JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Open Access on 02/11/2015

Parametrization of an Orbital-Based Linear-Scaling Quantum Force Field for Noncovalent Interactions

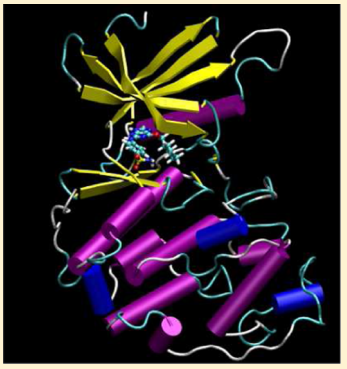
Timothy J. Giese,[†] Haoyuan Chen,[†] Ming Huang,[†,‡] and Darrin M. York*[,†]

[†]BioMaPS Institute and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, New Jersey 08854-8087, United States

[‡]Scientific Computation, University of Minnesota, 207 Pleasant Street SE, Minneapolis, Minnesota 55455−0431, United States

Supporting Information

ABSTRACT: We parametrize a linear-scaling quantum mechanical force field called mDC for the accurate reproduction of nonbonded interactions. We provide a new benchmark database of accurate ab initio interactions between sulfur-containing molecules. A variety of nonbond databases are used to compare the new mDC method with other semiempirical, molecular mechanical, ab initio, and combined semiempirical quantum mechanical/molecular mechanical methods. It is shown that the molecular mechanical force field significantly and consistently reproduces the benchmark results with greater accuracy than the semiempirical models and our mDC model produces errors twice as small as the molecular mechanical force field. The comparisons between the methods are extended to the docking of drug candidates to the Cyclin-Dependent Kinase 2 protein receptor. We correlate the protein−ligand binding energies to their experimental inhibition constants and find that the mDC produces the best correlation. Condensed phase simulation of mDC water is performed and shown to produce O−O radial distribution functions similar to TIP4P-EW.



1. INTRODUCTION

Linear-scaling quantum-mechanical (LS-QM) methods[1] overcome the computational bottlenecks of traditional QM methods without resorting to the use of additional parametrized functions. As a result, they offer a means for modeling system sizes much larger than would otherwise be practical while largely retaining the accuracy of a standard implementation. There are numerous recent examples that further develop LS-QM methods in an effort to make them more affordable;[2−7] however, the widespread use of LS-QM methods continue to be hampered by their high cost relative to combined quantum mechanical/molecular mechanics (QM/MM) or traditional force fields. Furthermore, the LS-QM literature has generally placed greater focus on the benchmark timings of ad hoc systems or on the application of LS-QM methods without having convincingly demonstrated that their added benefit justifies their higher cost. For example, if one must consider relatively large systems to reap the benefits of a LS-QM method, then one will pragmatically choose to use small basis sets and inexpensive Hamiltonians, such as B3LYP/6-31G* or a semiempirical/tight-binding model. But even if a LS-QM method at this level was demonstrated to be computationally feasible for a real application, one may find the accuracy unacceptable or inferior to cheaper parametric alternatives. The GAFF MM force field[8,9] after all, is many times faster than even the fastest semiempirical method and can, for many problems, model intermolecular interactions with acceptable accuracy even though it does not explicitly treat electronic polarization. The utility of LS-QM methods have thus been highlighted by applications that require a detailed description of the electron density or molecular orbitals (MOs), including, the description of biomolecular electrostatic potentials,[10] electron transfer[11−21] and excitation[22] in large systems, and inferring enzyme specificity from frontier orbitals.[23] Some of the most promising examples that demonstrate the utility of LS-QM methods have been provided by Merz, whom has used them to identify misfolded proteins,[24] compute NMR chemical shifts,[25,26] decompose protein−ligand interactions,[27] and perform ligand-binding and drug design.[28−31] Others have used LS-QM methods to examine interactions within clusters,[32−36] compute chemical reaction energies,[37−40] and perform drug screening.[41−43]

Linear-scaling quantum mechanical force fields (QMFFs) are an attempt to circumvent the above criticism by supplementing the LS-QM energy with parametrized interactions. The parametrized functional form provides the means to make a concerted improvement to the method's accuracy and performance while significantly reducing, if not entirely eliminating, their "break-even point"; i.e., the system size required to yield a performance benefit.

There are two categories of QMFFs that have found chemical application to molecular systems: (1) MO-based models, such as X-Pol[44−48] and the water-specific XP3P model,[49] and (2) electron density-based models, e.g. SIBFA, GEM, and QMPFF.[50−61] The MO-based QMFFs, such as those proposed by Gao,[44] achieve their performance by approximating the wave function with a Hartree product of

Received: November 28, 2013
Published: February 11, 2014

antisymmetrized fragment determinants.[62] This approximation results in a block diagonal Fock matrix that can be diagonalized fragment-by-fragment. The number of blocks increases with the size of the system, but the size of each block does not, so the scaling of the Fock matrix diagonalization reduces from $O(N^3)$ to $O(N)$. Lennard-Jones (LJ) or Buckingham potentials are then used to recuperate the energetic effect of interfragment MO coupling.[19,45,46,62−66] The density-based QMFFs avoid Fock matrix diagonalization altogether; the interfragment interactions are computed from the overlap and Coulomb interactions of an accurate density prefitted to ab initio and is allowed to respond to the environment according to the principle of chemical potential equalization. We recently introduced[67] a modified divide-and-conquer (mDC) QMFF model which made an attempt at unifying the orbital- and density-based QMFFs. In that model, the density resulting from fragment MOs were used to interact fragments with a density-overlap van der Waals (vdW) model.[67,68]

The present work is a parametrization of a MO-based mDC QMFF which, like X-Pol, models the interfragment vdW interactions with simple LJ functions; however, there are two distinguishing features that differentiate mDC from X-Pol. The interfragment X-Pol interactions are treated with a QM/MM interaction potential; whereas, the mDC interactions are not. A QM/MM potential does not produce symmetric interactions in the sense that the interaction depends upon which residue is considered the "QM region." X-Pol therefore recomputes the interfragment interactions for each QM fragment and averages the energy and forces. The interfragment mDC interactions are symmetric and are therefore computed once. Second, the mDC fragment densities are not limited to being those used in a QM/MM potential. A QM/MM treatment of DFTB2 and DFTB3 densities would yield atomic charges only; however, this severely degrades the accuracy of these methods. The ability of DFTB2 and DFTB3 to reproduce hydrogen bond angles is a result of their tight-binding coupling matrix elements, not their charges. The Hartree−Product approximation removes these coupling elements and their effect must be recovered by performing electrostatics with higher-order atomic multipoles.[67]

In this work, we parametrize the mDC method for nonbonded interactions; we supply a new database, Sulfur Set 7 (SS7), of accurate nonbond interactions involving sulfur; we apply the parametrized mDC method to the screening of Cyclin-Dependent Kinase 2 (CDK2) protein receptor drug candidates;[41] and we explore the ability of mDC to simultaneously reproduce small-to-medium sized water clusters and the condensed phase water radial distribution function (RDF) computed from molecular dynamics simulation. A series of standard nonbond interaction databases are used to compare the parametrized mDC model to MM force fields, semiempirical models, inexpensive ab initio Hamiltonians, and a semiempirical QM/MM method. The comparisons quantify the poor accuracy of standard semiempirical models relative to a MM force field and reveal the high accuracy of mDC. The CDK2 drug screening exercise is repeated with MM and semiempirical QM/MM to further demonstrate the benefits of our QMFF. The variety of molecular interactions examined in this work and the number of other methods we compare mDC against extend significantly beyond other recent investigations.[48]

2. METHODS

2.1. Linear-Scaling mDC Method. The mDC total energy is

$$E = \sum_A E_A(\mathbf{C}_A^\sigma; \mathbf{R}_A) + \frac{1}{2}\sum_{\substack{a \\ lm\in a}} q_{lm}p_{lm} + {\sum_{b>a}}'\left(\frac{C_{12,ab}}{R_{ab}^{12}} - \frac{C_{6,ab}}{R_{ab}^{6}}\right) + E_{\text{bonded}}(\mathbf{R}) \quad (1)$$

where the first term is a summation of fragment ab initio energies, $\mathbf{C}_A^\sigma$ are the σ-spin MO coefficients for the A'th fragment, and $\mathbf{R}_A$ are the nuclear positions of the atoms in fragment A. The remaining terms describe the interactions between the fragments. The second term is the interfragment electrostatic energy, where

$$q_{lm\in a} = Z_a\delta_{l0}\delta_{m0} - \int \rho_a(\mathbf{r})C_{lm}(\mathbf{r} - \mathbf{R}_a)\, \mathrm{d}^3r \quad (2)$$

are atomic multipole moments on atom a, $C_{lm}(\mathbf{r})$ is a real regular solid harmonic, $\rho_a(\mathbf{r})$ is an atom-partitioned density, Z_a is a nuclear plus core electron charge, and

$$p_{lm\in a} = {\sum_{\substack{b\neq a \\ jk\in b}}}' q_{jk}\frac{C_{lm}(\nabla_a)}{(2l-1)!!}\frac{C_{jk}(\nabla_b)}{(2j-1)!!}\frac{1}{R_{ab}} \quad (3)$$

is a "multipolar potential" arising from point-multipole Coulomb interactions. The primed summation indicates that intrafragment electrostatics are excluded because those Coulomb interactions are already considered in the ab initio calculation of E_A.

The last two terms in eq 1 are an interfragment LJ energy and a standard molecular mechanical (MM) bonded energy, respectively. Unlike a standard MM Hamiltonian, E_{bonded} includes corrections only for those bonds, angles, and dihedrals that cross the boundary between two covalently bonded fragments. The systems examined in the present work involve intermolecular interactions between nonbonded molecules and the intermolecular interaction between drug ligands and a rigid protein; therefore, there is no need to further elaborate on E_{bonded} here.

