

Large molecules in SIBFA are constructed from libraries of smaller fragments. As parameters are fitted for the optimized geometry of each fragment, fragments are traditionally kept rigid and conformational flexibility of the larger molecule is achieved using torsional degrees of freedom about interfragment bonds. Fragments within a molecule interact with one another according to eq. (1), describing intramolecular polarization, electrostatic and steric terms. The force field is currently being extended to allow full relaxation of all degrees of freedom for molecular dynamics simulation. While the original SBK SIBFA parametrization has performed well in many different applications to date, recent extensions to more demanding systems[24] have highlighted the advantages of using more accurate, and robust reference data. Establishing a semiautomated workflow that would allow robust and efficient reparametrization of the whole force field is the ultimate goal of this work.

Force field terms

SIBFA uses multipole moments derived from the *ab initio* charge density to describe electrostatic interactions. Multipole moments give a greater degree of anisotropy than can be achieved using atom-centered point charges alone. This anisotropy can be essential to describe, for example, the out-of-plane charge concentrations associated with oxygen atom lone pairs.[25] Multipolar force fields are therefore starting to find wider application in molecular mechanics studies and in molecular dynamics simulations.[4],[26–31] The partitioning scheme originally used to divide the molecular multipole moments into atom- and bond-centered contributions in SIBFA was developed by Vigné-Maeder and Claverie,[32] although SIBFA is compatible with essentially any partitioning scheme and is also used in conjunction with refined distributed multipole analysis[33] (DMA) multipole moments in this work (see Methods).

The multipolar interaction energy E_{mtp} in SIBFA is truncated at quadrupole–quadrupole interactions. The sharing of multipoles between both atom and bond centers aims to improve short-range convergence of the electrostatic interaction energy as a function of multipolar rank, as convergence is not guaranteed at all for points within the maximum radius of the charge density volume that is described by the multipole moments.[34] Increasing the number of multipolar sites reduces the volume of charge density described by each site, which can improve the short-range convergence properties of E_{mtp}.[35] A more recent extension[10] to the electrostatic term in SIBFA also adds a penetration contribution to account for the overlap of molecular charge densities at very close range. This is particularly important to describe interactions such as H-bonds and ligand–metal binding, where atom–atom distances are short and there is significant overlap of the electron clouds associated with each monomer. This penetration energy function comprises parameters which cannot be calculated directly, and need to be fitted.

Short-range electron–electron repulsion is evaluated between bond and lone-pair sites in SIBFA, yielding an anisotropic description that more closely matches the total repulsion energy than the simpler atom-centered Lennard–Jones terms more commonly used in classical force fields.[9] The form of the function encapsulates electron-pair overlap, and is evaluated between all bonds and lone pairs in the interacting molecules. Overlap integrals are approximated using tabulated orbital coefficients, and effective radii that need to be fitted to RVS reference data.

Charge transfer can also amount to several kcal/mol of the total interaction energy in metal–ligand complexes in particular. Electron lone pair positions are again used in SIBFA to represent the positions of electron donor sites. A characteristic acceptor strength of the cation is used to evaluate the degree of charge transfer between each lone pair and the central metal acceptor ion. Atomic radii again need to be fitted to RVS data to accurately evaluate this term.

Finally, the polarization energy in SIBFA uses a 3×3 polarizability matrix for each bond barycenter and electron lone pair. As a second-order correction to the electrostatic interaction energy, polarization affects primarily the valence electrons so placement of polarizable sites away from the nuclear positions offers a more natural description of the response of the molecular charge density to an external electric field. Significant effort has been invested in the development of polarizable force fields over recent years[2,3,5,36] as more accurate descriptions of intermolecular and intramolecular interaction energies are sought. Without this term a molecule does not respond to its environment, so in a nonpolarizable force field a water molecule in the gas phase is identical to a water molecule in bulk liquid, at an interface or in the presence of an ion.

In SIBFA, the polarization energy is refined iteratively until self consistency of the electric field between molecules is reached. In other words, monomer A polarizes monomer B, the now polarized monomer B repolarizes monomer A in

return and so on. While the necessary polarizability matrices can be directly calculated and are not fitted, the so-called "polarization catastrophe" that occurs at close range as polarizable sites come close to one another requires the use of a further damping function. The parameters involved in this damping function are again fitted using RVS data.

Methodology

RVS reference data

Unless otherwise stated, all electronic structure calculations, energy decomposition analyses, and calculation of physical properties were carried out using GAMESS.[37] RVS energy decomposition calculations were performed for complexes and dimers of H_2O, formamide, and imidazole with Mg^{2+}, chosen for their relevance to biological systems. The inclusion of a metal cation is important for parametrization of polarization and charge-transfer terms. RVS analysis is currently restricted to Hartree-Fock Self Consistent Field (HF SCF) calculations, although there is an explicit dispersion energy term in SIBFA which will be added subsequently to yield agreement with post-HF methods. This term is not dealt with in the current work, meaning that binding energies reported here can differ slightly from post-HF reference values. The aug-cc-pVTZ basis set was taken from the basis set exchange.[38] *f*-functions were removed to improve stability of the RVS calculations, and were found to have negligible effect on the different energy components. HF geometry optimizations for monomers and ligands were performed using the Gaussian09[39] program with the same aug-cc-pVTZ basis set.

RVS analysis for larger systems can be extremely time-consuming. A large basis set in combination with more than three or four monomers can make calculations prohibitive for even quite small monomers (5–6 heavy atoms). In addition, no "direct" method is implemented in GAMESS for RVS calculations, meaning that all necessary integrals are written to disk and reread as needed subsequently. This approach can be very efficient for small systems, but for larger systems the hard-disk requirements become prohibitive and the time devoted to I/O operations greatly reduces the gains made by rereading, rather than recalculating the relevant integrals. There is still significant insight to be gained from a decomposition of the total complex binding energy into first-order (Coulomb plus exchange repulsion) and second-order (charge-transfer plus polarization) terms, however, to help identify which terms in SIBFA are performing well and which may be responsible for deviations from the total HF binding energy. To resolve this issue a method is introduced here that obtains the total first-order and second-order RVS energy for large complexes.

First, the complex to be examined is broken into its *N* constituent monomers. The co-ordinates of each monomer *i* correspond to those in the original complex, so that there are no axis-system incompatibilities when the complex is rebuilt. A single-point energy calculation is performed for each isolated monomer to obtain a gas-phase monomer wave function, described by the coefficients and exponents of the corresponding basis functions. The individual monomer wave functions $|\psi_i\rangle$ are then combined into a supermolecular "pseudowave function" for the entire complex $|\phi\rangle$, composed of gas-phase monomer wave functions. The "pseudowave function" is then read as an initial guess for an SCF calculation of the entire complex. The energy E_0 of the first SCF cycle, representing the ground-state wave function of the complex in the presence of the frozen, gas phase monomer charge densities, is obtained by diagonalizing the transformed Fock matrix F' which was constructed using $|\phi\rangle$ as a guess density [eq. (2)]. E_0 is used to yield the total first-order RVS energy E_1 according to eq. (4). Then, following convergence of the SCF cycles, a wave function $|\Psi\rangle$ is obtained with energy E_{sup} [eq. (5)]. The difference between E_{sup} and the first-order energy E_1 gives the total RVS second-oder energy E_2 [eq. (6)].

$$F'C' = C'\varepsilon \tag{2}$$

$$E_i = \langle \psi_i | \mathrm{H} | \psi_i \rangle \tag{3}$$

$$E_1 = E_0 - \sum_{i=1}^{\mathrm{N}} E_i \tag{4}$$

$$E_{\mathrm{sup}} = \langle \Psi | \mathrm{H} | \Psi \rangle \tag{5}$$

$$E_2 = E_{\mathrm{sup}} - E_0 \tag{6}$$

$$E_{\mathrm{bind}} = E_{\mathrm{sup}} - \sum_{i=1}^{N} E_i = E_1 + E_2 \tag{7}$$

The sums run over the energy E_i of each gas-phase monomer wave function. C' is the transformed coefficient matrix of the SCF procedure[40] and ϵ is the vector of eigenvalues.

Calculated properties

Polarizabilities were calculated directly for use in SIBFA. 3 × 3 polarizability matrices for each site, as well as the spatial positions of the lone pairs and bond barycenters were calculated using Boys and Foster[41,42] orbital analysis and an aug-cc-pVTZ basis set. Fragment geometries were optimized at the same level of theory using Gaussian09.

Traditionally, multipole moments up to quadrupole were obtained for SIBFA according to the procedure of Claverie and Vigné-Maeder[32] for each site to describe its apportioned electron density. As is the case for all purely orbital-based methods, however, this procedure was found to suffer from a dependence on the underlying basis set used. The very diffuse basis functions included in the large aug-cc-pVTZ basis set caused significant problems in the conjugated systems studied here, and so generalized distributed multipole analysis (GDMA) multipole moments were incorporated as an alternative. While also an orbital-based approach, recent adaptations[33] that divide up very diffuse functions using a spatial grid help to overcome the issues associated with very large basis sets. Precisely how the spatial partitioning takes place is based on the relative sizes of the atoms, as defined by the user. The result is an essentially infinite number of different partitioning possibilities for a given molecule, each reproducing the total

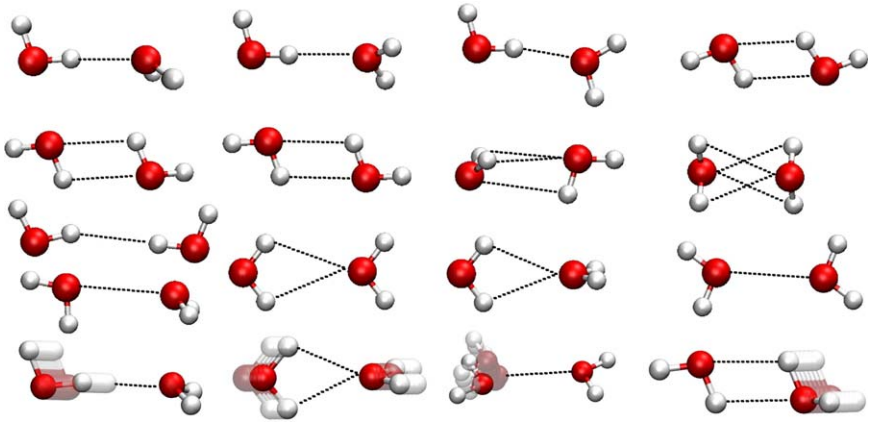

Figure 1. Figure showing the different water dimers used in the training set. Scans of a particular coordinate are indicated by overlaying all sampled points as transparent images on top of the initial structure.

molecular multipole moments correctly but each with different short-range convergence properties. We are not aware of any method to systematically identify the GDMA atomic radii with the best convergence properties, and different combinations of radii produced significantly different results. In the case of formamide, no satisfactory scheme was found that could accurately represent the electrostatic interaction energy when atomic and bond multipole moments were truncated at quadrupole, including at medium range where penetration effects are small.

