JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

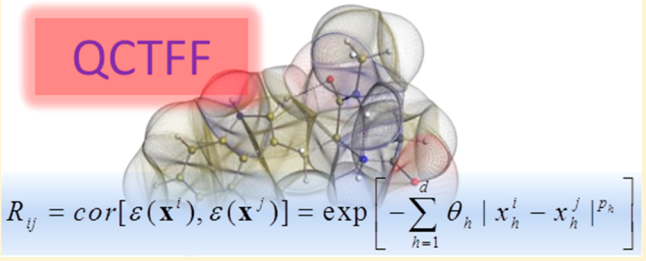
Prediction of Intramolecular Polarization of Aromatic Amino Acids Using Kriging Machine Learning

Timothy L. Fletcher, Stuart J. Davie, and Paul L. A. Popelier*

Manchester Institute of Biotechnology (MIB), 131 Princess Street, Manchester, M1 7DN, Great Britain

School of Chemistry, University of Manchester, Oxford Road, Manchester, M13 9PL, Great Britain

ABSTRACT: Present computing power enables novel ways of modeling polarization. Here we show that the machine learning method kriging accurately captures the way the electron density of a topological atom responds to a change in the positions of the surrounding atoms. The success of this method is demonstrated on the four aromatic amino acids histidine, phenylalanine, tryptophan, and tyrosine. A new technique of varying training set sizes to vastly reduce training times while maintaining accuracy is described and applied to each amino acid. Each amino acid has its geometry distorted via normal modes of vibration over all local energy minima in the Ramachandran map. These geometries are then used to train the kriging models. Total electrostatic energies predicted by the kriging models for previously unseen geometries are compared to the true energies, yielding mean absolute errors of 2.9, 5.1, 4.2, and 2.8 kJ mol^{-1} for histidine, phenylalanine, tryptophan, and tyrosine, respectively.



1. INTRODUCTION

The set of 20 naturally occurring amino acids constitute the building blocks of almost every protein in existence,[1] with protein structure determined by interactions between amino acids and between amino acids and water. Interactions involving CH···π and π···π stacking are ubiquitous in biochemical systems, with pivotal roles in the structure and function of DNA[2] and RNA,[3] molecular recognition,[4] self-assembly,[5] and even chemical synthesis.[6] In a study of the Protein Database (PDB) it was found that the aromatic amino acids tyrosine, tryptophan, and phenylalanine, make up less than 9% of bodily amino acids but are greatly over-represented at active sites. Moreover, there is one π···cation interaction for every 77 amino acids in the PDB.[7] The fundamental importance of the π-stacking interaction has led to it being called the *fourth key force in macromolecular structure* alongside hydrophobic effects, hydrogen bonding, and salt bridges,[7] as well as a *deus ex machina, intervening in reactions, stabilizing complexes, and influencing structure*.[8]

Commonly, the stacking interaction is understood to be caused by competition between a London dispersion interaction[9] (attractive) and quadrupole−quadrupole[10] interactions (repulsive). Ideally, these contributions would be included in a force field but many currently popular force fields lack multipole moments entirely[11] as well as sophisticated (or even well-parametrized) methods of describing van der Waals forces, which may contribute to errors in structure determination.[12] A study[13] of CHARMM, AMBER FF99, MM3,[14] and OPLS[15] showed that all force fields qualitatively reproduce interaction energies compared with experimental trends, but quantitatively in an inconsistent manner, especially at short-range. AMBER was shown to overestimate stacked dimer energies at equilibrium distance by 25%, with overestimation further increasing below this distance.[12] Even when Lennard-Jones parameters were optimized for benzene dimers, no close fit to quantum chemical data was obtained, but this could be rectified by including polarization and charge-penetration terms. In another study,[16] which also included OPLSAA, MMFF94/s, and MM2, it was found that all force fields struggled to reproduce intermolecular interaction energies for hydrogen bonded and some stacked complexes, with difficulties further increasing the shorter the range. AMOEBA[17] also experienced difficulty with π-stacking and it has become a restricting factor in modeling proteins featuring stacked cores, as they occur in the protein GB3.[17] The force field XED[18] fared better with these complexes, most likely attributed to its ability to represent electronic anisotropy,[19−22] a requirement for the π-electrons above and below the aromatic rings.[23] The natural inclusion of anisotropy is a feature of the current work, in order to represent the electrostatics of an aromatic ring accurately.

Aromatic atoms tend to carry different charges, even differing significantly from a double-bond. In many force fields this fact warrants the definition of additional atom types. Their parametrization can be difficult because of a lack of experimental data or understanding of the parameter set,[24] especially for atoms in a biological environment. In AMBER,[25] dealing with aromatic carbons alone involves the introduction of many new atom types.[26] However, even when properly

Received: May 14, 2014
Published: July 15, 2014

parametrized, merely adding new atom types does not help guide intermolecular binding potentials. CHARMM[27] reproduces these binding potentials at 85−99% of the experimental value, although the accuracy of these potentials deviates by 77−116% from the experimental value for stacked molecules beyond those commonly parametrized for (such as benzene−benzene complexes).[28] Ultimately, it has been suggested that different force fields be used for different systems, depending on which interactions one suspects might be encountered and indeed what structure a protein is expected to take. While using different force fields to model separate parts of a single system is indeed possible, perhaps one can argue that this represents a dystopian future for modeling. Requiring detailed knowledge of a system prior to its modeling actually precludes the construction of accurate force fields to begin with. Starting afresh with a force field architecture that fully embraces the physics of an atom and how it interacts with another atom is a strategy that should not lead to such a situation. The current work, which furthers the novel protein force field Quantum Chemical Topology Force Field (QCTFF), should been seen in this light.

It is clear that with a good description of polarization,[29,30] aromatic systems can be satisfactorily modeled and many groups are incorporating advanced polarization for such reasons.[31] We note that, as with hydrogen bonding, the inclusion of aromatic interactions into a force field is often treated *ad hoc* by introducing extra (and often costly) potential energy terms. It is with this in mind that we here investigate the four aromatic (standard) amino acids to show that the successful methodology we developed and tested on alanine[32] and histidine[33] also works for tryptophan, tyrosine and phenylalanine. It is known that multipole moments have a direct relationship with the molecular geometry[34,35] and thus a machine learning method can be used to model this relationship, especially if it is complicated. As our method intrinsically and automatically captures polarization and charge transfer effects[36,37] without the need for additional terms, it is reasonable to expect that aromatic atoms are handled seamlessly along with nonaromatic atoms. We show particular interest in tryptophan, which is a known C−H···π acceptor and shows more CH···π interactions than any other amino acid. This is attributed to tryptophan's relatively large aromatic surface area (an indole) and its potential to form stronger interactions than other aromatic amino acids with one of its two aromatic sites.[28] We recently reported our method[38] for predicting electrostatic energies built from multipole moments for histidine,[33] showing marked improvements in prediction accuracy over our work on alanine[32] due to improvements in methodology.

The first time that we believe kriging was used in the design of a potential was in 2009,[39] followed by excellent work[40] in solid state physics about half a year later, a significant contribution in the area of molecular atomization energies[41] in 2012, and nice work on approximating density functionals[42] by kriging. It should also be noted that neural networks potentials[43] can now[44] also be successfully constructed for high-dimensional potential energy surfaces.

Here we present electrostatic energy predictions for the four standard aromatic amino acids with a considerable reduction of computational cost compared to the work on histidine. No special treatment is given to the multipole moments of the aromatic atoms, without loss of accuracy when predicting these multipole moments or their resultant electrostatic interactions. Many of the principles of the design of QCTFF have been laid out before[45,46] in great detail. This is why we only review some essential elements here highlighting features that could otherwise be missed.

2. BACKGROUND AND METHODS

2.1. Topological Atoms. At the heart of the QCTFF approach is the topological atom.[47,48] Figure 1 shows how the

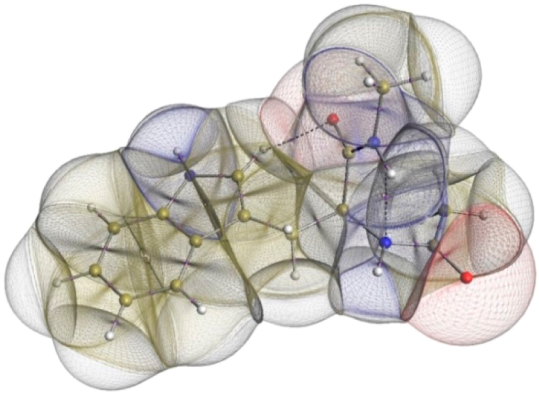


Figure 1. Peptide-capped tryptophan with topological atoms (C: gold, N: blue, O: red, H: white) superimposed over the molecular graph, showing the intramolecular hydrogen bonds via dotted lines. This picture was generated using in-house methodology developed earlier.[50,51]

electron density ρ of a molecular configuration of tryptophan (capped at both sides with peptide-bonded terminal groups) is represented as a collection of topological atoms.[49] These atoms arise in a parameter-free way, as a product of the gradient vector field of ρ. Gradient paths, or paths of steepest ascent, trace themselves. A bundle of gradient paths attracted to a given nucleus, carve out a finite volume that constitutes the topological atom corresponding to that nucleus. The interatomic surfaces are clearly visible as dividers within the molecule, whereas the outer boundary of each atom consist of a constant electron density contour surface set at $\rho = 0.001$ au.

