JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

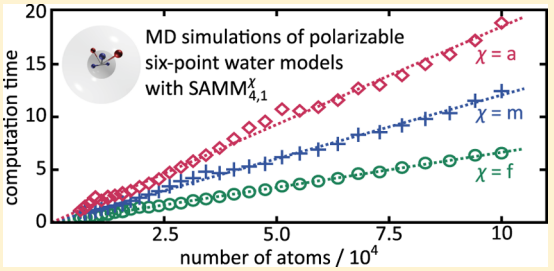
Including the Dispersion Attraction into Structure-Adapted Fast Multipole Expansions for MD Simulations

Konstantin Lorenzen, Christoph Wichmann, and Paul Tavan*

Lehrstuhl für Biomolekulare Optik, Ludwig-Maximilians-Universität, Oettingenstr. 67, 80538 München, Germany

Supporting Information

ABSTRACT: Molecular dynamics (MD) simulations of protein−solvent systems, which are modeled by polarizable or nonpolarizable all-atom force fields and are enclosed by periodic boundaries, require accurate and efficient algorithms for the computation of the long-range interactions. A possible choice is the fast structure-adapted multipole method called SAMM$_p$/RF (Lorenzen et al. *J. Chem. Theory Comput.* **2012**, *8*, 3628−3636). It is based on pth order Cartesian Taylor expansions of the electrostatic interactions, on an adaptive and hierarchical decomposition of a macromolecular simulation system into a quaternary tree of nested atom clusters, and on a reaction field (RF) correction originating from a distant dielectric continuum. Here, we substantially extend this method by adding qth order Taylor expansions of the dispersion attraction and by formulating an interaction acceptance criterion for cluster−cluster interactions, which is based on substance-specific accuracy estimates. As a result, we obtain with the default expansion orders $(p,q) = (4,1)$ a family of MD algorithms SAMM$^{\chi}_{4,1}$, which comprises carefully balanced compromises χ between accuracy and efficiency ranging from "accurate" (χ = a) to "fast" (χ = f). Issues of accuracy and efficiency are discussed by sample simulations of liquid water and methanol using simple nonpolarizable and complex polarizable model potentials. Here, it is shown that the computational effort scales linearly with the number N of atoms. For a complex polarizable water model, these simulations also show that SAMM$^{\chi}_{4,1}$ is by factors between 2 (χ = a) and 5 (χ = f) faster than its predecessor SAMM$_4$. Other benefits, which arise in simulations employing polarizable force fields with a high degree of local complexity, are also discussed.



1. INTRODUCTION

The calculation of the long-range forces is the computational bottleneck in molecular dynamics (MD) simulations of biomolecular systems[1−4] described by molecular mechanics (MM) force fields such as CHARMM,[5] AMBER95,[6] or GROMOS.[7] Accurate and efficient algorithms for force evaluation are even more urgently needed, if the effects of electrostatic polarizability are explicitly included in the description, because then several self-consistency iterations have to be carried out for the computation of the electrostatics at each integration step of the equations of motion. Corresponding polarizable molecular mechanics (PMM) force fields have been suggested not only for the key solvent water (see e.g. ref 8 and references quoted therein) but also for polypeptides[9−11] and nucleic acids.[12]

There are two conceptually different approaches to the computation of long-range forces, i.e., the lattice-summations (LS) of the Ewald type[13−15] and the related multilevel summation (MLS),[16] on the one hand, and fast multipole methods[17−33] (FMM), on the other. Most of these approaches were originally restricted to electrostatic interactions and applied a short-range cutoff at distances $r_c \approx 1.0-1.5$ nm to the dispersion attraction[5−7,34,35] (early FMM-exceptions are refs 19 and 21). Because this cutoff entails algorithmic artifacts such as cooling,[36] several LS[37−39] and FMM approaches[40,41] were more recently extended toward the dispersion interaction.

LS and MLS methods naturally take advantage of periodic boundary conditions (PBC), which avoid surface artifacts and, thus, enable the control of the density or of the pressure within the simulated system of typical size L. Less straightforward is the combination of FMM methods with PBC. Early FMM implementations[19,21,27] were restricted to molecular clusters surrounded by a vacuum. More recent implementations employ a moving boundary reaction field (RF) approach[28−30] or the isotropic periodic sum[22,23,42] to account, in a mean-field fashion, for interactions at distances larger than the cutoff radius $d_{MIC} = L/2$, which is dictated by the minimum image convention[43] (MIC). These methods work with nonperiodic electrostatic potentials and, thus, actually implement toroidal boundary conditions,[43] which are well-suited for nonperiodic liquid-phase systems. On the other hand, combinations[20,24] of FMM with LS concepts have also been developed and can be employed to describe the periodic potentials of crystalline structures.

Our choice of a FMM/RF approach for toroidally closed systems is the structure adapted multipole method (SAMM)[25−28] and its recent extension[29,30,44] toward the balanced inclusion of multipole and Taylor expansions up to pth order (SAMM$_p$/RF), where the default is $p = 4$. Note that

Received: April 14, 2014
Published: June 19, 2014

   dx.doi.org/10.1021/ct500319a | *J. Chem. Theory Comput.* 2014, 10, 3244−3259

SAMM$_p$/RF is not restricted to partial point charges as sources of the electrostatic potential but can also efficiently treat inducible (Gaussian) dipoles.[30,44] SAMM differs from other FMM approaches by the hierarchical decomposition of a simulation system into an adaptive and quaternary tree of nested atomic clusters, which replaces the commonly employed[18−24] geometric and octal tree.

It is one of the aims of this paper to explain the favorable properties of adaptive quaternary trees and the algorithms employed for their reliable and computationally efficient construction. Concurrently, such trees enable an optimized exploitation of computational resources on parallel computers. Because these issues were largely omitted in the previous descriptions of SAMM[25−30] and because the underlying algorithms were repeatedly optimized during the past decade, a thorough presentation seems necessary.

A more important aim, however, is the demonstration that the advantages offered by quaternary trees can be fully exploited only if the dispersion interaction is also included in the FMM scheme. A first benefit of such an inclusion is, of course, that the short-range cutoff (at ∼1 nm) of the dispersion attraction and the associated cooling[36] and other artifacts[37,38] can be avoided. As mentioned above, such a cutoff has been common practice in biomolecular MD simulations. Extensions of LS[37−39] and FMM approaches[21,40,41] toward the inclusion of the long-range parts of the dispersion attraction represent a more recent development. Also, the first implementation of SAMM$_p$/RF provided by the parallelized MD program IPHIGENIE[29,30,44] applied a short-range cutoff to the dispersion.

Correspondingly, we here present the extension of SAMM$_p$/RF toward SAMM$_{p,q}$/RF, where q defines the highest order of the additional FMM expansion employed for the dispersion attraction (a cutoff is still applied to the shorter-range Pauli-repulsion). For the implementation of this extension, the computational strategy of SAMM has been thoroughly revised. As a result, the MD program IPHIGENIE now enables one to choose among several different and carefully tuned compromises between accuracy and efficiency. Note that one of these compromises has already been applied to extended MD simulations, which served to characterize a recent polarizable six-point model potential for water.[8,45]

The explanation of the revised computational strategy starts with the formal presentation of the qth order Cartesian FMM expansions used for the dispersion. Subsequently, we introduce the SAMM cluster hierarchy employed for the decomposition of a toroidally closed simulation system into a nested hierarchy of atomic clusters. In particular, we review the predefined molecular structures forming the lowest level of the hierarchy and sketch the algorithms employed for combining these lowest level atomic clusters into higher order clusters such that also the parallelization strategy is clarified. Next we explain the top-down procedure of interaction list generation, which rests on a novel acceptance criterion for the computation of interactions at a given cluster level or, alternatively, for the decomposition of clusters at this level into their constituent subclusters. Using models for water and methanol as examples, we develop a strategy to optimize the associated compromise between accuracy and efficiency in such a way that the chosen level of accuracy applies to different chemical compositions. For systems subject to toroidal boundary conditions, we then introduce the highest cluster level, whose interactions are still compatible with the MIC cutoff $d_{\mathrm{MIC}} = L/2$, beyond which the electrostatic and dispersion interactions are approximated by mean-field expressions. MD simulations of liquid water and methanol illustrate the resulting compromises between accuracy and efficiency and demonstrate the overall linear scaling.

2. THEORY

The electrostatics FMM approach SAMM$_p$/RF suggested in ref 29 is readily extended toward the dispersion attraction giving rise to a method called SAMM$_{p,q}$/RF, where q is the order of the resulting FMM expansion of the dispersion energy. Figure 1 introduces the associated concept.

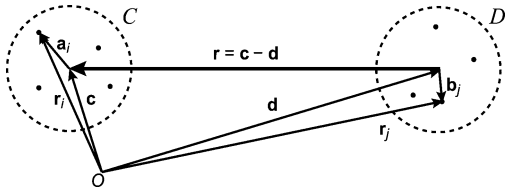


Figure 1. FMM geometry for two interacting clusters C and D (dashed spheres) of atoms $i \in C$ at $\mathbf{r}_i$ and $j \in D$ at $\mathbf{r}_j$ (dots) carrying dispersion charges B_i and B_j. The dispersion interaction of i and j depends on the connecting vector $\mathbf{r}_i - \mathbf{r}_j$ and is evaluated by a Taylor expansion around the vector $\mathbf{r}$ linking the two cluster centers. The positions of these centers are denoted by $\mathbf{c}$ and $\mathbf{d}$, respectively, those of the atoms within the respective clusters by $\mathbf{a}_i$ and $\mathbf{b}_j$.

2.1. Balanced FMM for the Dispersion.

The total dispersion energy

$$E_{\mathrm{d}}(C, D) = \sum_{i \in C} B_i \phi^D(\mathbf{r}_i) \tag{1}$$

of the two clusters C and D depicted in Figure 1 is given by the dispersion charges B_i of all atoms $i \in C$ and by the dispersion potential

$$\phi^D(\mathbf{r}_i) \equiv -\sum_{j \in D} \frac{B_j}{|\mathbf{r}_i - \mathbf{r}_j|^6} \tag{2}$$

which is generated at the positions $\mathbf{r}_i$ of the atoms i by the dispersion charges B_j of all atoms $j \in D$. Here, we have assumed that the parameters $B_{i,j}$, specifying in MM force fields the dispersion attraction between atoms i and j, obey the product decomposition

$$B_{i,j} = B_i B_j \tag{3}$$

According to this rule,[7,34,46,47] the pair parameters $B_{i,j}$ are calculated as geometric means of the van der Waals parameters σ and ε, which define the dispersion attraction between atoms i and j of the same type through $4\varepsilon\sigma^6/(\mathbf{r}_i - \mathbf{r}_j)^6$. Thus, the dispersion charge for an atom i of this type is $B_i = 2\,\sigma^3\sqrt{\varepsilon}$. There are, however, force fields like CHARMM22[5] or AMBER95,[6] which combine the van der Waals diameters σ by the arithmetic mean. The differences of the parameters $B_{i,j}$ obtained by the two rules are usually very small,[48] such that the geometric combination rule should be applicable also in combination with the latter force fields.

With the geometry explained by Figure 1 the potential (eq 2) may be equivalently written as

$$\phi^D(\mathbf{r}_i) \equiv -\sum_{j \in D} \frac{B_j}{|\mathbf{r} + (\mathbf{a}_i - \mathbf{b}_j)|^6} \tag{4}$$

If we denote the geometrical centers of the two clusters C and D, which comprise $|C|$ and $|D|$ atoms, respectively, by

$$\mathbf{c} \equiv \frac{1}{|C|}\sum_{i\in C}\mathbf{r}_i \tag{5}$$

and $\mathbf{d}$, the qth order Taylor expansion of the potential $\phi^D(\mathbf{r}_i)$ around the connecting vector $\mathbf{r} = \mathbf{c} - \mathbf{d}$ is

$$\phi^{D,q}(\mathbf{r}_i) = -\sum_{n=0}^{q}\frac{1}{n!}\left(\partial_{(n)}\frac{1}{r^6}\right)\odot\sum_{j\in D}B_j(\mathbf{a}_i - \mathbf{b}_j)^{(n)} \tag{6}$$

where $r \equiv |\mathbf{r}|$ is the center−center distance of the two clusters and $\mathbf{a}_i - \mathbf{b}_j$ the difference of the local coordinates $\mathbf{a}_i$ and $\mathbf{b}_j$. In eq 6, we have used, just like in the preceding paper,[29] the tensorial notation of Warren and Salmon.[31] Rearranging terms,[29] one finds the equivalent qth order Taylor expansion

$$\phi^{D,q}(\mathbf{r}_i) = -\sum_{n=0}^{q}\frac{1}{n!}\mathbf{a}_i^{(n)}\odot\mathbf{T}^{D,n,q}(\mathbf{c}) \tag{7}$$

around the center $\mathbf{c}$ of cluster C. The expansion coefficients

$$\mathbf{T}^{D,n,q}(\mathbf{c}) \equiv \partial_{(n)}\sum_{m=0}^{q-n}\phi^{m,D}(\mathbf{c}),\ n = 0, ..., q \tag{8}$$

derive from the potentials

$$\phi^{m,D}(\mathbf{c}) = \frac{1}{m!}\frac{1}{r^{2m+6}}\mathbf{r}^{(m)}\odot\mathbf{M}^{m,\mathbf{d}} \tag{9}$$

generated by the mth order multipole moments $\mathbf{M}^{m,\mathbf{d}}$, which characterize the distribution of dispersion charges B_j in cluster D with respect to the reference point $\mathbf{d}$. With the local distances $b_j = |\mathbf{b}_j|$, they are given by

$$\mathbf{M}^{m,\mathbf{d}} = \sum_j B_j b_j^{m+6}\left(\partial_{(m)}\frac{1}{b_j^6}\right)(-1)^m \tag{10}$$

Explicit expressions for these multipole moments are given in sections S1 and S2 of the Supporting Information (SI) for m = 0, 1, 2, and 3. Similarly, section S3 in the SI lists the corresponding explicit expressions for the potentials $\phi^{m,D}(\mathbf{c})$, which originate from these multipole moments and are evaluated at the center $\mathbf{c}$ of cluster C.

The potential generated at an atomic position $\mathbf{r}_i$ within cluster C by a set of other clusters D then follows from a Taylor expansion analogous to eq 7, in which the nth order expansion coefficients are simply the sums of the coefficients $\mathbf{T}^{D,n,q}(\mathbf{c})$ belonging to the clusters D and defined by eq 8.

