Schnelle Multipolmethoden für die langreichweitigen Wechselwirkungen in molekülmechanischen Molekulardynamik Simulationen

Konstantin Lorenzen

München 2015

Schnelle Multipolmethoden für die langreichweitigen Wechselwirkungen in molekülmechanischen Molekulardynamik Simulationen

Konstantin Lorenzen

Dissertation
an der Fakultät für Physik
der Ludwig-Maximilians-Universität
München

vorgelegt von
Konstantin Lorenzen
aus Lübeck

München, August 2015

Erstgutachter: Prof. Dr. Paul Tavan
Zweitgutachter: Prof. Dr. Martin Zacharias
Tag der mündlichen Prüfung: 9. Oktober 2015

Zusammenfassung

Bei Molekulardynamik (MD) Simulationen von Proteinen in Lösung, welche durch polarisierbare molekularmechanische (PMM) Kraftfelder beschrieben werden, wird der Rechenaufwand durch die Auswertung der langreichweitigen Wechselwirkungen (Elektrostatik, Lennard-Jones) bestimmt. Ihre exakte Darstellung als atomare Paarwechselwirkungen scheidet wegen des quadratisch mit der Größe des Simulationssystems skalierenden Rechanaufwands aus. Deshalb müssen Näherungsalgorithmen verwendet werden, welche immer einen Kompromiss aus Effizienz und Genauigkeit repräsentieren. Höchste Genauigkeitsanforderungen stellen dabei Hybridrechnungen, welche die quantenmechanische (QM) Beschreibung eines kleinen Teils des Simulationssystems mit einem PMM Modell der restlichen Umgebung verbinden und auf die Berechnung der Schwingungsspektren des QM-Fragments zielen.

Aufbauend auf die in der Arbeitsgruppe für molekulare Biophysik am Lehrstuhl für BioMolekulare Optik zu Beginn des Jahrtausends für MD Simulationen entwickelte schnelle "Struktur-Adaptierte Multipol-Methode" (SAMM) und auf deren Kombination mit der QM Dichtefunktionaltheorie (DFT) wird in dieser Arbeit ein neuer SAMM Algorithmus entwickelt, der linear mit der Systemgröße skaliert und die geforderte hohe Genauigkeit und Effizienz bietet.

Die neue Methode bezieht alle langreichweitigen Wechselwirkungen in die SAMM Entwicklungen ein. Bei abgeschlossenen Systemen (molekularen Clustern) erhalten die neuen SAMM Entwicklungen numerisch exakt den Gesamtimpuls, den Gesamtdrehimpuls und, sieht man von sehr kleinen algorithmischen Rauschartefakten ab, die SAMM Energie /1,5,7/. Bei Systemen in periodischen Randbedingungen ist der Gesamtdrehimpuls dagegen keine Erhaltungsgröße, da sie nicht abgeschlossen sind.

Durch den Vorschlag eines genauigkeitsgewichteten Akzeptanzkriteriums, welches diejenigen SAMM Hierarchieebenen selektiert, auf welchen die Wechselwirkungen von Atomclustern beschrieben werden, wird der Kompromiss zwischen Genauigkeit und Effizienz optimiert /5/. Insbesondere sind mit dem neuen SAMM Verfahren PMM Kraftfelder effizient behandelbar /3,4/. Seine Kombination mit DFT Beschreibungen eines Teilsystems ermöglicht DFT/PMM Simulationen, deren Genauigkeit und Effizienz jeweils um mehr als eine Größenordnung gesteigert ist /2,6/. Seine Verbindung mit der von Bauer et al. (*J. Chem. Phys.* **140** 104102, 2014) entwickelten Hamiltonschen Kontinuums-MD Methode HADES macht dieses Verfahren, bei Beibehaltung seines Hamiltonschen Charakters, linear skalierend und damit auf große, in ein Dielektrikum eingebettete Proteine effizient anwendbar.

Die skizzierten Entwicklungen neuer Rechenmethoden werden in der vorliegenden Dissertation mittels dreier publizierter Arbeiten /1,5,7/ vorgestellt. Das zugehörige Computerprogramm steht der wissenschaftlichen Öffentlichkeit über das Internet zur freien Verfügung.

Verzeichnis der im Rahmen dieser Arbeit entstandenen Publikationen

/1/ **Konstantin Lorenzen, Magnus Schwörer, Philipp Tröster, Simon Mates, and Paul Tavan. Optimizing the Accuracy and Efficiency of Fast Hierarchical Multipole Expansions for MD Simulations. *J. Chem. Theory Comput.* 8, 3628-3636 (2012).**

/2/ Magnus Schwörer, Benedikt Breitenfeld, Philipp Tröster, Sebastian Bauer, Konstantin Lorenzen, Paul Tavan, and Gerald Mathias. Coupling DFT to Polarizable Force Fields for Efficient and Accurate Hamiltonian Molecular Dynamics simulations. *J. Chem. Phys.* **138**, 244103 (2013).

/3/ Philipp Tröster, Konstantin Lorenzen, Magnus Schwörer, and Paul Tavan. Polarizable Water Models from Mixed Computational and Empirical Optimization. *J. Phys. Chem. B* **117**, 9486-9500 (2013).

/4/ Philipp Tröster, Konstantin Lorenzen, and Paul Tavan. Polarizable Six-Point Water Models from Computational and Empirical Optimization. *J. Phys. Chem. B* **118**, 1589-1602 (2014).

/5/ **Konstantin Lorenzen, Christoph Wichmann and Paul Tavan. Including the Dispersion Attraction into Structure-Adapted Fast Multipole Expansions for MD Simulations. *J. Chem. Theory Comput.* 10, 3244-3259 (2014).**

/6/ Magnus Schwörer, Konstantin Lorenzen, Gerald Mathias, and Paul Tavan. Utilizing Fast Multipole Expansions for Efficient and Accurate Quantum-Classical Molecular Dynamics Simulations. *J. Chem. Phys.* **142**, 104108 (2015).

/7/ **Konstantin Lorenzen, Gerald Mathias, and Paul Tavan. Linearly Scaling and Almost Hamiltonian Dielectric Continuum Molecular Dynamics Simulations through Fast Multipole Expansions. *J. Chem. Phys.* 143, 184114 (2015).**

Die **fett** hervorgehobenen Arbeiten sind in den Text der Dissertation eingearbeitet und dort nachgedruckt.

Inhaltsverzeichnis

1 Einleitung

Seit mehr als 20 Jahren war es ein zentrales Anliegen der Arbeitsgruppe für theoretische Biophysik am Lehrstuhl für BioMolekulare Optik (BMO) der LMU, die zur genauen Berechnung der Infrarot-(IR-)Spektren von Molekülen in kondensierter Phase erforderlichen theoretischen Methoden zu schaffen.

Hinsichtlich möglicher Anwendungen standen dabei biologische Farbstoffe im Fokus der Aufmerksamkeit, wie beispielsweise die protonierte Schiffsche Base des Retinal [1, 2], diverse Chinone [3, 4] und Flavine [5–7], welche in photosynthetischen Proteinen, wie dem Bakteriorhodopsin [8, 9] oder den Reaktionszentren der bakteriellen Photosynthese [10], oder in Lichtrezeptoren, wie dem Rhodopsin [11] oder den BLUF-Domänen (BLUF: blue light sensors using flavine-adenine-dinucleotide) [12], die biologische Funktion dieser Eiweißstoffe ermöglichen.

Daneben waren auch die IR-Spektren der Amigdgruppen, die das Rückgrat der Polypeptide bilden, von Interesse, weil diese Spektren sowohl über lokale Strukturmotive in Proteinen als auch über deren Faltungs- und Umfaltungsprozesse Auskunft geben können. Diese Systeme wurden etwa seit Beginn des neuen Jahrtausends parallel in der Arbeitsgruppe von Wolfgang Zinth am BMO untersucht. Dabei kamen Mittel der ultraschnellen Laserspektroskopie zum Einsatz, bei denen Umfaltungsprozesse im optischen Spektralbereich ausgelöst und im IR abgetastet wurden. Die avisierten theoretischen Beschreibungen sollten daher auch zum Verständnis dieser Experimente beitragen [13–17].

1.1 Theoretischer Ansatz

Die Berechnung der Schwingungsspektren derartiger Moleküle erfordert die Verwendung quantenmechanischer (QM) Methoden, da molekülmechanische (MM) Kraftfeldansätze nicht die erforderliche Genauigkeit bieten [1, 18]. Andererseits sind hochpräzise QM Verfahren, die beispielsweise durch das Programmpaket GAUSSIAN [19] zur Verfügung gestellt werden und unter Verwendung großer Gaußscher Basissätze die Effekte der Elektronenkorrelation durch Störungstheorie mindestens zweiter Ordnung näherungsweise erfassen können, zwar hinreichend genau, aber für derart große Moleküle, wie Peptide und Chromoproteine, aus Gründen des nicht beherrschbaren Rechenaufwandes ungeeignet. Selbst einzelne Retinalfarbstoffe waren bis vor wenigen Jahren für derartige Ansätze zu groß.

Hier stellte die Entwicklung der Dichtefunktionaltheorie (DFT) [20–22] einen Durchbruch dar, weil sie es gestattete die Schwingungsspektren auch größerer Moleküle, wie etwa der Retinalfarbstoffe, bei hinreichend geringem Rechenaufwand mit überraschender Genauigkeit zu berechnen [3, 23]. Andererseits sind Berechnungen der IR-Spektren isolierter Biofarbstoffe

von nur begrenztem Nutzen für das Verständnis ihrer *in situ* gemessenen Spektren, da die jeweilige Proteinumgebung, wie auch schon ein umgebendes polares Lösungsmittel, zu großen spektralen Veränderungen führt, die nur unzureichend durch vereinfachte Umgebungsmodelle erfasst werden können [1]. Daher war klar geworden, dass genauere, atomar aufgelöste Umgebungsmodelle mit der QM Beschreibung der in kondensierter Phase eingebetteten Moleküle kombiniert werden müssen. Hierfür boten sich MM Kraftfelder wie CHARMM [24], AMBER [25] oder GROMOS [26] an, die zur Beschreibung der Konformationsdynamik von in wässriger Lösung eingebetteten Proteinen entwickelt worden waren.

Anknüpfend an den im Jahre 2013 mit dem Nobelpreis ausgezeichneten Vorschlag [27] von Warshel und Levitt aus dem Jahre 1976, nach dem in einem hybriden Ansatz ein kleiner Teil eines Simulationssystems, etwa ein Chromophor, mit QM Verfahren und seine große Protein/Lösungsmittel-Umgebung mit einem MM Kraftfeld beschrieben werden kann, stellte sich also die Aufgabe, zur *in situ* Beschreibung der IR Spektren von Chromophoren eine neue QM/MM Methode durch Kombination der DFT mit einem geeigneten MM Kraftfeld zu konstruieren.

1.2 Eine neue DFT/MM Hybridmethode

Markus Eichinger hat sich Ende der 90'iger Jahre des letzten Jahrhunderts im Rahmen seiner Doktorarbeit am BMO dieser Herausforderung gestellt und ein auf den genannten Zweck zugeschnittenes DFT/MM Hybridverfahren konstruiert [28]. In technischer Hinsicht hat er dabei das recheneffiziente und parallelisierte DFT-Programm CPMD [29], welches die Kohn-Sham Orbitale des DFT Fragments des Simulationssystems auf einem Gitter in einer Basis ebener Wellen darstellt, mit dem hauseigenen MM-Molekulardynamik (MD) Programm EGO [30] kombiniert, welches für das CHARMM Kraftfeld [24] ausgelegt war.

Das Simulationsprogramm EGO war dabei zwar für reine MM-MD Simulationen im Parallelbetrieb nutzbar, bei DFT/MM Hybridrechnungen konnte das MM Fragment aber nur im sequentiellen Betrieb genutzt werden. Angesichts des viel größeren Rechenaufwands, der seinerzeit für das DFT Fragment bei Hybridrechnungen aufgewendet werden musste, schien diese Vereinfachung des Programmieraufwands jedoch vertretbar. Ferner nutzte EGO mit der Verwendung einer schnellen Multipolmethode {engl. *fast multipole method* (FMM) [31]} ein damals bei MM-MD Programmen sehr ungewöhnliches, aber linear mit der Atomzahl N skalierendes Verfahren zur approximativen Berechnung der langreichweitigen Anteile der elektrostatischen Wechselwirkungen. Dieses Verfahren war als "Struktur-Adaptierte Multipol-Methode" (SAMM) von Christoph Niedermeier am BMO vorentwickelt worden [32, 33] und nutzte Multipolentwicklungen atomarer Ladungsverteilungen bis zur dipolaren Ordnung.

In der zitierten Arbeit [28] konnte Markus Eichinger zeigen, dass die von ihm entworfene DFT/MM-Kopplung, die technisch als Schnittstelle zwischen den Programme EGO und CPMD realisiert war, die Wechselwirkung auch kovalent miteinander verbundener DFT- und MM-Fragmente so gut beschreibt, dass starke Störungen der hohen Qualität der DFT Resultate durch das sehr viel ungenauere MM Umgebungskraftfeld ausgeschlossen werden konnten. Insbesondere zeigte sich, dass die polarisierende Wirkung der starken elektrostatischen

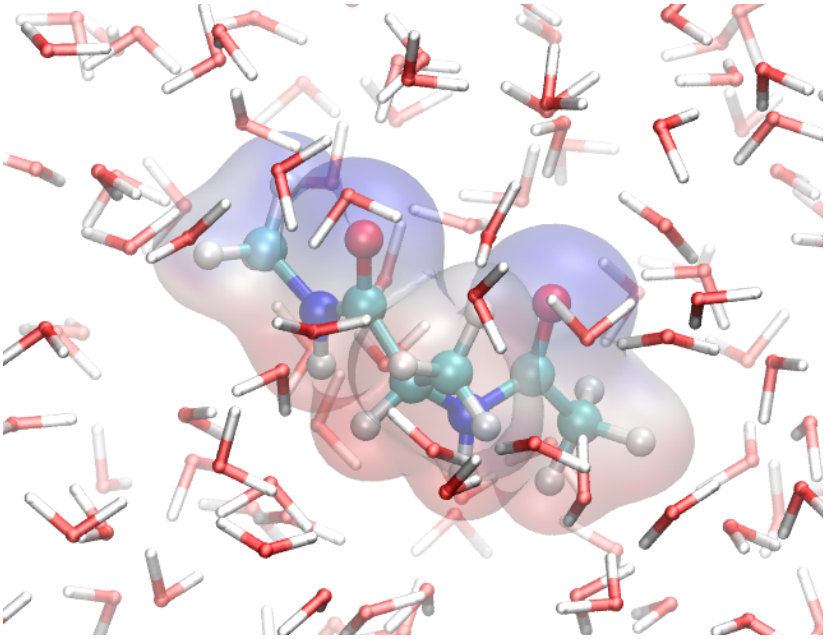


Abbildung 1.1: Typisches Beispiel für DFT/MM Rechnungen. Gezeigt ist ein Alanindipeptid, welches als DFT Fragment in wässriger MM Umgebung gelöst ist. Um das Alanindipeptid ist eine Isofläche der Elektronendichte gezeigt, auf welcher die Farbskala die Stärke des äußeren Potentials kodiert.

Felder, die in kondensierter Phase von polaren und geladenen Molekülen in der Umgebung des DFT Fragments erzeugt werden, anscheinend zutreffend von dem neuen Hybridverfahren erfasst werden können [34]. Entsprechend wurde in der Zusammenfassung der zitierten Publikation zufrieden festgestellt: „*The results demonstrate that our QM/MM hybrid method is especially well suited for the vibrational analysis of molecules in condensed phase*“.

1.3 Verdienste und Probleme der DFT/MM Hybridmethode

Tatsächlich stellte sich in Nachfolgeuntersuchungen an Chinonen in wässriger Lösung [4] und in den Reaktionszentren der Photosynthese [35] sowie an Phosphationen in Wasser [36] heraus, dass beobachtete Effekte der Umgebung auf die Schwingungsspektren der jeweiligen Moleküle quantitativ (Chinone) oder zumindest qualitativ (Phosphate) durch die DFT/MM Simulationsrechnungen beschrieben werden konnten.

Die Abbildung 1.1 zeigt in diesem Zusammenhang ein typisches Beispiel für DFT/MM Rechnungen. In der Abbildung ist ein DFT Fragment (Alanindipeptid) in einer wässrigen Umgebung, die durch ein einfaches MM Dreipunktmodell beschrieben wird, gelöst.

Im Falle der angesprochenen Phosphate in Lösung lieferten dabei verbleibende Abweichungen der Vorhersagen von den beobachteten IR-Spektren den Hinweis, dass dafür Ungenauigkeiten des MM Kraftfeldes für das Umgebungswasser verantwortlich sein könnten. Tat-

sächlich haben spätere, sehr aufwändige und sehr genaue sogenannte "first-principles" Berechnungen der IR Spektren von Phosphaten in Wasser [37] gezeigt, dass die dort erkennbaren Solvatisierungsstrukturen im DFT/MM Hybridszenario nicht vorhanden waren, weil die MM Wassermodelle in der Umgebung der DFT-Phosphate zu schwach gebunden waren. Daher waren die DFT-Phosphate dort zu schwachen elektrischen Umgebungsfeldern ausgesetzt, was die erwähnten Abweichungen zwischen experimentell beobachteten und berechneten IR-Spektren der Phosphate erklärt [37].

Diese und andere [38] Untersuchungen haben also gezeigt, dass IR Spektren von Molekülen in kondensierter Phase sehr empfindlich auf die Details der Modellierung der Umgebungsstrukturen und ihrer polarisierenden elektrostatischen Wirkung reagieren. Entsprechend müssen theoretische Beschreibungen sehr sorgfältig konzipiert werden [39]. Aber selbst dann waren mit der von Eichinger konstruierten DFT/MM Methode häufig nur unter größten Mühen und unter Aufwendung beträchtlicher Ressourcen akzeptable Resultate zu erzielen [15], weil beispielsweise im Falle eines in MM Wasser gelösten DFT β-hairpin-Peptids die für Normalmodenanalysen nötigen Minimierungen des Peptids im Lösungsmittelkäfig große Konvergenzprobleme aufwiesen.

Dieser Befund lieferte den Hinweis, dass die von Eichinger konstruierte DFT/MM Schnittstelle einer gründlichen Überarbeitung im Hinblick auf Recheneffizienz und Genauigkeit der Beschreibung bedurfte. Dabei musste insbesondere das durch die Schnittstelle repräsentierte DFT/MM Modell durch Verwendung von Hellmann-Feynman Kräften zu einer Hamiltonschen Energiefunktion erweitert werden, weil die vom bisher verwendeten Näherungsverfahren tolerierten Verletzungen des Newtonschen Reaktionsprinzips als Ursachen der schlechten Konvergenz identifiziert werden konnten [40]. Ferner musste auch die gemeinsame Parallelisierung der MM und DFT Programmpakete EGO und CPMD in Angriff genommen werden, um eine effiziente Nutzung der Hybridmethode auch auf hochparallelen *high performance computing* (HPC) Systemen zu ermöglichen. Mein Kollege Magnus Schwörer hat in seiner laufenden Doktorarbeit dieses Problems erfolgreich gelöst, wobei er sich auf meine Ergebnisse zur Verbesserung des SAMM Verfahrens [41–43] stützen konnte [44, 45], weshalb ich als Koautor auch zu diesen Arbeiten beigetragen habe.

Einen weiteren grundlegenden Mangel des DFT/MM Hybridansatzes haben die Untersuchungen von Galina Babitzki [46] zu den Schwingungsspektren des Retinalchromophors von Bakteriorhodopsin aufgedeckt. Dieser Mangel besteht, wie von ihr und anderen gezeigt werden konnte [2, 47] in der Vernachlässigung der elektronischen Polarisierbarkeit durch das für die Modellierung der Chromophor-Umgebung verwendete MM Kraftfeld CHARMM [24].

In dieser Modellierung werden nämlich die elektronischen Signaturen der molekularen Bausteine eines Protein-Lösungsmittel Simulationssystems durch statische atomare Partialladungen beschrieben. Tatsächlich herrschen aber in der kondensierten Phase bio-molekularer Systeme aufgrund des Vorkommens von Ionen und der großen Polarität vieler Moleküle und molekularen Bausteinen überall starke lokaler elektrische Felder $\mathbf{E}(\mathbf{r}_i)$, welche zu einer wechselseitigen Polarisation der Ionen, Moleküle und molekularen Bausteine führen. Solche Polarisationseffekte lassen sich bei Verwendung der linearen Antwortnäherung beispielsweise durch induzierte atomare Dipole $\mathbf{p}_i = \alpha_i \mathbf{E}(\mathbf{r}_i)$ modellieren, wobei die Konstanten α_i geeignet zu wählende atomare Polarisierbarkeiten sind. Dieser Zugang wurde in den bisherigen bio-

molekularen MM Standardkraftfeldern [24–26] wegen des erhöhten Rechenaufwands, der mit den selbstkonsistent zu berechnenden induzierten Dipolen $\mathbf{p}_i$ verbunden ist, vermieden. Der im MM-MD Programm EGO zur approximativen Berechnung der langreichweitigen Elektrostatik verwendete FMM-Algorithmus SAMM bietet hier im Prinzip die Chance den Rechenaufwand in Grenzen zu halten.

Die Vernachlässigung der Polarisierbarkeit der MM Umgebung führte beispielsweise schon bei reinen MM-MD Simulationen der Chromophor-Bindungstaschen des Bakteriorhodopsins [47] und der BLUF-Domäne von AppA [7] zu einem Kollaps der experimentell gut charakterisierten Proteinstrukturen. Dies konnte durch ausgedehnte iterative DFT/MM Rechnungen an den Proteinresiduen, welche in der Bindungstasche den jeweiligen Chromophor umgeben, und die Ableitung strukturadaptierter Partialladungen q_i, d.h. durch die Berechnung eines "polarisierten MM Kraftfeldes" gezeigt werden, weil sich der sonst zu beobachtende strukturelle Zerfall der Chromophor-Bindungstaschen erst durch ein derart angepasstes Kraftfeld verhindern ließ.

Selbstverständlich führen inkorrekte Proteinstrukturen, wie sie etwa aus MM-MD Simulationen mit Standardkraftfeldern folgen, auch zu sehr stark von den Beobachtungen abweichenden Vorhersagen der IR Spektren der jeweiligen Chromophore [2, 7]. Darüber hinaus hat sich aber auch bei Einsatz korrekter (d.h. zu hochauflösenden Messungen passender) Proteinstrukturen gezeigt, dass die DFT/MM Beschreibungen der IR Spektren noch besser mit den Beobachtungen übereinstimmen, wenn die Polarisation der Umgebung berücksichtigt wird [2, 7].

Insgesamt haben diese Ergebnisse also demonstriert, dass die große Empfindlichkeit, mit der die IR Spektren von Molekülen in kondensierter Phase auf kleinste Variationen der Umgebungselektrostatik reagieren, den Einsatz polarisierbarer molekülmechanischer (PMM) Kraftfelder bei Hybrid-Berechnungen solcher Spektren erzwingt. Diese Einsicht setzte nun, über die sowieso schon nötige Überarbeitung der DFT/MM Schnittstelle hinaus, ihre Erweiterung zu einer DFT/PMM Kopplung auf die Tagesordnung. Auch diese Aufgabe wurde mittlerweile erfolgreich gelöst [44]. Als Voraussetzung dazu musste das hauseigene MM-MD Programm EGO gründlich überarbeitet und für die Verwendung von PMM Kraftfeldern erweitert werden. Da PMM-MD Simulationen aber eine selbst-konsistente Berechnung der induzierten Dipole $\mathbf{p}_i$ erfordern, die während der simulierten Dynamik glatt variieren sollten, damit man zur Effizienzsteigerung Gedächtniseffekte nutzen kann [48–50], benötigen sie eine erhöhte Genauigkeit der eingesetzten FMM Approximationen.

1.4 Integration von PMM Kraftfeldern in SAMM

Für den Einsatz von PMM Kraftfeldern und den effizienten Betrieb einer DFT/PMM Kopplung musste die ursprüngliche SAMM Methode [30, 32, 33, 51] in vielerlei Hinsicht gründlich überarbeitet werden. So musste zur Steigerung der Recheneffizienz und zum Erzielen einer homogenen Genauigkeit der Approximation das ursprüngliche Entfernungsklassen-Schema aufgegeben und durch ein Wechselwirkungs-Akzeptanz-Kriterium ersetzt werden, welche das

grundlegende Näherungskonzept von FMM Methoden durch Verwendung der scheinbaren Größe von wechselwirkenden Objekten konsequent realisiert [42].

Als Voraussetzung dafür war die Einbeziehung aller nicht-bindenden Wechselwirkungen, d.h. der $\sim 1/r^6$ Dispersions-Attraktion und der $\sim 1/r^{12}$ Lennard-Jones Repulsion in das SAMM Verfahren nötig [42, 43]. Ferner mussten, den FMM Konzepten von Dehnen [52] folgend, die FMM Entwicklungen als zweiseitige Taylor-Entwicklungen gestaltet werden, um die Einhaltung des Newtonschen Reaktionsprinzips bei der Berechnung aller Kraftbeiträge garantieren zu können [41]. Dabei sollte die erzielbare Genauigkeit durch Verwendung von Entwicklungen vierter Ordnung für die Elektrostatik soweit gesteigert werden, dass das im Bereich des DFT Fragments wirkende und von der PMM Umgebung erzeugte elektrostatische Potential den bei DFT/PMM-MD Simulationen für schnelle Konvergenz erhöhten Anforderungen an die Genauigkeit der Berechnung genügt [45]. Schließlich musste, um speziell bei DFT/PMM Hybridrechnungen den Hamiltonschen Charakter der Energiefunktion zu wahren, auch die im SAMM Algorithmus verwendeten FMM Entwicklungen energieerhaltend gestaltet werden [43]. Diesen Herausforderungen habe ich mich während meiner Doktorarbeit gestellt und sie, wie die vorliegende Dissertation hoffentlich zeigen kann, auch erfolgreich bewältigt.

Insgesamt musste also das MM-MD Programm EGO gründlichst überarbeitet werden, was unter der Federführung von Gerald Mathias dankenswerterweise geschah und zu dem fast völlig neu geschriebenen PMM-MD Programm IPHIGENIE [53] führte. Ich habe dabei die Aufgabe der Neugestaltung des SAMM Elektrostatikteils und der zugehörigen Clusterhierarchie (genauere Erläuterungen folgen weiter unten) übernommen [42] und für eine effiziente Parallelisierung der Algorithmen gesorgt.

1.5 Erste Anwendungen des neuen PMM-MD Verfahrens

Zum Testen der in mehreren Stufen erfolgten Entwicklung des neuen PMM-MD Verfahrens benötigte ich ein polarisierbares Kraftfeldmodell, wobei die Wahl auf das wichtigste biologische Lösungsmittel, das Wasser, fiel.

In einer parallel zu meinen methodischen Entwicklungen laufenden Doktorarbeit hat sich mein Kollege Philipp Tröster dieser Aufgabe angenommen und mehrere PMM Modellpotentiale zunehmender Komplexität entwickelt [54, 55].

Diese Modelle waren durch einen Gaußschen induzierbaren Dipol am Sauerstoffatom des Wassermoleküls, positiven Partialladungen an Wasserstoffatomen und ein bis drei negativen Partialladungen an masselosen Orten in der Nähe des Sauerstoffatoms charakterisiert. Bei Verwendung des von mir überarbeiteten SAMM Verfahrens übersteigt der Rechenaufwand auch für das komplexeste Wassermodell den eines einfachen nicht-polarisierbaren Modells, wie etwa des bekannten sogenannten Dreipunktmodells TIP3P [56], um höchstens den Faktor fünf [42].

In jüngster Zeit wurden, anknüpfend an die Trösterschen Entwicklungen, noch weiter verbesserte Wassermodelle entwickelt, welche die punktförmigen statischen Partialladungen durch

Gaußsche Ladungsverteilungen ersetzen und speziell auf den Einsatz in DFT/PMM Hybridsimulationen ausgelegt sind. Erste Anwendungen sind hier DFT/PMM Hybridsimulationen einfacher Modelle für Amidgruppen, wie des Monomers N-Methyl-Acetamid und des Dimers Ac-Ala-NHMe in wässriger Lösung, welche auf die Berechnung der Schwingungsspektren dieser Modelle zielen (laufende Arbeiten von C. Wichmann und M. Schwörer). Diese Rechnungen dienen ferner der Entwicklung eines neuartigen "spektroskopischen" polarisierbaren Kraftfelds für das Rückgrat von Polypeptiden, das entsprechende Vorarbeiten [38, 57] substanziell erweitern soll (C. Wichmann, laufende Doktorarbeit).

Während meiner Doktorarbeit zeigte sich zudem, dass die zeitgleich durch Sebastian Bauer entwickelte *Hamiltonian Dielectric Solvent* (HADES) Kontinuumsmethode [48, 58] für hocheffiziente, Hamiltonsche und physikalisch fundierte — da auf der Lösung der Poisson-Gleichung beruhende — Simulation von Proteinen in implizitem Lösungsmittel algorithmisch eng verwandt mit der Berechnung polarisierbarer Kraftfelder ist. Dies ist der Fall, da in HADES die Polarisation des Kontinums über eine selbstkonsistent zu berechnende Antipolarisierbarkeit der Proteinatome ausgedrückt wird.

Der Kopplung der HADES Methode mit dem SAMM Algorithmus kam zugute, dass ich, zunächst für die DFT/PMM Kopplung, eine Hamiltonsche SAMM Variante entwickelt hatte. In Kombination mit HADES erhält diese SAMM Variante den Hamiltonschen Charakter der Kontinuumsmethode und ermöglicht gleichzeitig einen lediglich linear mit der Atomzahl N steigenden Rechenaufwand. Wie es sich für ein isoliertes Hamiltonsches System gehört sind sowohl der gesamte Drehimpuls als auch der gesamte lineare Impuls bei HADES/SAMM-MD Simulationen Erhaltungsgrößen [43].

1.6 Überblick über die Präsentation des Materials

Die Ergebnisse, die ich bei der Überarbeitung des SAMM Verfahrens erzielt und als Erstautor publiziert habe [41–43], sind in den Unterkapiteln 2.1-2.3 dieser Arbeit nachgedruckt. Diese Kapitel enthalten, neben den publizierten Arbeiten [41–43], das jeweils zugehörige, on-line publizierte und recht umfangreiche unterstützende Material. Ferner skizziert Kapitel 3 meine Beiträge zu vier weiteren Publikationen [44, 45, 54, 55], an denen ich als Koautor mitgearbeitet habe.

Um auch Lesern, die nicht zutiefst mit den Techniken und Problemen der Simulation biomolekularer Systeme vertraut sind, den Zugang zu meiner Arbeit zu erleichtern habe ich im nun folgenden Rest der Einleitung zu meiner Arbeit einige, wie ich hoffe, zum Verständnis nützliche Fakten und Konzepte skizziert. Ferner gibt Abschnitt 1.15 einen etwas detaillierteren Überblick über die im Kapitel 2 nachgedruckten Publikationen.

1.7 Wichtige Eigenschaften bio-molekularer Systeme

Beginnen wir mit einem Überblick über zentrale Eigenschaften von Proteinen, um zu erklären, welche physikalischen Effekte der Strukturdynamik von Proteien in wässiger Lösung

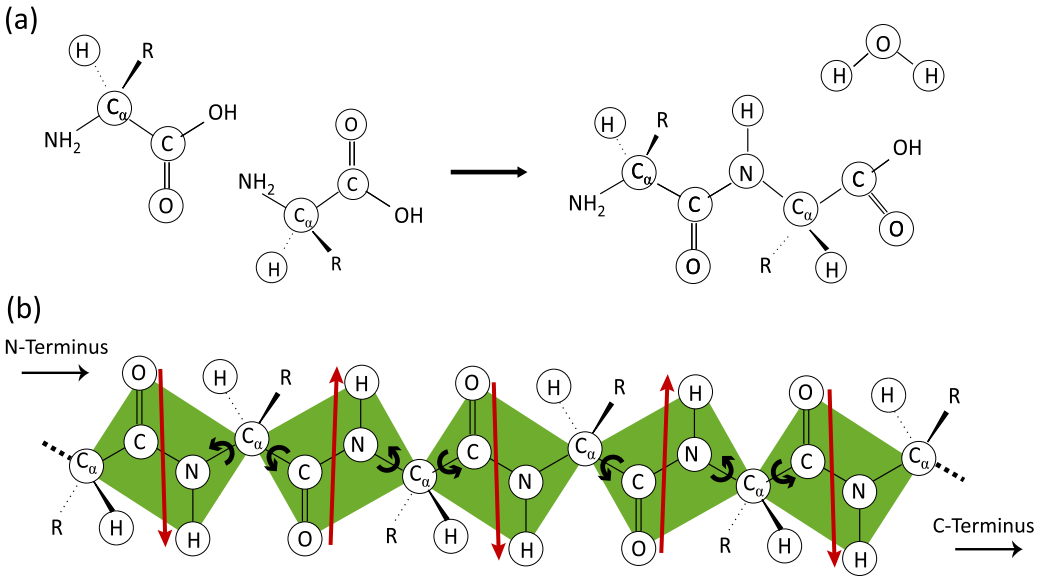


Abbildung 1.2: Die obere Teilabbildung (a) zeigt die Bildung einer Peptidbindung zwischen zwei α-Aminosäuren unter Wasserabscheidung. In der unteren Abbildung (b) ist ein Proteinrückgrat gezeigt, das durch Wiederholung des obigen Mechanismus gebildet wurde. Die grünen Flachen symbolisieren hier die planaren und rigiden Amidgruppen, der sog. Peptidplättchen, welche ausgeprägte Dipole (rote Pfeile) bilden.

zu Grunde liegen und warum gerade die Berücksichtigung von Polarisationseffekten dafür essenziell ist.

Proteine werden in den Ribosomen durch Ablesen der in den m-RNA Molekülen kodierten Primärsequenz als Kettenmoleküle durch Polymerisation aus 20 kanonischen α-Aminosäuren synthetisiert [59, 60]. Posttranslatorisch können gelegentlich noch einzelne dieser Aminosäuren auf biochemischem Wege modifiziert werden. Das menschliche Genom beinhaltet beispielsweise die Baupläne von insgesamt ca. 20000-25000 verschiedenen Proteinen [61].

Abbildung 1.2 zeigt in der oberen Teilabbildung (a) exemplarisch die Bildung einer Peptidbindung zwischen zwei α-Aminosäuren. In jeder Aminosäure folgt auf ein Carboxy-Kohlenstoffatom C ein weiteres Kohlenstoffatom C_α, an das eine Aminogruppe und ein charakteristischer Rest gebunden sind. Dieser Rest spezifiziert dabei die unterschiedlichen Aminosäuren [59].

Ein Charakteristikum der so gebildeten Ketten sind die planaren Amidgruppen, d.h. das C_α Atom und die C=O Gruppe der vorderen und die N-H Gruppe sowie das C_α Atom der hinteren Gruppe formen fast rigide planare Einheiten [62]. Dazu stellt die Abbildung 1.2 im unteren Bildabschnitt (b) einen Ausschnitt eines so gebildeten Polypeptids dar. Aufgrund der Rigidität der Amidgruppen (grüne Flächen) sind die maßgeblichen Freiheitsgrade eines Proteinrückgrats die Dihedralwinkel ϕ und ψ der Peptidbindungen. Die roten Pfeile in der Abbildung stellen die Dipolmomente dar, welche durch die Ladungsverteilung in den Amidgruppen hervorgerufen werden.

Die Amidgruppen sind stark polarisierbar, d.h. in äußeren elektrischen Feldern, die in der

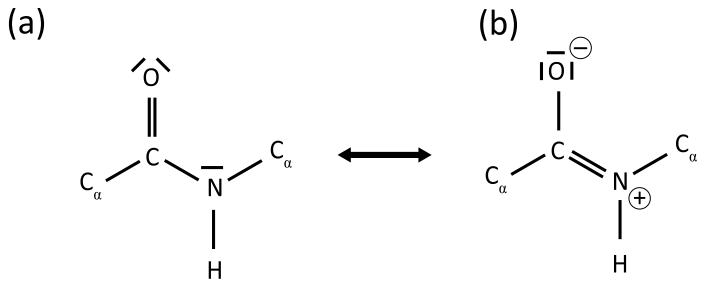


Abbildung 1.3: Gezeigt sind die zwei π-Resonanzstrukturen einer Amidgruppe (vgl. Abbildung 1.2). Die linke Teilabbildung (a) zeigt die neutrale, die rechte Abbildung (b) die zwitterionische Struktur. Bild nach [62], siehe dort für mehr Information.

kondensierten Phase anliegen, werden diese Dipolmomente in der Regel verstärkt [62]. Abbildung 1.3 zeigt in diesem Zusammenhang die π-Elektronen Resonanzstrukturen, aus welchen das starke Dipolmoment, die große Polarisierbarkeit und die Planarität der Peptidgruppen erklärt werden können [62].

Diese starken Dipolmomente der Amidgruppen führen durch attraktive Dipol-Dipol Wechselwirkungen zu der Ausbildung gewisser häufig anzutreffender Anordnungsmuster, sogenannter Sekundärstrukturmotive, in Teilabschnitten von Proteinen, welche mit charakteristischen räumlichen Anordnungen der Peptidplättchen einhergehen. Von größter Bedeutung sind hier die α-Helix und das β-Faltblatt [62]. Die Stärke der zugrundeliegenden Dipolwechselwirkungen ist aufgrund der Polarisierbarkeit der Peptidgruppen in α-Helix- und β-Faltblattstrukturen unterschiedlich stark ausgeprägt. Aus dem Blickwinkel der (P)MM/MD-Simulationen ist dieser Effekt nicht zu vernachlässigen und dafür verantwortlich zu machen, dass nicht-polarisierbare MM Kraftfelder zumeist entweder α-helikale oder β-Faltblattstrukturen ausreichend gut vorhersagen können aber nicht beide gleichzeitig [62].

Da die Wechselwirkungen zwischen dem Protein und dem umgebenden wässrigen Medium aufgrund des hydrophoben Effekts die Grobstruktur des jeweiligen Proteins bestimmen [59, 62], ist es wichtig das Lösungsmittel gut zu modellieren. Wassermoleküle sind klein, stark polar, verfügen über ein ausgeprägtes Quadrupolmoment und sind durch äußere Felder stark polarisierbar. Zum Beispiel erhöht sich das Dipolmoment eines Wassermoleküls beim Übergang von der Gasphase in die flüssige Phase von ca. 1.85 auf ca. 2.5 Debye [28]. Aufgrund des großen Dipolmonents des Wassers in der flüssigen Phase hat diese bei Normalbedingungen die sehr große dielektrische Konstante $\epsilon \approx 78$, die den Haupteffekt zur Solvatisierung von Proteinen beisteuert. In welchem Maße auch spezifische Bindungen von einzelnen Wassermolekülen an Proteinoberflächen für die Eigenschaften dieser Moleküle wichtig sind, ist bislang nicht im Einzelnen geklärt.

Aufschluss könnten hier Vergleiche der MD Simulationen von Proteinen in explizit und atomar beschriebenem PMM Wasser mit entsprechenden Simulationen dieser Moleküle im dielektrischen Kontiunuum liefern. Die neue HADES Methode [48, 58, 63] könnte dazu ebenso beitragen wie die neuentwickelten PMM Wassermodelle [54, 55], falls die Proteine im jeweiligen Lösungsmittel effizient und mit guter Statistik beschrieben werden können. Vorausset-

zung dafür sind linear skalierende Simulationsalgorithmen wie das von mir in dieser Arbeit entwickelte neue SAMM Verfahren.

1.8 Molekulardynamik

Computersimulationen von Proteinen in Lösung werden seit etwa 40 Jahren zur Beschreibung der Struktur und Dynamik dieser Makromoleküle eingesetzt [62, 64, 65]. Dabei wird zumeist das Verfahren der MD Simulation verwendet, welches auf der Born-Oppenheimer Näherung [66] zur Trennung der Kern- und Elektronenbewegung beruht und die Atome i als klassische, an den Kernorten $\mathbf{r}_i$ lokalisierte Massenpunkte darstellt. Die Wechselwirkungen der Atome werden dabei durch das von den Elektronen erzeugte effektive Potential $E(\mathbf{R})$ der Kernbewegung vermittelt, wobei $\mathbf{R} \equiv (\mathbf{r}_1, \ldots, \mathbf{r}_N)^T \in \mathbb{R}^{3N}$ die Konfiguration eines Simulationssystems mit N Atomen bezeichnet

Die zur Bestimmung des effektiven Potentials $E(\mathbf{R})$ nötige Lösung der elektronischen Schrödingergleichung ist für Simulationssysteme mit mehr als einigen hundert Atomen, trotz der Entwicklung recheneffizienter DFT-Methoden [20–22, 29], zu rechenaufwändig. Deshalb wird $E(\mathbf{R})$ oft durch ein analytisches MM Kraftfeld ersetzt [62, 67], welches, wie oben dargestellt wurde, Effekte der elektronischen Polarisation vernachlässigt. Erst in jüngerer Zeit wurde damit begonnen auch PMM Kraftfelder für Proteine zu entwickeln [68–70]. Ein solches Kraftfeld beschreibt die zwischen Atomen wirkenden Kräfte aus kurzreichweitigen chemischen Bindungen sowie aus langreichweitigen elektrostatischen und van der Waals Wechselwirkungen durch geeignet parametrisierte Modellpotentiale. Durch Bildung der negativen Gradienten

$$\mathbf{F}_i = -\nabla_i E(\mathbf{R}) \tag{1.1}$$

für die Koordinaten $\mathbf{r}_i$ der Atome erhält man die auf diese Atome in der Systemkonfiguration $\mathbf{R}(t)$ zum Zeitpunkt t wirkenden Kräfte, sodass die zugehörigen Newtonschen Bewegungsgleichungen numerisch, z.B. vermittels des velocity Verlet Algorithmus [71], mit einem vorgegebenen Zeitschritt Δt integriert werden können, wenn die atomaren Geschwindigkeiten $\dot{\mathbf{R}} \equiv (\dot{\mathbf{r}}_1, \ldots, \dot{\mathbf{r}}_N)^T$ in einer Anfangskonfiguration (Zeitpunkt $t = 0$) bekannt sind. Der finite Zeitschritt Δt, der die Zeitspanne zwischen zwei aufeinander folgenden Zeitpunkten der numerischen Integration angibt, muss klein genug gewählt sein, damit die Kerne in der sich ergebenden diskreten Trajektorie $\mathbf{R}(n\Delta t)$, $n = 1, 2, \ldots, \tau$, auch die schnellsten Schwingungen des Systems ausreichend abbilden. Δt wird üblicherweise in der Größenordnung einer Femtosekunde gewählt.

Die Anfänge der MD Simulationen gehen auf die Simulationen von harten Kugeln durch Alder und Wainwright [72] im Jahr 1957 und die Simulation von flüssigem Argon, das durch weiche Lennard-Jones Kugeln beschrieben wurde, durch Rahman [73] im Jahre 1964 zurück. Eine erste bio-physikalische Anwendung der Technik der MM/MD Simulationen an einem komplexen Protein stellt die Simulation des Trypsin-Inhibitors BPTI durch McCammon, Gelin und Karplus [74] von 1977 dar. Ihre Simulationen, welche noch ohne Lösungsmittelumgebung durchgeführt wurden, konnten zeigen, dass Proteine keineswegs starr sind, sondern sich teilweise eher wie Flüssigkeiten verhalten.

Aufbauend auf diesen Anfängen wurden die MM und PMM Modelle immer weiter verfeinert. Heutige MM Standardkraftfelder, wie z.B. CHARMM22 [24], AMBER95 [25] oder GROMOS [26], ähneln sich stark in ihrer funktionelle Form und bringen spezifische Parametrisierungen für bestimmte Anwendungsbereiche ein. Zur Illustration sollen nun einige wichtige Charakteristiken solcher Modelle beschrieben werden.

1.9 MM Kraftfelder für Proteine in Lösung

Standard MM Kraftfelder enthalten zum einen Terme langreichweitiger nicht-gebundener (nb) und kurzreichweitiger gebundener (b) Wechselwirkungen

$$E(\mathbf{R}) = E_{\text{nb}}(\mathbf{R}) + E_{\text{b}}(\mathbf{R}). \tag{1.2}$$

Hierbei sind die langreichweitigen Terme, welche die Form von Paarwechselwirkungen zwischen den Atomen des molekularen Systems annehmen, als Summen von elektrostatischen und van der Waals Wechselwirkungen gegeben. Sie werden durch Coulomb- und Lennard-Jones-Paarpotentiale beschrieben als

$$E_{\text{nb}}(\mathbf{R}) = \sum_i \sum_{j \le i} f_{\text{near}}(i,j) \left(\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right), \tag{1.3}$$

wobei die Doppelsumme über alle Atompaare (i, j) läuft. Der Faktor $f_{\text{near}}(i, j)$ beschreibt die kraftfeldspezifische Handhabung der langreichweitige Wechselwirkungen für chemisch gebundene Nachbaratome. Üblicherweise ist $f_{\text{near}}(i, j)$ für solche Nachbarn und für Nachbarn von Nachbarn gleich 0 und sonst gleich 1.

Der erste Term in Gl. (1.3) beschreibt die elektrostatische Wechselwirkung zwischen den Partialladungen q_i und q_j der Atome i und j, welche an den Orten $\mathbf{r}_i$ und $\mathbf{r}_j$ vorzufinden sind und somit zueinander den Abstand $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ haben. Diese Partialladungen q_i bilden (im Rahmen der jeweiligen Parametrisierungansätze) näherungsweise die Differenz zwischen der positiven Kernladung des betreffenden Atoms und der negativen Ladung der Elektronendichte in der Umgebung des Atoms ab. Somit beschreiben die Partialladungen in erster Linie Elektronegativitätsdifferenzen von Atomen in den betrachteten molekularen Verbünden. Daneben beschreiben sie auch Effekte der elektronischen Polarisation im mittleren äußeren Feld. Da die elektrostatischen Wechselwirkungen für alle Atompaare ausgewertet werden müssen, skaliert der Rechenaufwand quadratisch ($\sim N^2$) mit der Atomzahl N, wenn er nicht durch geeignete Näherungsalgorithmen reduziert wird.