It is important to realize that topological atoms do not overlap, which means that they do not cause penetration effects, and hence do not need damping functions to correct for these effects. Moreover, there are no gaps between the atoms. From this assertion one deduces that each point in space belongs to an atom, even points in "empty" regions inside large rings, for example. This feature opens a new paradigm for protein−ligand interaction and drug design.[52] Combining the feature of no interatomic gaps and that of no overlap is the hallmark of a "space filling" or "exhaustive" partitioning. As a result, the net charge of a functional group is simply the sum of the net charges of the constituent topological atoms. Finally, it should be clarified that QCT in QCTFF stands for Quantum Chemical Topology, which is an approach built on the central idea of partitioning a quantum system by the gradient vector field of a quantum mechanical function of relevance. This idea was pioneered in the Quantum Theory of Atoms in Molecules (QTAIM),[47,53] which covers only the electron density and its Laplacian. Details of which other quantum functions have been partitioned according to QCT can be found elsewhere.[49,54,55] As a collection of topologically partitioned data, QCT is a powerful and original way to recover chemical insight from

dx.doi.org/10.1021/ct500416k | J. Chem. Theory Comput. 2014, 10, 3708−3719

modern wave functions, an attribute that QCTFF benefits from, at least at the level of ρ.

In this work a molecular configuration is distorted via its normal modes of vibration, achieved by the in-house program EROS. This program generates concerted perturbations to the molecular geometry, affecting the whole nuclear skeleton with changes in bond lengths and angles. A complete methodology for such distortions along with relevant considerations is given by Ochterski.[56] Such a perturbation is illustrated in Figure 2

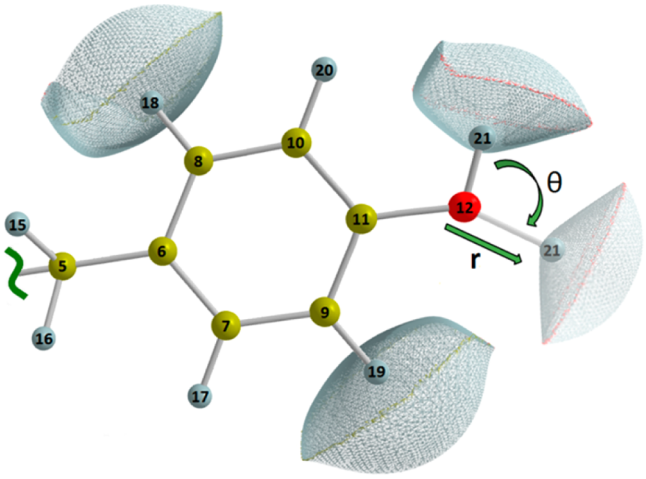


Figure 2. Change in molecular geometry causes change in the shape of each topological atom due to polarization, illustrated by H21 in the phenolic side chain of tyrosine. The pairwise interactions between the atoms give the electrostatic interaction energy.

where atom H21 undergoes a strongly localized geometric change, for demonstration. Guided by the realism of normal modes (as opposed to arbitrary combinations of values of internal coordinates), this atom ends up in a new position represented by the faded version of the atom. This geometric change also causes a change in the multipole moments of this and the other atoms. Through polarization and charge transfer, which QCT treats on a par (without extra, special terms), each perturbation also affects the electrostatic properties of every other atom. Distorting through all normal modes simultaneously yields as many unique geometries as desired, each geometry resulting in a unique set of topological atoms, atomic multipole moments and thus interatomic electrostatic energies.

A local energy minimum of the molecule is used as a template (or starting point) for distortion. EROS applies energy to all normal modes of vibration in random amounts, giving a limitless supply of unique distortions (although distortions are filtered if their bond length or angle surpasses 20% deviation of its original value, to avoid broken bonds). Normal coordinates q_k are given[57] by eq 1

$$q_k = \sum_i \sum_\alpha \sqrt{m_i}\, l_{i\alpha k}(r_{i\alpha} - r^0_{i\alpha}) \tag{1}$$

where m_i is the mass associated with the ith atom at the instantaneous position $r_{i\alpha}$ ($\alpha \in \{x, y, z\}$) in the Eckart frame. The nuclear coordinates of the stationary reference structure are given by $r^0_{i\alpha}$ and $l_{i\alpha k}$ specify the elements of the eigenvectors of the mass-weighted Hessian. Following the work of Watson,[58] one can express the instantaneous positions as a function of the normal coordinates

$$r_{i\alpha} = r^0_{i\alpha} + \sqrt{1/m_i} \sum_k l_{i\alpha k} q_k \tag{2}$$

After generating the distorted geometries, a wave function for each geometry is calculated using GAUSSIAN03[59] at the B3LYP/a-pc1[60] level of theory and the molecular electron density is evaluated. This augmented basis set is invoked here because it has been optimized for use with density functionals. From its wave function, the electron density of a molecule can be expressed as atom-centered multipole moments, calculated by the program AIMAll[61] using default settings and integration error control.[62] The most basic moment, the monopole, is a shapeless, isotropic moment that is equal to the charge. As moments increase in rank, the monopole being rank zero, their shape becomes more complex; the dipole moment has a shape analogous to a p-orbital while quadrupole moments can be considered similar to d-orbitals. In this work we use all moments up to hexadecapole (rank 4),[63] providing an anisotropic and polarized description of an atom's electrostatics and electronic shape.

When all interactions between multipole moments are summed, the total interatomic electrostatic interaction energy is reached. The electrostatic interaction E_{AB} between atoms A and B is calculated[64] using eq 3

$$E_{AB} = \sum_{l_A l_B m_A m_B} Q_{l_A m_A} T_{l_A l_B m_A m_B} Q_{l_B m_B} \tag{3}$$

where l and m respectively denote the rank and component of multipole moment Q while T is the interaction tensor between the two multipole moments, one on atom A and one on atom B.

2.2. Treatment of Polarization. Before going into technical details, the overall strategy should be explained. First, it should be highlighted that the partitioning scheme that is *QCT makes no topological distinction between intermolecular and intramolecular interactions*. In other words, topological atoms are malleable boxes with boundaries obtained by an *identical* prescription, independently of whether they interact in the same molecule or across two different molecules. This means that the method laid out in this paper is also applicable for intermolecular polarization; in fact, this is where it was developed[65] and applied in the first place. This also means that QCTFF does not use the ideas of long-range Rayleigh–Schrödinger perturbation theory. Instead, QCTFF starts from the electron density of any given system, which can be a single molecule or a molecular cluster (including a solvated ion,[66] for example). The second major difference with traditional treatments of polarization, even in the context of multipolar electrostatics, is that we focus on the end result of the polarization process rather than the polarizability itself. In other words, QCTFF does not work with explicit polarizabilities, but strives to predict the response itself, i.e. the new multipole moment of interest, after polarization has occurred. QCTFF focuses directly on what a multipole moment of a given atom is, only knowing the nuclear positions of the atoms surrounding this given atom. Because this relationship can be complicated it is best to capture it via machine learning.