2.2. FMM Forces. Besides the dispersion energy (eq 1), MD simulations also require the associated atomic forces, which are the negative gradients $-\nabla_i E_d$. Differentiating eq 1 after inserting the qth order Taylor expansion (eq 7), one finds

$$\mathbf{f}_d^q(\mathbf{r}_i) = B_i\sum_{n=0}^{q-1}\frac{1}{n!}\mathbf{a}_i^{(n)}\odot\mathbf{T}^{D,n+1,q}(\mathbf{c}) \tag{11}$$

which is a Taylor expansion of the dispersion forces $\mathbf{f}_d(\mathbf{r}_i)$ up to order $q - 1$, whose coefficients $\mathbf{T}^{D,k,q}$, k = 1, ..., q, contain by eq 8 the kth order derivatives of the multipole potentials (eq 9). The forces calculated by the approximate expression (eq 11) obey Newton's third law, as one can see by repeating the arguments given in section 2.2 of ref 29 for the given case.

Due to the truncation of the Taylor expansion (eq 11), the resulting SAMM$_q$ dispersion forces will deviate from the exact values. For two clusters C and D, which are separated by the distance r, one can estimate this deviation by taking the first neglected term in the expansion (eq 11) as a measure and by averaging over all mutual orientations of C and D. A similar estimate can be gained from eq 15 in ref 29 for the error of the SAMM$_p$ electrostatic forces. When the variables

$$(\alpha, b) \in \{(p, \mathrm{e}), (q, \mathrm{d})\} \tag{12}$$

which discriminate the SAMM$_\alpha$ descriptions of electrostatics (e) and dispersion (d) and the function

$$\gamma(b) = \begin{cases} 2 & \text{for } b = \mathrm{e} \\ 7 & \text{for } b = \mathrm{d} \end{cases} \tag{13}$$

are introduced, the resulting estimates of the SAMM$_\alpha$ force errors are

$$\Delta\tilde{f}_{C,D,b}^{(\alpha)}(r) = A_{C,D,b}^{(\alpha)}\frac{(2\langle R\rangle_{C,D})^\alpha}{r^{\alpha+\gamma(b)}} \tag{14}$$

Here, $A_{C,D,b}^{(\alpha)}$ are constants, which can be estimated by the procedures explained further below in connection with eq 37, and $\langle R\rangle_{C,D} \equiv (R_C + R_D)/2$ is the average radius of gyration of C and D. For the cluster C of atoms i at the local positions $\mathbf{a}_i$, this radius is

$$R_C = \left[\frac{1}{|C|}\sum_{i\in C}\mathbf{a}_i^2\right]^{1/2} \tag{15}$$

A comparison of the error estimate (eq 14) for the average dispersive and electrostatic forces acting on the atoms of two clusters C and D separated by the distance r shows that the error of the dispersive forces decays much more quickly with r than that of the electrostatic forces mainly because $\gamma(\mathrm{d}) \gg \gamma(\mathrm{e})$. For a given r, one thus expects that errors of a comparably small size can be achieved by choosing an order q of the dispersion expansion that is much smaller than the order p of the electrostatics expansion. In fact, it will turn out that an expansion of the dispersion energy up to dipolar order q = 1 usually suffices in combination with an electrostatics expansion up to hexadecapolar order p = 4. With the notation introduced at the beginning of this section, the resulting FMM/RF algorithm will be called SAMM$_{4,1}$/RF.

3. SAMM CLUSTER HIERARCHY

FMM methods like SAMM$_{p,q}$ decompose a molecular simulation system, which is made up by the set S of all N atoms i, i = 1, ..., N, into a nested hierarchy of spatially compact subsets $C_{j,l} \subset S$, which are called clusters. Here, the index j, j = 1, ..., N_l, counts the clusters $C_{j,l}$ within a given hierarchy level l. The level index l may assume the values l = 0, 1, ..., λ, with λ marking the topmost hierarchy level, which is the highest level containing more than one cluster (usually $N_\lambda \approx 100$). Formally, one may add a further level $\lambda + 1$, which combines all atoms into the single cluster $C_{1,\lambda+1} = S$ comprising the whole simulation system.

At each hierarchy level $l \le \lambda + 1$, the clusters $C_{j,l}$ form a disjoint decomposition

$$\cup_{j=1}^{N_l}C_{j,l} = S \text{ and } C_{j,l}\cap C_{k,l} = \emptyset, \text{ if } j \ne k \tag{16}$$

of the atom set S. For an upper level $l > 0$, each cluster $C_{j,l}$ comprises according to

dx.doi.org/10.1021/ct500319a | J. Chem. Theory Comput. 2014, 10, 3244−3259

$$C_{j,l} = \cup_{\{k|C_{k,l-1}\cap C_{j,l}\neq\emptyset\}}\ C_{k,l-1} \tag{17}$$

a set of $M_{j,l-1}$ clusters $C_{k,l-1}$ at the next lower level, which are called children of the parent $C_{j,l}$. Adding the numbers $M_{j,l-1}$ of children of all parents $C_{j,l}$ at the level $l > 0$ yields the number

$$\sum_{j=1}^{N_l} M_{j,l-1} = N_{l-1} \tag{18}$$

of all clusters $C_{k,l-1}$ at the children level $l - 1$. Thus, eqs 17 and 18 define a unique parent−children relation for each pair $(l,l-1)$ of hierarchy levels with $0 < l \leq \lambda + 1$.

At all upper hierarchy levels $0 < l \leq \lambda$, common FMM algorithms[18−20,22,24] split each parent $C_{j,l}$ into $M_{j,l-1} = 8$ children $C_{k,l-1}$. This parent−children relation derives from the refinement of a cubic grid with the lattice constant a by means of a subgrid with the lattice constant $a/2$ implying that each original cube of volume a^3 splits into eight subcubes of volume $(a/2)^3$. All atoms found in cube j at level l are elements of the cluster $C_{j,l}$. Thus, the resulting hierarchical decomposition of S represents an octal tree.

Figure 2 sketches the alternative splitting scheme employed by SAMM$_{p,q}$ at the intermediate hierarchy levels $0 < l \leq \lambda$.

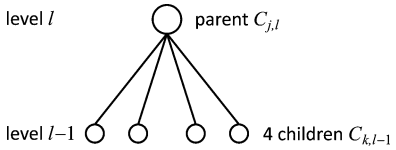


Figure 2. Local splitting motif typical for SAMM$_{p,q}$. A parent cluster $C_{j,l}$ is split into four children $C_{k,l-1}$ at an intermediate level $0 < l \leq \lambda$ within a quaternary tree, which represents the SAMM$_{p,q}$ decomposition of a simulation system into a nested hierarchy of compact clusters.

Here, each parent cluster $C_{j,l}$ is split only into $M_{j,l-1} = 4$ children $C_{k,l-1}$, thus mapping the cluster hierarchy on a quaternary tree. In our implementation, the quaternary tree structure is exact for all levels with $1 < l \leq \lambda$ and an average property at the level $l = 1$ (due to the applied top-down decomposition described further below).

In a quaternary tree, the radius of gyration of compact clusters increases with the level height l according to $R_l \approx 4^{l/3}R_0 \approx 1.5874^l R_0$, whereas in an octal tree it grows much more rapidly, because here one has $R_l = 2^l R_0$. As is demonstrated by the FMM error estimate (eq 14) for the dispersion and electrostatic forces, the slower increase of R_l is advantageous, because these errors scale at a given cluster−cluster distance r with $(2R_l)^q$ and $(2R_l)^p$, respectively. Hence, for achieving a given accuracy, one may calculate cluster−cluster interactions at a given hierarchy level l with SAMM$_{p,q}$ already at much smaller distances than with common FMM.

However, if one wants to exploit this difference, then one must find tools which can decompose a simulation system into a quaternary hierarchy of compact clusters. As will be described further below, such quaternary trees can be reliably and efficiently constructed at all levels $0 < l \leq \lambda$ with the help of neural clustering algorithms.[49,50] Then, only the clusters at the bottom level $l = 0$ and the decomposition of the system level $l = \lambda + 1$ require special considerations.

3.1. Bottom-Level: Structural Units. In SAMM, the N_0 clusters $C_{j,0}$ at the lowest hierarchy level $l = 0$ consist of chemically stable and predefined groups of $|C_{j,0}| = 3, 4, ..., 16$ atoms, which include at most seven and on average about three to four non-hydrogen atoms. These clusters $C_{j,0}$ are called structural units (SUs). For molecular solvents like water or methanol, for instance, the SUs are the solvent molecules. The positions and sizes of the SUs are given by their centers of geometry $\mathbf{r}_{j,0}$ (cf. eq 5) and radii $R_{j,0}$ of gyration (cf. eq 15), respectively.

Using a CHARMM-type nomenclature,[5] Tables S3 and S4 in section S4.1 of the SI specify the chemical compositions and radii $R_X \equiv R_{j,0}$ of gyration of typical SUs X, into which one must decompose protein/solvent simulation systems for applications of SAMM$_{p,q}$/RF.

3.2. Choice of the Top Level. There are two conflicting aims guiding the choice of the height λ of the hierarchy. The first aim is a balanced distribution of the computational load, when executing a MD simulation on a parallel computer. Because SAMM$_{p,q}$/RF has been implemented in the program package IPHIGENIE[29,30,44] with a MPI parallelization, the computation of the long-range interactions can take advantage of N_c CPUs.

If the number N_λ of top-level clusters is chosen according to

$$N_\lambda = \mu_c N_c \tag{19}$$

as an integer multiple μ_c of N_c, then the same number μ_c of top-level clusters $C_{j,\lambda}$ can be assigned to each CPU. Figure 3

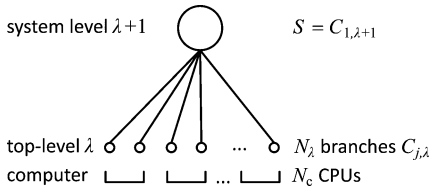


Figure 3. Top level: System $S = C_{1,\lambda+1}$ split into N_λ top-level clusters $C_{j,\lambda}$, the so-called branches, which form the roots of quaternary trees extending toward the lower levels $0 < l < \lambda$ and hierarchically decomposing the $C_{j,\lambda}$. Identical numbers N_λ/N_c of branches are assigned to the N_c CPUs of a parallel computer.

illustrates such an assignment of top-level clusters $C_{j,\lambda}$ to the N_c CPUs of a parallel computer (for $\mu_c = 2$). Because of the quaternary tree structure, each top-level cluster $C_{j,\lambda}$ contains on average 4^λ SUs $C_{k,0}$ each comprising on average $\langle|C_0|\rangle$ atoms. Thus, load-balance requires that the number of atoms per CPU is approximately given by $N/N_c \approx \mu_c 4^\lambda \langle|C_0|\rangle$, which is the atom number expected from assigning μ_c top-level clusters to each CPU. Because N can be expressed by $N = N_0\langle|C_{k,0}|\rangle$ in terms of the number N_0 of SUs, the load-balance requirement becomes $\mu_c 4^\lambda \approx N_0/N_c$. This condition approximately holds, if

$$\mu_c = \lfloor N_0/(4^\lambda N_c)\rfloor \tag{20}$$

where $\lfloor\cdots\rfloor$ denotes the floor operation. Equation 20 determines the integer multiple μ_c, which is necessary to compute by eq 19 the number N_λ of top-level clusters, from the number N_c of CPUs, the height λ of the tree, and the number N_0 of structural units.

For a system of size N and a computer with N_c CPUs, the quality of the above approximation can be expected to be better for smaller λ, because then the integer multiple μ_c can be chosen larger (implying by eq 19 that the number N_λ of top-level clusters is also large) and the relative deviations from load-balance, whose upper limit is $1/\mu_c$ %, become smaller. In summary, for an optimal load balance, the height λ of the

hierarchy and, hence, the top-level cluster sizes as measured by the number 4^λ of enclosed SUs should be small.

On the other hand, the FMM computation of interactions loses efficiency, if the height λ of the hierarchy is chosen small, because then the top-level comprises many clusters, for which interactions have to be calculated. Conversely, the choice of a larger λ entails fewer and larger clusters $C_{j,\lambda}$ at the correspondingly elevated top level. Then, a substantial part of all interactions can be evaluated at a reduced computational effort at this elevated top level. Furthermore, the computational procedure for the determination of the top-level clusters, which will be introduced below in section 3.3 and in section S4.2 of the SI, scales approximately with $(N_\lambda)^\eta \times N_0$, where $1 < \eta < 2$, such that a small number N_λ of correspondingly large top-level clusters can save much of the computational effort spent on this clustering step.

As experience has shown, a reasonable compromise between these conflicting optimization targets can be obtained by the following empirical formula, which determines the height

$$\lambda = \begin{cases} 0 & \text{if } 100 > N \\ 1 & \text{if } 700 > N \geq 100 \\ \lfloor \ln(2 \times N/6.8^3)/1.4586 \rfloor & \text{else} \end{cases} \tag{21}$$

of the SAMM$_{p,q}$ hierarchy from the number N of atoms. Accordingly and as depicted in Figure 4, λ grows for $\log_{10} N \gtrsim 3$ logarithmically with N.

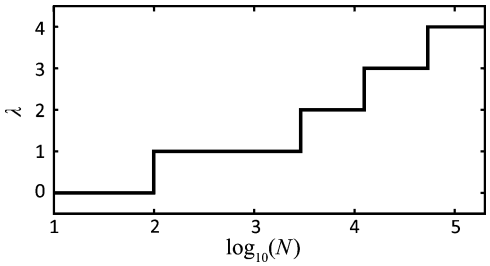


Figure 4. Height λ of the SAMM$_{p,q}$ hierarchy (cf. eq 21) for systems with N atoms.

Assume now one has a parallel computer with 32 CPUs and has found out that assigning $N_a \equiv 576$ atoms to each CPU yields a good performance in parallelized MD simulations.[51] Then, the number N_c of CPUs to be employed for a simulation of N atoms should be

$$N_c = \begin{cases} 32 & \text{if } 32 \times N_a < N \\ \lfloor N/N_a \rfloor & \text{if } 32 \times N_a \geq N \geq N_a \\ 1 & \text{else} \end{cases} \tag{22}$$

Figure 5 illustrates for the computational scenario defined by eq 22 and for the choice of eq 21 of the top-level index λ the number N_λ of top-level clusters obtained through eqs 19 and 20 for systems of size N containing either small SUs (blue, three atoms, e.g., H_2O) or medium sized SUs (red, six atoms, e.g. MeOH). Here, the system sizes are chosen from the range $N \in [10^1, 2 \times 10^5]$.