Der zweite Term in Gl. (1.3) nähert die über A_{ij} parametrisierte repulsive Pauli-Abstoßung zwischen den Elektronenhüllen chemisch nicht gebundener Atome. Der dritte Term erfasst vermittels der Parameter B_{ij} die attraktive Dispersionswechselwirkung, welche auf Korrelationseffekten der Elektronen beruht. Das relativ schnelle Abfallen der $\sim r_{ij}^{-6}$ Dispersionswechselwirkung und des $\sim r_{ij}^{-12}$ Repulsionspotentials ermöglicht es, die Auswertung dieser Paarinteraktionen, im Gegensatz zu den elektrostatischen Wechselwirkungen, in MM/MD-Simulationen für Abstände $|\mathbf{r}_{ij}| \gtrsim 10$ Å in akzeptabel guter Näherung zu vernachlässigen.

Dann skaliert der für die hiermit spezifizierte (12-6) Lennard-Jones Wechselwirkung nötige Rechenaufwand nur linear mit N.

Neben den langreichweitigen Wechselwirkungen $E_{\text{nb}}(\mathbf{R})$ müssen in MM Kraftfeldern noch explizit quantenmechanische Effekte, nämlich die chemischen Bindungen zwischen den Atomen, modelliert werden. Dies geschieht durch die Definition der Bindungspotentiale $E_{\text{b}}(\mathbf{R})$ aus Gl. (1.2), deren explizite Form hier nicht näher behandelt werden soll. Es sei hier nur erwähnt, dass diese Potentiale die Elastizität von Bindungslängen zwischen zwei Atomen und Bindungswinkeln zwischen drei Atomen in harmonischer Näherung beschreiben. Ferner sind in $E_{\text{b}}(\mathbf{R})$ Potentiale für vieratomige Gruppen enthalten. Dies sind zum einen Torsionspotentiale um Diederwinkel, welche für vier linear angeordnete gebundene Atome definiert sind und z.B. die Rotation um die mittlere der drei chemischen Bindungen einschränken können. Andererseits sind auch Torsionspotentiale für Gruppen bestehend aus einem zentralen und drei daran gebundenen Atomen definiert, die es ermöglichen, die Planarität einer solchen Gruppe zu gewährleisten. In Bezug auf den Rechenzeitaufwand stellt die Auswertung der gebundenen Wechselwirkungen $E_{\text{b}}(\mathbf{R})$ im Vergleich zu den langreichweitigen Wechselwirkungen $E_{\text{nb}}(\mathbf{R})$ ein kleines Problem dar, da sie immer linear (und noch dazu mit einem kleinen Vorfaktor) skaliert.

Die so definierten MM Kraftfelder dienen der Beschreibung der Konformationsdynamik von Makromolekülen. Aufgrund des eingeschränkten, zumeist harmonischen, Ansatzes für die Bindungspotentiale eignen sich solche Kraftfelder nicht für die Beschreibung der Schwingungsspektren von Proteinen im mittleren IR Bereich [18, 75].

Trotz der stark vereinfachten und aufgrund der vernachlässigten Polarisationseffekte in vielerlei Hinsicht ungenauen MM Kraftfelder, stellen MM-MD Simulationen von bio-molekularen Systemen bislang den Hauptzugang zur Beschreibung der Konformationsdynamik dieser Systeme dar. Sie werfen anspruchsvolle Rechenzeitprobleme auf, was trotz des technischen Fortschritts der verfügbaren Rechner noch immer die Anwendungen begrenzt. So sind etwa Faltungsprozesse größerer Proteine mit Relaxationszeiten im Sekundenbereich durchaus üblich. Bei den benötigten Zeitschritten Δt im Bereich einer Femtosekunde sind Simulationszeiten im Bereich oberhalb einiger Mikrosekunden auch mit den heutigen Rechnern bei weitem noch nicht zu erreichen.

1.10 Komplexere Kraftfelder

Wie ich schon mehrfach betont habe, verzichten die oben diskutierten MM-Kraftfelder zugunsten relativer algorithmischer Einfachheit auf die Genauigkeit der Beschreibung. Ein Beispiel ist die Vernachlässigung von Polarisationseffekten in Gl. (1.2), wobei vor allem die Polarisation der Peptidgruppen für die Struktur und Dynamik von Proteinen von besonderer Bedeutung ist [62].

Kritisch kann ferner gesehen werden, dass die elektrostatischen Signaturen von Molekülen und molekularen Bausteinen näherungsweise durch Punktladungen beschrieben werden, während Atome in Molekülen tatsächlich durch ausgedehnte Ladungsverteilungen charakterisiert

sind. Dieser Unterschied kann bei größeren interatomaren Distanzen sicher vernachlässigt werden, bei van der Waals Kontaktdistanzen sollte die Partialladungsnäherung jedoch aufgrund der Überschätzung der elektrostatischen Wechselwirkungen Schwächen zeigen.

Eine Abhilfe könnte hier die Verwendung von Gaußladungen bieten. Solche geglätteten Ladungsdichten $\rho_i(\mathbf{r}|\mathbf{r}_i,\sigma) = q_i G(\mathbf{r}|\mathbf{r}_i,\sigma)$, welche vermittels einer Gaußschen Verteilung

$$G(\mathbf{r}|\mathbf{r}_i,\sigma) = (2\pi\sigma^2)^{-3/2}\exp\left(-\frac{|\mathbf{r}-\mathbf{r}_i|^2}{\sigma^2}\right) \tag{1.4}$$

der Breite σ definiert sind, schwächen die Nahfeldwechselwirkungen ab und vermeiden Divergenzen der Potentialfunktionen, was z.B. bei DFT/PMM Simulationen wichtig ist [28, 44].

Kritisch kann schließlich gesehen werden, dass die Partialladungen q_i in üblichen MM Kraftfeldern lediglich an den Atomorten lokalisiert sind, so dass die höheren Multipolmomente der dadurch beschriebenen Ladungsverteilungen von den tatsächlichen Multipolmonenten stark abweichen können. Dies führt auf kurze Distanzen zu fehlerhaften Beschreibungen der elektrostatischen Wechselwirkungen. Insbesondere im Fall des Wassermoleküls erzeugt die herkömmliche elektrostatische Architektur gröblich falsche höhere Multipolmomente [54].

Um diese Nachteile zu überwinden, wurden für das Wassermolekül fortgeschrittenere Modelle parametrisiert, die adäquatere höhere Multipolmomente vermittels einer komplexeren Modellierung der statischen Ladungsverteilung erreichen, wobei hier Partialladungen an masselose Orte im Molekülmodell positioniert werden. Solche Modelle sind schon lange insbesondere zur Beschreibung von Wasser verbreitet, wie z.B. die häufig genutzten MM Wassermodelle TIP4P [56] und TIP5P [76] zeigen. Analog hat auch mein Kollege Philipp Tröster diesen Ansatz bei der Konstruktion seiner PMM Wassermodelle mit vier bis sechs Kraftansatzpunkten gewählt [54, 55].

Um ferner die angesprochenen Polarisationseffekte in die MM Beschreibung einzubinden, wenden PMM Kraftfelder [54, 68–70] verschiedene Methoden an [77]. In dieser Arbeit stehen die eingangs erwähnten Modelle, welche induzierbare atomare Dipole $\mathbf{p}_i$ verwenden, im Mittelpunkt, da die von Philipp Tröster entwickelten Vier- bis Sechspunkt PMM-Modelle zu dieser Klasse gehören [54, 55]. Basierend auf der Analyse der Polarisationseigenschaften von Wassermolekülen durch Bernhard Schropp [57, 78] setzen diese Modelle Gaußsche Dipoldichten $\tilde{\mathbf{p}}_i = \mathbf{p}_i G(\mathbf{r}|\mathbf{r}_i,\sigma)$ anstatt der Punktdipole $\mathbf{p}_i$ ein, was den Aufwand zur Berechnung von Feldern und Potentialen nur bei kurzen Abständen ($r_{ij}/\sigma \leq 6$) erhöht. Bei größeren Abständen sind die Potentiale von Punkt- und Gaußdipolen nämlich numerisch identisch (einfache Genauigkeit).

Hier sei erwähnt, dass der zusätzliche Programmieraufwand zur Beschreibung induzierter Dipole vermittels sog. Drudeoszillatoren umgangen werden kann [77]. Drudeoszillatoren modellieren nämlich induzierte Dipole durch eine masselose Gegenladung, die vermittels eines harmonischen Potentials an ein geladenes und polarisierbares Atom gekoppelt ist. Wegen der damit verbundenen Einführung eines weiteren Punktes, an dem die elektrischen Felder ausgewertet werden müssen, erhöht dieses Vorgehen aber den Rechenaufwand.

Zusammenfassend lässt sich also feststellen, dass solche PMM Kraftfelder durch das Einführen von Ladungen an masselosen Punkten die Anzahl N^2 der auszuwertenden Wechselwirkungen vergrößern und dass komplexere elektrostatische Objekte, wie etwa Gaußladungen

anstatt herkömmlicher Partialladungen q_i, die Komplexität der Wechselwirkungsberechnung erhöhen. Der in dieser Doktorarbeit entwickelte Algorithmus zur Berechnung der langreichweitigen Wechselwirkungen hat sich als geeignet erwiesen, gerade solche komplexen Modellierungen mit vielen masselosen Ladungspunkten und Gaußschen Verteilungen effizient zu handhaben [42, 43].

1.11 Kontinuumselektrostatik

Im Gegensatz zu den realistischeren und damit aufwändigeren PMM Kraftfeldern zielt ein weiterer wichtiger Ansatz zur Simulation von Proteinen in wässriger Lösung in erster Linie auf die Steigerung der Effizienz. Anstatt die Proteinumgebung in Simulationen atomar aufgelöst darzustellen, wird in den sogenannten Kontinuumsmethoden gleich komplett auf die atomistische Beschreibung des Lösungsmittels zugunsten einer Kontinuumsnäherung verzichtet und so die Anzahl der auszuwertenden inter-atomaren Paarwechselwirkungen stark verkleinert.

Zum Verständnis sollte hier erwähnt werden, dass ein Simulationsmodell, welches den physiologischen Verhältnissen eines gelösten Proteins angemessen Rechnung trägt und Artefakte der verwendeten periodischen Randbedingungen vermeidet, bei expliziter Lösungsmittelbeschreibung zu mehr als 90 Prozent aus Wasser bestehen sollte [51, 62]. Bei der Simulation eines solchen Systems wird also ein Großteil der benötigten Rechenzeit ausschließlich auf die Berechnung der Wechselwirkungen zwischen den Atomen des Lösungsmittels aufgewendet. Der Wegfall dieses Rechenaufwands kann zu stark erhöhter Effizienz führen, wenn, wie bei HADES, die Einbeziehung der Wirkung des Kontinuums mit hinreichend kleinem Rechenaufwand verbunden ist [48].

Die physikalische Aufgabe, die sich impliziten Lösungsmittelmodellen stellt, ist die in jedem Intergrationszeitschritt einer Dynamiksimulation wiederholte Lösung der Poisson-Gleichung (PG)

$$\nabla[\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{1.5}$$

für ein Protein im Kontinuum. Hier beschreibt $\varepsilon(\mathbf{r})$ ein inhomogenes dielektrisches Medium, welches das Lösungsmittelkontinuum und das eingebettete Protein umfasst, $\Phi(\mathbf{r})$ ist das gesuchte elektrostatische Potential und $\rho(\mathbf{r})$ die Ladungsdichte des Proteins am Ort $\mathbf{r}$.

Viele verbreitete implizite Lösungsmittelmodelle lösen die gestellte Aufgabe nicht, sondern versuchen durch heuristische Ansätze Näherungslösungen zu raten. Dies ist z.B. für die weit verbreitete Generalized Born Methode der Fall [79–81], welche das Born-Modell eines einzelnen Ions im Dielektrikum auf Ansammlungen von Ladungen, wie sie z.B. bei der Modellierung eines Proteins durch (P)MM Kraftfelder definiert werden, verallgemeinert. Numerische Lösungen der PG liefern dagegen zwar das Potential auf dem bei der Rechnung verwendeten und das gelöste Protein überdeckenden Gitter, aber weder hinreichend genaue Ausdrücke für die vom zugehörigen elektrischen Feld erzeugten und auf die Ladungen wirkenden Kräfte noch irgendwelche Auskünfte für die vom Kontinuum ausgehenden Reaktionskräfte [82, 83]. Daher erlauben solche Methoden prinzipiell keine energieerhaltenden Kontinuums-MD Simulationen.

Wie ich schon in Abschnitt 1.5 erwähnt habe, wurde von Sebastian Bauer während seiner Doktorarbeit in der Arbeitsgruppe für theoretische molekulare Biophysik am BMO eine neue Hamiltonsche Methode zur genauen und hinreichend effizienten Lösung der PG für Proteine im dielektrischen Kontinuum entwickelt [48, 58, 63]. Sebastian Bauer führte damit frühere Ansätze [84, 85] zur Vollendung, welche Teillösungen des oben skizzierten Problems geliefert hatten. Das neue Verfahren liegt in Form des im Programm IPHIGENIE implementierten HADES Algorithmus vor und wurde mit diesem Programm der wissenschaftlichen Öffentlichkeit zur Verfügung gestellt [53].

Der Bauersche Ansatz behebt, durch die Berechnung analytischer und Hamiltonscher Kraftfunktionen, viele frühere Probleme [58, 85–89], da die Reaktionskräfte, welche durch das Kontinuum auf die Proteinatome ausgeübt werden und den sogenannten dielektrischen Druck verursachen, durch HADES adäquat beschrieben werden. Damit wird ein Protein im Kontinuum als isoliertes, wechselwirkendes Vielteilchensystem dargestellt, für das die üblichen Erhaltungssätze (Energie, Gesamtimpulse) gelten.

Der HADES Algorithmus beruht auf einer Reformulierung [84, 85, 90] der PG, in welcher eine Antipolarisationsdichte innerhalb des Proteinvolumens, für das eine kleine Dielektrizitätskonstante ε_s angenommen wird, die Polarisation des umgebenden Kontinuums mit der Dielektrizitätskonstante ε_c ersetzt. Diese Antipolarisation erzeugt das Reaktionsfeld. Sie wird durch Gaußsche Reaktionsfelddipoldichten $\tilde{\mathbf{p}}_i$ und Gaußsche Abschirmladungsdichten mit der Gesamtladung $\hat{q}_i = -q_i(1 - \varepsilon_s/\varepsilon_c)$ annähernd dargestellt.

Die induzierten Gaußschen Antipolarisations-Dipole $\tilde{\mathbf{p}}_i$ werden in HADES selbstkonsistent aus den über die jeweiligen atomaren Volumina gemittelten elektrischen Feldern $\langle\tilde{\mathbf{E}}(\mathbf{r}_i)\rangle_{\sigma_i}$ über die Selbstkonsistenzbedingung $\tilde{\mathbf{p}}_i = -\alpha_i\langle\tilde{\mathbf{E}}(\mathbf{r}_i)\rangle_{\sigma_i}$ mit $\alpha_i > 0$ abgeleitet [48, 58]. Damit ist klar, dass HADES, sieht man von dem Vorzeichen in der Selbstkonsistenzbedingung für die induzierten Dipole ab, auf einer zu PMM Kraftfeldern sehr ähnlichen elektrostatischen Beschreibung basiert.

Der Effizienzgewinn, der durch das Ersetzen eines expliziten Lösungsmittelmodells durch den HADES Algorithmus zu erreichen ist, ist im Prinzip groß. Für ein kleines α-helikales Peptid mit $N = 150$ Atomen wurde von Bauer et al. [63] im Vergleich zu einer Beschreibung mit explizitem Lösungsmittel eine Beschleunigung der Simulation des Peptids um einen Faktor 20 erreicht.

Bei größeren Proteinen geht dieser Effizienzvorteil jedoch auf Grund des quadratischen Skalierungsverhaltens $\sim N^2$ rasch verloren, da bei HADES-MD Simulationen die elektrostatischen Wechselwirkungen als atomare Paarwechselwirkungen ausgewertet werden. Ich konnte nun zeigen (vgl. Unterkapitel 2.3), dass durch die Einbindung Hamiltonscher SAMM-Kräfte in HADES-MD [43], d.h. durch die Entwicklung eines HADES/SAMM-MD Verfahrens, ein linear mit der Anzahl der Proteinatome N skalierender Algorithmus erzeugt werden kann, der mit Drehimpuls- und Impulserhaltung sowie bei sehr guter, annähernder Energieerhaltung die grundlegenden und vorteilhaften Eigenschaften von HADES bewahrt.

1.12 Berechnung langreichweitiger Wechselwirkungen

Zur Reduktion der bei PMM/MD-Simulationen für die Auswertung der langreichweitigen Wechselwirkungen benötigten Rechenzeit sind verschiedene Algorithmen verbreitet. Gemeinsam ist diesen Methoden, dass sie die auftretenden Summationen über die interatomaren Paarwechselwirkungen [vgl. Gl. (1.3)] näherungsweise umformulieren, um durch kontrolliertes Zusammenfassen oder Weglassen von Summanden effiziente Algorithmen zu generieren.

Der einfachste Algorithmus, der in Abschnitt 1.8 bereits für Dispersions- und Repulsionswechselwirkungen erwähnt wurde, besteht darin, ab einer gewissen Entfernung R_c, dem Abschneideradius, jegliche Wechselwirkung zu vernachlässigen. Jedem Versuch, über diesen trivialen Ansatz für die Elektrostatik einen mit der Anzahl N der Atome linear skalierenden Algorithmus zu erhalten, steht der langreichweitige Charakter dieser Wechselwirkung entgegen, da hier ein einfaches Abschneiden üblicherweise zu signifikanten Artefakten in der Dynamik der betrachteten Systeme führt [51, 91, 92].

1.13 Gittersummenmethoden

Ein weit verbreiteter Ansatz zur Milderung des Rechenzeitproblems durch approximative Methoden sind die Gittersummenmethoden, welche auf der Ewald-Summation [93] basieren. Diese wurde ursprünglich für Kristallsysteme entwickelt und vervielfältigt ein meist kubisches Simulationssystem durch periodische Anlagerung von identischen Systemen. Die Auswertung der Wechselwirkung dieser unendlichen Summe von Bildern mit dem zentralen System erfolgt über eine geschickte Reformulierung. Eine ausführliche Behandlung dieser Methode findet sich z.B. in [94] und für Gittersummen, die nicht nur Partialladungen q_i sondern auch multipolare Beiträge enthalten, in [95], so dass hier nur die grundlegende Idee skizziert werden soll.

Startpunkt ist die Gittersumme

$$\Phi^{\mathrm{GS}}(\mathbf{r}_i) = \sum_{n=0}^{\infty} \sum_{\substack{j=1 \\ j\neq i \vee n\neq 0}}^{N} \frac{q_j}{|\mathbf{r}_i - \mathbf{r}_j - \mathbf{t}_n|} \tag{1.6}$$

für das elektrostatische Potential Φ^{GS} an den Atomorten $\mathbf{r}_i$ im zentralen Bild, welches durch die Atome j im zentralen ($n = 0$) und allen periodischen Bildern ($n > 0$) erzeugt wird. Hier beschreibt der Vektor $\mathbf{t}_n$ den Verbindungsvektor von dem Zentrum des n-ten Bildes zu dem Zentrum des zentralen Bildes.

Zum Zwecke der Umformulierung werden (i) alle auftretenden Ladungen q_j durch Gaußsche Ladungsdichten $\rho_{j,\mathrm{s}}(\mathbf{r}|\mathbf{r}_j, \sigma) = -q_j G(\mathbf{r}|\mathbf{r}_j, \sigma)$ der Breite σ abgeschirmt [vgl. Gl. (1.4)]. Um (ii) den Einfluss der $\rho_{j,\mathrm{s}}(\mathbf{r})$ auf das Gesamtpotential wieder zu eliminieren, wird zusätzlich noch ein weiter Satz Gaußscher Ladungsdichten $\rho_{j,\mathrm{g}}(\mathbf{r}) = -\rho_{j,\mathrm{s}}(\mathbf{r})$ eingeführt, so dass sich

die Beiträge $\rho_{j,\mathrm{s}}(\mathbf{r})$ und $\rho_{j,\mathrm{g}}(\mathbf{r})$ zur Ladungsdichte gegenseitig aufheben. Zur eleganten Auswertung von Gl. (1.6) werden nun die Ladungen q_j und die zugehörigen Abschirmladungen $\rho_{j,\mathrm{s}}(\mathbf{r})$ zusammengefasst. Jedes Paar (q_j, $\rho_{j,\mathrm{s}}$) aus einem Bild n erzeugt nun ein auf der Skala der Breiten σ ein sehr kurzreichweitges Potential

$$\phi_{j,n}(\mathbf{r}_i) = q_j \frac{\mathrm{erfc}(\sigma|\mathbf{r}_{ij} - \mathbf{t}_n|)}{|\mathbf{r}_{ij} - \mathbf{t}_n|}, \tag{1.7}$$

so dass die Gittersumme (siehe [95]) für diese Beiträge

$$\Phi^{\mathrm{k}}(\mathbf{r}_i) = \sum_{n=0}^{\infty} \sum_{\substack{j=1 \\ j\neq i \vee n\neq 0}}^{N} \phi_{j,n}(\mathbf{r}_i) \tag{1.8}$$

im Ortsraum schon für relativ kleine Abstände $r_{\mathrm{ij,n}} = |\mathbf{r}_{ij} - \mathbf{t}_n|$ abgebrochen werden kann. Dies beschleunigt die Auswertung stark, müssen nun doch für jedes Atom i nur Wechselwirkungen mit wenigen benachbarten Atomen j ausgewertet werden. Da dieser Abschneideradius in MD-Simulationsalgorithmen oft auch gleichzeitig für die Lennard-Jones Wechselwirkungen Anwendung findet, wird er meist in der Größenordung 10 Å gewählt.

Das periodische Potential $\Phi^{\mathrm{r}}(\mathbf{r}_i)$ der Gaußschen Gegenladungen $\rho_{i,\mathrm{g}}(\mathbf{r})$ kann per Fourier-Transformationen im reziproken k-Raum ausgewertet werden. Die Summe über die Bilder n übersetzt sich im k-Raum in eine Summe über reziproke Vektoren $\mathbf{k}$. Auch die sich so ergebende Summe (siehe [95])

$$\Phi^{\mathrm{r}}(\mathbf{r}_i) = \frac{1}{\pi V} \sum_{\mathbf{k}\neq 0} \sum_{j=1}^{N} q_j \exp\left(\frac{-\pi^2\mathbf{k}^2}{\sigma^2}\right) \frac{1}{\mathbf{k}^2} \exp\left[2\pi i\, \mathbf{k} \odot (\mathbf{r}_j - \mathbf{r}_i)\right], \tag{1.9}$$

in der V das Volumen des kubischen Systems ist, kann in Abhängigkeit der Gaußbreite σ abgebrochen werden. Im Gegensatz zur Summation im Ortsraum, in dem kleine Breiten σ ein effizienteres Abbrechen der Gittersumme ermöglichen, sind für die Summation im k-Raum mit kleineren Breiten σ mehr Summationsterme nötig um eine vorgegebene Genauigkeit zu erreichen. Um nicht erwünschte Wechselwirkungsbeiträge, wie etwa die der elektrostatischen Wechselwirkungen zwischen chemisch gebundenen Atomen, auszuschließen, müssen diese schließlich noch explizit von dem periodischen Potential Φ^{r} abgezogen werden (siehe [95]).

Durch ein optimales Tarieren der Breiten σ sowie des Abschneideradius R_c und der Anzahl der im k-Raum ausgewerteten Vektoren $\mathbf{k}$ kann ein Skalierungsverhalten von $N^{3/2}$ erzielt werden [96]. Insbesondere für den Bereich der MM-MD Simulationen wurde die Ewald-Summation durch den Einsatz von schnellen Fourier-Transformationen weiterentwickelt. Durch diese Methoden wurde der Rechenaufwand für die approximative Auswertung der periodischen Potentiale $\Phi^{\mathrm{r}}(\mathbf{r}_i)$ bzw. Kräfte weiter reduziert, so dass die resultierenden Algorithmen mit $N \log N$ skalieren [97–99].

Durch den Einsatz periodischer Randbedingungen ist bei Gittersummenmethoden der Druck in Simulationssystemen kontrollierbar. Andererseits verwenden sie ein periodisches elektrostatisches Potential, welches bei Simulationen von nicht-periodischen Systemen, wie z.B.

Protein-Lösungsmittelsystemen, artifizielle Effekte erzeugen kann [100]. Offensichtlich sind Gittersummenmethoden wegen ihrer Periodizität nicht mit Methoden wie HADES koppelbar, da in HADES-MD Rechnungen geschlossene Systeme in nicht-periodischen Randbedingungen simuliert werden.

1.14 Schnelle Multipolmethoden

Eine weitere Methode, die ein noch vorteilhafteres Skalierungsverhalten als die genannten Gittermethoden bietet, ist die eingangs schon erwähnte schnelle Multipolmethode. Die herausragendste Eigenschaft der FMM-Algorithmen ist, dass mit ihnen der Rechenaufwand für die näherungsweise Auswertung von N^2 atomaren Paarwechselwirkungen linear mit der Anzahl der Proteinatome N skaliert [31]. Um dieses Skalierungsverhalten zu erreichen, berechnen die FMM-Algorithmen die langreichweitigen Wechselwirkungen eines System S abstandsabhängig. Wechselwirkungen zwischen zwei voneinander separierten Mengen von Atomen — hier Cluster genannt — werden, wenn ihr Abstand r im Verhältnis zu ihrer Ausdehnung groß genug ist, durch speziell formulierte Multipol-Wechselwirkungen der atomaren Ladungsverteilungen genähert.

1.14.1 Hierarchische Systempartionierung in schnellen Multipolmethoden

FMM Methoden benötigen als algorithmische Infrastruktur also zunächst die Abbildung des Simulationssystems S auf einen mehrstufigen, ineinander hierarchisch geschachtelten Baum atomarer Cluster, deren Ladungsverteilungen durch Multipolmomente repräsentiert werden können. Um diese notwendige Hierarchie zu erstellen, wenden die meisten FMM-Varianten eine regelmäßige geometrische Zerlegung des Simulationsvolumens an und bilden das System S so auf den benötigten Baum ab [52, 101–107]. Für ein kubisches Simulationssystem der Kantenlänge L in drei Raumdimensionen bedeutet dies z.B. die Zerlegung des Systems in acht disjunkte Kuben der Kantenlänge $L/2$ auf der obersten Ebene l_{t}. Die so generierten kleineren Volumina werden mit dem selben Verfahren wieder in acht Subvolumina zerlegt und dadurch eine weitere Hierarchieebene $l_{\text{t-1}}$ generiert. Dieser Vorgang wird wiederholt, bis eine gewünschte Feinheit der Zerlegung des Systems erreicht ist. Da jeder Eltern-Cluster in dieser Systematik in acht Kind-Cluster zerfällt, wird der resultierende Baum auch als Oktalbaum bezeichnet.

In SAMM [30, 32, 33, 51] wird im Gegensatz zu einer solchen geometrischen Zerlegung ein alternatives Verfahren eingesetzt, das die speziellen Eigenschaften bio-molekularer Systeme, z.B. die Existenz chemisch gebundener Atomgruppen, ausnutzt. Hier werden zunächst die Cluster der untersten Ebene, als sogenannte Strukturelle Einheiten definiert. Typische Beispiele solcher Einheiten sind Wassermoleküle, Amidgruppen und kleine Seitengruppen in Proteinen. Daraufhin wird durch ein hocheffizientes Clusteringverfahren [108] eine mehrstufige Hierarchie kompakter, atomarer und ineinander geschachtelter Cluster erstellt. Wie bei der geometrischen Zerlegung des Simulationsvolumens gilt hier, dass alle Clusterpaare

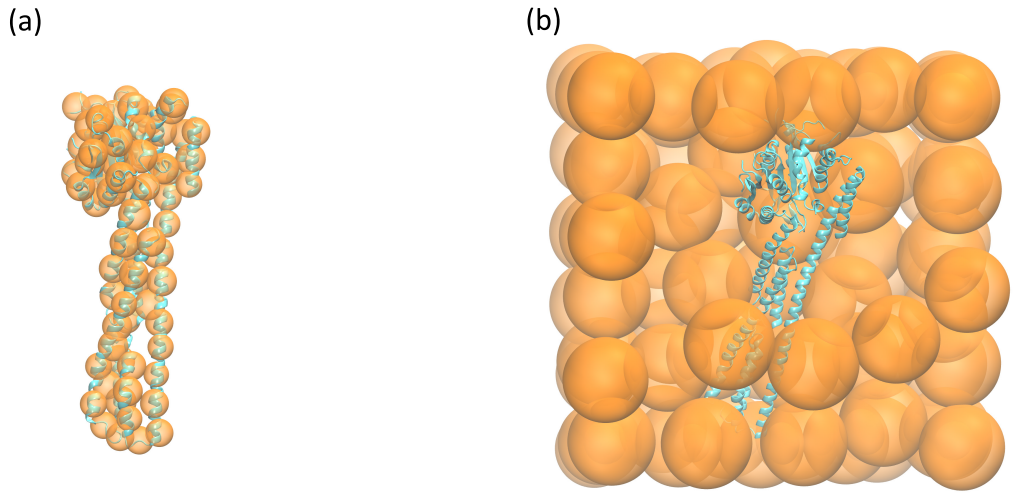


Abbildung 1.4: In der linken Abbildung (a) ist das Ergebnis des adaptiven Clustering-Verfahren in SAMM für ein typisches Beispiel abgeschlossener Systeme ohne periodische Randbedingungen gezeigt. Die orangenen Kugeln stellen die atomaren Cluster der obersten Hierarchieebene dar. In der rechten Teilabbildung (b) ist ein analoges Ergebnis für ein typisches kubisches System als Basis für Simulationen unter periodischen Randbedingungen gezeigt.

(C_i, C_j) einer Hierarchieebene disjunkt sind ($C_i \cap C_j = \emptyset$ für $i \neq j$) und dass die Vereinigung $\cup_i C_i$ aller Cluster C_i einer Ebene gerade wieder alle Atome des Systems S beinhaltet. Allerdings erstellt SAMM keinen oktalen, sondern einen quaternären Baum, was zur Folge hat, dass sich die mittleren Radien der Cluster auf zwei benachbarten Hierarchieebenen nicht um einen Faktor 2 sondern nur um ca. $4^{1/3} \approx 1.587$ unterscheiden, so dass SAMM für ein gegebenes System in der Regel eine feinere Auflösung mit mehr Hierarchieebenen erreicht [30, 51].

Abbildung 1.4 zeigt exemplarisch die Ergebnisse dieses Clusteralgorithmus für zwei typische Anwendungsbeispiele. In der linken Teilabbildung (a) sind die resultierenden Cluster der höchsten Hierarchieebene als orange gefärbte Kugeln und das geclusterte Protein GBP1 [109] gezeigt. Diese Teilabbildung verdeutlicht die hervorragende Adaptivität des Clusterings für Systeme in offenen Randbedingungen, wie sie z.B. im Rahmen der HADES-MD auftreten. In der rechten Teilabbildung (b) ist analog das Ergebnis eines Clusterings für ein kubisches System gezeigt, in dem das Protein von explizitem Lösungsmittel umgeben ist (die umgebenden Wassermoleküle sind aus Gründen der Übersichtlichkeit nicht abgebildet). Die in (b) gezeigte Menge von Clustern kann typischerweise als das zentrale System in Simulationen mit periodischen Randbedingungen genutzt werden.

1.14.2 Toroidale Randbedingungen in SAMM MD-Simulationen

Abbildung 1.5 illustriert den von Mathias et al. [51] entwickelten SAMM/RF Ansatz für die Simulation von Systemen in explizitem Lösungsmittel unter periodischen Randbedingungen.

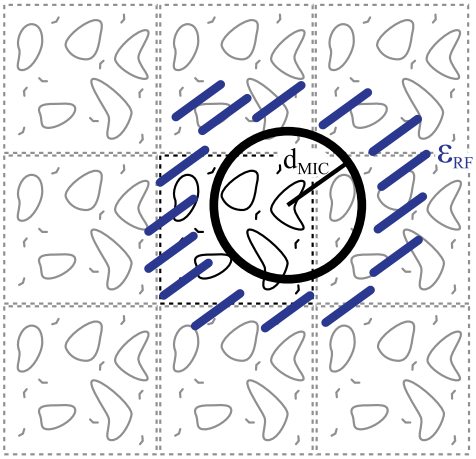


Abbildung 1.5: Die Abbildung beschreibt die toroidalen Randbedingungen, welche in Verbindung mit einem effizienten Reaktionsfeldansatz die Grundlage für explizite Lösungsmittelsimulationen in SAMM/RF [51] darstellen. Innerhalb eines Radius d_{MIC} um ein Molekül werden alle Wechselwirkung explizit berechnet. Für größere Abstände wird das Lösungsmittel durch ein Dielektrikum (skizziert durch blaue Striche) mit der Dielektrizitätskonstante $\varepsilon_{\mathrm{RF}}$ beschrieben.

Im speziellen sind sogenannte toroidale Randbedingungen [110] dargestellt, in denen ein zentrales kubisches System von genau einer Schicht aus Spiegelbildern umgeben ist. Es werden also nicht wie in den Gittermethoden die Wechselwirkungen aller periodischer Bilder mit dem zentralen Bild ausgewertet werden. Der Kreis in der Abbildung, dessen Radius d_{MIC} gerade gleich der halben Kantenlänge $L/2$ des kubischen Systems ist, illustriert den Bereich in dem sichergestellt ist, dass jeweils nur maximal ein periodisches Bild der das System zusammensetzenden Objekte vorhanden ist. Im Geltungsbereich dieser sogenannten *minimum image convention* (MIC) [94] werden in SAMM alle Wechselwirkungen der Atome explizit, sei es atomar-paarweise oder über schnelle Multipolmethoden, berechnet. Die blauen Striche in der Abbildung symbolisieren ein dielektrisches Kontinuum. Die in SAMM/RF nicht explizit ausgewerteten Interaktionen werden durch dieses Kontinuum repräsentiert und durch einen effizienten Reaktionsfeldansatz ausgewertet [51].

1.14.3 Approximation langreichweitiger Wechselwirkungen in SAMM$_{2003}$

Die schnelle Multipolmethode SAMM$_{2003}$ [30, 32, 33, 51] stellt den direkten Vorgänger der in den Kapiteln 2.1-2.3 beschriebenen algorithmischen Entwicklungen dar. Um dem Leser leichter zu ermöglichen, die Neuentwicklungen von Eigenschaften der Vorgängerversion zu unterscheiden, soll die folgende Darstellung des typischen Vorgehens zur Berechnung von Wechselwirkungen in Schnellen Multipolmethoden explizit am Beispiel der Version SAMM$_{2003}$ erfolgen. Zur Einordnung in die veröffentlichte Literatur über schnelle Multipolmethoden soll hier noch betont werden, dass sich die verschiedenen FMM-Varianten nach der Wahl der Ko-

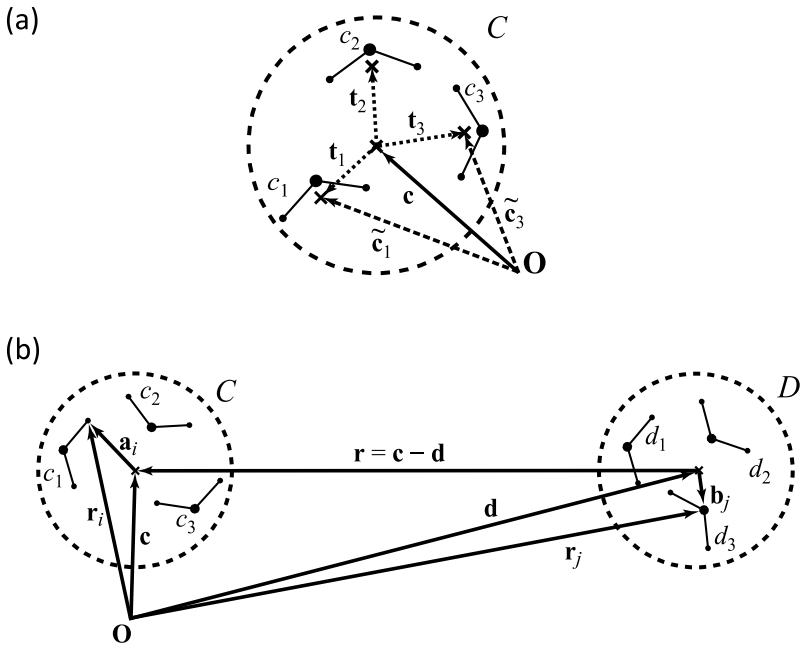


Abbildung 1.6: Die Abbildung stellt die maßgeblichen Eigenschaften der Clustergeometrie in den SAMM dar. Oben (a) ist für einen Cluster C die geometrische Beziehung zu seinen Kind-Clustern auf dem der nächst niedrigeren Hierarchieebene dargestellt. In dem unteren Teil (b) der Abbildung ist die maßgebliche Geometrie zweier Cluster C und D, deren Wechselwirkung durch den SAMM Algorithmus berechnet wird, dargestellt. Außerdem sind hier die Koordinaten der Atome $i \in C$ und $j \in D$ in globalen ($\mathbf{r}_i$ bzw. $\mathbf{r}_j$) und cluster-lokalen Koordinaten ($\mathbf{a}_i$ bzw. $\mathbf{b}_j$ gezeigt.

ordinatenbasis in zwei Klassen unterscheiden lassen. Zum einen werden, wie auch in der ursprüngliche Variante von Greengard [31], sphärische Koordinaten für die Multipoldarstellungen genutzt [31, 101–103], zum anderen — zu dieser Klasse gehört auch SAMM — werden diese Entwicklungen in kartesischen Koordinaten durchgeführt [33, 52, 104–107, 111, 112].

Der SAMM Algorithmus zur Berechnung der Wechselwirkungen unterteilt sich in mehrere Phasen. Die Abbildung 1.6 bildet die zur mathematischen Beschreibung notwendige Geometrie ab. In der oberen Hälfte (a) der Abbildung sind die geometrischen Beziehungen eines Clusters C mit dem Zentrum $\mathbf{c}$ und seiner Kind-Cluster c_i $(i, 2, 3)$ an den Orten $\tilde{\mathbf{c}}_i$. dargestellt. Im internen Koordinatensystem des Clusters C sind die Positionen der Kind-Cluster c_i durch die Verschiebevektoren $\mathbf{t}_i = \tilde{\mathbf{c}}_i - \mathbf{c}$ gegeben. Die untere Hälfte (b) zeigt schematisch eine Wechselwirkungsgeometrie zweier Cluster C und D mit den Positionen $\mathbf{c}$ und $\mathbf{d}$, welche sich durch den Abstand $\mathbf{r} = \mathbf{c} - \mathbf{d}$ von einander getrennt sind. Außerdem zeigt (b) die Koordinaten der Atome $i \in C$ und $j \in D$ welche in dem globalen Koordinatensystem durch $\mathbf{r}_i$ bzw. $\mathbf{r}_j$ gegeben sind. In den jeweiligen lokalen Koordinaten der Cluster C und D sind die Positionen der Atome durch $\mathbf{a}_i = \mathbf{r}_i - \mathbf{c}$ sowie durch $\mathbf{b}_j = \mathbf{r}_j - \mathbf{d}$ gegeben.

Als Basis des Algorithmus werden in zwei Schritten die jeweiligen Multipolmomente m-ter Ordnung M^m für alle Cluster auf allen Hierarchieebenen berechnet. Hierzu werden für die niedrigste Ebene der Hierarchie die jeweiligen Multipolmomente M^m über die atomaren

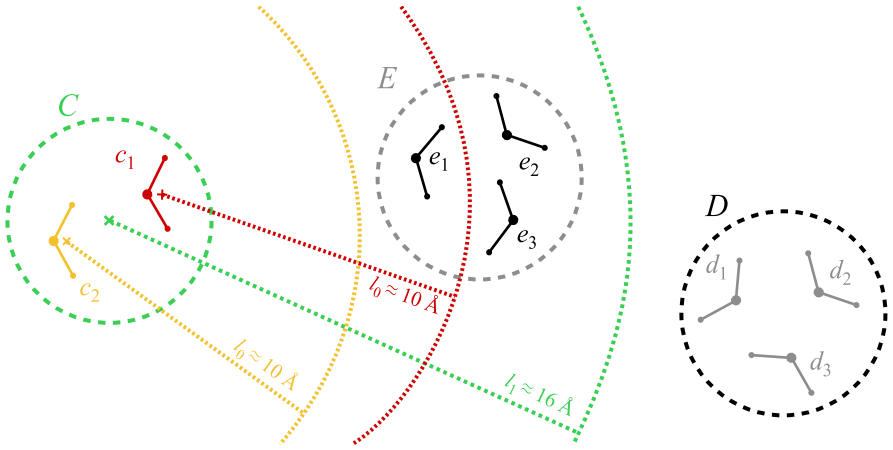


Abbildung 1.7: Schematische Darstellung der SAMM$_{2003}$ Distanzklassen. Gezeigt sind die drei Cluster C, D, E der Hierarchieebene $l = 1$, sowie deren jeweilige Kind-Cluster c_i, d_i, e_i auf Ebene $l = 0$ (Strukturelle Einheiten, Wassermoleküle).

Partialladungen q_i innerhalb der Cluster berechnet. Aus den q_i ergeben sich so exemplarisch die zwei niedrigsten Momente eines Clusters C zu $\mathbf{M}^0 = \sum_i q_i$ (Monopolmoment) und $\mathbf{M}^1 = \sum_i q_i \mathbf{a}_i$ (Dipolmoment). Formeln für die Berechnung von Multipolmomenten $\mathbf{M}^m$ höherer Ordnungen finden sich z.B. in [41]. Daraufhin werden die Multipolmomente der Cluster C an den Orten $\mathbf{c}$ auf höheren Ebenen jeweils aus den Multipolmomenten der zugehörigen Kind-Cluster c zentriert um $\tilde{\mathbf{c}}$ auf der nächst niedrigeren Ebene berechnet, z.B. $\mathbf{M}^{0,C} = \sum_c \mathbf{M}^{0,c}$ und $\mathbf{M}^{1,C} = \sum_c [\mathbf{M}^{1,c} + \mathbf{M}^{0,c}(\tilde{\mathbf{c}} - \mathbf{c})]$.

Um nun mit Hilfe dieser Multipolmomente die exakte Berechnungen der atomaren Wechselwirkungen durch geeignete Multipolnäherungen zu ersetzen, wird zunächst eine Regel benötigt um zu entscheiden, ob der Abstand $r = |\mathbf{c} - \mathbf{d}|$ zwischen zwei Clustern C und D dafür groß genug ist. In SAMM$_{2003}$ wurde diese Entscheidung anhand eines Distanzenklassenkriteriums gefällt. Hierbei wurde für jede Ebene l ein Abstand l_l definiert. Solange für den Abstand r zwischen einem Paar C, D auf der l-ten Ebene $r \geq l_l$ gilt, können die Wechselwirkung zwischen den Atomen in diesem Paar über die im Folgenden näher beschrieben Multipolwechselwirkungen zwischen den Clustern C und D auf Ebene l dargestellt werden. Um zu verhindern, dass atomare Wechselwirkungen mehrfach durch Paare von Clustern auf verschiedenen Ebenen berechnet werden, wird das obige Abstandskriterium für Paare C, D beginnend auf der höchsten Ebene und dann Ebene für Ebene absteigend ausgewertet. Sollte das Kriterium für ein Paar C, D erfüllt sein, wird die Auswertung für alle Paare von Kind-Clustern $c \in C, d \in D$ auf den niedrigen Ebenen übersprungen.

Abbildung 1.7 beschreibt dieses Kriterium für die in SAMM$_{2003}$ verwendeten untersten zwei Ebenen $l = 0, 1$ anhand einer Beispielgeometrie von drei Clustern C, D und E wie auch ihrer jeweiligen Kind-Cluster c_i, d_i und e_i. In der Abbildung werden für den Cluster C und seine Kind-Cluster c_i diese Klassen dargestellt um zu verdeutlichen, auf welcher Hierarchieebene diese mit zwei entfernten Clustern D und E sowie ihren Kind-Clustern d_i und e_i wechselwirken. Aus Sicht des Cluster C ist eine Cluster-Cluster Wechselwirkung nur mit Cluster D möglich, da Cluster E nicht den geforderten Abstand von $l_1 = 16$Å hat. Die Wechsel-

wirkung zwischen C und E muss also auf der nächst niedrigeren Ebene überprüft werden. Der Kind-Cluster c_2 ist weit genug von allen Clustern e_i ($i = 1, 2, 3$) entfernt, so dass diese Wechselwirkungen auf Ebene $l = 0$ durch schnelle Multipolmethoden ausgewertet werden können. Für Cluster c_1 gilt dies nur für die Wechselwirkungspartner e_2 und e_3. Die Interaktionen zwischen c_2 und e_1 müssen somit paarweise über exakte atomare Beiträge berechnet werden.