Machine learning is used to directly map the relationship between a distorted geometry and the atomic multipole moments in that geometry. In this work, we use a powerful machine learning method, called kriging,[67] which is elaborated in section 2.3. In machine learning language the name *feature* refers to an input variable, which here is one of the $3N - 6$

internal coordinates (where N is the number of atoms). In order to reduce a full description of the geometry to $3N - 6$ coordinates an Atomic Local Frame (ALF) is installed on a given atom, with respect to which the entire molecule is expressed in terms of ALF coordinates. For example, let us choose an atom in a molecule, denote it atom A, and take it as the origin of the local frame. Then this atom's ALF is defined by the two highest-priority (using Cahn−Ingold−Prelog rules) neighbors, denoted atoms X and Y. The distance from A to each of these neighbors constitutes the first two ALF components (R_{AX} and R_{AY}), while the angle between the two neighbors and the origin atom gives the third ALF component (θ_{XAY}). The ALF then establishes a right-handed coordinate system, centered at the origin atom, from which every other atom's position can be defined using spherical polar coordinates (R_{Ak}, ϕ_{Ak}, θ_{Ak} where k is any and each atom except the origin atom A and its two ALF neighbors X and Y). This string of coordinates (of the 3 ALF atoms and the $3(N - 3) = 3N - 9$ non-ALF atoms) constitute the input features for atom A's kriging model. Each model is built using $3N - 6$ inputs for each molecular geometry in the training set. Every atom is at the origin of its own kriging model and this fact may be exploited for transferability of kriging models in future work. No global rotation or translation is considered by the kriging procedure. Even more details and a figure can be found elsewhere.[68]

Kriging is capable of modeling the relationship between geometry and corresponding multipole moments in a high-dimensional feature space, where multiple features all influence the same multipole moment. We create a kriging model for each atomic multipole moment occurring in the molecule, up to hexadecapole. When a previously unseen molecular geometry is taken, these kriging models are used to determine its multipole moments, without the need for any ab initio calculation. The training data for each kriging model consists of at least 400 examples, i.e. molecular geometries with known ab initio atomic multipole moments. As the training set size increases, the prediction accuracy for multipole moments for atoms in unknown geometries typically increases. Previously unseen test geometries that lie near the training set geometries are generally predicted better than those further away. Hence, it is advantageous to include as many training points as is practical.

An in-house program called FEREBUS is used to predict the multipole moments for a new geometry. The moments are then used by the in-house program NYX to calculate pairwise interatomic electrostatic energies, given in eq 3, that can then be summed to give a total intramolecular electrostatic energy. The error in the electrostatic energy of a system geometry is defined as the difference between its predicted energy and its actual energy. The former is calculated through the sum of pairwise interactions of predicted moments between atoms A and B, the latter through the similar pairwise interactions of the system's ab initio moments. Both are used to gauge the kriging prediction accuracy. The absolute electrostatic energy prediction error of a system is given by eq 4

$$\begin{aligned} |\Delta E_{\text{system}}| &= |E_{\text{system}}^{\text{actual}} - E_{\text{system}}^{\text{predicted}}| \\ &= |\sum_{\text{AB}} E_{\text{AB}}^{\text{actual}} - \sum_{\text{AB}} E_{\text{AB}}^{\text{predicted}}| \\ &= |\sum_{\text{AB}} \Delta E_{\text{AB}}| \end{aligned} \tag{4}$$

One can report this error as an average over all configurations, but it is more transparent and informative to display the full spectrum of errors encountered for all configurations of the test set. A so-called S-curve, called after its typically sigmoidal shape, serves this purpose (first example see Figure 6, to be discussed later).

2.3. Machine Learning: Kriging. When many geometries of a molecule are fully assigned atomic multipole moments (by time-consuming ab initio and atomic partitioning calculation), a cause-and-effect relationship is constructed between the molecular geometry and the atomic multipole moments. To model this relationship we use the kriging machine learning method,[67] also named "Gaussian process regression".[69] Kriging is a stochastic interpolation technique based on work by Krige.[70] The method is rigorously elaborated upon in a previous work[33] but is only summarized here based on the treatment of Jones et al.[71,72] Kriging can map an input (here, descriptions of molecular geometry) to an output (multipole moments) so long as these values lie within the bounds of the data used to train with.

The kriging prediction process is given by eq 5

$$\hat{y}(\mathbf{x}^*) = \hat{\mu} + \sum_{i=1}^{n} a_i \varphi(\mathbf{x}^* - \mathbf{x}^i) \tag{5}$$

where $\hat{\mu}$ is a constant "background" term modeling the global trend of the trained outputs, $\hat{y}(\boldsymbol{x}^*)$ is the output from a new input point $\mathbf{x}^*$ (i.e., not used for training), while $\mathbf{x}^i$ is a known input (i.e., the set of features of the ith training example), and $\varphi(\boldsymbol{x}^* - \boldsymbol{x}^i)$ and a_i are quantities calculated below. The global trend $\hat{\mu}$ is similar to the mean value of the outputs used for training. All outputs used for training are considered "errors" (deviations) from this global trend, instigated by a change in the input. For example, if we use kriging to model an atomic charge, the global trend is a typical background charge that is most generally applicable to that atom while changes in bond lengths and angles (i.e., inputs) cause a deviation from this in the charge (i.e., outputs). Finally, a_i is the ith element of the vector $\boldsymbol{a} = \boldsymbol{R}^{-1}(\boldsymbol{y} - \mathbf{1}\hat{\mu})$ while $\varphi(\boldsymbol{x}^* - \boldsymbol{x}^i)$ is the ith element of the row vector $\mathbf{r}$, which is calculated via eq 6

$$r = \{\text{cor}[\varepsilon(\mathbf{x}^*), \varepsilon(\mathbf{x}^1)], \text{cor}[\varepsilon(\mathbf{x}^*), \varepsilon(\mathbf{x}^2)], ..., \text{cor}[\varepsilon(\mathbf{x}^*), \varepsilon(\mathbf{x}^n)]\}' \tag{6}$$

where the correlation kernel is defined as

$$R_{ij} = \text{cor}[\varepsilon(\mathbf{x}^i), \varepsilon(\mathbf{x}^j)] = \exp[-\sum_{h=1}^{d} \theta_h |x_h^i - x_h^j|^{p_h}] \tag{7}$$

where d is the dimensionality of the feature space (i.e., number of features or inputs), and $\mathbf{x}^i$ and $\mathbf{x}^j$ describe the ith and jth molecular configuration, respectively, using the set of features detailed in the paragraph on the ALF in Section 2.2. The principle of kriging is the assumption that if a new input data point is of a similar value to an existing input, their corresponding outputs should also be similar. The correlation between any two points is determined by the distance in feature space that separates them: as the distance between two points increases, the correlation approaches zero. This also means that the kriging predictor passes exactly through each training point, with the prediction error at these points being zero and their correlation 1. Thus, is it beneficial to have a high density of training data to ensure prediction inputs are close to training points. Note that the parameters θ_h ($\theta_h \geq 0$) and p_h ($1 < p_h \leq 2$) can be written as d-dimensional vectors, $\boldsymbol{\theta}$ and $\mathbf{p}$. These

dx.doi.org/10.1021/ct500416k | J. Chem. Theory Comput. 2014, 10, 3708−3719

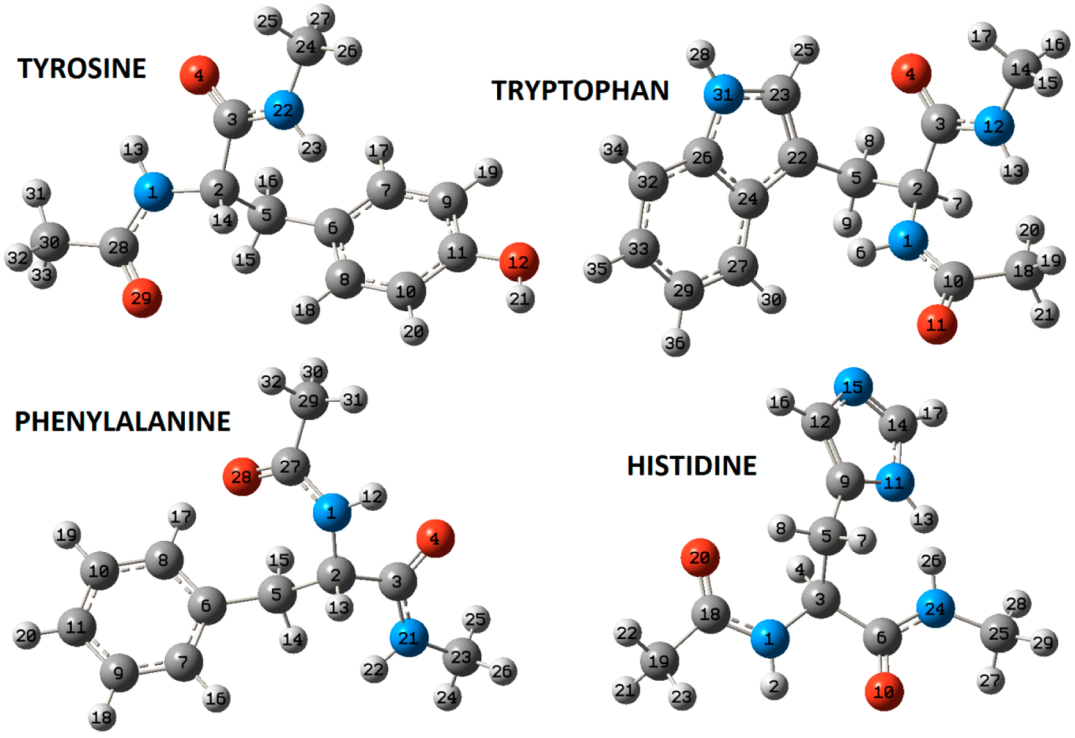


Figure 3. Full set of four doubly peptide-capped aromatic amino acids. Atoms are numbered as they are referred to in later figures. Atoms are color-coded according to element (C: dark gray; O: red; N: blue; N: blue; and H: light gray).