Figure 5 demonstrates that N_λ fluctuates for $N \gtrsim 10^3$ and for each of the two SU sizes around an average value, which is about 150 for the small SUs with $N/N_0 = 3$ and about 75 for the 2 times larger SUs. Whenever the index λ of the top level

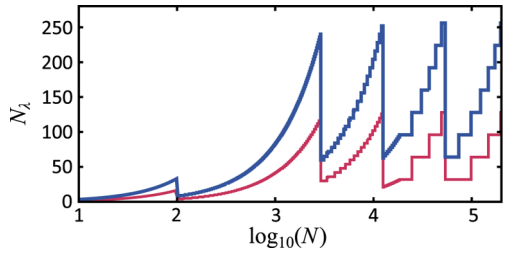


Figure 5. Number N_λ of top-level clusters $C_{j,\lambda}$ as a function of the atom number N for N_c parallel CPUs (cf. eq 22) and for SUs, which comprise three (blue) or six (red) atoms.

increases by one (cf. Figure 4), N_λ shows a sudden drop, which brings N_λ for water (blue) from about 250 down to 60. Thus, the values N_λ are range bound. Therefore, the computational effort $[\sim(N_\lambda)^\eta \times N_0]$ of top-level clustering scales linearly with the number N_0 of SUs or, equivalently, with the number N of atoms.

3.3. Top-Level Clustering. As we have seen above, a reasonable number N_λ of top-level clusters is readily chosen for a given simulation system and parallel computer. Next, the N_0 SUs together with the enclosed atoms must be assigned to the various top-level clusters $C_{j,\lambda}$. For this purpose, SAMM$_{p,q}$ applies[25] the neural algorithm suggested by Martinetz et al.[49] for vector quantization (VQ) and clustering.

This algorithm is described in section S4.2 of the SI. It manages to represent a large d-dimensional data set $\mathcal{X} \equiv \{\mathbf{x}_i | i = 1, ..., N\} \in \mathbb{R}^d$ by a much smaller so-called codebook $\mathcal{W} \equiv \{\mathbf{w}_r | r = 1, ..., M\} \in \mathbb{R}^d$ in such a way that the distribution $p(\mathbf{w})$ of the codebook vectors closely resembles the distribution $p(\mathbf{x})$ of the data.[49,50]

Assigning then each data vector $\mathbf{x}_i$ uniquely to the closest codebook vector $\mathbf{w}_{r'}$, i.e. the one obeying $\min_{\mathbf{w}_r \in \mathcal{W}} |\mathbf{x}_i - \mathbf{w}_r|$, partitions the data set $\mathcal{X}$ into M mutually disjoint and optimally compact subsets $C_r \subset \mathcal{X}$, whose centers of geometry (eq 5) are the associated code book vectors $\mathbf{w}_r$.

The application to the calculation of optimally compact SAMM$_{p,q}$ top-level clusters is now straightforward. For this purpose, the data set $\mathcal{X}$ is identified with the set $\mathcal{X}_0(t)$ collecting the N_0 geometrical centers $\mathbf{r}_{k,0}(t)$ of the SUs $C_{k,0}$ at a certain time point t of the simulation. Furthermore, the codebook $\mathcal{W}$ is identified with the set $\mathcal{W}_\lambda(t)$ comprising all N_λ geometrical centers $\mathbf{r}_{j,\lambda}(t)$ of the top-level clusters $C_{j,\lambda}(t)$.

After the VQ by the Martinetz algorithm, the SUs $C_{k,0}$ are assigned to the top-level clusters $C_{j,\lambda}(t)$ by the minimum distance criterion. Thus, the set $\mathcal{X}_0(t)$ is decomposed into N_λ disjoint subsets $\mathcal{X}_{0,j,\lambda}(t)$ containing all those geometrical centers $\mathbf{r}_{k(j),0}(t)$ of SUs $C_{k(j),0}$, for which $\mathbf{r}_{j,\lambda}(t)$ is the closest top-level cluster center. The subsets $\mathcal{X}_{0,j,\lambda}(t)$ will then contain on average N_0/N_λ vectors $\mathbf{r}_{k(j),0}(t)$.

In condensed phase systems, the centers $\mathbf{r}_{k,0}(t)$ of the SUs are uniformly distributed, if their radii of gyration are sufficiently similar. Therefore, also the codebook vectors $\mathbf{r}_{j,\lambda}(t)$ have this property. As a result, the minimum distance criterion yields a Voronoi tesselation of the simulation system into cells of approximately equal volumes, and the top-level clusters $C_{j,\lambda}(t)$ will have similar radii $R_{j,\lambda}$ of gyration.

Figure 6 shows the results of a top-level clustering for two liquid model systems (blue, H_2O; red, MeOH) each comprising $N \approx 26\,000$ atoms. The systems had been

dx.doi.org/10.1021/ct500319a | J. Chem. Theory Comput. 2014, 10, 3244−3259

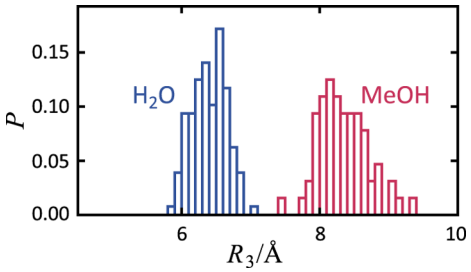


Figure 6. Normalized histograms $P(R_3)$ of radii R_3 of gyration, which characterize the top-level ($\lambda = 3$) clusters calculated for two homogeneous liquid model systems with $N \approx 26\,000$ atoms. Blue: water, $N_0 = 8737$, $N_3 = 128$; red: methanol, $N_0 = 4275$, $N_3 = 64$.

equilibrated by MD employing SAMM$_{4,1}$/RF at the temperature $T_0 = 298.15$ K and the corresponding experimental densities[52,53] (for more details, see section 5.2). Equation 21 yields, for the given values of N, the top-level index $\lambda = 3$. Equation 19 then determines the numbers N_3 specified in the caption of Figure 6. Because MeOH is larger than H_2O ($R_0^{\mathrm{MeOH}} \approx 1.77\ R_0^{\mathrm{TIP3P}}$), the average radii $\langle R_3^{\mathrm{MeOH}} \rangle = 8.4$ Å of gyration of the top-level methanol clusters are larger than their counterparts $\langle R_3^{\mathrm{TIP3P}} \rangle = 6.4$ Å in the aqueous system. According to the distributions shown in Figure 6, the standard deviations are about 4% of the average value in both cases, demonstrating that the cluster sizes actually exhibit only small standard deviations as has to be expected from a reasonable clustering algorithm.

Section S4.3 of the SI describes how the continuity and slowness of the SU motions can be exploited in the course of a MD simulation for an efficient computation of the top-level clusters. A *de novo* clustering is executed only once at the beginning of a simulation. Afterward, one keeps the top-level codebook vectors $\mathbf{r}_{j,\lambda}(t)$ adjusted to the distribution of the SU centers $\mathbf{r}_{k,0}(t)$ through an adaptation procedure, which is typically by a factor of 10 faster. For the H_2O system presented in Figure 6, for instance, the *de novo* clustering takes 0.92 s on a single CPU of a current PC, whereas an adaptive reclustering takes only 0.09 s. By default, the adaptation is regularly repeated every 256 integration steps.

3.4. Top-Down Clustering at the Intermediate Levels $1 \le l \le \lambda - 1$. At all intermediate levels $1 \le l \le \lambda - 1$ of the quaternary tree, the four children $C_{i,l}(t)$ of the parent clusters $C_{j,l+1}(t)$ are determined by the Martinetz[49] algorithm in a top-down fashion. Here, the four codebook vectors $\mathbf{r}_{i,l}(t)$, which are the geometric centers of the children $C_{i,l}(t)$, are calculated from the centers $\mathbf{r}_{k(j),0}(t)$ of all those SUs, which were associated in the preceding clustering to the parent cluster $C_{j,l+1}(t)$ and are collected in the disjoint data sets $\mathcal{X}_{0,j,l+1}$. Whereas the sizes of the codebooks $\mathcal{W}_{j,l}$ remain constant at four, the sizes of these data sets are approximately given by $|\mathcal{X}_{0,j,l+1}| \approx 4^{l+1}$ and, hence, become rapidly smaller with decreasing hierarchy level l. Therefore, the clustering of the complete quaternary tree consumes about as little computer time as the adaptive reclustering. Section S4.4 of the SI presents algorithmic details and safeguards used in top-down tree-clustering.

4. TOP-DOWN COMPUTATION OF INTERACTION LISTS

Starting at the top level, the quaternary tree is used for decisions, whether interactions should be calculated for clusters $C_{i,l}$ and $C_{j,l}$ at a given level $l \le \lambda$ or for their children (or grandchildren etc.) at the lower levels. The results of these decisions are interaction lists comprising for a cluster $C_{i,l}$ at level l the labels j of interacting clusters $C_{j,l}$ at the same level. The decisions try to optimize the compromise between accuracy and efficiency by considering the absolute errors (eq 14) of the FMM computation of the electrostatic forces. Up to a factor $1/r^2$, where r is the distance between $C_{i,l}$ and $C_{j,b}$ these errors depend on the pth power of the average

$$\langle \vartheta_l(r) \rangle_{i,j} \equiv \frac{1}{2}[\vartheta_{i,l}(r) + \vartheta_{j,l}(r)] \tag{23}$$

accuracy weighted apparent sizes

$$\vartheta_{j,l}(r) \equiv \frac{1}{a_{j,l}} \frac{2R_{j,l}}{r} \tag{24}$$

of the two clusters. Here, $R_{j,l}$ is the radius of gyration (eq 15) of $C_{j,l}$ and $a_{j,l} \ge 1$, an accuracy correction, which derives from the constant $A_{i,j,l,b}^{(\alpha)}$ appearing in eq 14 for the electrostatics case $[(\alpha,b) = (p,e)]$. Further below, we will provide reasonable estimates for the $a_{j,l}$.

A cluster $C_{j,l}$ is added to the interaction list of cluster $C_{i,l}$ (and *vice versa*), if the "interaction acceptance criterion" (IAC)

$$\langle \vartheta_l(r) \rangle_{i,j} \le \Theta \tag{25}$$

is fulfilled. Cluster pairs $C_{i,l}$ and $C_{j,l}$ missing the IAC are decomposed into their respective children, for which the IAC is checked at the next lower level $l - 1$. This top-down process of interaction list computation is continued until, at the lowest level $l = 0$, closely neighboring cluster pairs $C_{i,0}$ and $C_{j,0}$ are decomposed into individual atoms, whose interactions are computed by the exact expressions for the electrostatic and the van der Waals pair interactions. At this atomic level also all modifications of nonbonded interactions, which are dictated by the applied force field for covalently linked atoms, are properly applied.

Small values of Θ in the IAC (eq 25) lead to accurate but slow algorithms, because they exclude the computation of interactions among relatively close and large clusters. Large values of the IAC threshold Θ have the opposite effect. From a series of SAMM$_{4,1}$/RF test calculations on different systems, we have deduced the three reasonable choices Θ_χ with $\chi \in \{\mathrm{a}, \mathrm{m}, \mathrm{f}\}$ listed in Table 1. The letters χ mean "slow but very accurate" (a), "intermediate" (m), and "fast but still reasonably accurate" (f). We will denote the corresponding algorithms from now on as SAMM$_{4,1}^{\chi}$/RF.

Table 1. SAMM$_{4,1}$: Reasonable Values for Θ

name	Θ_a	Θ_m	Θ_f
value	0.17	0.20	0.25

4.1. Top Level $\tilde{\lambda}(\Theta_\chi)$ in Periodic Systems. The top-down procedure of interaction list generation must be modified for reasons explained in detail by Mathias et al.,[28] if toroidal boundary conditions are applied and if the electrostatics is described by a RF approach for interaction distances beyond the MIC cutoff $d_{\mathrm{MIC}} = L/2$ (cf. section 1), i.e. if a SAMM$_{p,q}^{\chi}$/RF algorithm is applied. Before entering this issue, we would like to note that the MIC cutoff is efficiently implemented at the top level λ by replicating the N_λ top-level clusters $C_{j,\lambda}$ in the periodic cells surrounding the central one and by checking, for all clusters $C_{j,\lambda}$ in the central cell, the MIC cutoff condition with

respect to complete ensemble of top-level clusters, which also covers all periodically replicated cells.

To sketch the necessary modification of interaction list generation, which is thoroughly motivated and explained in section S5 of the SI, we assign for a given IAC threshold Θ_χ to each level l the distance

$$d_l(\Theta_\chi) \equiv 2(\langle R_l\rangle + \langle \tilde{R}_l\rangle/\Theta_\chi) \tag{26}$$

where $\langle R_l\rangle$ and $\langle \tilde{R}_l\rangle$ are ensemble averages of the radii $R_{i,l}$ of gyration and of their accuracy weighted counterparts

$$\tilde{R}_{i,l} \equiv R_{i,l}/a_{i,l} \tag{27}$$

at level l. $d_l(\Theta_\chi)$ measures the typical interaction distances of clusters at level l. Starting at the top level λ and descending the tree, it is checked after each clustering step whether the level-associated distance $d_l(\Theta_\chi)$ complies through

$$d_l(\Theta_\chi) \le d_{\mathrm{MIC}} \tag{28}$$

with the MIC. The first level $l > 0$, for which the inequality (eq 28) holds, will be called the "interaction top level" and denoted by $\tilde{\lambda}(\Theta_\chi)$. Note that section S5 of the SI also discusses precautions for very small systems, for which one may get $\tilde{\lambda}(\Theta_\chi) = 0$. Furthermore, the SI compares in Figure S13 for increasing system sizes the growth of λ (cf. eq 4) with that of $\tilde{\lambda}(\Theta_f)$ taking the pure liquid systems water and methanol as examples.