It was, therefore, considered important to introduce a technique to refine the initial GDMA multipole moments, improving their convergence properties and allowing truncation of the atomic multipole expansion at the quadrupole. Following earlier work,[22],[43–45] and other, similar approaches,[46,47] the multipole moments were treated as parameters in a fitting scheme and the electrostatic potential across a grid outside the 0.001 a.u. isodensity surface of the molecule was used as target data. For convenience, I-NoLLS[48] (an Interactive Non-Linear Least Squares fitting program, see later) was used as the fitting engine. Grid points that lay inside the molecular volume were discarded, as the electrostatic potential in this region is relevant to molecular stability and not to intermolecular interactions. The excluded molecular volume was defined using the 10^{-3} a.u. electron isodensity surface of the molecule, chosen as it lies roughly at the boundary of the interaction region of interest. The minimum distance from each atom to this surface defined an atomic radius, and points lying within the radius of any atom in the molecule were discarded. The total molecular charge, dipole, and quadrupole moments were strictly maintained during fitting by dividing any excess or deficiency between all N atomic and bond centers, that is, taking only the z-axis components as an example:

$$dq = \frac{1}{N}\left[\left(\sum_{i=1}^{N} q_i\right) - q_{\mathrm{Mol}}\right] \tag{8}$$

$$d\mu_z = \frac{1}{N}\left[\left(\sum_{i=1}^{N} r_{z,i}q_i + \mu_{z,i}\right) - \mu_{z,\mathrm{Mol}}\right] \tag{9}$$

$$dQ_{zz} = \frac{1}{N}\left[\left(\sum_{i=1}^{N} 2r_{z,i}\mu_{z,i} - r_{x,i}\mu_{x,i} - r_{y,i}\mu_{y,i} + 0.5(3r_{z,i}^2 - r_i^2)q_i + Q_{zz,i}\right) - Q_{zz,\mathrm{Mol}}\right] \tag{10}$$

where dq is the correction added to each partial charge (placed at nuclei and bond barycenters) to maintain the correct total molecular charge, $d\mu_z$ is the dipole z-component correction and dQ_{zz} is the quadrupole zz-component correction. The sum in each case runs over all N atom or bond centers. $r_{z,i}$ is the z-coordinate of center i relative to the molecular origin used to define the reference molecular multipole moments μ_{Mol} and Q_{Mol}. r_i is the distance of center i from the molecular origin and the molecular charge, q_{Mol}, is zero for the neutral ligands studied here.

All atomic and bond-centered multipole moments were additionally loosely constrained to their initial GDMA values to avoid problems associated with "buried atoms" that are known to affect the analogous "CHELPG"[49] point-charge fitting approach. The result is a set of atom- and bond-centered multipole moments that optimally describes the electrostatic potential in the important interaction region between molecules. The number of sites necessary (atoms and bonds or bonds only) and the maximum rank of each site needed (truncation at charge, dipole, or quadrupole) was not investigated and was considered beyond the scope of this investigation.

Parameter fitting

The goal of this work is to devise a semiautomated approach to parameter fitting for SIBFA using the more detailed HF/aug-cc-pVTZ level of theory for reference data and to compare this procedure with manual fitting techniques.

The manual fitting approach previously used[12,50] for SIBFA involves first obtaining parameters for water, then monoligated and diligated complexes of the divalent cation, and finally further organic ligands including formamide and imidazole. The water parameters are based on dimers and include H···O, H···H, and O···O interactions (Fig. 1). This combination unambiguously determines H- and O-parameters, whereas if

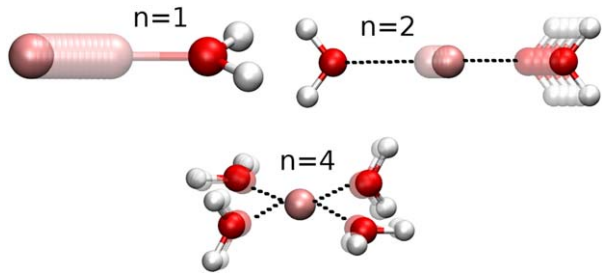

Figure 2. Figure showing the different $[Mg(H_2O)_n]^{2+}$ complexes used in the training set. Scans of a particular coordinate are indicated by overlaying all sampled points as transparent images on top of the initial structure.

only the more chemically important H···O interactions were probed then many parameters such as atomic radii would not be uniquely defined. Only one radius correctly defines the O···O distance, whereas many combinations would correctly define O···H. Mg^{2+} is then used as a standard probe, with its parameters determined by the now fixed O and H parameters (Fig. 2). The Mg^{2+} probe parameters are then fixed and used to fix parameters for formamide (Fig. 3) and imidazole (Fig. 4).

This approach was adapted for semiautomated fitting. Fitting algorithms are designed to find the combination of parameters that minimize the difference between observed (reference) and calculated data. As recently shown in a study of Iridium complexes described by the VALBOND-TRANS force

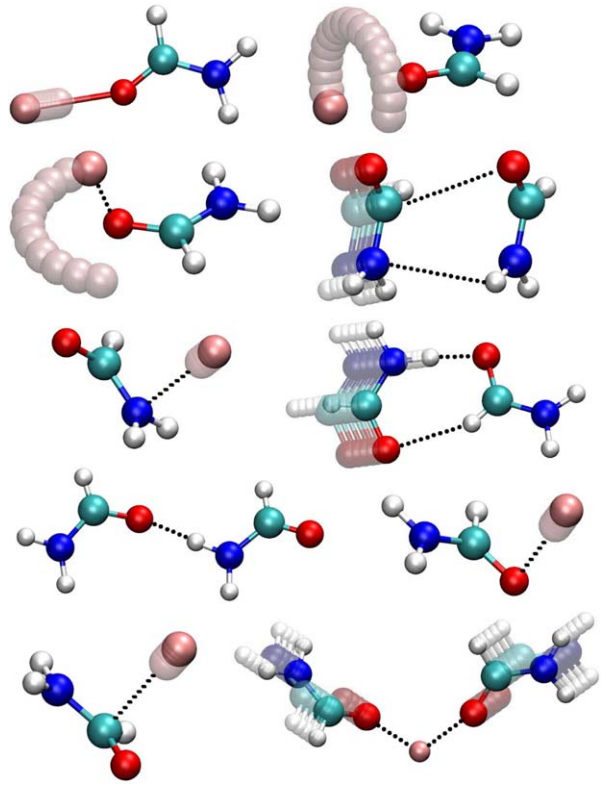

Figure 3. Figure showing the different $[Mg(HCONH_2)]^{2+}$ complexes and $HCONH_2$ dimers used in the training set. Scans of a particular coordinate are indicated by overlaying all sampled points as transparent images on top of the initial structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

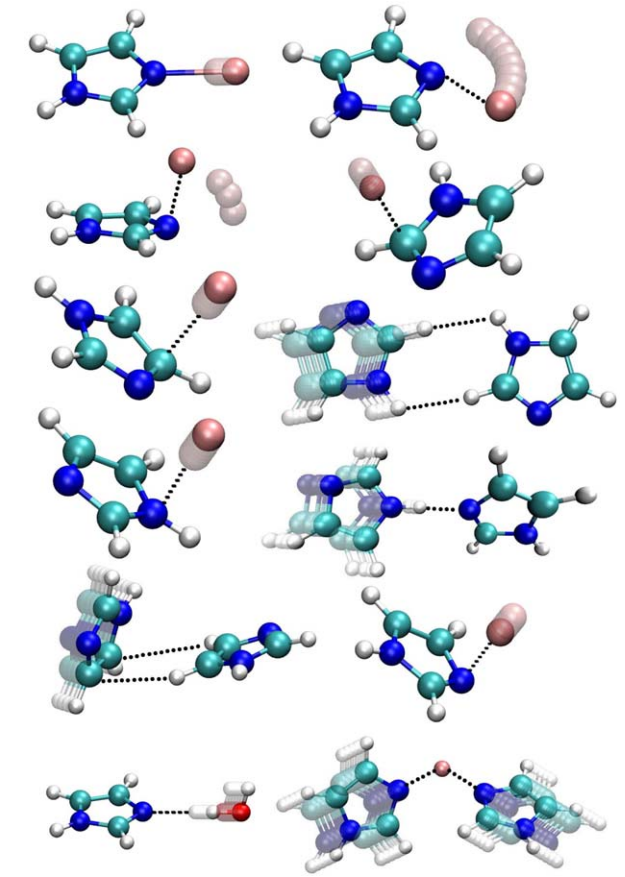

Figure 4. Figure showing the different $[Mg(C_3N_2H_4)]^{2+}$, $C_3N_2H_4$ dimers and $C_3N_2H_4\cdots H_2O$ complexes used in the training set. Scans of a particular coordinate are indicated by overlaying all sampled points as transparent images on top of the initial structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

field[51] a restricted training set can be fitted with a very high degree of accuracy using automated approaches, but parameters are no longer applicable to even quite closely related complexes with different ligand combinations. In other words, the parameters are not "transferable." Fitting across several types of complexes simultaneously leads to larger errors in the description of each complex individually, but a much more robust set of parameters is obtained that is equally applicable to cases outside of the original training set. Many aspects can affect transferability when a training set is too small. For this work, fitting initially to water complexes where polarization is relatively weak can lead to parameters that try to account for small shortcomings in the SIBFA methodology (such as the absence of quadrupolar polarizabilities or limitations at short range) in an averaged way. This small improvement in the description of the polarization energy for $H_2O\cdots H_2O$ interactions can lead to problems in the presence of a cation with a much larger charge where significant short-range screening is required, limiting the transferability of the water parameters severely. During manual fitting, a combination of judgment and experience can be used to avoid unfavorable parameters, and where problems do arise they can be corrected subsequently by revisiting the original training data in the light of

new information. For a semiautomated approach, it is much more effective to use a diverse training set from the outset, as parameter values that are suitable for one chemical environment but not for another one are easily detected with both chemical environments present in the training data.