Die näherungsweise SAMM$_{2003}$ Wechselwirkung für ein Paar C, D ergibt für die Atome $i \in C$ die durch Cluster D generierten Potentiale

$$\Phi^D(\mathbf{r}_i) = \sum_{n=0}^{2} \frac{1}{n!} \mathbf{a}_i^{(n)} \odot \partial_{(n)} \sum_{m=0}^{2} \Phi^{m,D}(\mathbf{c}) \tag{1.10}$$

und Felder

$$\mathbf{E}^D(\mathbf{r}_i) = -\sum_{n=0}^{1} \frac{1}{n!} \mathbf{a}_i^{(n)} \odot \partial_{(n+1)} \sum_{m=0}^{2} \Phi^{m,D}(\mathbf{c}) \tag{1.11}$$

an den Atomorten $\mathbf{r}_i$ aus einer lokalen Taylorentwicklung vermittels der lokalen Atomkoordinaten $\mathbf{a}_i = \mathbf{r}_i - \mathbf{c}$ um das Zentrum $\mathbf{c}$ des Clusters C und den n-ten Ableitungen $\partial_{(n)} \equiv \nabla_{\mathbf{r}}^{(n)}$ der Multipolpotentiale $\Phi^{m,D}(\mathbf{c})$ (SAMM$_{2003}$: $n = 0, 1, 2$) des m-tem Multipolmoments $\mathbf{M}^m$ (SAMM$_{2003}$: $m = 0, 1, 2$) nach $\mathbf{r}$ als Entwicklungskoeffizienten. Hier ist $\nabla^{(n)}$ als n-faches äußeres Produkt des Vektors ∇ zu verstehen. Explizite Darstellungen der Multipolpotentiale $\Phi^{m,D}(\mathbf{c})$ finden sich z.B. in [41].

Hier ist zu betonen, dass die in den Gln. (1.10) und (1.11) beschriebene Art der Multipol-Atom Wechselwirkung als zweifache Taylorentwicklung zu verstehen ist. Neben den Multipolentwicklungen in Cluster D werden die resultierenden Multipolpotentiale $\Phi^{m,D}$ und Multipolfelder $-\partial_{(1)}\Phi^{m,D}$ nicht direkt an den Atomorten $\mathbf{r_i}$ ausgewertet, sondern über eine lokale Taylorentwicklung um das Zentrum $\mathbf{c}$ des Clusters C approximiert. Das ist die zentrale Eigenschaft der FMM, die das lineare Skalierungsverhalten erst ermöglicht. Der Barnes-Hut Algorithmus, ein logischer Vorgänger der FMM, welcher die Monopolpotentiale entfernter Cluster direkt an Atomorten $\mathbf{r}_i$ für gravitativ wechselwirkende N-Teilchen Systeme berechnet, skaliert z.B. nur mit $N\log(N)$ [113].

Um das lineare Skalierungsverhalten der FMM bei der effizienten Berechnung der atomaren Größen $\Phi^D(\mathbf{r}_i)$ und $\mathbf{E}^D(\mathbf{r}_i)$ aus den Gln. (1.10) und (1.11) zu erreichen, darf jedoch nicht für jedes Paar C, D eine lokale Taylorentwicklung ausgeführt werden. Aus der linearen Struktur dieser Gleichungen wird klar, das erst alle Beiträge unterschiedlicher Cluster zu den Entwicklungskoeffizienten, also den Potentialen Φ und den n-ten Ableitungen $\partial_{(n)}\Phi$ dieser Potentiale, aufsummiert werden sollten, bevor die lokale Taylorentwicklung ausgeführt wird.

Hierzu werden die Entwicklungskoeffizienten in zwei Schritten in einem Top-down (engl. von oben nach unten) Prozess zusammengefasst. Zunächst werden die jeweiligen lokalen Entwicklungskoeffizienten n-ter Ordnung aller mit einem Cluster C auf der gleichen Hierarchieebene wechselwirkenden Cluster D aufsummiert zu

$$\mathbf{T}^n_{\text{interact}}(\mathbf{c}) \equiv \sum_D \partial_{(n)} \sum_{m=0}^{2} \Phi^{m,D}(\mathbf{c}), \tag{1.12}$$

wobei die Multipolpotentiale $\Phi^{m,D}(\mathbf{c})$ m-ter Ordnung Berücksichtigung finden.

Diese Koeffizienten für die Cluster C werden danach, beginnend auf der höchsten Ebene [für diese Ebene gilt: $\mathbf{T}^{\text{all},n}(\mathbf{c}) \equiv \mathbf{T}^{n}_{\text{interact}}(\mathbf{c})$], auf die jeweiligen Kind-Cluster c der nächst niedrigeren Ebene durch die Taylorentwicklung

$$\tilde{\mathbf{T}}^{n}_{\text{inherit}}(\tilde{\mathbf{c}}) = \sum_{l=0}^{2-n} \frac{1}{l!} \mathbf{t}^{(l)} \odot \mathbf{T}^{\text{all},n+l}(\mathbf{c}) \tag{1.13}$$

vererbt, wobei $\mathbf{t} = \tilde{\mathbf{c}} - \mathbf{c}$ den Abstandsvektor zwischen dem Eltern-Cluster C und dem Kind-Cluster c wie in Abbildung 1.6 dargestellt beschreibt. So ergeben sich auf der nächstniedrigeren Ebenen die gesammelten Koeffizienten

$$\tilde{\mathbf{T}}^{n}_{\text{all}}(\tilde{\mathbf{c}}) = \tilde{\mathbf{T}}^{n}_{\text{interact}}(\tilde{\mathbf{c}}) + \tilde{\mathbf{T}}^{n}_{\text{inherit}}(\tilde{\mathbf{c}}) \tag{1.14}$$

für die Cluster c an den Positionen $\tilde{\mathbf{c}}$ als eine Summe aus direkten Wechselwirkungsbeiträgen $\tilde{\mathbf{T}}^{n}_{\text{interact}}(\tilde{\mathbf{c}})$ und den vererbten Beiträgen höherer Ebenen $\tilde{\mathbf{T}}^{n}_{\text{inherit}}(\tilde{\mathbf{c}})$. Wenn so alle Koeffizienten auf bis auf die niedrigste Ebene vererbt wurden, werden aus diesen gesammelten lokalen Koeffizienten $\tilde{\mathbf{T}}^{n}_{\text{all}}(\tilde{\mathbf{c}})$, welche alle Wechselwirkungsbeiträge aller Cluster-Ebenen beinhalten, atomare Potentiale

$$\Phi^{\text{all}}(\mathbf{r}_i) = \sum_{n=0}^{2} \frac{1}{n!} \mathbf{a}_i^{(n)} \odot \tilde{\mathbf{T}}^{n}_{\text{all}}(\tilde{\mathbf{c}}) \tag{1.15}$$

und Felder

$$\mathbf{E}^{\text{all}}(\mathbf{r}_i) = -\sum_{n=0}^{1} \frac{1}{n!} \mathbf{a}_i^{(n)} \odot \tilde{\mathbf{T}}^{n+1}_{\text{all}}(\tilde{\mathbf{c}}) \tag{1.16}$$

berechnet.

Schließlich werden noch alle Wechselwirkungen, die nach dem Distanzkriterium auch für Strukturelle Einheiten, also auf der niedrigsten Hierarchieebene, nicht per Multipolentwicklung ausgewertet werden können, über die Berechnung von exakten Ausdrücken für die atomaren Paarwechselwirkungen berechnet und aufsummiert.

Die Kombination aus den Ordnungen n und m (SAMM$_{2003}$: $n \leq 2$ und $m \leq 2$) und dem Kriterium, welches entscheidet, ob ein Paar von Clustern C, D im Abstand r wechselwirken kann, bestimmt die Güte der SAMM Approximation.

1.14.4 Kritik an SAMM$_{2003}$

Vergleicht man die Formulierung der Taylorentwicklungen des SAMM$_{2003}$ Potentials und der zugehörigen Felder in Gln. (1.10) und (1.11) mit der von Dehnen [52] vorgeschlagenen Ableitung von FMM Methoden aus zweiseitigen Taylorentwicklungen, so fällt ein wichtiger Unterschied auf.

Während bei Dehnenschen Taylorentwicklungen alle Terme bis zur p-ten Ordnung ($p = 2, 3, 4$) berücksichtigt werden, was als Summenbedingung $p = m + n$ für die Ordnungen

m der Multipolentwicklungen im Quellcluster und n der einseitigen Taylorentwicklungen im Zielcluster ausgedrückt werden kann, fehlt eine derartige Systematik in SAMM$_{2003}$. Hier werden nämlich die Ableitungen n-ter Ordnung der Multipolpotentiale m-ter Ordnung für alle Kombinationen mit $n = 0, 1, 2$ und $m = 0, 1, 2$ berücksichtigt. In den Taylorentwicklungen für die Wechselwirkungen zwischen Paaren von Clustern, welche durch einen Abstand r separiert sind, skalieren die einzelnen Beiträge der Paare (n, m) der Ableitungs- bzw. Multipolordnung dann proportional zu $r^{-(m+n+1)}$.

In SAMM$_{2003}$ werden also bespielsweise zwei Wechselwirkungsterme berücksichtigt, die mit $\sim r^{-4}$ skalieren, nämlich die Dipol-Quadrupolwechselwirkungen der beiden Cluster, zwei weitere Wechselwirkungsterme, die ebenfalls mit $\sim r^{-4}$ skalieren und Monopol-Oktupolwechselwirkungen darstellen, fehlen aber. Wie ich in Unterkapitel 2.1 zeigen werde, verletzten die so berechnete SAMM$_{2003}$ Kräfte das Newtonsche Reaktionsprinzip.

Der Restfehler der berechneten Potentiale und Felder hängt neben den Clusterabständen r auch von den Radien R der wechselwirkenden Cluster ab. So ergeben sich für die Potentiale insgesamt Restfehler $\sim R^{p+1}/r^{-(p+2)}$ und für die Kräfte Restfehler $\sim R^{p}/r^{-(p+2)}$. Im bisherigen Distanzklassenschema streuten nun die Radien der Cluster auf einer Hierarchieebene recht stark. Daher streute auch die für Qualität der Näherung entscheidene scheinbare Clustergröße $2R/r$.

Konsequenter, im Sinne des FMM Konzepts, wäre es, die scheinbaren Clustergrößen unmittelbar zur Entscheidung zu verwenden, ob eine Wechselwirkung auf einer gegebenen höheren Hierarchieebene l oder ob die Cluster in Subcluster aufgelöst werden sollen, deren Wechselwirkungen dann auf tieferliegenden Ebenen berechnet werden. Dieses Konzept habe ich bei meiner, in Unterkapitel 2.2 ausführlich dargestellten, Revision des SAMM Algorithmus realisiert.

Das neue Konzept eröffnet unmittelbar die Möglichkeit, von den rechenaufwändigen exakten Paarwechselwirkungen schon bei im Vergleich zu SAMM$_{2003}$ sehr viel geringeren Abständen zur Näherung durch Clusterwechselwirkungen überzugehen, wenn gleichzeitig auch die Lennard-Jones Wechselwirkungen durch FMM Entwicklungen genähert und nicht, wie früher bei einen Cutoffabstand $R_c \approx 10$ Å abgeschnitten werden. Ein solches Verfahren habe ich realisiert. Dort setzen FMM Näherungen schon bei Abständen von 5,5-7 Å ein. Entsprechend sollte der Rechenaufwand hier auch deutlich kleiner sein als der von Gittersummenverfahren, bei denen atomare Paarwechselwirkungen, ähnlich wie bei der Entfernungsklassenmethode SAMM$_{2003}$, bis zu Abständen von 10-15 Å ausgewertet werden müssen.

Schließlich erwiesen sich im Lauf meiner Arbeit die im SAMM$_{2003}$ unabhängig gewählten FMM-Entwicklungen für Felder und Potentiale als suboptimal, da dann die Felder nicht negative Gradienten der Potentiale sind. Entsprechend können solche Kräfte die SAMM Energie nicht erhalten. In Unterkapitel 2.3 wird gezeigt, wie energierhaltende SAMM Kräfte zu formulieren sind, so dass die Entwicklung eines weiterhin Hamiltonschen und linear skalierenden HADES/SAMM-MD Verfahrens für Proteine im dielektrischen Kontinuum möglich wurde.

1.15 Ziele und Überblick

Das nun folgende Kapitel 2 führt die soeben angesprochenen Lösungen der Probleme des $SAMM_{2003}$ Algorithmus, welche in das PMM-MD Programm IPHIGENIE [53] implementiert wurden und damit der wissenschaftlichen Öffentlichkeit zugänglich sind, genauer aus.

Die in Unterkapitel 2.1 nachgedruckte Veröffentlichung [41] beinhaltet eine effiziente Neuformulierung der Elektrostatikbeschreibung des SAMM Algorithmus. Diese Neuformulierung, welche auf den von Dehnen [52] vorgeschlagenen, ausgewogenen, zweiseitigen Taylorentwicklungen des Potentials und der Felder beruht, birgt den Vorteil, dass in den bis zu einer gegebenen Ordnung p ausgeführten Approximationen der Cluster-Cluster Wechselwirkungen alle Terme bis zu Potenzen $r^{-(p+1)}$ des Clusterabstandes r berücksichtigt werden. Dies ist der Fall, da dort nur die Terme der Ordnung m der Multipolpotentiale und der Ordnung n der Ableitungen dieser Potentiale, für die $p \leq m + n$ gilt, aufsummiert werden. Damit garantiert die Neuformulierung des SAMM Algorithmus das Newtonsche Reaktionsprinzip, was zur Erhaltung des Gesamtimpulses in SAMM-MD Simulationen führt. Zur Steigerung der Genauigkeit wird ferner die Approximation der elektrostatischen Wechselwirkungen bis maximal zur Hexadekapolordnung ($p = 4$) erweitert.

Um eine möglichst effiziente Formulierung des SAMM Algorithmus zu ermöglichen, wurde in der in Unterkapitel 2.2 nachgedruckten Veröffentlichung [42] ein sogenanntes *interaction acceptance criterion* (IAC) entwickelt. Das IAC-Kriterium orientiert sich für ein gegebenes Paar von Clustern an ihren Radien und an empirisch bestimmten Parametern. Diese Parameter hängen von der physiko-chemischen Natur der eingeschlossenen molekularen Bestandteile ab. Sie wandeln damit das IAC Kriterium von einem geometrischen Genauigkeitsmaß in ein Maß um, das auch die unterschiedliche Polarität bzw. Ladung der molekularen Gruppen zur Erzielung einer möglichst gleichmäßigen Approximationsgenauigkeit berücksichtigt.

Die Arbeit zeigt ferner, wie durch die Einbeziehung der Dispersionsinteraktionen in die SAMM Entwicklungen (bis zur maximalen Ordnung $q = 3$), die mit dem früheren Abschneiden verbundenen Artefakte selbst dann vermieden werden können, wenn schon bei sehr viel kleineren Abständen von der exakten Auswertung der Paarwechselwirkung zur SAMM Näherung übergegangen wird.

Anhand von Beispielsimulationen an flüssigem Wasser, das sowohl durch das einfache TIP3P Modell [56] als auch durch das komplexe und polarisierbare Sechspunkt-Wassermodell TL6P [55] beschrieben wird, wird schließlich das lineare Skalierungsverhalten mit der Systemgröße N der resultierenden SAMM Algorithmen nachgewiesen. Damit wird insbesondere gezeigt, dass SAMM sehr gut für die Simulation komplexer polarisierbarer MM Modelle geeignet ist.

Die Vorteile des Übergangs von atomaren Paarwechselwirkungen auf SAMM Näherungen werden in der im Unterkapitel 2.3 nachgedruckten Veröffentlichung [43] auf die HADES-MD Methode [48, 58] für Kontinuumssimulationen übertragen. Zu diesem Zweck werden dort durch Bildung exakter Gradienten der SAMM Energien (Elektrostatik, Lennard-Jones) Hamiltonsche SAMM Kräfte abgeleitet, die, wie gezeigt wird, insbesondere den Gesamtdrehimpuls des im dielektrischen Kontinuum simulierten Proteins erhalten.

Diese Drehimpulserhaltung ist ein besonderes Merkmal des FMM-Algorithmus SAMM. Im

Gegensatz dazu ist die Drehimpulserhaltung bei Schnellen Multipolmethoden, welche auf einer Zerlegung des Simulationssystems durch ein hierarchisches raumfestes Gitter beruhen, ausgeschlossen, da das Gitter die Isotropie des Rauems bricht.

Es wird gezeigt, dass der zusätzliche Rechenaufwand zur Ermittlung der Hamiltonschen SAMM Kräfte gering ist und das Hauptziel, nämlich die Konstruktion eines linear mit der Systemgröße skalierenden HADES/SAMM-MD Algorithmus, erreicht wird. Somit wird demonstriert, dass HADES/SAMM-MD für die Simulation von großen Proteinen im dielektrischen Kontinuum geeignet ist.

Kapitel 3 fasst die Beiträge meiner Koautorschaften an den Veröffentlichungen von Magnus Schwörer und Philipp Tröster zusammen.

Kapitel 4 fasst die Ergebnisse der Arbeit zusammen und gibt einen Ausblick auf anschließende Möglichkeiten zur Weiterentwicklung des SAMM Algorithmus.

2 Der neue SAMM Algorithmus

In den nun folgenden Unterkapiteln 2.1-2.3 sind meine Beiträge zu dem SAMM Algorithmus, bei denen ich die Position des Erstautors einnehme, abgedruckt.

2.1 Effizienzoptimierung durch zweiseitige Taylorentwicklungen

Die nachfolgende Publikation[1]

„Optimizing the Accuracy and Efficiency of Fast Hierarchical Multipole Expansions for MD Simulations“

Konstantin Lorenzen, Magnus Schwörer, Philipp Tröster, Simon Mates, and Paul Tavan

J. Chem. Theory Comput. **8**, 3628-3636 (2012),

die ich zusammen mit Magnus Schwörer, Philipp Tröster, Simon Mates und Paul Tavan verfasst habe, beschreibt die Erweiterung der berücksichtigten Multipolordnungen von $p = 2$ auf $p = 4$ im Kontext einer Neuformulierung hin zu zweiseitigen Taylorentwicklungen in der schnellen Multipolmethode SAMM, welche das Newtonsche Reaktionsprinzip erfüllen.

[1] Mit freundlicher Genehmigung der Verlags

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

Optimizing the Accuracy and Efficiency of Fast Hierarchical Multipole Expansions for MD Simulations

Konstantin Lorenzen, Magnus Schwörer, Philipp Tröster, Simon Mates, and Paul Tavan*

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

Supporting Information

ABSTRACT: Based on p'th order Cartesian Taylor expansions of Coulomb interactions and on hierarchical decompositions of macromolecular simulation systems into hierarchies of nested, structure-adapted, and adaptively formed clusters of increasing size, fast multipole methods are constructed for rapid and accurate calculations of electrostatic interactions. These so-called SAMM$_p$ algorithms are formulated through totally symmetric and traceless tensors describing the multipole moments and the coefficients of local Taylor expansions. Simple recursions for the efficient evaluation and shifting of multipole moments are given. The required tensors are explicitly given up to order $p = 4$. The SAMM$_p$ algorithms are shown to guarantee the reaction principle. For systems with periodic boundaries, a reaction field (RF) correction is applied, which introduces at distances beyond the "minimum image convention" boundary a dielectric continuum surrounding each cluster at the top level of coarse graining. The correctness of the present SAMM$_p$ implementation is demonstrated by analyzing the scaling of the residuals and by checking the numerical accuracy of the reaction principle for a pair of distant molecular ions in vacuum. Molecular dynamics simulations of pure water and aqueous solutions containing artificial ions, which are enclosed by periodic boundaries, demonstrate the stability and low-noise behavior of SAMM$_p$/RF.

1. INTRODUCTION

Molecular dynamics (MD) all-atom simulations of large biomolecular systems[1,2] pose a challenging computational problem mainly due to the long-range nature of electrostatic interactions. If one applies one of the common molecular mechanics (MM) force fields[3−5] for the description of such systems, the electrostatic signatures of the molecular components are specified by partial charges q_i localized at essentially all N atoms i in the simulated system, where N is typically in the range between 10^4−10^6. Because the electrostatic interactions cannot be truncated at the corresponding system sizes L of about 5−20 nm without introducing sizable artifacts,[6−8] the effort for the exact computation of the electrostatic forces scales like N^2 and, thus, rapidly becomes intractable with increasing N. Therefore, various schemes for the approximate and more efficient computation of the electrostatic forces were designed.

The most popular approximation schemes are the so-called lattice sum (LS) methods.[9−11] These methods 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. LS methods enable for PBC systems the approximate computation of the correspondingly periodic electrostatic potential at a computational effort scaling with N log N. On the other hand, the assumption of a periodic electrostatic potential can introduce periodicity artifacts into the description of nonperiodic systems, such as liquids and proteins in solution.[12]

As an alternative, a combination of a fast structure-adapted multipole method (SAMM)[13−15] with a moving boundary reaction field (RF) approach[8] has been suggested, which applies the minimum image convention (MIC)[16] to the explicitly computed electrostatic forces and approximates the electrostatic forces for distances larger than the MIC distance $R_{MIC} = L/2$ through the Kirkwood RF.[17] Thus, this approach adds to toroidal boundary conditions (TBCs)[16] a suitable RF correction and, therefore, avoids the use of an artificially periodic potential as well as corresponding artifacts. Because the SAMM scheme is applied to the electrostatics computation within the MIC sphere surrounding every charge and because its hierarchically nested ternary tree structure is efficiently exploited in a top-down fashion for the generation of the interaction lists at all hierarchy levels, this TBC/RF approach scales linearly with N.[8,15] A corresponding parallelized MD code called EGO has been successfully applied in various large-scale biomolecular simulations (see refs 18 and 19 for recent examples).

The linear scaling achieved by SAMM is typical for such fast multipole methods (FMM),[20] which employ hierarchical decompositions of the total system into a tree of nested subsystems[21,22] of decreasing sizes. Whereas most FMM methods employ regular and nested real-space grids for the construction of this tree (see, e.g., refs 20 and 23−31), SAMM applies a partially adaptive bottom-up clustering of atoms into a nested hierarchy of molecular groups and clusters of such groups, which are formed by neural clustering algorithms.[32,33] A key difference between the real-space grid approaches and the SAMM is the structure of the resulting trees onto which the respective nested hierarchies of charge groups are mapped. The grid approaches employ octal trees, whereas SAMM uses ternary trees, within which the cluster sizes increase much more

Special Issue: Wilfred F. van Gunsteren Festschrift

Received: January 31, 2012
Published: March 21, 2012

slowly from a lower to the next higher level.[8,15] This leads to more efficient but more complex algorithms.

In the past, a variety of other FMM algorithms were suggested for the efficient simulation of periodic biomolecular systems, which usually added LS sum methods to the FMM description of the local electrostatics with the aim of including the periodic images into the computation of the periodic electrostatic potential.[23−25,27,28] Markedly different is the recent combination[34] of a FMM tree code[26] with the isotropic periodic sum approach,[35] which includes the effects of the periodic images in a mean-field sense and, just like SAMM/RF, avoids the generation of an artificially periodic electrostatic potential.

FMM algorithms employ either spherical[20,23−25] or Cartesian[14,26−31,36] coordinates for the computation of the required multipole and Taylor expansions. While the expansions based on spherical harmonics generally were extended to relatively high orders p, the Cartesian methods usually truncate the expansions of the electrostatic potential at lower orders, e.g., $p = 2$ or 3 (quadrupole or octopole, respectively), and choose the depth of the grid hierarchy accordingly for a reasonable compromise between the conflicting aims of accuracy and efficiency. In the case of SAMM, e.g., the multipole and local Taylor expansions were both truncated at second order.[8]

Extending an earlier analysis of Cartesian FMM approaches by Warren and Salmon,[29] Dehnen[31] made in 2002 the important observation that the resulting approximate FMM forces obey Newton's reaction principle, if the multipole and local Taylor expansions of the electrostatic potential are truncated at those levels m and n, respectively, which obey the sum rule $p = m + n$, where p is the highest multipole order considered. Within the thus specified FMM design, the order p defines for the energy of two interacting charges q_i and q_j the approximate magnitude of the expected FMM error through $O[(q_iq_j/r)(d/r)^{p+1}]$. Here, r is the distance between the respective clusters A and B, to which the charges q_i and q_j belong, and d is the typical radius of these clusters.

As an important consequence of this FMM design, the local Taylor expansion of the potential Φ^m, which is generated by a multipole moment of order m, has to be carried out only up to the order $n = p - m$ to reach a desired accuracy (as defined by the highest multipole order p included in the treatment). Accounting for higher order terms $n > p - m$ in this Taylor expansion, like in the previous implementation of SAMM/RF,[8] considerably increases the computational effort without adding a substantial gain of accuracy.

Therefore the quoted result of Dehnen[31] represents a guideline for the construction of Cartesian FMM methods, which not only yield dynamically reasonable interatomic forces but also additionally represent optimal compromises between accuracy and efficiency. In both respects the existing SAMM/RF version,[8] which we will call $SAMM_{2003}$/RF from now on, was clearly suboptimal. Note that $SAMM_{2003}$/RF extended the even simpler predecessor[15] $SAMM_{1997}$, which had been solely applicable to systems under fixed boundary conditions.

Because the molecular polarizability has to be included into MM-MD simulations[2] and into quantum-classical hybrid simulations[37] and because this inclusion requires an enhanced accuracy of the electrostatics calculations, a revision of our parallelized MD code became necessary. In view of the quoted result of Dehnen,[31] we decided to revise $SAMM_{2003}$/RF accordingly. For a most efficient and flexible description of polarization effects we decided to employ atomic polarization dipoles, which are easily integrated into Cartesian FMM approaches.

It is the purpose of this contribution, to sketch the resulting SAMM algorithms which will be called $SAMM_p$, $p = 2−4$, where p is the highest multipole order employed for the Cartesian FMM expansion of the electrostatic potential. We start with a sketch of the theory in a form that closely matches the actual implementation and, therefore, can serve as a guideline for future users of our revised MD code, which will be called IPHIGENIE from now on. Using two simple sample systems, i.e., a pair of the molecular ions H_3O^+ and $H_2PO_4^-$ at varying distances and a periodic box filled with 1500 MM water molecules at ambient temperature and pressure, we study the properties of the new $SAMM_p$ and $SAMM_p$/RF algorithms.

2. THEORY

The revised $SAMM_p$ algorithms announced above, which have been implemented in our parallelized MM-MD simulation program IPHIGENIE, are based on Cartesian multipole and Taylor expansions approximating the electrostatic potentials and fields caused by distant charge distributions.

2.1. FMM from Taylor Expansions. Figure 1 identifies the geometry for the derivation of the $SAMM_p$ approximations to

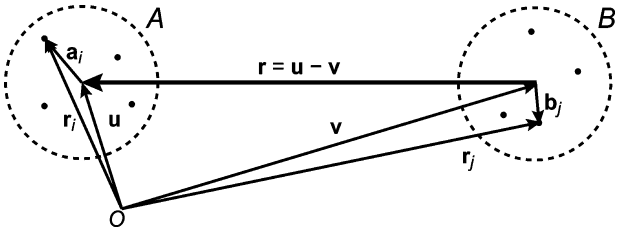


Figure 1. SAMM geometry for two interacting clusters A and B (dashed spheres) of charges $q_i \in A$ and $q_j \in B$ (dots). The interaction of q_i and q_j at $\mathbf{r}_i$ and $\mathbf{r}_j$, respectively, depends on the connecting vector $\mathbf{r}_i - \mathbf{r}_j$ and is evaluated by a Taylor expansion around the vector $\mathbf{r}$ linking the two cluster centers. The positions of these centers are denoted by $\mathbf{u}$ and $\mathbf{v}$, respectively, those of the charges within the respective clusters by $\mathbf{a}_i$ and $\mathbf{b}_j$.

the electrostatic potential and field. The Coulomb potential Φ^B at the position r_i of the charge $q_i \in A$, which is generated by all charges $q_j \in B$, is given (in Gaussian CGS units) by

$$\Phi^B(\mathbf{r}_i) = \sum_{j \in B} \frac{q_j}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{1}$$

For the geometry displayed in Figure 1 this expression can be rewritten as

$$\Phi^B(\mathbf{r}_i) = \sum_{j \in B} \frac{q_j}{|\mathbf{r} + (\mathbf{a}_i - \mathbf{b}_j)|} \tag{2}$$

where we have used the cluster-local coordinates $\mathbf{a}_i = \mathbf{r}_i - \mathbf{u}$ and $\mathbf{b}_j = \mathbf{r}_j - \mathbf{v}$ together with the cluster−cluster connection vector $\mathbf{r} = \mathbf{u} - \mathbf{v}$.

Employing the tensor notation of Warren and Salmon,[29] which is explained in Section 1 of the Supporting Information by providing relevant examples for inner and outer tensor

products, the p'th order Taylor expansion of eq 2 around $\mathbf{r}$ reads

$$\Phi^{B}(\mathbf{r}_i) = \Phi^{B,p}(\mathbf{r}_i) + R^{B,p}(\mathbf{r}_i) \tag{3}$$

with the residual $R^{B,p}(\mathbf{r}_i)$ and the expansion

$$\Phi^{B,p}(\mathbf{r}_i) = \sum_{j\in B} q_j \sum_{n=0}^{p} \frac{1}{n!}\left(\partial_{(n)}\frac{1}{r}\right) \odot (\mathbf{a}_i - \mathbf{b}_j)^{(n)} \tag{4}$$

Here $\partial_{(n)}(1/r)$ is a tensor of rank n composed of the n'th order partial derivatives of $1/r$. Section 2 of the Supporting Information explicitly lists its components for $n \leq 4$. The symbol $\odot$ denotes the inner contraction product of two tensors. The tensor $(\mathbf{a}_i - \mathbf{b}_j)^{(n)}$ of rank n is the n-fold outer product of the vector $\mathbf{a}_i - \mathbf{b}_j$ with itself. In the notation of these outer products $\partial_{(n)}(1/r)$ may be equivalently written as $\nabla_{\mathbf{r}}^{(n)}(1/|\mathbf{r}|)$.

Denoting the outer tensor product by $\otimes$, applying the binomial law

$$(\mathbf{a}_i - \mathbf{b}_j)^{(n)} = \sum_{m=0}^{n} (-1)^m \binom{n}{m} \mathbf{a}_i^{(n-m)} \otimes \mathbf{b}_j^{(m)} \tag{5}$$

to the n-fold outer product in eq 4, sorting the resulting linear combination according to increasing powers of $\mathbf{a}_i$ and exchanging the order of summations, the p'th order Taylor expansion eq 4 becomes

$$\Phi^{B,p}(\mathbf{r}_i) = \sum_{n=0}^{p} \frac{1}{n!}\mathbf{a}_i^{(n)} \odot \partial_{(n)} \sum_{m=0}^{p-n} \Phi^{m,B}(\mathbf{u}) \tag{6}$$

where

$$\Phi^{m,B}(\mathbf{u}) = \frac{(-1)^m}{m!}\left(\partial_{(m)}\frac{1}{r}\right) \odot \sum_{j\in B} q_j \mathbf{b}_j^{(m)} \tag{7}$$

are the potentials $\Phi^{m,B}(\mathbf{u})$ of m'th order multipole moments localized at the reference point $\mathbf{v}$ of cluster B (cf., Figure 1). Equation 6 is one of the basic FMM equations, because it describes a p'th order Taylor expansion of multipole potentials generated by cluster B around the center $\mathbf{u}$ of cluster A.

However, in eq 7 the representation of the m'th order multipole moments through the outer products $\mathbf{b}_j^{(m)}$ is suboptimal, because the associated totally symmetric tensors have $(m + 2)(m + 1)/2$ independent components.[38] A more compact representation is achieved, if one employs the reduced totally symmetric multipole tensors

$$\mathbf{M}^{m,\mathbf{v}} = \sum_{j\in B} q_j (-1)^m b_j^{2m+1}\left(\partial_{(m)}\frac{1}{b_j}\right) \tag{8}$$

which have only $2m + 1$ independent components, because they are traceless with respect to every pair of tensor components.[39,38] Here, the symbols b_j denote the absolute values of the local coordinates $\mathbf{b}_j$ of the charges q_j making up cluster B localized around $\mathbf{v}$. Section 3 of the Supporting Information lists explicit expressions for the components of these multipole moments up to order $m = 4$. With the reduced moments, eq 8, the multipole potentials, eq 7, can be alternatively expressed as

$$\Phi^{m,B}(\mathbf{u}) = \frac{(-2)^m}{(2m)!}\left(\partial_{(m)}\frac{1}{r}\right) \odot \mathbf{M}^{m,\mathbf{v}} \tag{9}$$

The local Taylor expansion eq 6 of the multipole potentials within cluster A can be compactly rewritten by introducing the expansion coefficient tensors

$$\mathbf{T}^{B,n,p}(\mathbf{u}) \equiv \partial_{(n)} \sum_{m=0}^{p-n} \Phi^{m,B}(\mathbf{u}), \qquad n = 0, ..., p \tag{10}$$

which account for the contributions of all multipole moments $\mathbf{M}^{m,\mathbf{v}}$ localized at the center $\mathbf{v}$ of cluster B to a given order n of the local Taylor expansion in cluster A in such a way that the maximal order p of the original Taylor expansion eq 4 is preserved. Correspondingly, all multipole moments $\mathbf{M}^{m,\mathbf{v}}$ up to the maximal order $m = p$ contribute to the zeroth order term of the FMM potential eq 6, whereas only the potential of cluster B's total charge $\mathbf{M}^{0,\mathbf{v}}$ contributes to the p'th order term in eq 6 through its p'th partial derivatives.

We note that the expansion coefficient tensors $\mathbf{T}^{B,n,p}(\mathbf{u})$ have only $2n + 1$ independent components [just like the reduced multipole moment tensors eq 8]. With eq 10 the local Taylor expansion eq 6 of the potential reads

$$\Phi^{B,p}(\mathbf{r}_i) = \sum_{n=0}^{p} \frac{1}{n!}\mathbf{a}_i^{(n)} \odot \mathbf{T}^{B,n,p}(\mathbf{u}) \tag{11}$$

For dynamics simulations one needs the electrostatic forces $\mathbf{f}_i = q_i\mathbf{E}(r_i)$ acting on the charges q_i. FMM offers two ways to calculate the required fields $\mathbf{E}(\mathbf{r}_i)$. One can either use the negative gradient

$$\mathbf{E}^{B,p}(\mathbf{r}_i) = -\nabla_i \Phi^{B,p}(\mathbf{r}_i) \tag{12}$$

of the FMM potential eq 11, which gives the p'th order field $\mathbf{E}^{B,p}(\mathbf{r}_i)$ generated by the charge cluster B. On the other hand we can start with the field

$$\mathbf{E}^{B}(\mathbf{r}_i) = \sum_{j\in B} \frac{q_j(\mathbf{r} + (\mathbf{a}_i - \mathbf{b}_j))}{|\mathbf{r} + (\mathbf{a}_i - \mathbf{b}_j)|^3} \tag{13}$$

associated to the original electrostatic potential eq 1 and apply a FMM Taylor expansion analogous to that in eq 4 but limited to order $p - 1$, i.e.

$$\mathbf{E}^{B,p}(\mathbf{r}_i) = \sum_{j\in B} q_j \sum_{n=0}^{p-1} \frac{1}{n!}\left(\partial_{(n)}\frac{\mathbf{r}}{r^3}\right) \odot (\mathbf{a}_i - \mathbf{b}_j)^{(n)} \tag{14}$$

to find with eqs 8 and 10 the same final expression

$$\mathbf{E}^{B,p}(\mathbf{r}_i) = -\sum_{n=0}^{p-1} \frac{1}{n!}\mathbf{a}_i^{(n)} \odot \mathbf{T}^{B,n+1,p}(\mathbf{u}) \tag{15}$$

Like eq 11 for the potential, this expansion also accounts for all terms up to order $(1/r)^{p+1}$.

2.2. FMM Forces Fulfill the Reaction Principle. As mentioned in the Introduction and pointed out by Dehnen,[30,31] electric fields calculated by eqs 10 and 15 ensure Newton's

third law. This is seen by considering the field, eq 14, at the location $\mathbf{r}_i$ in A for a single generating charge q_j in B, i.e.

$$\mathbf{E}^{j,p}(\mathbf{r}_i) = -q_j \sum_{n=0}^{p-1} \frac{1}{n!}\left(\partial_{(n+1)}\frac{1}{r}\right) \odot (\mathbf{a}_i - \mathbf{b}_j)^{(n)} \tag{16}$$

and the field $\mathbf{E}^{i,p}(\mathbf{r}_j)$ generated by q_i in A at the location $\mathbf{r}_j$ in B. With the geometry in Figure 1 and the vector $\mathbf{r}' \equiv -\mathbf{r}$, this FMM field is

$$\mathbf{E}^{i,p}(\mathbf{r}_j) = -q_i \sum_{n=0}^{p-1} \frac{1}{n!}\left(\partial_{(n+1)}\frac{1}{r'}\right) \odot (\mathbf{b}_j - \mathbf{a}_i)^{(n)} \tag{17}$$

With $\partial_{(n+1)}(1/r') = (-1)^{n+1}\partial_{(n+1)}(1/r)$ and $(\mathbf{b}_j - \mathbf{a}_i)^{(n)} = (-1)^n(\mathbf{a}_i - \mathbf{b}_j)^{(n)}$, one immediately arrives at the reaction principle $\mathbf{f}_{ij} = q_i\mathbf{E}^{j,p}(\mathbf{r}_i) = -q_j\mathbf{E}^{i,p}(\mathbf{r}_j) = -\mathbf{f}_{ji}$ for the pair forces generated by the FMM field eq 15.

2.3. The SAMM Algorithm. A characteristic feature of FMM methods, like SAMM, is the decomposition of the whole system into a hierarchically nested tree of charge clusters. In SAMM, 3−5 partially charged atoms are combined into clusters at the lowest level $l = 0$ of the hierarchy following local chemical motifs.[13] Applying neural clustering algorithms[32,33] (in predefined time intervals, which are large compared to the integration step), on average, 4 of these clusters are combined into compact clusters at the next higher level $l = 1$, and this procedure is repeated until a certain top-level t is reached.[14,15]

Next, interaction lists L_c^l are calculated from distance criteria for each cluster c in every level l of the tree in a top-down fashion.[14] If the system is enclosed by periodic boundaries, only those clusters c' are included into L_c^t, whose surface-to-surface distance $d_{c'c}$ complies with the MIC ($d_{c'c} < R_{\mathrm{MIC}}$) and is larger than a predefined threshold d_t. Here, the top-level t is chosen in such a way that $L_c^t \neq 0$. Clusters c and c' with $d_{c'c} < d_t$ are decomposed into their children. Pairs of these children are included into the interaction lists of level $t-1$, if their distances exceed the threshold $d_{t-1} < d_t$ belonging to this level. This procedure is repeated until the lowest cluster level $l = 0$ is reached. Clusters violating the distance threshold d_0 are finally decomposed into individual atoms, whose interactions are treated with the usual Coulomb expressions. The interaction lists are updated in predefined time intervals. Note that the accuracy of SAMM can be steered by the choice of the distance class boundaries d_l.

2.4. Calculation of Multipole Moments. Using the tree structure in a bottom-up fashion the multipole moments $\mathbf{M}_C^{m,\mathbf{0}}$ of parent clusters C are calculated from those of their children $c(C)$ by a simple iterative procedure:

At level $l = 0$, the multipole moments $\mathbf{M}_c^{m,\mathbf{0}}$ of all clusters c are calculated directly by eq 8 from the partial charges of the embedded atoms taking the origin $\mathbf{0}$ as the reference point. Collecting all clusters $c \in C$, which belong to a parent cluster C at the next higher level, the multipole moments

$$\mathbf{M}_C^{m,\mathbf{0}} = \sum_{c\in C} \mathbf{M}_c^{m,\mathbf{0}} \tag{18}$$

of parent C with respect to $\mathbf{0}$ are simply the sums of the moments $\mathbf{M}_c^{m,\mathbf{0}}$ of its children c. This procedure is repeated until the top-level t is reached.

As is apparent from Figure 1 and eqs 9 and 11, a computation of FMM interactions between two clusters A and B belonging to a given level l requires multipole moments localized at the respective cluster centers $\mathbf{u}$ and $\mathbf{v}$. Thus, at all hierarchy levels the multipole moments $\mathbf{M}^{m,\mathbf{0}}$ of the various clusters have to be shifted from the origin $\mathbf{0}$ to the associated cluster centers $\mathbf{c}$. Using auxiliary tensors $\mathbf{H}_{m,\mathbf{c}}^i$ of rank $i \in \{0, ..., m\}$, which are recursively calculated for $k \in \{0, ..., m-1\}$ through

$$\mathbf{H}_{m,\mathbf{c}}^{k+1} = \mathbf{M}^{k+1,\mathbf{0}} - \frac{(k+1)}{m-k}\hat{S}_{k+1}[(2k+1)(\mathbf{c}\otimes\mathbf{H}_{m,c}^k) - k(\mathbf{c}\odot\mathbf{H}_{m,\mathbf{c}}^k)\otimes\mathbf{I}] \tag{19}$$

from the multipole moments $\mathbf{M}^{k+1,\mathbf{0}}$ at the origin, where the starting point $k = 0$ of the recursion

$$\mathbf{H}_{m,\mathbf{c}}^0 = \mathbf{M}^{0,\mathbf{0}} \tag{20}$$

is the total charge $\mathbf{M}^{0,\mathbf{c}} = \mathbf{M}^{0,\mathbf{0}}$ of the cluster, the shifted multipole tensors

$$\mathbf{M}^{m,\mathbf{c}} = \mathbf{H}_{m,\mathbf{c}}^m \tag{21}$$

are given by the auxiliary tensors of rank m. In eq 19 the operator $\hat{S}_n$ is the symmetrizer

$$\hat{S}_n(A_{1,2,...,n}^n) = \frac{1}{n!}\sum_{p\in\mathcal{I}^n} A_{p(1,2,...,n)}^n \tag{22}$$

for the components $A_{1,2,...,n}^n$ of a tensor of rank n, where $\mathcal{I}^n$ denotes the symmetric group of permutations p for n objects. Note here that the recursion eq 19 enables a sequential and highly efficient evaluation of the multipole moments $\mathbf{M}^{m,\mathbf{c}}$ for $m = 1, ..., p$, in which only the $2(m+1)$ nonredundant tensor components have to be considered at each rank m.

2.5. FMM Interactions in SAMM. Starting at the top-level t, the Taylor expansion coefficients eq 10, which are eventually required to compute the electrostatic fields and potentials at the positions $\mathbf{r}_i$ of the atoms, are computed by descending the levels of the tree. At each level l of the hierarchy, the multipole moments $\mathbf{M}^{m,\mathbf{v}}$ of all clusters B centered at the positions $\mathbf{v}$, which belong to the interaction list L_A^l of a cluster A, contribute up to the rank $m = p - n$ to the tensor $\mathbf{T}^{B,n,p}(\mathbf{u})$ of coefficients in the Taylor expansion eq 11 around the center $\mathbf{u}$ of A. As is illustrated in Figure 2 by a dashed arrow, this action of B on A is inherited to the children $c \in A$ at level $l-1$, i.e., to local Taylor expansions centered at the reference points $\mathbf{c}$, by a shifting operation. The shifting can be carried out without loss of information. The children have then, of course, additional direct contributions to their local Taylor expansions from all

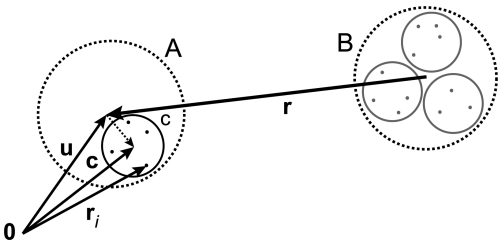


Figure 2. At the parent level l, the potential, which is generated by the charge distribution B, is given at the point r_i by the p'th order Taylor expansion eq 11 around the center $\mathbf{u}$ of cluster A. This potential can be equivalently expressed by a p'th order Taylor expansion around the center $\mathbf{c}$ of the child cluster c using the shifted expansion coefficient eq 24.

clusters in their respective interaction lists L_c^{l-1}, and all these contributions are shifted to the next lower level.

For a proof that the shifting operation sketched above preserves information, one considers the approximate potential $\Phi^{B,p}(\mathbf{r}_i)$ in eq 11, which is generated by a distant charge distribution B and is given as a p'th order Taylor expansion around a center $\mathbf{u}$ with the coefficients $\mathbf{T}^{B,n,p}(\mathbf{u})$ calculated by means of eq 10. Then one replaces in eq 11 the local atomic coordinates $\mathbf{a}_i$, which refer to the cluster center $\mathbf{u}$, by the vectors $\mathbf{d} + \tilde{\mathbf{a}}_i$, where the $\tilde{\mathbf{a}}_i$ is the local atomic coordinate with respect to the new reference point $\mathbf{c}$ and where $\mathbf{d} = \mathbf{c} - \mathbf{u}$ is the translation from $\mathbf{u}$ to $\mathbf{c}$. In Figure 2, which illustrates the geometry, $\mathbf{d}$ is drawn as a dashed arrow. Using, as in eq 5, the binomial law to evaluate the powers $(\mathbf{d} + \tilde{\mathbf{a}}_i)^{(n)}$ appearing in the resulting expression and sorting according to ascending powers of $\tilde{\mathbf{a}}_i$ yields

$$\Phi^{B,p}(\mathbf{r}_i) = \sum_{n=0}^{p} \frac{1}{n!}\tilde{\mathbf{a}}_i^{(n)} \odot \tilde{\mathbf{T}}^{B,n,p}(\mathbf{c}) \tag{23}$$

with the shifted Taylor expansion coefficients

$$\tilde{\mathbf{T}}^{B,n,p}(\mathbf{c}) = \sum_{l=0}^{p-n} \frac{1}{l!}\mathbf{d}^{(l)} \odot \mathbf{T}^{B,l+n,p}(\mathbf{u}) \tag{24}$$

In this way, electrostatic interactions between higher level clusters are inherited to the lowest level, where the resulting Taylor expansions are used to compute the contributions of distant charges to the electrostatic potential and field acting on the individual atoms.