4.2. Smooth Transitions Across the MIC Boundary. Cluster pairs with $r \approx d_{\mathrm{MIC}}$ may move during simulated dynamics into or out of the dielectric continuum extending at distances beyond d_{MIC}. Mathias et al.[28] have suggested an algorithm, which smoothly handles such transitions. For this purpose, they defined an effective size

$$R_{i,j,l} \equiv [(R_{i,l}^3 + R_{j,l}^3)/2)]^{1/3} \tag{29}$$

for a cluster pair with the radii $R_{i,l}$ and $R_{j,l}$ of gyration. The SAMM$^\chi_{p,q}$ calculation of long-range interactions is smoothly replaced by a RF description, if the cluster distance r obeys

$$d_{\mathrm{MIC}} - \Delta_{\tilde{\lambda}} - R_{i,j,l} \le r \le d_{\mathrm{MIC}} - \Delta_{\tilde{\lambda}} + R_{i,j,l} \tag{30}$$

where $\Delta_{\tilde{\lambda}}$ is the maximal half-width of a transition region.[28] Here, $\Delta_{\tilde{\lambda}}$ is determined by

$$\Delta_{\tilde{\lambda}} = \min\left[\langle R_{\tilde{\lambda}}\rangle \frac{d_{\mathrm{MIC}}}{d_{\tilde{\lambda}}(\Theta)}, R_{\tilde{\lambda}}^{\max}\right] \tag{31}$$

and, thus, is given in terms of quantities, which characterize the interaction top-level $\tilde{\lambda}$. These are the distance $d_{\tilde{\lambda}}(\Theta)$ defined by eq 26 as well as the average and the maximal radii of gyration $\langle R_{\tilde{\lambda}}\rangle$ and $R_{\tilde{\lambda}}^{\max} \equiv \max_i\{R_{i,\tilde{\lambda}} \mid i = 1, ..., N_{\tilde{\lambda}}\}$, respectively.

For cluster pairs obeying $R_{i,j,l} \le \Delta_{\tilde{\lambda}}$, which is likely for top-level cluster pairs because of the MIC condition (eq 28) selecting $\tilde{\lambda}$, the actual half-width of the transition region is $R_{i,j,l}$. Interactions of cluster pairs with $R_{i,j,l} > \Delta_{\tilde{\lambda}}$ are treated at the next lower level (cf. section S5 in the SI).

Clusters, which occupy the transition region, smoothly fade away into or reappear out of the dielectric continuum with changing distance r, and so do their SAMM$^\chi_{p,q}$/RF interactions.[28] At the distance $d_{\mathrm{MIC}} - \Delta_{\tilde{\lambda}}$, for instance, which marks the center of the transition region, the SAMM$^\chi_{p,q}$ cluster−cluster interactions are scaled down by a factor one-half and the RF model of the electrostatics acts at half of its full strength.[28] For energy and pressure evaluations, the dispersion interactions with atoms more distant than $d_{\mathrm{MIC}} - \Delta_{\tilde{\lambda}}$ are included by a mean field term.[43]

4.3. Bottom-Up Calculation of Multipole Moments $\mathbf{M}^{m,\mathbf{c}}$. As is explained in section 2.4 of ref 29 for the SAMM$_p$ electrostatics treatment, FMM algorithms calculate the mth order multipole moments $\mathbf{M}^{m,\mathbf{c}}$ with respect to the reference point $\mathbf{c}$ of a parent cluster C on a level $l > 0$ from the multipole moments of its children $c \subset C$ (c.f. also ref 30 for the treatment of dipole distributions). For this purpose, the multipole moments of C are first calculated with respect to the origin $\mathbf{0}$ as simple sums of the corresponding moments of its children. Shifting then the reference point from $\mathbf{0}$ to $\mathbf{c}$ by a procedure that is specified by eqs 19−22 in ref 29 yields the desired moments $\mathbf{M}^{m,\mathbf{c}}$ from the $\mathbf{M}^{m,\mathbf{0}}$.

For the dispersion, this procedure is almost identical to that of electrostatics. Solely the recursion relation (eq 19 in ref 29) for the auxiliary tensors $\mathbf{H}^i_{m,\mathbf{c}}$, $i \in \{0, ..., m\}$ is replaced by the slightly modified expression

$$\mathbf{H}^{k+1}_{m,c} = \mathbf{M}^{k+1,0} - \frac{(k+1)}{m-k} S_{k+1}[(2k+6)(\mathbf{c} \otimes \mathbf{H}^k_{m,c}) - k(\mathbf{c} \odot \mathbf{H}^k_{m,c}) \otimes \mathbf{I}] \tag{32}$$

in which S_k denotes the symmetrizer for the components of rank k tensors (eq 22 in ref 29). The recursion starts with $\mathbf{H}^0_{m,\mathbf{c}} \equiv \mathbf{M}^{0,0}$, and the shifted moments are $\mathbf{M}^{m,\mathbf{c}} = \mathbf{H}^m_{m,\mathbf{c}}$.

4.4. Top-Down Calculation of Expansion Coefficients $T^{D,n,q}(\mathbf{c})$. In contrast, the Taylor expansion coefficients $\mathbf{T}^{D,n,q}(\mathbf{c})$, which are defined by eq 8, are computed in a top-down fashion. At each level $l \le \tilde{\lambda}(\Theta_\chi)$, the multipole moments of all clusters D, which fulfill for a given cluster C the IAC (eq 25), contribute through eq 8 to the coefficients $\mathbf{T}^{D,n,q}(\mathbf{c})$. Furthermore, the action of the clusters D is inherited by the children c of C through a procedure that shifts the reference point of the Taylor expansion from the center $\mathbf{c}$ of C to the centers of the children $c \subset C$.

Because the computation and inheritance of the Taylor expansion coefficients is formally identical for dispersive and electrostatic interactions, a reference to the detailed description in section 2.5 of ref 29 must suffice here. The quoted methods then guarantee that all dispersive interactions between higher level clusters are inherited to the lowest level, where the resulting Taylor expansions are used to compute the contributions of distant dispersion charges to the potential and force acting on an individual atom.

5. METHODS

The computations carried out within this study served for two different purposes, that is the fine-tuning of SAMM$^\chi_{p,q}$ and the thorough evaluation of the compromises $\chi \in \{a, m, f\}$ between efficiency and accuracy.

5.1. Fine Tuning. The calculation of the accuracy weighted apparent size $\vartheta_{j,l}(r)$ of a cluster $C_{j,l}$ defined by eq 24 requires estimates for accuracy corrections $a_{j,l}$ for clusters of all sizes and chemical compositions. These estimates should guarantee an approximately homogeneous accuracy at all levels of a SAMM$^\chi_{p,q}$ description.

As the reference cluster, we take the TIP3P[54] model of a water molecule j. This cluster is a SU of the type $X = T \equiv$ TIP3P (cf. Table S3 in the SI) and is localized at the level $l = 0$ of the SAMM hierarchy. For a pure TIP3P water system, we define

$$a_{j,0} = a_T \equiv 1 \tag{33}$$

such that the accuracy weighted apparent size $\vartheta_T(r) \equiv \vartheta_{j,0}(r)$ of a TIP3P model j solely depends on its radius of gyration $R_T \equiv R_{j,0}$ and on the distance r, i.e. reduces to the common apparent size. Because the average apparent size of a pair of TIP3P models is simply $\langle \vartheta_l(r) \rangle_{i,j} = \vartheta_T(r)$, the IAC (eq 25) becomes for $\Theta = \Theta_f$ the distance criterion $r \geq d_T(\Theta_f)$, where

$$d_T(\Theta_f) \equiv 2R_T/\Theta_f = 5.42\text{Å} \tag{34}$$

marks the boundary between a $\text{SAMM}^f_{p,q}$ and the exact description.

Thus, in a pure TIP3P water system, the $\text{SAMM}^f_{p,q}$ approximations are replaced by the exact computation of the electrostatic and dispersive pair interactions as soon as two molecules cross the boundary at $d_T(\Theta_f)$ upon mutual approach. This change of description causes random errors, which represent algorithmic noise.

One can empirically estimate for a pair of clusters C and D the size of such errors by computing the root-mean-square deviation (RMSD)

$$\Delta f^{(p,q)}_{C,D}(r) \equiv \left\{ \left\langle \frac{1}{3|C|} \sum_{i \in C} [\mathbf{f}(\mathbf{r}_i) - \mathbf{f}^{p,q}(\mathbf{r}_i)]^2 \right\rangle_{\mathcal{A}} \right\}^{1/2} \tag{35}$$

between the exact $[\mathbf{f}(\mathbf{r}_i)]$ and approximate $[\mathbf{f}^{p,q}(\mathbf{r}_i)]$ force components. This RMSD is evaluated for an ensemble $\mathcal{A}$ of 2×10^4 randomly chosen mutual orientations of the two clusters, which are separated by a fixed distance r. The considered forces act on the atoms i of cluster C and originate from the electrostatic and dispersion charges of the atoms j in cluster D. To estimate the algorithmic noise in $\text{SAMM}^f_{p,q}$/RF simulations of TIP3P water, the RMSD $\Delta f^{(p,q)}_T(r) \equiv \Delta f^{(p,q)}_{C,D}(r)$ between exact and approximate force components should be calculated at the IAC boundary $r = d_T(\Theta_f)$.

One can calculate such RMSDs $\Delta f^{(\alpha)}_{C,D,b}(r)$ also separately for the electrostatic $[(\alpha,b) = (p,\text{e})]$ or the dispersive $[(\alpha,b) = (q,\text{d})]$ forces, if one wants to judge the relative sizes of the errors. Furthermore, one can vary the orders p and q of the respective FMM expansions, if one wants to identify reasonable combinations (p,q) of expansion orders. Finally, one can check to what extent the empirical errors $\Delta f^{(\alpha)}_{C,D,b}(r)$ are covered by the first neglected terms $\Delta \tilde{f}^{(\alpha)}_{C,D,b}(r)$ of the SAMM_α expansions of the forces. These SAMM_α error estimates are given by eq 14.

We have extensively studied these issues not only for pairs of water molecules but also for many pairs of other SUs X commonly occurring in protein solvent systems. In analogy to eq 34, which applies to TIP3P, we chose also here the distance

$$r_X \equiv 2R_X/\Theta_f \tag{36}$$

for the computation of the empirical errors $\Delta f^{(\alpha)}_{X,\beta}(r_X) \equiv \Delta f^{(\alpha)}_{C,D,\beta}(r_X)$ (cf. eq 35). The radii R_X of gyration (cf. Tables S3 and S4 in the SI) as well as the electrostatic and dispersion charges were taken from CHARMM22[5] and from other sources quoted in the tables.

Assuming now that the empirical errors $\Delta f^{(\alpha)}_{X,\beta}(r)$ are well represented at all distances by the SAMM_α estimates $\Delta \tilde{f}^{(\alpha)}_{X,b}(r)$ (c.f. eq 14), the free parameter $A^{(\alpha)}_{X,b} \equiv A^{(\alpha)}_{C,D,b}$ of $\Delta \tilde{f}^{(\alpha)}_{X,b}(r)$ can be calculated from setting $\Delta f^{(\alpha)}_{X,b}(r_X) = \Delta \tilde{f}^{(\alpha)}_{X,b}(r_X)$. Inserting eq 14 yields

$$A^{(\alpha)}_{X,b} = \Delta f^{(\alpha)}_{X,b}(r_X) r_X^{\alpha+\gamma(b)}/(2R_X)^\alpha \tag{37}$$

with r_X defined by eq 36 and $\gamma(b)$ by eq 13. The thus determined SAMM_α error estimates $\Delta \tilde{f}^{(\alpha)}_{X,b}(r)$ now enable us to address the question at which distance r these force errors become equal to the reference errors $\Delta \tilde{f}^{(\alpha)}_{T,b}[d_T(\Theta)]$ of TIP3P at the boundary $d_T(\Theta_f)$ between the exact and $\text{SAMM}_{p,q}$ descriptions.

If we assume that the errors are dominated by the electrostatics and that the order of the electrostatic SAMM_p expansion is $p = 4$, then this question amounts with eq 14 to the equation

$$\frac{A^{(4)}_{T,e} R_T^4}{A^{(4)}_{X,e} R_X^4} = \left(\frac{d_T(\Theta_f)}{r} \right)^6 \tag{38}$$

where the TIP3P boundary distance $d_T(\Theta_f)$ is given by eq 34. Now we additionally require that the distance r is just the IAC distance

$$d_X(\Theta_f) = \frac{1}{a_X} \frac{2R_X}{\Theta_f} \tag{39}$$

for a SU pair of type X, which follows for $\Theta = \Theta_f$ from eqs 25 and 24. Setting $r = d_X(\Theta_f)$ and inserting eqs 34 and 39 into eq 38 yields, after a few rearrangements, the accuracy corrections

$$a_X = \left[\frac{A^{(4)}_{T,e}}{A^{(4)}_{X,e}} \left(\frac{R_X}{R_T} \right)^2 \right]^{1/6} \tag{40}$$

of the SUs X as functions of the constants $A^{(4)}_{X,e}$ and R_X. On the basis of the assumptions noted above, the thus determined accuracy corrections a_X guarantee that the electrostatic SAMM_4 force errors at the IAC boundary $d_X(\Theta_f)$, which separates the exact and $\text{SAMM}_{p,q}$ descriptions, resemble the corresponding errors encountered in a TIP3P reference system. By construction, the a_X should be transferable to other choices of Θ. The above procedure can be extended toward larger clusters, and we have carried out corresponding experiments with pairs of pure methanol and TIP3P clusters, each of which comprised four SUs and was selected from corresponding simulation systems.

5.2. Evaluation of the $\text{SAMM}^x_{p,q}$/RF Accuracy by MD Simulations. Whenever a pair of clusters crosses an IAC, a MIC, or a cutoff boundary during a dynamics simulation, the approximation and, hence, the detailed values of the interatomic forces experience small sudden changes. Efficient MD programs check and realize boundary crossings at the regular time points $t_T \equiv T\tau$, $T = 0, 1, \ldots$ of the interaction list updates, where $\tau = u\Delta t$ is an integer multiple of the integration time step Δt (IPHIGENIE: $u = 64$). Depending on the nature of the forces, the changes may either cause a heating or a cooling of the system.

We studied the size of these artifacts for various $\text{SAMM}_{p,q}$/RF algorithms using two liquid systems enclosed by periodic cubic boxes as test beds. System $\mathcal{T}$ was filled with $N_\mathcal{T} = 2133$ TIP3P[54] water models and system $\mathcal{M}$ with $N_\mathcal{M} = 952$ CHARMM22[5] methanol models. All lengths of bonds involving hydrogen atoms and the bond angle of the TIP3P model were kept at their equilibrium values by applying the MSHAKE[55] and RATTLE[56] algorithms with a relative tolerance of 10^{-10}. The chosen experimental densities[52,53] at the standard temperature $T_0 = 298.15$ K and pressure $p_0 = 1$ bar yielded box-lengths $L \approx 40$ Å. The dielectric constants ε_{RF} of the surrounding continua were set to the experimental values[57,58]

78 (H_2O) and 32.7 (MeOH). Keeping the particle numbers N and volumes V fixed, the systems were equilibrated for 1 ns at T_0 in the NVT ensemble using a Bussi[59] thermostat (with a coupling time of 0.5 ps) for temperature control, the $SAMM^f_{4,1}$/RF algorithm for the long-range interactions, and (as always) a time step $\Delta t = 1$ fs for the integration of the dynamics with the velocity Verlet algorithm.[60] Note that the above construction procedure was analogously applied to the systems discussed in Figure 6.