I-NoLLS. I-NoLLS is an interactive, nonlinear least squares program that allows supervision of a fitting process to speed up convergence and avoid nonphysical parameter values.[48] Each step in parameter space is selected based on the Jacobian matrix J of partial derivatives of each data element y_i (in this case RVS energies) with respect to each of the available parameters p_i:

$$J_{ij} = \frac{\partial y_i^{\text{calc}}}{\partial p_j} \tag{11}$$

Singular value decomposition (SVD) of J (with the elements of J weighted according to the "importance" assigned by the user to each of the data points) can be used in combination with the Levenberg-Marquardt algorithm in I-NoLLS to effectively deal with nonlinear, highly correlated fitting problems. Here, "nonlinear" indicates that the Jacobian matrix depends on the current position in parameter space, requiring iterative refinement of the error by taking small steps and re-evaluating J each time. Force field fitting is often highly nonlinear, for example, due to the characteristic exponential terms involved in the short-range repulsion and damping functions in SIBFA. "Correlated" signifies that two or more parameters are nonorthogonal with respect to their relationship to reference data. For example, atomic radii tend to be highly correlated as an increase in one is often compensated by a decrease in another.

The singular values from SVD of J in I-NoLLS can be used to determine the best-defined step in parameter space, which is the step that will allow the largest decrease in the total error for the smallest change in parameter values. This is especially important in highly correlated fitting problems to increase the stability of the fit when far from a minimum in the total error, by discarding large changes in correlated parameter values that have little effect on the total error. The Levenberg–Marquardt algorithm then offers a means to prevent large changes in any given parameter by effectively applying soft constraints that tie parameters to their current values. Restricting step size is important in nonlinear fitting problems that are far from a minimum. A final, simpler option in I-NoLLS allows users to scale the current parameter step by a factor less than one, which can again improve stability and convergence in highly nonlinear fitting scenarios.

The final concern in a highly correlated fitting problem is to impose suitable constraints on the values of parameters that have physical meaning. A negative atomic radius, for example, is clearly nonphysical. It may also be advantageous to build-in human experience into the fitting procedure, such as recognizing that exceeding a certain limit for a given parameter, such as the damping parameters of an atomic polarizability, could lead to instability in certain circumstances, such as occurrence of the so-called "polarization catastrophe" at close range. Soft constraints, which bias parameters to remain within a certain range but do not enforce this range rigidly, can be implemented in I-NoLLS by including initial parameter values as observable data. Deviating from initial values then incurs an error in the corresponding "data point" that is proportional to the amount of deviation, with a proportionality constant defined by the user-defined precision of the data point.

No "hard constraints" are currently implemented in I-NoLLS. Pseudohard-constraints were, therefore, devised using an exponential error term of the form:

$$Err = 10^{10(|p_i - p_{0,i}| - \delta p_i + 0.1)} + k|p_i - p_{0,i}| \tag{12}$$

where p_i is the current value of parameter i, $p_{0,i}$ is the center of the range of values to which it should be constrained, δp_i is the maximum allowed deviation of parameter i from $p_{0,i}$, and k is a constant currently set to $2/\delta p_i$. The first term is an exponential function that increases rapidly when $|p_i - p_{0,i}| > \delta p_i$. The second term is a linear constraint that weights parameters to their initial values, increasing stability of the fit by constraining p_i within $p_{0,i} \pm \delta p_i$. The value of Err for each constrained parameter was then passed to I-NoLLS as a calculated data point with target value zero.

Fitting took place by communicating a set of trial parameters from I-NoLLS to SIBFA via a series of wrapper scripts. Calculated energies using these parameters were then returned from SIBFA to I-NoLLS to evaluate either the current quality of the fit or J. J is evaluated numerically, requiring two evaluations of each data point per parameter (one for $p_i + \delta p_i$, one for $p_i - \delta p_i$).

Different adaptations of the I-NoLLS approach outlined above have already been successfully applied with CHARMM[52] and VALBOND-TRANS[51] for a number of different molecular mechanics and molecular dynamics fitting applications. The methodology is generally applicable to any fitting problem where it is computationally demanding to numerically evaluate the matrix of partial derivatives of data points with respect to parameters, J. Especially in nonlinear fitting scenarios where gradient-following methods are inefficient, human intervention can prevent erroneous steps in parameter space and reduce the number of costly evaluations of J. This is especially true where time-consuming molecular dynamics simulations are required to evaluate observable data but also for the current application where a sizeable number of complexes are combined with a large number of parameters.

Results

Multipole fitting

Before SIBFA parameter refinement could begin, the necessary calculable properties (multipole moments and polarizabilities) needed to be obtained. These were calculated directly for H_2O and imidazole, while multipolar fitting for formamide proved to be relatively linear, and converged rapidly to yield a

significant improvement in the description of the electrostatic potential around the formamide molecule (Fig. 5).

The figure focuses on the 10^{-3} a.u. isodensity surface, which lies near the chemically important region where H-bonding and ligand–metal interactions take place. Lying slightly inside the molecular volume, it is where values of the molecular electrostatic potential (MEP) are at their largest and the largest errors tend to be found. The improvement in reproducing the MEP when fitting the MTP coefficients is clearly visible, as there are marked reductions in the number of dark-blue areas of highest error ($\geq$16 kcal/mol). There is also a much more even distribution of error across the surface after fitting, and fewer of the red rings that correspond to transitions between overestimation and underestimation of the MEP. The evenly distributed residual error of around 8 kcal/mol can additionally be more easily corrected using SIBFA's penetration energy term.

Beyond this isodensity surface, the mean absolute error (MAE) across all grid points lying between the 10^{-3} a.u. isodensity surface and another surface a factor of 1.6 times further from the molecule[53] is reduced from 1.1 to 0.5 kcal/mol. In the region between this surface and a surface 2.2 times the distance from the molecule the MAE is reduced from 0.2 to 0.04 kcal/mol. In the long-range region lying beyond this surface the MAE is reduced from 0.06 to 0.02 kcal/mol, demonstrating that improvement in the description of the MEP is found across the whole volume surrounding the molecule and not just the region closest to it. As penetration energy corrections in SIBFA apply to short-range interactions only, satisfactory performance of the multipole expansion for long-range interactions is particularly important. The refined formamide multipole moments were therefore adopted for use in this work.

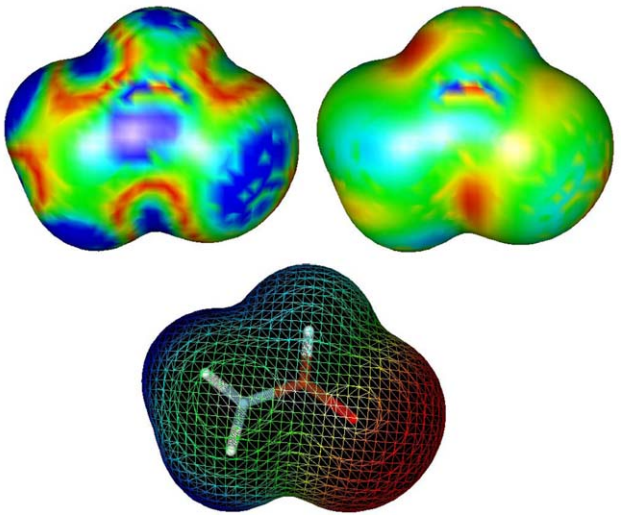

Figure 5. Figure comparing the absolute error in the molecular electrostatic potential (MEP) across the 10^{-3} a.u. isodensity surface of formamide using original GDMA multipole moments (top-left) and I-NoLLS-refined GDMA moments (top-right). Colors correspond to 0.0 (red), 4.0 (yellow), 8.0 (green), 12.0 (cyan), and $\geq$ 16.0 (blue) kcal/mol. The lower image shows the RHF/aug-ccpVTZ reference MEP mapped onto the same surface, with the molecular orientation visible within. Colors range from −0.06 a.u. (red) to 0.06 a.u. (blue). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

SIBFA parameter fitting

An unsupervised fit (with all parameters free to vary) was initially attempted, but the highly correlated nature of the parameters made this approach unfeasible. Nonphysical parameter values were quickly reached and the complex, highly nonlinear relationship between parameter values and RVS energies resulted in unsuitable parameter combinations and unstable energy evaluations in SIBFA.

Fits to a single subset of the data, for example, starting with water dimers alone, were found to be highly successful in reducing the total error but suffered greatly from poor transferability. Over-fitting for the chemically restricted training set, despite using soft constraints to bias parameters towards their initial values, produced parameters that were not compatible with other fragment types or systems outside the training set.

The most successful strategy was to include a diverse training set from the outset and rigidly imposing carefully chosen constraints. It was found that parameters from the original, SBK parametrization of SIBFA (henceforth parametrization 1, P1) performed reasonably well in many cases and needed only minimal adaptation to describe the new reference data. Parameters were therefore biased to remain close to their SBK-fitted values using soft constraints. To allow parameters to adjust relatively freely within a short range around their initial values, while avoiding problems due to correlated and anticorrelated parameters that tend to push pairs of parameters far from their initial values, pseudohard constraints were applied that only started to take effect at the edge of an allowed range [eq. (12)]. In general, the same constraints were applied to all parameters, but background knowledge could be used to adapt the tolerances for a given parameter where appropriate. Some degree of normalization of complex energies, so that fitting did not focus entirely on complexes with the largest energies and therefore largest absolute errors, was achieved using the "experimental error" term in I-NoLLS that is translated into a weighting-matrix during calculation of the next step in parameter space.[48] Fitting then took place for the entire training set consisting of 1085 data points simultaneously, comprising the electrostatic, exchange repulsion, polarization, and charge-transfer terms for each complex. The total binding energy was not used for fitting purposes to avoid the danger of I-NoLLS deliberately compensating one energy term against another to reduce the total error. Such compensation was considered to constitute "over-fitting" and may harm transferability to systems outside the training set. This approach was used to fit "P3."

For comparison purposes, a manually fitted parameter set (henceforth P2) was also prepared using a subset of the same training data. A subset was chosen as it quickly becomes impractical to use a large training set while fitting in this way. A second set of I-NoLLS-fitted parameters (P4) was subsequently obtained by starting from the manually fitted parameters, and constraining around these values instead of SBK-fitted values.