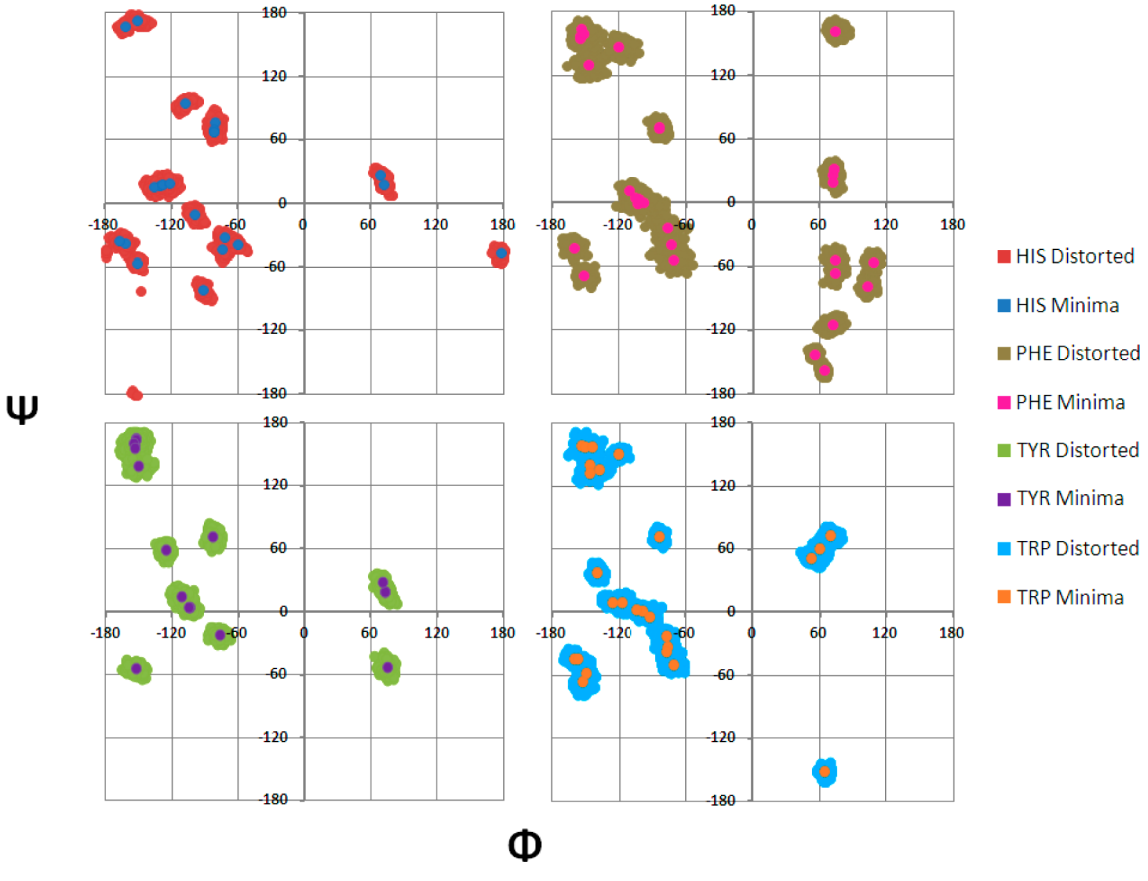


Figure 4. Ramachandran plots for the four aromatic amino acids. Solid, bright points on the plot represent energy minima where the distorted molecules tend to form their "islands" in a different color marked "distorted".

dx.doi.org/10.1021/ct500416k | J. Chem. Theory Comput. 2014, 10, 3708−3719

parameters can be optimized in order to maximize the likelihood function[69] associated with the kriging process. This is achieved by a Particle Swarm Optimizer, following the methods of Kennedy and Eberhart,[73] which optimizes all parameters θ_l, while previous work by the group found that setting $p_h = 2$ is an excellent approximation, reducing overall computation time by removing the need to optimize it.

3. RESULTS AND DISCUSSION

Figure 3 shows the four amino acids under study, each chosen for being a naturally occurring amino acid with an aromatic side chain. Together they represent a subset of amino acids that are of special interest to the development of QCTFF, for the reasons outlined in the Introduction. Each molecule is distorted pseudorandomly according to its normal modes of vibration; however, it is difficult to visualize the extent of these distortions due to their high dimensionality, multiple energy minima and the large number of distortions themselves.

Figure 4 gives an idea of the extent of distortion but seen through a Ramachandran plot, which only displays two dihedral angles associated with the protein backbone geometry around the C_α. In each plot, the local energy minima are represented by uniquely colored points, each in an "island" of distortions that spread out from the minima. The islands are typically quite circular, indicating that there is no large bias toward particular dihedral values from the normal mode distortions. The islands may join one another where the dihedral values are similar between minima. It is important to note that we only see two degrees of freedom through the Ramachandran plots and two conformations with similar (or even identical) dihedral angles can still be very different geometrically. It appears that some amino acids are more easily distorted than others. For example, the islands for phenylalanine are noticeably larger than those of histidine. In addition, it tends to be true that all minima for any given molecule are quite evenly distorted and larger islands are found where multiple minima gather. Some of the more isolated minima present a potential problem for kriging, apparently being distant from other training data points. This is not necessarily true, as the conformations in an isolated island may yet have similar features (dihedral angles are not used as features) to other molecules, although it is logically less likely. In this investigation, both training and test examples are taken from the same pool of data, shown in Figure 4. As such, we do not note significantly poorer predictions for conformations in isolated islands.

Each atomic multipole moment in a molecule has its own kriging model (independently obtained from another model). An important question is whether each kriging model needs the same number of training examples, under the hypothesis that higher-rank multipole moments may not contribute as much to the total error in electrostatic interaction energy than lower-rank ones. Therefore, we test a more efficient method of data selection, where the number of data points trained for depends on the rank of the multipole moment. Higher-rank multipole moments can be trained for with fewer data points. This greatly reduces the overall time taken to train for all multipole moments of an atom, with minimal sacrifice of prediction quality in the total energy error.

Figure 5 shows the effect of altering training set size for different atomic multipole moments for a single carbon in *N*-methylacetamide, which is used as a preliminary test system. The blue data points show the traditional methodology where all multipole moments (referred to as the "full series") are

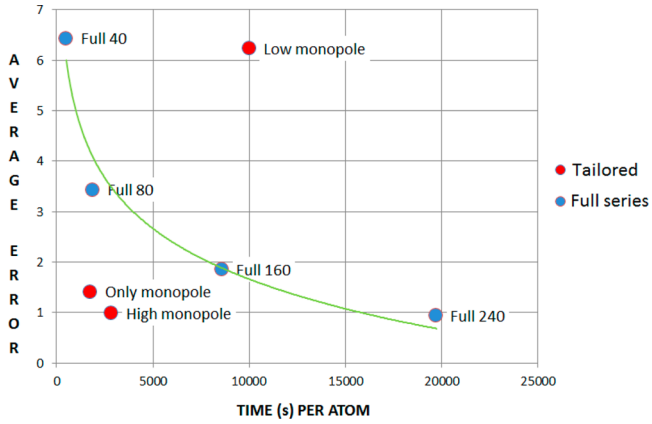


Figure 5. Effect of using tailored training set sizes for *N*-methylacetamide (NMA) in terms of absolute prediction error (kJ mol^{-1}) and training time required. The Pareto front (green) connects the results for the traditional training size ("full", blue), where all moments are trained for with the same number of training points. The tailored training sets are given in red, and their respective training set sizes are displayed in Table 1.

treated with training sets of identical size. Meanwhile, the red points mark the results for multipole moments trained with different set sizes, "size-tailored" or just "tailored" toward specific moments by committing more training examples to them. Each point gives the mean absolute electrostatic error of 100 NMA geometries as given in eq 4.