The computational advantage afforded by eq 1 arises from having treated the MOs of each fragment as if they did not overlap with those of any other fragment. This approximation allows one to solve for the MO coefficients from a series of small generalized eigenvalue problems (proportional to the size of a fragment)

$$\mathbf{F}_A^\sigma\cdot\mathbf{C}_A^\sigma = \mathbf{S}_A\cdot\mathbf{C}_A^\sigma\cdot\mathbf{E}_A^\sigma \quad (4)$$

rather than solving a single eigenvalue problem for the entire system. Although the lack of interfragment coupling between the MOs is a significant approximation, the remaining interactions in eq 1 are parametrized to reproduce high-level ab initio data. The lack of explicit interfragment MO coupling matrix elements does not imply that the fragments are uncoupled. The interfragment coupling occurs through the interaction of their atomic multipoles which are determined from the fragment electron densities within the self-consistent-field (SCF) procedure. The σ-spin Fock matrix for region A with inclusion of this coupling is

$$F^{\sigma}_{A,ij} = \left.\frac{\partial E_A}{\partial P^{\sigma}_{A,ij}}\right|_{\mathbf{q},\mathbf{p},\mathbf{R}} + \sum_{\substack{a\in A\\ lm\in a}} plm \left.\frac{\partial qlm}{\partial P_{A,ij}}\right|_{\mathbf{p},\mathbf{R}} \tag{5}$$

where

$$P^{\sigma}_{A,ij} = \sum_k n^{\sigma}_{A,k} C^{\sigma}_{A,ik} C^{\sigma}_{A,jk} \tag{6}$$

is the spin-resolved AO density matrix of fragment A, and $n^{\sigma}_{A,k}$ is the occupation number of σ-spin orbital k in fragment A.

Our linear-scaling framework minimizes the mDC energy by performing a double-SCF procedure.[44,45,69,70] In brief, each fragment is provided a set of atom-centered external multipolar potentials $\mathbf{p}$ and is instructed to compute a converged fragment energy E_A and return a new set of atomic multipole moments $\mathbf{q}$. The linear-scaling framework evaluates the interfragment electrostatics to transform the moments $\mathbf{q}$ into potentials $\mathbf{p}$, and the procedure is repeated until the total energy and $\mathbf{q}$ stop changing. Once convergence is met, the atomic gradients are computed from

$$\frac{\mathrm{d}E}{\mathrm{d}X_a} = \sum_A \left.\frac{\partial E_A}{\partial X_a}\right|_{\mathbf{q},\mathbf{p}} + \frac{1}{2}\sum_{lm\in a} q_{lm} \left.\frac{\partial p_{lm}}{\partial X_a}\right|_{\mathbf{q}} + {\sum_{b\neq a}}' \left(3\frac{C_{6,ab}}{R^8_{ab}} - 6\frac{C_{12,ab}}{R^{14}_{ab}}\right) X_{ab} + \frac{\mathrm{d}E_{\mathrm{bonded}}}{\mathrm{d}X_a} \tag{7}$$

The interface between the linear-scaling framework and the underlying Hamiltonian must extract or calculate the atomic multipole moments $\mathbf{q}$ from the ab initio density, and it must correct the model's Fock matrix [eq 5] by evaluating the derivative $\partial q_{lm}/\partial P^{\sigma}_{A,ij}\,|_{\mathbf{p},\mathbf{R}}$.

In the present work, we use the DFTB3 Hamiltonian (parametrization "3ob") described in ref 71. DFTB3 happens to use atomic charges in the calculation of the energy, but our interface to DFTB3 is free to choose a different set of $\mathbf{q}$'s that are used to interact a fragment with the outside world. This is a particularly useful feature when using DFTB3 because its success in achieving good hydrogen bond (H-bond) angles is largely a result of the intermolecular tight-binding matrix elements[67] (e.g., Figure 1). The linear-scaling framework eliminates these elements, so our approach is to recapture the angular dependence via the electrostatic interactions by increasing the order of the atomic multipoles. A mDC fragment can thus be thought as having two representations of the atomic multipoles: a set of atomic charges that DFTB3 happens to use internally and a set of $\mathbf{q}$'s that our interface constructs from the fragment's total density matrix $\mathbf{P}_A = \mathbf{P}^{\alpha}_A + \mathbf{P}^{\beta}_A$.

We decompose the DFTB3 density into atomic components as follows:

$$\rho_a(\mathbf{r}) = \sum_{ij\in a} P_{A,ij}\chi_i(\mathbf{r})\chi_j(\mathbf{r}) + \sum_{b\neq a} f_{ab}(b_{ab}) \sum_{\substack{i\in a\\ j\in b}} P_{A,ij}\chi_i(\mathbf{r})\chi_j(\mathbf{r}) \tag{8}$$

where $\chi_i(\mathbf{r}) = \chi_i(r)Y_{l,m_i}(\Omega)$ is an AO basis function, and

$$f_{ab}(b_{ab}) = f^{\mathrm{s}}_{ab} + S_{on}\left(\frac{b_{ab} - b^{\mathrm{s}}_{ab}}{b^{\mathrm{d}}_{ab} - b^{\mathrm{s}}_{ab}}\right)\left(f^{\mathrm{d}}_{ab} - f^{\mathrm{s}}_{ab}\right) \tag{9}$$

is a fraction between 0 and 1 and holds the property $f_{ab} = 1 - f_{ba}$. If $f_{ab} = 1/2$, then one would say that eq 8 is a Mulliken partitioning; however, we allow for a bias as a function of atom-pair and Mulliken bond-order

$$b_{ab} = 2\sum_{\substack{i\in a\\ j\in b}} P_{A,ij}S_{A,ij} \tag{10}$$

f^{s}_{ab} and f^{d}_{ab} are parameters corresponding to the fractions that we found most agreeable when atoms a and b are connected by a single bond and double bond, respectively; and b^{s}_{ab} and b^{d}_{ab} are the Mulliken single- and double-bond orders, i.e., $f^{\mathrm{s}}_{ab} \equiv f_{ab}(b^{\mathrm{s}}_{ab})$ and $f^{\mathrm{d}}_{ab} \equiv f_{ab}(b^{\mathrm{d}}_{ab})$. Had we expressed the partition using Wigner bond indices, b^{s}_{ab} and b^{d}_{ab} would take values 1 and 2, respectively; however, Mulliken bond orders yield convenient expressions for the Fock matrix corrections.

$$S_{\mathrm{on}}(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 1 \\ 10x^3 - 15x^4 + 6x^5 & \text{otherwise} \end{cases} \tag{11}$$

is a smooth polynomial used to switch f_{ab} from f^{s}_{ab} to f^{d}_{ab} as the bond order increases.

The atomic multipoles are obtained by inserting eq 8 into eq 2 while restricting the contributions of the two-center densities to charge only

$$q'_{lm} = \begin{cases} Z_a - b_{aa}/2 - \sum_{\substack{b\in A\\ b\neq a}} f_{ab}(b_{ab})b_{ab} & \text{if } l = 0 \\ \sum_{ij\in a} P^A_{ij} M^{(l)}_{ij} \sqrt{\dfrac{4\pi}{2l+1}} \times & \text{if } l > 0 \\ \quad \int Y_{lm}(\Omega)Y_{l,m_i}(\Omega)Y_{l,m_j}(\Omega)\,\mathrm{d}\Omega & \end{cases} \tag{12}$$

where the

$$M^{(l)}_{ij} = \int_0^{\infty} \chi_i(r)\chi_j(r)r^{2+l}\,\mathrm{d}r \tag{13}$$

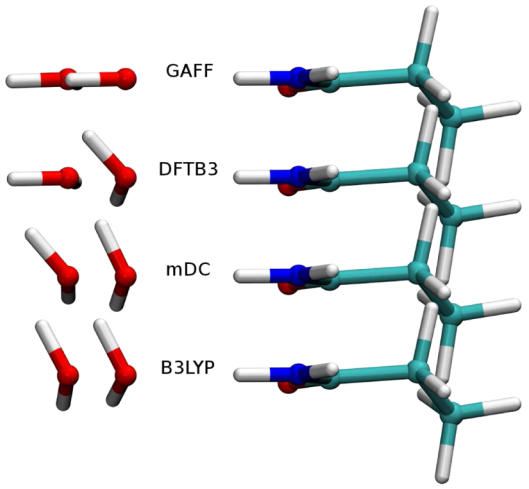


Figure 1. Comparison between optimized TIP3P/GAFF, DFTB3, mDC, and B3LYP/6-31++G** water−aspargine H-bond angle as a function of N···O separation (3 and 4 Å). The TIP3P water is coplanar with the ASN amine group. The DFTB3 water is angled in a manner similar to B3LYP near the energy minimum but reverts to a TIP3P-like planar structure when the inter-residue overlap becomes small. The mDC angle is in close agreement with B3LYP even at larger separations than those shown.

dx.doi.org/10.1021/ct401035t | J. Chem. Theory Comput. 2014, 10, 1086−1098

are treated as a parameters. For an *sp*-basis, there are two parameters: $M_{sp}^{(1)}$ and $M_{pp}^{(2)}$, which control the magnitude of the dipole and quadrupole contributions, respectively. When the multipole expansion is limited to quadrupoles, an *spd*-basis formally requires three additional parameters: $M_{sd}^{(2)}$, $M_{pd}^{(1)}$, and $M_{dd}^{(2)}$; however, we constrain these additional parameters to the values of $M_{pp}^{(2)}$, $M_{sp}^{(1)}$, and $M_{pp}^{(2)}$, respectively.