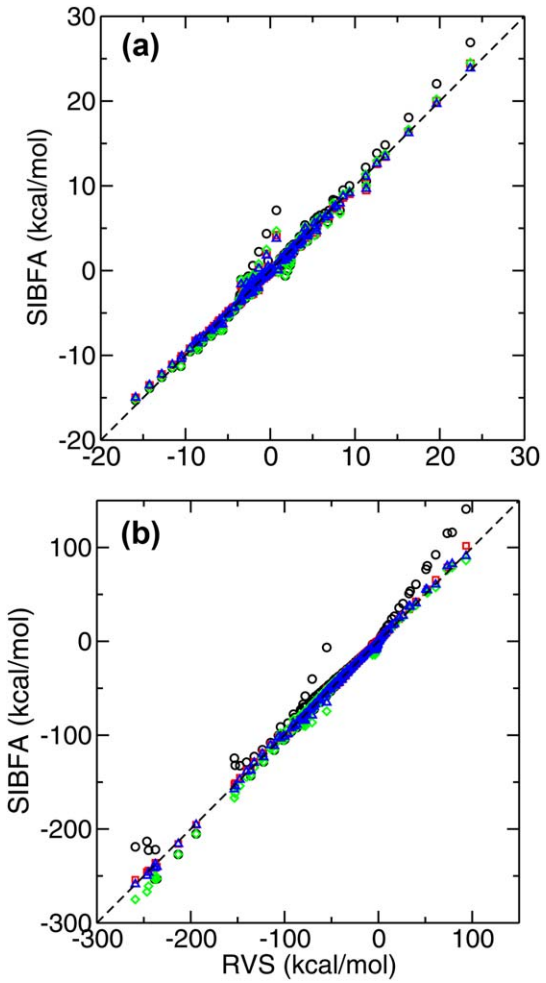

Figure 6. SIBFA vs. RVS energies for the series of water dimers (top) and $[Mg(H_2O)_n]^{2+}$ complexes (bottom) used in the training set. Parameters used in SIBFA were from P1 (black circles), P2 (green diamonds), P3 (red squares), and P4 (blue triangles). All energy components are included (electrostatic, repulsion, polarization, charge transfer, and total binding energy). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The protocol used to fit P3 and P4 using I-NoLLS is given in detail as Supporting Information (Section 1.1), along with the optimized parameter values. Fitting was halted when no further reduction in the total error was possible after recalculation of J, as evaluated using the variance (σ^2) reported in I-NoLLS.

Water Complexes. Although fitting of all training set complexes took place simultaneously in I-NoLLS, results will be broken down and discussed individually here. The training set of 17 water dimers (44 conformations including scanned coordinates) shown in Figure 1 was chosen from the work of Tschumper et al.[54] to include a broad range of monomer–monomer interactions.

H_2O parameters are not fitted to bulk water properties in SIBFA, as a versatile model is required that is applicable to bulk, interfacial, and binding-site water molecules. Second-order polarization and charge-transfer terms allow water molecules to adapt to their environment in a way that nonpolarizable force fields cannot, making a single set of parameters applicable to many different environments. The lack of further higher order response terms such as hyperpolarizability and structural relaxation, however, mean that the final potential may not be as finely tuned to describe bulk properties as potentials designed specifically for this purpose.[55] The approach has been applied to various static energetic studies involving small numbers of monomers,[17,56] although validation of bulk properties through molecular dynamics simulation has yet to be performed.

The results of fitting are shown in Figure 6. A MAE of 0.44 kcal/mol was found for the whole set of water dimers using P1 (the original SBK-parametrization), in quite reasonable agreement with RVS data. P2 (manual fitting) reduced the MAE to 0.35 kcal/mol, while P3 (fitting from the SBK-parametrization using I-NoLLS) reduced the error still further to 0.24 kcal/mol. P4 (fitting with I-NoLLS with the manually fitted values as a starting point) yielded an MAE of 0.23 kcal/mol.

Fitting of the $[Mg(H_2O)_n]^{2+}$ complexes yielded larger MAEs due to the larger absolute binding energies involved (Fig. 6). Using P1 the MAE was 5.1 kcal/mol, while P2 reduced this to 3.0 kcal/mol. I-NoLLS fitting reduced the error noticeably to 1.1 kcal/mol for P3 and 1.5 kcal/mol for P4.

Formamide Complexes. P1 was also found to perform reasonably well for neutral formamide dimers extracted from the high-pressure crystal structure.[57] An MAE of 1.8 kcal/mol was obtained while Figure 7 shows the quality of the data before and after fitting. P2 reduced the MAE to 1.3 kcal/mol, while I-NoLLS reduced the MAE further to 0.8 kcal/mol for P3 and 0.9 kcal/mol for P4. Similarly to the water dimers, there are a few notable outliers that are poorly described using P1 but which are well described after I-NoLLS fitting.

The Mg-formamide complexes again lead to larger MAEs than the neutral dimers due to the larger energies involved (Fig. 7). This time P1 yields an MAE of 7.8 kcal/mol, while P2 to P4 improve to 3.9, 2.2, and 2.3 kcal/mol, respectively, (Table 3). Notable outliers are again found for P1, which are much better described after both hand- and I-NoLLS fitting.

Imidazole Complexes. P1 performed slightly better for describing imidazole dimers taken from the low-temperature crystal structure[58] than for the formamide dimers, with an MAE of 0.6 kcal/mol and noticeably fewer outliers. (Fig. 8). P2 reduced the MAE to 0.5 kcal/mol, while P3 and P4 both yielded an MAE of 0.4 kcal/mol.

The charged Mg-complexes yielded substantially larger errors, however, (Fig. 8) and notable outliers with P1 deliver an MAE of 11.5 kcal/mol. P2 left a non-negligible MAE of 4.9 kcal/mol, while I-NoLLS fitting was able to reduce this value to 2.4 kcal/mol in P3 and 2.5 kcal/mol for P4 (Table 4).

Validation set

Water Complexes. Validation began with the set of 5 $(H_2O)_n$ and 10 $[Mg(H_2O)_n]^{2+}$ complexes (including scans) shown in Figures 9 and 10. Water clusters were taken from the work of

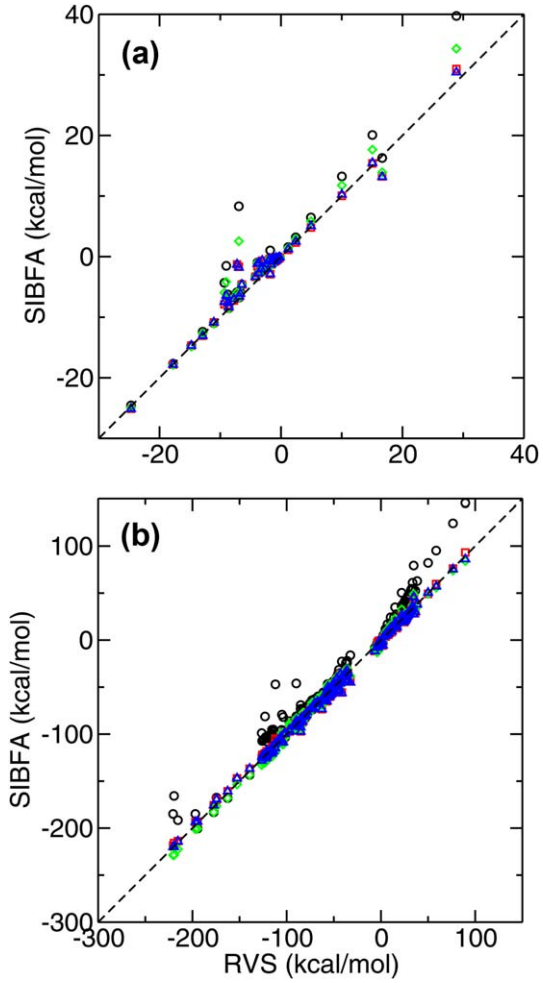

Figure 7. SIBFA vs. RVS energies for the series of formamide dimers (top) and Mg-formamide complexes (bottom) used in the training set. Parameters used in SIBFA were from P1 (black circles), P2 (green diamonds), P3 (red squares), and P4 (blue triangles). All energy components are included (electrostatic, repulsion, polarization, charge transfer, and total binding energy). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Wales and Hodges.[59] The larger size of these complexes with respect to those in the training set leads to increased steric interaction, and an accumulation of any ligand–metal or ligand–ligand errors that were present in the dimers. The performance of the different parametrizations when applied to this set are included graphically as Supporting Information (Section 2.1), with values listed for selected $(H_2O)_n$ complexes in Table 1 and for selected $[Mg(H_2O)_n]^{2+}$ complexes in Table 2. Complexes included in Table 2 are those with shortest Mg···O distance for each set shown in Figure 10, the planar tetraligated complex is a geometrical isomer of that used in the training set (Fig. 2). The improved performance of the fitted parameters P2–P4 over P1 is particularly visible in the $[Mg(H_2O)_n]^{2+}$ complexes, where the 13.7 kcal/mol MAE associated with the SBK-derived parameters, is reduced to 7.9 kcal/mol with P2 and just 1.8 kcal/mol for P3 and 2.3 kcal/mol for P4. It should be noted that P2 systematically overestimates the charge transfer energy relative to RVS reference data. As the P4 parameters are loosely constrained to P2, P4 also overestimates this term. P3 is loosely constrained to P1, yielding better performance. In addition, we note that even small differences in parametrization of a force field can change the description of data outside of a common training set.[60] The $(H_2O)_n$ clusters are well described using all of the parameter sets, including P1.

The final validation step was to randomly perturb the relative positions of all H_2O ligands of complex $[Mg(H_2O)_6]^{2+}$ relative to the central cation. 100 distinct conformers were generated, all lying within an energy range of −310 and −285 kcal/mol, with the exception of seven higher-energy outliers. This set offers an extensive sampling of the important dynamically accessible conformational space around the global minimum energy structure, with a few higher-energy structures that could be of interest in, for example, thermodynamic studies. The results are shown in Figure 11. This time systematic errors in the total complex binding energy are clearly

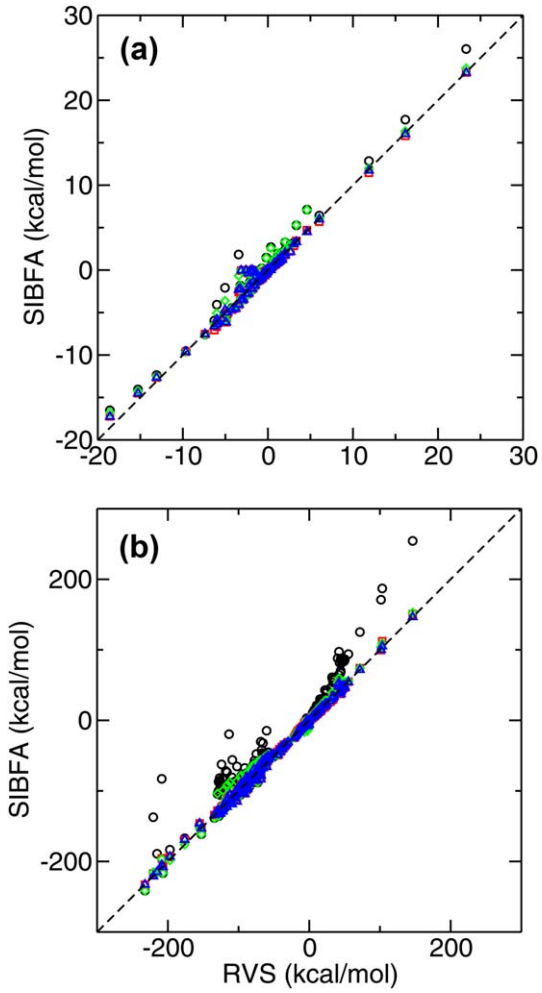

Figure 8. SIBFA vs. RVS energies for the series of imidazole dimers (top) and Mg-imidazole complexes (bottom) used in the training set. Parameters used in SIBFA were from P1 (black circles), P2 (green diamonds), P3 (red squares), and P4 (blue triangles). All energy components are included (electrostatic, repulsion, polarization, charge transfer, and total binding energy). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