The NVT simulations of the systems $\mathcal{G} \in \{\mathcal{T}, \mathcal{M}\}$ were continued for another 2 ns. Snapshots drawn at 10 ps delays generated for each $\mathcal{G}$ an ensemble $\mathcal{I}_\mathcal{G}$ of 200 statistically independent initial conditions. Each ensemble $\mathcal{I}_\mathcal{G}$ was the common starting point for several ensembles $\mathcal{T}_\mathcal{G}(\Theta, P)$ of short NVE simulations, each of which covered the time span $\delta t \equiv 10$ ps. These ensembles differed by the choices of the IAC threshold Θ and of the three-parametric simulation settings $P \equiv (p, q, c_r)$, which signify a specific choice of the $SAMM_{p,q}$/RF expansion orders and of the Pauli repulsion cutoff distance c_r. This repulsion is represented, in the given cases, by the $1/r^{12}$ contribution to the Lennard-Jones potentials.[61]

With the aim of singling out specific sources of algorithmic noise, which may transfer heat into or out of a simulation system, the simulation settings $P = (p, q, c_r)$ were grouped into comparative pairs, which are listed and named in Table 2 (they will be explained in detail further below). Each pair consists of a supposedly more exact reference simulation P_{ref} and a specific simulation P differing from P_{ref} usually in only one (but sometimes also more than one) of the three parameters (p, q, c_r). The differing parameters mark specific sources of algorithmic noise. Therefore, measurements of heat production differences

Table 2. Parameter Sets P and P_{ref} for Heating Rate Comparisons

comparison	$P = (p, q, c_r)$	$P_{ref} = (p, q, c_r)_{ref}$
$c_r = d_X$	(∞, ∞, d_X)	$(\infty, \infty, d_{MIC})$
$q = -1$	$(\infty, -1, d_X)$	(∞, ∞, d_X)
$q = 3$	$(4, 3, d_X)$	$(4, \infty, d_X)$
$q = 2$	$(4, 2, d_X)$	$(4, \infty, d_X)$
$q = 1$	$(4, 1, d_X)$	$(4, \infty, d_X)$
$p = 4$	$(4, 3, d_X)$	$(\infty, 3, d_X)$
$p = 3$	$(3, 3, d_X)$	$(\infty, 3, d_X)$
$SAMM_{4,3}$	$(4, 3, d_X)$	$(\infty, \infty, d_{MIC})$
$SAMM_{4,2}$	$(4, 2, d_X)$	$(\infty, \infty, d_{MIC})$
$SAMM_{4,1}$	$(4, 1, d_X)$	$(\infty, \infty, d_{MIC})$

$$\Delta\dot{Q}(\Theta, P, P_{ref}) = \dot{Q}(\Theta, P) - \dot{Q}(\Theta, P_{ref}) \quad (41)$$

which were observed in these pairs of NVE simulations, identify the amount of heat produced by these and only these sources at each given IAC threshold Θ. Other possible sources of heat, like, e.g., the choice of the SHAKE tolerance or of the time-step of the dynamics integration, are eliminated by the formation of the heating rate differences according to eq 41. The required heating rates per solvent molecule

$$\dot{Q}(\Theta, P) \equiv \langle E(\delta t) - E(0)\rangle_{\mathcal{T}(\Theta,P)}/\delta t \quad (42)$$

were calculated as ensemble averages from the total energies $E(t)$ per molecule observed at the beginning ($t = 0$) and end ($t = \delta t = 10$ ps) of the NVE trajectories contained in the simulation ensembles $\mathcal{T}_\mathcal{G}(\Theta, P)$.

In these simulations, the IAC threshold Θ was sampled over the range [0.14, 0.26] by 13 regularly spaced values. Furthermore, the interaction top level was confined to $\tilde{\lambda} = 0$; i.e., the $SAMM_{p,q}$ description was solely applied to the SUs $X =$ T (= TIP3P) or $X =$ M (≡ MeOH).

As is apparent from the characterization of the comparisons in Table 2 through the parameter sets P and P_{ref}, the SAMM expansion orders p and q were usually chosen for the electrostatics as $p \in \{3, 4\}$ and for the dispersion as $q \in \{1, 2, 3\}$. In the corresponding simulations, the SUs X were resolved into individual atoms for the exact computation of the long-range interactions as soon as the inter-SU distance r became smaller than the IAC boundary $d_X(\Theta)$ associated with Θ by eq 39. Transitions of SUs across this IAC boundary will then cause a certain amount of algorithmic noise. However, besides the just quoted expansion orders p and q, one also detects the strange expansion orders $p = \infty$, $q = \infty$, and $q = -1$.

Here, $p = \infty$ denotes the limiting algorithm $\lim_{p\to\infty} SAMM_{p,q}$/RF, in which the IAC distance $d_X(\Theta)$ is selectively shifted for the electrostatic interactions to d_{MIC} (≈ 20 Å). Thus, the electrostatics is calculated exactly within a sphere of radius $d_{MIC} - 2R_X$ (c.f. section 4.2), while beyond that sphere the cluster-based smooth transition into and out of the dielectric continuum is maintained. $q = \infty$ analogously signifies that the dispersion is calculated exactly up to $d_{MIC} - 2R_X$ and experiences a smooth cutoff in the following transition zone of the width $2R_X$. Finally, $q = -1$ indicates the complete neglect of the dispersion at distances $r \geq d_X(\Theta)$, i.e. the common short-range dispersion cutoff.

According to Table 2, the Pauli repulsion cutoff distance c_r was usually chosen as $d_X(\Theta)$, except in several reference simulations, in which this distance was shifted outward up to d_{MIC}, implying that, here, the effects of the repulsion cutoff are negligibly small.

The table starts with the comparison denoted by "$c_r = d_X$" and shows that the associated simulation parameters P and P_{ref} solely differ by the choice of the repulsion cutoff c_r, which is shifted from usually small values $d_X(\Theta)$ to d_{MIC}, where the repulsion cutoff can be neglected. Hence the associated heating rate difference $\Delta\dot{Q}(\Theta, P, P_{ref})$ measures the contribution of the repulsion cutoff at $d_X(\Theta)$ to the overall violation of energy conservation.

Similarly, in the next entry "$q = -1$", the only difference between P and P_{ref} is that the use of a dispersion cutoff at $d_X(\Theta)$ in P is abandoned in favor of an exact computation of the dispersion in a range up to d_{MIC}. Thus, this comparison can reveal the contribution of a short-range dispersion cutoff to the overall algorithmic heat production $\dot{Q}(\Theta, P)$.

The following comparisons "$q = 3, 2, 1$" and "$p = 4, 3$" measure to what extent the cutoff of the $SAMM_p$ electrostatics or the $SAMM_q$ dispersion expansion after the indicated orders p and q, respectively, contributes to the overall algorithmic noise. Here, the reference simulations are either carried out with exact dispersion ($q = \infty$) or with exact electrostatics ($p = \infty$) such that the associated difference eq 41 actually yields the announced insight.

Finally, the last three rows characterize comparisons, which serve to identify the combined contributions of the $SAMM_{4,q}$ expansions ($q = 3, 2, 1$) and of the repulsion cutoff at $d_X(\Theta)$ to the total algorithmic heat production. For this purpose, these comparisons suppress all those contributions, which are due to

transitions of distant clusters into or out of the dielectric continuum extending beyond d_{MIC}. The latter contributions, which we call $\dot{Q}_{RF}$, are independent of Θ, solely depend on the system size, and decrease with $1/d_{MIC}$, because their sources are confined to a spherical surface of radius d_{MIC}.

5.3. Check of Linear Scaling. We have prepared a series of periodic simulation boxes $\mathcal{G}_i$, i = 0, 1, ..., 30, with the side lengths L_i = 40 + 2i Å. They were filled either with the nonpolarizable TIP3P[54] or with the recent so-called TL6P[8] polarizable six-point water models at the experimental density[52] n = 0.997 g/cm^3 for T_0 = 298.15 K and p_0 = 1 bar. Note that TL6P features an inducible Gaussian dipole distribution centered at the oxygen, two positive point charges at the hydrogens, and three negative mass-less point charges near the oxygen. Correspondingly, we call the simulation systems $\mathcal{G}_i$ either $\mathcal{T}_i$ (TIP3P) or $\mathcal{B}_i$ (TL6P). The $\mathcal{G}_i$ were equilibrated by SAMM$^f_{4,1}$/RF-MD simulations for about 100 ps in the NVT_0 ensemble controlling T by a Berendsen[62] thermostat with the coupling time τ = 0.5 ps. In the $\mathcal{B}_i$ simulations, the threshold for the self-consistency iteration of the components of the induced dipoles was set to 10^{-4} D. For information on the various methods implemented in IPHIGENIE to speed up the self-consistency iterations of the induced dipoles in the $\mathcal{B}_i$ simulations, see Section III.B in ref 44.

Computing times t per time step of the SAMM$^\chi_{4,1}$/RF dynamics integration were measured by averaging over 30 integration steps for the three accuracy/efficiency choices $\chi \in$ {a, m, f} and for each equilibrated box on a single core of a 3 GHz Intel Core2 Duo CPU E8400.

6. RESULTS

With the aim of generating similarly accurate SAMM$^\chi_{4,1}$ descriptions for all types of SUs and higher order clusters occurring in protein−solvent systems, we have introduced in section 5.1 several assumptions that finally led to eq 40, from which one can calculate the accuracy corrections a_X of SUs and corresponding corrections $a_{j,l}$ of higher order clusters. In SAMM$^\chi_{p,q}$, these corrections are required for the evaluation of the IAC condition eqs 25 by means of the accuracy weighted apparent sizes eq 24. The following presentation of results concentrates on the default value p = 4 for the order of the SAMM$_p$ electrostatics expansion.

6.1. Verification of Fine Tuning Assumptions. *Assumption 1.* The arguments leading to eq 37 essentially rest on the assumption that the empirical errors $\Delta f^{(4)}_{X,e}(r)$ of the SAMM$_4$ electrostatics expansion are well described at all distances r by the analytical estimates $\Delta\tilde{f}^{(4)}_{X,e}(r)$, whose sole parameters $A^{(4)}_{X,e}$ are calculated by eq 37. These analytical estimates are defined by eq 14 and are empirically parametrized at $r = r_X$ (cf. eq 36). If one expresses these estimates as functions of the dimensionless distances $\tilde{r}_X \equiv r/2\tilde{R}_X$, where $\tilde{R}_X$ is the accuracy weighted radius of gyration (eq 27) of a SU X, then one finds by eq 40 that the estimates are given by the master formula

$$\Delta\tilde{f}^{(4)}_{X,e}(\tilde{r}_X) = A^{(4)}_{T,e}/(2R_T)^2\tilde{r}_X^6 \quad (43)$$

which solely depends on parameters belonging to the TIP3P reference SU (X = T). If one plots the empirical force errors (eq 35) as functions of the dimensionless distances $\tilde{r}_X$, then they all should fall onto the master curve given by eq 43.

Figure 7 demonstrates that the empirical SAMM$_4$ force errors of TIP3P (blue circles) and MeOH (red crosses), whose

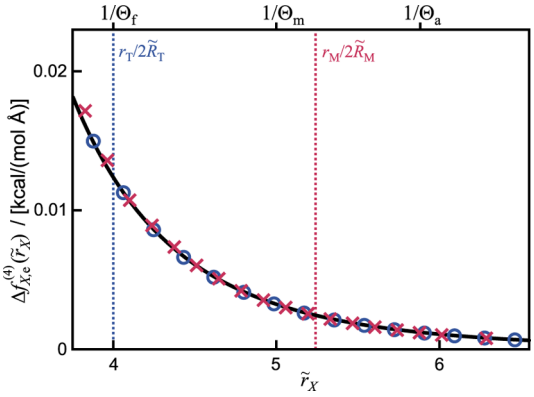


Figure 7. Empirical force errors $\Delta f^{(4)}_{X,e}(\tilde{r}_X)$ calculated by means of eq 35 are compared on the scale of the dimensionless distances $\tilde{r}_X$ for TIP3P water (blue circles) and MeOH (red crosses) with the predictions of master formula eq 43 (black) expressing the error estimates eq 14 for the SAMM$_4$ electrostatics expansion.

numerical values are listed in Table S6 of the SI, are very well described by the analytical estimate eq 43 (black) over the shown range of distances $\tilde{r}_X$. These distances are relevant, because the IAC criterion eq 25 may be rewritten as $\tilde{r}_X \geq 1/\Theta$ and because the reciprocal values of the standard IAC thresholds Θ listed in Table 1 are, as is indicated in the figure, in the given range. The blue and red dashed lines mark the locations of the distances r_T and r_M (cf. eq 36), at which the analytical estimates were parametrized, on the $\tilde{r}_X$ scale. In the shown range, the relative force errors are all below 5%. For $\tilde{r}_X \geq$ 4, they are even smaller than 2%. Thus, assumption 1 holds with great accuracy.

Assumption 2. The next key point of the arguments leading to eq 38 was the assumption that the empirical errors (eq 35) at the boundary $d_T(\Theta_f)$ between the exact and SAMM$_{p,q}$ descriptions are dominated for TIP3P by the contributions $\Delta f^{(4)}_{T,e}[d_T(\Theta_f)]$ of the SAMM$_4$ electrostatics expansion. For its check, we have additionally calculated the force errors $\Delta f^{(4)}_{T,d}[d_T(\Theta_f)]$ of the SAMM$_q$ dispersion expansions for q = 1, 2, and 3 and the ratios

$$E_{4,q} \equiv \Delta f^{(q)}_{T,d}[d_T(\Theta_f)]/\Delta f^{(4)}_{T,e}[d_T(\Theta_f)] \quad (44)$$

between the empirical errors of these expansions.

For TIP3P, one gets the ratio $E_{4,1}$ = 0.34 demonstrating that, at the IAC boundary $d_T(\Theta_f)$ = 5.42 Å, the errors of the SAMM$_1$ dispersion expansion are by a factor of 0.34 smaller than those of the SAMM$_4$ electrostatics expansions. Next, the ratios $E_{4,2}$ = 0.11 and $E_{4,3}$ = 0.04 prove that the quality of the SAMM$_q$ dispersion expansion gets rapidly better with increasing q. We have checked this issue also for other SUs (data not shown) and found similar ratios and dependences on q. Hence, also assumption 2 clearly holds such that the validity of the arguments leading to formula 40 for the accuracy corrections a_X has been demonstrated.