For systems in periodic boundaries, also the top-level clusters inherit electrostatic interactions from a higher level, which is a dielectric continuum starting at the MIC distance R_{MIC} from the center of every top-level cluster and is modeled by the Kirkwood RF[17] as described in ref 8. In such cases the electrostatics treatment will be called SAMM$_p$/RF. We strongly advice readers interested in the concepts and details of this highly efficient combination of FMM with a RF correction, which include, e.g., provisions for smooth transitions of distant top-level clusters into and out of the "Kirkwood sphere" surrounding each of these clusters, to study the original paper.[8]

2.6. Aspects of the Current Implementation. As mentioned in the Introduction, the SAMM$_p$/RF algorithm sketched above has been implemented in a parallelized fashion in the C-program IPHIGENIE for the expansion orders $p = 2-4$. This program represents a thorough revision and extension of an earlier MD simulation code called EGO,[8,15] which employed the FMM approach SAMM$_{2003}$/RF. As opposed to the SAMM$_p$ algorithms, SAMM$_{2003}$ generally violated the reaction principle, because the orders of the multipole and Taylor expansions were not properly balanced. For electrostatic forces originating from charged clusters, the residual error scaled as $1/r^4$, just like that of SAMM$_2$. For neutral clusters, however, the residual of the SAMM$_{2003}$ force calculation scaled as $1/r^5$ (like SAMM$_3$) with a computational effort somewhat larger than that of SAMM$_3$. Note that SAMM$_{1997}$ employed a more balanced combination of multipole and Taylor expansions for the computation of electric fields; therefore, it fulfilled the reaction principle. Like in SAMM$_2$, its residual force error scaled with $1/r^4$ despite a larger computational effort.

Beyond a systematic, balanced, and more accurate combination of multipole and Taylor expansions achieved particularly through SAMM$_4$, the extensions of the code include the use of polarizable force fields, in which the individual atoms can carry inducible dipoles in addition to the static partial charges. Furthermore, they involve the interface[40] to the density functional theory program CPMD,[41] which meanwhile enables fully Hamiltonian MD simulations[42] through the use of Hellmann–Feynman forces (evaluated by FMM at larger distances). For the polarizable degrees of freedom, a separate FMM tree and self-consistent field iterations are available.

However some features are still inherited from its predecessor EGO. For instance, IPHIGENIE still employs the same scheme of SAMM distances d_l ($l = 0, 1, ..., t$), which define the various distance classes and associated cluster levels l, as EGO.[8] This choice is most certainly suboptimal because SAMM$_4$ enables, as we will show now, a much more accurate treatment of larger clusters at even much smaller distances.

3. METHODS

To check the accuracy of the SAMM$_p$ electrostatics calculation and the dynamic stability of SAMM$_p$/RF MD simulations of condensed phase systems subject to toroidal boundary conditions,[16] we chose three model systems. The first consists of the molecular ion pair $H_2PO_4^-\cdots H_3O^+$ in vacuum and the other two of polar and ionic liquids enclosed by periodic boundaries.

3.1. Accuracy Checks. For the SAMM$_p$ accuracy checks we chose, like in Figure 1, two charge clusters A and B separated by a center-to-center distance r. Because the charge distributions should carry multipole moments of all orders, we chose the two nontrivial molecular ions $H_2PO_4^-$ and H_3O^+ as representatives for A and B. Figure 3 illustrates one of the many relative arrangements of these ions.

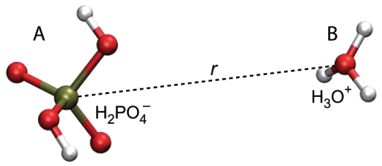


Figure 3. The molecular ions $H_2PO_4^-$ and H_3O^+ at a distance r as representatives for the charge clusters in Figure 1.

Rigid and purely electrostatic MM models were derived for each of these ions by density functional theory (DFT) calculations with the program Gaussian[43] using the B3LYP functional and the 6-31G** basis set. The resulting electrostatic potential (ESP)[44] partial charges and optimized geometries are listed in Section 5 of the Supporting Information. Like in Figure 1, we denote the atomic coordinates of $H_2PO_4^-$ by $\mathbf{r}_i$ and those of H_3O^+ by $\mathbf{r}_j$. Furthermore, the geometric centers $\mathbf{u}$ of $H_2PO_4^-$ and $\mathbf{v}$ of H_3O^+ were chosen as reference points such that the cluster–cluster distance is $r = |\mathbf{u} - \mathbf{v}|$.

The accuracy, by which the SAMM$_p$ algorithms describe the electrostatic interactions, depends on r. To characterize this dependence, r was varied in the range 5–15 Å in steps of 0.2 Å. At each step each of the rigid molecules was randomly rotated and an ensemble $\mathcal{A}$ of 10 000 relative arrangements was generated. For each arrangement $a \in \mathcal{A}$ the total electrostatic energy of the system U and the forces $\mathbf{f}_n$ on the various atoms $n \in A \cup B$ were calculated approximately by SAMM$_p$, $p = 2-4$ and exactly through the Coulomb expressions.

Because the error $O[(1/r)(d/r)^{p+1}]$ of a p'th order FMM expansion does not only depend on the cluster–cluster distance

dx.doi.org/10.1021/ct300080n | J. Chem. Theory Comput. 2012, 8, 3628–3636

r but also on the typical radius d of the interacting clusters, we designed a second test, focusing on the influence of d. Choosing r fixed at 10 Å, the local coordinates $\mathbf{a}_i = \mathbf{r}_i - \mathbf{u}$ and $\mathbf{b}_j = \mathbf{r}_j - \mathbf{v}$ were scaled by factors $g \in [1.0, 2.0]$, which were varied in steps $\Delta g = 0.05$.

For each value of the parameters r and g, which we jointly denote by x, the absolute SAMM$_p$ errors $\xi(x|p)$ of the energies U and forces $\mathbf{f}_n$ were measured by the root-mean-square deviations

$$\xi_U(x|p) = \sqrt{\langle (U - U^p)^2 \rangle_{\mathcal{A}}} \tag{25}$$

and

$$\xi_{\mathbf{f}}(x|p) = \sqrt{\left\langle \frac{1}{3|A \cup B|} \sum_{n \in A \cup B} (\mathbf{f}_n - \mathbf{f}_n^p)^2 \right\rangle_{\mathcal{A}}} \tag{26}$$

where the brackets $\langle ... \rangle_{\mathcal{A}}$ denote the arithmetic mean over the structural ensemble $\mathcal{A}$ at a given value of the respective parameter $x \in \{r,g\}$ and where $(U,\mathbf{f}_n)$ denote the exact quantities and $(U^p,\mathbf{f}_n^p)$ the respective SAMM$_p$ approximations. Relative errors $\rho(x|p)$ are then defined by

$$\rho_U(x|p) = \sqrt{\frac{\langle (U - U^p)^2 \rangle_{\mathcal{A}}}{\langle U^2 \rangle_{\mathcal{A}}}} \tag{27}$$

and

$$\rho_{\mathbf{f}}(x|p) = \sqrt{\frac{\langle \sum_{n \in A \cup B} (\mathbf{f}_n - \mathbf{f}_n^p)^2 \rangle_{\mathcal{A}}}{\langle \sum_{n \in A \cup B} \mathbf{f}_n^2 \rangle_{\mathcal{A}}}} \tag{28}$$

3.2. Dynamics Checks. In SAMM$_p$/RF, the algorithm for electrostatics calculations discontinuously changes with distance switching, e.g., at about $D_0 = 10$ Å from pairwise Coulomb interactions to the level $l = 0$ of the FMM treatment, at a larger distance of about $D_1 = 16$ Å to level $l = 1$, etc., until the computation eventually smoothly switches[8] to the RF description at about R_{MIC}. In a liquid, the diffusive dynamics induces constant distance boundary crossings of clusters. The discontinuities resulting for the associated forces cause algorithmic noise, which heats the simulated system. Note here that the symbols D_l are center-to-center distances of clusters or atoms, whereas the symbols d_l used in Section 2.3 for the construction of the interaction lists are surface-to-surface distances.

To estimate the decrease of algorithmic noise with increasing SAMM order p, we have collected ensembles $\mathcal{T}$ of 200 short ($\Delta t = 10$ ps) MD trajectories at constant volume V, number of molecules N, and total energy E (i.e., in the NVE setting) for two periodic liquid systems ($\mathcal{L}_1$, $\mathcal{L}_2$) and have measured heating rates per molecule

$$\dot{Q}_p = \frac{1}{N} \left\langle \frac{E_p(\Delta t) - E_p(0)}{\Delta t} \right\rangle_{\mathcal{T}} \tag{29}$$

as ensemble averages $\langle ... \rangle_{\mathcal{T}}$ from the total energies $E(t)$ at $t = 0$ and $t = \Delta t$. Here, statistically independent initial conditions had been drawn every 5 ps from 1 ns NVT trajectories of the systems in which a Berendsen thermostat[45] was used for control of the temperature T ($\tau = 100$ fs, $T_0 = 295$ K).

The test system $\mathcal{L}_1$ consisted of 1500 TIP3P[46] water models, which were kept rigid using MSHAKE[47] with a relative tolerance of 10^{-6}. To generate a solution $\mathcal{L}_2$ with a few ions, we changed the partial charges, which are assigned to the oxygen atoms, at two molecules by $+1e$ and at another two by $-1e$. After an initial embedding into cubic boxes with periodic boundaries, $\mathcal{L}_1$ and $\mathcal{L}_2$ were equilibrated by MD for 1 ns in the NpT ensemble using a Berendsen thermostat ($T_0 = 295$ K, $\tau = 100$ fs) and barostat ($p_0 = 1$ atm, $\tau = 5$ ps, $\kappa_T = 4.6 \times 10^{-5}$ atm^{-1}).[45] Equations of motion were integrated by the Verlet algorithm[48] with a time step of 1 fs. For both systems, R_{MIC} was about 18 Å, and the dielectric constant of the surrounding continuum was set to $\varepsilon = 78$. Because the boundary D_1 to the cluster level $l = 1$ was too close[8] to R_{MIC}, only predefined level 0 clusters (i.e., the molecules) were used for SAMM.

Beside the SAMM treatment of the electrostatics, the heating $\dot{Q}_p$ can have further sources like, e.g., the cutoff of the Lennard-Jones interactions at D_0, the disappearance of molecules into and their reappearance from the RF continuum or delayed updates of the interaction lists. To uniquely identify the SAMM contribution to $\dot{Q}_p$, we updated the interaction lists at every integration step and implemented an "exact" reference, which calculates all interactions within R_{MIC} by the Coulomb expressions and treats the RF boundary in exactly the same way as SAMM$_p$/RF. Hence, the reference method mimics SAMM$_p$/RF for $p \to \infty$ and, therefore, is denoted as SAMM$_\infty$/RF. The associated heating rate is called $\dot{Q}_\infty$. The differences $\dot{Q}_p - \dot{Q}_\infty$ then uniquely characterize the contribution of the SAMM$_p$ electrostatics approximation to the algorithmic noise.

4. RESULTS AND DISCUSSION

For a first verification of our implementation, the model calculations with the charge clusters $H_2PO_4^-$ and H_3O^+ depicted in Figure 3 serve us to check whether the residuals $R^{B,p}(\mathbf{r})$ [cf. eq 3] and $-\nabla R^{B,p}(\mathbf{r})$ of the SAMM$_p$ approximations, eqs 11 and eq 15, for the potential and field, respectively, show the expected scaling behaviors with the distance r and with the scaling factor g of the cluster sizes (cf. Section 3.1). From $|R^{B,p}(\mathbf{r})| = O[(1/r)(gd/r)^{p+1}]$ one expects that the scalings are $r^{-(p+2)}$ and g^{p+1}. Similarly, from $|-\nabla R^{B,p}(\mathbf{r})| = O[(1/r)^2(gd/r)^p]$, one gets $r^{-(p+2)}$ and g^p.

4.1. Actual Scalings. The log–log scale chosen in Figure 4 for the presentation of the absolute errors eq 26, which are connected with the SAMM$_p$ force calculations for our sample clusters (cf. Figure 3) at varying $r \in [5, 15$ Å$]$, allows us to extract the exponents of distance-scaling from the slopes of the shown linear regressions. Instead of the expected values

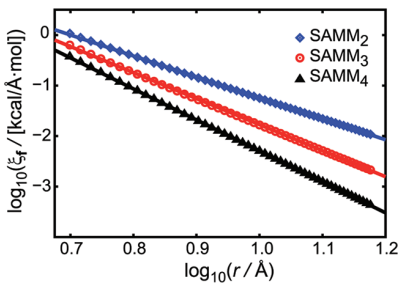


Figure 4. Log–log plot for the absolute errors eq 26 generated by SAMM$_p$ calculations of atomic forces $\mathbf{f}_n$ for the charge clusters $H_2PO_4^-$ and H_3O^+ shown in Figure 3 at cluster–cluster distances $r \in [5, 15$ Å$]$.

dx.doi.org/10.1021/ct300080n | J. Chem. Theory Comput. 2012, 8, 3628–3636

$-(p+2)$, $p = 2-4$, the regressions have slightly different slopes of −4.15, −5.15, and −6.13, respectively. For the absolute errors eq 26, which are associated with the SAMM$_p$ computation of the electrostatic energies U (data not shown), one obtains similar slopes of −4.07, −5.13, and −6.14. The small deviations from the expected values indicate that not only the neglected terms of order $(1/r)^{p+2}$ but also higher order and, thus, more rapidly vanishing terms contribute to the absolute SAMM$_p$ errors in the studied distance range.

Similarly, Figure 5 serves to check whether the observed absolute SAMM$_p$ errors eqs 25 and 26 show the expected

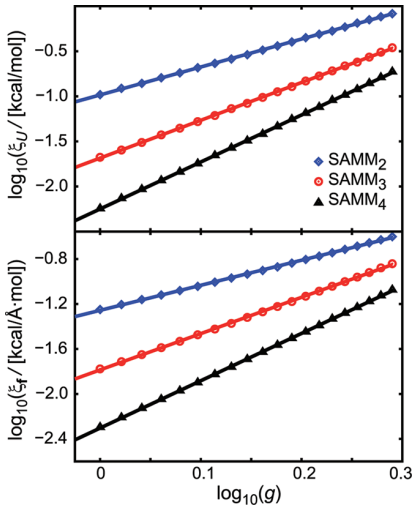


Figure 5. Log−log plots for the absolute errors (top) eq 25 and (bottom) eq 26 of potential energies U and atomic forces $\mathbf{f}_n$, respectively, at a cluster−cluster distance $r = 10.0$ Å when the cluster sizes are scaled by factors $g \in [1.0, 2.0]$; see the caption to Figure 4 for further information.

scaling with the cluster size. For the energy error one expects the slopes $p + 1$, whereas the regressions in Figure 5 (top) yield 3.10, 4.19, and 5.21 for $p = 2-4$, respectively. The force error depends less critically than that of the electrostatic energy on the cluster size. Here one expects that the slopes have the values p and finds from the linear regressions in Figure 5 (bottom) the values 2.23, 3.23, and 4.21, which are again very close to the expectations. As a result, the scaling behavior of our implementation complies with theory strongly suggesting that it is correct (note that we have successfully scrutinized the correctness of the implementation by a series of further tests including, for instance, the check whether Newton's reaction principle is obeyed at numerical accuracy).

However, the absolute SAMM$_p$ errors eqs 25 and 26 considered above are less indicative for the quality of a SAMM$_p$ electrostatics treatment in a condensed phase MD simulation than the relative errors, eq 28 because the relative errors allow us to estimate the relative sizes of the inevitable algorithmic discontinuities at distance class boundaries. In the original SAMM implementation,[15] the first distance class boundary is at $D_0 \approx 10$ Å, where the computation switches from exact Coulomb interactions to a FFM approximation for predefined molecular groups comprising 3−5 atoms. The size of the phosphate ion $H_2PO_4^-$ employed for our error estimates thus represents an upper limit for the typical size of molecular groups used in SAMM at cluster level $l = 0$. Correspondingly, the relative errors eq 28 shown in Figure 6 for our sample clusters at $r = 10$ Å (dashed lines) represent upper limits for the

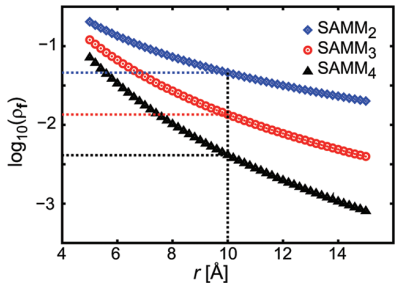


Figure 6. The relative errors eq 28 of the SAMM$_p$ force computations are represented on a logarithmic scale as functions of the cluster−cluster distance r; the dashed lines associated to the distance $r = 10$ Å give upper limits for relative discontinuities of force computation at the distance class boundary $D_0 = 10$ Å.

relative discontinuities of calculated forces, which are encountered during a MD simulation whenever a molecular group crosses D_0.

According to the data presented in Figure 6, the relative discontinuities of force computation at D_0 are smaller than 5, 1.4, and 0.4% for $p = 2-4$, respectively. Here, the relative energy errors eq 27 are about 1 order of magnitude smaller (data not shown). The noted discontinuities lead to the expectation that algorithmic noise will be a serious issue for SAMM$_2$, may be tolerable for SAMM$_3$, and should be small for SAMM$_4$.

In fact, SAMM$_{1997}$ and SAMM$_{2003}$ were plagued with substantial algorithmic noise, whenever the system contained charged clusters, because the residual force error scaled in this case as $1/r^4$ just like that of SAMM$_2$ (cf. Section 2.6). Furthermore, the reaction principle was generally violated in SAMM$_{2003}$. Hence, algorithmic noise should be reduced already when using SAMM$_3$ with its 1.4% relative errors near D_0 and even more at SAMM$_4$. Note that the relative errors of the SAMM$_4$ computation become comparable to those of SAMM$_3$ only at the very small center−center distance $r = 7.5$ Å, which approximately corresponds to a distance of about 1.8 Å between the van der Waals surfaces of the two molecular ions.

4.2. Algorithmic Noise of SAMM$_p$/RF. As explained in Section 3 we have set up two periodic cubic simulation systems filled either with 1500 simple TIP3P water models ($\mathcal{L}_1$) or with a variant, in which four of the water models were artificially converted into ions ($\mathcal{L}_2$). Using eq 29 we have measured method-specific heating rates $\dot{Q}_p$, $p = 2-4, \infty$ from many short *NVT* MD simulations on these systems. Here, $p = \infty$ labels the reference simulations with SAMM$_\infty$/RF, in which the SAMM$_p$ approximations were replaced by the exact Coulomb expressions. Table 1 lists for our sample liquid models $\mathcal{L}_1$ and $\mathcal{L}_2$ the differences between the various $\dot{Q}_p$ and the associated reference heating rates $\dot{Q}_\infty$ together with the statistical standard deviations measured for the SAMM$_p$/RF and SAMM$_\infty$/RF methods.

According to the values shown in Table 1, the heating rates observed in the two systems $\mathcal{L}_1$ and $\mathcal{L}_2$ are statistically indistinguishable at all levels p of approximation. The very crude approximations of SAMM$_2$/RF yield heating rates that are by about 57 kcal/(mol ns) larger than the heating rates $\dot{Q}_\infty \approx 1.4$ kcal/(mol ns) of the reference method SAMM$_\infty$/RF.

dx.doi.org/10.1021/ct300080n | J. Chem. Theory Comput. 2012, 8, 3628−3636

Table 1. SAMM$_p$/RF Heating Rates, $\dot{Q}_p$, $p = 2-4$, Relative to the Reference Rate, $\dot{Q}_\infty$, Evaluated for the Two Test Systems $\mathcal{L}_1$ (polar solvent) and $\mathcal{L}_2$ (ionic solution)[a]

	$\mathcal{L}_1$	$\mathcal{L}_2$
$\dot{Q}_2 - \dot{Q}_\infty$	56.87 ± 0.22	57.36 ± 0.22
$\dot{Q}_3 - \dot{Q}_\infty$	2.67 ± 0.05	2.59 ± 0.05
$\dot{Q}_4 - \dot{Q}_\infty$	0.05 ± 0.04	0.01 ± 0.04

[a]Rates are given per molecule in kcal/(mol ns); for explanations see the text.

This difference is strongly reduced to about 2.6 kcal/(mol ns) by the transition to SAMM$_3$/RF and essentially vanishes for SAMM$_4$/RF, whose heating rates $\dot{Q}_4$ are statistically indistinguishable from the reference rates $\dot{Q}_\infty$ despite the large statistical ensembles of 200 MD trajectories (Δt = 10 ps) spent for the computation of each $\dot{Q}$ in each of the two systems.

On the one hand, the heating rate data largely confirm the expectations derived from the relative force computation errors analyzed in connection with Figure 6. Like the discontinuities of the force computation at the distance class boundary D_0, also the heating rates become strongly suppressed with increasing order p of the SAMM expansions. On the other hand they surprisingly demonstrate that the relative discontinuities (<0.4%) remaining with SAMM$_4$ actually lead to a negligible heating.

The latter finding suggests that with SAMM$_4$/RF the distance class boundaries[8,15] D_l, l = 0, 1, ..., can be chosen considerably smaller without introducing large algorithmic artifacts. Such a change will then cause large efficiency gains, because the numbers of interaction partners will be greatly reduced at each level l of the cluster hierarchy. However, such a change requires that the van der Waals dispersion interaction is additionally included into the FMM scheme of computing nonbonded interactions for the following reason: Currently, the dispersion is truncated at the distance $D_0 \approx 10$ Å, at which the electrostatics computation switches from the Coulomb expressions to the SAMM$_p$ approximation. A choice of a smaller D_0 (e.g., 7 Å), which is compatible with a quite accurate SAMM$_4$ electrostatics computation, would cause with the current code a correspondingly short-range truncation of the dispersion interaction and, therefore, a sizable amount of algorithmic noise.

This noise can be avoided and quite small values D_l can be used as soon as the dispersion will be included into our FMM scheme, which is ongoing work.[26,36] Its completion will subsequently enable systematic studies of how the conflicting aims of accuracy and efficiency can be optimally reached with SAMM$_4$/RF. Already at the present stage of the implementation SAMM$_4$/RF is, for the sample systems $\mathcal{L}_1$ and $\mathcal{L}_2$, only by 24% less efficient than SAMM$_3$/RF, whereas the extremely crude SAMM$_2$/RF is only by 17% more efficient than SAMM$_3$/RF. As a result, increasing the order of p to p = 4 should be capable of shifting the compromise between accuracy and efficiency to a higher level.

Finally we would like to stress that the use of the Kirkwood RF correction[8] in connection with SAMM$_p$ is necessary for low-noise MD simulations. If one switches off the RF correction (e.g., by choosing the dielectric constant ε = 1 for the continuum surrounding each top-level cluster beyond R_{MIC}, thereby[8] implementing a smooth electrostatics cutoff at R_{MIC}), then the heating rates become very large for all p (data not shown). On the other hand, one observes an effective cooling for our sample systems with SAMM$_4$/RF and SAMM$_\infty$/RF characterized, e.g., by a rate $\dot{Q}_4 = -3.11$ kcal/(mol ns), if interaction lists are updated only every 64 time steps. As we have checked by switching off the electrostatics, this cooling is caused by the fact that particles are always diffusively spreading within the Lennard-Jones cutoff spheres and, due to the rare updates, even beyond ,while the new particles entering these spheres are identified only with a delay. As a result, deceleration by the dispersion attraction on average dominates the acceleration. Therefore, rare interaction list updates imply a "dispersion cooling".

4.3. Summary. We have presented a careful revision of the SAMM/RF algorithm.[8] This revision has been designed for rapid and accurate MD simulations of periodic condensed phase systems. IPHIGENIE, the new SAMM$_p$/RF implementation, employs systematic p'th order Cartesian FMM expansions, which, by construction, guarantee the reaction principle and scale linearly with the system size. Through the use of the Cartesian, totally symmetric and traceless multipole tensors eq 8 featuring at rank m only $2m + 1$ independent components and through the use of the coefficient tensors eq 10, which have analogous properties, the computational and storage costs are kept minimal. In particular, the shifting of the multipole moment tensors to new reference points can be effected through the efficient recursion, eq 19. As a corollary, this recursion reduces to an algorithm for the sequential computation of rank m multipole tensors for distributions composed of point charges, point dipoles, etc., as is immediately clear if one assumes that all these electrostatic point objects are initially located at the origin (cf. the Supporting Information).

SAMM$_4$ yields very accurate forces and electrostatic energies showing the theoretically expected scaling of the residuals. With SAMM$_4$/RF the algorithmic noise turned out to be negligible even for a relatively small sample simulation system with an inner radius R_i of only 18 Å. Because previous and much less accurate implementations of SAMM/RF showed a strong reduction of algorithmic noise with increasing R_i, we expect further improvements also for SAMM$_4$/RF at larger R_i.

The computational scenario outlined above is easily generalized toward the efficient treatment of polarizable force fields and of DFT/MM hybrid simulations (ongoing work). On the other hand, not all options for algorithmic optimizations have been exhausted so far such that there is, as always, ample room for further optimizations. Several of these issues have been identified in this work like, e.g., the choice of smaller distance classes with SAMM$_4$ and the inclusion of the dispersion into the FMM scheme, the use of polarizable force fields (PFF) in this setting, or the efficient electrostatics computation in fully Hamiltonian DFT/PFF simulations.

■ ASSOCIATED CONTENT

Supporting Information

Explanation of the tensorial notation, explicit expressions and recursion relations for the n'th derivatives of $1/r$, for the m'th order totally symmetric and traceless multipole moments and for the electrostatic potentials generated by these moments for $n, m \leq 4$. Furthermore parameters for the charge distributions $H_2PO_4^-$ and H_3O^+ are given. 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). Like all our work it has been guided by the desire to approach with our own contributions those standards, which have been set by Wilfred and his group in the field of macromolecular simulation through decades of outstanding scientific efforts and achievements.

REFERENCES

(1) 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.

(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) 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.

(4) Ponder, J.; Case, D. *Adv. Protein Chem.* **2003**, *66*, 27−85.

(5) Oostenbrink, C.; Villa, A.; Mark, A.; Van Gunsteren, W. *J. Comput. Chem. B* **2004**, *25*, 1656−1676.

(6) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *102*, 5451−5459.

(7) Hünenberger, P. H.; van Gunsteren, W. F. *J. Chem. Phys.* **1998**, *108*, 6117−6134.

(8) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847−10860.

(9) Darden, T. A.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(10) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577−8593.

(11) Luty, B. A.; Tironi, I. G.; van Gunsteren, W. F. *J. Chem. Phys.* **1995**, *103*, 3014−3021.

(12) Hünenberger, P. H.; McCammon, J. A. *Biophys. Chem.* **1999**, *78*, 69−88.

(13) Niedermeier, C.; Tavan, P. *J. Chem. Phys.* **1994**, *101*, 734−748.

(14) Niedermeier, C.; Tavan, P. *Mol. Simul.* **1996**, *17*, 57−66.

(15) Eichinger, M.; Grubmüller, H.; Heller, H.; Tavan, P. *J. Comput. Chem.* **1997**, *18*, 1729−1749.

(16) Allen, M. P.; Tildesley, D. *Computer Simulations of Liquids*; Clarendon: Oxford, U.K., 1987.

(17) Kirkwood, J. G. *J. Chem. Phys.* **1934**, *2*, 351−361.

(18) Lingenheil, M.; Denschlag, R.; Tavan, P. *Eur. Biophys. J.* **2010**, *39*, 1177−1192.

(19) Denschlag, R.; Schreier, W. J.; Rieff, B.; Schrader, T. E.; Koller, F. O.; Moroder, L.; Zinth, W.; Tavan, P. *Phys. Chem. Chem. Phys.* **2010**, *12*, 6204−6218.

(20) Greengard, L.; Rokhlin, V. *J. Comput. Phys.* **1987**, *73*, 325 −348.

(21) Appel, A. A. *SIAM J. Sci. Stat. Comput.* **1985**, *6*, 85−103.

(22) Barnes, J.; Hut, P. *Nature* **1986**, *324*, 446−449.

(23) Figueirido, F.; Levy, R. M.; Zhuo, R.; Berne, B. J. *J. Chem. Phys.* **1997**, *106*, 9835−9849.

(24) Challacombe, M.; White, C.; Head-Gordon, M. *J. Chem. Phys.* **1997**, *107*, 10131−10139.

(25) Amisaki, T. *J. Comput. Chem.* **2000**, *21*, 1075−1087.

(26) Ding, H.-Q.; Karasawa, N.; Goddard, W. A. III *J. Chem. Phys.* **1992**, *97*, 4309−4315.

(27) Ding, H.-Q.; Karasawa, N.; Goddard, W. A. III *Chem. Phys. Lett.* **1992**, *196*, 6−10.

(28) Shimada, J.; Kaneko, H.; Takada, T. *J. Comput. Chem.* **1994**, *15*, 28−43.

(29) Warren, M. S.; Salmon, J. K. *Comput. Phys. Commun.* **1995**, *87*, 266 −290.

(30) Dehnen, W. *Astrophys. J.* **2000**, *536*, L39 −L42.

(31) Dehnen, W. *J. Comput. Phys.* **2002**, *179*, 27 −42.

(32) Martinetz, T.; Berkovich, S.; Schulten, K. *IEEE Trans. Neural Networks* **1993**, *4*, 558−569.

(33) Dersch, D. R.; Tavan, P. In Proceedings of the International Conference on Artificial Neural Networks 1994, (ICANN'94), , Sorrento, Italy, May 26-29, 1994; Moreno, M., Morasso, P., Eds.; Springer: London, 1994; pp 1067−1070.

(34) Takahahi, K. Z.; Narumi, T.; Yasuoka, K. *J. Chem. Phys.* **2011**, *135*, 174108.

(35) Wu, X.; Brooks, B. R. *J. Chem. Phys.* **2005**, *122*, 044107.

(36) Shanker, B.; Huang, H. *J. Comput. Phys.* **2007**, *226*, 732−753.

(37) Schmitz, M.; Tavan, P. In *Modern methods for theoretical physical chemistry of biopolymers*; Tanaka, S., Lewis, J., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Chapter 8, pp 157−177.

(38) Hinsen, K.; Felderhof, B. U. *J. Math. Phys.* **1992**, *33*, 3731−3735.

(39) Buckingham, A. D. *Adv. Chem. Phys.* **1967**, *12*, 107−147.

(40) Eichinger, M.; Tavan, P.; Hutter, J.; Parrinello, M. *J. Chem. Phys.* **1999**, *110*, 10452−10467.

(41) Hutter, J.; Alavi, A.; Deutsch, T.; Bernasconi, M.; Goedecker, S.; Marx, D.; Tuckerman, M.; Parrinello, M. *CPMD: Car−Parinello Molecular Dynamics*, version 3.10; IBM Corporation and Max-Planck Institut, Stuttgart: Armonk, NY and Stuttgart, Germany, 1997; www.cpmd.org.

(42) Laio, A.; VandeVondele, J.; Rothlisberger, U. *J. Chem. Phys.* **2002**, *116*, 6941−6947.

(43) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

(44) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129−145.

(45) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684−3690.

(46) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926−935.

(47) Kräutler, V.; van Gunsteren, W. F.; Hünenberger, P. H. *J. Comput. Chem.* **2001**, *22*, 501−508.

(48) Verlet, L. *Phys. Rev.* **1967**, *159*, 98−103.

Der folgende Abdruck[2]

„Supplementary Information for
Optimizing the accuracy and efficiency of fast hierarchical
multipole expansions for MD simulations“

Konstantin Lorenzen, Magnus Schwörer, Philipp Tröster, Simon Mates,
and Paul Tavan
J. Chem. Theory Comput. **8**, 3628-3636 (2012)

enthält zusätzliche Informationen zu Herleitung, Implementierung und Validierung der entwickelten Theorie.

[2] Mit freundlicher Genehmigung des Verlags.

Supplementary Information for Optimizing the accuracy and efficiency of fast hierarchical multipole expansions for MD simulations

Konstantin Lorenzen, Magnus Schwörer, Philipp Tröster, Simon Mates, and Paul Tavan*

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

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

*To whom correspondence should be addressed

1 Inner and outer tensor products

The main text employs the inner ($\odot$) and outer ($\otimes$) tensor products for a most compact notation of tensorial entities. To exemplify the meaning of these products we consider totally symmetric tensors $\mathbf{A}^m$ and $\mathbf{B}^n$ of ranks m and n respectively. Choosing $m = 1$ and $n = 3$ and denoting the Cartesian components of these tensors by A^1_α and $B^3_{\beta\gamma\delta}$ with $\alpha, \beta, \gamma, \delta \in \{x, y, z\}$, the inner product $\mathbf{A}^m \odot \mathbf{B}^n$ is a tensor of rank $|n - m| = 2$ given by

$$\left(\mathbf{A}^1 \odot \mathbf{B}^3\right)_{\gamma\delta} = \sum_{\alpha\in\{x,y,z\}} A^1_\alpha B^3_{\alpha\gamma\delta}$$

whereas the outer product $\mathbf{A}^m \otimes \mathbf{B}^n$ is a tensor of rank $m + n = 4$ given by

$$\left(\mathbf{A}^1 \otimes \mathbf{B}^3\right)_{\alpha\beta\gamma\delta} = A^1_\alpha B^3_{\beta\gamma\delta}.$$

This outer product becomes a symmetric tensor upon applying the symmetrization operator $\hat{S}_4$ defined by Eq. (22), i.e.,

$$\left[\hat{S}_4 \left(\mathbf{A}^1 \otimes \mathbf{B}^3\right)\right]_{\alpha\beta\gamma\delta} = \frac{1}{4!} \sum_{\mathfrak{p}\in\mathscr{S}^4} \left(\mathbf{A}^1 \otimes \mathbf{B}^3\right)_{\mathfrak{p}(\alpha\beta\gamma\delta)}.$$

2 Components of the tensors $\partial_{(n)}(1/r)$ for $n \leq 4$

The tensors $\partial_{(n)}(1/r)$ of rank $n = 1, \ldots, p$ can be calculated from the following recursion

$$\partial_{(n)} \frac{1}{r} = \frac{-1}{r^2} \hat{S}_n \left[(2n-1) \left(\mathbf{r} \otimes \partial_{(n-1)} \frac{1}{r} \right) - (n-1) \left(\mathbf{r} \odot \partial_{(n-1)} \frac{1}{r} \right) \otimes \mathbf{I} \right], \tag{30}$$

with $\partial_{(0)}(1/r) = 1/r$ and where $\hat{S}_n$ is the symmetrizer given in Eq. (22).

For $n = 1,2,3,4$ the Cartesian components $\alpha, \beta, \gamma, \varepsilon \in \{x, y, z\}$ of the tensors $\partial_{(n)}(1/r)$ are given explicitly by

$$\left(\partial_{(1)} \frac{1}{r} \right)_{\alpha} = \frac{-1}{r^3} r_\alpha, \tag{31}$$

$$\left(\partial_{(2)} \frac{1}{r} \right)_{\alpha\beta} = \frac{1}{r^5} \left(3 r_\alpha r_\beta - r^2 \delta_{\alpha\beta} \right), \tag{32}$$

$$\left(\partial_{(3)} \frac{1}{r} \right)_{\alpha\beta\gamma} = \frac{-3}{r^7} \left[5 r_\alpha r_\beta r_\gamma - r^2 \left(r_\alpha \delta_{\beta\gamma} + r_\beta \delta_{\gamma\alpha} + r_\gamma \delta_{\alpha\beta} \right) \right], \tag{33}$$

$$\left(\partial_{(4)} \frac{1}{r} \right)_{\alpha\beta\gamma\varepsilon} = \frac{3}{r^9} \left[\begin{array}{l} 35 r_\alpha r_\beta r_\gamma r_\varepsilon - \\ 5r^2 \left(\begin{array}{l} r_\alpha r_\varepsilon \delta_{\beta\gamma} + r_\beta r_\varepsilon \delta_{\gamma\alpha} + r_\gamma r_\varepsilon \delta_{\alpha\beta} + \\ r_\alpha r_\gamma \delta_{\beta\varepsilon} + r_\beta r_\gamma \delta_{\alpha\varepsilon} + r_\alpha r_\beta \delta_{\gamma\varepsilon} \end{array} \right) \\ + r^4 \left(\delta_{\alpha\beta} \delta_{\gamma\varepsilon} + \delta_{\alpha\gamma} \delta_{\beta\varepsilon} + \delta_{\alpha\varepsilon} \delta_{\beta\gamma} \right) \end{array} \right]. \tag{34}$$

3 Components of the multipole moment tensors $\mathbf{M}^{m,\mathbf{0}}$ for $m \leq 4$

Taking the origin $\mathbf{0}$ of a global coordinate system as the reference point, the totally symmetric and traceless multipole tensors $\mathbf{M}^{m,\mathbf{0}}$ of rank $m = 1,\ldots,p$ can be calculated for a distribution of point charges q_j at positions $\mathbf{r}_j$ from the recursion

$$\mathbf{M}^{m,\mathbf{0}} = \sum_j \hat{S}_m \left[(2m-1) \left(\mathbf{r}_j \otimes \mathbf{M}^{m-1,\mathbf{0}} \right) - (m-1) \left(\mathbf{r}_j \odot \mathbf{M}^{m-1,\mathbf{0}} \right) \otimes \mathbf{I} \right], \tag{35}$$

where the lowest moment is the total charge

$$M^{0,\mathbf{0}} = \sum_{j \in B} q_j. \tag{36}$$

For $m = 1,2,3,4$ the Cartesian components of the tensors $\mathbf{M}^{m,\mathbf{0}}$ are given explicitly by

$$M_\alpha^{1,\mathbf{0}} = \sum_j q_j r_{j\alpha} \tag{37}$$

$$M_{\alpha\beta}^{2,\mathbf{0}} = \sum_j q_j \left(3 r_{j\alpha} r_{j\beta} - r_j^2 \delta_{\alpha\beta} \right) \tag{38}$$

$$M_{\alpha\beta\gamma}^{3,\mathbf{0}} = 3 \sum_j q_j \left[5 r_{j\alpha} r_{j\beta} r_{j\gamma} - r_j^2 \left(r_{j\alpha}\delta_{\beta\gamma} + r_{j\beta}\delta_{\gamma\alpha} + r_{j\gamma}\delta_{\alpha\beta} \right) \right] \tag{39}$$

$$M_{\alpha\beta\gamma\varepsilon}^{4,\mathbf{0}} = 3 \sum_j q_j \left[\begin{array}{l} 35 r_{j\alpha} r_{j\beta} r_{j\gamma} r_{j\varepsilon} \\ -5 r_j^2 \left(\begin{array}{l} r_{j\alpha} r_{j\varepsilon} \delta_{\beta\gamma} + r_{j\beta} r_{j\varepsilon} \delta_{\gamma\alpha} + r_{j\gamma} r_{j\varepsilon} \delta_{\alpha\beta} + \\ r_{j\beta} r_{j\gamma} \delta_{\alpha\varepsilon} + r_{j\alpha} r_{j\gamma} \delta_{\beta\varepsilon} + r_{j\alpha} r_{j\beta} \delta_{\gamma\varepsilon} \end{array} \right) \\ +r_j^4 \left(\delta_{\alpha\varepsilon}\delta_{\beta\gamma} + \delta_{\beta\varepsilon}\delta_{\gamma\alpha} + \delta_{\gamma\varepsilon}\delta_{\alpha\beta} \right) \end{array} \right]. \tag{40}$$

4 Potentials of the multipole moments $\mathbf{M}^{m,\mathbf{v}}$ for $m \leq 4$

The multipole potentials Eq. (9) are given in terms of the reduced totally symmetric and traceless multipole tensors $\mathbf{M}^{m,\mathbf{v}}$ defined by Eq. (8), which have only $2m+1$ independent components. Using the notation

$$\mathbf{r} = \begin{pmatrix} r_x \\ r_y \\ r_z \end{pmatrix} \equiv \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

for the position vector $\mathbf{r}$ in Figure 1, here we list explicit expressions of these multipole potentials for $m \leq 4$.

$$\Phi^{0,B}(\mathbf{u}) = \frac{1}{r} M^{0,\mathbf{v}} \tag{41}$$

$$\Phi^{1,B}(\mathbf{u}) = \frac{1}{r^3}\left(xM_x^{1,\mathbf{v}} + yM_y^{1,\mathbf{v}} + zM_z^{1,\mathbf{v}}\right) \tag{42}$$

$$\Phi^{2,B}(\mathbf{u}) = \frac{1}{2}\frac{1}{r^5}\begin{bmatrix} \left(x^2-z^2\right)M_{xx}^{2,\mathbf{v}} + \left(y^2-z^2\right)M_{yy}^{2,\mathbf{v}} \\ +xyM_{xy}^{2,\mathbf{v}} + xzM_{xz}^{2,\mathbf{v}} + yzM_{yz}^{2,\mathbf{v}} \end{bmatrix} \tag{43}$$

$$\Phi^{3,B}(\mathbf{u}) = \frac{1}{6}\frac{1}{r^7}\begin{bmatrix} \left(3x^2-y^2\right)yM_{xxy}^{3,\mathbf{v}} + \left(3x^2-z^2\right)zM_{xxz}^{3,\mathbf{v}} + \\ \left(3y^2-x^2\right)xM_{yyx}^{3,\mathbf{v}} + \left(3y^2-z^2\right)zM_{yyz}^{3,\mathbf{v}} + \\ \left(3z^2-x^2\right)xM_{zzx}^{3,\mathbf{v}} + \left(3z^2-y^2\right)yM_{zzy}^{3,\mathbf{v}} + \\ 6xyzM_{xyz}^{3,\mathbf{v}} \end{bmatrix} \tag{44}$$

$$\Phi^{4,B}(\mathbf{u}) = \frac{1}{24}\frac{1}{r^9}\begin{bmatrix} \left(6x^2y^2-x^4-y^4\right)M_{xxyy}^{4,\mathbf{v}} + \left(6x^2z^2-x^4-z^4\right)M_{xxzz}^{4,\mathbf{v}} + \\ \left(6y^2z^2-y^4-z^4\right)M_{yyzz}^{4,\mathbf{v}} + \\ 4xy\left\{\left(x^2-3z^2\right)M_{xxxy}^{4,\mathbf{v}} + \left(y^2-3z^2\right)M_{xyyy}^{4,\mathbf{v}}\right\} + \\ 4xz\left\{\left(x^2-3y^2\right)M_{xxxz}^{4,\mathbf{v}} + \left(z^2-3y^2\right)M_{xzzz}^{4,\mathbf{v}}\right\} + \\ 4yz\left\{\left(y^2-3x^2\right)M_{yyyz}^{4,\mathbf{v}} + \left(z^2-3x^2\right)M_{yzzz}^{4,\mathbf{v}}\right\} \end{bmatrix} \tag{45}$$

5 The charge distribution models $H_2PO_4^-$ and H_3O^+

The following table lists the ESP[1] partial charges and atomic Cartesian coordinates of the molecular ions $H_2PO_4^-$ and H_3O^+, which were calculated by DFT with the program GAUSSIAN[2] using the B3LYP functional and a 6-31G** basis set. The thus defined charge distributions were used for our SAMM$_p$ accuracy checks.

Table 2: Atomic coordinates and partial charges q of the $H_2PO_4^-$ and H_3O^+ charge clusters.

Molecule	Atom	x[Å]	y[Å]	z[Å]	$q[e]$
$H_2PO_4^-$	P	0.000002	0.000244	-0.165372	0.996988
	O_1	-0.868651	-1.012979	-0.859174	-0.720624
	O_2	0.868643	1.015482	-0.856230	-0.720317
	O_3	1.010060	-0.804788	0.907120	-0.661509
	O_4	-1.010064	0.802149	0.909440	-0.660392
	H_1	-1.833510	0.301761	0.836155	0.382994
	H_2	1.833575	-0.304320	0.835177	0.382859
H_3O^+	O	0.077246	-0.100080	0.071652	-0.563729
	H_1	-0.054307	0.082040	1.027628	0.521424
	H_2	0.994622	0.069795	-0.234847	0.521223
	H_3	-0.272665	-0.973108	-0.210744	0.521083

References

(1) Singh, U. C.; Kollman, P. A. *J. Comput. Chem.* **1984**, *5*, 129–145.

(2) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03, Revision C.02*; Gaussian, Inc.: Wallingford, CT, 2004.

2.2 Einbettung der Lennard-Jones Dispersion in SAMM und ein neues Akzeptanzkriterium für SAMM-Wechselwirkungen

Die nachfolgende Publikation[3]

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

Konstantin Lorenzen, Christoph Wichmann and Paul Tavan
J. Chem. Theory Comput. **10**, 3244-3259 (2014),

die ich zusammen mit Christoph Wichmann und Paul Tavan verfasst habe, beinhaltet die Einbettung der $\sim 1/r^6$ Lennard-Jones Dispersionswechselwirkung in die schnelle Multipolmethode SAMM. Zusätzlich wird die Entwicklung des Akzeptanzkriteriums *interaction accaptance criterion* (IAC) beschrieben. Dieses Kriterium, das entscheidet, ob ein gegebenes Paar von Clustern über schnelle Multipolmethoden ausgewertet werden darf, orientiert sich an einer Vorhersage der zu erwartenden absoluten Kraftfehler einer solchen Cluster-Cluster Wechselwirkung. Schließlich werden die Effizienzgewinne dargestellt, die durch diese beiden Neuerungen erreicht werden, und es wird das lineare Skalierungsverhalten des Rechenaufwandes demonstriert.

[3]Mit freundlicher Genehmigung der Verlags

JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

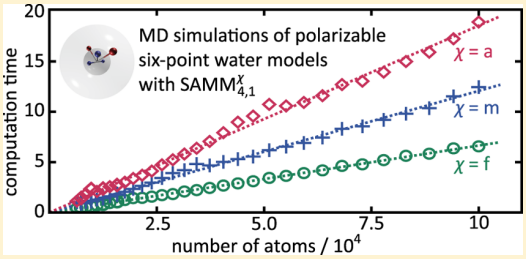
Including the Dispersion Attraction into Structure-Adapted Fast Multipole Expansions for MD Simulations

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

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.