If the fragment is not covalently bonded to any other fragment, which is the predominant scenario in the present work, then eq 12 is the expression for the multipoles, i.e., $q_{lm} = q'_{lm}$. The CDK2 protein docking results described in this paper, however, use a fixed protein structure that has been fragmented by residue. Connection atoms are used to cap the resulting "dangling bonds", and we have found it necessary to treat the atomic multipoles at the fragmentation boundaries specially, as described in Figure 2. For the purposes of the present work, we

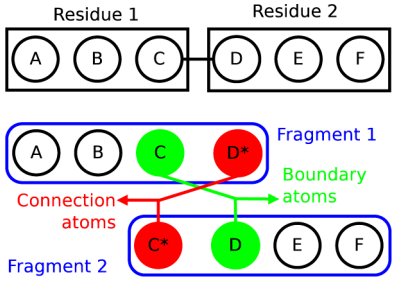


Figure 2. Illustration of the relationship between residues and fragments and the relationship between boundary atoms and connection atoms in the mDC method. Atoms C and D are connected by a covalent bond; therefore, fragments 1 and 2 include pseudoatom D* and C*, respectively. When fragment 1 is queried for the new set of multipole moments of atoms A, B, and C; the charges of atoms C and D* are summed; the multipole moments of atom D* are set to zero; the charge of atom C is assigned a predefined fixed MM charge; and the excess charge from the sum is evenly distributed to atoms A and B. An analogous procedure is performed for fragment 2 to fix the charges of atoms C* and D. Thus, the charges of atoms C and D are fixed and independent of the ab initio density.

found it acceptable to use hydrogens for the connection atoms. The charges of the frontier atoms, i.e., the connection and boundary atoms, need to be constrained to yield reliable SCF convergence. Although it is possible to modify the ab initio method by introducing additional constraints into the density matrix, it is much simpler and less intrusive to compute the unconstrained atomic multipole moments from eq 12, manually change the charges of the frontier atoms, and renormalize the remaining charges in the fragment, i.e.,

$$q_{lm\in a} = \begin{cases} 0 & a \text{ is a connection atom} \\ \delta_{l0}\delta_{m0} q_{\mathrm{mm},a} & a \text{ is a boundary atom} \\ q'_{lm} + \delta_{l0}\delta_{m0}\langle \Delta q\rangle_A & \text{otherwise} \end{cases} \tag{14}$$

where $q_{\mathrm{mm},a}$ is a fixed molecular mechanical charge and

$$\langle \Delta q\rangle_A = \frac{\sum_b^{N_{\mathrm{b},A}} (q'_{(00)_b} - q_{\mathrm{mm},b}) + \sum_c^{N_{\mathrm{c},A}} q'_{(00)_c}}{N_{\mathrm{a},A} - N_{\mathrm{b},A} - N_{\mathrm{c},A}} \tag{15}$$

is the charge excess to be evenly distributed among the nonfrontier atoms in the region after altering the charges of the frontier atoms. $N_{\mathrm{a},A}$, $N_{\mathrm{b},A}$, and $N_{\mathrm{c},A}$ are the total number of atoms, boundary atoms, and connection atoms in fragment *A*, respectively. The summations over *b* and *c* in eq 15 are meant to encompass the boundary and connection atoms, respectively. With this scheme, the derivative required to correct the Fock matrix requires one additional chain-rule:

$$\left.\frac{\partial q_{lm\in a}}{\partial P^{\sigma}_{A,ij}}\right|_{\mathbf{p},\mathbf{R}} = \sum_{\substack{b \\ jk\in b}} \frac{\mathrm{d}q_{lm}}{\mathrm{d}q'_{jk}} \left.\frac{\partial q'_{jk}}{\partial P^{\sigma}_{A,ij}}\right|_{\mathbf{p},\mathbf{R}} . \tag{16}$$

2.2. Model Hamiltonians. *DFTB3.* This method is described in ref 71 and performed using a standard SCF procedure with the recently developed "3ob" parametrization. We do not use the "H−H-mod", "H−N-mod", nor spin-polarization corrections described in ref 71 because these are intended to improve atomization energies, proton affinities, and high-spin open-shell electronic configurations that are not relevant here.

DFTB2. This method (sometimes called SCC-DFTB) is described in ref 72 and performed using the "mio" parametrization. We do not use the γ_{h} correction that appears in more recent works[73−75] which has been shown in improve H-bond strengths. Nor do we supplement DFTB2 with a London dispersion model for select nonbonded interactions, which has been shown to improve the description of weakly bound complexes.[76]

PM6 and AM1. These semiempirical methods are described in refs 77 and 78, respectively; as implemented in Gaussian 09.[79]

B3LYP. This method refers to the B3LYP/6-31G* density functional model implemented in Gaussian 09.[79]

GAFF. This comprises the general Amber force field[8,9] and TIP3P water. The CDK2 protein structure made use of the ff99 Amber force field[80] supplemented with parameters from ref 81 for the phosphotyrosine with a single protonated phosphate group. Figure 7 makes use of the TIP4P-EW water model.[82]

MNDO/MM. Semiempirical QM/MM calculations performed with Amber 12.[9] Clusters of *N* monomers were geometry optimized *N* times to produce *N* relative energies and coordinate root-mean-square deviations relative to the reference structure. The average relative energy and coordinate root-mean-square is then used to avoid the ambiguity assigning the QM region. This procedure is *not* used by X-Pol; it is merely our attempt to compare to a traditional semiempirical QM/MM method. The CDK2 calculations treat the ligand and protein with MNDO and Amber ff99, respectively.

2.3. Reference Data. *S22.* A database of 22 dimers is taken from ref 83. The set contains small to relatively large complexes chosen to represent hydrogen bonding, dispersion, and mixed electrostatic-dispersion interactions. The S22 reference is not a collection from a consistent level of theory but is instead a collection of the best available calculations that could be afforded at the time of its construction, and the reader should refer to ref 83 for full details. Generally speaking, most of the energies are an approximate CCSD(T) complete basis extrapolation (CBS) from counterpoise (CP) corrected calculations. Both the monomers are dimers are geometry optimized.

JSCH-2005. A collection of 124 nucleobase and amino acid complexes compiled from several publications.[83] In this work, we refer only to a small subset of the JSCH-2005 database where optimized geometries of hydrogen bonded or stacked nucleobases are available. In particular, we use the A···T S1, A···

dx.doi.org/10.1021/ct401035t | J. Chem. Theory Comput. 2014, 10, 1086−1098

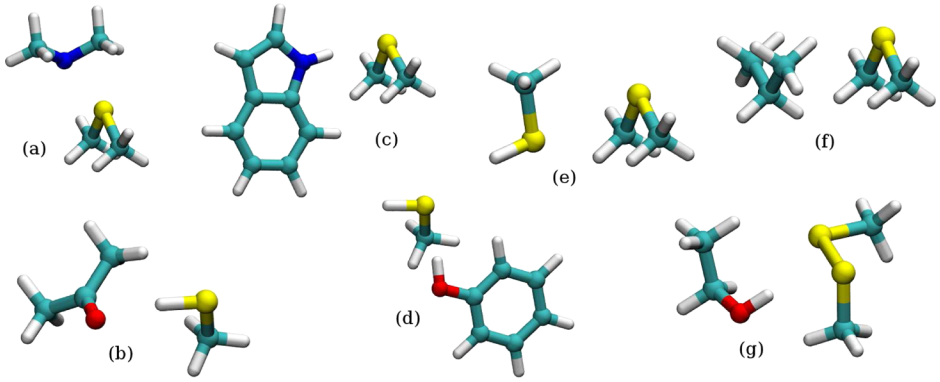


Figure 3. Dimers used in the SS7 database. The interaction energies (kcal/mol) using the multilevel method described in the text are (a) −3.652, (b) −3.737, (c) −6.036, (d) −6.473, (e) −3.683, (f) −2.643, and (g) −5.191. Coordinates are provided in the Supporting Information.

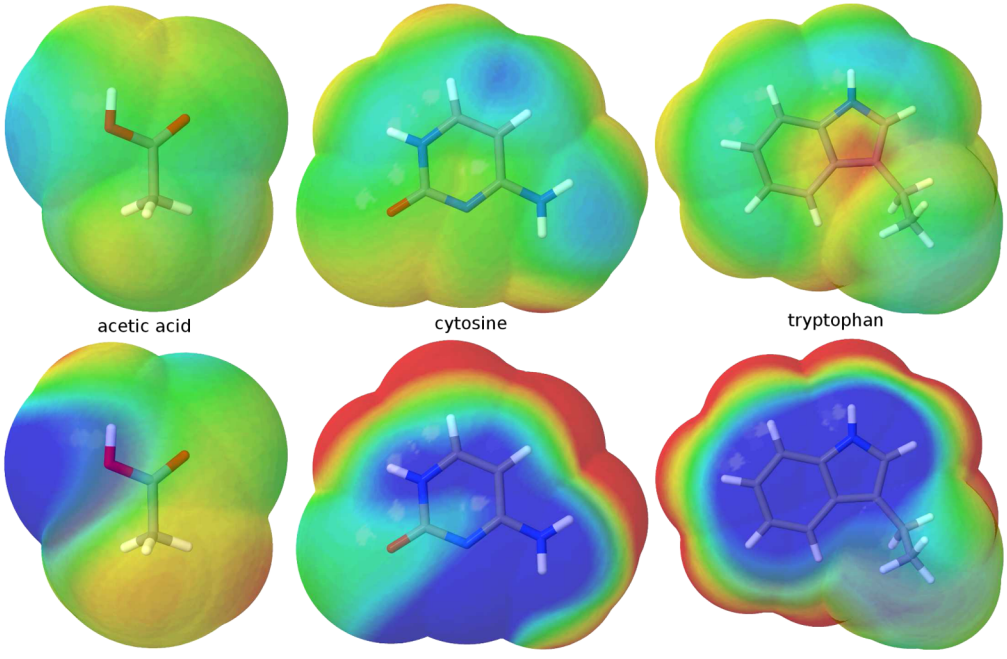


Figure 4. Electrostatic potential difference maps between B3LYP/6-311++G** and mDC (top row) and DFTB3 (bottom row) evaluated on the solvent accessible surface of acetic acid, cytosine, and tryptophan. Blue and red indicate that the model requires more negative charge and more positive charge, respectively, relative to B3LYP/6-311++G**. Green indicates agreement between the electrostatic potential of the model and B3LYP/6-311++G**. The colors are bounded by the range of ±0.003 au.

T WC, G···C S, G···C wc, mA···mT S, and mG···mC S complexes from ref 84 [CCSD(T)/CBS//MP2/TZVPP] and the G···A1, G···A1 pl, G···A2, G···A2 pl, G···A3, and G···A4 complexes from ref 85 [MP2/CBS//MP2/cc-pVTZ]. Both the monomers and dimers are geometry optimized.

SCAI. A set of 24 pairs of amino acid side chain interactions whose structures as taken from X-ray crystal structures.[86] The reference data was obtained[86] at CCSD(T)/CBS//TPSS/TZVP. The dimer geometries in this database are constructed by optimizing the hydrogen positions while fixing the heavy atom positions to the experimental structure. The monomer geometries are taken from the partially optimized dimer structure. This procedure is performed for each of the methods compared, as opposed to having used the reference geometries without further optimization.