6.1.1. Limiting the Range of the Accuracy Corrections a_X. Applying the procedures explained in connection with eq 35, we have calculated empirical force errors $\Delta f^{(4)}_{X,e}(r_X)$ at the reference distances r_X given by eq 36 for a series of SUs X typically occurring in protein solvent systems. These SUs are listed in Tables S3 and S4 in the SI. Subsequently, we have calculated through eqs 37 and 40 accuracy corrections a_X (cf. eq 40) for all these SUs.

For several large and hardly polar SUs we found quite large values $a_X \gtrsim 2$, which imply that $\tilde{R}_X$ becomes smaller than $R_X/2$. Although the electrostatics description remains sufficiently accurate at the correspondingly close IAC boundary $d_X = 2\tilde{R}_X/\Theta_f$, this may not be the case for the SAMM$_q$ dispersion expansion, whose accuracy depends on $2R_X$ but not on $\tilde{R}_X$. Here, the IAC boundary should be moved closer to $2R_X/\Theta$.

Therefore, we decided to introduce reasonable upper and lower bounds for a_X by the function

$$f(a_X) = \begin{cases} 1 & \text{if } a_X < 1 \\ a_X & \text{if } 1 \le a_X \le 1.8 \\ 1.8 & \text{else} \end{cases} \quad (45)$$

and subsequently identified the accuracy correction with its bounded values $[a_X \leftarrow f(a_X)]$. The resulting bounded values are listed in Tables S3 and S4 of the SI.

6.1.2. Accuracy Corrections a_C of Clusters at Levels $l > 0$. The check of the IAC condition (eq 25) for higher level clusters C requires accuracy weighted apparent sizes $\vartheta_C(r)$ as defined by eq 24 and, hence, accuracy corrections a_C. With the aim of getting an idea whether the a_C's are related to the a_X's of the SUs $X \in C$, we randomly chose from TIP3P and MeOH simulation systems 10 clusters each comprising four SUs. Applying the procedures explained in section 5.1 and averaging over the 10 pairs composed of identical clusters, we found by the analysis of the electrostatic force errors at the distances $r_C = 2R_C/\Theta_f$ the ratios $a_{C(T)}/a_T = 1.47$ and $a_{C(M)}/a_M = 1.44$ of the cluster corrections to the (bounded) corrections of the enclosed SUs X with $X = T$ or $X = M$.

Because one cannot possibly calculate accuracy corrections for all kinds of clusters occurring in protein solvent systems, we decided to convert the apparent similarity of the above ratios into a rule. Thus, we define the accuracy correction for a cluster C at level $l > 0$

$$a_C = \langle a_c \rangle_C \times \begin{cases} 1.45 & \text{for } l = 1 \text{ and } |C| > 1 \\ 1 & \text{else} \end{cases} \quad (46)$$

as the given multiples of the average (bounded) accuracy correction of the children $c \in C$. Hence, the factor 1.45 applies only to the transition from SUs, which generally contain covalently connected atoms, to clusters at level $l = 1$, which mainly contain noncovalently attached atoms. For all further transitions, the a_C's of higher level ($l > 1$) clusters are simply averages of the a_c of their children c.

6.1.3. Accuracy Corrections $a_C > 1$ Enhance the Efficiency. Considering a homogeneous system consisting of SUs X with (bounded) accuracy corrections $a_X > 1$, one recognizes that the spheres $S(a_X, \Theta)$ of radius $d_X(\Theta, a_X) = 2R_X/a_X\Theta$, within which the SUs have to be resolved by the IAC criterion (eq 25) into atoms, contain rapidly much fewer atoms with increasing a_X. Correspondingly, much of the costly evaluation of atomic pair interactions can be saved.

If we denote the number of atoms within $S(a_X, \Theta)$ by $N(a_X, \Theta)$ and assume a homogeneous density within the simulation system, then we get

$$N(a_X, \Theta) = N(1, \Theta)/a_X^3 \quad (47)$$

where $N(1, \Theta)$ is the number of atoms within the reference sphere $S(1, \Theta)$ defined by $a_X = 1$. For a given value of Θ, the atom numbers within the spheres $S(a_X, \Theta)$ vary in the range $N(1, \Theta) \ge N(a_X, \Theta) \ge 0.17N(1, \Theta)$, because $1 \le a_X \le 1.8$. Therefore, one expects that the cost of computing the exact atomic pair interactions within the spheres $S(a_X, \Theta)$ is for $a_X = 1.0$ by a factor 5.8 larger than for $a_X = 1.8$.

Similar considerations apply to clusters at levels $l > 1$, because here the accuracy corrections are within the range $1.45 \le a_C \le 2.61$, which shift the IAC boundaries to much smaller values than those resulting for $a_C = 1$. Correspondingly, much of the SAMM$_{p,q}$ description of interactions is shifted toward the more efficient treatment at the higher levels of the cluster hierarchy. As a result, accuracy corrections $a_C > 1$ not only serve to ensure a homogeneous accuracy of the SAMM$_4$ electrostatics description but additionally entail substantial speedups.

6.1.4. Also Large IAC Thresholds Θ Enhance the Efficiency. Because the radius $d_X(\Theta, a_X)$ of the spheres $S(a_X, \Theta)$ depends in the same way on the IAC threshold Θ as on a_X, the above arguments analogously apply to Θ. Choosing the IAC threshold Θ_f as the reference (cf. Table 1), the enclosed atom numbers are $N(a_X, \Theta) = N(a_X, \Theta_f)(\Theta_f/\Theta)^3$. For the two more accurate choices $\Theta_\chi < \Theta_f$, $\chi \in \{m, a\}$, one gets atom numbers $N(a_X, \Theta_\chi)$, which are larger by the factors 1.95 (m) and 3.18 (a) than $N(a_X, \Theta_f)$. For large systems, these efficiency reductions are repeated at the higher hierarchy levels.

6.2. Evaluation of SAMM$^\chi_{p,q}$/RF Accuracy by MD Simulations. It will now be interesting to see to what extent the decreasing accuracy of SAMM$^\chi_{p,q}$, which is caused by an increasing IAC threshold Θ_χ, affects macroscopic properties observable in MD simulations. But before we consider this fine point of SAMM$^\chi_{p,q}$, we first want to highlight the progress achieved by including the SAMM$_q$ dispersion expansion into the computation of the long-range interactions.

For these and related purposes, we use the comparative MD simulations on the two liquid phase simulation systems $\mathcal{T}$ (TIP3P water) and $\mathcal{M}$ (methanol) described in section 5.2. The parameters P and P_{ref} of these comparative simulations are listed in Table 2. The simulations yield heating rate differences $\Delta\dot{Q}(\Theta, P, P_{ref})$ as defined by eq 41, which represent our main observables and selectively identify the various algorithmic sources of noise.

6.2.1. SAMM$_q$ Suppresses Dispersion Cutoff Cooling. Without the approximate inclusion of the long-range dispersion by SAMM$_q$ one would have to apply a short-range cutoff to the dispersion at the IAC distance $d_X(\Theta)$ defined by eq 39. In combination with interaction list updates, which are regularly repeated after time delays $\tau \gg \Delta t$, the dispersion cutoff is known to cause a cooling of the simulation system.[36] Upon the use of SAMM$_q$, a cutoff at $d_X(\Theta)$ has to be applied solely to the Pauli repulsion. For small $d_X(\Theta)$, the repulsion cutoff is expected to cause some heating

Figure 8 quantifies the cooling and heating, which is caused by the cutoff of the dispersion and of the Pauli repulsion, respectively, in the system $\mathcal{T}$ as a function of the IAC threshold Θ (recall that $d_X(\Theta) \sim 1/\Theta$). The heat transfers are represented by the heating rate differences $\Delta\dot{Q}(\Theta, P, P_{ref})$ per molecule (cf. eq 41), whose parameters P and P_{ref} are specified by the entries "$q = -1$" and "$c_r = d_X$" in Table 2 (see section 5.2 for further explanations).

The solid line in Figure 8 shows that the dispersion cooling rapidly grows for IAC thresholds $\Theta > 0.14$. In the neighborhood of this minimal value, which corresponds to a IAC distance $d_T(0.14) = 9.7$ Å and, hence, to a dispersion cutoff distance used in many MD simulations, the cooling is acceptably small. For larger IAC thresholds Θ, however, which fall into the range $[\Theta_a, \Theta_f]$ of our standard values, the

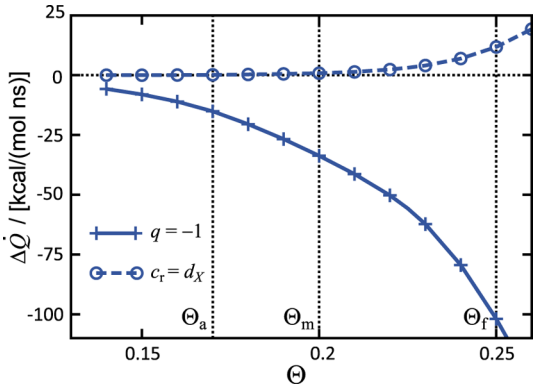


Figure 8. Contributions $\Delta\dot{Q}$ of the dispersion cutoff (solid line) and Pauli repulsion cutoff (dashed line) at $d_T(\Theta)$ to the total heating rate $\dot{Q}$ per molecule in system $\mathcal{T}$ as functions of the IAC threshold Θ. The parameters of the displayed heating rate differences $\Delta\dot{Q}$ are labeled as "$q = -1$" and "$c_r = d_X$," respectively, in Table 2. For explanation, see the text.

dispersion cutoff cooling would be very large. Therefore, the inclusion of the dispersion is mandatory, if one wants to use correspondingly small IAC distances $d_T(\Theta)$.

Note here that section S8 in the SI explains, by a short discussion of temperature control in MD simulations,[63] why we classify algorithmic cooling or heating as "acceptably small," if it has at most a power of ±2 kcal/(mol ns) per degree of freedom, i.e. ± 12 kcal/(mol ns) per TIP3P water and ±28 kcal/(mol ns) per partially stiff MeOH, and "as almost negligible," if it is by more than 1 order of magnitude smaller.

In contrast to the large dispersion cutoff cooling and as demonstrated by the dashed line in Figure 8, the heating caused by the repulsion cutoff is acceptably small over the whole range of IAC thresholds Θ and almost vanishes for $\Theta \leq \Theta_m$. Note that we have obtained quite similar results also for the system $\mathcal{M}$ (data not shown).

Figure 9 serves to show to what extent the approximate inclusion of the dispersion by SAMM$_q$ can repair the dispersion cutoff cooling artifact. The figure has been constructed by evaluating the heating rate differences (eq 41) with the parameters given in Table 2 by the entries "$q = 1$," "$q = 2$," and "$q = 3$." As explained in section 5.2, the heating rate differences

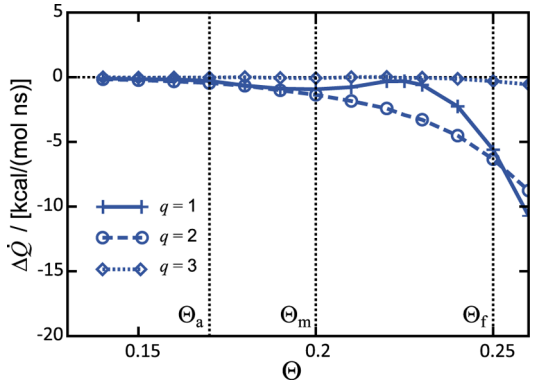


Figure 9. Cooling rates $\Delta\dot{Q}$ remaining in MD simulations of the system $\mathcal{T}$, if the dispersion cutoff is replaced by a SAMM$_q$ dispersion expansion of order order $q = 1$ (solid line), $q = 2$ (dashed line), or $q = 3$ (dotted line) for varying IAC thresholds Θ. For explanation, see the text.

then exclusively represent the contributions to the total heating rate $\dot{Q}$ per molecule in system $\mathcal{T}$, which are caused by the transition from the exact to the approximate SAMM$_q$ description of the dispersion at the IAC distance $d_T(\Theta)$.

A comparison of the dotted line in Figure 9 with the solid line in Figure 8 demonstrates that already the SAMM$_1$ dispersion expansion, which solely includes monopoles for the calculation of the forces, largely repairs the dispersion cutoff artifact. Even at the large IAC threshold Θ_f, the remaining cooling is acceptably small and by a factor of 20 smaller than with the dispersion cutoff (cf. the solid line in Figure 8). At Θ_a the algorithmic cooling is almost negligible for all orders of the SAMM$_q$ dispersion expansion. For the SAMM$_3$ dispersion expansion, the cooling remains almost negligible up to Θ_f (cf. the dotted line in Figure 8). Figure S14 in the SI demonstrates that these arguments also apply to the methanol system $\mathcal{M}$.

These results suggest that one may very well choose the most simple and computationally efficient SAMM$_1$ approximation for the long-range dispersion as a default for large scale simulations. The SAMM$_3$ dispersion expansion can be chosen, if very accurate forces are required like in quantum-classical hybrid simulations (see e.g. ref 30).

6.2.2. Third Order Electrostatics Does Not Suffice. Figure 10 displays the cooling or heating, which is caused by the

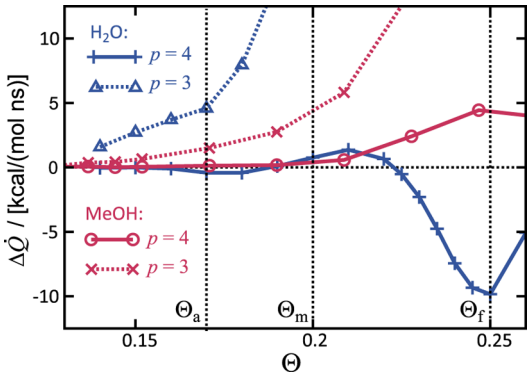


Figure 10. Heating and cooling caused by the SAMM$_p$ electrostatics expansions in the systems $\mathcal{T}$ (blue) and $\mathcal{M}$ (red) as functions of the IAC threshold Θ. The $\Delta\dot{Q}$ data represent those contributions to the total heating rate $\dot{Q}$ per molecule, which are caused by the transition from the exact calculation of the electrostatics to the SAMM$_p$ expansion at the IAC distance $d_T(\Theta)$ for the orders $p = 4$ (solid lines) and $p = 3$ (dotted lines). For explanation, see the text.

switch of the electrostatics description at $d_T(\Theta)$ from exact interatomic Coulomb interactions to approximate intermolecular SAMM$_p$ expansions of order $p = 3$, 4. The associated heating rate differences $\Delta\dot{Q}(\Theta, P, P_{ref})$, whose parameters P and P_{ref} are given by the entries "$p = 3$" and "$p = 4$" in Table 2, are depicted for the systems $\mathcal{T}$ (blue) and $\mathcal{M}$ (red) as functions of the IAC threshold Θ.