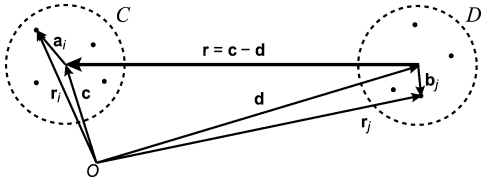


Figure 1. FMM geometry for two interacting clusters C and D (dashed spheres) of atoms $i \in C$ at $\mathbf{r}_i$ and $j \in D$ at $\mathbf{r}_j$ (dots) carrying dispersion charges B_i and B_j. The dispersion interaction of i and j depends on the connecting vector $\mathbf{r}_i - \mathbf{r}_j$ and is 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

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

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

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

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

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 $0 < l \leq \lambda$.

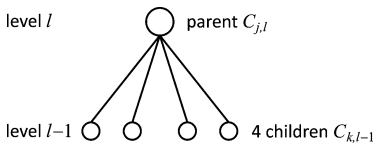


Figure 2. Local splitting motif typical for SAMM$_{p,q}$. A parent cluster $C_{j,l}$ is split into four children $C_{k,l-1}$ at an intermediate level $0 < l \leq \lambda$ within a quaternary tree, which represents the SAMM$_{p,q}$ decomposition of a simulation system into a nested hierarchy of compact 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 $|C_{j,0}| = 3, 4, \ldots, 16$ 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

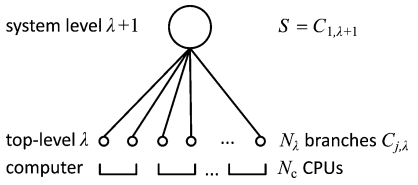


Figure 3. Top level: System $S = C_{1,\lambda+1}$ split into N_λ top-level clusters $C_{j,\lambda}$, the so-called branches, which form the roots of quaternary trees extending toward the lower levels $0 < l < \lambda$ and hierarchically decomposing the $C_{j,\lambda}$. Identical numbers N_λ/N_c of branches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

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

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.

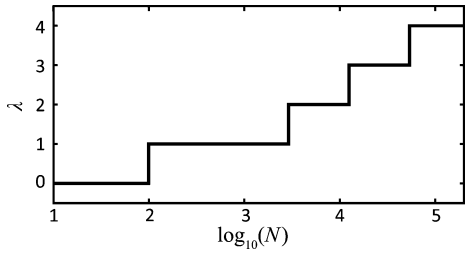


Figure 4. Height λ of the SAMM$_{p,q}$ hierarchy (cf. eq 21) for systems with N atoms.

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

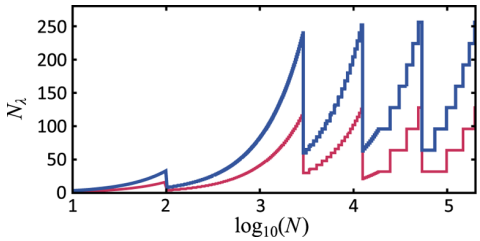


Figure 5. Number N_λ of top-level clusters $C_{j,\lambda}$ as a function of the atom number N for N_c parallel CPUs (cf. eq 22) and for SUs, which comprise three (blue) or six (red) atoms.

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 code-book $\mathcal{W} \equiv \{\mathbf{w}_r | r = 1, ..., M\} \in \mathbb{R}^d$ 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

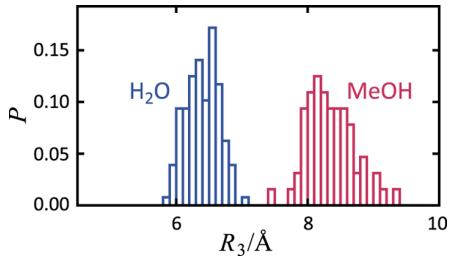


Figure 6. Normalized histograms $P(R_3)$ of radii R_3 of gyration, which characterize the top-level ($\lambda = 3$) clusters calculated for two homogeneous liquid model systems with $N \approx 26\,000$ atoms. Blue: water, $N_0 = 8737$, $N_3 = 128$; red: methanol, $N_0 = 4275$, $N_3 = 64$.

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 $C_{j,l}$, 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 $[(\alpha,b) = (p,\mathrm{e})]$. 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 $\mathbf{H}^0_{m,\mathbf{c}} \equiv \mathbf{M}^{0,\mathbf{0}}$, 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 $\langle \vartheta_i(r) \rangle_{i,j} = \vartheta_T(r)$, 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 $[(\alpha,b) = (p,e)]$ or the dispersive $[(\alpha,b) = (q,d)]$ 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 $\Delta f^{(\alpha)}_{X,b}(r)$ 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]

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

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 $\mathrm{SAMM}^{f}_{4,1}$/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 $\mathrm{SAMM}_{p,q}$/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	$P = (p, q, c_r)$	$P_{\mathrm{ref}} = (p, q, c_r)_{\mathrm{ref}}$
$c_r = d_X$	(∞, ∞, d_X)	$(\infty, \infty, d_{\mathrm{MIC}})$
$q = -1$	$(\infty, -1, d_X)$	(∞, ∞, d_X)
$q = 3$	$(4, 3, d_X)$	$(4, \infty, d_X)$
$q = 2$	$(4, 2, d_X)$	$(4, \infty, d_X)$
$q = 1$	$(4, 1, d_X)$	$(4, \infty, d_X)$
$p = 4$	$(4, 3, d_X)$	$(\infty, 3, d_X)$
$p = 3$	$(3, 3, d_X)$	$(\infty, 3, d_X)$
$\mathrm{SAMM}_{4,3}$	$(4, 3, d_X)$	$(\infty, \infty, d_{\mathrm{MIC}})$
$\mathrm{SAMM}_{4,2}$	$(4, 2, d_X)$	$(\infty, \infty, d_{\mathrm{MIC}})$
$\mathrm{SAMM}_{4,1}$	$(4, 1, d_X)$	$(\infty, \infty, d_{\mathrm{MIC}})$

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

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

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

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 $\mathrm{SAMM}_{p,q}$ 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 $\lim_{p\to\infty} \mathrm{SAMM}_{p,q}$/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 $d_{\mathrm{MIC}} - 2R_X$ (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 $d_{\mathrm{MIC}} - 2R_X$ 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 $\Delta\dot{Q}(\Theta, P, P_{\mathrm{ref}})$ 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 $\mathrm{SAMM}_{4,q}$ 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

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

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

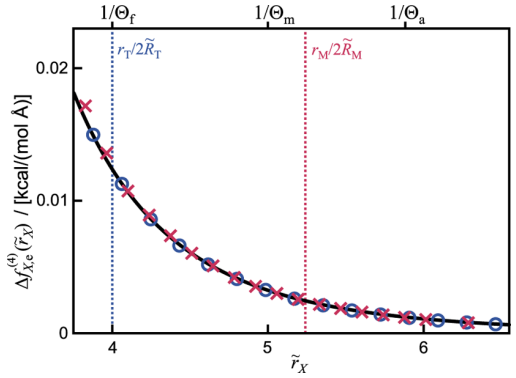


Figure 7. Empirical force errors $\Delta f^{(4)}_{X,e}(\tilde{r}_X)$ calculated by means of eq 35 are compared on the scale of the dimensionless distances $\tilde{r}_X$ for TIP3P water (blue circles) and MeOH (red crosses) with the predictions of master formula eq 43 (black) expressing the error estimates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

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

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.

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

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

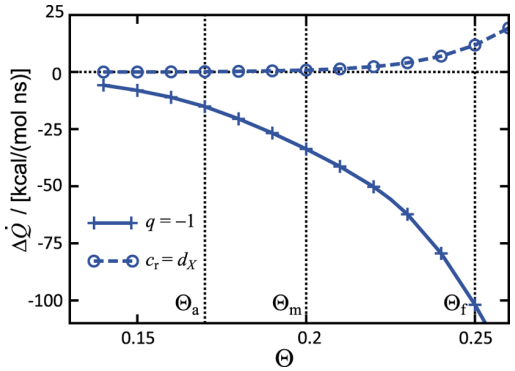


Figure 8. Contributions $\Delta\dot{Q}$ of the dispersion cutoff (solid line) and Pauli repulsion cutoff (dashed line) at $d_T(\Theta)$ to the total heating rate $\dot{Q}$ per molecule in system $\mathcal{T}$ as functions of the IAC threshold Θ. The parameters of the displayed heating rate differences $\Delta\dot{Q}$ 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

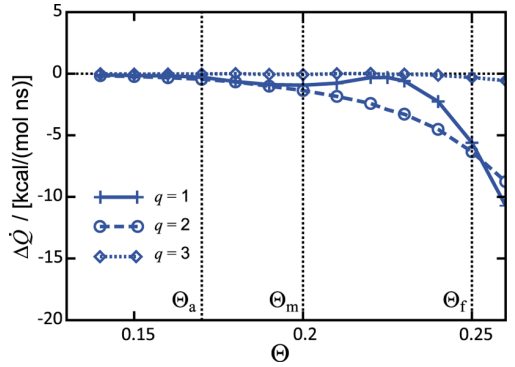


Figure 9. Cooling rates $\Delta\dot{Q}$ remaining in MD simulations of the system $\mathcal{T}$, if the dispersion cutoff is replaced by a SAMM$_q$ dispersion expansion of order order $q = 1$ (solid line), $q = 2$ (dashed line), or $q = 3$ (dotted line) for varying IAC thresholds Θ. For explanation,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

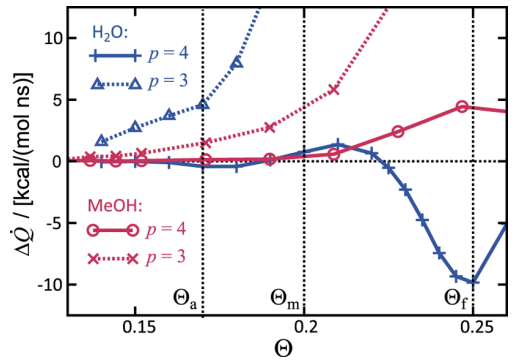


Figure 10. Heating and cooling caused by the SAMM$_p$ electrostatics expansions in the systems $\mathcal{T}$ (blue) and $\mathcal{M}$ (red) as functions of the IAC threshold Θ. The $\Delta\dot{Q}$ data represent those contributions to the total heating rate $\dot{Q}$ per molecule, which are caused by 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 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".

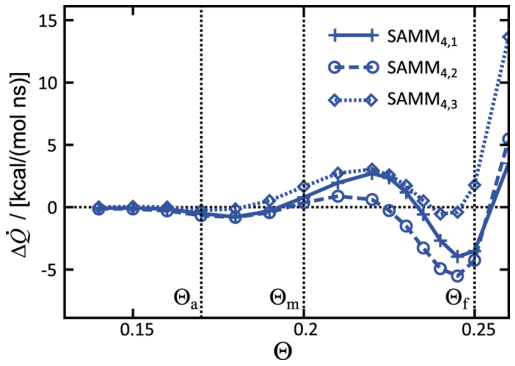


Figure 11. $SAMM_{4,q}$ algorithmic noise for $q = 1$ (solid line), $q = 2$ (dashed line), and $q = 3$ (dotted line) as a function of the IAC threshold Θ measured in the system $\mathcal{T}$ by the heating rate differences $\Delta\dot{Q}$, which are defined by the entries $SAMM_{4,q}$ in Table 2

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.

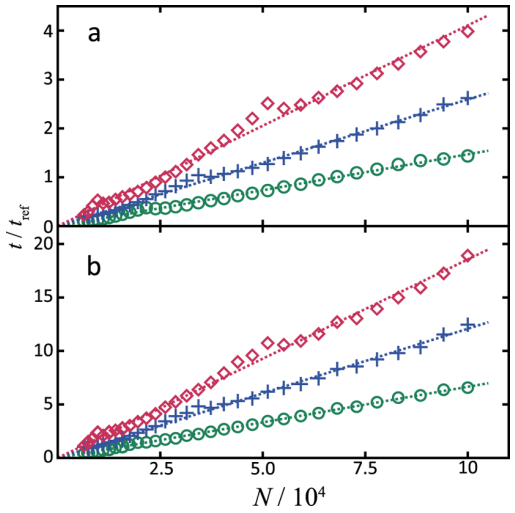


Figure 12. Relative computation times t/t_{ref} measured by applying SAMM$^{\chi}_{4,1}$/RF to (a) the TIP3P systems $\mathcal{T}_i$ and (b) the TL6P systems $\mathcal{B}_i$ with the accuracy/efficiency choices χ = a (red), χ = m (blue), and χ = f (green). Also shown are corresponding regression lines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

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 $[r \le d_X(\Theta)]$ 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 $d_X(\Theta)$ 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 $\mathbf{M}^{m,\mathbf{0}}$ 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 $\Delta f^{(4)}_{X,e}(\tilde{r}_X)$ at various distances $\tilde{r}_X$ 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 ($x = \mathbf{a}$) and geometric ($x = \mathbf{g}$) 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 $B_{i,j}^{\mathbf{a}}$ and $B_{i,j}^{\mathbf{g}}$. 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.

Der folgende Abdruck[4]

Supporting Information to:

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

Konstantin Lorenzen, Christoph Wichmann and Paul Tavan

J. Chem. Theory Comput. **10**, 3244-3259 (2014)

enthält zusätzliche Informationen zur Implementierung der $\sim 1/r^6$ Dispersion sowie zur Parametrisierung des IAC. Außerdem sind die Grundlagen der Implementierung des Clustering-Verfahrens beschrieben, welches in IPHIGENIE zur Erstellung einer geschachtelten, hierarchischen Zerlegung des Simulationssystems in kompakte atomare Cluster implementiert ist.

[4] Mit freundlicher Genehmigung des Verlags.

Supporting Information to:

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

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

*To whom correspondence should be addressed

S1 Components of the Tensors $\partial_{(n)}\left(1/r^6\right)$ for $n \leq 3$

For $n = 1, 2, 3$ the Cartesian components $\alpha, \beta, \gamma, \in \{x, y, z\}$ of the tensors $\partial_{(n)}\,(1/r^6)$ are given explicitly by

$$\left(\partial_{(1)}\frac{1}{r^6}\right)_{\alpha} = \frac{-6}{r^8}\, r_\alpha \tag{S47}$$

$$\left(\partial_{(2)}\frac{1}{r^6}\right)_{\alpha\beta} = \frac{6}{r^{10}}\left(8\, r_\alpha r_\beta - r^2\delta_{\alpha\beta}\right) \tag{S48}$$

$$\left(\partial_{(3)}\frac{1}{r^6}\right)_{\alpha\beta\gamma} = \frac{-48}{r^{12}}\left[10\, r_\alpha r_\beta r_\gamma - r^2\left(\delta_{\alpha\beta}r_\gamma + \delta_{\beta\gamma}r_\alpha + \delta_{\gamma\alpha}r_\beta\right)\right] \tag{S49}$$

S2 Components of the Tensors $\mathbf{M}^{m,\mathbf{0}}$ for $m \leq 3$

Taking the origin $\mathbf{0}$ of a global coordinate system as the reference point, the totally symmetric dispersion multipole tensors $\mathbf{M}^{m,\mathbf{0}}$ of rank $m = 1, \ldots, 3$ can be calculated from Eq. (10) for a distribution D of dispersion charges B_j at the positions $\mathbf{r}_j$. This calculation requires the knowledge of the m'th order derivatives $\partial_{(m)}(1/b_j^6)$, for which Section S1 lists explicit expressions (with $r \equiv b_j$).

For $m = 0, 1, 2, 3$ the Cartesian components of these tensors $\mathbf{M}^{m,\mathbf{0}}$ are explicitly given by

$$M^{0,\mathbf{0}} = \sum_j B_j \tag{S50}$$

$$M_\alpha^{1,\mathbf{0}} = \sum_j B_j\, 6\, r_{j,\alpha} \tag{S51}$$

$$M_{\alpha\beta}^{2,\mathbf{0}} = \sum_j B_j\, 6\left(8\, r_{j,\alpha}r_{j,\beta} - r_j^2\,\delta_{\alpha\beta}\right) \tag{S52}$$

$$M_{\alpha\beta\gamma}^{3,\mathbf{0}} = \sum_j B_j\, 48\left[10\, r_{j,\alpha}r_{j,\beta}r_{j,\gamma} - r_j^2\left(\delta_{\alpha\beta}r_{j,\gamma} + \delta_{\beta\gamma}r_{j,\alpha} + \delta_{\gamma\alpha}r_{j,\beta}\right)\right] \tag{S53}$$

S3 Potentials $\phi^{m,D}(\mathbf{c})$ for $m \leq 3$

The dispersion multipole potentials (9) at the center $\mathbf{c}$ of a cluster C are given in terms of the totally symmetric multipole tensors $\mathbf{M}^{m,\mathbf{d}}$ of a cluster D defined by Eq. (10). Using the notation $\mathbf{r} = (r_x, r_y, r_z)^T \equiv (x, y, z)^T$ for the vector $\mathbf{r}$ connecting in Figure 1 the centers of the clusters D and C leads for $m \leq 3$ to the following explicit expressions of these multipole potentials.

$$\phi^{0,D}(\mathbf{c}) = \frac{1}{r^6}\,\mathbf{M}^{0,\mathbf{d}} \tag{S54}$$

$$\phi^{1,D}(\mathbf{c}) = \frac{1}{r^8}\,\left(x\,\mathbf{M}^{1,\mathbf{d}}_x + y\,\mathbf{M}^{1,\mathbf{d}}_y + z\,\mathbf{M}^{1,\mathbf{d}}_z\right) \tag{S55}$$

$$\phi^{2,D}(\mathbf{c}) = \frac{1}{2}\frac{1}{r^{10}}\begin{pmatrix} x^2\,\mathbf{M}^{2,\mathbf{d}}_{xx} + y^2\,\mathbf{M}^{2,\mathbf{d}}_{yy} + z^2\,\mathbf{M}^{2,\mathbf{d}}_{zz} \\ 2\,x\,y\,\mathbf{M}^{2,\mathbf{d}}_{xy} + 2\,y\,z\,\mathbf{M}^{2,\mathbf{d}}_{yz} + 2\,x\,z\,\mathbf{M}^{2,\mathbf{d}}_{xz} \end{pmatrix} \tag{S56}$$

$$\phi^{3,D}(\mathbf{c}) = \frac{1}{6}\frac{1}{r^{12}}\begin{pmatrix} x^3\,\mathbf{M}^{3,\mathbf{d}}_{xxx} + y^3\,\mathbf{M}^{3,\mathbf{d}}_{yyy} + z^3\,\mathbf{M}^{3,\mathbf{d}}_{zzz} + \\ 3\,x^2\,y\,\mathbf{M}^{3,\mathbf{d}}_{xxy} + 3\,x^2\,z\,\mathbf{M}^{3,\mathbf{d}}_{xxz} + 3\,y^2\,x\,\mathbf{M}^{3,\mathbf{d}}_{yyx} + \\ 3\,y^2\,z\,\mathbf{M}^{3,\mathbf{d}}_{yyz} + 3\,z^2\,x\,\mathbf{M}^{3,\mathbf{d}}_{zzx} + 3\,z^2\,y\,\mathbf{M}^{3,\mathbf{d}}_{zzy} + \\ 6\,x\,y\,z\,\mathbf{M}^{3,\mathbf{d}}_{xyz} \end{pmatrix} \tag{S57}$$

S4 Clustering

For all levels $0 < l \leq \lambda$, the nested cluster hierarchy described in Section 3 of the main text is generated by the neural clustering algorithm suggested by Martinetz et al.,[1] which will be described in Sec. S4.2. The lowest level $l = 0$ is made up of so-called structural units (SUs), which are predefined chemical motifs occurring in protein-solvent systems.

S4.1 List of Units

Tables S3 and S4 characterize SUs occurring in proteins and solvents by CHARMM-type[2] names of the enclosed heavy atoms. The backbone SUs are the amide groups (AG) linking consecutive amino acids. Names of protein SUs X derive from CHARMM-type residue names with appended "P" marking protonated and "N" neutral states.

Table S3: Protein SUs, Bounded Accuracy Corrections a_X, and Radii R_X of Gyration.

SU X	a_X	R_X[Å]	atoms	SU X	a_X	R_X[Å]	atoms
			Amide Groups				
AG	1.35	1.429	C O +N +C_α	AGGly	1.36	1.470	C O +N +C_α
AGPro	1.34	1.342	C O +N +C_α				
			Residues				
Ala	1.80	0.909	C_β	Arg_1	1.80	1.523	C_β C_γ C_δ
Asn	1.43	1.572	C_β C_γ $O_{\delta 1}$ $N_{\delta 2}$	Arg_2	1.32	1.672	N_ϵ C_ζ $N_{\eta 1}$ $N_{\eta 2}$
Asp	1.15	1.394	C_β C_γ $O_{\delta 1}$ $O_{\delta 2}$	Cys	1.82	1.418	C_β S_γ
Gln	1.60	1.867	C_β C_γ C_δ $O_{\epsilon 1}$ $N_{\epsilon 2}$	Glu	1.20	1.627	C_β C_γ C_δ $O_{\epsilon 1}$ $O_{\epsilon 2}$
$His_{\delta 1}$	1.80	1.499	C_β C_γ $C_{\delta 2}$	$His_{\epsilon 1}$	1.80	1.499	C_β C_γ $C_{\delta 2}$
$His_{\delta 2}$	1.16	1.264	$N_{\delta 1}$ $C_{\epsilon 1}$ $N_{\epsilon 2}$	$His_{\epsilon 2}$	1.16	1.264	$N_{\delta 1}$ $C_{\epsilon 1}$ $N_{\epsilon 2}$
Ile	1.80	1.859	C_β $C_{\gamma 1}$ $C_{\gamma 2}$ C_δ	Leu	1.80	1.799	C_β C_γ $C_{\delta 1}$ $C_{\delta 2}$
Lys	1.56	2.189	C_β C_γ C_δ C_ϵ N_ζ	Met	1.80	1.905	C_β S_δ C_γ C_ϵ
Phe_1	1.80	1.686	C_β C_γ $C_{\delta 1}$ $C_{\delta 2}$	Pro	1.80	1.631	C_β C_γ C_δ
Phe_2	1.80	1.538	$C_{\epsilon 1}$ $C_{\epsilon 2}$ C_ζ	Ser	1.28	1.134	C_β O_γ
Thr	1.45	1.574	C_β $O_{\gamma 1}$ $C_{\gamma 2}$	Val	1.80	1.661	C_β $C_{\gamma 1}$ $C_{\gamma 2}$
Trp_1	1.66	1.918	C_β C_γ $C_{\delta 1}$ $N_{\epsilon 1}$	Tyr_1	1.80	1.689	C_β C_γ $C_{\delta 1}$ $C_{\delta 2}$
Trp_2	1.80	1.863	$C_{\delta 2}$ $C_{\epsilon 2}$ $C_{\epsilon 3}$ $C_{\zeta 2}$ $C_{\zeta 3}$ $C_{\eta 2}$	Tyr_2	1.67	1.624	$C_{\epsilon 1}$ $C_{\epsilon 2}$ C_ζ O_η
			Residues with Alternative Protonation/Redox States				
AspP	1.35	1.537	C_β C_γ $O_{\delta 1}$ $O_{\delta 2}$	GluP	1.44	1.786	C_β C_γ C_δ $O_{\epsilon 1}$ $O_{\epsilon 2}$
$HisP_1$	1.80	1.499	C_β C_γ $C_{\delta 2}$	DiSu	1.80	1.132	C_β S_γ
$HisP_2$	1.28	1.468	$N_{\delta 1}$ $C_{\epsilon 1}$ $N_{\epsilon 2}$	LysN	1.73	2.153	C_β C_γ C_δ C_ϵ N_ζ
			N-termini				
Nt	1.49	1.208	N C_α	NtPro	1.80	1.815	N C_α C_β C_γ C_δ
			Neutral N-termini				
NtN	1.44	1.175	N C_α	Ace	1.78	0.831	$C_{\alpha Y}$
			C-terminus				
Ct	1.00	0.899	C Ot_1 Ot_2				
			Neutral C-termini				
CtN	1.03	0.929	C Ot_1 Ot_2	Ct_1	1.20	1.152	C Ot_1 Ot_2 Ct
Ct_2	1.10	1.090	C O Nt	Ct_3	1.21	1.213	C O Nt Ct_α

Table S4: Solvent/Ion SU Names, Accuracy Corrections a_X, and Radii R_X of Gyration.

SU X	a_X	R_X[Å]	atoms	SU X	a_X	R_X[Å]	atoms
			Solvents				
TIP3P[3]	1.00	0.677	O	TL6P[4]	1.01	0.701	O_{pol}
MeOH	1.31	1.198	C_β Oγ	DMSO[5]	1.53	1.463	S O Me_1 Me_2
			Small Ions				
H_3O^+[6]	1.01	0.828	O	$H_2PO_4^-$[6]	1.23	1.603	P $O_{H,1}$ $O_{H,2}$ O_1 O_2

S4.2 The Vector Quantization Algorithm by Martinetz et al.[1]

The purpose of this neural clustering algorithm[1] is a so-called vector quantization[7] (VQ), which is a density-oriented representation of a d-dimensional data set

$$\mathcal{X} \equiv \{\mathbf{x}_n \,|\, n = 1, \ldots, N\} \in \mathbb{R}^d$$

of N data vectors $\mathbf{x}_n$ by a much smaller so-called codebook

$$\mathcal{W} \equiv \{\mathbf{w}_r \,|\, r = 1, \ldots, M\} \in \mathbb{R}^d$$

comprising $M \ll N$ codebook vectors $\mathbf{w}_r$. Here, density-orientation means that the distribution $p(\mathbf{w})$ characterizing the statistics of the codebook $\mathcal{W}$ closely resembles the distribution $p(\mathbf{x})$, from which the data set $\mathcal{X}$ is drawn.[1] The advantage of the Martinetz algorithm over related neural clustering algorithms[8,9] is its safer and faster convergence, which is bought by a slightly enhanced computational complexity of $N \times M^\alpha$ with $1 < \alpha < 2$, as compared to $\alpha = 1$ applicable to the related algorithms.

Starting with an initial codebook $\mathcal{W}$ which may be, e.g., randomly drawn from $\mathcal{X}$, a sequential and stochastic learning process is executed aiming at the optimization of $\mathcal{W}$. Here a vector $\mathbf{x}$ is randomly chosen from $\mathcal{X}$ and all squared distances

$$\forall_{r=1}^{M} \qquad d_r^2(\mathbf{x}) \equiv (\mathbf{x} - \mathbf{w}_r)^2 \tag{S58}$$

to the codebook vectors $\mathbf{w}_r$ are calculated. Then the square distances $d_r^2(\mathbf{x})$ are ordered according to increasing size, i.e., the codebook indices r are mapped to ranking numbers $k_r(\mathbf{x} \,|\, \mathcal{W}) \in \{0, 1, \ldots, M-1\}$ such that

$$\forall_{r=1}^{M} \forall_{r'=1}^{M} \qquad \left[d_r^2(\mathbf{x}) \leq d_{r'}^2(\mathbf{x}) \quad \Rightarrow \quad k_r(\mathbf{x} \,|\, \mathcal{W}) \leq k_{r'}(\mathbf{x} \,|\, \mathcal{W})\right]. \tag{S59}$$

Then the codebook vectors are shifted toward $\mathbf{x}$ according to the learning rule

$$\mathbf{w}_r^{\text{new}} = \mathbf{w}_r + \varepsilon\, a_r\,(\mathbf{x} \,|\, \mathcal{W}, \kappa)\,(\mathbf{x} - \mathbf{w}_r)\,. \tag{S60}$$

Here the overall magnitude of the shifts depends on the size of the so-called learning parameter $0 < \varepsilon \leq 1$. Furthermore the selection, how many and which codebook vectors $\mathbf{w}_r$ experience sizable shifts, depends on the size of the so-called activation function

$$a_r\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right) \equiv \exp\left[-\frac{k_r\left(\mathbf{x} \,|\, \mathcal{W}\right)}{\kappa}\right] \tag{S61}$$

whose range on the ranking scale is given by the so-called cooperativeness parameter $\kappa > 0$.

Independently of κ one always has $a_r\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right) = 1$ for the codebook vector $\mathbf{w}_r$ with the smallest $d_r^2(\mathbf{x})$, which gets shifted by $\varepsilon\left(\mathbf{x} - \mathbf{w}_r\right)$ toward $\mathbf{x}$. Values $\kappa << 1$ signify the case of maximally competitive and local learning, because then $a_{r'}\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right) \approx 0$ for all other $\mathbf{w}_{r'}$. For a large cooperativeness $\kappa = (M-1)/2$, in contrast, all activations are sizable $[e^{-2} < a_r\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right) \leq 1]$ and, therefore, all $\mathbf{w}_r$ participate with sizable weights $\varepsilon a_r\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right)$ in the learning process (S60).

For small values of the parameters κ and ε one can significantly speed up the learning by introducing the finite size cutoff

$$a_r\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right) < \vartheta/\varepsilon \quad \Rightarrow \quad a_r\left(\mathbf{x} \,|\, \mathcal{W}, \kappa\right) \equiv 0 \tag{S62}$$

with a threshold $0 < \vartheta < \varepsilon$ for the learning weights. A comparison with Eq. (S61) now shows that only those codebook vectors $\mathbf{w}_r$ with ranking numbers $k_r(\mathbf{x} \,|\, \mathcal{W})$ smaller or equal to

$$k_{\max}\left(\kappa, \varepsilon\right) = \min\left(M-1, \left\lfloor \kappa \ln \frac{\varepsilon}{\vartheta} \right\rfloor\right) \tag{S63}$$

participate in learning. Therefore the sorting of the square distances $d_r^2(\mathbf{x})$ as expressed by Eq. (S59) can be stopped as soon as the $k_{\max} + 1$ smallest values are determined and are ordered by size. For values $\kappa = 0.5$ and $\vartheta = 0.0005 < \varepsilon = 0.06$, which indicate slow ($\varepsilon << 1$) and competitive ($\kappa << M/2$) learning one gets $k_{\max} = 2$, such that only three codebook vectors participate in a learning step (S60).

The learning process sketched above is embedded into a so-called annealing schedule,[9] during which the parameters $(\varepsilon,\, \kappa)$ are reduced within $A \in \mathbb{N}$ steps $\alpha = 1,\, 2,\, \ldots,\, A$, from

large initial values ε_1 and κ_1 to small final values ε_A and κ_A. A good choice is the exponential annealing defined by

$$\epsilon_\alpha = \epsilon_1 \left(\epsilon_A/\epsilon_1\right)^{(\alpha-1)/(A-1)} \quad \text{and} \quad \kappa_\alpha = \kappa_1 \left(\kappa_A/\kappa_1\right)^{(\alpha-1)/(A-1)} . \tag{S64}$$

This schedule rapidly switches from an initial fast and cooperative learning to a slow and competitive adaptation.

As we have seen, the learning rule (S60) reduces at end of the annealing process ($\alpha = A$, $\kappa_A << 1$) and for the selected winner $\mathbf{w}_r$ to the simple form $\mathbf{w}_r^{\text{new}} = (1 - \varepsilon_A)\mathbf{w}_r + \varepsilon_A \mathbf{x}$, which is the sequential computation of a moving average. This moving average is defined by a normalized memory kernel, which exponentially decays on the scale $1/\varepsilon_A$ toward the past. Hence, as soon as one of the codebook vectors has seen about $3/\varepsilon_A$ data points, the moving average should be well relaxed toward the stationary target average. If one executes in the last annealing step A a total of $B_A \in \mathbb{N}$ learning iterations over the data set $\mathcal{X}$, then every codebook vector will see $B_A N/M$ data points for learning, and this number should be equal to $3/\varepsilon_A$. This consideration fixes the number B_A of iterations over $\mathcal{X}$, which is required in the annealing step $\alpha = A$ for approximate convergence, to $B_A = \max\left(1, \lfloor 3M/\varepsilon_A N \rfloor\right)$.

At earlier stages of the annealing, learning is cooperative and the size $1/\varepsilon_\alpha$ of the memory becomes smaller. Therefore less data points have to be presented and the number of iterations over the data set may be chosen as

$$B_\alpha = \max\left(1, \left\lfloor \frac{3M}{\varepsilon_A N} \frac{\alpha - 1}{A - 1} \right\rfloor\right) . \tag{S65}$$

As a result, at each annealing step there is at least one iteration over the data $\mathcal{X}$. Typically, B_α becomes larger than one toward the end of the annealing ($\alpha \to A$ and ε_α) and for small compression rates N/M of the data $\mathcal{X}$ by the codebook $\mathcal{W}$.

The algorithmic considerations expressed by Eqs. (S58)-(S65) are combined into the Algorithm 1, which is implemented in our MD program package IPHIGENIE.

Algorithm 1 Vector Quantization in IPHIGENIE

$\varepsilon_\alpha \leftarrow \varepsilon_1$
$\kappa_\alpha \leftarrow \kappa_1$
$\varepsilon_{\text{fac}} \leftarrow (\varepsilon_A/\varepsilon_1)^{1/(A-1)}$ ▷ exponential annealing Eq. (S64)
$\kappa_{\text{fac}} \leftarrow (\kappa_A/\kappa_1)^{1/(A-1)}$
$\vartheta \leftarrow 0.0005$
for $\alpha \leftarrow 1$ **to** A **do** ▷ loop over A annealing steps α
  $k_\alpha^{\max} = \min\left[M-1, \lfloor \kappa_\alpha \ln(\varepsilon_\alpha/\vartheta) \rfloor\right]$ ▷ Eq. (S63), # of learning $\mathbf{w}_r$
  $\kappa_\alpha^{\text{eff}} \leftarrow \kappa_\alpha$ ▷ set cooperativeness range $\kappa_\alpha^{\text{eff}}$
  if $M < 6$ **then** ▷ reduce it for quaternary clusters
    $\kappa_\alpha^{\text{eff}} \leftarrow \kappa_\alpha/2$
  end if
  for $k \leftarrow 0$ **to** $k_\alpha^{\max}$ **do** ▷ compute all required activations
    $a[k] \leftarrow \varepsilon_\alpha \exp\left(-k/\kappa_\alpha^{\text{eff}}\right)$ ▷ Eq. (S61)$\times\varepsilon_\alpha$
  end for
  $B_\alpha \leftarrow \max\left(1, \lfloor 3M(\alpha-1)/\varepsilon_A N(A-1) \rfloor\right)$ ▷ Eq. (S65), # of iterations over $\mathcal{X}$
  for $\beta \leftarrow 1$ **to** B_α **do** ▷ iterate B_α times over data set $\mathcal{X}$
    generate random permutation $\tilde{\mathcal{X}}$ of $\mathcal{X}$
    for $n \leftarrow 1$ **to** N **do** ▷ loop over data set $\tilde{\mathcal{X}}$
      $\mathbf{x} \leftarrow \mathbf{x}_n \in \tilde{\mathcal{X}}$
      for $r \leftarrow 1$ **to** M **do**
        $\Delta\mathbf{w}_r \leftarrow \mathbf{w}_r - \mathbf{x}$
        $d_r^2 \leftarrow \Delta\mathbf{w}_r^2$ ▷ Eq. (S58)
      end for
      select smallest values $d_{r(s)}^2$, $s = 0, \ldots, k_\alpha^{\max}$ ▷ apply Floyd&Rivest[10] algorithm $\Rightarrow$ reduces codebook $\mathcal{W}$ to $\tilde{\mathcal{W}}$
      generate ranking numbers $k_{r(s)}\ \forall\, \mathbf{w}_{r(s)} \in \tilde{\mathcal{W}}$ ▷ Eq. (S59) with $M \rightarrow k_\alpha^{\max}+1$ and $\mathcal{W} \rightarrow \tilde{\mathcal{W}}$
      for $s \leftarrow 0$ **to** $k_\alpha^{\max}$ **do**
        $\mathbf{w}_{r(s)} \leftarrow \mathbf{w}_{r(s)} + a[k_{r(s)}]\Delta\mathbf{w}_{r(s)}$ ▷ apply learning rule Eq. (S60)
      end for
    end for
  end for
  $\varepsilon_\alpha \leftarrow \varepsilon_{\text{fac}}\varepsilon_\alpha$ ▷ exponential annealing Eq. (S64)
  $\kappa_\alpha \leftarrow \kappa_{\text{fac}}\kappa_\alpha$
end for

S4.3 Top-Level Clustering

As is indicated in the main text, the data set $\mathcal{X}$, which is used at a certain time point t of a MD simulation to generate the N_λ clusters $C_{j,\lambda}(t)$ at the highest level λ of the hierarchy, is the set $\mathcal{X}_0(t)$ of all N_0 SU centers $\mathbf{r}_{k,0}(t)$. Similarly, the codebook vectors $\mathbf{w}_r \in \mathcal{W}$ are the geometrical centers $\mathbf{r}_{j,\lambda}(t) \in \mathcal{W}_\lambda(t)$ of the top-level clusters $C_{j,\lambda}(t)$, i.e. $r \equiv j$, $M \equiv N_\lambda$, and $\mathbf{w}_r \equiv \mathbf{r}_{j,\lambda}(t)$.

After the computation of the top-level codebook $\mathcal{W}_\lambda(t)$, the SUs $C_{k,0}$ are assigned to the top-level clusters $C_{j,\lambda}(t)$ by selecting the minimal distance $|\mathbf{r}_{k,0}(t) - \mathbf{r}_{j,\lambda}(t)|$. As a result the set $\mathcal{X}_0(t)$ is decomposed into N_λ disjoint subsets $\mathcal{X}_{0,j,\lambda}(t)$, which contain all vectors $\mathbf{r}_{k(j),0}(t)$ belonging to $\mathbf{r}_{j,\lambda}(t)$.

IPHIGENIE distinguishes two different modes of top-level VQ. These are

1. the *de novo* generation of the top-level codebook $\mathcal{W}_\lambda(0)$, which is executed once at the initialization ($t = 0$) of a MD simulation and is called "thorough" ($\mathfrak{T}$) learning and

2. the "adaptive" ($\mathfrak{A}$) learning of $\mathcal{W}_\lambda(T\tau)$, which is executed at every subsequent re-clustering step $T = 1, 2, \ldots$. By default re-clusterings are regularly carried out during a MD simulation at temporal distances $\tau = 256\Delta t$, where Δt is the size of the integration time step.

Table S5: Parameters of "thorough" ($\mathfrak{T}$), "adaptive" ($\mathfrak{A}$), and "normal" ($\mathfrak{N}$) VQ. A is the number of annealing cycles, $(\varepsilon_1, \kappa_1)$ and $(\varepsilon_A, \kappa_A)$ are the initial and final values of the learning parameter ε and of the cooperativeness range κ, respectively.

	A	ε_1	ε_A	κ_1	κ_A
$\mathfrak{T}$	10	0.95	0.03	$M/2$	0.25
$\mathfrak{A}$	5	0.06	0.03	0.5	0.25
$\mathfrak{N}$	8	0.30	0.03	$M/2$	0.25

An inspection of Table S5 shows that the two VQ modes $\mathfrak{T}$ and $\mathfrak{A}$ differ in the choice of the initial learning parameter ε_1, of the initial scale κ_1, and of the number A of annealing steps. For $\mathfrak{T}$ all these numbers are much larger than for $\mathfrak{A}$ indicating that also the computational

effort is much larger. VQ in mode $\mathfrak{A}$ is always highly competitive. Here at most three codebook vectors participate in each elementary learning step (S60).

Furthermore, the VQ modes $\mathfrak{T}$ and $\mathfrak{A}$ differ in the choice of the initial codebook $\mathcal{W}_\lambda^1(t)$. At $t = 0$ there is no previous codebook known. Thus, for a VQ in mode $\mathfrak{T}$ one has to guess $\mathcal{W}_\lambda^1(0)$ by randomly selecting N_λ SU positions $\mathbf{r}_{k,0}(0)$. After execution the result is taken as new initial codebook and the VQ in mode $\mathfrak{T}$ is once repeated to obtain the final codebook $\mathcal{W}_\lambda^A(0)$. At the time points $t = T\tau$ of re-clustering, however, the result of the preceding VQ is known and can be used as initial codebook for a VQ in mode $\mathfrak{A}$, i.e. $\mathcal{W}_\lambda^1(T\tau) = \mathcal{W}_\lambda^A[(T-1)\tau]$.

Because the spatial distribution of the SUs within a simulation volume slowly changes during a MD simulation and because the VQ yields a distribution of cluster centers closely reflecting the SU distribution, a previously calculated top-level codebook is a very good guess in a re-clustering step. Therefore, the codebook can be kept adaptively up to date by the computationally cheap VQ in mode $\mathfrak{A}$.

As is explained in Sec. 3.2, for the optimally load-balanced use of N_c CPUs of a parallel computer, the N_λ top-level clusters are partitioned into N_c groups each containing μ_c clusters $C_{j,\lambda}$, where μ_c is given by Eq. (20). The atom numbers $|C_{j,\lambda}|$ generally fluctuate around the average value $N/\mu_\mathrm{c}N_\mathrm{c}$. Similarly, the atoms assigned to the various CPUs will fluctuate around N/N_c. A simple algorithm for redistributing the $\mu_\mathrm{c}N_\mathrm{c}$ top-level clusters $C_{j,\lambda}$ among the N_c groups has been implemented in IPHIGENIE to keep the latter fluctuations small and, thus, to further optimize the load-balance.

S4.4 Top-Down Clustering along the Quaternary Tree

As is described in Sec. 3.4 of the main text, the four children $C_{i,l}(t)$ of the parent clusters $C_{j,l+1}(t)$ are calculated by the Martinetz[1] algorithm in a top-down fashion at all intermediate levels $1 \leq l \leq \lambda-1$ of the quaternary tree. For this purpose the four codebook vectors $\mathbf{r}_{i,l}(t)$, which define the children $C_{i,l}(t)$ and the codebook $\mathcal{W}_{j,l+1}$, are calculated from the centers $\mathbf{r}_{k(j),0}(t)$ of all those SU's, which were associated in the clustering at level $l+1$ to the parent

cluster $C_{j,l+1}(t)$ and are collected in the disjoint data sets $\mathcal{X}_{0,j,l+1}$.

Each of these clustering steps employs the parameters of "normal" VQ, which in Table S5 are labeled by the symbol $\mathfrak{N}$. During re-clustering the starting values $\mathbf{r}^1_{i,l}(t) \in \mathcal{W}^1_{j,l+1}(t)$ are derived from the preceding clustering result $\mathbf{r}^A_{i,l}(t-\tau) \in \mathcal{W}^A_{j,l+1}(t-\tau)$ by accounting for the motion of the parent cluster $C_{j,l+1}$ through $\mathbf{r}^1_{i,l}(t) = \mathbf{r}^A_{i,l}(t-\tau) + [\mathbf{r}^A_{j,l+1}(t) - \mathbf{r}^A_{j,l+1}(t-\tau)]$. The initialization at $t = 0$ is similar to that applied at the top-level.

Special care has to be taken at level $l = 1$ (if $\lambda > 1$), because the data sets $\mathcal{X}_{0,j,2}(t)$ will contain on the average only 16 data vectors $\mathbf{r}_{k(j),0}(t)$ such that its decomposition into four subsets $\mathcal{X}_{0,i,1}(t)$ by the minimum distance criterion may leave some of the resulting clusters empty, i.e. it may happen that $|C_{i,1}(t)| = 0$. In this case we fill this child cluster with the single SU, whose center $\mathbf{r}_{k(j),0}(t) \in \mathcal{X}_{0,j,2}(t)$ is maximally remote from the centers $\mathbf{r}^A_{i',1}(t)$ of the other three children $C_{i',1}(t)$. Hence, it is guaranteed that each cluster $C_{i,1}$ contains at least one SU. Because the numbers of SUs assigned to the clusters $C_{i,1}$ sizeably fluctuate around the average number of four, the strict ternary structure of the tree gets lost in the transition from level $l = 1$ to the SU level $l = 0$.

S5 Top-Level $\tilde{\lambda}(\Theta_\chi)$ of Interactions

According to Section 4.1 a top-down check of the inequality (28) determines, whether level l should become the so-called interaction top-level $\tilde{\lambda}(\Theta_\chi)$. This check compares the quantity $d_l(\Theta_\chi)$, which characterizes by Eq. (26) the typical interaction distances of clusters at level l, with the upper limit d_{MIC} of all interaction distances.

To motivate the condition (28) recall that the IAC Eq. (25) determines for a cluster pair $C_{i,l}$ and $C_{j,l}$ by

$$r^{\min}_{i,j,l}(\Theta_\chi) = \frac{1}{\Theta_\chi}\left(\tilde{R}_{i,l} + \tilde{R}_{j,l}\right) \tag{S66}$$

a minimal distance, at which the clusters are still added to their respective interaction lists.

Here, we have used the definition

$$\tilde{R}_{i,l} \equiv R_{i,l}/a_{i,l} \tag{S67}$$

for the accuracy weighted gyration radius of cluster $C_{i,l}$. For the clusters on level l one immediately gets the average boundary

$$\langle r_l^{\min}\rangle(\Theta_\chi) = 2\langle \tilde{R}_l\rangle/\Theta_\chi, \tag{S68}$$

where $\langle \tilde{R}_l\rangle$ is the ensemble average of $\tilde{R}_{i,l}$. Obviously, many of the $r_{i,j,l}^{\min}(\Theta_\chi)$ are smaller than $\langle r_l^{\min}\rangle(\Theta_\chi)$. Thus, many cluster pairs with $r \geq \langle r_l^{\min}\rangle(\Theta_\chi)$ will fulfill also the IAC (25).

When using SAMM$_{p,q}$/RF, a transition region, which starts at distances beyond $\langle \tilde{r}_l^{\min}\rangle(\Theta_\chi)$, has additionally to fit into the simulation system, whose inner radius is given by d_{MIC}. For narrowly distributed cluster sizes the average half-width of this region is with Eq. (29) approximately given by the average gyration radius $\langle R_l\rangle$ of the clusters at level l. Thus, with Eq. (S68) the system size should obey

$$\langle R_l\rangle + \langle \tilde{R}_l\rangle/\Theta_\chi \leq d_{\mathrm{MIC}}/2. \tag{S69}$$

If this condition is met at a clustering level $l > 0$ in a periodic system, then there is a chance to find a substantial number of clusters on that level, whose distances and accuracy weighted gyration radii are compatible with the IAC (25). The remaining clusters are decomposed into their children, whose interactions are handled on level $l - 1$.