S66. A set of 66 complexes[87] considered to be a more representative sampling of the distribution of interaction motifs than the S22 database. Unlike S22, S66 uses a consistent reference methodology for geometry optimizations (MP2/cc-pVTZ) and single point energy refinement using a procedure described in detail below. Furthermore, the S66 monomer structures are evaluated using the dimer optimized structures instead of having included the monomer deformation energy. Comparison to S66 will be made in two different ways: "S66 (Fixed)" and "S66 (Procedure)", where S66 (Fixed) computes the interactions using the reference geometries and S66 (Procedure) optimizes the dimers for each method and uses the monomer structures from the resulting optimizations.

SS7. A new set of 7 sulfur-containing complexes (sulfur set 7) presented as a part of this work (Figure 3). These complexes were chosen by extracting sulfur-containing residues from a variety of crystal structures deposited in the PDB databank, modeling the residues with small molecules, and geometry optimizing for a minimum. In other words, the complexes are not guaranteed to represent a global minimum configuration nor do they remain in the orientation observed in the crystal structures. The protocol we follow is analogous to S66: the dimers are optimized at MP2/cc-pVTZ and the interaction

Table 1. Multipole Moment Comparison between B3LYP/6-311++G//B3LYP/6-31++G** and Various Models for Amino Acids, Their Side Chains, and Nucleobases[a]**

			$\langle\lvert\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(1)}_{\mathrm{ref}}\rvert\rangle$					$\langle\lvert\boldsymbol{\mu}^{(2)} - \boldsymbol{\mu}^{(2)}_{\mathrm{ref}}\rvert\rangle$			
category	N	$\langle\mu^{(1)}_{\mathrm{ref}}\rangle$	mDC	DFTB3	DFTB2	GAFF	$\langle\mu^{(2)}_{\mathrm{ref}}\rangle$	mDC	DFTB3	DFTB2	GAFF
side-chains	20	1.36	**0.11**	0.13	0.17	0.16	8.68	**1.19**	1.90	2.08	1.22
amino acids	20	2.12	**0.15**	0.17	0.19	0.48	29.90	**1.97**	2.45	2.78	5.51
nucleobases	10	1.98	**0.07**	0.21	0.36	0.25	14.58	**1.22**	3.26	3.77	2.29
total	50	1.79	**0.12**	0.16	0.22	0.31	18.35	**1.51**	2.39	2.70	3.15

[a]The model multipole moments are computed at the B3LYP/6-31++G** optimized geometries. $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ are the molecular dipole and quadrupole multipole moment vectors of length 3 and 5, respectively. $\langle\mu^{(n)}_{\mathrm{ref}}\rangle$ is the magnitude of the multipole moment vector averaged over molecules and $\langle\lvert\boldsymbol{\mu}^{(n)} - \boldsymbol{\mu}^{(n)}_{\mathrm{ref}}\rvert\rangle$ is the magnitude of the multipole moment error vector averaged over molecules. Boldface is used to highlight the best value among the models.

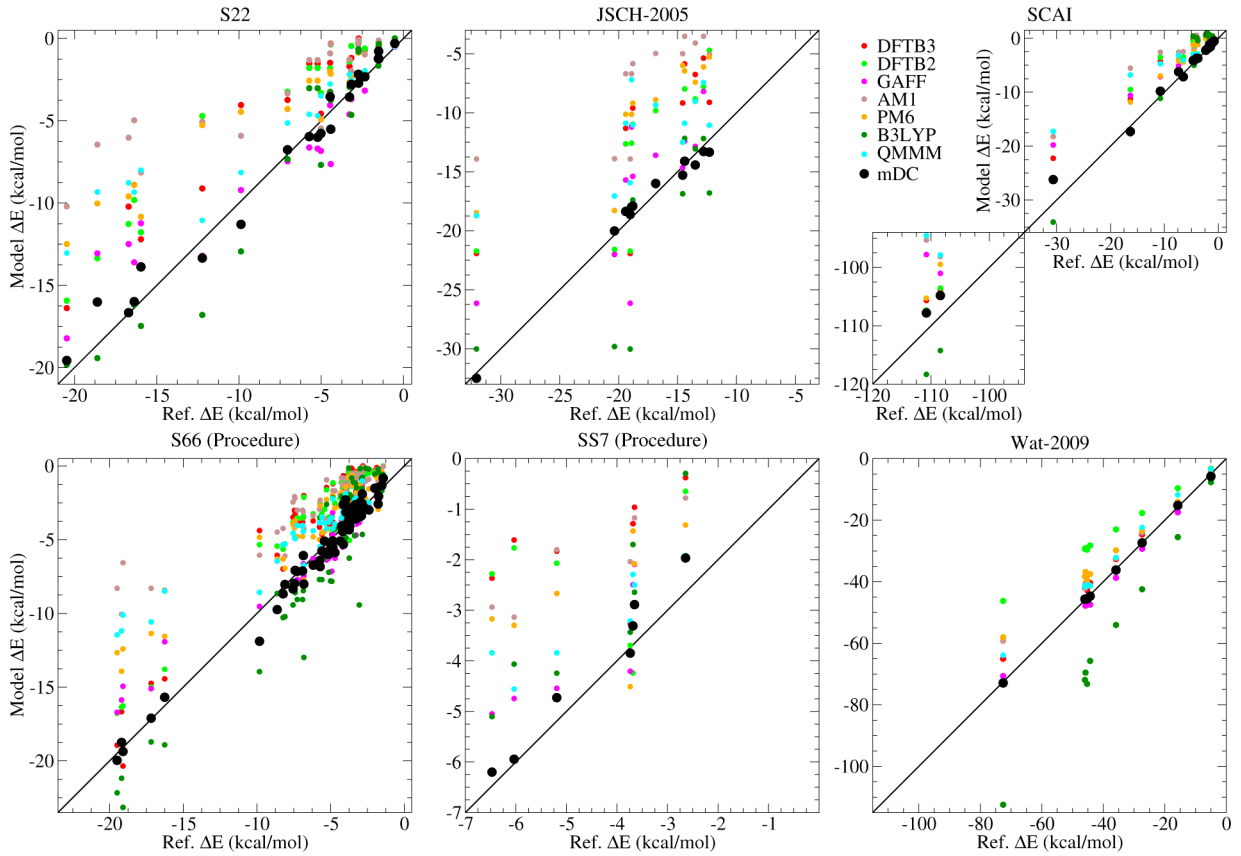


Figure 5. Comparison of mDC with other Hamiltonians using several databases of intermolecular interactions.

energy is computed using the monomer structures from the optimized dimer. The single point refinement energy is

$$E = \mathrm{HF/aQZ} + \Delta\mathrm{MP2/CBS} + \Delta\Delta\mathrm{CCSD(T)/aDZ} \tag{17}$$

where all energies are CP-corrected; a*X*Z denotes the aug-cc-pV*X*Z basis; ΔΔCCSD(T) is the correlation energy difference between CCSD(T) and MP2; and ΔMP2/CBS is the MP2 correlation energy in the CBS limit, as estimated from Halkier's two-point extrapolation formula[88]

$$\Delta\mathrm{MP2/CBS} = (64\Delta\mathrm{MP2/aQZ} - 27\Delta\mathrm{MP2/aTZ})/37 \tag{18}$$

Comparisons to SS7 will be denoted "SS7 (Fixed)" and "SS7 (Procedure)" using the same protocol described for S66.

Wat-2009. The reference structures and binding energies (CP-corrected MP2/CBS with a CCSD(T)/aDZ correction) of $(H_2O)_n$ n = 2−6 and 8 were taken from the database compiled[89−91] in ref 92.

Wat-2011. The reference structures and binding energies (RI-MP2/CBS with a CCSD(T)/aDZ correction) of $(H_2O)_n$ n = 2−10 where taken from ref 93. Unlike Wat-2009, these energies do not explicitly include CP-corrections, but the CBS extrapolation was parametrized to well-reproduce CP-corrected Δ*E* energies for a select set of clusters. As a result, the water cluster Δ*E* binding energies in this database are similar to those in the Wat-2009 database. The purpose for using Wat-2011 in the present work is to examine the relative ordering of clusters of a given cluster size (ΔΔ*E*'s); a purpose for which the Wat-2009 database does not adequately fullfill. It is not immediately clear to what extent explicit CP-corrections might effect the ΔΔ*E*'s, but we suspect that the general trends in ordering will be similar. We therefore limit our analysis to Pearson

dx.doi.org/10.1021/ct401035t | J. Chem. Theory Comput. 2014, 10, 1086−1098

Table 2. Statistical Summary of mDC and Other Hamiltonians for Several Databases of Intermolecular Interactions[a]

			model							
DB	*N*	property	mDC	DFTB3	DFTB2	GAFF	AM1	PM6	B3LYP	MNDO/MM
S22	22	mue	**0.699**	3.012	3.338	1.588	4.401	3.226	1.336	2.504
		mse	0.184	3.012	3.338	0.546	4.361	3.226	0.016	2.393
		R	**0.989**	0.964	0.973	0.953	0.910	0.973	0.972	0.939
		crms	**0.180**	0.579	0.277	0.255	0.924	0.407	0.843	0.467
S66 (Fixed)	66	mue	**0.545**	2.737	2.985	0.913	5.141	2.666	2.252	2.195
		mse	0.275	2.737	2.985	0.708	5.141	2.666	1.332	2.183
		R	**0.982**	0.957	0.973	0.958	0.521	0.965	0.930	0.899
S66 (Procedure)	66	mue	**0.532**	2.363	2.399	0.825	3.061	2.006	1.836	1.543
		mse	−0.130	2.313	2.353	0.192	3.041	1.934	−0.142	1.505
		R	**0.988**	0.961	0.968	0.960	0.890	0.949	0.940	0.952
		crms	**0.237**	1.774	1.484	0.330	0.976	0.693	0.723	0.532
SS7 (Fixed)	7	mue	**0.657**	3.184	2.885	1.547	3.573	3.015	2.121	2.034
		mse	0.657	3.184	2.885	1.547	3.573	3.015	2.121	2.034
		R	**0.964**	0.452	0.346	0.836	0.799	0.585	0.869	0.815
SS7 (Procedure)	7	mue	**0.393**	2.817	2.248	1.039	2.549	2.070	1.418	1.321
		mse	0.361	2.817	2.089	0.904	2.549	1.848	1.418	1.321
		R	**0.984**	0.433	−0.041	0.874	0.903	0.472	0.901	0.908
		crms	**0.251**	0.361	0.363	0.620	0.707	0.523	0.516	0.701
SCAI	23	mue	**0.819**	2.843	2.745	2.011	4.249	2.518	2.361	2.990
		mse	0.668	2.843	2.745	1.940	4.249	2.518	0.646	2.987
		R	**1.000**	0.998	0.998	0.997	0.995	0.999	0.999	0.995
JSCH-2005	12	mue	**0.697**	6.582	6.293	3.417	10.488	7.346	3.052	5.991
		mse	0.096	5.825	5.637	1.788	10.488	7.346	−1.493	5.991
		R	**0.989**	0.754	0.784	0.749	0.756	0.786	0.750	0.697
		crms	**0.183**	1.005	0.644	0.507	1.294	0.713	1.023	0.509
Wat-2009	9	mue	**0.352**	3.424	13.547	1.890	4.354	6.223	20.499	4.338
		mse	−0.100	3.424	13.547	−1.456	4.253	6.223	−20.499	4.338
		R	**1.000**	0.999	1.000	0.998	0.995	0.998	0.998	0.998
		crms	**0.073**	0.121	0.106	0.309	1.029	0.554	0.130	0.449