A Pareto front is constructed from the "full series" data and is used to assess the success of tailored size training sets (Figure 5, in red). When the atomic charge (i.e., monopole moment) has been trained for with a small set size and the other moments trained with a large set, the prediction quality is not significantly better than that for a traditional small training set size (such as "full 40"), despite dramatically larger training times (displayed by the "low monopole" point). Instead, using large training set sizes for charge, but small training set sizes for the higher order moments, gives a prediction quality almost identical to that of traditional large training set sizes but at only one-tenth of the required training time (displayed by the "only monopole" and "high monopole" points).

The tailoring of training sets presents new options for the kriging process and is especially relevant to larger molecules that would have otherwise been too costly to train at a large set size. When S-curves are constructed for these predictions, we have observed that it is advantageous to use a medium-size training set for monopole and dipole moments while higher-rank multipole moments can be trained by few data points. From now on, kriging models using the "high monopole" training set sizes will be discussed for the four naturally occurring aromatic amino acids. The electrostatic energy prediction of histidine has been reported previously[33] and is used to compare the full training set kriging models with the proposed tailored training set sizes. Note that training times for moments fluctuate even within the same atom and multipole moment rank, thus mean times are presented.

It is seen in Table 2 that the 50-example training sets train about 20 times faster than the 400-example training sets. While electrostatic energy models using 50-example training sets for each moment would produce very poor-quality predictions, Figure 5 demonstrates the applicability of smaller training sets to higher order multipole moments with negligible loss in

Table 1. Training Set Sizes for Each *N*-Methylacetamide Data Set Used in Figure 5 in Order To Test the Effect of Tailoring Specific Moments for Kriging Training[a]

	monopole	dipole	quadrupole	octopole	hexadecapole	ERROR
full 40	40	40	40	40	40	6.4
full 80	80	80	80	80	80	3.4
full 160	160	160	160	160	160	1.9
full 240	240	240	240	240	240	0.9
low monopole	40	80	160	240	240	6.2
high monopole	240	160	80	40	40	1.4
only monopole	240	40	40	40	40	1.0

[a]Absolute electrostatic energy prediction errors for each set are given in kJ mol^{-1}.

Table 2. Comparison between Training Sets of Size 400 and 50 Examples: Training Times (s) for all 25 Multipole Moments of C18 in Histidine[a]

	monopole	dipole	quadrupole	octopole	hexadecapole	total
50 time	401	1102	2066	2960	3898	10427
50 mean	401	367	413	422	433	417
400 time	8211	17656	28607	52139	63004	169617
400 mean	8211	5885	5721	7448	7000	6785
speed up	20.5×	16.0×	13.9×	17.6×	16.2×	16.3×

[a]Mean times are obtained by dividing the time for a multipole moment by its multiplicity (two times the rank plus one). All speed increases are calculated from the reported mean values.

accuracy. Table 2 breaks down timings for atom C18 of histidine, illustrating the difference in training times between the 400-example and 50-example training sets.

Application of the "only monopole" model (Table 1) for histidine results in an estimated training time of approximately 5 h, as opposed to 47 h for the "full 240" training set. Based on what can be learnt from Figure 5 (for the case of *N*-methylacetamide) we expect the application of the "only monopole" model to cause only a slight increase in energy error in spite of a speed up about 9 times (~47/5). In our previous work we used a "full 600" training set size of 600 for all multipole moments and atoms. Here we use 600 examples for the monopole only (accounting for 1/25th of the total number of models to be constructed). Here we also use 400 examples for the dipole (3/25th), 200 for quadrupole (5/25th), and 50 for all other moments (16/25th). As the kriging models of each atom are computed in parallel, the training time for the whole molecule is that of the atom that needs the longest training time. Note that if high-throughput computation is available all individual moments could be trained in parallel, further reducing the training time for an entire molecule to that of the "slowest" single atomic multipole moment. By employing tailored training set sizes, we reduce the previously reported training time of histidine from 87 h to only 10.5 h, which is 12% of the initial training cost, with an increase in absolute prediction error of only 0.4 kJ mol^{-1}.

Figure 6 shows the electrostatic energy predictions for histidine using the "tailored" training set sizes for both single and multiple minima. Each point in an S-curve corresponds to the absolute energy error obtained for a molecular geometry of the test set, containing 200 examples. An S-curve displays the full performance in terms of absolute prediction error, for all the kriging models contributing to the total electrostatic energy, i.e. 25 kriging models for each atom. Note that no geometry of the training set ever occurs in the test set, so the latter is *external*. It is clear from the blue "single energy minimum" curve that about 90% of the geometries in the test set have an absolute error of less than 1 kcalmol^{-1}. It should also be clear

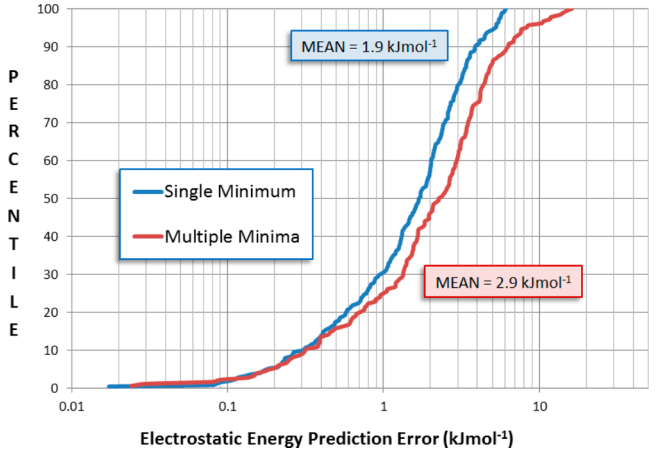


Figure 6. Absolute electrostatic energy prediction errors for histidine. The red curve uses multiple energy minima for training data and test geometries while the blue curve uses only a single energy minimum. Both S-curves are built from 200 test geometries external to the training set of 600.

that the maximum incurred error can be read off where the S-curve hits the ceiling of 100%, and that an S-curve lying to the left of another S-curve is superior to the latter.

The S-curve covering all 24 energy minima of histidine is marked in red in Figure 6 and has a mean error of 2.9 kJ mol^{-1}. Although this error is 0.4 kJ mol^{-1} larger than previously reported,[33] it is achieved in a fraction of the time using tailored training set sizes. Furthermore, when just a single energy minimum is trained for and predicted (blue curve, Figure 6), the errors are significantly lower than for the multiple minima set. This effect is due to the reduced range of geometries occurring in the training for a single-minimum, which is just a single island out of the 24 possible minima present in the Ramachandran plot of Figure 4. Note that it is possible to train each minimum separately, where the appropriate kriging model is automatically selected when needed for predicting, which presents an accessible route for further reduction of prediction

dx.doi.org/10.1021/ct500416k | J. Chem. Theory Comput. 2014, 10, 3708−3719

errors and computation times. However, this approach has not been used here, in favor of the "completeness" offered by a single kriging model encapsulating all energy minima.

Figure 7 gathers the S-curves for phenylalanine, tryptophan, and tyrosine. Phenylalanine, tryptophan, and tyrosine have 200

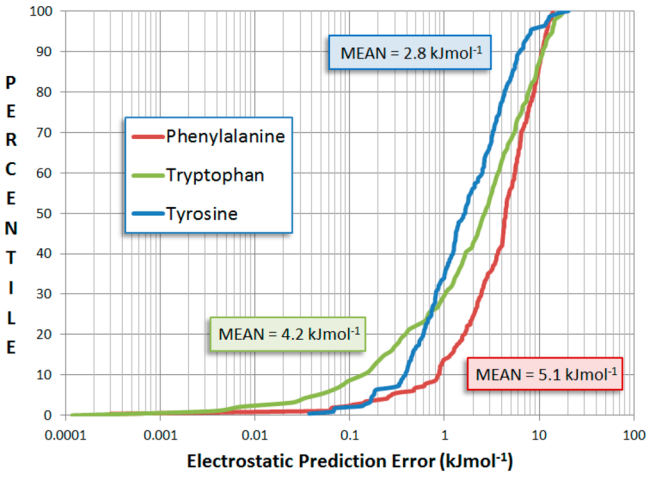


Figure 7. S-curves for phenylalanine, tryptophan, and tyrosine and mean absolute energy prediction errors.

test predictions for all of their 30, 26, and 17 energy minima, respectively. The training set sizes were tailored in the same way as for histidine. The mean absolute errors obtained for phenylalanine, tryptophan, and tyrosine are respectively 5.1, 4.2, and 2.8 kJ mol^{-1}. While tyrosine shows a mean absolute error closely in line with that for histidine, phenylalanine and tryptophan returned comparatively poor predictions. This result is in accordance with Figure 6, with the greater number of minima possessed by phenylalanine (30) and tryptophan (26), compared to tyrosine (17), resulting in a larger volume of feature space for the kriging model to predict for. It is interesting to note that regardless of mean absolute error, all amino acids have very similar worst-case predictions (i.e., maximum absolute errors found at the 100% percentile) of approximately 17 kJ mol^{-1}. These maximum errors are outliers, and tend to be rare. For phenylalanine, with the highest mean absolute error, 90% of conformation prediction errors lie below 10 kJ mol^{-1}, while tyrosine, with similar maximum absolute errors to the other amino acids but the lowest mean absolute error, has 90% of errors below 6 kJ mol^{-1}. It is important to note that for the systems considered, increasing molecular size does not have a notable effect on prediction accuracy. Indeed, the results suggest that small changes in the dimensionality of feature space are less significant to the accuracy of a kriging model than the value range within each dimension that is being sampled.