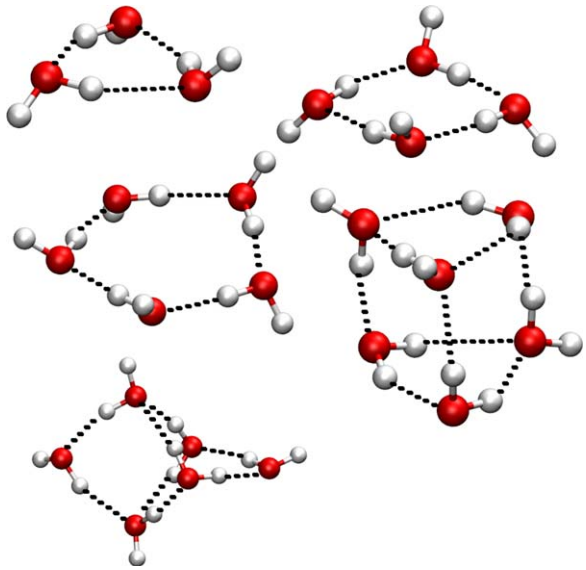

Figure 9. Figure showing the different water clusters used in the validation set. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

visible. P1 systematically and significantly underestimates the total binding energy, P2 overestimates to a smaller degree, P3 is consistently close to Restricted Hartree Fock (RHF) values and P4 has a tendency to underestimate the total energies. The degree of shift is reflected in the MAEs of 28.1, 22.3, 5.5, and 13.1 kcal/mol (of ca. 280 kcal/mol) for P1, P2, P3, and P4, respectively. This shift is important for quantitative comparison of binding energies, for example, between different complexes, and P3 clearly performs best. Such quantitative comparison of complex stability with different numbers and arrangements of ligands is not possible using simpler force fields due to the replacement of explicit ligand–metal interaction energies with harmonic bonded terms. Another measure of performance is the scatter of the points, which quantifies how reliably each

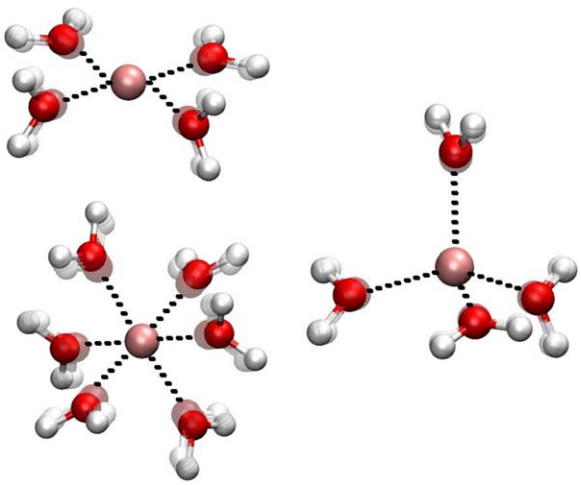

Figure 10. Figure showing the different $[Mg(H_2O)_n]^{2+}$ complexes used in the validation set. Scans of a particular coordinate are indicated by overlaying all sampled points as transparent images on top of the initial structure. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 1. Table comparing RVS energy components for $(H_2O)_n$ complexes with corresponding SIBFA terms using different parameter sets.

		E_{MTP}	E_{REP}	E_1	E_{CT}	E_{POL}	E_2	E_{TOT}
$(H_2O)_5$	P1	−47.7	40.7	−7.1	−3.1	−14.1	−17.2	−24.2
	P2	−47.7	38.4	−9.3	−3.1	−14.2	−17.3	−26.6
	P3	−46.1	36.8	−9.3	−5.3	−14.3	−19.6	−28.9
	P4	−46.0	37.6	−8.4	−4.7	−14.3	−19.1	−27.5
	RVS	−45.9	36.9	−9.1	−5.5	−12.4	−17.9	−27.0
$(H_2O)_6$ prism	P1	−55.0	40.6	−14.4	−3.1	−12.6	−15.8	−30.2
	P2	−55.0	39.0	−16.0	−3.1	−12.8	−15.9	−31.8
	P3	−53.1	36.7	−16.4	−5.5	−12.8	−18.4	−34.7
	P4	−53.0	38.1	−14.9	−4.9	−12.8	−17.8	−32.7
	RVS	−54.8	40.1	−14.7	−5.5	−11.3	−16.8	−31.5
$(H_2O)_6$ cage	P1	−55.8	43.5	−12.2	−3.3	−13.3	−16.6	−28.9
	P2	−55.8	41.7	−14.1	−3.3	−13.4	−16.7	−30.8
	P3	−53.8	39.4	−14.4	−5.8	−13.5	−19.4	−33.8
	P4	−53.8	40.8	−13.0	−5.3	−13.5	−18.9	−31.8
	RVS	−55.1	41.2	−13.9	−5.9	−11.9	−17.8	−31.6
$(H_2O)_n$ MAE	P1	0.9	2.2	1.5	2.5	1.5	2.0	1.6
	P2	0.9	1.0	1.0	2.5	1.6	2.0	1.1
	P3	1.1	1.8	1.4	0.1	1.7	0.9	1.2
	P4	1.1	1.1	1.1	0.6	1.7	1.2	0.9

The electrostatic (E_{MTP}), exchange repulsion (E_{REP}), charge transfer (E_{CT}), polarization after iteration to self-consistency (E_{POL}), total first (E_1) and second order (E_2), and total binding energies (E_{TOT}) are listed for P1–P4. Mean absolute errors for the individual energy components of each water cluster are included at the bottom of the table.

parameter set can predict relative energies for different conformers. A linear fit of RVS versus SIBFA energies was produced (the gradient of the fitted line was not constrained) and the correlation coefficient "r^2" of the data points relative to this line was evaluated. r^2 of 0.90, 0.97, 0.96, and 0.95 were obtained for P1, P2, P3, and P4, respectively. Again P1 is least

Table 2. Table comparing RVS energy components for $[Mg(H_2O)_n]^{2+}$ minimum-energy complexes with corresponding SIBFA terms using different parameter sets.

		E_{MTP}	E_{REP}	E_{CT}	E_{POL}	E_{TOT}
$[Mg(H_2O)_4]^{2+}$ planar	P1	−251.3	117.1	−3.2	−72.4	−209.8
	P2	−251.3	79.5	−14.4	−77.5	−263.8
	P3	−238.3	81.3	−3.7	−81.8	−242.4
	P4	−238.4	83.2	−9.6	−81.9	−246.6
	RVS	−234.2	79.3	−6.3	−82.8	−244.0
$[Mg(H_2O)_4]^{2+}$ pyramidal	P1	−244.0	101.5	−3.1	−76.1	−221.7
	P2	−244.0	67.1	−13.9	−81.3	−272.1
	P3	−232.6	69.1	−3.5	−85.7	−252.8
	P4	−232.7	71.1	−9.3	−85.8	−256.7
	RVS	−229.2	65.1	−5.6	−85.9	−255.5
$[Mg(H_2O)_6]^{2+}$ octahedral	P1	−374.6	188.9	−3.0	−72.2	−260.9
	P2	−374.6	132.8	−15.9	−77.9	−335.6
	P3	−351.0	133.8	−3.3	−82.7	−303.2
	P4	−351.2	137.3	−10.4	−82.8	−307.1
	RVS	−343.3	131.8	−8.1	−88.7	−308.3
$[Mg(H_2O)_n]^{2+}$ MAE	P1	16.9	30.3	2.8	9.5	13.7
	P2	16.9	2.4	7.6	5.3	7.9
	P3	3.1	2.0	2.3	1.4	1.8
	P4	3.2	5.4	3.1	1.4	2.3

The electrostatic (E_{MTP}), exchange repulsion (E_{REP}), charge transfer (E_{CT}), polarization after iteration to self-consistency (E_{POL}), and total binding energies (E_{TOT}) are listed for P1–P4. Mean absolute errors for each energy component over all $[Mg(H_2O)_n]^{2+}$ complexes in the validation set are included at the bottom of the table.

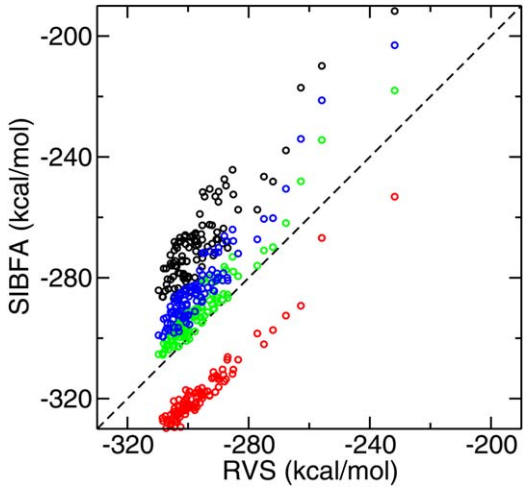

Figure 11. SIBFA vs RVS binding energies for the series of 100 $[Mg(H_2O)_6]^{2+}$ conformers used in the validation set. Parameters used in SIBFA were P1 (black circles), P2 (red circles), P3 (green circles), and P4 (blue circles). The black dashed line is included at $y = x$ to guide the eye. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

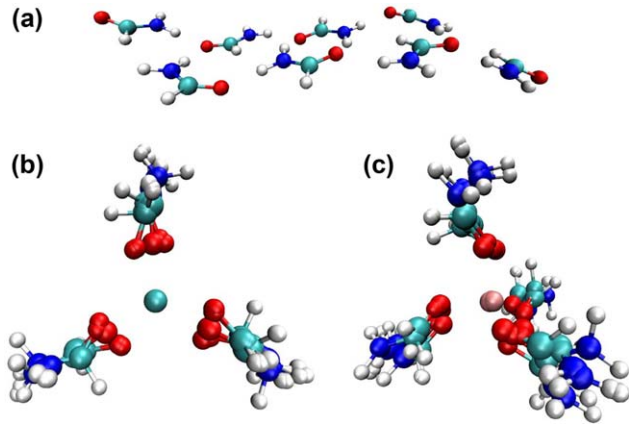

Figure 12. Figure showing a unit cell taken from the crystal structure and used for validation of P1-P4 against RVS data (top) and the results of geometry optimization of $[Mg(HCONH_2)_3]^{2+}$ (bottom-left) and $[Mg(HCONH_2)_4]^{2+}$ (bottom-right) using P1-P4. Structures are overlaid on the RHF/aug-cc-pVTZ optimized geometry. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

reliable and P3 is similar to P4 but Constrained Space Orbital Variations P2 now performs slightly better than P3 and P4. The relative energies of different conformers relative to the global minimum are particularly significant for molecular dynamics simulations, and allow the study of strained ligand arrangements in protein binding sites.