[a]The terms mue and mse are the mean unsigned and mean signed error in ΔE (kcal/mol). *R* is the Pearson correlation coefficient between the reference and model ΔE (unitless). crms is the average coordinate root mean square deviation of the optimized dimer structures to the reference structures (Å). Boldface is used to highlight the best value amongst the models.

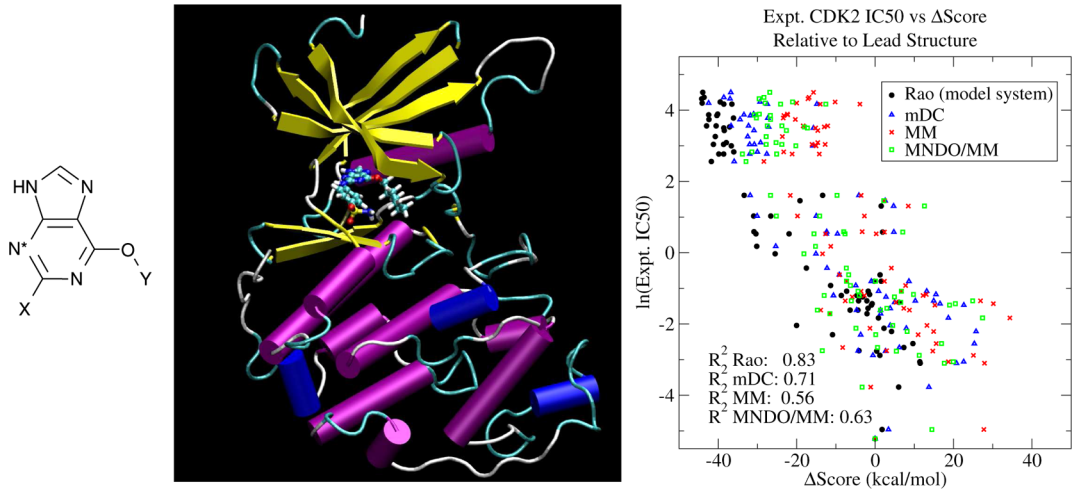


Figure 6. (left) Base structure common to all ligands. The atom marked "N*" is used to compare the position of a bound ligand to the lead structure, as explained in the text. (center) Lead ligand bound to the CDK2 protein. (right) Correlation between the experimental protein inhibition constant measured for each ligand and the assigned score relative to the lead structure. Rao indicate the gas-phase DC-DFT results taken from ref 41 computed using a small model system of CDK2.

correlations of $\Delta\Delta E$'s which tend to be either "good" or "very bad."

2.4. Parametrization. The mDC model makes use of A_{ab} and B_{ab} LJ parameters (two parameters per atom-type using Lorentz−Berthelot combining rules); $M_{sp}^{(1)}$ and $M_{pp}^{(2)}$ multipole parameters (two parameters per atomic number); f_{ab}^{s}, f_{ab}^{d}, b_{ab}^{s}, and b_{ab}^{d} charge mapping parameters (four parameters per atomic number pair); and the boundary atom $q_{mm,a}$ MM point

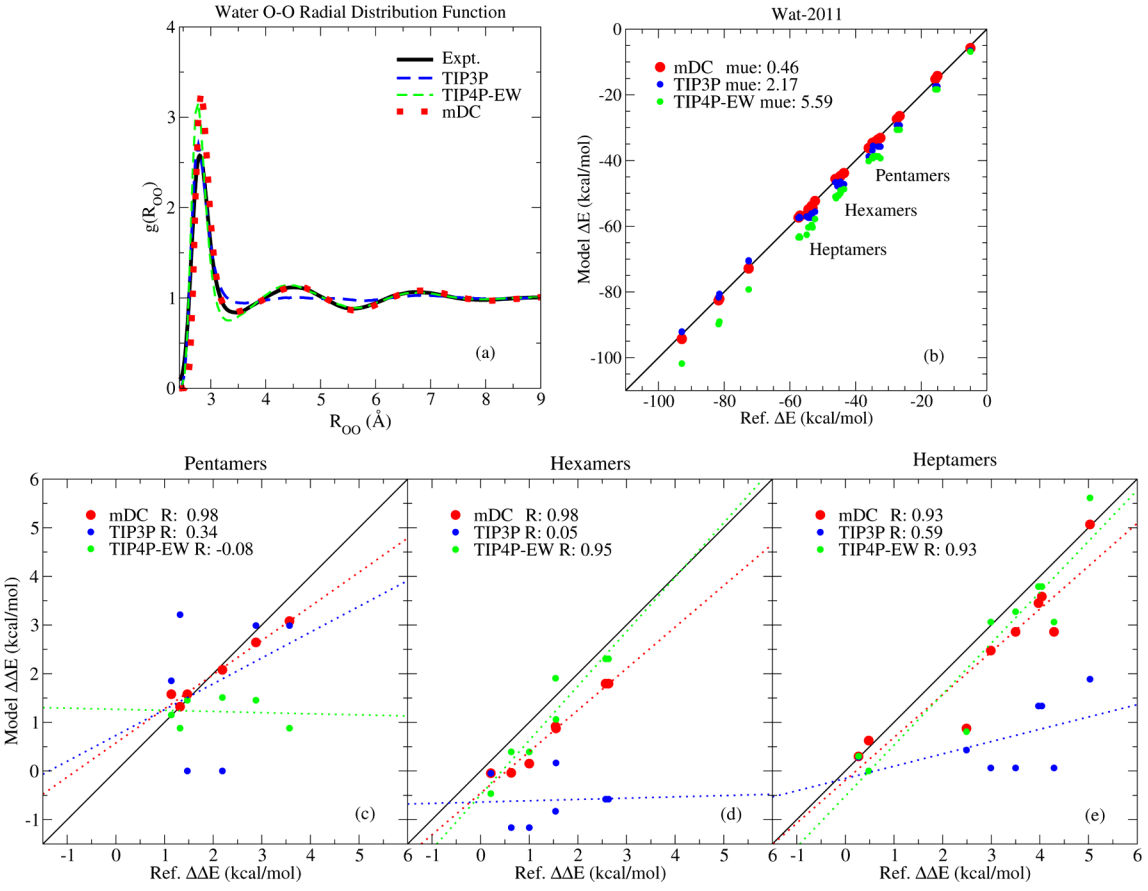


Figure 7. Comparison of water O−O radial distribution functions (a), water cluster binding energies (b), and the relative ordering of clusters (c−e). The mue's in b are mean unsigned errors of ΔE (kcal/mol). The R's in c−e are Pearson correlation coefficients.

charges, which are obtained along with the atom-type assignments from the AmberTools suite of packages.[9] Although we refer to b^{s}_{ab} and b^{d}_{ab} as parameters, their values are the Mulliken single- and double-bond orders for the pair of atoms and are of use only if the fractions f^{s}_{ab} and f^{d}_{ab} are different. A full list of H, C, N, O, and S parameters are provided as Supporting Information.

The multipole and charge mapping parameters were adjusted by comparing the mDC and B3LYP/6-311++G** electrostatic potentials evaluated on the solvent accessible surface for a large set of molecules. Some examples are shown in Figure 4. Fitting the parameters "by eye" to match the electrostatic potentials on the solvent accessible surface proved more reliable than having adjusted them to minimize the molecular dipole and quadrupole moments. The molecular multipole moments do improve with this procedure, however, as is shown in Table 1.

The LJ parameters were optimized to reduce the errors in the databases shown in Figure 5 and Table 2.

2.5. Drug Docking. Figure 6 displays the correlation between calculated protein−drug binding affinity scores and the experimentally determined[94−96] protein inhibition constant (IC50) for a set of drug ligands to the Cyclin-Dependent Kinase 2 (CDK2) protein receptor. The CDK2 structure[97] was taken from chain A of PDBID: 1H1S. Hydrogens were added and geometry optimized with Amber ff99. The autodock program[98] was used to dock 73 ligands to the protein structure using autodock's default options. The ligands were chosen as a subset of those listed in ref 41 subject to the availability of DFTB3 parameters. One of these ligands, ligand 29 in ref 41 is chosen to be a "lead", i.e., a ligand that has been identified as well-inhibiting the function of the protein. The goal then, is to identify additional candidate ligands that are structurally similar to the lead ligand but are predicted to inhibit the protein function with greater efficacy. If one assumes that ligands with a higher binding affinity to the receptor are generally more potent, then one would expect to observe a strong correlation between the ligand−receptor binding and the experimentally determined IC50. We compute the protein−ligand gas-phase interaction energy and compare the result to the experimental IC50s. These calculations ignore the desolvation of the drug upon binding to the receptor and the change in entropy associated with sterically inhibiting rotatable bonds. We also ignore the configurational averaging that would occur in dynamics and we freeze the CDK2 atoms to the 1H1S X-ray crystal structure during geometry optimization.