Finally we gauge kriging's ability to predict the multipole moments for aromatic atoms. Figure 8 shows absolute prediction errors for various carbon and hydrogen atoms throughout tryptophan; the atom numbers are the same as those in Figure 3.

When absolute prediction errors (au) for monopole moments (or charges) of each atom are compared, the aromatic atoms (in blue) show similar errors to other atom types of the same element. Hydrogen atoms (circles) tend to be better predicted than carbon atoms (squares) but are not exclusively so, and all atoms are well predicted. Because

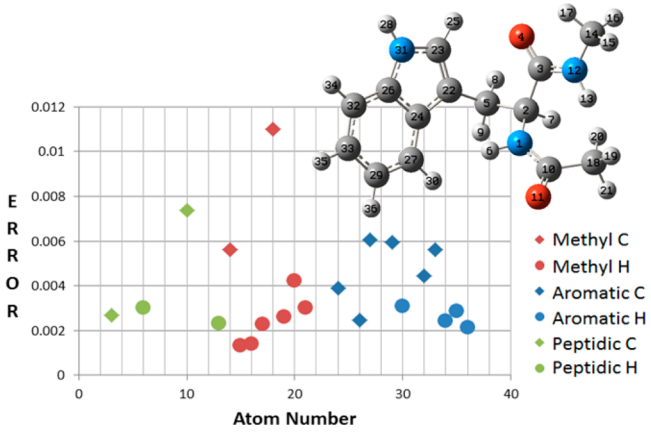


Figure 8. Absolute errors (au) for the prediction of the monopole moment (charge) on specific atoms in tryptophan. Atoms of similar "type" share the same color, while all hydrogens are circles and all carbons are diamonds.

aromatic atoms are conformationally locked and have less freedom than other atom types, they are well modeled by kriging. The small range of features and moments for these atoms means a higher density of training data and thus reliably accurate predictions. We emphasize that the aromatic atoms have their charges predicted to the same accuracy as other atom types without any special treatment.

Finally, a brief discussion of the potential of the current method in future work is in place here. First, work is underway that explicitly hydrates amino acids with an eye on QCTFF handling proteins in aqueous solution. Some time ago the effects on atomic volumes, energy, and multipole moments of intra- and intermolecular hydrogen bond formation in [amino acid···water] clusters has been worked out[74] within the framework of QCT. The in-house "pipeline" is mature enough to now construct kriging models for hydrated amino acids. A second important application of the current method is fuelled by scale enlargement, where unpublished work on cholesterol and $(Ala)_{10}$ shows that kriging can handle systems larger than single amino acids. Here feature selection becomes a priority, driven by a kriging-based computation of an atom type, which had been *computed* for the first time[75−77] but from intrinsic QCT atomic properties (without using kriging but cluster analysis instead). A third advance is the recently derived and implemented analytical forces of kriging models (submitted), which will be incorporated in a molecular simulation package, where we have chosen the FORTRAN90 program DL_POLY[78,79] as an appropriate vehicle. The prospect of the QCTFF methodology becoming applicable in the area of protein folding is challenging but exciting. A fourth research strand is the incorporation of nonelectrostatics terms (e.g., kinetic energy,[80] which is well-defined for a topological atom) within the QCT framework of Interacting Quantum Atoms (IQA).[81] The topological energy partitioning method IQA was made possible by the first ever calculation of an interatomic Coulomb potential energy,[82] achieved by a six-dimensional integration simultaneously over two topological atoms (now covering the case of a molecule not in an equilibrium geometry). This scheme provides novel and rigorous chemical insight into hydrogen bonding,[83] the origin of barriers[84,85] and quantifies covalency of bonds (and any other interatomic interaction) by the so-called exchange (correlation) potential energy.[86] This energetic quantity is linked[87] to the QCT feature

of bond critical points in an increasingly rigorous manner.[88,89] The prospect of incorporating this insight into QCTFF is tantalising because it would eliminate longstanding issues on how to deal with stereoelectronic effects in force fields. A proper treatment of conformational isomerism should then naturally follow from QCTFF's architecture in terms of the intra-atomic ("self-energy") and interatomic energy of the topological energy partitioning. Finally, a fifth avenue of QCTFF methodology expansion is that into chemical reactions. The recent momentum that QCT has gathered in terms of bond characterization via exchange (correlation) potential energy puts QCTFF in a strong position to tackle bond making and breaking, the hallmark of a chemical reaction. All of the activities discussed above are currently moving forward.

4. CONCLUSION

Next generation force fields need a reliable representation of the electrostatics of aromatic rings in amino acids. Here we demonstrate that current computational power enables the construction of a novel polarization method for such systems. The machine learning method kriging successfully predicts a multipole moment of a given atom *directly* from the positions of the atoms surrounding it. Thus, this method focuses on the outcome of the polarization process, avoiding the calculation of an explicit polarizability. After training, kriging predicts atomic multipole moments for previously unseen molecular geometries. The resulting total intramolecular electrostatic energies are predicted with mean absolute errors of 2.9, 5.1, 4.2, and 2.8 kJ mol^{-1} for histidine, phenylalanine, tryptophan and tyrosine, respectively. One set of kriging models covers all local energy minima found for each amino acid. Much computation time can be saved in the training process by tailoring the set sizes to the rank of the atomic multipole moments, by greatly reducing the training set size by increasing multipole moment rank. The current method shows that aromatic atoms do not require a special treatment compared to other atoms.

APPENDIX: DETAILS OF THE QCTFF METHODOLOGY

1. Introduction

"Pipeline" is a program, written in Perl, around 2500 lines in length, which aids the user in progressing through the methods described in the main text. By consolidating the entire method and all relevant programs into a single script, several benefits have been reaped:

- High level of automation gives a greatly improved turnover rate for results.
- Automation protects against user errors.
- New users can successfully go through the entire method with little training.
- Version control becomes more practical.
- Compatibility between different QCTFF sections is assured.

Pipeline is a necessity: for example, if an amino acid of 30 atoms is modelled and predicted, a standard run through the method would utilise 2000 geometries. This means 2000 GJF and WFN files, 60 000 INT files, and 750 kriging models, all of which must be moved and manipulated—without mistakes—to be used in the next part of the method.

The Pipeline program is split into five major parts (menus 1−5, see figure above), each activated by a single numerical keystroke that each performs a number of tasks in an order that is most standard to the method. A brief overview of each menu is given just below.

2.0. General Workings

2.1. Menu 1: Generation of Distorted Geometries. A machine learning method such as kriging needs a sufficient number of molecular geometries for use in training as well as many more unique geometries for testing. An in-house FORTRAN90 code named "EROS" is used by Pipeline to generate these geometries by applying energy to the molecule's normal modes of vibration, distorting it pseudorandomly.

Using option 1 in Pipeline will cause all WFN/GJF/FREQ files (wave function, Gaussian input, and frequency files, respectively) in the home directory to be moved to a new directory. Pipeline will then create input for EROS and run this program, creating distortions. The number of distortions (and thus, unique geometries) produced is given by the first number in Pipeline's input file and is typically over 1000, using half of these for training. When multiple energy minima are used, this number is automatically spread evenly across the number of minima.

2.2. Menu 2: Wave Function Generation. During this phase, a wave function is calculated for each of the geometries attained in menu 1. GAUSSIAN09 is the default application used for this, and Pipeline will submit GJF files as an array job after shuffling and reformatting them. Pipeline will first move all GJFs to a new directory and randomly shuffle their file numbers. The user is prompted to choose a level of theory. Each GJF retains the charge that the original GJFs were assigned (it is important for this to be correct at this stage as it cannot be easily corrected later). Any additional required custom levels of theory are written into the GJF files and each is formatted so as to reduce any errors from GAUSSIAN, which tends to be strict on its input. Jobs are submitted to backend nodes for calculation using a job array specifying the use of GAUSSIAN09 or GAUSSIAN03.