Formamide Complexes. A similar approach was taken for formamide with the crystal structure unit cell[57] shown at the top of Figure 12 and with data presented in Table 3. The very short monomer–monomer distances in the high-pressure structure led to an accumulated total error of roughly 10 kcal/mol for the total first- and second-order energy terms. Relaxing the monomer–monomer distances with a constrained HF optimization led to better agreement, with P1–P4 all within 6 kcal/mol of the first- and second-order energy terms. P3 and P4 both slightly overestimate the first-order energy, and underestimate the second-order energy, but benefit from a cancellation of errors to yield agreement with the total interaction energy. P2 performs better for the first-order energy but also underestimates the second-order terms, resulting in a larger error in the total energy, while P1 underestimates both terms.

The next validation step was to geometry optimize the $[Mg(HCONH_2)_n]^{2+}$ complexes shown in the bottom half of Figure 12 using each parameter set, but keeping monomer

Table 3. Table comparing RVS energy components for the $(HCONH_2)_8$ crystal structure unit cell and the $[Mg(HCONH_2)_3]^{2+}$ and $[Mg(HCONH_2)_4]^{2+}$ RHF minimum-energy structures with corresponding SIBFA terms using different parameter sets.

		E_{MTP}	E_{CT}	E_1	E_{CT}	E_{POL}	E_2	E_{tot}
$(HCONH_2)_8$	P1	−76.6	53.4	−23.2	−1.7	−20.0	−21.8	−45.0
	P2	−77.7	47.6	−30.1	−1.7	−20.1	−21.8	−51.8
	P3	−76.8	41.0	−35.8	−3.1	−20.1	−23.2	−59.0
	P4	−76.9	42.9	−34.0	−2.1	−20.1	−22.2	−56.2
	RVS	–	–	−29.6	–	–	−28.2	−57.8
$[Mg(HCONH_2)_3]^{2+}$	P1	−268.4	122.8	−145.6	−2.9	−100.5	−103.4	−249.0
	P2	−268.4	73.4	−194.9	−13.5	−103.0	−116.5	−311.5
	P3	−257.0	74.6	−182.4	−3.3	−110.3	−113.6	−296.0
	P4	−256.9	75.4	−181.5	−7.0	−110.3	−117.3	−298.8
	RVS	−260.3	71.9	−189.6	−2.8	−117.9	−120.7	−310.3
$[Mg(HCONH_2)_4]^{2+}$	P1	−332.4	134.8	−197.6	−2.7	−99.0	−101.7	−299.3
	P2	−332.4	82.1	−250.3	−13.7	−101.6	−115.3	−365.6
	P3	−317.7	80.1	−237.6	−3.0	−109.0	−112.0	−349.6
	P4	−317.5	83.6	−233.9	−7.0	−109.0	−116.0	−349.9
	RVS	–	–	−238.7	–	–	−120.5	−359.3
$[Mg(HCONH_2)_n]^{2+}$ MAE	P1	2.7	15.7	9.2	0.5	5.1	2.8	7.8
	P2	2.7	2.4	2.6	4.9	4.3	4.6	3.9
	P3	1.8	2.2	2.0	0.6	2.5	1.6	2.2
	P4	1.9	2.1	2.0	1.7	2.5	2.1	2.3

The electrostatic (E_{MTP}), exchange repulsion (E_{REP}), charge transfer (E_{REP}), polarization after iteration to self-consistency (E_{POL}), and total binding energies (E_{TOT}) are listed for P1–P4. Only total first-order (E_1) and second-order (E_2) energies are available as RVS reference data for the larger systems. The mean absolute error of each energy component for the training set $[Mg(HCONH_2)]^{2+}$ and $[Mg(HCONH_2)_2]^{2+}$ complexes using P1–P4 are shown at the bottom of the table.

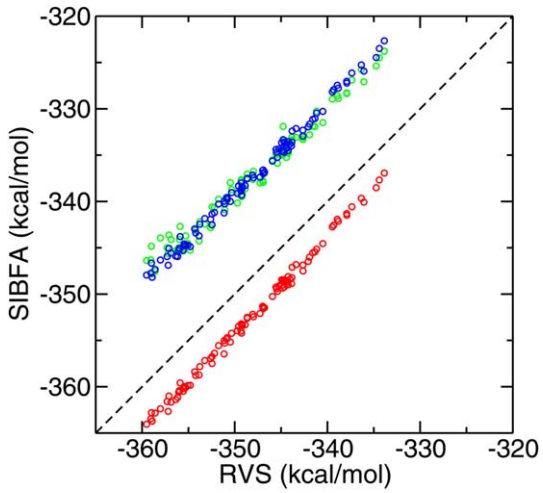

Figure 13. SIBFA vs RVS binding energies for the series of 100 $[Mg(CONH)_4]^{2+}$ conformers used in the validation set. Parameters used in SIBFA were P2 (red circles), P3 (green circles) and P4 (blue circles). The black dashed line is included at $y = x$ to guide the eye. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

geometries rigid. As seen in the figure, all parameter sets were able to find a similar geometry to the RHF reference structures, although with some perturbation of the precise orientation of the ligands. This is encouraging, as parameters were not directly fitted to geometrical data, and there are no bonded or angular terms in SIBFA to maintain a particular geometrical arrangement of the atoms. This is in contrast to approaches such as VALBOND which treat metal–ligand interactions as chemical bonds.[15] In the case of $[Mg(HCONH_2)_3]^{2+}$ the best description comes from P2, with an optimized geometry differing by an RMSD of just 0.3 Å from the RHF reference. A similar level of accuracy (RMSD = 0.4 Å) is found using P4. P1 also performs well, with an RMSD of 0.5 Å, however P3, the parameters fitted with I-NoLLS from SBK values lead to a larger error, with an RMSD of 1.1 Å. Closer inspection reveals that this parameter set leads to a structure with all ligand and metal atoms lying in a plane, whereas the reference structure shows the ligands should be rotated somewhat. Significantly, the energy surface is relatively shallow around the RHF reference structure, and the planar geometry is just 5 kcal/mol higher in energy at the RHF/aug-cc-pVTZ level of theory than the RHF-optimized reference structure. Further investigation revealed that the hydrogen atom attached to the carbonyl carbon of formamide is not well represented in the training data. Refinement of these parameters led to further improvement in polyligated and crystal structure complexes including the octamer, where CH groups come into contact with neighboring molecules, and significantly improved geometries for P3 with $[Mg(HCONH_2)_n]^{2+}$ in particular.

In the $[Mg(HCONH_2)_4]^{2+}$ complex, perhaps surprisingly, the closest match to the RHF reference structure is obtained using the original SBK parametrization, which yields an RMSD of just 0.2 Å. The remaining parameter sets yield larger errors, with P2 producing a geometry that deviated by an RMSD of 1.1 Å from RHF reference data, P3 deviating by an RMSD of 1.1 Å, and P4 deviating by an RMSD of 1.4 Å.

The energetic description of the RHF-optimized minimum energy structure of each complex using each parameter set is shown in Table 3. Due to the quite large size of the systems, again only the total first- and second-order energies were obtained by energy decomposition for the tetra-ligated complex using the procedure outlined in eq. (7). Apart from P1, all parameter sets yield quite reasonable agreement with energy decomposition data. P2 performs best in both cases for the total binding energy, although there appears to be some cancellation of errors with a somewhat overestimated charge-transfer energy balancing a slightly underestimated polarization energy term, and an overestimated first-order energy balancing an underestimated second-order contribution. The I-NoLLS-fitted parameter sets both give good estimates for the first-order energy, and slightly underestimate the second-order energy, leaving them just below the reference values in the triligated and tetra-ligated complexes. In both cases, it is encouraging that all fitted parameter sets lead to significantly better agreement with RVS validation data than the original P1.

The seemingly strong performance of P1 when describing complex geometries despite a poor energetic description can be explained by the original RHF/SBK level of theory that was used for parametrizion of P1. Although RHF/aug-cc-pVTZ and RHF/SBK do not agree well on total energies or on corresponding RVS energy components, they tend to agree much better on structural data. As such, P1 is able to recover the correct aug-cc-pVTZ geometry but not the corresponding aug-cc-pVTZ total binding energy or RVS energy components.

The final validation step again involved evaluating the total binding energy of 100 conformers of $[Mg(HCONH_2)_4]^{2+}$, generated by making random perturbations to the relative positions of the ligands. As seen in Figure 13, P2, P3, and P4 all give excellent relative energies of the different conformers (r^2 of 0.998, 0.988, and 0.996, respectively), suggesting that they all describe the energy surface around the global minimum-energy conformation well. As with the H_2O complexes, different systematic shifts in the total binding energy of the complex are again visible, this time with P2 exhibiting the smallest shift (MAE 4.2 kcal/mol) and P3 and P4 performing similarly to one another (MAE 10.7 and 10.8 kcal/mol, respectively). In the case of formamide P1 does not perform well in either the absolute binding energy (MAE 45.6 kcal/mol) or the relative energies ($r^2 = -0.52$) of the different conformers, with data shifted outside the range of the figure.

Imidazole Complexes. Next, the crystal structure unit cell of imidazole[58] (top of Fig. 14) was modeled using each parameter set, with data presented in Table 4. Better results are achieved than for the formamide crystal structure, in part due to the smaller number of monomers included. Particularly good results are obtained using P3 and P4, with P2 slightly overestimating the repulsion energy and P1 overestimating this term still further.

A further geometry-optimization study was performed for the $[Mg(C_3N_2H_4)_3]^{2+}$ and $[Mg(C_3N_2H_4)_4]^{2+}$ complexes shown in Figure 14. As is clearly visible in the figure, all parameter sets perform well in optimizing the geometry of $[Mg(C_3N_2H_4)_3]^{2+}$.