What remains to be described is the procedure that we performed to generate the bound structures used to compute the binding affinity scores. This procedure begins with the use of autodock, which binds a ligand to the receptor many times, clusters the results, and returns the coordinates of a docked ligand that is representative of each cluster. Each ligand is allowed to produce up to 20 clusters and thus we obtain up to 20 structures for each ligand and 632 structures in total for the set of 73 ligands. The ligand geometry for each of the 632 structures is gas-phase optimized with Amber ff99 for 500 steps while freezing the protein coordinates. The resulting Amber ff99 protein−ligand interaction energy is then stored. 100 steps of gas phase minimization are then performed for each of the

dx.doi.org/10.1021/ct401035t | J. Chem. Theory Comput. 2014, 10, 1086−1098

632 structures with mDC, and the final mDC gas-phase protein−ligand interaction energy is stored. The MNDO/MM results were computed using this same procedure. Although we have, for a given method, 632 geometry optimized structures, not all of them are geometrically relevant. For example, a structure is not relevant if it is not bound in the receptor pocket nor if the orientation of the ligand has been flipped so that the ligand substituents have occupied the receptor pocket. In other words, we only consider a structure to be relevant if it binds to the protein in the same way as the lead structure. To quantify this, we choose an atom common to all ligands (labeled N* in Figure 6) and measure the displacement of this atom relative to the minimum energy lead structure. We assign one structure to each of the 73 ligands for each method by choosing the structure that binds most similarly to the lead structure as measured by the N* atom unless the ligand has two or more structures that are within 3 Å of the lead structure, in which case we choose the lower energy structure.

Each of the 73 ligands is now represented by a single bound structure which is used to compute a "score"

$$\begin{aligned}\mathrm{Score}(n) &\equiv \Delta E_{\mathrm{ligand}\ n} \\ &= E_{\mathrm{receptor:ligand}\ n} - E_{\mathrm{receptor}} - E_{\mathrm{ligand}\ n}\end{aligned} \tag{19}$$

The score relative to the lead ligand, ligand 29, is

$$\Delta\mathrm{Score}(n) = \mathrm{Score}(29) - \mathrm{Score}(n) \tag{20}$$

which are shown in Figure 6.

The data labeled "Rao" in Figure 6 are the gas phase results reported in ref 41, who represented CDK2 with a 1000−1100 atom model system (in comparison to the approximately 4900 atom system used in this work) treated with an extended ONIOM method.[99,100] In their work, 300−380 atoms were computed with ωB97X-D/6-311++G** or XYG3/6-311++G** and the remaining atoms computed with PM6.

2.6. Simulation Protocol. Figure 7 displays the O−O RDF of water computed from molecular dynamics simulation using TIP3P, TIP4P-EW, and mDC. Each model was simulated with the same protocol. The system is composed of 512 waters in a cubic box. The simulation was performed using the canonical ensemble with a water density of 0.996 g/cm^3 and a temperature of 298 K for 3.2 ns with a time step of 1 fs. The first 200 ps of simulation was discarded as equilibration. An 8 Å nonbond cutoff was employed along with particle mesh Ewald (PME) to account for the long-range electrostatic interactions. The PME method used a 1 point/Å fast Fourier transform grid and interpolation was performed with sixth-order Cardinal B-splines. Shake was used to constrain the TIP3P and TIP4P models to the experimental gas-phase water geometry and the mDC model to the gas-phase DFTB3 geometry. Langevin dynamics was used to control the temperature with a collision frequency of 5 ps^{-1}. All simulations were performed using the SANDER program within Amber.[9] The implementation of mDC within SANDER required us to implement a new PME method generalized for use with multipolar charge densities. The generalization is based on the spherical tensor gradient operator, but we must defer a complete description of the new PME method to future work.

3. RESULTS AND DISCUSSION

Atomic multipoles are included into the mDC model to reproduce H-bond angles (Figure 1 and Wat-2009) and improve electrostatic potentials (Figure 4) and molecular multipole moments (Table 1). In ref 67, we noted that traditional DFTB methods produced good H-bond angles even though their electrostatics involved atomic charges only. We were thus lead to hypothesize that this was achieved through the multipolar character of the tight-binding coupling matrix elements. The DFTB3 hydrogen bond geometries shown in Figure 1 support this. When the hydrogen bond distance is small, the DFTB3 H-bond angle agrees with B3LYP/6-31++G**; however, at larger separations where the tight-binding matrix elements become small, it adopts a geometry similar to TIP3P.

Figure 4 illustrates a few examples where the use of multipoles improves the description of electrostatic potentials. The largest discrepancies between DFTB3 and B3LYP/6-311++G** occur in the description of π-bond electrons, sp^3 oxygen and sulfur lone pairs, and sp^2 nitrogen lone pairs. These differences are largely eliminated with mDC and result in an overall improved description of molecular multipole moments, as shown in Table 1.

The mDC LJ interactions were parametrized by perturbing the values used in GAFF to reduce the errors in the databases shown in Figure 5 and Table 2. mDC thus exhibits the smallest mue and largest correlation to the high-level reference data among the methods shown. One observes a significant amount of variation between the other models in Figure 5 for interactions stronger than 10 kcal/mol, which involve either multiple H-bonds or nucleobase stacking.

The S66 (Procedure) crms value in Table 2 for DFTB3 and DFTB2 are skewed because the lack of a dispersion model causes the benzene−ethene and benzene−benzene π···π complexes to dissociate until the intermolecular forces become sufficiently small to terminate the geometry optimizer. If we exclude those two molecules from the DFTB3 and DFTB2 statistics, their crms drop to 0.516 and 0.506 Å, respectively. The semiempirical models cause several of the stacked nucleobases in the JSCH-2005 database to devolve into hydrogen bonded complexes upon geometry optimization. The inclusion of a dispersion model could aid those cases as well.[101] Empirical dispersion corrections for DFTB2,[76,102] NDDO-based semiempirical methods,[103,104] and even ab initio methods[105,106] have been proposed; however, we limit the scope of our comparisons to the commonly used variant of these models to make the comparisons more useful to the wider set of users instead of the niche set of model developers. Furthermore, the series of databases that we consider are not biased toward dispersion-dominated complexes; we feel our comparisons are made "fair" by using the models consistently, as opposed to affording special treatment to the outliers in an attempt to make the models appear artificially tidy.

Among the semiempirical models, AM1 produces the worst nonbonded interaction energies, whereas DFTB3, DFTB2, and PM6 are of comparable quality. The reader may also note that the GAFF mue's are roughly twice as small as the semiempirical models. This is not unexpected; it has been previously noted that nonbond interactions are problematic for these methods.[107] The MNDO/MM relative energies are slightly better than the semiempirical methods, but their geometrical errors are similar.

The parametrized mDC method is applied to the prototype study of CDK2 drug screening, shown in Figure 6. CDK2 has been extensively studied due to the role it plays in tumor formation. The attention that it thus has received provides the computational community with an array of experimental results

which make this system attractive to those developing new linear-scaling quantum models. The experimental results are used to validate a computational procedure by presuming that a drug has been identified as a good candidate for protein inhibition and that one is tasked with identifying similar compounds that one predicts to inhibit protein function with greater efficacy.

The work of Rao[41] is the most recent study to use CDK2 to validate their computational protocols. With their method, they found an amazing R^2 correlation of 0.86 upon considering solvation and entropic effects. When only gas-phase interaction energies are used for the subset of drugs that we consider in this work, their R^2 drops to 0.83; however, this is still exceptionally good. For comparison, the fragment molecular orbital method (FMO) has recently been applied with MP2/6-31G* in a related study that, among other things, correlates the CDK2 experimental binding affinities to ab initio single-point calculations for a set 14 ligands with known X-ray crystal structures.[43] In that work,[43] the authors found a good R^2 correlation of 0.68 between their calculations and the experimental free energy of binding. A similar study of another system[42] employed the ONIOM3 method B3LYP/6-31G*:HF/3-21G:PM3 and produced a R^2 correlation of 0.64. AM1-DC has also been used[30] in the screening of zinc metalloenzymes to yield a correlation of 0.69.

We observe a R^2 correlation of 0.71 using mDC. MM and MNDO/MM calculations were performed to compare mDC with other methods that use the same CDK2 model and computational protocol. MM and MNDO/MM produce R^2 correlations of 0.56 and 0.63, respectively. There does not appear to be a clear relationship between the quality of nonbond interactions (Figure 5 and Table 2) and the model's correlation to the experimental CDK2 IC50s; the MNDO/MM energies correlate better than MM to the experimental IC50s even though MM more accurately reproduces the databases of high-level ab initio nonbond interactions. One may attribute this observation to the explicit treatment of polarization within mDC and MNDO/MM. Such attribution from this sole observation is speculative, in part, because semiempirical model polarizabilities are underestimated[108] by approximately 30%. Previous studies have also concluded, however, that the polarization energy plays a key role on the relative ordering of inhibitor interactions with focal adhesion kinase[109] and upon the binding of ligands to the trypsin protein.[110] Our results add support to their conclusions. QMFF's strive to achieve high accuracy in environments of varying degrees of heterogeneity, so we are encouraged to find that mDC yields more accurate small molecule cluster interactions *and* produces greater correlation with experiment when applied to large biomolecular systems.

A mDC calculation of CDK2 takes less than 10 s on a desktop computer. Rao's method requires about a day of computation;[41] however, we emphasize in this comparison that we use a system with more than four times as many atoms than used by Rao. A more detailed analysis of mDC timings is provided in ref 67. The computational efficiency of mDC offers the possibility of performing configurational sampling, which may improve the correlation between the computed CDK2 interactions with the experimental IC50s. In order to perform those simulations, one would need to parametrize the MM bonded energy terms that cross the fragment boundaries. We must defer the parametrization and analysis of the bonded interactions to future work.

Given that semiempirical models underestimate molecular polarizabilities, one may question if the quality of the mDC interactions degrade upon examining larger clusters or when used in condensed phase environments. To address this, Figure 7b displays the ΔE binding energy of water clusters up to ten waters. We limit our comparisons in Figure 7 to mDC and the MM water models because the errors observed in the Wat-2009 database are only amplified as more and larger clusters are examined. By this, we mean that each method in Figure 5 is observed to have a characteristic slope and that slope is largely unchanged when extended to Wat-2011. We therefore observe similar mDC and TIP3P mue's with Wat-2011 and Wat-2009. The TIP4P-EW clusters are bound too strongly because of its use of an enhanced dipole moment.