2.3. Menu 3: Topological Partitioning and Calculation of Atomic Properties. Here, in phase 3, the goal is to take the output from GAUSSIAN (wave function files) as input for AIMAll. AIMAll uses the electron density based on the molecule's wave function to define atomic basins and integrates over them to find each atom's multipole moments. Using option 3 will take all completed *.WFN files and copy them to the WFN directory for safe keeping and then to the AIMALL directory for calculation using AIMAll. Submitting these files to backend nodes will then create a job array containing all jobs and it is recommended to let these complete before continuing. Once all files are completed, using this option again will cause Pipeline to reorganize the output from AIMAll so that all files are sorted by atom and will offer to resubmit any missing files as a new job array. When resubmitting jobs, one should not continue to the kriging stage until those jobs are completed. If one chooses to continue through to kriging by gathering data instead of resubmitting, Pipeline will begin the reorganization process and this can take significant time due to the high volume of file handling operations. Pipeline will expect all atoms listed in the input to be present at this point and will flag an error if this is not the case.

2.4. Menu 4: Kriging Training. The training phase is the heart of the electrostatic prediction process, making kriging models that encapsulate the relationship between changes in molecular geometry to changes in atomic multipole moments. RAW files are created by looping through every WFN file and INT file (GAUSSIAN and AIMALL output, respectively) and matching molecular geometries to each atoms multipole moments. A single RAW file is created for each atom in the system and consists one line for each molecular geometry. Each line contains $3n + 25 + 2$ columns (Cartesian coordinates of the n atoms per molecular geometry + multipole moments + integration error and file number). The training examples in the RAW files must be converted to a format useful to FEREBUS, the program that trains the kriging models. The RAW files are used as inputs for an in-house FORTRAN program named "NORMART". This program will convert the Cartesian coordinates into "features" defined by the ALF. The atomic multipole moments are then rotated to the ALF after which both features and moments are normalized. The normalized features and normalized, rotated moments are printed into a training set file and their minimum, maximum and mean values are recorded in a STAND file so that NYX can later denormalize all values. NORMART requires no input from the user: only the RAW file and a parameter file that Pipeline will generate. The user may, however, need to load the latest Intel FORTRAN compiler modules if using a Linux machine.

The training sets can now be "scrubbed" by Pipeline. This is a three-step process.

1. Removing all geometries where the integration error is above the defined limit.
2. Removing all "incomplete" geometries so all atom training sets contain the same geometries.
3. Removing all geometries whose atomic charges do not equal what is expected (e.g., zero for a neutral molecule).

The training sets are now ready to be used by FEREBUS, an in-house FORTRAN program that creates kriging models from the training data. Pipeline will generate input files for FEREBUS and pair these with training sets in their own directories where FEREBUS can be executed. It is possible to split these files up quite significantly. FEREBUS could potentially run separately for every moment on every atom, creating 750 jobs for a hypothetical 30-atom system. However, this amount of parallelization is not commonly required and Pipeline's default approach is to have a single job for every atom in the system. The user is not prompted for advice on the FEREBUS input as Pipeline uses a set of default options along with a tailored training set. It is possible for the user to intercept FEREBUS's input and change the options before Pipeline continues to submit the jobs but this is not recommended under general use. Selecting to submit FEREBUS jobs will also rewrite the FEREBUS input files, giving users an opportunity to alter the training set size (Pipeline input file) and resubmit without having to go through a lengthy file-writing process again.

2.5. Menu 5: Prediction of Interatomic Electrostatic Energies. Finally, electrostatic energy predictions are ready to be made. This involves comparing "true" ab initio multipole moments to those predicted by the kriging models. For this, an in-house FORTRAN program, called NYX, is used. NYX is given an unseen molecular geometry and uses the ab initio data generated earlier to look up its multipole moments obtained by ab initio calculation (from AIMAll).

NYX will also go one step further than this and use the true multipole moments and geometry to calculate electrostatic interaction energies for specified atom pairings. The summation of these pairwise energies is then presented as the total electrostatic energy. NYX will then do the same using the predicted moments and compare the two energies, giving a final electrostatic interaction energy prediction error, as seen in the S-curves in this paper. While Pipeline is capable of performing the entirety of this step (using a Linux-compiled version of NYX), it is highly recommended that NYX is run manually and will require manual input from the user for the interactions list, regardless. This interactions list should only include 1−4 and higher (that is, atoms connected by 3 or more bonds) interactions. Should intermolecular energies be required, only intermolecular interactions should be included (in short, NYX

will only calculate interactions that it is explicitly asked to calculate).

AUTHOR INFORMATION

Corresponding Author

*E-mail: pla@manchester.ac.uk.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank EPSRC for funding this work through the Established Career Fellowship grant EP/K005472/1.