For very small systems, in which no level $l > 0$ fulfills the condition (S69), the criterion has to be slightly modified to ensure a smooth transition of all SU pairs across the MIC boundary. The modified condition is

$$R_0^{\max} + \tilde{R}_0^{\max}/\Theta_\chi \leq d_{\mathrm{MIC}}/2 \tag{S70}$$

where $R_0^{\max}$ and $\tilde{R}_0^{\max}$ are the maximal values of the gyration radii $R_{i,0}$ and their accuracy

weighted relatives $\tilde{R}_{i,0}$ found at the SU level $l = 0$. In this case the maximal half-width Δ_0 of the transition region is

$$\Delta_0 = R_0^{\max}. \tag{S71}$$

As a result of the finite SU sizes (cf. Tables S3 and S4), there is a lower limit $d_0^{\min}(\Theta)$ for the system size d_{MIC}, which is still compatible with a SAMM$_{p,q}$/RF treatment. For a pure TIP3P water system ($R_{\mathrm{TIP3P}} = \tilde{R}_{\mathrm{TIP3P}} = 0.677$ Å) and for $\chi = \mathrm{f}$ this size is $d_0^{\min}(\Theta_{\mathrm{f}}) = 6.77$ Å. As a cubic box such a system can harbor $N_0 = 83$ TIP3P molecules containing $N = 249$ atoms. For $\chi = \mathrm{a}$ the size is $d_0^{\min}(\Theta_{\mathrm{a}}) = 9.32$ Å compatible with $N_0 = 216$ molecules and $N = 648$ atoms. If one tries to simulate a water/peptide mixture containing at least one lysine residue, which has the largest gyration radius of all SUs (with and without accuracy weighting), then $d_0^{\min}(\Theta_{\mathrm{f}}) = 15.58$ Å corresponding to a pure water system with $N_0 = 1008$ molecules or $N = 3024$ atoms. With a typical tenfold water excess such a system could contain a peptide with about 20 residues.

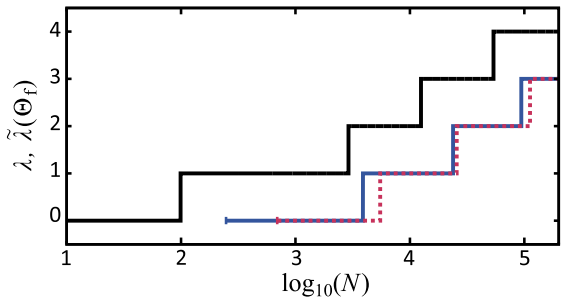


Figure S13: Heights λ (black) and $\tilde{\lambda}(\Theta_{\mathrm{f}})$ of the SAMM$_{p,q}$/RF hierarchies for pure H_2O (blue) and MeOH (red dashed) simulation systems with N atoms.

Figure S13 compares for different system sizes N the height λ (black) of the tree resulting from clustering (cf. Fig. 4) with the effective heights $\tilde{\lambda}(\Theta_{\mathrm{f}})$ admissible in SAMM$_{p,q}$/RF simulations of water (blue) and methanol (red dashed). The toroidal boundary conditions in conjunction with the transition zone, which enables smooth transitions of clusters from an explicit SAMM$_{p,q}$ to an implicit RF description (and *vice versa*), lead to values $\tilde{\lambda}(\Theta_{\mathrm{f}})$ of the interaction top-level, which are usually one level and rarely also two levels below the

top-level λ resulting from clustering. While the height λ of the tree obtained by clustering remains relevant for the purposes of load balancing, the effective height $\tilde{\lambda}(\Theta_{\mathrm{f}})$ steers the bottom-up and top-down computation of interactions. The figure also indicates the minimal system sizes accessible to SAMM$_{p,q}$/RF simulations for the two solvents considered here.

S6 Empirical Errors $\Delta f^{(4)}_{X,\mathrm{e}}(\tilde{r}_X)$ at Various Distances $\tilde{r}_X$.

Table S6 lists the values of the empirical errors $\Delta f^{(4)}_{X,\mathrm{e}}(\tilde{r}_X)$, which were calculated at the dimensionless distances $\tilde{r}_X = r/2(R_X/a_X)$ by means of (35) for SU dimers of the types X = T,M and are depicted in Figure 7 as blue circles (T) and red crosses(M), respectively. Additionally given are the associated distances r (in Å).

Table S6: Empirical Errors $\Delta f^{(4)}_{X,\mathrm{e}}(\tilde{r}_X)$, Dimensionless Distances $\tilde{r}_X$ and Distances r.

Methanol			TIP3P Water		
r[Å]	$\tilde{r}_\mathrm{M}$[Å]	$\Delta f^{(4)}_{\mathrm{M,e}}(\tilde{r}_\mathrm{M})$ [kcal/mol Å]	r[Å]	$\tilde{r}_\mathrm{T}$[Å]	$\Delta f^{(4)}_{\mathrm{T,e}}(\tilde{r}_\mathrm{T})$ [kcal/mol Å]
7.00	3.827	0.01716	5.25	3.877	0.01498
7.25	3.964	0.01359	5.50	4.062	0.01127
7.50	4.101	0.01073	5.75	4.247	0.00860
7.75	4.237	0.00891	6.00	4.431	0.00662
8.00	4.374	0.00736	6.25	4.616	0.00518
8.25	4.511	0.00604	6.50	4.801	0.00410
8.50	4.647	0.00508	6.75	4.985	0.00327
8.75	4.784	0.00421	7.00	5.170	0.00261
9.00	4.921	0.00354	7.25	5.355	0.00211
9.25	5.057	0.00300	7.50	5.539	0.00172
9.50	5.194	0.00256	7.75	5.724	0.00141
9.75	5.337	0.00218	8.00	5.908	0.00116
10.00	5.467	0.00187	8.25	6.093	0.00097
10.25	5.604	0.00159	8.50	6.278	0.00081
10.50	5.741	0.00138	8.75	6.462	0.00068
10.75	5.878	0.00120			
11.00	6.014	0.00103			
11.50	6.288	0.00079			

S7 Effects of Heating and Cooling in System $\mathfrak{M}$.

As a counterpart to Fig. 9, which applies to the TIP3P water system $\mathfrak{T}$, here we present the analogous Figure S14, which belongs to the methanol system $\mathfrak{M}$. Here the cooling per molecule, which remains at large values of Θ for all considered expansion orders q, is seen to be larger than for $\mathfrak{T}$. However, MeOH has 14 flexible degrees of freedom (DOF) whereas TIP3P has only six. Thus when measuring the heating power per DOF instead per molecule, the heating rates of MeOH at Θ_f are almost identical to those of TIP3P. Measuring heating rates per DOF is even more appropriate, because thermostats use these quantities for control.[11] As a result, the heating rates shown in the Figs. 9 and S14 are equivalent and everywhere at least acceptably (see Section S8) small.

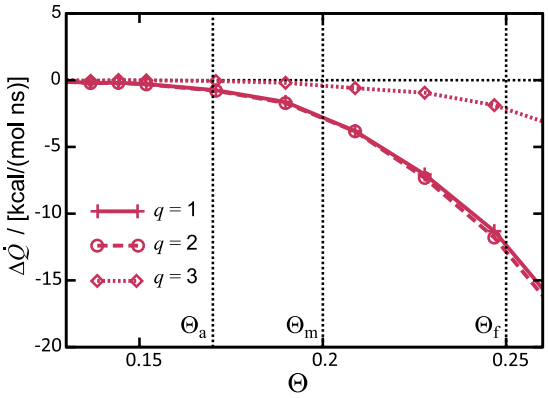


Figure S14: The contributions $\Delta\dot{Q}$ of the SAMM$_q$ approximation for the dispersion to the total heating rate $\dot{Q}$ as a function of the IAC threshold Θ for the orders $q = 1$ (solid line), $q = 2$ (dashed line), and $q = 3$ (dotted line). These contributions are caused by the transition from the exact calculation to the SAMM$_q$ treatment at the IAC distance $d_\mathrm{M}(\Theta)$. For explanation see the text.

Also Figure S15 is simply the MeOH counterpart to the TP3P water Figure 11 in the main text. Thus, Fig. S15 illustrates the total algorithmic noise of SAMM$_{4,q}$ for increasing orders $q = 1$, 2, 3 of the dispersion expansion. The additional heat production $\dot{Q}_\mathrm{MIC}$ occurring in SAMM$_{4,q}$/RF simulations, in which objects may dynamically vanish into or reappear out of the RF continuum near d_MIC, has been subtracted by the formation of the difference $\Delta\dot{Q}$. In system $\mathfrak{M}$ this source of algorithmic noise has the power $\dot{Q}_\mathrm{MIC} = 1.05$ kcal/(mol ns) per

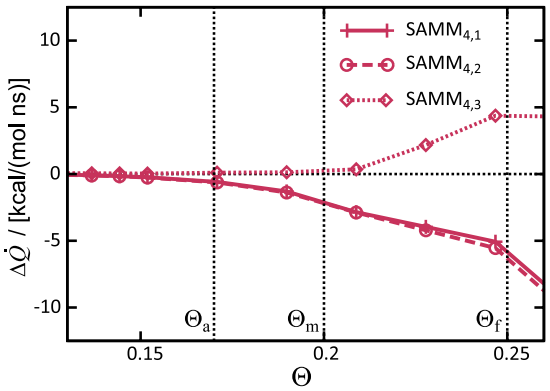


Figure S15: $\mathrm{SAMM}_{4,q}$ algorithmic noise for $q = 1$ (solid line), $q = 2$ (dashed line), $q = 3$ (dotted line) as a function of the IAC threshold Θ measured in the system $\mathfrak{M}$ by the heating rate differences $\Delta\dot{Q}$, which are defined by the entries $\mathrm{SAMM}_{4,q}$ in Table 2.

molecule.

According to Fig. S15, the small $\mathrm{SAMM}_{4,q}$ cooling at Θ_a is almost negligible for all q and remains negligible up to Θ_m for $q = 3$. For $q \leq 2$ the cooling gets more pronounced with increasing Θ and reaches about -5 kcal/(mol ns) at Θ_f, whereas for $q = 3$ the cooling turns into a small heating. However, up to Θ_f all these algorithmic artifacts remain acceptably small.

The small heating rate observed for $\mathrm{SAMM}_{4,3}$ in system $\mathfrak{M}$ at Θ_f (dotted line) is partially the result of cancellations, although in this case the cancellations are less significant than for system $\mathfrak{T}$ (cf. the discussion of Fig. 11). For system $\mathfrak{M}$ the SAMM_3 dispersion expansion causes according to the dotted line in Fig. S14 at Θ_f a small amount of cooling, which cancels the repulsion cutoff heating (data not shown) such that the total $\mathrm{SAMM}_{4,3}$ heating rate closely resembles the heating contribution of the SAMM_4 electrostatics expansion drawn as a red solid line in Fig. 10. Because the lower order SAMM_q dispersion expansions ($q = 1, 2$) do not completely remove in system $\mathfrak{M}$ the dispersion cutoff cooling (cf. the dashed and solid lines in Fig. S14), their residual coolings overcompensate the SAMM_4 electrostatics heating (cf. the solid red line in Fig. 10) and lead to the total cooling seen for $\mathrm{SAMM}_{4,1}$ and $\mathrm{SAMM}_{4,2}$ at Θ_f in Fig. S15, which is however acceptably small.

S8 Control of Algorithmic Heating and Cooling.

Efficient MD algorithms do not exactly conserve the total energy, because they usually change the approximation of the long range interactions with increasing distance. There are, of course, also energy conserving computational schemes; however, they work with artificially modified potentials and forces. With physical potentials, in contrast, all atoms, which dynamically cross approximation boundaries, are sources or sinks of heat. As a result, the temperature must be controlled by some thermostat, which in the case of solvated proteins should be exclusively coupled to the solvent.[11] If one wants to characterize dynamical properties, a minimally invasive[11] (MI) or weakly invasive (WI) thermostat should be applied.

For the construction of such a MI or WI thermostat one must measure the algorithmic heat production in the given simulation system by a series of short and statistically independent NVE simulations, whose initial conditions belong to the desired simulation temperature T. If one now sets the thermostat power to $\beta = -\dot{Q}$ and chooses a target temperature T_0 well above (for cooling $\dot{Q} < 0$) or below (for heating $\dot{Q} < 0$) the simulation temperature T (such that T_0 is well outside the range of the micro-canonical temperature fluctuations of the system), then the relaxation time of a MI Berendsen thermostat[12] is given by[11]

$$\tau(T_0|\beta, T) = \frac{k_{\mathrm{B}}\,(T_0 - T)}{2\beta}, \tag{S72}$$

where k_{B} is the Boltzmann constant.

Assume now that a simulation system is cooling with a power of -2 kcal/(mol ns) per flexible degree of freedom. For rigid TIP3P models this number translates into a cooling rate $\dot{Q} = -12$ kcal/(mol ns) per molecule and for partially stiff MeOH models into the rate $\dot{Q} = -28$ kcal/(mol ns) per molecule. Choosing as the desired simulation temperature $T_{\mathrm{ref}} = 298.15$ K and for the Berendsen thermostat the target temperature $T_0 = 400$ K, which

should be well outside the range of temperature fluctuations, then one finds

$$\tau(400\text{ K}|2\text{ kcal/(mol ns)}, 298.15\text{ K}) = 50.54\text{ ps}.$$

This quite long coupling time will suffice to keep the system near T_{ref}, if the heating rate $\dot{Q}(\Theta, T)$ of the system has a negative or vanishing temperature derivative

$$\left.\frac{\partial \dot{Q}(\Theta, T)}{\partial T}\right|_{T_{\text{ref}}} \leq 0. \tag{S73}$$

If, however, this derivative is positive, the MI thermostat cannot balance the system at T_{ref}. Then one needs a WI thermostat, whose target temperature T_0 is just outside the range of the NVE temperature fluctuations. If their standard deviation σ_T is about 3 K, then a target temperature outside a $3\sigma_T$ range, i.e. outside ± 10 K, which defines a WI thermostat, should work. In this case one finds

$$\tau(308.15\text{ K}|2\text{ kcal/(mol ns)}, 298.15\text{ K}) = 4.96\text{ ps}.$$

Eq. (S72) now suggests that decreasing $T_0 - T$ by another factor 10 to about 1 K means that τ must also be reduced by a factor 10 to about 0.5 ps. Then the WI scheme will not anymore apply, because T_0 will be within the range of temperature fluctuations, and one gets a usual Berendsen thermostat. In this case the simulation temperature will be at least by 1 K smaller than the target temperature T_0. More strongly cooling systems would thus require even shorter coupling times and, hence, stronger modifications of the simulated dynamics. Therefore we consider heating rates of $|2|$ kcal/(mol ns) per DOF as "acceptable" and classify heating rates, which are by more than one order of magnitude smaller, as "almost negligible".

In the SAMM$_{4,1}$/RF simulations of the systems $\mathfrak{T}$ and $\mathfrak{M}$ we found for $\Theta = \Theta_{\text{f}}$ the overall cooling rates of -0.18 kcal/(mol ns) and of -0.29 kcal/(mol ns) per DOF, which are both

about one order of magnitude smaller than the |2| kcal/(mol ns) reference rate and hence are almost negligibly small. On the other hand the temperature derivative of the heating rate turned out to be strongly positive as opposed to the requirement (S73). Hence, a WI thermostat had to be applied. With $T_0 = 308.15$ K, the coupling times of 55.9 ps and 34.3 ps, respectively, suffice to control T in $\mathfrak{T}$ and $\mathfrak{M}$ at $T_{\mathrm{ref}} = 298.15$ K (data not shown). We have used this terminology for qualifying the strengths of heat sources in Section 6.

S9 Check of Linear Scaling

For a better visibility we have replotted in Figure S16 the scaling data presented in Fig. 12 as relative computation times $(t/t_{\mathrm{ref}}) \times (10^4/N)$ per atom over the number $N/10^4$ of atoms in the system. The figure demonstrates that the fluctuations of the relative computation

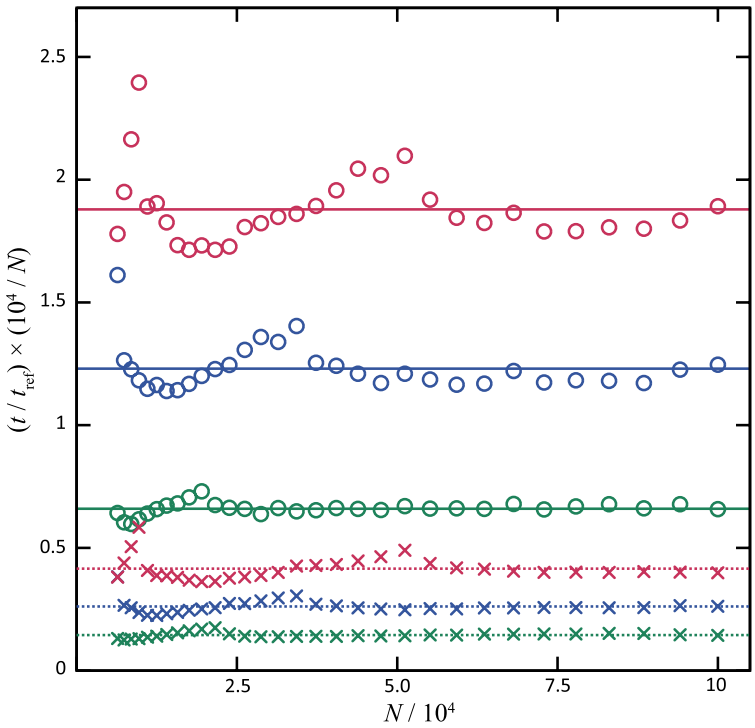


Figure S16: Relative computation times $(t/t_{\mathrm{ref}}) \times (10^4/N)$ per atom measured by applying $\mathrm{SAMM}^{\chi}_{4,1}$/RF to the systems $\mathfrak{T}_i$ (crosses) and $\mathfrak{P}_i$ (circles) with the accuracy/efficiency choices $\chi = \mathrm{a}$ (red), $\chi = \mathrm{m}$ (blue), and $\chi = \mathrm{f}$ (green). Also shown are corresponding averages ρ_χ ($\mathfrak{T}$: dotted, $\mathfrak{P}$: solid).

times per atom around the respective average values ρ_χ are qualitatively similar for $\mathfrak{T}$ and $\mathfrak{P}$ at each level χ of the accuracy/efficiency compromise. The amplitude of these fluctuations are seen to increase with the computation time per atom.

S10 Pressure Calculation Using SAMM$_{p,q}$

MD simulations in the NpT ensemble require a barostat and the calculation of the pressure in the given system. As is common in MD simulations, the dynamic contribution to the instantaneous pressure $p(t)$ is calculated from the virial expression,[13] which sums up the scalar products of the atomic positions and forces. Hence, all approximations applied to force computations affect the value of $p(t)$. In particular, also the orders (p, q), at which the SAMM expansions of the electrostatic and dispersive forces, respectively, are truncated, will modify $p(t)$.

For a quantification we apply a procedure, which is analogous to that presented in Sec. 5.2 for the measurement of algorithmic noise through heating rate differences. Hence we define the ensemble average pressure difference

$$\langle \Delta p(\Theta, P, P_{\mathrm{ref}}) \rangle \equiv \langle p(\Theta, P) \rangle - \langle p(\Theta, P_{\mathrm{ref}}) \rangle, \tag{S74}$$

where the brackets $\langle \ldots \rangle$ denote the average over an ensemble of 2000 snapshots taken at temporal distances of 10 fs from 20 ps MD-NVT-trajectories, which were executed with the parameters Θ, P and P_{ref}, respectively. Here Θ is the applied IAC threshold, and the pairs P and P_{ref} of parameter sets (p, q, c_{r}) are explained in the discussion of Table 2. Accordingly, each pair serves to identify the influence of a specific parameter choice on the pressure. As a test bed we used the system $\mathfrak{T}_{12}$ comprising 8737 TIP3P water models. Table S7 lists the pressure differences $\langle \Delta p(\Theta, P, P_{\mathrm{ref}}) \rangle$ obtained for the various comparisons with the three standard IAC thresholds Θ_{f}, Θ_{m}, and Θ_{a}. Judging from the 114 bar standard deviation of the 2000 pressure values $p(t)$, which amounts to 9.9 % of the average virial contribution to the total pressure, the statistical errors of the averages should be in the range of about $\pm$2.5 bar.

The first line ($c_{\mathrm{r}} = d_X$) of Table S7 shows that truncating the r^{-12} model of the Pauli repulsion at $c_{\mathrm{r}} \sim 1/\Theta_\chi$ entails pressure underestimates, which increase from 6 bar to 36 bar

Table S7: Pressure differences $\langle \Delta p(\Theta_\chi, P, P_{\mathrm{ref}}) \rangle$ in bar identifying the effects of specific approximations to the long-range inter-atomic forces.

Comparison	$\Theta_{\mathrm{a}} = 0.17$	$\Theta_{\mathrm{m}} = 0.20$	$\Theta_{\mathrm{f}} = 0.25$
$c_{\mathrm{r}} = d_X$	-6.4	-10.0	-36.4
$q = -1$	-3.5	7.3	18.4
$q = 3$	0.3	0.9	4.9
$q = 2$	0.3	0.9	4.5
$q = 1$	7.7	17.9	54.6
$p = 4$	0.3	0.8	11.4
$p = 3$	5.4	14.6	77.9
$\mathrm{SAMM}_{4,3}$	-5.8	-8.4	-20.0
$\mathrm{SAMM}_{4,2}$	-5.9	-8.4	-20.5
$\mathrm{SAMM}_{4,1}$	1.6	8.6	29.6

with decreasing c_{r}. Pressure underestimates are to be expected whenever repulsive forces are partially neglected.

The second line ($q = -1$) quantifies the effect of replacing the explicit calculation of the dispersion attraction at distances beyond $d_X \sim 1/\Theta_\chi$ by a continuum model.[14] Accordingly, the missing explicit calculation of the dispersion attraction is overcompensated by the continuum approximation for large transition distances d_X, whereas at smaller d_X (Θ_{m}, Θ_{f}) the continuum model does not completely compensate the missing explicitly calculated attractive forces. This instance becomes apparent in the remaining pressure overestimates of 7 bar and 18 bar, respectively. Without the continuum correction, i.e. with a simple truncation of the dispersion attraction, the pressure overestimates are much larger and range from 378 bar to 1231 bar (data not shown).

All truncations of the dispersion FMM expansion at finite values q lead to overestimates of the pressure, which tend to decrease with an increasing value of the expansion order q and with an increasing IAC distance $d_X \sim 1/\Theta_\chi$. Correspondingly, the pressure overestimates are largest (55 bar) for $q = 1$ and Θ_{f} and smallest (0.3 bar) for $q = 3$ and Θ_{a}. As a result, for the dispersion attraction the FMM truncations all lead to more or less significant overestimates of the pressure.

Also the truncations of the electrostatics FMM expansion (cases "$p = 3$" and "$p = 4$") entail pressure overestimates indicating that they cause a neglect of attractive force contributions. The transition from $p = 3$ to $p = 4$ substantially reduces the magnitude of the pressure overestimates. At $p = 4$ they have a size comparable to the overestimates caused by the truncation of the dispersion expansion at $q = 3$ and are much smaller than the ones observed for $q = 1$.

Thus, in $\mathrm{SAMM}^{\chi}_{p,q}$ the pressure overestimates, which are due to the truncations of the dispersion and of the electrostatics FMM expansions, add up, whereas the truncation of the Pauli repulsion causes a compensation.

This expectation is corroborated by the entries "$\mathrm{SAMM}_{4,q}$" at the bottom of Table S7. In the case of the more accurate versions ($\mathrm{SAMM}_{4,3}$ and $\mathrm{SAMM}_{4,2}$ combined with Θ_a and Θ_m) the pressure is underestimated by less than 10 bar. A comparison with the first line of the table shows, that these underestimates are almost exclusively due to the truncation of the Pauli repulsion at $c_\mathrm{r} = d_X$. Thus, setting up a hypothetical FMM expansion also for these repulsive forces promises to reduce the pressure errors to about 1-2 bar with Θ_a and Θ_m. The absolute values of the pressure errors, which are observed for the combinations of $\mathrm{SAMM}_{4,q}$ with Θ_f, are smaller than 30 bar. In combination with a FMM expansion for the Pauli repulsion the $\mathrm{SAMM}_{4,2}$ and $\mathrm{SAMM}_{4,3}$ algorithms promise to render pressure errors below 10 bar even with the most short range IAC threshold Θ_f. However, in view of the 140 bar standard deviation of the pressure fluctuations, all the above systematic errors can be considered as quite small and the possible repair by a hypothetical FMM expansion of the Pauli repulsion is not urgently needed.

References

(1) Martinetz, T.; Berkovich, S.; Schulten, K. *IEE Trans. Neur. Networks* **1993**, *4*, 558–569.

(2) 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.

(3) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(4) Tröster, P.; Lorenzen, K.; Tavan, P. *J. Phys. Chem. B* **2014**, *118*, 1589–1602.

(5) Bordat, P.; Sacristan, J.; Reith, D.; Girard, S.; Glättli, A.; Müller-Plathe, F. *Chem. Phys. Lett.* **2003**, *374*, 201 – 205.

(6) Lorenzen, K.; Schwörer, M.; Tröster, P.; Mates, S.; Tavan, P. *J. Chem. Theory Comput.* **2012**, *8*, 3628–3636.

(7) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification, 2nd Edition*; Wiley: Hoboken, NJ, 2000.

(8) Kohonen, T. *Self-Organizing Maps*; Springer: Berlin, 1995.

(9) Kloppenburg, M.; Tavan, P. *Phys. Rev. E* **1997**, *55*, R2089–R2092.

(10) Floyd, R. W.; Rivest, R. L. *Commun. ACM* **1975**, *18*, 173.

(11) Lingenheil, M.; Denschlag, R.; Reichold, R.; Tavan, P. *J. Chem. Theory Comput.* **2008**, *4*, 1293–1306.

(12) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(13) Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*; Oxford University Press: New York, 2010.

(14) Allen, M. P.; Tildesley, D. *Computer Simulations of Liquids*; Clarendon: Oxford, 1987.

2.3 Hamiltonsche Kombination von HADES-MD mit SAMM

Das nachfolgende Publikation[5]

„Linearly Scaling and Almost Hamiltonian Dielectric Continuum Molecular Dynamics Simulations through Fast Multipole Expansions “

Konstantin Lorenzen, Gerald Mathias, and Paul Tavan
J. Chem. Phys. **143** 184114 (2015),

die ich zusammen mit Gerald Mathias und Paul Tavan verfasst habe, behandelt die Kombination der durch Sebastian Bauer entwickelten *Hamiltonian Dielectric Solvent* (HADES) Kontinuumsmethode [48, 58] für hocheffiziente Simulationen von Proteinen im dielektrischen Kontinuum mit der schnellen Multipolmethode SAMM. Hierzu werden Hamiltonsche SAMM-Kräfte entwickelt, welche die Erhaltung des linearen Impulses und des Drehimpulses ermöglichen. Es wird gezeigt, dass der Rechenaufwand der HADES/SAMM-MD linear mit der Anzahl N der Proteinatome skaliert und dadurch auch große Proteine effizient zu simulieren sind.

[5]Mit freundlicher Genehmigung der Verlags

THE JOURNAL OF CHEMICAL PHYSICS **143**, 184114 (2015)

Linearly scaling and almost Hamiltonian dielectric continuum molecular dynamics simulations through fast multipole expansions

Konstantin Lorenzen, Gerald Mathias, and Paul Tavan[a)]
Lehrstuhl für BioMolekulare Optik, Ludig–Maximilians Universität München, Oettingenstr. 67, 80538 München, Germany

(Received 17 August 2015; accepted 30 October 2015; published online 13 November 2015)

Hamiltonian Dielectric Solvent (HADES) is a recent method [S. Bauer *et al.*, J. Chem. Phys. **140**, 104103 (2014)] which enables atomistic Hamiltonian molecular dynamics (MD) simulations of peptides and proteins in dielectric solvent continua. Such simulations become rapidly impractical for large proteins, because the computational effort of HADES scales quadratically with the number N of atoms. If one tries to achieve linear scaling by applying a fast multipole method (FMM) to the computation of the HADES electrostatics, the Hamiltonian character (conservation of total energy, linear, and angular momenta) may get lost. Here, we show that the Hamiltonian character of HADES can be almost completely preserved, if the structure-adapted fast multipole method (SAMM) as recently redesigned by Lorenzen *et al.* [J. Chem. Theory Comput. **10**, 3244-3259 (2014)] is suitably extended and is chosen as the FMM module. By this extension, the HADES/SAMM forces become exact gradients of the HADES/SAMM energy. Their translational and rotational invariance then guarantees (within the limits of numerical accuracy) the exact conservation of the linear and angular momenta. Also, the total energy is essentially conserved—up to residual algorithmic noise, which is caused by the periodically repeated SAMM interaction list updates. These updates entail very small temporal discontinuities of the force description, because the employed SAMM approximations represent deliberately balanced compromises between accuracy and efficiency. The energy-gradient corrected version of SAMM can also be applied, of course, to MD simulations of all-atom solvent-solute systems enclosed by periodic boundary conditions. However, as we demonstrate in passing, this choice does not offer any serious advantages. © *2015 AIP Publishing LLC.* [http://dx.doi.org/10.1063/1.4935514]

I. INTRODUCTION

Molecular dynamics (MD) simulations of protein-solvent systems, whose inter-atomic forces are calculated from an all-atom molecular mechanics (MM) force field,[1–3] still pose a computational challenge.[4–6] Here, the atomistic description of the aqueous solvent is a major factor limiting the conformational sampling of the solute proteins, because the number of solvent atoms should exceed the number of protein atoms by at least one order of magnitude for physically adequate descriptions.[7–9] Therefore, most of the computational effort is spent on the calculation of the interactions among the water molecules.

This effort can be saved, if the solvent can be replaced by a computationally inexpensive and physically correct continuum model. For this purpose, the dielectric Poisson equation (PE) must be solved at each step of the numerical integration of the protein dynamics.[10] A corresponding approach neglects the dielectric relaxation[11] (fs–ps) of the water and its structure near a protein surface. Here, the neglect of the water relaxation should be of minor importance, because the conformational dynamics of proteins is much slower (>ns). In contrast, the significance of the water structure at protein surfaces is still unclear[12,13] and can be assessed only if an efficient and accurate continuum approach is available.

Such a continuum approach to MM-MD simulations has been recently constructed[10,14] and has been shown to yield a Hamiltonian dynamics, which explains its name "Hamiltonian dielectric solvent" (HADES). Particularly because it includes, by construction, the reaction forces, which generate the so-called dielectric boundary pressure and are exerted by the continuum on the protein atoms, it removes many of the practical and conceptual difficulties[10,15–19] posed by earlier continuum models. Generalized Born methods (see, e.g., Refs. 20–22), in contrast, fail to solve the PE[10,21,23] and, therefore, cannot provide accurate expressions for the reaction forces. A similar critique[24] applies to the "inducible multipole solvation" (IMPS) model.[25,26] Atomic forces derived from grid-based solutions of the PE[27,28] violate energy conservation not only because of numerical inaccuracies but, more importantly, because they have to introduce *ad-hoc* models for the dielectric boundary pressure, which do not comply with Newton's reaction principle.

On the other hand, a free energy functional approach,[29] which actually yields a Hamiltonian dynamics, turned out to be slower[18] than explicit solvent simulations.

HADES[10,14] rests on a reformulation of the PE,[19,24,30] which exactly replaces the polarization of an aqueous continuum featuring a high dielectric constant ϵ_c by an

[a)]Electronic mail: tavan@physik.uni-muenchen.de

anti-polarization within the embedded solute protein, whose interior is characterized by a low dielectric constant ϵ_s. This anti-polarization is then approximately expressed for all atoms i by Gaussian reaction field (RF) dipoles $\tilde{\mathbf{p}}_i$ and for atoms carrying partial charges q_i by additional Gaussian shielding charges $\hat{q}_i = -q_i(1 - \epsilon_s/\epsilon_c)$, which jointly generate the RF contribution to the electrostatic potential. Up to a constant factor, the widths σ_i of these Gaussian charge and dipole distributions are identical to the widths of Gaussian atom models, which collectively define the space occupied by the protein. The RF dipoles self-consistently derive from the electric field $\langle\tilde{\mathbf{E}}(\mathbf{r}_i)\rangle_{\sigma_i}$ within an atom at the position $\mathbf{r}_i$, where the field $\tilde{\mathbf{E}}(\mathbf{r})$ is averaged over the Gaussian atomic volume specified by σ_i, through the anti-polarization relation $\tilde{\mathbf{p}}_i = -\alpha_i\langle\tilde{\mathbf{E}}(\mathbf{r}_i)\rangle_{\sigma_i}$. Here, α_i is the RF polarizability of atom i.[10]

As a result, HADES approximately expresses the continuum electrostatics of a protein, whose atoms carry static partial charges, as an isolated, rotationally and translationally invariant, and Hamiltonian many body system, in which the electric field is generated by partial charges q_i, by Gaussian shielding charges $\hat{q}_i$ and Gaussian RF dipoles $\tilde{\mathbf{p}}_i$. For distances larger than about 5 Å the Gaussian charges and dipoles can be safely replaced by point charges and dipoles, because the Gaussian widths σ_i can be chosen[10] smaller than about 0.8 Å. Thus, the computational effort of HADES is comparable to that of a polarizable force field featuring partial point charges and inducible point dipoles. This effort thus scales with N^2, where N is the number of protein atoms.

For a small ($N = 150$) α-helical decapeptide, the simulation speed of HADES-MD turned out to exceed that of an explicit solvent simulation by a factor of about 20.[31] For larger proteins, the quadratic scaling will rapidly remove this advantage of HADES-MD, because explicit solvent MD simulations can be executed in a linearly scaling fashion, if one employs, e.g., a fast multipole method[32–43] (FMM) like the structure adapted multipole method[8,44–49] (SAMM) for the description of the electrostatic and dispersive interactions. Thus, the extension of HADES-MD toward linear scaling is mandatory,[31] if one wants to preserve its computational advantages also for large proteins.

With this contribution, we want to demonstrate that HADES-MD can be actually converted into a linearly scaling simulation approach by combining it with a suitable extension of the FMM method SAMM.[44–49] With the purpose of preserving the Hamiltonian character, the key issue will be the use of SAMM forces, which are exact negative gradients of the associated SAMM energies. The required mathematics will be sketched in Sec. II and presented in more detail in several sections of the supplementary material.[50] Subsequent sample simulations will serve to demonstrate the (almost) Hamiltonian character and the linear scaling of HADES/SAMM-MD.

II. THEORY

As mentioned further above, in HADES the electrostatic potential $\Phi(\mathbf{r})$ is generated by partial charges q_i, Gaussian shielding charges $\hat{q}_i$, and induced Gaussian RF dipoles $\tilde{\mathbf{p}}_i$. When combining HADES with the most recent version[45] of SAMM, only interactions at distances smaller than a lower limit $d_0 \approx 5.5$-7.0 Å (with about 100 nearby atoms) are calculated from exact pair expressions. For larger distances, the interactions are approximated by FMM expansions. At such distances, the Gaussian shielding charges and RF dipoles can be safely treated by SAMM as point-like objects, because the sizes of the Gaussian widths σ_i employed by HADES are sufficiently small ($d_0/\sigma_i \gtrsim 7$). Hence, a combination of HADES with SAMM requires FMM expansions of point charge and point dipole distributions and energy expressions for interactions among these objects.

Mathematical derivations, which cover charges and dipoles instead of solely charges, take a lot of space without adding too much insight. Therefore, we decided to restrict the following derivation of SAMM forces as exact gradients of the SAMM energy to the case of point charges. This presentation is complemented by the supplementary material,[50] which provides the corresponding mathematics for electrostatic charge-dipole and dipole-dipole interactions (Section S1.2), for the $\sim 1/r^6$ dispersion attraction (Section S1.3), and for a $\sim 1/r^{12}$ soft core repulsion (Section S1.4).

The scenario considered by those FMMs, which are based on Cartesian two-sided Taylor expansions,[41–45] is sketched in Figure 1, which addresses the interactions among the atoms of two distant atomic clusters C and D. Electrostatic charges q_j, which belong to the atoms $j \in D$ located at the positions $\mathbf{s}_j$, generate at positions $\mathbf{s}$ the Coulomb potentials $q_j/|\mathbf{s} - \mathbf{s}_j|$, which add up to the total potential $\Phi^D(\mathbf{s})$ caused by cluster D. The total electrostatic energy $E(C, D)$ of the cluster C, which is exposed to this potential and contains charges q_i at positions $\mathbf{r}_i$, is $\sum_{i\in C} q_i\Phi^D(\mathbf{r}_i)$. $E(C, D)$ consists of pair contributions depending on the distances $r_{ij} \equiv |\mathbf{r}_i - \mathbf{s}_j|$ between the atoms $i \in C$ and $j \in D$. With the local coordinates $\mathbf{a}_i$ and $\mathbf{b}_j$, which are defined (cf. Fig. 1) with respect to the centers of geometry,

$$\mathbf{c} \equiv \frac{1}{|C|}\sum_{i\in C}\mathbf{r}_i \qquad \text{and} \qquad \mathbf{d} \equiv \frac{1}{|D|}\sum_{j\in D}\mathbf{s}_j, \tag{1}$$

of the two clusters, and with the connecting vector $\mathbf{r} = \mathbf{c} - \mathbf{d}$ the inter-atomic distances may be rewritten as

$$r_{ij} = |\mathbf{r} + (\mathbf{a}_i - \mathbf{b}_j)|. \tag{2}$$

Note here that most other FMM approaches cover the simulation system with a spatially fixed hierarchical grid,

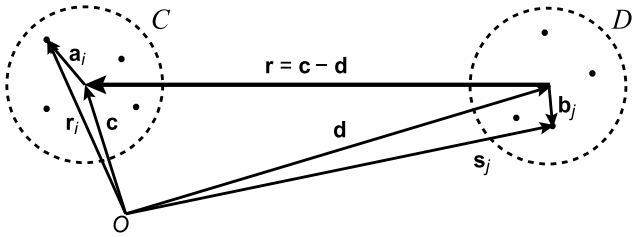


FIG. 1. FMM geometry for two interacting clusters C and D (dashed spheres) of atoms $i \in C$ at $\mathbf{r}_i$ and $j \in D$ at $\mathbf{s}_j$ (dots) carrying electrostatic charges q_i and q_j, respectively. The interactions depend on the connecting vectors $\mathbf{r}_i - \mathbf{s}_j$ and are evaluated by two-sided Taylor expansions around the vector $\mathbf{r} = \mathbf{c} - \mathbf{d}$ linking the two cluster centers. Relative to the centers, the atomic positions are given by $\mathbf{a}_i$ for $i \in C$ and $\mathbf{b}_j$ for $j \in D$.

which then partitions the simulation system into a nested hierarchy of atomic clusters. Thus, the cluster centers $\mathbf{c}$ and $\mathbf{d}$ are constants. SAMM, in contrast, employs a structure-adapted and adaptive decomposition of the system into a quaternary tree of nested and optimally compact atomic clusters, which represents the required FMM hierarchy.[45] Therefore, in SAMM the cluster centers move with the enclosed atoms.

The above expression, (2), for r_{ij} suggests that a two-sided Cartesian Taylor expansion of the potential $\Phi^D(\mathbf{r}_i)$ around the connecting vector $\mathbf{r}$ should represent a rapidly converging approximation, if the cluster-cluster distance,

$$r \equiv |\mathbf{r}| = |\mathbf{c} - \mathbf{d}|, \tag{3}$$

is large compared to typical distances $|\mathbf{a}_i - \mathbf{b}_j|$. This is the key concept underlying the associated FMM approaches[41–43] to which also SAMM belongs.[44,45]

Correspondingly, SAMM applies such Taylor expansions[41–43] of order $p \in \{3,4\}$ to the electrostatic potential, i.e.,

$$\Phi^{D,p}(\mathbf{r}_i) = \sum_{n=0}^{p} \frac{1}{n!} \partial_{(n)} \frac{1}{r} \odot \sum_{j \in D} q_j (\mathbf{a}_i - \mathbf{b}_j)^{(n)}, \tag{4}$$

where $\partial_{(n)}(1/r)$ is a tensor of rank n composed of the nth order partial derivatives of $1/r$, where $\odot$ denotes the inner contraction product of two tensors, and where $(\mathbf{a}_i - \mathbf{b}_j)^{(n)}$ is the n-fold outer product of the vector $\mathbf{a}_i - \mathbf{b}_j$ with itself. In particular, $\partial_{(n)}(1/r)$ is the n-fold outer product $(\partial/\partial\mathbf{r})^{(n)}$ of the gradient operator $\partial/\partial\mathbf{r}$ applied to $(1/|\mathbf{r}|)$. Note that the employed tensorial notation is thoroughly explained in the quoted papers.[41–44]

Replacing the exact potential in the interaction energy $E(C,D)$ by its approximate counterpart (4) yields the total electrostatic cluster-cluster interaction energy

$$E^p(C,D) \equiv \sum_{i \in C} q_i \Phi^{D,p}(\mathbf{r}_i) \tag{5}$$

of the pth order SAMM algorithm.[44,45] With Taylor expansion (4) this energy is given by

$$E^p(C,D) = \sum_{i \in C} q_i \sum_{n=0}^{p} \frac{1}{n!} \partial_{(n)} \frac{1}{r} \odot \sum_{j \in D} q_j (\mathbf{a}_i - \mathbf{b}_j)^{(n)}. \tag{6}$$

By computing the gradient

$$\mathbf{f}^p(\mathbf{r}_k) \equiv -\frac{\partial}{\partial \mathbf{r}_k} E^p(C,D) \tag{7}$$

of $E^p(C,D)$ with respect to the position $\mathbf{r}_k$ of an atom $k \in C$, one obtains the associated electrostatic force $\mathbf{f}^p(\mathbf{r}_k)$, which acts on k and is generated by the cluster D.

An inspection of Eq. (6) shows that $E^p(C,D)$ does not explicitly depend on the position vectors $\mathbf{r}_i$ of the atoms $i \in C$ but only indirectly through (i) the local coordinates $\mathbf{a}_i$, which according to Figure 1 are given by

$$\mathbf{a}_i(\mathbf{r}_1, \ldots, \mathbf{r}_{|C|}) = \mathbf{r}_i - \mathbf{c}(\mathbf{r}_1, \ldots, \mathbf{r}_{|C|}), \tag{8}$$

through (ii) the geometric center $\mathbf{c}(\mathbf{r}_1, \ldots, \mathbf{r}_{|C|})$ specified by Eq. (1), and through (iii) the cluster-cluster distance r, which is defined by Eq. (3) and also depends on $\mathbf{c}(\mathbf{r}_1, \ldots, \mathbf{r}_{|C|})$. Applying the chain rule to the gradient in Eq. (7), one thus gets

$$\mathbf{f}^p(\mathbf{r}_k) = -\sum_{i=1}^{|C|} \frac{\partial E^p(C,D)}{\partial \mathbf{a}_i} \frac{\partial \mathbf{a}_i}{\partial \mathbf{r}_k} - \frac{\partial E^p(C,D)}{\partial r} \frac{\partial r}{\partial \mathbf{r}_k}. \tag{9}$$

The required gradients $\partial \mathbf{a}_i / \partial \mathbf{r}_k$ and $\partial r / \partial \mathbf{r}_k$ follow from Eqs. (8), (1), and (3). With the unit matrix $\mathbf{I} \in \mathbb{R}^3 \times \mathbb{R}^3$ they are $\partial \mathbf{a}_i / \partial \mathbf{r}_k = (\delta_{ik} - 1/|C|)\mathbf{I}$ and $\partial r / \partial \mathbf{r}_k = \mathbf{r}/(r|C|)$. Then Eq. (9) reduces to

$$\mathbf{f}^p(\mathbf{r}_k) = -\frac{\partial E^p(C,D)}{\partial \mathbf{a}_k} - \frac{1}{|C|} \left[\frac{\partial E^p(C,D)}{\partial \mathbf{r}} - \sum_{i=1}^{|C|} \frac{\partial E^p(C,D)}{\partial \mathbf{a}_i} \right]. \tag{10}$$

Inserting explicit energy expression (6) for $E^p(C,D)$, one first derives the identity

$$\sum_{i=1}^{|C|} \frac{\partial E^p(C,D)}{\partial \mathbf{a}_i} = \frac{\partial}{\partial \mathbf{r}} E^{p-1}(C,D). \tag{11}$$

Next, one collects the second and third term on the r.h.s. of Eq. (10), up to the common weighting factor $1/|C|$, into the cluster-cluster force

$$\mathbf{F}^p(C,D) \equiv -\frac{\partial}{\partial \mathbf{r}} \left[E^p(C,D) - E^{p-1}(C,D) \right]. \tag{12}$$

This force is generated by D, acts on C, is directed along the vector $\mathbf{r}$ connecting the two cluster centers, and becomes uniformly distributed [as is witnessed by the constant weight $1/|C|$ in Eq. (10)] among all the atoms in C. For the first term on the r.h.s. of Eq. (10), one finds that it is essentially given by the two-sided Taylor expansion $\mathbf{E}^{D,p}(\mathbf{r}_k)$ of the electric field specified by Eq. (14) in Ref. 44, i.e.,

$$-\frac{\partial E^p(C,D)}{\partial \mathbf{a}_k} = q_k \mathbf{E}^{D,p}(\mathbf{r}_k). \tag{13}$$

Thus, force (7), which was defined as the negative gradient of the approximate energy $E^p(C,D)$, is finally given by

$$\mathbf{f}^p(\mathbf{r}_k) = q_k \mathbf{E}^{D,p}(\mathbf{r}_k) + \frac{1}{|C|} \mathbf{F}^p(C,D). \tag{14}$$

Despite the approximate character of the SAMM energy expressions, the use of such forces in HADES/SAMM-MD simulations should lead to a dynamics, which exactly conserves the total energy as well as the total linear and angular momenta as long as the so-called interaction lists are unchanged.