Figure 10 demonstrates that the electrostatic algorithmic heating or cooling, which is caused by transition from the exact description to the SAMM$_4$ expansion at $d_X(\Theta)$, is almost negligible for $\Theta \leq \Theta_a$ in both systems $\mathcal{T}$ and $\mathcal{M}$ (solid lines). At this rather small IAC threshold, the algorithmic artifacts of the SAMM$_3$ electrostatics expansion (dotted lines) are sizable but still acceptably small. For $\Theta > \Theta_a$, however, SAMM$_3$ feeds increasing amounts of algorithmic heat into the system. For $\mathcal{T}$ this heating rapidly becomes already intolerable as Θ approaches Θ_m (blue dotted line). Also system $\mathcal{M}$ shows

dx.doi.org/10.1021/ct500319a | J. Chem. Theory Comput. 2014, 10, 3244−3259

such a transition, which occurs, however, at slightly larger values of Θ (red dotted line). Thus, the $SAMM_3$ electrostatics expansion is incompatible with IAC thresholds as large as Θ_f, i.e. with correspondingly short IAC distances $d_X(\Theta_f)$ for the transition from the time-consuming exact to the much more cost-effective SAMM description. The $SAMM_4$ electrostatics approximation, in contrast, features acceptable heating ($\mathcal{M}$) or cooling ($\mathcal{T}$) rates up to Θ_f.

These results suggest to choose $SAMM_4$ as the default for the electrostatics approximation, because it should enable relatively short IAC distances $d_X(\Theta_f)$ at acceptably small heating rates. Combined with the substantial suppression of the dispersion cutoff cooling (Figures 8 and 9) and with the acceptably small repulsion cutoff heating (Figure 8), one expects that the total heating rate of $SAMM_{4,1}$, which includes the repulsion cutoff heating, is still acceptable at Θ_f for the system $\mathcal{T}$.

This expectation is verified by Figure 11. The figure shows the total algorithmic heating rate of $SAMM_{4,1}$ (solid line) in the

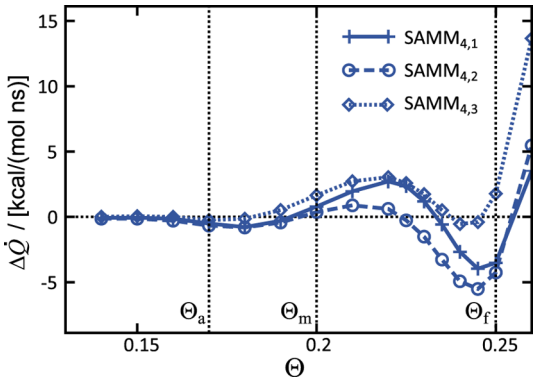


Figure 11. $SAMM_{4,q}$ algorithmic noise for $q = 1$ (solid line), $q = 2$ (dashed line), and $q = 3$ (dotted line) as a function of the IAC threshold Θ measured in the system $\mathcal{T}$ by the heating rate differences $\Delta\dot{Q}$, which are defined by the entries $SAMM_{4,q}$ in Table 2

system $\mathcal{T}$ and solely excludes the algorithmic noise, which is caused by transitions into and out of the RF continuum at d_{MIC}. At Θ_a, the $SAMM_{4,q}$ heating rates are seen to be almost negligible for all three values of q. Therefore, the $SAMM^a_{4,q}$ algorithms are rightfully called "accurate".

At Θ_f, the remaining cooling rate is about −3.5 kcal/(mol ns) for $SAMM_{4,1}$ and, hence, acceptably small. Here, $SAMM_{4,3}$ is more accurate and features a small and almost negligible heating rate up to Θ_f. As one can see by comparing the $SAMM_4$ electrostatics cooling apparent in Figure 10 (blue solid line) with the absence of any significant $SAMM_3$ dispersion cooling documented by Figure 9 (dotted line) and with the Pauli repulsion cutoff heating shown in Figure 9 (dashed line), the almost negligible heating rate of $SAMM_{4,3}$ at Θ_f is caused by a cancellation of the electrostatic cooling with the Pauli repulsion heating. Carrying out the same comparison at Θ_a demonstrates that here all individual algorithmic heating rates are negligibly small.

The $SAMM_{4,q}$ descriptions of the methanol system $\mathcal{M}$, which are presented and discussed in section S7 of the SI, are slightly different concerning certain details but overall lead to the same conclusions. The conclusions are that the most efficient combination $SAMM_{4,1}$ exhibits acceptably small algorithmic noise even with the large IAC threshold Θ_f and an almost negligibly small noise with Θ_a. If an even lower noise level is desired, $SAMM_{4,3}$ is a viable alternative.

As mentioned at the bottom of section 5.2, in $SAMM^\chi_{4,1}$/RF simulations top-level clusters may move into or out of the dielectric continuum extending beyond d_{MIC} and, thereby, cause the additional algorithmic heat $\dot{Q}_{RF}$. This heat is independent of Θ and decreases with system size. Although the systems $\mathcal{T}$ and $\mathcal{M}$ are quite small ($d_{MIC} \approx 20$ Å), this additional heat source has powers of only 2.38 kcal/(mol ns) and 1.05 kcal/(mol ns) per molecule, respectively.

All simulations in this work were carried out in the *NVT* ensemble. Therefore, we did not mention the effects of the various approximations on the computation of the pressure. However, for interested readers we have added to the SI with section S10 a short study of the errors of pressure computation resulting from approximations such as the finite distance truncation of the van der Waals forces and the truncation of the FMM expansions for the dispersion and the electrostatics. The associated results essentially corroborate those of the above study on algorithmic noise. Here, a single exception is provided by the fact that the pressure calculation becomes substantially more accurate, if one increases the order of the FMM dispersion expansion from $q = 1$ to $q = 2$. No comparable improvement has been observed for the algorithmic noise.

6.3. Check of Linear Scaling. The arguments in section 6.1.4, which concluded the presentation of the $SAMM^\chi_{p,q}$ fine-tuning, suggested that $SAMM^m_{p,q}$ and $SAMM^a_{p,q}$ should be by factors of 1.95 and 3.18, respectively, slower than $SAMM^f_{p,q}$. With the aim of checking this suggestion together with the linear scaling, which is expected for $SAMM^\chi_{4,1}$/RF, we have carried out the test simulations characterized in section 5.3. These simulations were executed for 31 liquid water systems $\mathcal{G}_i$ of increasing size using either the nonpolarizable TIP3P[54] ($\mathcal{G}_i = \mathcal{T}_i$) or the complex polarizable six-point potential[8] TL6P ($\mathcal{G}_i = \mathcal{B}_i$).

As reference t_{ref} for the computation times t per step of the dynamics integration, we chose the TIP3P system $\mathcal{T}_{12}$, which contained $N = 26\,211$ atoms, and the $SAMM^a_{4,1}$/RF simulation. Thus, we introduced the dimensionless computing times t/t_{ref} and plotted them as functions of the number N of atoms contained in the systems $\mathcal{G}_i \in \{\mathcal{T}_i, \mathcal{B}_i\}$. The results are shown in Figure 12a and b.

The expected linear scaling of the $SAMM^\chi_{4,1}$/RF computation time t/t_{ref} with increasing size N of the systems (a) $\mathcal{T}_i$ and (b) $\mathcal{B}_i$ is verified by Figure 11 for each of the three accuracy/efficiency choices χ: "a" (red), "m" (blue), and "f" (green). Thus, the linear scaling also applies to complex polarizable force fields.

The data in Figure 12 match the regression lines $t_\chi(N)$ only in an average sense. One clearly recognizes sudden jumps to lower computation times t/t_{ref} at certain transitions from a system $\mathcal{G}_i$ to the next larger system $\mathcal{G}_{i+1}$. One such jump occurs for instance in the green curves belonging to the most efficient computation near $N \approx 22\,500$. With $\log_{10}(22\,500) \approx 4.4$, a comparison with the blue curve in Figure S13 of the SI demonstrates that at this system size the effective height $\tilde{\lambda}(\Theta_f)$ of the interaction hierarchy jumps from 1 to 2, which then enables the inclusion of larger level $l = 2$ clusters into the computation of the large-distance electrostatics and dispersion. The other jumps seen also in the red and blue data have analogous origins.

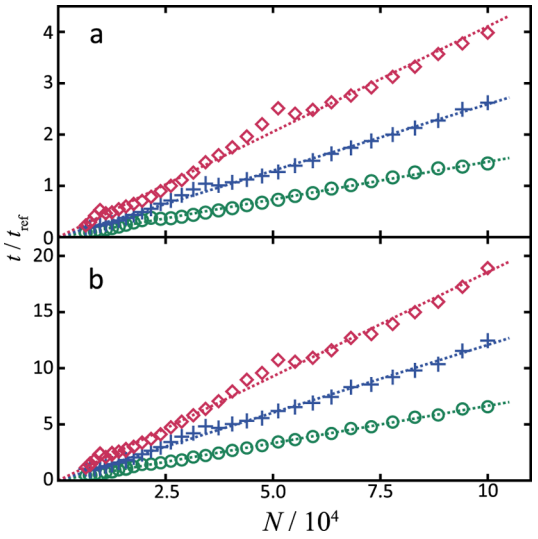


Figure 12. Relative computation times t/t_{ref} measured by applying SAMM$^{\chi}_{4,1}$/RF to (a) the TIP3P systems $\mathcal{T}_i$ and (b) the TL6P systems $\mathcal{B}_i$ with the accuracy/efficiency choices χ = a (red), χ = m (blue), and χ = f (green). Also shown are corresponding regression lines $t_\chi(N) = \rho_\chi N$.

For TL6P, a statistical scatter $\sigma_t(N, \chi)$ of the depicted average computation times occasionally masks these jumps. This scatter $\sigma_t(N, \chi)$ is due to varying numbers of self-consistency iterations during a simulation. Thus, the 30 measured integration times are random variables drawn from a broad distribution, and the depicted average integration times inherit this property. The standard deviation $\sigma_t(N, \chi)$ increases linearly with N, i.e. $\sigma_t(N, \chi) \approx s_\chi N$ with $s_a = 0.45 \times 10^{-5}$, $s_m = 0.30 \times 10^{-5}$, and $s_f = 0.15 \times 10^{-5}$. For a better visibility of the deviations from the linear regressions, the data shown in Figure 12 are replotted in Figure S16 of the SI as (approximately constant) computation times per atom.

The SAMM$^{\chi}_{4,1}$/RF simulations of TIP3P water yield for the relative slopes ρ_χ/ρ_f, $\chi \in \{a, m\}$, of the linear regressions depicted as dotted lines in Figure 12a the values 2.80 (a) and 1.76 (m). For the complex TL6P water models, one gets the almost identical values 2.79 and 1.82, respectively, showing that the more accurate SAMM$^{a}_{4,1}$/RF and SAMM$^{m}_{4,1}$/RF algorithms are by these factors slower than SAMM$^{f}_{4,1}$/RF in simulations of TIP3P and of TL6P water. Hence, the efficiency reduction accompanying the accuracy enhancement does not change with the use of polarizable force fields. Furthermore, the slow-down factors measured here are even slightly smaller than the factors of 3.18 and 1.95 expected from the estimates in section 6.1.4.

6.4. Further Efficiency Comparisons. At this point, the reader might ask how SAMM$^{f}_{4,1}$/RF compares with the predecessor algorithm SAMM$_4$/RF, which employed fixed distance classes[27,28] and a 10 Å dispersion cutoff with its sizable cooling artifact. Employing the simulation system $\mathcal{B}_9$ with its 6503 TL6P models as an example, we found that the use of the IAC criterion (eq 25) combined with Θ_f entails a speedup by a factor of 5. Even the most accurate version SAMM$^{a}_{4,1}$/RF is still by a factor of 1.8 faster than SAMM$_4$/RF. These speedups are the key benefits rendered by the inclusion of the dispersion attraction into our carefully revised SAMM scheme.

As compared to earlier SAMM/RF versions,[27,28] which truncated the electrostatics expansion at $p \lesssim 3$ and, therefore, were plagued by considerable algorithmic noise (cf. Figure 10), the speedups are still factors 3.8 and 1.4 for SAMM$^{f}_{4,1}$/RF and SAMM$^{a}_{4,1}$/RF, respectively. As a result, the now completed redesign of SAMM/RF has eventually enhanced not only the accuracy but also the efficiency of the algorithms.

If one wants to take advantage of the enhanced accuracy generated by the increase of the dispersion expansion order q from 1 to 3, then one has to accept that the computational effort increases by 6−30%. We have measured these slow-downs for the water systems $\mathcal{B}_9$ and $\mathcal{T}_9$, which both comprise 6503 molecules. Here, the small 6% increase relates to TL6P, of course, and the larger 30% value to TIP3P.

Comparing now Figure 12a and b, the average ratio $\langle \rho_\chi(\mathcal{B})/\rho_\chi(\mathcal{T})\rangle_{\chi\in\{a,m,f\}}$ of corresponding slopes is 4.6 with a standard deviation of only 0.1. Thus, SAMM$^{\chi}_{4,1}$/RF is, independently of χ, for the very complex TL6P model only 4.6 times slower than for TIP3P. This is an excellent performance, because one has to compute 4 times more interactions in the innermost interaction shell $[r \le d_\chi(\Theta)]$ for TL6P than for TIP3P. Most of the additional computational effort appears to be caused by the 2 times larger number of force points and only a little by the self-consistency iteration, which involves solely the update of one polarizable degree of freedom per TL6P model at an otherwise static configuration of the system. Note here that the strongly enhanced simulation power of SAMM$^{f}_{4,1}$/RF was a key technical prerequisite for the 20 ns replica exchange simulations on TL6P water,[45] which enabled a sampling of the density−temperature profile with a hitherto unprecedented statistical accuracy.

The computational performance of other polarizable models is much worse. For instance, the data displayed by Table S2 in the SI of ref 64 indicate that the polarizable AMOEBA model is by factors of 20−30 slower than TIP3P. Even the recent iAMOEBA model (which cannot be qualified as polarizable, because it skips the self-consistency iteration) is still by factors of 5.5−7 slower than TIP3P.