REFERENCES

(1) Hertweck, C. *Angew. Chem., Int. Ed.* **2011**, *50*, 2.
(2) Hunter, C. A. *J. Mol. Biol.* **1993**, *230*, 1025−1054.
(3) Waters, M. L. *Curr. Op. Chem. Biol.* **2002**, *6*, 15.
(4) Asensio, J. L.; Vacas, T.; Corzana, F.; Jimenez-Oses, G.; Gonzalez, C.; Gomez, A. M.; Bastida, A.; Revuelta, J. *J. Am. Chem. Soc.* **2010**, *132*, 17.
(5) Shen, Z.; Wang, T.; Liu, M. *Chem. Commun.* **2014**, *50*, 4.
(6) Stoddart, J. F.; Tseng, H. R. *Proc. Nat. Acad. Sci. USA* **2002**, *99*, 4797−4800.
(7) Dougherty, D. A. *J. Nutr.* **2007**, *137*, 5.
(8) Waller, M. P.; Robertazzi, A.; Platts, J. A.; Hibbs, D. E.; Williams, P. A. *J. Comput. Chem.* **2006**, *27*, 491−504.
(9) Rothlisberger, U.; Lin, I.-C.; von Lilienfeld, A.; Coutinho-Neto, M. D.; Tavernelli, I. *J. Phys. Chem. B* **2007**, *111*, 9.
(10) Battaglia, M. R.; Buckingham, A. D.; Williams, J. H. *Chem. Phys. Lett.* **1981**, *78*, 3.
(11) Cardamone, S.; Hughes, T. J.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2014**, *16*, 10367−10387.
(12) Hobza, P.; Kolar, M.; Jurecka, P. *ChemPhysChem* **2010**, *11*, 10.
(13) Spiwok, V.; Lipovová, P.; Skálová, T.; Vondráčková, E.; Dohnálek, J.; Hašek, J.; Králová, B. *J. Comput. Aid Mol. Design* **2006**, *19*, 15.
(14) Allinger, N. L. Force Fields: MM3. In *Encyclopedia of computational chemistry*; von Rague Schleyer, P., Ed.; John Wiley & Sons: New York, 1998; Vol. *2*; pp 1028−1033.
(15) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(16) Okamoto, Y.; Yoda, T.; Sugita, Y. *Chemical Physical Letters* **2004**, *386*, 18.
(17) Ponder, J. W.; Wu, C.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio, R. A. J.; Head-Gordon, M.; Clark, G. N. I.; Johnson, M. E.; Head-Gordon, T. *J. Phys. Chem. B* **2010**, *114*, 2549−2564.
(18) Vinter, J. G. *J. Comput. Aided Mol. Des.* **1994**, *8*, 653−668.
(19) Gresh, N.; Kafafi, S. A.; Truchon, J.-F.; Salahub, D. R. *J. Comput. Chem.* **2004**, *25*, 823−834.
(20) Gresh, N. *J. Comput. Chem.* **1995**, *16*, 856−882.
(21) Piquemal, J.-P.; Chelli, R.; Procacci, P.; Gresh, N. *J. Phys. Chem. A* **2007**, *111*, 8170−8176.
(22) Piquemal, J.-P.; Cisneros, G. A.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.
(23) Hunter, C. A.; Chessari, G.; Low, C. M. R.; Packer, M. J.; Vinter, J. G.; Zonta, C. *Chem.—Eur. J.* **2002**, *10*, 8.
(24) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712−725.
(25) Case, D. A.; Darden, T. A.; Cheatham, T. E., III; Simmerling, C. L.; Wang, J.; Duke, R.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S.; Tsui, V.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Mathews, D. H.; Schafmeister, C.; Ross, W. S.; Kollman, P. A. *AMBER 9*; University of California: San Francisco, 2006.
(26) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 14.
(27) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187−217.
(28) Mackerell, J. A. D.; Macias, A. T. *J. Comput. Chem.* **2005**, *26*, 1452−1463.
(29) Sherrill, C. D. *Acc. Chem. Res.* **2012**, *46*, 1020−1028.
(30) Cubero, E. F.; Luque, J.; Orozco, M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5.
(31) Jorgensen, W. L. *J. Chem.Theory Comp.* **2007**, *3*, 1877.
(32) Mills, M. J. L.; Popelier, P. L. A. *Theor. Chem. Acc.* **2012**, *131*, 1137−1153.
(33) Kandathil, S. M.; Fletcher, T. L.; Yuan, Y.; Knowles, J.; Popelier, P. L. A. *J. Comput. Chem.* **2013**, *34*, 1850−1861.
(34) Faerman, C. H.; Price, S. L. *J. Am. Chem. Soc.* **1990**, *112*, 4915−4926.
(35) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem.Phys.Lett.* **1995**, *238*, 253−260.
(36) Tafipolsky, M.; Engels, B. *J. Chem. Theory Comput.* **2011**, *7*, 1791−1803.
(37) Freitag, M. A.; Gordon, M. S.; Jensen, J. H.; Stevens, W. J. *J. Chem. Phys.* **2000**, *112*, 7300−7306.
(38) Mills, M. J. L.; Popelier, P. L. A. *Comput.Theor.Chem.* **2011**, *975*, 42−51.
(39) Handley, C. M.; Hawe, G. I.; Kell, D. B.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6365−6376.
(40) Bartok, A.; Payne, M. C.; Kondor, R.; Csanyi, G. *Phys. Rev. Lett.* **2010**, *104*, 136403.
(41) Rupp, M.; Tkatchenko, A.; Mueller, K.-R.; von Lilienfeld, O. A. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(42) Snyder, J. C.; Rupp, M.; Hansen, K.; Mueller, K. R.; Burke, K. *Phys. Rev. Lett.* **2012**, *108*, 253002−1−5.
(43) Handley, C. M.; Popelier, P. L. A. *J. Phys. Chem. A* **2010**, *114*, 3371−3383.
(44) Behler, J. *J. Phys.: Condens. Matter* **2014**, *26*, 183001.
(45) Popelier, P. L. A. *AIP Conf.Proc.* **2012**, *1456*, 261−268.
(46) Popelier, P. L. A. A generic force field based on Quantum Chemical Topology. In *Modern Charge-Density Analysis*; Gatti, C., Macchi, P., Eds.; Springer: Germany, 2012; Vol. *14*; pp 505−526.
(47) Bader, R. F. W. *Atoms in Molecules. A Quantum Theory*; Oxford Univ. Press: Oxford, Great Britain, 1990.
(48) Popelier, P. L. A. *Atoms in Molecules. An Introduction*; Pearson Education: London, Great Britain, 2000.
(49) Popelier, P. L. A. The Quantum Theory of Atoms in Molecules. In *The Nature of the Chemical Bond Revisited*; Frenking, G., Shaik, S., Eds.; Wiley-VCH, Chapter 8, 2014; pp 271−308.
(50) Rafat, M.; Devereux, M.; Popelier, P. L. A. *J. Mol. Graphics Modell.* **2005**, *24*, 111−120.
(51) Rafat, M.; Popelier, P. L. A. *J. Comput. Chem.* **2007**, *28*, 2602−2617.
(52) Popelier, P. L. A. Quantum Chemical Topology: on Descriptors, Potentials and Fragments. In *Drug Design Strategies: Computational Techniques and Applications*; Banting, L., Clark, T., Eds.; Roy. Soc. Chem., 2012; Vol. *20*, Chapter 6, pp 120−163.
(53) Matta, C. F.; Boyd, R. J. *The Quantum Theory of Atoms in Molecules. From Solid State to DNA and Drug Design.*; Wiley-VCH: Weinheim, Germany, 2007.
(54) Popelier, P. L. A.; Brémond, É. A. G. *Int. J. Quantum Chem.* **2009**, *109*, 2542−2553.
(55) Popelier, P. L. A. Quantum Chemical Topology: on Bonds and Potentials. *In Structure and Bonding. Intermolecular Forces and Clusters*; Wales, D. J., Ed.; Springer: Heidelberg, Germany, 2005; Vol. *115*; pp 1−56.
(56) Ochterski, J. W. *Vibrational Analysis in Gaussian*. http://www.gaussian.com/g_whitepap/vib.htm, 1999.
(57) Neff, M.; Rauhut, G. *Spectrochimica Acta Part A* **2014**, *119*, 100−106.
(58) Watson, J. K. G. *Mol. Phys.* **1968**, *15*, 479−490.

(59) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A.; , J.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, Revision C.02 ed.; Gaussian, Inc.: Wallingford, CT, 2004.
(60) Jensen, F. *J. Chem. Phys.* **2002**, *117*, 9234−9240.
(61) Keith, T. A. *program AIMAll*, 11.04.03 ed.; 2011; aim.tkgristmill.com.
(62) Aicken, F. M.; Popelier, P. L. A. *Can. J. Chem.* **2000**, *78*, 415−426.
(63) Rafat, M.; Popelier, P. L. A. *J. Chem. Phys.* **2006**, *124*, 144102−1−7.
(64) Popelier, P. L. A.; Joubert, L.; Kosov, D. S. *J. Phys. Chem. A* **2001**, *105*, 8254−8261.
(65) Handley, C. M.; Popelier, P. L. A. *J. Chem. Theory Comput.* **2009**, *5*, 1474−1489.
(66) Mills, M. J. L.; Hawe, G. I.; Handley, C. M.; Popelier, P. L. A. *Phys. Chem. Chem. Phys.* **2013**, *15*, 18249−18261.
(67) Matheron, G. *Economic Geology* **1963**, *58*, 21.
(68) Yuan, Y.; Mills, M. J. L.; Popelier, P. L. A. *J. Mol. Model.* **2014**, *20*, 2172−2186.
(69) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press: Cambridge, USA, 2006.
(70) Krige, D. G. *J. Chem., Metal. Mining Soc. S. Afr.* **1951**, *52*, 119−139.
(71) Jones, D. R. *J. Global Optim.* **2001**, *21*, 345−383.
(72) Jones, D. R.; Schonlau, M.; Welch, W. J. *J. Global Optim.* **1998**, *13*, 455−492.
(73) Kennedy, J.; Eberhart, R. C. *Proc. IEEE Int. Conf. Neural Networks* **1995**, *4*, 1942−1948.
(74) Devereux, M.; Popelier, P. L. A. *J. Phys. Chem. A* **2007**, *111*, 1536−1544.
(75) Popelier, P. L. A.; Aicken, F. M. *ChemPhysChem* **2003**, *4*, 824−829.
(76) Popelier, P. L. A.; Aicken, F. M. *J. Am. Chem. Soc.* **2003**, *125*, 1284−1292.
(77) Popelier, P. L. A.; Aicken, F. M. *Chem.—Eur. J.* **2003**, *9*, 1207−1216.
(78) Smith, W.; Leslie, M.; Forester, T. R. *DLPOLY*; CCLRC, Daresbury Lab: Daresbury, Warrington, England, 2003.
(79) Todorov, I. T.; Smith, W. *Philos. Trans. R. Soc. London A* **2004**, *362*, 1835−1852.
(80) Fletcher, T. L.; Kandathil, S. M.; Popelier, P. L. A. *Theor. Chem. Acc.* **2014**, *133*, 1499:1−10.
(81) Blanco, M. A.; Pendas, A. M.; Francisco, E. *J. Chem. Theor. Comput.* **2005**, *1*, 1096−1109.
(82) Popelier, P. L. A.; Kosov, D. S. *J. Chem. Phys.* **2001**, *114*, 6539−6547.
(83) Pendás, A. M.; Blanco, M. A.; Francisco, E. *J. Chem. Phys.* **2006**, *125*, 184112.
(84) Pendas, A. M.; Blanco, M. A.; Francisco, E. *J. Comput. Chem.* **2009**, *30*, 98−109.
(85) Darley, M. G.; Popelier, P. L. A. *J. Phys. Chem. A* **2008**, *112*, 12954−12965.
(86) Garcia-Revilla, M.; Francisco, E.; Popelier, P. L. A.; Martin-Pendas, A. M. *ChemPhysChem* **2013**, *14*, 1211−1218.
(87) Pendas, A. M.; Francisco, E.; Blanco, M. A.; Gatti, C. *Chem.—Eur. J.* **2007**, *13*, 9362−9371.
(88) Tognetti, V.; Joubert, L. *J. Chem. Phys.* **2013**, *138*, 024102.
(89) Tognetti, V.; Joubert, L. *Phys. Chem. Chem. Phys.* **2014**, DOI: 10.1039/c3cp55526g.