All FMM algorithms use lists, which sort the atoms into a hierarchy of nested clusters. This sorting may change upon atomic motions. For reasons of efficiency, it is periodically updated in MD algorithms after a (rather large and) predefined number of steps during the numerical integration of the dynamics. If the interactions among the clusters are calculated by FMM expansions, which sacrifice the achievable limit of numerical accuracy in favor of an enhanced computational efficiency, then the force description will exhibit, at the time points of the interaction list updates, small temporal discontinuities, which act as algorithmic noise. Thus, such efficient FMM algorithms cannot exactly conserve the energy. Whether the total momenta are conserved remains to be checked.

Note here that the usual FMM approaches, which employ spatially fixed grids to set up the required nested hierarchy of atomic clusters, cannot conserve the total angular momentum (even if they operate at numerical accuracy), because these grids break the required isotropy of the systems. No such grid is employed by SAMM for the construction of the FMM hierarchy. Instead, a structure adapted partitioning is applied[45,46,48] to a given molecular simulation system, which generates a quaternary tree of nested atomic clusters moving with the atoms. Therefore, the rotational symmetry of an isotropic system (as provided, e.g., by HADES for a protein in a dielectric continuum) is preserved.

Note, furthermore, that Eq. (7) does not represent the only way how one can compute reasonably accurate FMM forces for efficient MD simulations. One may, for instance, also directly use the forces[44]

$$\boldsymbol{\varphi}^p(\mathbf{r}_k) \equiv q_k \mathbf{E}^{D,p}(\mathbf{r}_k), \tag{15}$$

which derive from the separate Taylor expansion $\mathbf{E}^{D,p}(\mathbf{r}_k)$ of the electric field $\mathbf{E}^D(\mathbf{r}_k)$ generated by cluster D and acting on the atoms $k \in C$. Compared with the energy conserving forces $\mathbf{f}^p(\mathbf{r}_k)$, the resulting SAMM algorithm shows in simulations of periodic condensed phase systems a strongly enhanced efficiency without significant losses of accuracy. Section S2 of the supplementary material[50] illustrates this issue using liquid water as a relevant example.

For HADES/SAMM-MD simulations, however, the use of forces obeying Eq. (7) is vital, because they can, as we will show, quite accurately preserve the conservation laws, although the long-distance SAMM approximations are applied to speed up continuum simulations of large systems. If one employs, instead, non-Hamiltonian forces like the forces $\boldsymbol{\varphi}^p(\mathbf{r}_k)$ in HADES/SAMM-MD, the simulated protein will start to rotate as we will show below in Section IV. Most importantly, however, and as we will also demonstrate there too, the use of HADES/SAMM-MD leads to a linear scaling of the computational effort with the number of protein atoms.

The above arguments on the advantage of the use of electrostatic forces (14), which, by construction, conserve the SAMM energies $E^p(C,D)$, obviously apply to other non-bonded interactions (electrostatic dipoles, dispersion attraction, soft core repulsion), if they are also treated by symmetric two-sided Taylor expansions analogous to Eq. (4) and if the respective forces $\mathbf{f}(\mathbf{r}_k)$ are derived like in Eq. (7) from associated SAMM energy expressions as negative gradients.

As is shown in Sections S1.2-S1.4 of the supplementary material,[50] all these other forces assume for the geometry depicted in Fig. 1 the form $\mathbf{f}(\mathbf{r}_k) = \boldsymbol{\varphi}(\mathbf{r}_k) + \mathbf{F}(C,D)/|C|$. With Eqs. (13) and (15), one may recognize that this force expression is analogous to that provided by Eq. (14) for the pth order electrostatics force $\mathbf{f}^p(\mathbf{r}_k)$. The general expression for the SAMM force $\mathbf{f}(\mathbf{r}_k)$ implies that a scaled cluster-cluster force $\mathbf{F}(C,D)/|C|$, which is analogous to the one defined by Eq. (12) for clusters C and D of charges, always serves to correct the local force $\boldsymbol{\varphi}(\mathbf{r}_k) \equiv -\partial E_{\mathrm{SAMM}}(C,D)/\partial \mathbf{a}_k$ toward energy conservation.

Here, we finally should emphasize that SAMM does not directly work with two-sided Taylor expansions of potentials and fields, for which Eq. (4) represents a relevant example. Instead, these expansions are transformed into multipole expansions for cluster D and one-sided Taylor expansions of the multipole potentials generated by cluster D around the center $\mathbf{c}$ of cluster C. The mathematics of this transformation has been previously specified for the electrostatic potential $\Phi^{D,p}(\mathbf{r}_k)$ and field $\mathbf{E}^{D,p}(\mathbf{r}_k)$ in Ref. 44. For the electrostatic cluster-cluster force $\mathbf{F}^p(C,D)$, it is given in Section S1.1 of the supplementary material.[50] This section additionally explains how the computation of the electrostatic forces $\mathbf{F}^p(C,D)$ is beneficially integrated into the organization of computations up and down a logical tree structure, which characterizes FMM algorithms.

III. METHODS

Algorithmic properties of HADES/SAMM-MD were studied using CHARMM22[2] simulation models of the proteins, which are listed together with their protein data bank[51] (PDB) codes in Table I. The models were constructed from the PDB files with the help of the program package VMD.[52]

All HADES/SAMM-MD simulations were based on fourth order ($p = 4$), third order ($q = 3$), and first order ($r = 1$) FMM expansions for the electrostatics, $1/r^6$ Lennard-Jones attraction, and $1/r^{12}$ repulsion, respectively. The vacuum dielectric constant $\epsilon_s = 1$ was assigned to the interior of the considered proteins and the constant $\epsilon_c = 80$, which models the dielectrics of an aqueous solvent, to their surroundings. The dynamics was numerically integrated by the velocity Verlet (vV) algorithm[58] with a time step $\Delta t_0 = 1$ fs. Lengths of bonds involving hydrogen atoms were constrained by the RATTLE[59] and MSHAKE[60] algorithms with relative tolerances of 10^{-10}.

We chose the set $\chi^{\max}$ of convergence thresholds defined in Ref. 14 for the various HADES self-consistency iterations. In particular, we chose for the convergence threshold χ_p, which provides for all induced RF dipoles $\tilde{\mathbf{p}}_i$ an upper limit for the absolute change of any Cartesian component during the self-consistency iteration, a value of 10^{-7} D. Here, the polarizing field $\langle \tilde{\mathbf{E}}(\mathbf{r}_i) \rangle_{\sigma_i}$ generated by sufficiently distant atomic sources is calculated by FMM. Because the molecule is static during the iterations, only the contributions of the induced RF dipoles to the electric field are iteratively updated. Further details on the algorithms applied to speed up the dipole convergence are given in Ref. 14.

The threshold Θ, which enters the "interaction acceptance criterion" (IAC) and steers the SAMM accuracy,[45] was chosen as $\Theta_{\mathrm{m}} = 0.20$. This value represents a reasonable

TABLE I. Simulated proteins.

Protein	PDB entry	N	Reference
BLUF domain (AppA)	1YRX	1692	53
H-Ras p21	2CE2	2609	54
LpxA	1LXA	3945	55
RhoA	1OW3	6003	56
GBP1	1DG3	8769	57

compromise between accuracy and efficiency.[45] During the HADES/SAMM-MD simulations, the interaction list updates and re-clusterings[45] were usually performed every 64 fs and 256 fs, respectively.

To quantify the violation of conservation laws, which could arise in HADES/SAMM-MD simulations of proteins from the periodic interaction list updates and re-clusterings[45] (cf. the discussion in Section II), we carried out several sets $\mathcal{S}$ of HADES-MD simulations with durations of $\Delta T \equiv 100$ ps on H-Ras p21 at 300 K. Each set $\mathcal{S}$ comprised 5 independent MD trajectories, employed either the original HADES forces (reference set $\mathcal{R}$), the HADES/SAMM forces $\mathbf{f}(\mathbf{r}_k)$ ("Hamiltonian" set $\mathcal{H}$), or the more efficient HADES/SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ ("efficient" set $\mathcal{E}$). All sets $\mathcal{S}$ started at the same initial conditions (proteins at 300 K, exactly vanishing total linear and angular momenta).

We extracted from each simulation the trajectories $O(t)$ of various observables O, which comprised the total and rotational energies per atom (E and E_{rot}) and the total linear and angular momenta ($\mathbf{P}$ and $\mathbf{L}$). For the energies we calculated linear regressions, which yielded trajectory averages of the energy drifts $\langle dO/dt \rangle_{\Delta T}$. By further averaging over the elements of each set $\mathcal{S}$, we obtained energy drift data $\langle\langle dO/dt \rangle_{\Delta T}\rangle_{\mathcal{S}}$, which characterize the underlying simulation algorithm. For the total energy E, in particular, the average drift $\langle\langle dE/dt \rangle_{\Delta T}\rangle_{\mathcal{S}}$ is the algorithmic heating rate $\dot{Q}_{\mathcal{S}}$ per atom. From the trajectories of the momenta, we calculated by numerical differentiation the absolute values of the average total force $\langle|\langle \mathbf{F}\rangle_{\Delta T}|\rangle_{\mathcal{S}} \equiv \langle|\langle d\mathbf{P}/dt\rangle_{\Delta T}|\rangle_{\mathcal{S}}$ and torque $\langle|\langle \mathbf{M}\rangle_{\Delta T}|\rangle_{\mathcal{S}} \equiv \langle|\langle d\mathbf{L}/dt\rangle_{\Delta T}|\rangle_{\mathcal{S}}$ acting on average in the simulation set $\mathcal{S}$ on H-Ras p21. For $\mathbf{L}$, we additionally calculated set average trajectories $\langle|\mathbf{L}(t)|\rangle_{\mathcal{S}}$.

The scaling behavior of the computational effort was determined from short (1 ps) HADES/SAMM-MD simulations on each of the proteins listed in Table I. As a reference, we chose the original HADES forces (simulation set $\mathcal{R}_{\text{scal}}$; here the subscript "scal" points to the aim of these simulations, which is the determination of the scaling behavior). Furthermore, we considered the HADES/SAMM forces $\mathbf{f}(\mathbf{r}_k)$ (set $\mathcal{H}_{\text{scal}}$) and their supposedly more efficient approximations $\boldsymbol{\varphi}(\mathbf{r}_k)$ (set $\mathcal{E}_{\text{scal}}$). We extracted from each trajectory the average computing time T per integration step. The trajectory of the smallest protein (BLUF domain of AppA), which comprises, according to Table I, $N_{\text{min}} = 1692$ atoms, yields in the most efficient simulation set $\mathcal{E}_{\text{scal}}$ the shortest time T_{min} per integration step. Computation times and protein sizes are then conveniently measured by the dimensionless quantities

$$\tau \equiv T/T_{\text{min}} \quad \text{and} \quad \nu \equiv N/N_{\text{min}}, \tag{16}$$

respectively. The scaling behavior of a simulation method can then be determined by plotting the quotient τ/ν as a function of ν for each studied simulation set $\mathcal{S}_{\text{scal}}$ with $\mathcal{S} \in \{\mathcal{R},\mathcal{H},\mathcal{E}\}$ (see Fig. 4 further below).

MD simulations of a strictly conservative system, which employ the vV algorithm for the integration of the dynamics, do not conserve the energy but a different quantity called "shadow Hamiltonian."[61] This numerical conservation law can be checked by considering the fluctuations

$$F(t\,|\,\Delta t,\Delta t_0) \equiv [E(t\,|\,\Delta t) - \langle E(t\,|\,\Delta t)\rangle]\,/(\Delta t/\Delta t_0)^2 \tag{17}$$

of the total energy $E(t\,|\,\Delta t)$ scaled by the square of the time step Δt (measured in units of a reference step size Δt_0) as a function of the simulation time t, which are given by[61]

$$F(t\,|\,\Delta t,\Delta t_0) = f(t)\Delta t_0^2 + O(\Delta t^2), \tag{18}$$

i.e., are represented for identical initial conditions by a universal fluctuation function $f(t)\Delta t_0^2$ (up to corrections vanishing with Δt^2). Thus, energy conservation can be checked by superimposing the plots of $F(t\,|\,\Delta t,\Delta t_0)$ from two simulations with small but different time steps and identical initial conditions. For reasonably short simulation time spans (e.g., 0.5 ps), the results should be identical.

In Ref. 14, the Hamiltonian character of HADES-MD has been validated by this approach for the small dipeptide Ac-Ala-NHMe. Here, the much stricter set χ_{ini} of HADES self-consistency convergence thresholds[14] has been employed in combination with the MSHAKE[60] and RATTLE[59] tolerances of 10^{-14}. We adopted these settings also for our most sensitive energy conservation check concerning the SAMM forces $\mathbf{f}(\mathbf{r}_k)$. Its database was formed by two 0.5 ps HADES/SAMM-MD simulations of H-Ras p21 at 300 K, which started at identical initial conditions and used the time steps $\Delta t_0/2$ and $\Delta t_0/4$, respectively. In both simulations, the interaction lists were kept fixed thus avoiding algorithmic noise induced by updates. We denote these simulations as the set $\mathcal{H}_{\text{check}}$.

IV. RESULTS

For our most sensitive check of energy conservation in HADES/SAMM-MD, we chose the H-Ras p21 protein. Its diameter of ≈40 Å is large enough that the employed IAC threshold Θ_{m} assigns a large fraction of the non-bonded interactions to the approximate description by SAMM expansions. Scaled energy fluctuations $F(t\,|\,\Delta t,\Delta t_0)$ as defined by Eq. (17) were extracted from the simulation set $\mathcal{H}_{\text{check}}$ introduced above. They are plotted in Figure 2 on top of each

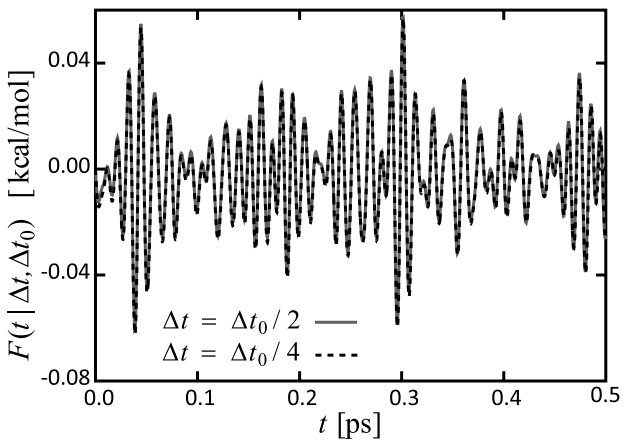


FIG. 2. Scaled energy fluctuations $F(t\,|\,\Delta t,\Delta t_0)$ in the two HADES/SAMM-MD simulations of H-Ras p21 with the time steps $\Delta t_0/4$ (black dotted) and $\Delta t_0/2$ (gray solid). The simulations used the supposedly energy conserving forces $\mathbf{f}(\mathbf{r}_k)$ and started at identical conditions.

other (black dotted: $\Delta t_0/4$; gray solid: $\Delta t_0/2$) as functions of the simulation time t.

Figure 2 demonstrates by the almost prefect match of the black dotted and gray solid curves that the scaled energy fluctuations $F(t\,|\,\Delta t, \Delta t_0)$ of the two HADES/SAMM-MD simulations represent, independently of the time step size Δt, the same function $f(t)\Delta t_0^2$. Thus, they fulfill the condition, which is posed by Eq. (18) to a Hamiltonian system, whose dynamics is numerically integrated by the vV algorithm. This result proves the energy conservation by the SAMM forces $\mathbf{f}(\mathbf{r}_k)$. Note that the check documented by Fig. 2 has been shown to be extremely sensitive[14,49] to violations of energy conservation. As a corollary we may, therefore, state that the algebraic derivations of the electrostatic and Lennard-Jones SAMM forces $\mathbf{f}(\mathbf{r}_k)$, which are presented in Section II and in Sections S1.1-S1.4 of the supplementary material,[50] are as correct as their implementation in the MD program IPHIGENIE.[62]

As follows from the discussion of interaction list updates in Section II, the strict energy conservation documented by Figure 2 is a consequence of the fixed interaction lists employed in the very short trajectories of $\mathcal{H}_{\text{check}}$. In a realistic simulation scenario, which covers time spans that are by many orders of magnitude larger, the interaction lists and the cluster structure must be frequently updated. For a check of the conservation laws under these conditions, we now discuss the results of the simulation sets $\mathcal{S} \in \{\mathcal{R}, \mathcal{H}, \mathcal{E}\}$, which were also introduced in Section III and likewise deal with H-Ras p21.

Figure 3 compares the average trajectories $\langle|\mathbf{L}(t)|\rangle_{\mathcal{S}}$ of the total angular momentum's absolute value obtained from the simulation sets $\mathcal{S} \in \{\mathcal{H}, \mathcal{E}\}$. The black dashed line in the figure belongs to the simulation set $\mathcal{E}$ and demonstrates that the more efficient forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ exert an almost constant torque of size $\langle|\langle\mathbf{M}\rangle_{\Delta T}|\rangle_{\mathcal{E}} \approx 320$ uÅ^2/ps^2 on the protein, where u is the atomic mass unit. This sizable torque leads to an increase of the rotational energy per atom by about $\langle\langle dE_{\text{rot}}/dt\rangle_{\Delta T}\rangle_{\mathcal{E}} \approx 0.002$ kcal/(mol ns). The forces $\mathbf{f}(\mathbf{r}_k)$, in contrast, quite accurately conserve the total angular momentum as is witnessed by the almost constant gray line extracted from the simulation set $\mathcal{H}$. Its slope nearly vanishes and corresponds to an absolute value $\langle|\langle\mathbf{M}\rangle_{\Delta T}|\rangle_{\mathcal{H}}$ of the torque, which is smaller than 10^{-5} uÅ^2/ps^2. Correspondingly,

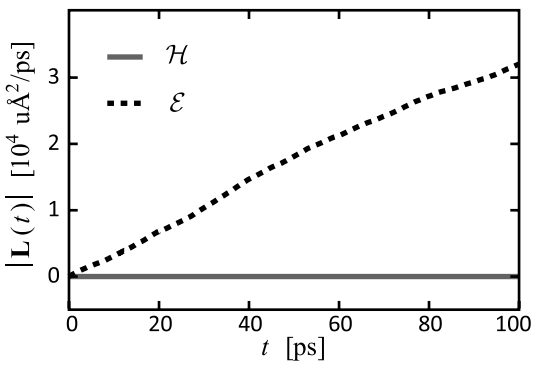


FIG. 3. Average trajectories $\langle|\mathbf{L}(t)|\rangle_{\mathcal{S}}$ of the absolute value of the angular momentum obtained for H-Ras p21 by HADES/SAMM-MD for the SAMM energy conserving forces $\mathbf{f}(\mathbf{r}_k)$ (gray solid) and the more efficient approximations $\boldsymbol{\varphi}(\mathbf{r}_k)$ (black dashed), respectively.

the forces $\mathbf{f}(\mathbf{r}_k)$ conserve the rotational energy per atom with numerical accuracy $[\langle\langle dE_{\text{rot}}/dt\rangle_{\Delta T}\rangle_{\mathcal{E}} < 10^{-17}$ kcal/(mol ns)]. This result clearly indicates that the newly derived forces $\mathbf{f}(\mathbf{r}_k)$ are highly useful for HADES/SAMM-MD simulations.

We would like to note that the average trajectory $\langle|\mathbf{L}(t)|\rangle_{\mathcal{R}}$, which we obtained from the HADES-MD reference simulations $\mathcal{R}$, cannot be distinguished from the gray solid trajectory representing $\mathcal{H}$ and, therefore, is not shown in Figure 3. For the reference method, the absolute value $\langle|\langle\mathbf{M}\rangle_{\Delta T}|\rangle_{\mathcal{R}}$ of the torque and the average increase $\langle\langle dE_{\text{rot}}/dt\rangle_{\Delta T}\rangle_{\mathcal{R}}$ of the rotational energy were also always smaller than 10^{-5} uÅ^2/ps^2 and 10^{-17} kcal/(mol ns), respectively. Thus, these upper bounds are identical to those obtained for the Hamiltonian HADES/SAMM forces $\mathbf{f}(\mathbf{r}_k)$ and, therefore, indicate the limits of numerical accuracy, which can be achieved with any HADES-MD approach, concerning the conservation of the total angular momentum.

As explained in Section III, we extracted further observables from the simulation sets $\mathcal{S}$. Accordingly, the average absolute value $\langle|\langle\mathbf{F}\rangle_{\Delta T}|\rangle_{\mathcal{S}}$ of the total force turned out to be similarly small for all $\mathcal{S} \in \{\mathcal{R}, \mathcal{H}, \mathcal{E}\}$ as is demonstrated by the upper bounds 10^{-11}, 10^{-9}, and 10^{-9} uÅ/ps^2, respectively. Here, the HADES-MD result ($\mathcal{R}$) had to be expected, because all atomic forces derive from pair forces $\mathbf{f}_{ij}$ obeying Newtons third law,[14] i.e., $\mathbf{f}_{ij} = -\mathbf{f}_{ji}$. Apart from inevitable numerical errors they, therefore, do not generate a total force on the whole protein and conserve the total linear momentum. Also, the approximate atomic SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ can be expressed in terms of pair forces[43,44] with $\boldsymbol{\varphi}_{ij} = -\boldsymbol{\varphi}_{ij}$ and, therefore, the total force should vanish here too. In the case of the forces $\mathbf{f}(\mathbf{r}_k)$, the reaction principle holds for each cluster-cluster interaction, i.e., in the nomenclature of Fig. 1 one has $\sum_{i\in C}\mathbf{f}(\mathbf{r}_i) = -\sum_{j\in D}\mathbf{f}(\mathbf{r}_j)$, as one can show by analyzing Eqs. (7)-(14). This condition suffices, however, to make the total force vanish also in this case. Consequently, the slightly smaller upper bound detected for $\mathcal{R}$ as compared to those of $\mathcal{H}$ and $\mathcal{E}$ solely indicates that the limiting numerical accuracy of HADES-MD slightly beats that of HADES/SAMM-MD, because the latter approach comprises even more complex algorithmic procedures, which all add little contributions to the deterioration of the achievable limit of numerical accuracy.

After the above consideration of the total momenta, solely the total energy remains to be studied to complete our check, whether or to what extent algorithmic artifacts violate the conservation laws, which characterize Hamiltonian many-body systems. Here, the relevant observable is the average heating rate $\dot{Q}_{\mathcal{S}}$ per atom defined in Section III. For $\mathcal{S} \in \{\mathcal{R}, \mathcal{H}, \mathcal{E}\}$, we found the values 0.0003, −0.0476, −0.0468 kcal/(mol ns) with the standard deviations 0.0006, 0.0033, and 0.0027 kcal/(mol ns), respectively. Hence, the heating rates $\dot{Q}_{\mathcal{H}}$ and $\dot{Q}_{\mathcal{E}}$ of the two HADES/SAMM-MD methods are within the limits of statistical accuracy identical, represent a very small cooling, and their absolute values are only by two orders of magnitude larger than the reference heating rate $\dot{Q}_{\mathcal{R}}$. Because the latter belongs to the strictly energy conserving HADES-MD method, it represents for the given choice $\chi_{\max}$ of self-consistency thresholds the limit of numerical accuracy. In particular, it proves that

the algorithmic noise, which is caused by the numerically incomplete convergence (only single precision enforced) of the induced RF dipoles, is negligible as compared to the numerical FMM artifacts.

As a result, the SAMM algorithm with its very small temporal discontinuities of the force computation, which are caused by the periodically repeated interaction list updates and the re-clusterings in combination with the deliberately limited accuracy of the FMM expansions, entails a larger (though still very small) energy drift, which is, interestingly, independent of the force approximation $\mathbf{f}(\mathbf{r}_k)$ vs. $\boldsymbol{\varphi}(\mathbf{r}_k)$. The absolute value of this drift may be diminished, of course, by choosing for the IAC parameter Θ of SAMM the smaller value $\Theta_a = 0.17$, which should enhance the accuracy and reduce the efficiency.[45] Then the absolute value $|\dot{Q}_{\mathcal{H}}|$ of the algorithmic heating rate becomes actually smaller by a factor 1/4 (data not shown).

Summarizing we may thus state that, in a HADES/SAMM-MD setting, the key advantage of the (almost) Hamiltonian SAMM forces $\mathbf{f}(\mathbf{r}_k)$ is the conservation of the total angular momentum. As compared to the HADES-MD reference method documented by the simulations $\mathcal{R}$, the only disadvantage of the SAMM forces $\mathbf{f}(\mathbf{r}_k)$ is the noted slight cooling, which requires a compensation by a thermostat. If a minimally invasive thermostat[63] is chosen for this purpose, the perturbation of the Hamiltonian dynamics can be kept extremely small. Hence, we expect that long-time HADES/SAMM-MD simulations of very large systems will stably run at the temperature controlled by the minimally invasive thermostat in an almost Hamiltonian fashion.

The key advantage of and the only reason for the use of SAMM in HADES-MD is the greatly enhanced computational efficiency, which will be documented now by comparing the results of the trajectory sets $\mathcal{S}_{\text{scal}}$, $\mathcal{S} \in \{\mathcal{R},\mathcal{H},\mathcal{E}\}$, described in Section III on the observables for the computation time (τ) and protein size (ν), which are defined by Eq. (16).

Figure 4 shows the quotients τ/ν of the dimensionless computation times τ and dimensionless protein sizes ν together with associated regression lines as functions of ν for the simulation sets $\mathcal{R}_{\text{scal}}$ (gray circles), $\mathcal{H}_{\text{scal}}$ (black symbols "×"), and $\mathcal{E}_{\text{scal}}$ (light gray symbols "+"). The linear increase of the gray circles ($\mathcal{R}_{\text{scal}}$) with the system size ν demonstrates that the computational effort of the HADES-MD reference approach actually shows the expected quadratic scaling ($\sim N^2$) with the number N of protein atoms. For the two HADES/SAMM-MD approaches, which are represented by the simulations $\mathcal{H}_{\text{scal}}$ and $\mathcal{E}_{\text{scal}}$, the values of τ/ν are apparently constants as is demonstrated by the slopes of the black and light gray regression lines, whose absolute values are smaller than 0.06. Thus, the computation time of HADES/SAMM-MD scales linearly with N.

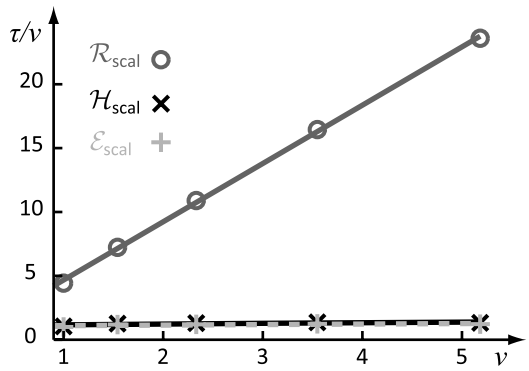


FIG. 4. Quotients τ/ν of computation times τ and system sizes ν defined by Eq. (16) were plotted over ν: the data were extracted from the HADES-MD reference simulation set $\mathcal{R}_{\text{scal}}$ (gray circles) and from the HADES/SAMM-MD simulation sets $\mathcal{H}_{\text{scal}}$ (black symbols "×") and $\mathcal{E}_{\text{scal}}$ (light gray symbols "+") explained in Section III. See the text for a discussion.

Additionally, one may recognize that the black line derived from $\mathcal{H}_{\text{scal}}$ is found at slightly larger values of τ/ν than the light gray line representing $\mathcal{E}_{\text{scal}}$. Consequently, the evaluation of the original SAMM forces[44] $\boldsymbol{\varphi}(\mathbf{r}_k)$ is actually more efficient than that of the "Hamiltonian" SAMM forces $\mathbf{f}(\mathbf{r}_k)$. Comparing the constant τ/ν values of $\mathcal{H}_{\text{scal}}$ and $\mathcal{E}_{\text{scal}}$, the efficiency reduction, which is caused by the additional computation of cluster-cluster forces (12), turns out to be 7.2%, i.e., the slope of the linear increase of the computation time τ with the protein size ν is by this percentage larger for $\mathcal{H}_{\text{scal}}$ than for $\mathcal{E}_{\text{scal}}$.

Interestingly, the computational advantage of using SAMM in HADES-MD can be detected already for the smallest protein ($N = 1692$) considered here, which is the BLUF domain of AppA. Here, even the use of the more costly SAMM forces $\mathbf{f}(\mathbf{r}_k)$ speeds up the calculation still by a factor of 4.2. This finding demonstrates that the computational overhead of the much more complex computational scheme of SAMM is surprisingly small. This advantage of HADES/SAMM-MD is largely preserved also for the more accurate variant of SAMM, which employs the smaller IAC threshold Θ_a. Then, the speed-up still has the sizable value of 3.1.

V. SUMMARY AND DISCUSSION

In principle, HADES-MD as developed by Bauer *et al.*[10,14] should be a viable and accurate method for the simulation of proteins embedded in dielectric solvent continua, because it can approximately solve the PE on the fly with the integration of the molecular dynamics and because it automatically includes the reaction forces exerted by the continuum on the protein atoms, i.e., because it represents the proteins as closed and fully Hamiltonian many-body systems. To become a continuum MD simulation approach, which is viable also for applications to large proteins or protein complexes, the problem of its computational complexity, which scales quadratically with the number of protein atoms,[31] had to be solved.

Following a corresponding suggestion and building upon earlier work,[31] we have therefore combined HADES-MD with the FMM approach[45] called SAMM. As expected, this combination transformed the quadratic scaling of the original HADES-MD algorithm[14] into a linear one. The computational advantage turned out to be sizable (factor 3-4) already for quite small proteins like the BLUF domain[53] of AppA. A simple combination of the existing SAMM approach[44,45] with HADES-MD would have, however, destroyed the Hamiltonian character of this simulation method.

Here, the most notable violation of the Hamiltonian conservation laws is that concerning the total angular momentum of the simulated protein, which we have demonstrated by the black dashed curve in Fig. 3 for HADES/SAMM-MD. By representing the SAMM forces $\mathbf{f}(\mathbf{r}_k)$ acting on the atoms as exact negative gradients of the SAMM potential energies, we could now develop a revised HADES/SAMM-MD approach, which preserves the Hamiltonian character of the simulation method and, in particular, the total angular momentum (cf. the gray solid curve in Fig. 3). The additional computational effort, which has to be spent for this purpose, turned out to be quite small (≈7%). Concerning the energy, the preservation of the Hamiltonian character is only approximate, if one tunes the SAMM approach for high efficiency, and becomes more and more strict by tuning SAMM for enhanced accuracy.

The herewith generated Hamiltonian version of SAMM applies to all non-bonded interactions, although the above text solely deals with the electrostatics of point charge distributions. The highly similar mathematics applying to distributions of electrostatic dipoles, to dispersion ($\sim 1/r^6$), and soft-core repulsion ($\sim 1/r^{12}$) interactions is presented in the supplementary material.[50] Particularly for the soft-core repulsion, Section S1.5 of the supplementary material[50] proves for the example of an aqueous simulation system that replacing the usual short distance cutoff by first order SAMM substantially reduces the cutoff-induced algorithmic heating.

The HADES/SAMM-MD simulation method still awaits long term applications to large soluble proteins, which go far beyond the unfolding simulation previously carried out with HADES-MD on a small peptide.[31] Despite the considerable insensitivity of HADES to the choice of its parameters,[14] which are[10] the Gaussian atomic widths σ_i and a certain scaling factor ζ, a further fine-tuning of these parameters should also be of interest. Following the arguments put forward in Ref. 64, one could try to jointly optimize these parameters by comparisons of RF forces computed for a protein by HADES/SAMM-MD with electrostatic forces exerted in SAMM-MD simulations of a periodic boundary simulation system by an explicitly modeled aqueous solvent on the partially charged atoms of this protein.

Corresponding simulations are open to anybody in the scientific community, because the MD program package IPHIGENIE,[62] which is available online for download, offers HADES/SAMM-MD as one of its options. Note that this download also covers two HADES/SAMM-MD examples, which are the CHARMM22[2] models of the 35 residue villin-headpeace[65] and of the 592 residue human guanylate-binding protein 1.[57]

The supplementary material[50] additionally shows that for all-atom condensed phase MD simulations, e.g., of proteins in aqueous solution, the SAMM energy conserving forces solely reduce the efficiency and enhance the accuracy by certain small amounts, but do not induce any sizable benefits. For HADES/SAMM-MD, in contrast, the use of these forces is highly beneficial, because it renders corrections of the otherwise steadily growing total angular momentum superfluous. Other important benefits arise in hybrid MD simulations, in which a subsystem is treated quantum-mechanically and the large remainder by a (polarizable) MM force field.[66]

ACKNOWLEDGMENTS

This work was supported by the Deutsche Forschungsgemeinschaft (Grant No. SFB749/C4) and by the Bavarian Competence Network for Technical and Scientific High Performance Computing (KONWIHR-III). We thank Magnus Schwörer and Christoph Wichmann for valuable discussions and technical support.

[1]W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell, and P. A. Kollman, J. Am. Chem. Soc. **117**, 5179 (1995).
[2]A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus, J. Phys. Chem. B **102**, 3586 (1998).
[3]C. Oostenbrink, A. Villa, A. Mark, and W. Van Gunsteren, J. Comput. Chem. **25**, 1656 (2004).
[4]A. D. MacKerell, J. Comput. Chem. **25**, 1584 (2004).
[5]W. F. van Gunsteren, D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt, and H. B. Yu, Angew. Chem., Int. Ed. **45**, 4064 (2006).
[6]G. A. Cisneros, M. Karttunen, P. Ren, and C. Sagui, Chem. Rev. **114**, 779 (2014).
[7]W. Weber, P. H. Hünenberger, and J. A. McCammon, J. Phys. Chem. B **104**, 3668 (2000).
[8]G. Mathias, B. Egwolf, M. Nonella, and P. Tavan, J. Chem. Phys. **118**, 10847 (2003).
[9]P. Tavan, H. Carstens, and G. Mathias, "Molecular dynamics simulations of proteins and peptides: Problems, achievements, and perspectives," in *Protein Folding Handbook*, edited by J. Buchner and T. Kiefhaber (Wiley-VCH, Weinheim, 2005), Vol. 1, pp. 1170–1195.
[10]S. Bauer, G. Mathias, and P. Tavan, J. Chem. Phys. **140**, 104102 (2014).
[11]R. Jimenez, G. Fleming, P. Kumar, and M. Maroncelli, Nature **369**, 471 (1994).
[12]R. Geney, M. Layten, R. Gomperts, V. Hornak, and C. Simmerling, J. Chem. Theory Comput. **2**, 115 (2006).
[13]J. Wang, C. Tan, E. Chanco, and R. Luo, Phys. Chem. Chem. Phys. **12**, 1194 (2010).
[14]S. Bauer, P. Tavan, and G. Mathias, J. Chem. Phys. **140**, 104103 (2014).
[15]R. Zhou and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **99**, 12777 (2002).
[16]H. Nymeyer and A. E. García, Proc. Natl. Acad. Sci. U. S. A. **100**, 13934 (2003).
[17]M. Feig, A. Onufriev, M. S. Lee, W. Im, D. A. Case, and C. L. Brooks III, J. Comput. Chem. **25**, 265 (2003).
[18]N. Levy, D. Borgis, and M. Marchi, Comput. Phys. Commun. **169**, 69 (2005).
[19]M. Stork and P. Tavan, J. Chem. Phys. **126**, 165105 (2007).
[20]D. Bashford and A. Case, Annu. Rev. Phys. Chem. **51**, 129 (2000).
[21]T. Grycuk, J. Chem. Phys. **119**, 4817 (2003).
[22]A. Fenley, J. C. Gordon, and A. Onufriev, J. Chem. Phys. **129**, 075101 (2008).
[23]A. Onufriev, D. A. Case, and D. Bashford, J. Comput. Chem. **23**, 1297 (2002).
[24]B. Egwolf and P. Tavan, J. Chem. Phys. **118**, 2039 (2003).
[25]M. E. Davis, J. Chem. Phys. **100**, 5149 (1994).
[26]L. David and M. J. Field, Chem. Phys. Lett. **245**, 371 (1995).
[27]W. Im, D. Beglov, and B. Roux, Comput. Phys. Commun. **111**, 59 (1997).
[28]W. Geng and G. W. Wei, J. Comput. Phys. **230**, 435 (2011).
[29]M. Marchi, D. Borgis, N. Levy, and P. Ballone, J. Chem. Phys. **114**, 4377 (2001).
[30]H. Sklenar, F. Eisenhaber, M. Poncin, and R. Lavery, in *Theoretical Biochemistry & Molecular Biophysics, 2. Proteins*, edited by D. L. Beveridge and R. Lavery (Adenine Press, New York, 1991), pp. 317–335.
[31]S. Bauer, P. Tavan, and G. Mathias, Chem. Phys. Lett. **612**, 20 (2014).

[32]J. Barnes and P. Hut, Nature **324**, 446 (1986).
[33]L. Greengard and V. Rokhlin, J. Comput. Phys. **73**, 325 (1987).
[34]H.-Q. Ding, N. Karasawa, and W. A. Goddard III, J. Chem. Phys. **97**, 4309 (1992).
[35]F. Figueirido, R. M. Levy, R. Zhuo, and B. J. Berne, J. Chem. Phys. **106**, 9835 (1997).
[36]K.-T. Lim, S. Brunett, M. Iotov, R. B. McClurg, N. Vaidehi, S. Dasgupta, S. Taylor, and W. A. Goddard, J. Comput. Chem. **18**, 501 (1997).
[37]K. Z. Takahashi, T. Narumi, and K. Yasuoka, J. Chem. Phys. **135**, 174108 (2011).
[38]K. Z. Takahashi, T. Narumi, D. Suh, and K. Yasuoka, J. Chem. Theory Comput. **8**, 4503 (2012).
[39]Y. Andoh, N. Yoshii, K. Fujimoto, K. Mizutani, H. Kojima, A. Yamada, S. Okazaki, K. Kawaguchi, H. Nagao, K. Iwahashi, F. Mizutani, K. Minami, S.-I. Ichikawa, H. Komatsu, S. Ishizuki, Y. Takeda, and M. Fukushima, J. Chem. Theory Comput. **9**, 3201 (2013).
[40]Y. Ohno, R. Yokota, H. Koyama, G. Morimoto, A. Hasegawa, G. Masumoto, N. Okimoto, Y. Hirano, H. Ibeid, T. Narumi, and M. Taiji, Comput. Phys. Commun. **185**, 2575 (2014).
[41]M. S. Warren and J. K. Salmon, Comput. Phys. Commun. **87**, 266 (1995).
[42]W. Dehnen, Astrophys. J. **536**, L39 (2000).
[43]W. Dehnen, J. Comput. Phys. **179**, 27 (2002).
[44]K. Lorenzen, M. Schwörer, P. Tröster, S. Mates, and P. Tavan, J. Chem. Theory Comput. **8**, 3628 (2012).
[45]K. Lorenzen, C. Wichmann, and P. Tavan, J. Chem. Theory Comput. **10**, 3244–3259 (2014).
[46]C. Niedermeier and P. Tavan, J. Chem. Phys. **101**, 734 (1994).
[47]C. Niedermeier and P. Tavan, Mol. Simul. **17**, 57 (1996).
[48]M. Eichinger, H. Grubmüller, H. Heller, and P. Tavan, J. Comput. Chem. **18**, 1729 (1997).
[49]M. Schwörer, B. Breitenfeld, P. Tröster, K. Lorenzen, P. Tavan, and G. Mathias, J. Chem. Phys. **138**, 244103 (2013).
[50]See supplementary material at http://dx.doi.org/10.1063/1.4935514 for which provides on 18 pages in two sections a total of two figures (S5 and S6) and 29 equations (S18-S46) as additional material to the derivation of the Hamiltonian SAMM forces in Section 2 and to the simulation results presented in Section 4.
[51]H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, Nucleic Acids Res. **28**, 235 (2000).
[52]W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[53]S. Anderson, V. Dragnea, S. Masuda, J. Ybe, K. Moffat, and C. Bauer, Biochemistry **44**, 7998 (2005).
[54]B. U. Klink, R. S. Goody, and A. J. Scheidig, Biophys. J. **91**, 981 (2006).
[55]C. R. H. Raetz and S. L. Roderick, Science **270**, 997 (1995).
[56]D. L. Graham, P. N. Lowe, G. W. Grime, M. Marsh, K. Rittinger, S. J. Smerdon, S. J. Gamblin, and J. F. Eccleston, Chem. Biol. **9**, 375 (2002).
[57]B. Prakash, G. J. K. Praefcke, L. Renault, A. Wittinghofer, and C. Herrmann, Nature **403**, 567 (2000).
[58]W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson, J. Chem. Phys. **76**, 637 (1982).
[59]H. C. Andersen, J. Comput. Phys. **52**, 24 (1983).
[60]V. Kräutler, W. F. van Gunsteren, and P. H. Hünenberger, J. Comput. Chem. **22**, 501 (2001).
[61]M. E. Tuckerman, in *Statistical Mechanics: Theory and Molecular Simulation*, 1st ed. (Oxford University Press, New York, USA, 2010), Chap. 3.13, pp. 121–124.
[62]IPHIGENIE is available for download free of charge under the GPL licence at http://sourceforge.net/projects/iphigenie.
[63]M. Lingenheil, R. Denschlag, R. Reichold, and P. Tavan, J. Chem. Theory Comput. **4**, 1293 (2008).
[64]M. Stork and P. Tavan, J. Chem. Phys. **126**, 165106 (2007).
[65]C. McKnight, P. Matsudaira, and P. Kim, Nat. Struct. Biol. **4**, 180 (1997).
[66]M. Schwörer, K. Lorenzen, G. Mathias, and P. Tavan, J. Chem. Phys. **142**, 104108 (2015).

Der folgende Abdruck[6]

„Supplementary Material to: Linearly Scaling and Almost Hamiltonian Dielectric Continuum Molecular Dynamics Simulations through Fast Multipole Expansions“

Konstantin Lorenzen, Gerald Mathias, and Paul Tavan
J. Chem. Phys. **143** 184114, (2015)

enthält zusätzliche Informationen zur Ableitung der Hamiltonschen SAMM-Kräfte und der Implementierung der Lennard-Jones Repulsion in das SAMM Schema. Außerdem wird der Nutzen der neuen Hamiltonschen SAMM-Kräfte auf Simulationen von expliziten Lösungsmittelsystemen in toroidalen Randbedingungen quantifiziert.

[6] Mit freundlicher Genehmigung des Verlags.

Supplementary Material to:

Linearly Scaling and Almost Hamiltonian Dielectric Continuum Molecular Dynamics Simulations through Fast Multipole Expansions

Konstantin Lorenzen, Gerald Mathias, and Paul Tavan*

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

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

S1 Energy Conserving Non-Bonded Forces in SAMM

The main text discusses solely those energy conserving SAMM forces $\mathbf{f}^p(\mathbf{r}_k)$, which are generated by point charge distributions. Furthermore, these p'th order atomic forces $\mathbf{f}^p(\mathbf{r}_k)$ are simply represented in terms of two-sided Taylor expansions. For their efficient use in computations, which take advantage of the hierarchically nested cluster structure by following the FMM tree bottom-up for the calculation of cluster multipole moments and top-down for the evaluation of cluster-local Taylor expansions, one needs a corresponding representation of the SAMM energies and forces.

*To whom correspondence should be addressed

For the p'th order electrostatic cluster-cluster interaction energies $E^p(C,D)$ [Eq. (6)] and the contributions $\boldsymbol{\varphi}^p(\mathbf{r}_k)$ [Eq. (15)] to the forces $\mathbf{f}^p(\mathbf{r}_k)$ [Eq. (14)], the required representation has been previously presented.[1] To complete the FMM mathematics of point charge distributions solely the new force contributions $\mathbf{F}^p(C,D)/|C|$ defined by Eq. (12) remain to be explicitly specified in terms of the n'th order multipole tensors $\mathbf{M}^{n,\mathbf{c}}$ and $\mathbf{M}^{n,\mathbf{d}}$ of the clusters C and D, respectively, and of the components of the vector $\mathbf{r}$ connecting these clusters (cf. Fig. 1 in the main text). This issue will be discussed in the next subsection.

Subsequently, we will complete the presentation of the energy conserving SAMM forces by first including the case of electrostatic dipole distributions, and by subsequently addressing the cases of the $\sim 1/r^6$ dispersion attraction and of the $\sim 1/r^{12}$ soft-core repulsion.

S1.1 The Electrostatic Cluster-Cluster Force $\mathbf{F}^p(C,D)$

For the specification of $\mathbf{F}^p(C,D)$ in terms of $\mathbf{M}^{n,\mathbf{c}}$, $\mathbf{M}^{n,\mathbf{d}}$, and $\mathbf{r}$ one firstly has to insert the energy expression Eq. (5) into Eq. (12). Using the definitions [see Eqs. (8) and (9) in Ref. 1] of the n'th order electrostatic multipole tensors $\mathbf{M}^{n,\mathbf{c}}$ of cluster C and of the m'th order multipole potentials $\phi^{m,D}(\mathbf{c})$ of cluster D one derives the identity

$$\frac{2^n}{(2n)!}\mathbf{M}^{n,\mathbf{c}} \odot \partial_{(n)}\phi^{p-n,D}(\mathbf{c}) = \frac{1}{n!}\sum_{i\in C} q_i\, \mathbf{a}_i^{(n)} \odot \partial_{(n)}\phi^{p-n,D}(\mathbf{c}). \tag{S18}$$

With this identity Eq. (12) can be rewritten in terms of multipole-multipole interactions, whose orders $n = 0,\, 1,\, \ldots, p$, and $p-n$ always add up to p. One finds

$$\mathbf{F}^p(C,D) = -\frac{\partial}{\partial \mathbf{r}}\sum_{n=0}^{p}\frac{2^n}{(2n)!}\mathbf{M}^{n,\mathbf{c}} \odot \partial_{(n)}\phi^{p-n,D}(\mathbf{c}). \tag{S19}$$

If one inserts the definition of the $(p-n)$'th order multipole potential $\phi^{p-n,D}(\mathbf{c})$ into Eq. (S19) and calculates the required derivatives $\partial_{(n+1)}\phi^{p-n,D}(\mathbf{c})$, then the multipole-multipole interactions of Eq. (S19) become expressed in terms of the unit vector $\hat{\mathbf{r}} = \mathbf{r}/r$ associated to the

connection vector $\mathbf{r}$ and of the multipole moments $\mathbf{M}^{n,\mathbf{c}}$ and $\mathbf{M}^{(p-n),\mathbf{d}}$.