If we finally compare the performance of SAMM$^{f}_{4,1}$/RF on TL6P with that of SAMM$^{a}_{4,1}$/RF on TIP3P, we recognize that the most efficient way of simulating TL6P is only 1.62 times slower than an accurate simulation of TIP3P. As a result, we may safely conclude that SAMM$^{\chi}_{4,1}$/RF as implemented in IPHIGENIE is particularly well suited for the simulation of complex polarizable force fields.

7. SUMMARY

We have complemented the pth order Cartesian FMM electrostatics expansion[29,30] SAMM$_p$ (p = 3, 4) by a qth order expansion of the dispersion attraction (q = 1, 2, 3), have designed accuracy corrected IAC thresholds Θ_χ, $\chi \in \{a, m, f\}$, representing different compromises between efficiency and accuracy, and have implemented the thus obtained SAMM$^{\chi}_{p,q}$/RF algorithms for the treatment of long-range interactions into the DFT/(P)MM program package IPHIGENIE.[29,30,44] The algorithms were optimized by studying a series of chemically different dimers of molecules or molecular fragments, which represent building blocks (SUs) of proteins in solution, and several liquid systems $\mathcal{G}$ modeling H_2O and MeOH, which were either described by conventional nonpolarizable energy functions or by the complex and polarizable water model[8] TL6P.

Upon systematically comparing the accuracy by which SAMM$_p$ describes the electrostatic forces acting between dimers of SUs X, we introduced a substance dependence a_X

into the acceptance criterion eq 25, which decides up to what minimal distance $d_\chi(\Theta)$ SUs are still treated by FMM before they are resolved into their constituent atoms. This substance dependence introduces a similarly accurate FMM description for all components of an inhomogeneous simulation system. In this context, the first neglected order (eq 14) of the multipole expansions was shown to reliably describe the distance dependence of the approximation errors.

The inclusion of the dispersion into the SAMM expansions was demonstrated to remove the algorithmic cooling artifact, which is caused by the usual short-range cutoff ($\approx$ 10 Å) of these interactions. Furthermore, it was shown to enable a transition from the exact interatomic calculation to a reasonably accurate FMM treatment of the long-range interactions at IAC distances $d(\Theta_f)$, which may become as short as 5.4 Å in the case of H_2O. Such a short distance is particularly important for the efficient treatment of very complex and polarizable molecular models (like TL6P), because beyond this distance the complexity difference vanishes. Fortunately, the heating artifact, which is caused by the associated 5.4 Å repulsion cutoff, turned out to be still sufficiently small.

Detailed studies of algorithmic artifacts carried out for the systems $\mathcal{T}$ and $\mathcal{M}$ showed that the expansion orders $p = 4$ for the electrostatics and $q = 1$ for the dispersion represent a nice balance between accuracy and efficiency, which may be fine-tuned by the choice of χ. As compared to the predecessor algorithm $SAMM_4$/RF, the inclusion of the dispersion and of the IAC criterion (eq 25) into the revised $SAMM^{\chi}_{4,1}$/RF algorithms eventually yielded speedups by factors of 1.8 (χ = a) to 5 (χ = f).

The thus established $SAMM^{\chi}_{4,1}$/RF family of MD algorithms showed the expected linear scaling with the number of atoms in the system as was demonstrated for bulk water systems $\mathcal{T}_i$ and $\mathcal{B}_i$ modeled by the nonpolarizable TIP3P[54] and polarizable TL6P[8] potentials, respectively. For a given χ, the computational effort of TL6P turned out to be only by a factor of 4.6 larger than for TIP3P, indicating that IPHIGENIE is a convenient choice for simulating protein−solvent systems modeled by complex polarizable force fields.

Note here that $SAMM^{\chi}_{4,1}$ can also be beneficially used for so-called Hamiltonian dielectric solvent[44] (HADES) MD simulations of proteins, in which the solvent is replaced by a dielectric continuum and the Poisson equation is speedily solved by a novel reaction field (RF) approach[65] during the integration of the atomic motion. The reason is that HADES is an integral part of IPHIGENIE and that the underlying RF approach has the form of an antipolarizable force field closely resembling, e.g., the polarizable force field of TL6P. Due to the use of $SAMM^{\chi}_{4,1}$, the computational effort of this new continuum method should scale linearly with the number of protein atoms.

■ ASSOCIATED CONTENT

Supporting Information

Explicit expressions for the components of the tensors $\partial_{(n)}(1/r^6)$ and $\mathbf{M}^{m,0}$ and for dispersion multipole potentials $\phi^{m,D}(\mathbf{c})$ for $m,n \leq 3$. The clustering algorithm is presented in detail, and the underlying structural units are listed and characterized. A thorough explanation and motivation for the otherwise magic distance (26) determining the MIC compliance (28) of level l is given. Empirical approximation errors $\Delta f^{(4)}_{\chi,e}(\tilde{r}_\chi)$ at various distances $\tilde{r}_\chi$ are listed for H_2O and MeOH. The effects of algorithmic cooling and heating observed in simulations of liquid MeOH are presented and discussed. The control of algorithmic cooling and heating artifacts is analyzed. The linear scaling of $SAMM^{\chi}_{4,1}$/RF is redrawn at an enhanced graphical resolution. The pressure effects of the short-range truncation of the van der Waals forces and of the finite order truncations of the FMM dispersion and electrostatics expansions are studied. This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: tavan@physik.uni-muenchen.de.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (SFB749/C). We thank S. Bauer and M. Schwörer for their critical reading of the manuscript.

■ REFERENCES

(1) MacKerell, A. D. *J. Comput. Chem.* **2004**, *25*, 1584−1604.
(2) Tavan, P.; Carstens, H.; Mathias, G. In *Protein Folding Handbook*; Buchner, J., Kiefhaber, T., Eds.; Wiley-VCH: Weinheim, Germany, 2005; Vol. *1*, pp 1170−1195.
(3) van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chem., Int. Ed.* **2006**, *45*, 4064−4092.
(4) Cisneros, G. A.; Karttunen, M.; Ren, P.; Sagui, C. *Chem. Rev.* **2014**, *114*, 779−814.
(5) MacKerell, A. D.; Bashford, D.; Bellott, M.; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; Karplus, M. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.
(6) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(7) Oostenbrink, C.; Villa, A.; Mark, A.; Van Gunsteren, W. *J. Comput. Chem. B* **2004**, *25*, 1656−1676.
(8) Tröster, P.; Lorenzen, K.; Tavan, P. *J. Phys. Chem. B* **2014**, *118*, 1589−1602.
(9) Kaminski, G. A.; Stern, H. A.; Berne, B. J.; Friesner, R. A.; Cao, Y. X.; Murphy, R. B.; Zhou, R.; Halgren, T. A. *J. Comput. Chem.* **2002**, *23*, 1515−1531.
(10) Harder, E.; Kim, B.; Friesner, R. A.; Berne, B. J. *J. Chem. Theory Comput.* **2005**, *1*, 169−180.
(11) Wang, Z.-X.; Zhang, W.; Wu, C.; Lei, H.; Cieplak, P.; Duan, Y. *J. Comput. Chem.* **2006**, *27*, 781−790.
(12) Baker, C. M.; Anisimov, V. M.; MacKerell, A. D. *J. Phys. Chem. B* **2011**, *115*, 580−596.
(13) Darden, T. A.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.
(14) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.
(15) Luty, B. A.; Tironi, I. G.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *103*, 3014−3021.
(16) Skeel, R. D.; Tezcan, I.; Hardy, D. J. *J. Comput. Chem.* **2002**, *23*, 673−684.
(17) Barnes, J.; Hut, P. *Nature* **1986**, *324*, 446−449.
(18) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325−348.

(19) Ding, H.-Q.; Karasawa, N.; Goddard, W. A., III *J. Chem. Phys.* **1992**, *97*, 4309−4315.
(20) Figueirido, F.; Levy, R. M.; Zhuo, R.; Berne, B. J. *J. Chem. Phys.* **1997**, *106*, 9835−9849.
(21) Lim, K.-T.; Brunett, S.; Iotov, M.; McClurg, R. B.; Vaidehi, N.; Dasgupta, S.; Taylor, S.; Goddard, W. A. *J. Comput. Chem.* **1997**, *18*, 501−521.
(22) Takahashi, K. Z.; Narumi, T.; Yasuoka, K. *J. Chem. Phys.* **2011**, *135*, 174108.
(23) Takahashi, K. Z.; Narumi, T.; Suh, D.; Yasuoka, K. *J. Chem. Theory Comput.* **2012**, *8*, 4503−4516.
(24) Andoh, Y.; Yoshii, N.; Fujimoto, K.; Mizutani, K.; Kojima, H.; Yamada, A.; Okazaki, S.; Kawaguchi, K.; Nagao, H.; Iwahashi, K.; Mizutani, F.; Minami, K.; Ichikawa, S.-i.; Komatsu, H.; Ishizuki, S.; Takeda, Y.; Fukushima, M. *J. Chem. Theory Comput.* **2013**, *9*, 3201−3209.
(25) Niedermeier, C.; Tavan, P. *J. Chem. Phys.* **1994**, *101*, 734−748.
(26) Niedermeier, C.; Tavan, P. *Mol. Simul.* **1996**, *17*, 57−66.
(27) Eichinger, M.; Grubmüller, H.; Heller, H.; Tavan, P. *J. Comput. Chem.* **1997**, *18*, 1729−1749.
(28) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847−10860.
(29) Lorenzen, K.; Schwörer, M.; Tröster, P.; Mates, S.; Tavan, P. *J. Chem. Theory Comput.* **2012**, *8*, 3628−3636.
(30) Schwörer, M.; Breitenfeld, B.; Tröster, P.; Lorenzen, K.; Tavan, P.; Mathias, G. *J. Chem. Phys.* **2013**, *138*, 244103.
(31) Warren, M. S.; Salmon, J. K. *Comput. Phys. Commun.* **1995**, *87*, 266−290.
(32) Dehnen, W. *Astrophys. J.* **2000**, *536*, L39−L42.
(33) Dehnen, W. *J. Comput. Phys.* **2002**, *179*, 27−42.
(34) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(35) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. *J. Chem. Theory Comput.* **2013**, *9*, 4046−4063.
(36) Sagui, C.; Darden, T. *Annu. ReV. Biophys. Biomol. Struct.* **1999**, *28*, 155−179.
(37) in't Veld, P. J.; Ismail, A. E.; Grest, G. S. *J. Chem. Phys.* **2007**, *127*, 144711.
(38) Isele-Holder, R. E.; Mitchell, W.; Ismail, A. E. *J. Chem. Phys.* **2012**, *137*, 174107.
(39) Tameling, D.; Springer, P.; Bientinesi, P.; Ismail, A. E. *J. Chem. Phys.* **2014**, *140*, 024105.
(40) Duan, Z.-H.; Krasny, R. *J. Comput. Chem.* **2001**, *22*, 184−195.
(41) Shanker, B.; Huang, H. *J. Computat. Phys.* **2007**, *226*, 732−753.
(42) Wu, X.; Brooks, B. R. *J. Chem. Phys.* **2005**, *122*, 044107.
(43) Allen, M. P.; Tildesley, D. *Computer Simulations of Liquids*; Clarendon: Oxford, 1987.
(44) Bauer, S.; Tavan, P.; Mathias, G. *J. Chem. Phys.* **2014**, *140*, 104103.
(45) Tröster, P.; Tavan, P. *J. Phys. Chem. Lett.* **2014**, *5*, 138−142.
(46) Good, R. J.; Hope, C. J. *J. Chem. Phys.* **1970**, *53*, 540−543.
(47) Peña, M. D.; Pando, C.; Renuncio, J. A. R. *J. Chem. Phys.* **1982**, *76*, 325−332.
(48) If the relative deviation $\delta_{i,j} \equiv |\sigma_i - \sigma_j|/\sigma_i$ of van der Waals diameters σ_i and σ_j is small as is common, e.g., for second row elements, then the relative deviation of the dispersion parameters $B_{i,j}^{\ x}$, calculated by the arithmetic (x = **a**) and geometric (x = **g**) mean, respectively, is to leading order $(3/4)\delta_{i,j}^{\ 2}$. Thus, a 12% deviation of van der Waals diameters translates into a 1% difference of the dispersion parameters $B_{i,j}^{\ a}$ and $B_{i,j}^{\ g}$. Thus **a** can be replaced by **g** without seriously changing the properties of a given force field.
(49) Martinetz, T.; Berkovich, S.; Schulten, K. *IEEE Trans. Neural Networks* **1993**, *4*, 558−569.
(50) Kloppenburg, M.; Tavan, P. *Phys. Rev. E* **1997**, *55*, R2089−R2092.
(51) For a specific parallel computer with 16 CPUs on one main board, we found out, e.g., that simulations of TIP3P water systems with $N_a \geq 1152$ atoms per CPU and with system sizes in the range $6399 \leq N \leq 99\,981$ parallelized with a negligible communication overhead, whereas for the complex polarizable water model TL6P, this overhead was negligible only for $N_a \geq 2304$ atoms per CPU (data not shown).
(52) Kell, G. S. *J. Chem. Eng. Data* **1975**, *20*, 97−105.
(53) Ortega, J. *J. Chem. Eng. Data* **1982**, *27*, 312−317.
(54) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.
(55) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Comput. Chem.* **2001**, *22*, 501−508.
(56) Andersen, H. C. *J. Comput. Phys.* **1983**, *52*, 24−34.
(57) Kaatze, U. *J. Chem. Eng. Data* **1989**, *34*, 371−374.
(58) Sastry, N. V.; Valand, M. K. *J. Chem. Eng. Data* **1998**, *43*, 152−157.
(59) Bussi, G.; Parrinello, M. *Comput. Phys. Commun.* **2008**, *179*, 26−29.
(60) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637−649.
(61) Lennard-Jones, J. E. *Proc. Phys. Soc.* **1931**, *43*, 461−482.
(62) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(63) Lingenheil, M.; Denschlag, R.; Reichold, R.; Tavan, P. *J. Chem. Theory Comput.* **2008**, *4*, 1293−1306.
(64) Wang, L.-P.; Head-Gordon, T.; Ponder, J. W.; Ren, P.; Chodera, J. D.; Eastman, P. K.; Martinez, T. J.; Pande, V. S. *J. Phys. Chem. B* **2013**, *117*, 9956−9972.
(65) Bauer, S.; Mathias, G.; Tavan, P. *J. Chem. Phys.* **2014**, *140*, 104102.