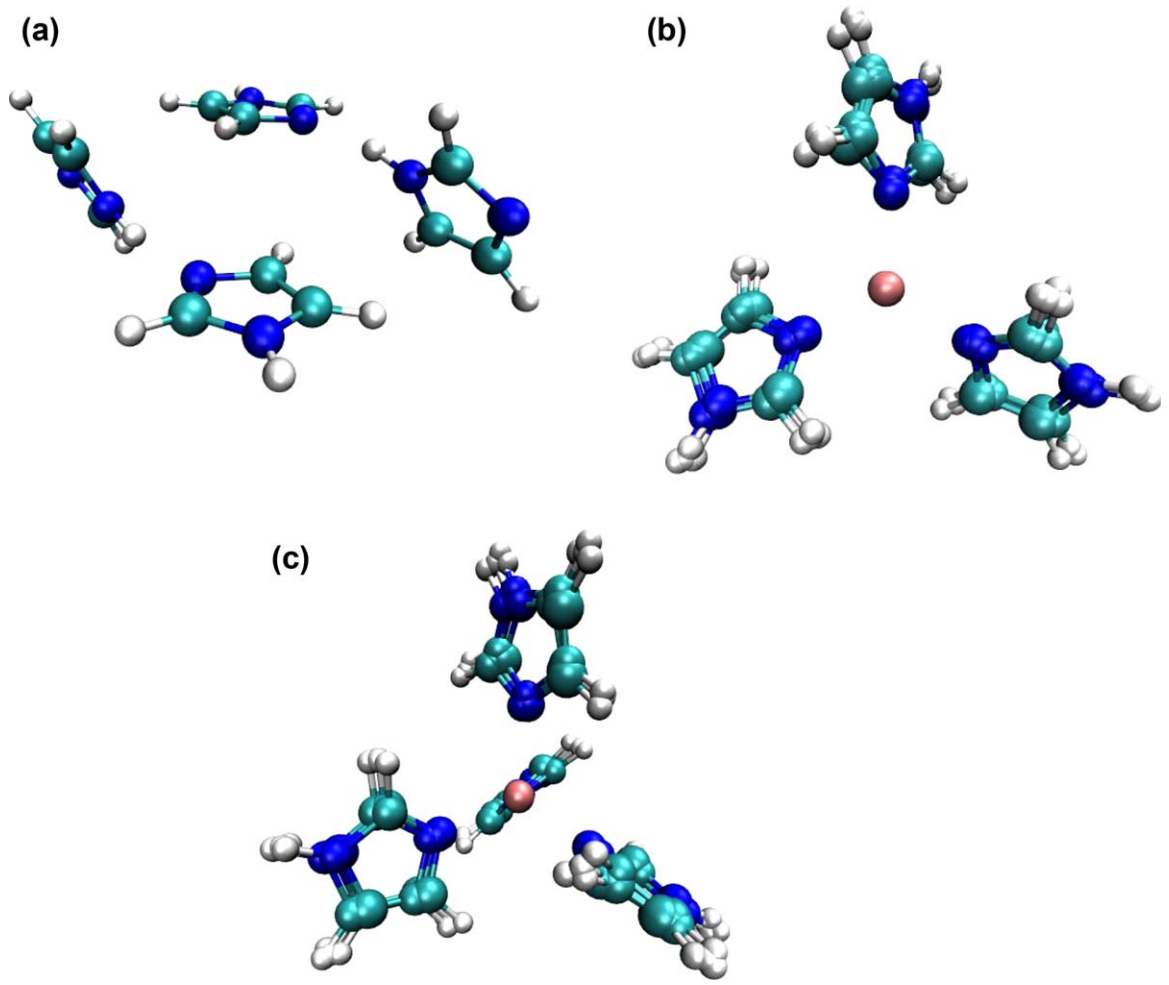

Figure 14. Figure showing the crystal structure unit cell used for comparison of the performance P1–P4 against RVS data (above) and the results of geometry optimization of $[Mg(C_3N_2H_4)_3]^{2+}$ (bottom-left) and $[Mg(C_3N_2H_4)_4]^{2+}$ (bottom-right) using P1-P4. Structures are overlaid on the RHF/aug-cc-pVTZ optimized geometry. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

P1 yields the largest RMSD, at 0.33 Å, while P3 and P4 yield a slightly improved RMSD of 0.29 Å, and the lowest RMSD of 0.26 Å is achieved using P2. A greater range in performance is again visible in the energetic description of the complex, however. P1 yields substantially different energetic components to the refitted parameter sets, leading to large errors in the first-order, second-order, and total interaction energies (Table 4). The much better geometric than energetic performance is again likely to be due to the structural similarity between the RHF/SBK- and RHF/aug-cc-pVTZ-optimized geometries. In contrast, there is good agreement between P2, P3, P4, and RVS energy-decomposition data for the first-order interaction energy, although there is some discrepancy in the distribution between the electrostatic and repulsion terms. P2, P3, and P4 also yield good agreement with the second-order interaction energy, with the largest discrepancy arising from P2. As a result both P3 and P4 are within 5 kcal/mol of the RHF total interaction energy, while P2 underestimates by around 10 kcal/mol.

$[Mg(C_3N_2H_4)_4]^{2+}$ geometries were similarly well predicted, with RMSDs of 0.29 Å for P1, 0.22 Å for P2, 0.38 Å for P3, and 0.42 Å for P4. The first-order energies were in good agreement with RVS values for all parameter sets except P1, although there was a tendency to underestimate the second-order (largely polarization) contribution by around 10–15 kcal/mol. A cancellation of errors leads to total errors also of around 10–15 kcal/mol (of −353 kcal/mol), with the best results from model P3.

Finally, a study of 100 randomly generated conformers of $[Mg(C_3N_2H_4)_4]^{2+}$ (Fig. 15) reveals that again P2, P3, and P4 are able to describe the conformational space around the global minimum quite satisfactorily while P1 fails both in describing the total binding energy (MAE 55.9 kcal/mol) and relative conformer energies ($r^2 = –0.54$). P2 again has the smallest absolute shift in estimated total binding energy of each conformer (MAE 5.5 kcal/mol of circa 335 kcal/mol), with P3 performing similarly well (MAE 7.6 kcal/mol) and P4 with a slightly larger systematic shift (MAE 10.2 kcal/mol). The relative energies of the conformers are best described using P4 ($r^2 = 0.991$), however, or P3 ($r^2 = 0.987$) with a noticeably broader distribution associated with P2 ($r^2 = 0.937$).

Conclusions

A supervised approach to force field parameter fitting has been introduced and applied to the SIBFA polarizable force

Table 4. Table comparing RVS energy components for $[Mg(C_3N_2H_4)_3]^{2+}$ and $[Mg(C_3N_2H_4)_4]^{2+}$ with corresponding SIBFA terms using different parameter sets.

		E_{MTP}	E_{CT}	E_1	E_{CT}	E_{POL}	E_2	E_{tot}
$(C_3N_2H_4)_4$	P1	−37.0	49.0	12.0	−3.1	−13.3	−16.4	−4.3
	P2	−37.5	46.2	8.8	−3.1	−13.3	−16.4	−7.6
	P3	−37.9	40.8	2.9	−4.2	−13.7	−17.9	−15.0
	P4	−37.9	40.9	3.0	−3.6	−13.7	−17.3	−14.3
	RVS	–	–	3.3	–	–	−16.3	−13.0
$[Mg(C_3N_2H_4)_3]^{2+}$	P1	−277.2	154.3	−122.9	−3.1	−100.3	−103.3	−226.2
	P2	−277.2	96.0	−181.2	−14.3	−99.8	−114.2	−295.4
	P3	−264.7	82.5	−182.2	−2.7	−118.1	−120.9	−303.1
	P4	−264.7	80.9	−183.9	−7.2	−118.1	−125.3	−309.2
	RVS	–	–	−174.2	–	–	−131.0	−305.3
$[Mg(C_3N_2H_4)_4]^{2+}$	P1	−345.7	178.1	−167.7	−2.7	−96.9	−99.6	−267.3
	P2	−345.7	114.2	−231.5	−14.0	−96.7	−110.8	−342.3
	P3	−330.1	101.9	−228.2	−2.4	−113.6	−115.9	−344.2
	P4	−330.1	109.7	−220.4	−7.0	−113.5	−120.5	−340.9
	RVS	–	–	−219.5	–	–	−133.7	−353.3
$[Mg(C_3N_2H_4)_n]^{2+}$ MAE	P1	3.8	22.7	13.2	1.0	7.4	4.2	11.5
	P2	3.8	2.9	3.4	4.5	7.7	6.1	4.9
	P3	1.6	2.4	2.0	1.0	2.9	2.0	2.4
	P4	1.7	1.7	1.7	1.7	2.9	2.3	2.5

The electrostatic (E_{MTP}), exchange repulsion (E_{REP}), charge transfer (E_{REP}), polarization after iteration to self-consistency (E_{POL}), and total binding energies (E_{TOT}) are listed for P1–P4. Only total first-order (E_1) and second-order (E_2) energies are available as RVS reference data due to the size of the systems. Mean absolute errors for each energy component are shown for the $[Mg(C_3N_2H_4)]^{2+}$ and $[Mg(C_3N_2H_2)_4]^{2+}$ complexes of the training set at the bottom of the table.

field. The approach will form the basis of a procedure to obtain libraries of SIBFA parameters to describe a wide range of metal complexes and biomolecules. The semiautomated nature of the tools additionally makes them suitable for end-users with little previous experience of the force field, who require missing parameters for particular applications.

Parameters were fitted for a series of H_2O, formamide, and imidazole complexes with Mg^{2+} that yield significantly reduced errors for aug-cc-pVTZ training data when compared to the original SBK parametrization of SIBFA, with a reduction in MAE over all energy components of all complexes in the training sets from 6.0 kcal/mol with the initial parameters to 1.5 kcal/mol in the case of the I-NoLLS-fitted parameter set P3 and to 1.6 kcal/mol for the I-NoLLS parameter set P4. P3 and P4 also offer some improvement over manual fitting, where a MAE of 2.9 kcal/mol was obtained for the training data. Total binding energies of different complexes, not available using simpler force field models that treat metal–ligand interactions with bonded terms, are generally well estimated using all of the fitted parameter sets for both the training sets and subsequent validation work. Relative energies of different conformers of a single complex are described better still, as evident from geometry-optimization and conformational surface sampling around global minima. The validation data confirms a marked improvement of I-NoLLS-fitted parameters with respect to the initial SBK parameter set, and a favorable comparison with manual fitting. Furthermore, fitting to a small set of carefully chosen interactions has proven sufficient to provide parameters that are transferable to new clusters, complexes and conformers well outside the original training data. For greater precision and transferability, the training set can be augmented further, for example, by including randomly generated dimer orientations. Overall P3, fitted from the original SBK parameters using I-NoLLS, yielded the most promising results and will be adopted as the starting point for parametrization of further fragments in SIBFA, building a robust library for a wide range of applications.