The TIP3P $\Delta\Delta E$'s in Figure 7c−e do not correlate well with ab initio because the higher-energy structures have a propensity to devolve into lower-energy configurations. This symptom may be directly related to the lack of structure observed in TIP3P RDFs. The TIP4P-EW model $\Delta\Delta E$'s correlate well for the larger clusters, but not for $(H_2O)_5$. One possible reason for this could be because the dipole moment of water increases as one goes from small to large clusters, thus making the enhanced dipole moment of TIP4P-EW inadequate for describing the higher-energy geometrical configurations of smaller-sized clusters. Devising an unambiguous computational experiment to test this hypothesis would require some clever partitioning of the electron density; however, we note that the average B3LYP/6-31G* Mulliken charge of oxygen computed for a series of water clusters increases with the size of the cluster and then begins to stabilize around 5 or 6 waters. This explanation therefore seems plausible. mDC correlates well with the reference data largely due to its tendency to maintain the desired H-bond geometry upon optimization.

Figure 7a compares the mDC, TIP3P, and TIP4P-EW O−O RDF to the recent experimental result taken from ref 111. Both TIP4P-EW and mDC display good agreement to experiment; whereas TIP3P suffers from a well-known lack of structure beyond the first solvation peak. The mDC first solvation peak is too high, but is similar to TIP4P-EW. Although the location of mDC's first solvation peak is in good agreement with experiment, it appears that its repulsive wall is slightly too steep at short separations. The use of an exponential-like repulsive wall, as was used in ref 67, may improve this. The mDC simulation results provided here are *very* preliminary. mDC extensively relies on the use of atomic multipoles and the implementation of mDC into the SANDER[9] program required us to derive and implement a new generalization of the PME method based on the spherical tensor gradient operator for multipolar charge densities. The details of the new PME method and further analysis of condensed phase properties is currently in preparation. The 3.2 ns mDC simulation of 512 waters required approximately 37 h per ns of dynamics using a 1 fs time step on a desktop workstation.

4. CONCLUSIONS

In this work, we parametrized a linear-scaling quantum force field for nonbonded interactions. Much effort was spent on the comparison of the method with standard MM force fields, semiempirical models, and relatively inexpensive ab initio Hamiltonians for a wide variety of nonbond interactions using a series of benchmark databases. As a part of this comparison, we provided a new benchmark database (SS7) of accurate nonbond interactions between sulfur-containing dimers. The

comparisons reveal that the GAFF MM force field significantly and consistently reproduced nonbond energies with far greater accuracy than the DFTB2-mio, DFTB3-3ob, AM1, and PM6 semiempirical models and the MNDO/MM method. The parametrized mDC model is shown to produce errors approximately half those of GAFF. The benefit of mDC relative to MM and MNDO/MM was further demonstrated by correlating CDK2 ligand binding energies to their experimental inhibition constants. It was shown that mDC correlated the data with a R^2 of 0.71; whereas MM and MNDO/MM produced correlations of 0.56 and 0.63, respectively.

In the process of parametrizing the mDC model, we made several observations using GAFF, DFTB3, and ab initio that may influence the future development tight-binding models. We observe that DFTB3's good hydrogen bond angles are only produced when there is significant overlap between fragments, but quickly decay into MM-like configurations when they are separated. We interpret this as the modeling of multipole interactions from within the tight-binding matrix instead of through second-order electrostatics. Furthermore, we observe systematic discrepancies between ab initio and DFTB3 electrostatic potentials in molecules containing π-bonds, sp^3 oxygen and sulfur lone pairs, and sp^2 nitrogen lone pairs. We are able to include the higher-order atomic multipoles within mDC without making intrusive modifications to DFTB3; whereas the extension of DFTB3 to atomic multipoles would be a significant undertaking.

■ ASSOCIATED CONTENT

Supporting Information

mDC parameters and the SS7 structures and energies are provided. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: york@biomaps.rutgers.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful for financial support provided by the National Institutes of Health (GM62248). Computational resources from the Minnesota Supercomputing Institute for Advanced Computational Research (MSI) were utilized in this work. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

■ REFERENCES

(1) Goedecker, S.; Scuseria, G. E. *IEEE Comput. Sci. Eng.* **2003**, *5*, 14−21.
(2) VandeVondele, J.; Borštnik, U.; Hütter, J. *J. Chem. Theory Comput.* **2012**, *8*, 3565−3573.
(3) Rudberg, E.; Rubensson, E. H.; Salek, P. *J. Chem. Theory Comput.* **2011**, *7*, 340−350.
(4) Khaliullin, R. Z.; VandeVondele, J.; Hütter, J. *J. Chem. Theory Comput.* **2013**, *9*, 4421−4427.
(5) He, X.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2010**, *6*, 405−411.
(6) Anisimov, V. M.; Bugaenko, V. L.; Bobrikov, V. V. *J. Chem. Theory Comput.* **2006**, *2*, 1685−1692.
(7) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 031103.
(8) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(9) Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Götz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, C., Sagui, R.; ; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*; University of California, San Francisco: San Francisco, CA, 2012.
(10) Khandogin, J.; York, D. M. *J. Phys. Chem. B* **2002**, *106*, 7693−7703.
(11) Mei, Y.; Li, Y. L.; Zeng, J.; Zhang, J. Z. H. *J. Comput. Chem.* **2012**, *33*, 1374−1382.
(12) Sena, A. M. P.; Miyazaki, T.; Bowler, D. R. *J. Chem. Theory Comput.* **2011**, *7*, 884−889.
(13) Kubar, T.; Elstner, M. *J. R. Soc. Interface* **2013**, *10*, 20130415.
(14) Lüdemann, G.; Woiczikowski, P. B.; Kubař, T.; Elstner, M.; Steinbrecher, T. B. *J. Phys. Chem. B* **2013**, *117*, 10769−10778.
(15) Nadig, G.; Van Zant, L. C.; Dixon, S. L.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **1998**, *120*, 5593−5594.
(16) van der Vaart, A.; Merz, K. M., Jr. *J. Chem. Phys.* **2002**, *116*, 7380−7388.
(17) Wang, B.; Brothers, E. N.; van der Vaart, A.; Merz, K. M., Jr *J. Chem. Phys.* **2004**, *120*, 11392−11400.
(18) Gao, J.; Garcia-Viloca, M.; Poulsen, T. D.; Mo, Y. *Adv. Phys. Org. Chem.* **2003**, *38*, 161−181.
(19) Isegawa, M.; Gao, J.; Truhlar, D. G. *J. Chem. Phys.* **2011**, *135*, 084107.
(20) Khaliullin, R. Z.; Bell, A. T.; Head-Gordon, M. *J. Chem. Phys.* **2008**, *128*, 184112.
(21) Pavanello, M.; van Voorhis, T.; Visscher, L.; Neugebauer, J. *J. Chem. Phys.* **2013**, *138*, 054101.
(22) Wu, F.; Liu, W.; Zhang, Y.; Li, Z. *J. Chem. Theory Comput.* **2011**, *7*, 3643−3660.
(23) Fukushima, K.; Wada, M.; Sakurai, M. *Proteins* **2008**, *71*, 1940−1954.
(24) Wollacott, A. M.; Merz, K. M., Jr. *J. Chem. Theory Comput.* **2007**, *3*, 1609−1619.
(25) He, X.; Wang, B.; Merz, K. M., Jr. *J. Phys. Chem. B* **2009**, *113*, 10380.
(26) Wang, B.; Raha, K.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2004**, *126*, 11430−11431.
(27) Raha, K.; van der Vaart, A. J.; Riley, K. E.; Peters, M. B.; Westerhoff, L. M.; Kim, H.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2005**, *127*, 6583−6594.
(28) Raha, K.; Peters, M. B.; Wang, B.; Yu, N.; Wollacott, A. M.; Westerhoff, L. M.; Merz, K. M., Jr. *Drug Discov. Today* **2007**, *12*, 725−731.
(29) Peters, M. B.; Raha, K.; Merz, K. M., Jr. *Curr. Opin. Drug Discov. Devel.* **2006**, *9*, 370.
(30) Raha, K.; Merz, K. M., Jr. *J. Am. Chem. Soc.* **2004**, *126*, 1020−1021.
(31) Raha, K.; Merz, K. M., Jr. *J. Med. Chem.* **2005**, *48*, 4558−4575.
(32) Richard, R. M.; Herbert, J. M. *J. Chem. Phys.* **2012**, *137*, 064113.
(33) Gordon, M. S.; Mullin, J. M.; Pruitt, S. R.; Roskop, L. B.; Slipchenko, L. V.; Boatz, J. A. *J. Phys. Chem. B* **2009**, *113*, 9646−9663.
(34) Li, W. *J. Chem. Phys.* **2013**, *138*, 014106.
(35) Bondesson, L.; Rudberg, E.; Luo, Y.; Salek, P. *J. Phys. Chem. B* **2007**, *111*, 10320−10328.
(36) Fox, S. J.; Pittock, C.; Fox, T.; Tautermann, C. S.; Malcolm, N.; Skylaris, C.-K. *J. Chem. Phys.* **2011**, *135*, 224107.
(37) Fracchia, F.; Filippi, C.; Amovilli, C. *J. Chem. Theory Comput.* **2013**, *9*, 3453−3462.
(38) Fracchia, F.; Filippi, C.; Amovilli, C. *J. Chem. Theory Comput.* **2012**, *8*, 1943−1951.
(39) Cembran, A.; Payaka, A.; Lin, Y.-L.; Xie, W.; Mo, Y.; Song, L.; Gao, J. *J. Chem. Theory Comput.* **2010**, *6*, 2242−2251.