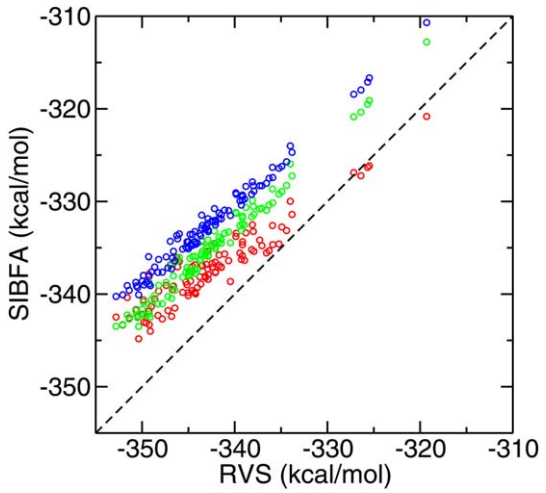

Figure 15. SIBFA vs. RVS binding energies for the series of 100 $[Mg(C_3N_2H_4)_4]^{2+}$ conformers used in the validation set. Parameters used in SIBFA were P2 (red circles), P3 (green circles), and P4 (blue circles). The black dashed line is included at $y = x$ to guide the eye. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The supervised, interactive fitting procedure is a promising compromise between fully automated fitting, which was found to be unstable for the highly nonlinear and correlated problem of force field parameter optimization, and manual fitting of parameters, which quickly becomes prohibitively time-

consuming and requires significant experience and expertise. Such work opens new directions to establish a robust parametrization of the approach for proteins, DNA, and RNA.

Finally, a recently developed approach for multipole refinement has been adapted and applied to create an improved electrostatic description of formamide. The approach gives a systematic means to obtain distributed multipole moments with superior convergence properties for future application of SIBFA.

Keywords: SIBFA · polarizable force field · parametrization · parameter fitting · I-NoLLS

How to cite this article: M. Devereux, N. Gresh, J.-P. Piquemal, M. Meuwly *J. Comput. Chem.* **2014**, *35*, 1577–1591. DOI: 10.1002/jcc.23661

Additional Supporting Information may be found in the online version of this article.

[1] D. M. Root, C. R. Landis, T. Cleveland, *J. Am. Chem. Soc.* **1993**, *115*, 4201.
[2] P. Y. Ren, J. Y. Ponder, *J. Phys. Chem. B* **2003**, *107*, 5933.
[3] M. J. L. Mills, P. L. A. Popelier, *Theor. Chem. Acc.* **2012**, *131*, 1137.
[4] N. Plattner, M. Meuwly, *Biophys. J.* **2008**, *94*, 2505.
[5] J.-P. Piquemal, G. A. Cisneros, P. Reinhardt, N. Gresh, T. A. Darden, *J. Chem. Phys.* **2006**, *124*, 104101.
[6] N. Gresh, G. A. Cisneros, T. A. Darden, J.-P. Piquemal, *J. Chem. Theory Comput.* **2007**, *3*, 1960.
[7] N. Gresh, P. Claverie, A. Pullman, *Theor. Chim. Acta* **1984**, *66*, 1.
[8] J.-P. Piquemal, B. Williams-Hubbard, N. Fey, R. J. Deeth, N. Gresh, C. Giessner-Prettre, *J. Comput. Chem.* **2003**, *24*, 1963.
[9] J.-P. Piquemal, H. Chevreau, N. Gresh, *J. Chem. Theory Comput.* **2007**, *3*, 824.
[10] J.-P. Piquemal, N. Gresh, C. Giessner-Prettre, *J. Phys. Chem. A* **2003**, *107*, 10353.
[11] N. Gresh, C. Policar, C. Giessner-Prettre, *J. Phys. Chem. A* **2002**, *106*, 5660.
[12] N. Gresh, J.-P. Piquemal, M. Krauss, *J. Comput. Chem.* **2005**, *26*, 1113.
[13] N. Gresh, B. de Courcy, J.-P. Piquemal, S. Courtiol-Legourd, L. Salmon, *J. Phys. Chem. B* **2011**, *115*, 8304.
[14] B. de Courcy, J.-P. Piquemal, N. Gresh, *J. Chem. Theory Comput.* **2008**, *4*, 1659.
[15] I. Tubert-Brohman, M. Schmid, M. Meuwly, *J. Chem. Theory Comput.* **2009**, *5*, 530.
[16] W. Stevens, M. Krauss, H. Basch, P. Jasien, *Can. J. Chem.* **1992**, *70*, 612.
[17] B. de Courcy, J.-P. Piquemal, C. Garbay, N. Gresh, *J. Am. Chem. Soc.* **2010**, *132*, 3312.
[18] W. J. Stevens, W. Fink, *Chem. Phys. Lett.* **1987**, *139*, 15.
[19] B. Jeziorski, R. Moszyński, K. Szalewicz, *Chem. Rev.* **1994**, *94*, 1887.
[20] P. S. Bagus, K. Hermann, C. W. Bauschlicher, Jr., *J. Chem. Phys.* **1984**, *80*, 4378.
[21] D. G. Fedorov, K. Kitaura, *J. Comput. Chem.* **2007**, *28*, 222.
[22] T. Bereau, C. Kramer, F. W. Monnard, E. S. Nogueira, T. R. Ward, M. Meuwly, *J. Phys. Chem. B* **2007**, *117*, 5460.
[23] N. Plattner, M. Meuwly, *Biophys. J.* **2008**, *94*, 2505.
[24] M. Devereux, M.-C. van Severen, O. Parisel, J.-P. Piquemal, N. Gresh, *J. Chem. Theory Comput.* **2011**, *7*, 138.
[25] M. Shaik, M. Devereux, P. Popelier, *Mol. Phys.* **2008**, *106*, 1495.
[26] P. G. Karamertzanis, S. L. Price, *J. Chem. Theory Comput.* **2006**, *2*, 1184.
[27] M. Devereux, N. Plattner, M. Meuwly, *J. Phys. Chem. A* **2009**, *113*, 13199.
[28] X. Zheng, C. Wu, J. W. Ponder, G. R. Marshall, *J. Am. Chem. Soc.* **2012**, *134*, 15970.
[29] M. W. Lee, M. Meuwly, *J. Phys. Chem. A* **2011**, *115*, 5053.
[30] M. W. Lee, J. K. Carr, M. Goellner, P. Hamm, M. Meuwly, *J. Chem. Phys.* **2013**, *139*, 54506.
[31] M. W. Lee, M. Meuwly, *PCCP* **2013**, *15*, 20303.
[32] F. Vigné-Maeder, P. Claverie, *J. Chem. Phys.* **1988**, *88*, 4934.
[33] A. Stone, *J. Chem. Theory Comput.* **2005**, *1*, 1128.
[34] D. Kosov, P. Popelier, *J. Chem. Phys.* **2000**, *113*, 3969.
[35] L. Joubert, P. Popelier, *Mol. Phys.* **2002**, *100*, 3357.
[36] C. M. Baker, V. M. Anisimov, A. D. MacKerell, Jr., *J. Phys. Chem. B* **2011**, *115*, 580.
[37] M. Schmidt, K. Baldridge, J. Boatz, S. Elbert, M. Gordon, J. Jensen, S. Koseki, N. Matsunaga, K. Nguyen, S. Su, T. L. Windus, M. Dupuis, J. A. Montgomery Jr., *J. Comput. Chem.* **1993**, *14*, 1347.
[38] K. Schuchardt, B. Didier, T. Elsethagen, L. Sun, V. Gurumoorthi, J. Chase, J. Li, T. Windus, *J. Chem. Inf. Model.* **2007**, *47*, 1045.
[39] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, D. J. Fox, Gaussian 09, Revision A.02; Gaussian, Inc.: Wallingford, CT, **2009**.
[40] A. Szabo, N. S. Ostlund, Modern Quantum Chemistry; Dover Publications, Inc.: Mineola, New York, **1996**.
[41] J. M. Foster, S. F. Boys, *Rev. Mod. Phys.* **1960**, *32*, 300.
[42] S. F. Boys, *Rev. Mod. Phys.* **1960**, *32*, 296.
[43] C. Kramer, P. Gedeck, M. Meuwly, *J. Comput. Chem.* **2012**, *33*, 1673.
[44] T. Bereau, C. Kramer, M. Meuwly, *J. Chem. Theory Comput.* **2013**, *9*, 5450.
[45] C. Kramer, P. Gedeck, M. Meuwly, *J. Chem. Theory Comput.* **2013**, *9*, 1499.
[46] C. I. Bayly, P. Cieplak, W. Cornell, P. A. Kollman, *J. Phys. Chem.* **1993**, *40*, 10269.
[47] J. C. Wu, G. Chattree, P. Ren, *Theor. Chem. Acc.* **2012**, *131*, 1138.
[48] M. M. Law, J. M. Hutson, *Comput. Phys. Commun.* **1997**, *102*, 252.
[49] C. Breneman, K. Wiberg, *J. Comput. Chem.* **1990**, *11*, 361.
[50] N. Gresh, D. Garmer, *J. Comput. Chem.* **1996**, *17*, 1481.
[51] F. Hofmann, M. Devereux, A. Pfaltz, M. Meuwly, *J. Comput. Chem.* **2014**, *35*, 18.
[52] M. Devereux, M. Meuwly, *J. Chem. Inf. Model.* **2010**, *50*, 349.
[53] F. M. Richards, B. Lee, *J. Mol. Biol.* **1971**, *55*, 379.
[54] G. S. Tschumper, M. L. Leininger, B. C. Hoffman, E. F. Valeev, H. F. Schaefer, M. Quack, *J. Chem. Phys.* **2002**, *116*, 690.
[55] L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, V. S. Pande, *J. Phys. Chem. B.* **2013**, *117*, 9956.
[56] N. Gresh, *J. Phys. Chem. A* **1997**, *101*, 8680.
[57] R. Gajda, A. Katrusiak, *Cryst. Growth Des.* **2011**, *11*, 4768.
[58] R. K. McMullan, J. Epstein, J. R. Ruble, B. M. Craven, *Acta Crystallogr.* **1979**, *B35*, 688.
[59] D. J. Wales, M. P. Hodges, *Chem. Phys. Lett.* **1998**, *286*, 65.
[60] P. Cazade, T. Bereau, M. Meuwly, *J. Phys. Chem.* **2014**, DOI: 10.1021/jp5011692.

Received: 23 January 2014
Revised: 14 April 2014
Accepted: 25 May 2014
Published online on Wiley Online Library