(40) Goyal, P.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2011**, *115*, 6790−6805.
(41) Rao, L.; Zhang, I. Y.; Guo, W.; Feng, L.; Meggers, E.; Xu, X. *J. Comput. Chem.* **2013**, *34*, 1636−1646.
(42) Alzate-Morales, J. H.; Caballero, J.; Jague, A. V.; Nilo, F. D. G. *J. Chem. Inf. Model.* **2009**, *49*, 886−899.
(43) Mazanetz, M. P.; Ichihara, O.; Law, R. J.; Whittaker, M. *J. Cheminform.* **2011**, *3*, 2−16.
(44) Gao, J. *J. Phys. Chem.* **1997**, *101*, 657−663.
(45) Xie, W.; Gao, J. *J. Chem. Theory. Comput.* **2007**, *3*, 1890−1900.
(46) Xie, W.; Song, L.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2008**, *128*, 234108.
(47) Song, L.; Han, J.; Lin, Y.-L.; Xie, W.; Gao, J. *J. Phys. Chem. A* **2009**, *113*, 11656−11664.
(48) Han, J.; Truhlar, D. G.; Gao, J. *Theor. Chem. Acc.* **2012**, *131*, 1161−1161.
(49) Han, J.; Mazack, M. J. M.; Zhang, P.; Truhlar, D. G.; Gao, J. *J. Chem. Phys.* **2013**, *139*, 054503.
(50) Elking, D. M.; Cisneros, G. A.; Piquemal, J.; Darden, T. A.; Pedersen, L. G. *J. Chem. Theory Comput.* **2010**, *6*, 190−202.
(51) Chaudret, R.; Ulmer, S.; van Severen, M.-C.; Gresh, N.; Parisel, O.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. In *Theory and Applications of Computational Chemistry*; Wei, D.-Q., Wang, X.-J., Eds.; American Institute of Physics, 2008; pp 185−192.
(52) Piquemal, J.; Chevreau, H.; Gresh, N. *J. Chem. Theory Comput.* **2007**, *3*, 824−837.
(53) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. *J. Chem. Theory Comput.* **2007**, *3*, 1960−1986.
(54) Cisneros, G. A.; Elking, D.; Piquemal, J.-P.; Darden, T. A. *J. Phys. Chem. A* **2007**, *111*, 12049−12056.
(55) Cisneros, G. A.; Piquemal, J.; Darden, T. A. *J. Chem. Phys.* **2006**, *125*, 184101.
(56) Piquemal, J.; Cisneros, G.; Reinhardt, P.; Gresh, N.; Darden, T. A. *J. Chem. Phys.* **2006**, *124*, 104101.
(57) Cisneros, G.; Piquemal, J.; Darden, T. A. *J. Chem. Phys.* **2005**, *123*, 044109.
(58) Piquemal, J.-P.; Gresh, N.; Giessner-Prettre, C. *J. Phys. Chem. A* **2003**, *107*, 10353−10359.
(59) Donchev, A. G.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, O. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 7829−7834.
(60) Donchev, A. G.; Galkin, N. G.; Illarionov, A. A.; Khoruzhii, O. V.; Olevanov, M. A.; Ozrin, V. D.; Subbotin, M. V.; Tarasov, V. I. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 8613−8617.
(61) Donchev, A. G.; Galkin, N. G.; Pereyaslavets, L. B.; Tarasov, V. I. *J. Chem. Phys.* **2006**, *125*, 244107.
(62) Wang, Y.; Sosa, C. P.; Cembran, A.; Truhlar, D. G.; Gao, J. *J. Phys. Chem. B* **2012**, *116*, 6781−6788.
(63) Dahlke, E. E.; Truhlar, D. G. *J. Chem. Theory Comput.* **2007**, *3*, 1342−1348.
(64) Xie, W.; Orozco, M.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2009**, *5*, 459−467.
(65) Zhang, P.; Truhlar, D. G.; Gao, J. *Phys. Chem. Chem. Phys.* **2012**, *14*, 7821−7829.
(66) Gao, J.; Wang, Y. *J. Chem. Phys.* **2012**, *136*, 071101.
(67) Giese, T. J.; Chen, H.; Dissanayake, T.; Giambaşu, G. M.; Heldenbrand, H.; Huang, M.; Kuechler, E. R.; Lee, T.-S.; Panteva, M. T.; Radak, B. K.; York, D. M. *J. Chem. Theory Comput.* **2013**, *9*, 1417−1427.
(68) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2007**, *127*, 194101.
(69) Gao, J. *J. Chem. Phys.* **1998**, *109*, 2346−2354.
(70) Otto, P.; Ladik, J. *Chem. Phys.* **1975**, *8*, 192−200.
(71) Gaus, M.; Goez, A.; Elstner, M. *J. Chem. Theory Comput.* **2013**, *9*, 338−354.
(72) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260−7268.
(73) Yang, Y.; Yu, H.; York, D. M.; Cui, Q.; Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 10861−10873.
(74) Hu, H.; Lu, Z.; Elstner, M.; Hermans, J.; Yang, W. *J. Phys. Chem. A* **2007**, *111*, 5685−5691.
(75) Gaus, M.; Cui, Q.; Elstner, M. *J. Chem. Theory Comput.* **2011**, *7*, 931−948.
(76) Kumar, A.; Elstner, M.; Suhai, S. *Int. J. Quantum Chem.* **2003**, *95*, 44−59.
(77) Stewart, J. J. P. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(78) Dewar, M. J. S.; Zoebisch, E.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(79) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, M.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Straverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; 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.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision A.02; Gaussian, Inc.: Wallingford, CT, 2009.
(80) Cheatham, T. E., III; Cieplak, P.; Kollman, P. A. *J. Biomol. Struct. Dyn.* **1999**, *16*, 845−862.
(81) Homeyer, N.; Horn, A. H. C.; Lanig, H.; Sticht, H. *J. Mol. Model.* **2006**, *12*, 281−289.
(82) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665−9678.
(83) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985−1993.
(84) Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2003**, *125*, 15608−15613.
(85) Šponer, J.; Jurečka, P.; Hobza, P. *J. Am. Chem. Soc.* **2004**, *126*, 10142−10151.
(86) Berka, K.; Laskowski, R.; Riley, K. E.; Hobza, P.; Vondrášek, J. *J. Chem. Theory Comput.* **2009**, *5*, 982−992.
(87) Řezàč, J.; Riley, K. E.; Hobza, P. *J. Chem. Theory Comput.* **2011**, *7*, 2427−2438.
(88) Halkier, A.; Helgaker, T.; Jørgensen, P.; Klopper, W.; Koch, H.; Olsen, J.; Wilson, A. K. *Chem. Phys. Lett.* **1998**, *286*, 243−252.
(89) Xantheas, S. S. Interaction Potentials for Water from Accurate Cluster Calculations. *Intermolecular Forces and Clusters II*; Structure and Bonding; Springer Berlin Heidelberg: Berlin, 2005; Vol. *116*; pp 119−148.
(90) Xantheas, S. S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493−1499.
(91) Xantheas, S. S.; Aprà, E. *J. Chem. Phys.* **2004**, *120*, 823−828.
(92) Bryantsev, V. S.; Diallo, M. S.; van Duin, A.; Goddard, W. A., III *J. Chem. Theory Comput.* **2009**, *5*, 1016−1026.
(93) Temelso, B.; Archer, K. A.; Shields, G. C. *J. Phys. Chem. A* **2011**, *115*, 12034−12046.
(94) Gibson, A. E.; Arris, C. E.; Bentley, J.; Boyle, F. T.; Curtin, N. J.; Davies, T. G.; Endicott, J. A.; Golding, B. T.; Grant, S.; Griffin, R. J.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Tucker, J. A.; Whitfield, H. J. *J. Med. Chem.* **2002**, *45*, 3381−3393.
(95) Hardcastle, I. R.; Arris, C. E.; Bentley, J.; Boyle, F. T.; Chen, Y.; Curtin, N. J.; Endicott, J. A.; Gibson, A. E.; Golding, B. T.; Griffin, R. J.; Jewsbury, P.; Menyerol, J.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Pratt, D. J.; Wang, L.-Z.; Whitfield, H. J. *J. Med. Chem.* **2004**, *47*, 3710−3722.
(96) Griffin, R. J.; Henderson, A.; Curtin, N. J.; Echalier, A.; Endicott, J. A.; Hardcastle, I. R.; Newell, D. R.; Noble, M. E. M.; Wang, L.-Z.; Golding, B. T. *J. Am. Chem. Soc.* **2006**, *128*, 6012−6013.

(97) Davies, T. G.; Bentley, J.; Arris, C. E.; Boyle, F. T.; Curtin, N. J.; Endicott, J. A.; Gibson, A. E.; Golding, B. T.; Griffin, R. J.; Hardcastle, I. R.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E. M.; Tucker, J. A.; Wang, L.; Whitfield, H. J. *Nat. Struct. Biol.* **2002**, *9*, 745−749.
(98) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. *J. Comput. Chem.* **2009**, *16*, 2785−2791.
(99) Guo, W.; Wu, A.; Xu, X. *Chem. Phys. Lett.* **2010**, *498*, 203−208.
(100) Guo, W.; Wu, A.; Zhang, I. Y.; Xu, X. *J. Comput. Chem.* **2012**, *33*, 2142−2160.
(101) Dabkowska, I.; Jurečka, P.; Hobza, P. *J. Chem. Phys.* **2005**, *122*, 204322.
(102) Zhechkov, L.; Heine, T.; Patchkovskii, S.; Seifert, G.; Duarte, H. A. *J. Chem. Theory Comput.* **2005**, *1*, 841−847.
(103) Zhang, P.; Fiedler, L.; Leverentz, H. R.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2011**, *7*, 857−867.
(104) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362−2370.
(105) Kruse, H.; Goerigk, L.; Grimme, S. *J. Org. Chem.* **2012**, *77*, 10824−10834.
(106) Antony, J.; Grimme, S. *J. Comput. Chem.* **2012**, *33*, 1730−1739.
(107) Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808−3816.
(108) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2005**, *123*, 164108.
(109) de Courcy, B.; Piquemal, J.; Garbay, C.; Gresh, N. *J. Am. Chem. Soc.* **2010**, *132*, 3312−3320.
(110) Jiao, D.; Golubkov, P. A.; Darden, T. A.; Ren, P. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 6290−6295.
(111) Skinner, L. B.; Huang, C.; Schlesinger, D.; Pettersson, L. G. M.; Nilsson, A.; Benmore, C. J. *J. Chem. Phys.* **2013**, *138*, 074506.