For the order $p = 3$ one obtains the four summands

$$
\begin{aligned}
\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{3,D}(\mathbf{c}) &= \frac{1}{2}\frac{1}{r^5}\left[-\frac{7}{3}\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{d}}\right) + \hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{d}}\right]\mathbf{M}^{0,\mathbf{c}} \\
\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{2,D}(\mathbf{c}) &= \frac{5}{r^5}\left[\begin{array}{l}\frac{7}{2}\hat{\mathbf{r}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{c}}\right)\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) - \\ \hat{\mathbf{r}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}} \odot \mathbf{M}^{1,\mathbf{c}}\right) - \frac{1}{2}\mathbf{M}^{1\mathbf{c}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) - \\ \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right)\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{c}}\right) + \frac{1}{5}\left(\mathbf{M}^{1,\mathbf{c}} \odot \mathbf{M}^{2,\mathbf{d}}\right)\end{array}\right] \\
\frac{1}{6}\mathbf{M}^{2,\mathbf{c}} \odot \partial_{(3)}\phi^{1,D}(\mathbf{c}) &= -\frac{5}{r^5}\left[\begin{array}{l}\frac{7}{2}\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right)\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right) - \\ \hat{\mathbf{r}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}} \odot \mathbf{M}^{1,\mathbf{d}}\right) - \frac{1}{2}\mathbf{M}^{1,\mathbf{d}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right) - \\ \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right)\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right) + \frac{1}{5}\left(\mathbf{M}^{1,\mathbf{d}} \odot \mathbf{M}^{2,\mathbf{c}}\right)\end{array}\right] \\
\frac{1}{90}\mathbf{M}^{3,\mathbf{c}} \odot \partial_{(4)}\phi^{0,D}(\mathbf{c}) &= -\frac{1}{2}\frac{1}{r^5}\left[-\frac{7}{3}\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{c}}\right) + \hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{c}}\right]\mathbf{M}^{0,\mathbf{d}}
\end{aligned}
$$

For the order $p = 4$ the terms are

$$
\begin{aligned}
\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{4,D}(\mathbf{c}) &= \frac{1}{24}\frac{1}{r^6}\left[-9\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(4)} \odot \mathbf{M}^{4,\mathbf{d}}\right) + 4\mathbf{M}^{4,\mathbf{d}} \odot \hat{\mathbf{r}}^{(3)}\right]\mathbf{M}^{0,\mathbf{c}} \\
\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{3,D}(\mathbf{c}) &= \frac{7}{2}\frac{1}{r^6}\left[\begin{array}{l} 3\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{d}}\right)\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right) - \\ \hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{d}} \odot \mathbf{M}^{1,\mathbf{c}}\right) + \frac{2}{7}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{3,\mathbf{d}} \odot \mathbf{M}^{1,\mathbf{c}}\right) - \\ \frac{1}{3}\mathbf{M}^{1,\mathbf{c}}\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{d}}\right) - \left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{d}}\right)\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right) \end{array}\right] \\
\frac{1}{6}\mathbf{M}^{2,\mathbf{c}} \odot \partial_{(3)}\phi^{2,D}(\mathbf{c}) &= \frac{5}{6}\frac{1}{r^6}\left[\begin{array}{l} -\frac{63}{2}\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right)\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) + \\ 14\hat{\mathbf{r}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right) \odot \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right) - \hat{\mathbf{r}}\left(\mathbf{M}^{2,\mathbf{c}} \odot \mathbf{M}^{2\mathbf{d}}\right) + \\ 7\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right)\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) - 2\,\mathbf{M}^{2,\mathbf{c}} \odot \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right) + \\ 7\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right)\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right) - 2\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right) \odot \mathbf{M}^{2,\mathbf{d}} \end{array}\right] \\
\frac{1}{90}\mathbf{M}^{3,\mathbf{c}} \odot \partial_{(4)}\phi^{1,D}(\mathbf{c}) &= \frac{7}{2}\frac{1}{r^6}\left[\begin{array}{l} 3\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right)\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{c}}\right) - \\ \hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{c}} \odot \mathbf{M}^{1,\mathbf{d}}\right) + \frac{2}{7}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{3,\mathbf{c}} \odot \mathbf{M}^{1,\mathbf{d}}\right) - \\ \frac{1}{3}\mathbf{M}^{1,\mathbf{d}}\left(\hat{\mathbf{r}}^{(3)} \odot\ \mathbf{M}^{3,\mathbf{c}}\right) - \left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{c}}\right)\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right) \end{array}\right] \\
\frac{2}{7!}\mathbf{M}^{4,\mathbf{c}} \odot \partial_{(5)}\phi^{0,D}(\mathbf{c}) &= \frac{1}{24}\frac{1}{r^6}\left[-9\,\hat{\mathbf{r}}\left(\mathbf{M}^{4,\mathbf{c}} \odot \hat{\mathbf{r}}^{(4)}\right) + 4\mathbf{M}^{4,\mathbf{c}} \odot \hat{\mathbf{r}}^{(3)}\right]\mathbf{M}^{0,\mathbf{d}}
\end{aligned}
$$

In SAMM, interactions calculated at upper levels of the hierarchy for large and distant clusters are transferred to the contained smaller clusters, the children, at the next lower level by a shifting of upper level Taylor expansions to the centers of the child clusters and by a subsequent addition to the Taylor expansions expressing the lower-level interactions.[1] This so-called inheritance process is particularly simple for the the cluster-cluster forces $\mathbf{F}^p(C, D)$, which are exerted according to Eq. (14) by large clusters D on the atoms of another large cluster C at a given hierarchy level. All these forces are inherited to the enclosed child clusters $c \in C$ according to

$$\sum_D \mathbf{F}^p(c, D) = \frac{|c|}{|C|}\sum_D \mathbf{F}^p(C, D) \tag{S20}$$

and are added to the cluster-cluster forces $\mathbf{F}^p(c,b)$ exerted by lower level clusters b on the children c. This inheritance process is repeated until the lowest (atomic) level of the hierarchy is reached.

S1.2 Extension of SAMM to Point Dipole Distributions

According to the introductory remarks in Section II of the main text, the FMM expansion of the long-range HADES electrostatics requires SAMM expansions for clusters not only of point charges q_j but also of point dipoles $\mathbf{p}_j$.

For the FMM interaction geometry depicted by Fig. 1 in the main text, the p'th order SAMM energy thus consists according to

$$E^p(C,D) = E^p_{qq}(C,D) + E^p_{q\mathbf{p}}(C,D) + E^p_{\mathbf{pp}}(C,D) \tag{S21}$$

of contributions for charge-charge (qq), charge-dipole ($q\mathbf{p}$), and dipole-dipole ($\mathbf{pp}$) interactions. With the potentials $\phi^{m,D}(\mathbf{c})$ [cf. Eq. (9) in Ref. 1] generated at the center $\mathbf{c}$ of cluster C by the m'th order multipole tensors $\mathbf{M}^{m,\mathbf{d}}$ of the charge distribution in cluster D, the SAMM energy $E^p_{qq}(C,D)$ of the charges q_i at the local positions $\mathbf{a}_i$ in cluster C is given by the p'th order Taylor expansion

$$E^p_{qq}(C,D) = \sum_{i\in C}\sum_{n=0}^{p} \frac{1}{n!}\mathbf{a}_i^{(n)} \odot \partial_{(n)} \sum_{m=0}^{p-n} q_i \Phi^{m,D}(\mathbf{c}) \tag{S22}$$

around the center $\mathbf{c}$ of cluster C. This expression is, of course, strictly equivalent to Eq. (6), which employs however two-sided Taylor expansions for $E^p_{qq}(C,D)$ and denotes this energy simply as $E^p(C,D)$.

In a similar fashion one may formulate the p'th order energy $E^p_{q\mathbf{p}}(C,D)$ of the charges q_i in the multipole potentials $\Phi^{m,D}_{\mathbf{p}}(\mathbf{c})$, which are generated by the m'th order multipole tensors $\tilde{\mathbf{M}}^{m,\mathbf{d}}$ of the dipole distribution in cluster D, and of the dipoles $\mathbf{p}_i$ in the fields

$-\partial_{(1)}\Phi^{m-1,D}(\mathbf{c})$, which originate from the $(m-1)$'th order multipole tensors $\mathbf{M}^{m,\mathbf{d}}$ of the charge distribution in D, as the Taylor expansion

$$E^p_{q\mathbf{p}}(C,D) = \sum_{i\in C}\sum_{n=0}^{p-1}\frac{1}{n!}\mathbf{a}_i^{(n)} \odot \partial_{(n)} \sum_{m=1}^{p-n}\left[q_i\Phi^{m,D}_{\mathbf{p}}(\mathbf{c}) + \mathbf{p}_i \odot \partial_{(1)}\Phi^{m-1,D}(\mathbf{c})\right]. \tag{S23}$$

Note here that the definition of the m'th order multipole potentials $\Phi^{m,D}_{\mathbf{p}}(\mathbf{c})$, which belong to the multipole tensors $\tilde{\mathbf{M}}^{m,\mathbf{d}}$ of dipoles, is formally identical to the definition [Eq. (9) in Ref. 1] of the potentials $\Phi^{m,D}(\mathbf{c})$, which are caused by the multipole tensors $\mathbf{M}^{m,\mathbf{d}}$ of charges. For the dipole-dipole interaction energy one analogously gets

$$E^p_{\mathbf{pp}}(C,D) = \sum_{i\in C}\sum_{n=0}^{p-2}\frac{1}{n!}\mathbf{a}_i^{(n)} \odot \partial_{(n)} \sum_{m=2}^{p-n}\mathbf{p}_i \odot \partial_{(1)}\Phi^{m-1,D}_{\mathbf{p}}(\mathbf{c}). \tag{S24}$$

Formulas, for the recursive computation of higher order multipole tensors $\mathbf{M}^{m,\mathbf{d}}$ for charge and $\tilde{\mathbf{M}}^{m,\mathbf{d}}$ for dipole distributions from lower order counterparts are given in Sec. 2.4 of Ref. 1 and in the appendix to Ref. 2, respectively. Ref. 1 furthermore explains, how multipole tensors of larger clusters are computed in a bottom-up fashion along the FMM tree from the tensors of the enclosed sub-clusters by simple addition.

The derivation of p'th order forces $\mathbf{f}^p(\mathbf{r}_k)$, which conserve the electrostatic SAMM cluster-cluster interaction energy $E^p(C,D)$, as negative gradients of that energy with respect to the atomic position $\mathbf{r}_k$ of an atom k in cluster C is identical to that presented in Section II of the main text for its charge-charge component $E^p_{qq}(C,D)$. Starting from the definition (7) one arrives at Eq. (10), which assumes, with the definition (12) of the cluster-cluster force $\mathbf{F}^p(C,D)$, the form

$$\mathbf{f}^p(\mathbf{r}_k) = \boldsymbol{\varphi}^p(\mathbf{r}_k) + \mathbf{F}^p(C,D)/|C|, \tag{S25}$$

where we have defined the local force $\boldsymbol{\varphi}^p(\mathbf{r}_k)$ more generally as

$$\boldsymbol{\varphi}^p(\mathbf{r}_k) \equiv -\frac{\partial E^p(C,D)}{\partial \mathbf{a}_k}. \tag{S26}$$

For the point charge case $E^p_{qq}(C,D)$ this definition reduces to $q\mathbf{E}$-force of Eq. (13), of course.

As a result, the electrostatic SAMM forces $\mathbf{f}^p(\mathbf{r}_k)$ and $\boldsymbol{\varphi}^p(\mathbf{r}_k)$ specified by Eqs. (S25) and (S26) have exactly the form, which was asserted at the bottom of Section II in the main text.

Note that the electrostatic cluster-cluster force $\mathbf{F}^p(C,D)$, which is given by Eq. (12) of the main text, can be expressed also in the more general case considered here in terms of multipole tensors and multipole potentials. The more restricted case of point charge distributions has been discussed above in Section S1.1 and has led to Eq. (S19). The analogous general expression is

$$\mathbf{F}^p(C,D) = -\frac{\partial}{\partial \mathbf{r}} \sum_{n=0}^{p} \frac{2^n}{(2n)!} \left[\mathbf{M}^{n,\mathbf{c}} + \tilde{\mathbf{M}}^{n,\mathbf{c}} \right] \odot \partial_{(n)} \left[\Phi^{p-n,D}(\mathbf{c}) + \Phi^{p-n,D}_{\mathbf{p}}(\mathbf{c}) \right], \tag{S27}$$

where we have assigned a vanishing monopole moment $\tilde{\mathbf{M}}^0 = 0$ to dipole distributions. The more explicit expressions given above in Section S1.1 remain likewise formally unchanged.

S1.3 Energy Conserving SAMM Forces for the $1/r^6$ Dispersion

A SAMM treatment of the $1/r^6$ dispersion attraction has been presented in Ref. 3 for the FMM cluster-cluster interaction scenario depicted by Fig. 1 of the main text. Accordingly, the atoms j in cluster D, carry dispersion charges B_j, which generate at the positions $\mathbf{r_i}$ of the atoms i in C the dispersion potential $\Phi^D(\mathbf{r}_i) = -\sum_{j\in D} B_j/|\mathbf{r}_i - \mathbf{s}_j|^6$, such that dispersive cluster-cluster energy is $E(C,D) = \sum_{i\in C} B_i \Phi^D(\mathbf{r}_i)$, where the B_i are the dispersion charges in cluster C. It can be approximated by the q'th order two-sided Taylor expansion

$$E^q(C,D) = -\sum_{i\in C} B_i \sum_{n=0}^{q} \frac{1}{n!} \partial_{(n)} \frac{1}{r^6} \odot \sum_{j\in D} B_j \left(\mathbf{a}_i - \mathbf{b}_j\right)^{(n)}, \tag{S28}$$

which is strictly analogous to Eq. (6) describing the electrostatics of point charge clusters.

In complete analogy to this electrostatics case (see Section II of the main text) one can

now represent the SAMM forces

$$\mathbf{f}^q(\mathbf{r}_k) \equiv -\frac{\partial}{\partial \mathbf{r}_k} E^q(C,D), \tag{S29}$$

conserving the q'th order SAMM dispersion cluster-cluster energy $E^q(C,D)$ through

$$\mathbf{f}^q(\mathbf{r}_k) = \boldsymbol{\varphi}^q(\mathbf{r}_k) + \mathbf{F}^q(C,D)/|C| \tag{S30}$$

as a combination of the local forces

$$\boldsymbol{\varphi}^q(\mathbf{r}_k) \equiv -\frac{\partial}{\partial \mathbf{a}_k} E^q(C,D) \tag{S31}$$

and of the dispersive cluster-cluster attraction

$$\mathbf{F}^q(C,D) \equiv -\frac{\partial}{\partial \mathbf{r}} \left[E^q(C,D) - E^{q-1}(C,D) \right]. \tag{S32}$$

The mathematics of the local forces $\boldsymbol{\varphi}^q(\mathbf{r}_k)$ has been presented in Ref. 3. Thus, solely the force $\mathbf{F}^q(C,D)$ remains to be specified and we may restrict our attention to expansion orders $q \leq 3$, because the dispersion is much more short ranged than the electrostatics.[3]

For the representation of $\mathbf{F}^q(C,D)$ in terms of m'th order dispersive multipole tensors $\mathbf{M}^{m,\mathbf{d}}$ [defined by Eq. (10) in Ref. 3] and of the connection vector $\mathbf{r}$ (cf. the above Section S1.1 for the electrostatics case) one has to obey one caveat, i.e. in the case of the dispersion the traces of the tensors $\mathbf{M}^{m,\mathbf{d}}$ do not vanish whereas in the electrostatics case[4,5] they do.

As a result, the expressions for the dispersive q'th order cluster-cluster force

$$\mathbf{F}^q(C,D) = \frac{\partial}{\partial \mathbf{r}} \sum_{i \in C} B_i \sum_{n=0}^{q} \frac{1}{n!} \mathbf{a}_i^{(n)} \odot \partial_{(n)} \phi^{q-n,D}(\mathbf{c}), \tag{S33}$$

where $\phi^{m,D}(\mathbf{c})$ denotes the dispersive potential at the center $\mathbf{c}$ of cluster C generated by the m'th order dispersive multipole moment $\mathbf{M}^{m,\mathbf{d}}$ of cluster D, become slightly more compli-

cated. One finds for the orders $q = 1, 2, 3$ the forces

$$\begin{aligned}
\mathbf{F}^1(C,D) &= \mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{1,D}(\mathbf{c}) + \frac{1}{6}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{0,D}(\mathbf{c}), && \text{(S34)}\\
\mathbf{F}^2(C,D) &= \mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{2,D}(\mathbf{c}) - \mathbf{M}^{0,\mathbf{d}} \odot \partial_{(1)}\phi^{2,C}(\mathbf{d}) + \frac{1}{6}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{1,D}(\mathbf{c}), && \text{(S35)}\\
\mathbf{F}^3(C,D) &= \mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{3,D}(\mathbf{c}) - \mathbf{M}^{0,\mathbf{d}} \odot \partial_{(1)}\phi^{3,C}(\mathbf{d})\\
&\quad + \frac{1}{6}\left[\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{2,D}(\mathbf{c}) - \mathbf{M}^{1,\mathbf{d}} \odot \partial_{(2)}\phi^{2,C}(\mathbf{d})\right], && \text{(S36)}
\end{aligned}$$

which solely contain low order ($m \leq 1$) multipole moments $\mathbf{M}^{m,\mathbf{d}}$ and $\mathbf{M}^{m,\mathbf{c}}$.

Using the notation of the above Section S1.1 we find for the terms in Eq. (S34) the explicit representations

$$\begin{aligned}
\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{1,D}(\mathbf{c}) &= \frac{1}{r^8}\left[-8\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right) + \mathbf{M}^{1,\mathbf{d}}\right]\mathbf{M}^{0,\mathbf{c}}\\
\frac{1}{6}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{0,D}(\mathbf{c}) &= -\frac{1}{r^8}\left[-8\,\hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right) + \mathbf{M}^{1,\mathbf{c}}\right]\mathbf{M}^{0,\mathbf{d}}.
\end{aligned}$$

For the terms in Eq. (S35) we get

$$\begin{aligned}
\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{2,D}(\mathbf{c}) &= \frac{1}{r^9}\left[-5\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) + \hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right]\mathbf{M}^{0,\mathbf{c}}\\
\mathbf{M}^{0,\mathbf{d}} \odot \partial_{(1)}\phi^{2,C}(\mathbf{d}) &= -\frac{1}{r^9}\left[-5\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right) + \hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right]\mathbf{M}^{0,\mathbf{d}}\\
\frac{1}{6}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{1,D}(\mathbf{c}) &= \frac{8}{6}\frac{1}{r^9}\begin{bmatrix} 10\,\hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right)\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right) - \mathbf{M}^{1,\mathbf{c}}\left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}}\right) \\ -\hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{c}} \odot \mathbf{M}^{1,\mathbf{d}}\right) - \mathbf{M}^{1,\mathbf{d}}\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right) \end{bmatrix}
\end{aligned}$$

and, finally, for the terms in Eq. (S36)

$$
\begin{aligned}
\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{3,D}(\mathbf{c}) &= \frac{1}{r^{10}}\left[-2\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{d}}\right) + \frac{1}{2}\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{d}}\right]\mathbf{M}^{0,\mathbf{c}} \\
\mathbf{M}^{0,\mathbf{d}} \odot \partial_{(1)}\phi^{3,C}(\mathbf{d}) &= \frac{1}{r^{10}}\left[-2\,\hat{\mathbf{r}}\left(\hat{\mathbf{r}}^{(3)} \odot \mathbf{M}^{3,\mathbf{c}}\right) + \frac{1}{2}\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{3,\mathbf{c}}\right]\mathbf{M}^{0,\mathbf{d}} \\
\frac{1}{6}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{2,D}(\mathbf{c}) &= \frac{5}{3}\frac{1}{r^{10}}\left[\begin{array}{l} 6\,\hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right)\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) - \\ \frac{1}{2}\mathbf{M}^{1,\mathbf{c}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{d}}\right) - \hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{c}} \odot \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right)\right) - \\ \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{d}}\right)\left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}}\right) + \frac{1}{10}\left(\mathbf{M}^{1,\mathbf{c}} \odot \mathbf{M}^{2,\mathbf{d}}\right) \end{array}\right] \\
\frac{1}{6}\mathbf{M}^{1,\mathbf{d}} \odot \partial_{(2)}\phi^{2,C}(\mathbf{d}) &= \frac{5}{3}\frac{1}{r^{10}}\left[\begin{array}{l} 6\,\hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{d}} \odot \hat{\mathbf{r}}\right)\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right) - \\ \frac{1}{2}\mathbf{M}^{1,\mathbf{d}}\left(\hat{\mathbf{r}}^{(2)} \odot \mathbf{M}^{2,\mathbf{c}}\right) - \hat{\mathbf{r}}\left(\mathbf{M}^{1,\mathbf{d}} \odot \left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right)\right) - \\ \left(\mathbf{M}^{1,\mathbf{d}} \odot \hat{\mathbf{r}}\right)\left(\hat{\mathbf{r}} \odot \mathbf{M}^{2,\mathbf{c}}\right) + \frac{1}{10}\left(\mathbf{M}^{1,\mathbf{d}} \odot \mathbf{M}^{2,\mathbf{c}}\right) \end{array}\right].
\end{aligned}
$$

S1.4 Including the $1/r^{12}$ Lennard-Jones Repulsion into SAMM

Assuming that the short range $1/r^{12}$ Lennard-Jones (LJ) repulsion is sufficiently small at distances r larger than the radius d_0, within which all non-bonded interactions are calculated from exact pair expressions, these interaction were simply cutoff at d_0 in previous versions[1,3,6,7] of SAMM. For the current revision of our MD program IPHIGENIE,[8] which was motivated by HADES, we decided to include a first order SAMM expansion also for the LJ repulsion.

Here we assume that the parameters A_{ij}, which specify in MM force fields the Lennard-Jones repulsion energy $A_{ij}/|\mathbf{r}_i - \mathbf{s}_j|^{12}$ of a non-bonded pair of atoms i and j, obey the product decomposition $A_{ij} = A_i\,A_j$ into atomic repulsion charges A_i and A_j, respectively. Assuming once again the FMM geometry depicted by Fig. 1, we see that the repulsion charges A_j in cluster D generate at the positions $\mathbf{r}_i$ of the repulsion charges A_i in cluster C the potentials $\Phi^D(\mathbf{r}_i) \equiv \sum_{j\in D} A_j/|\mathbf{r}_i - \mathbf{s}_j|^{12}$. The total repulsion energy of the cluster pair C and D then is $E(C,D) = \sum_{i\in C} A_i\,\Phi^D(\mathbf{r}_i)$.

Due to the rapid decay of $\Phi^D(\mathbf{r}_i)$ with the distances $r_{ij} \equiv |\mathbf{r}_i - \mathbf{s}_j|$, a first order two-sided

Taylor expansion of $E(C,D)$ around the vector $\mathbf{r}$ connecting the cluster centers $\mathbf{c}$ and $\mathbf{d}$ (cf. Fig. 1) should suffice for a reasonably accurate SAMM approximation, i.e.

$$E^1(C,D) = \sum_{i\in C} A_i \left[\frac{1}{r^{12}} \sum_{j\in D} A_j + \partial_{(1)} \frac{1}{r^{12}} \odot \sum_{j\in D} A_j \left(\mathbf{a}_i - \mathbf{b}_j\right) \right]. \tag{S37}$$

With the definition of the zeroth and first order repulsive multipole tensors

$$\mathbf{M}^{0,\mathbf{d}} \equiv \sum_{j\in D} A_j \quad \text{and} \quad \mathbf{M}^{1,\mathbf{d}} \equiv 12 \sum_{j\in D} A_j \mathbf{b}_j \tag{S38}$$

of cluster D, which generate the corresponding multipole potentials

$$\phi^{0,D}(\mathbf{c}) \equiv \frac{1}{r^{12}} \mathbf{M}^{0,\mathbf{d}} \quad \text{and} \quad \phi^{1,D}(\mathbf{c}) \equiv \frac{1}{r^{14}} \mathbf{r} \odot \mathbf{M}^{1,\mathbf{d}}, \tag{S39}$$

the first order Taylor expansion $E^1(C,D)$ can be rewritten as

$$E^1(C,D) = \sum_{i\in C} A_i \left[\phi^{0,D}(\mathbf{c}) + \phi^{1,D}(\mathbf{c}) + \mathbf{a}_i \odot \partial_{(1)}\, \phi^{0,D}(\mathbf{c}) \right]. \tag{S40}$$

This representation of $E^1(C,D)$ has the typical FMM form,[3,9] because the evaluation of the m'th order multipole potentials $\phi^{m,D}(\mathbf{c})$ of cluster D and of their derivatives at the center $\mathbf{c}$ of the target cluster C yields the coefficients of a one-sided Taylor expansion, which is evaluated at the positions $\mathbf{r}_i$ of the atoms $i \in C$.

In close analogy to Eq. (14) the Hamiltonian force $\mathbf{f}^1(\mathbf{r}_k) \equiv -\partial E^1(C,D)/\partial \mathbf{r}_k$ exerted by cluster D on an atom $k \in C$ is

$$\mathbf{f}^1(\mathbf{r}_k) \equiv \boldsymbol{\varphi}^1(\mathbf{r}_k) + \mathbf{F}^1(C,D)/|C|, \tag{S41}$$

where the local forces $\boldsymbol{\varphi}^1(\mathbf{r}_k) \equiv -\partial E^1(C,D)/\partial \mathbf{a}_k$ are given by

$$\boldsymbol{\varphi}^1(\mathbf{r}_k) = -A_k \partial_{(1)} \phi^{0,D}(\mathbf{c}). \tag{S42}$$

This is the force on the repulsion charge A_k in the monopole-field $-\partial_{(1)}\phi^{0,D}(\mathbf{c})$, which is generated by cluster D and is evaluated at the center $\mathbf{c}$ of cluster C. The equally weighted the cluster-cluster forces $\mathbf{F}^1(C,D)$ in Eq. (S41) are

$$\mathbf{F}^1(C,D) \equiv -\frac{\partial}{\partial \mathbf{r}} \left[E^1(C,D) - E^0(C,D) \right], \tag{S43}$$

where $E^1(C,D)$ is specified by Eq. (S40) and $E^0(C,D)$ by

$$E^0(C,D) \equiv \mathbf{M}^{0,\mathbf{c}} \phi^{0,D}(\mathbf{c}). \tag{S44}$$

Thus, one gets the more explicit expression

$$\mathbf{F}^1(C,D) = -\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{1,D} - \frac{1}{12}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{0,D}, \tag{S45}$$

with

$$\begin{aligned} -\mathbf{M}^{0,\mathbf{c}} \odot \partial_{(1)}\phi^{1,D} &= \frac{1}{r^{14}} \left[14\, \hat{\mathbf{r}} \left(\hat{\mathbf{r}} \odot \mathbf{M}^{1,\mathbf{d}} \right) - \mathbf{M}^{1,\mathbf{d}} \right] \mathbf{M}^{0,\mathbf{c}} \\ -\frac{1}{12}\mathbf{M}^{1,\mathbf{c}} \odot \partial_{(2)}\phi^{0,D} &= -\frac{1}{r^{14}} \left[14\, \hat{\mathbf{r}} \left(\mathbf{M}^{1,\mathbf{c}} \odot \hat{\mathbf{r}} \right) - \mathbf{M}^{1,\mathbf{c}} \right] \mathbf{M}^{0,\mathbf{d}}, \end{aligned}$$

where we have used the notation introduced further above in Section S1.1.

S1.5 The Benefit of Describing the LJ Repulsion by SAMM

We have carried out and evaluated sample MD simulations on TIP3P water[10] using exactly the methods described in Ref. 3. In short: We chose a cubic simulation system with a box-length $L \approx 40$ Å enclosed by toroidal[11] boundary conditions. We combined the explicit description of the electrostatics for distances smaller than the minimum image[11] radius $d_{\mathrm{MIC}} = L/2$ with a reaction field (RF) approach[7] modeling a distant dielectric continuum (dielectric constant $\epsilon_{\mathrm{RF}} = 78$) extending beyond d_{MIC}. The system was prepared at the

experimental density of water at $T_0 = 298.15$ K and normal pressure. It was equilibrated at T_0 for 1 ns by MD using the velocity Verlet algorithm[12] with an integration time step of 1 fs and a Bussi thermostat[13] with a relaxation time of 0.5 ps (cf. Ref. 3).

Also the measurements of differential algorithmic heat production rates per molecule

$$\Delta\dot{Q}(\Theta, \mathcal{S}) = \dot{Q}(\Theta, \mathcal{S}) - \dot{Q}(\Theta, \mathcal{R}), \tag{S46}$$

were executed in the same way as previously.[3] They were derived for different simulation setups $\mathcal{S}$ at varying values of the IAC threshold Θ, which steers the SAMM accuracy,[3] and for a reference setup $\mathcal{R}$ from ensembles of micro-canonical MD simulations, each covering 100 short 10 ps trajectories. The observed changes of the total energies yielded the heat production rates $\dot{Q}$.

In the reference setup $\mathcal{R}$ all non-bonded interactions were calculated (within the explicit interaction spheres of radius d_{MIC}) from the exact pair expressions. The simulation setups $\mathcal{S}$ were designed to quantify the impact of the inclusion of the LJ repulsion into SAMM on the algorithmic noise. Therefore, they differed from $\mathcal{R}$ only in the treatment of the LJ repulsion. Setup $\mathcal{C}_{\mathrm{rep}}$ applies the cutoff for molecular distances larger than $d_0 = 2R_{\mathrm{TIP3P}}/\Theta$, where $R_{\mathrm{TIP3P}} = 0.67$ Å is the radius of gyration of the TIP3P model. In the setups $\mathcal{E}_{\mathrm{rep}}$ and $\mathcal{H}_{\mathrm{rep}}$ the long range parts of the LJ repulsion forces were modeled by the efficient $\boldsymbol{\varphi}^1(\mathbf{r}_k)$ and by the Hamiltonian SAMM forces $\mathbf{f}^1(\mathbf{r}_k)$, respectively.

Figure S5 compares the differential heating rate $\Delta\dot{Q}(\Theta)$ of the the cutoff setup $\mathcal{C}_{\mathrm{rep}}$ (dashed curve) with that of the SAMM setup $\mathcal{E}_{\mathrm{rep}}$ (crosses) for a wide range of IAC thresholds Θ covering the standard values $\{\Theta_{\mathrm{a}}, \Theta_{\mathrm{m}}, \Theta_{\mathrm{f}}\} = \{0.17, 0.2, 0.25\}$. Apparently, the repulsive SAMM forces $\boldsymbol{\varphi}^1(\mathbf{r}_k)$ reduce the cutoff-induced differential heat production by about a factor of five over the whole range of shown IAC thresholds Θ. We did not add the results of setup $\mathcal{H}_{\mathrm{rep}}$, which refers to the Hamiltonian SAMM forces $\mathbf{f}^1(\mathbf{r}_k)$, to the figure, because they would be visually indistinguishable from the shown data (crosses), which represent the setup $\mathcal{E}_{\mathrm{rep}}$.

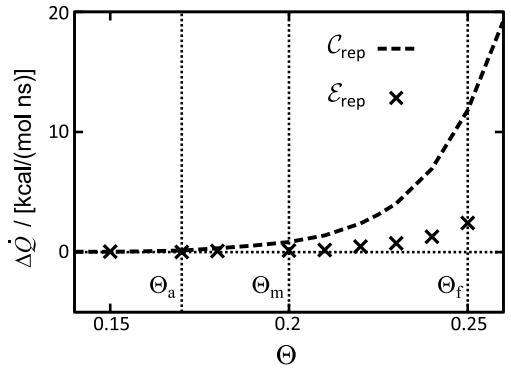


Figure S5: In the TIP3P test system the differential heat production $\Delta\dot{Q}(\Theta)$, which originates from the cutoff of the $1/r^{12}$ LJ repulsion at $d_0(\Theta)$ (dashed curve, $\mathcal{C}_{\rm rep}$), is diminished for all Θ by the first order SAMM forces $\boldsymbol{\varphi}^1(\mathbf{r}_k)$ acting at distances beyond $d_0(\Theta)$ (crosses, $\mathcal{E}_{\rm rep}$).

We conclude that our first order SAMM approximations of the repulsive LJ forces remove in our TIP3P test system at least 80 % of the cutoff-induced heat production and, in this respect, there is no difference between the forces $\boldsymbol{\varphi}^1(\mathbf{r}_k)$ and $\mathbf{f}^1(\mathbf{r}_k)$. Solely concerning the computational cost there is a slight difference between the two, because the computational cost[14] increases by about 0.5 % when the "efficient" forces $\boldsymbol{\varphi}^1(\mathbf{r}_k)$ are replaced by their Hamiltonian counterparts $\mathbf{f}^1(\mathbf{r}_k)$. Because of the strong suppression of algorithmic noise we chose the first order SAMM description of the repulsive LJ forces as our new default. Therefore, the HADES/SAMM simulations presented in the main text employed this new default.

S2 Water: No Benefits of Energy Conservation

Addressing HADES/SAMM-MD simulations of proteins embedded in a dielectric continuum, the main text of this paper has demonstrated that the conservation of the total angular momentum is the key advantage of the Hamiltonian forces $\mathbf{f}(\mathbf{r}_k)$ over their approximate counterparts $\boldsymbol{\varphi}(\mathbf{r}_k)$. Therefore the use of the $\mathbf{f}(\mathbf{r}_k)$ is mandatory here. Concerning the amount of algorithmic noise, however, no significant differences seemed to arise from using the simple forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ instead of the computationally more demanding forces $\mathbf{f}(\mathbf{r}_k)$ (cf. Section IV in the main text).

In contrast to the isolated HADES simulation systems, classical particle systems with periodic boundary conditions in general do not exhibit a conserved total angular momentum, because they do not represent isolated systems.[15] In view of the apparently similar amounts of algorithmic noise observed in HADES/SAMM-MD simulations with the two types of SAMM forces, the use of the more costly Hamiltonian forces $\mathbf{f}(\mathbf{r}_k)$ may seem superfluous. On the other hand, the similarity of algorithmic noise production may not be transferable to condensed phase all-atom solvent-solute simulation systems, whose electrostatics is dominated by the very small and polar water molecules. Therefore, we decided to address this issue using the TIP3P water model system introduced above in Section S1.5 and the differential heating rates $\Delta\dot{Q}(\Theta, \mathcal{S})$ from Eq. (S46) for a corresponding test.

Following once again the procedures described above in above in Section S1.5 we used the setup $\mathcal{R}$ as our reference. The reference heating rates $\dot{Q}(\Theta, \mathcal{R})$, which refer to the evaluation of all non-bonded interactions by exact pair expressions for distances smaller than d_{MIC}, were compared with two SAMM/RF-MD simulation setups $\mathcal{H}$ and $\mathcal{E}$, which employed the Hamiltonian $[\mathbf{f}(\mathbf{r}_k)]$ and efficient $[\boldsymbol{\varphi}(\mathbf{r}_k)]$ SAMM forces for distances smaller than d_{MIC}, respectively. The formation of the differential heating rates $\Delta\dot{Q}(\Theta, \mathcal{S})$ thus enabled precise estimates of the amount of algorithmic noise, which is caused by the two different SAMM forces in the TIP3P test system. The orders of the SAMM expansions were 4, 3, and 1 for the electrostatic ($\sim 1/r$), dispersive ($\sim 1/r^6$), and repulsive ($\sim 1/r^{12}$) interactions, respectively.

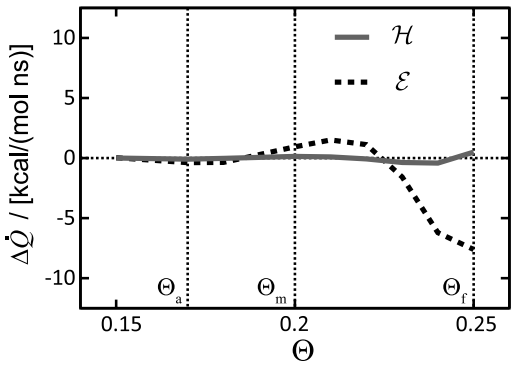


Figure S6: Algorithmic heating rates $\Delta\dot{Q}$ in the TIP3P test system for the Hamiltonian ($\mathcal{H}$, solid curve) and "efficient" ($\mathcal{E}$, dashed curve) SAMM forces at various values of Θ.

Addressing the TIP3P test system, Figure S6 demonstrates that the absolute values of the differential heating rates $|\Delta\dot{Q}(\Theta, \mathcal{H})|$, which belong to the Hamiltonian SAMM forces $\mathbf{f}(\mathbf{r}_k)$ (solid curve), are almost for all Θ smaller than those generated by the more approximate and efficient SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ (dashed curve).

For the maximal value $\Theta_{\rm f} = 0.25$, i.e. for the most inaccurate (and efficient) choice of the IAC parameter Θ, the Hamiltonian SAMM forces $\mathbf{f}(\mathbf{r}_k)$ produce a small residual heating of 0.5 Kcal/(mol ns) per molecule, which is comparable to the likewise small heating observed for the "efficient" SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ at the smaller value 0.225 of Θ. According to Figure 8 in Ref. 3 a dispersion cutoff at about 1.0 nm causes a cooling, whose absolute value is by one order of magnitude larger than these small heating rates. Because a dispersion cutoff at about 1.0 nm is common practice in other MD simulation packages implementing force fields like CHARMM22,[16] GROMOS,[17] OPLS,[18] or AMBER,[19] we are led to conclude that IPHIGENIE[8] produces with Hamiltonian SAMM forces $\mathbf{f}(\mathbf{r}_k)$ for $(p, q, r) = (4, 3, 1)$ a residual algorithmic noise in periodic simulation systems containing aqueous solutions, which is at least by one order of magnitude smaller than that produced by other MD simulation packages.

Interestingly, the computational effort, which has to be spent on the forces $\mathbf{f}(\mathbf{r}_k)$ at $\Theta_{\rm f}$, is still by 7 % larger than the one spent on the forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ for the smaller IAC parameter $\Theta = 0.225$ (which implies for these forces an enhanced accuracy and reduced efficiency in

comparison with the situation at Θ_f). We conclude that the Hamiltonian SAMM forces do not offer any significant performance gain over their more efficient predecessors, if the SAMM accuracy is properly tuned by the diligent choice of Θ. This statement proves the headline ("No Benefits of Energy Conservation") of this section.

Note that at Θ_f the computational effort of the SAMM forces $\mathbf{f}(\mathbf{r}_k)$ is by 30 % larger than that of the non-Hamiltonian SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$, which is a quite substantial efficiency loss. If one is willing to compensate the small residual cooling observed in Figure S6 at Θ_f by a suitable thermostat,[20] then the non-Hamiltonian SAMM forces even offer a distinct advantage.

The diligent reader may have noticed that the dashed curve measuring the differential heating rate of the "efficient" SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ as a function of Θ differs from the almost corresponding dotted curve in Fig. 11 of Ref. 3, which has been calculated with SAMM expansions of the orders 4 and 3 for the electrostatics and dispersion, respectively, and with a cutoff of the repulsion. The cutoff induced heating documented further above by Fig. S5 (dashed curve) compensates the cooling, which is documented by the dashed curve in Figure S6 for values $\Theta > 0.22$, and leads to the almost vanishing differential heating rate identified at Θ_f for the combination of the electrostatic and dispersive SAMM forces $\boldsymbol{\varphi}(\mathbf{r}_k)$ with a repulsion cutoff described in the predecessor paper.[3]

Finally, we would like to emphasize, that the enhanced accuracy of the Hamiltonian SAMM forces $\mathbf{f}(\mathbf{r}_k)$ becomes very useful[2,21] in DFT/PMM hybrid simulations. The somewhat larger computational cost required by these forces is irrelevant in this context, because the quantum-mechanical part of such a calculation typically consumes more than 95 % of the computer time.[21]

References

(1) Lorenzen, K.; Schwörer, M.; Tröster, P.; Mates, S.; Tavan, P. *J. Chem. Theory Comput.* **2012**, *8*, 3628–3636.

(2) Schwörer, M.; Breitenfeld, B.; Tröster, P.; Lorenzen, K.; Tavan, P.; Mathias, G. *J. Chem. Phys.* **2013**, *138*, 244103.

(3) Lorenzen, K.; Wichmann, C.; Tavan, P. *J. Chem. Theory Comput.* **2014**, *10*, 3244–3259.

(4) Hinsen, K.; Felderhof, B. U. *J. Math. Phys.* **1992**, *33*, 3731–3735.

(5) Buckingham, A. D. *Adv. Chem. Phys.* **1967**, *12*, 107–147.

(6) Eichinger, M.; Grubmüller, H.; Heller, H.; Tavan, P. *J. Comput. Chem.* **1997**, *18*, 1729–1749.

(7) Mathias, G.; Egwolf, B.; Nonella, M.; Tavan, P. *J. Chem. Phys.* **2003**, *118*, 10847–10860.

(8) IPHIGENIE is available for download free of charge under the GPL licence at `http://sourceforge.net/projects/iphigenie`.

(9) Dehnen, W. *J. Comput. Phys.* **2002**, *179*, 27–42.

(10) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(11) Allen, M. P.; Tildesley, D. *Computer Simulations of Liquids*; Clarendon: Oxford, 1987.

(12) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. *J. Chem. Phys.* **1982**, *76*, 637–649.

(13) Bussi, G.; Parrinello, M. *Comput. Phys. Commun.* **2008**, *179*, 26–29.

(14) Speed tests were performed on an Intel Core 2 Duo CPU (E8400) for periodic TIP3P water with an inner radius of $d_{\mathrm{MIC}} = 29$ Å. The systems was filled with 6503 water models and simulated at ambient temperature.

(15) Kuzkin, V. A. *Z. Angew. Math. Mech.* **2014**, doi: 10.1002/zamm.201400045.

(16) 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.

(17) Oostenbrink, C.; Villa, A.; Mark, A.; Van Gunsteren, W. *J. Comput. Chem. B* **2004**, *25*, 1656–1676.

(18) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. *J. Am. Chem. Soc.* **1996**, *118*, 11225–11236.

(19) 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.

(20) Lingenheil, M.; Denschlag, R.; Reichold, R.; Tavan, P. *J. Chem. Theory Comput.* **2008**, *4*, 1293–1306.

(21) Schwörer, M.; Lorenzen, K.; Mathias, G.; Tavan, P. *J. Chem. Phys.* **2015**, *142*, 104108.

3 Beiträge als Koautor

Im Rahmen meiner Doktorarbeit in der Arbeitsgruppe für theoretische Biophysik am Lehrstuhl für BioMolekulare Optik der LMU konnte ich durch meine Beiträge, die ich im Folgenden kurz skizzieren werde, Koautorenschaften bei Veröffentlichungen meiner Kollegen Magnus Schwörer [44, 45] und Philipp Tröster [54, 55] erlangen. Die beiden Veröffentlichungen von Magnus Schwörer betreffen hierbei die gründliche Neuentwicklung der DFT/PMM Kopplung für quantenmechanische/molekularmechanische Hybridsimulationen. Die Veröffentlichungen von Philipp Tröster beschreiben die Parametrisierung von zunehmend komplexen polarisierbaren Wassermodellen und stellen gleichzeitig eine Anwendung der genannten DFT/PMM Kopplung dar.

3.1 DFT/PMM Kopplung

In den zwei Veröffentlichungen [44, 45] wurde das (P)MM-MD Programm IPHIGENIE [53] mit dem gitterbasierten DFT Programm CPMD [29] kombiniert und damit ein modernes und Hamiltonsches DFT/(P)MM-MD Verfahren entwickelt.

In der ersten Veröffentlichung [44] konnten die von mir in [41] entwickelten SAMM Entwicklungen des elektrostatischen Potentials bis zur Hexadekapolordnung genutzt werden, um die Güte der Berechnung des externen Potentials im Bereich des DFT-Fragments stark zu verbessern und Unstetigkeiten im Potentialverlauf zu reduzieren. Bei der Implementierung der entsprechenden SAMM Entwicklungen in die IPHIGENIE/CPMD Schnittstelle konnte ich hierbei Magnus Schwörer zur Seite stehen.

In der zweiten Veröffentlichung [45] wurde nun der SAMM-Ansatz, der auf dem in [42] entwickelten Wechselwirkungskriterium IAC beruht, konsequent auf die Gitterpunkte des Raumgitters im Bereich des DFT-Fragments ausgedehnt. Dies schien geboten, da die Ausdehnung der Gitterpunkte, die den Atomen einer strukturellen Einheit zugeordnet werden, in der Regel sehr viel größer ist als die auf Atomkernpositionen beruhende Ausdehnung der strukturellen Einheit selbst. Um dieses Problem zu umgehen, hat Magnus Schwörer im Bereich des DFT-Gitters die SAMM-Hierarchie unterhalb der Ebene der strukturellen Einheiten ($l = 0$) um zwei weitere Ebenen erweitert. Zum einen werden sogenannte Voxel auf der Ebene $l = -2$ definiert, die das DFT-Gitter geometrisch disjunkt zerlegen und mehrere Gitterpunkte enthalten. Auf der Ebene $l = -1$ werden Gruppen von Voxeln gebildet, die die Vereinigung derjenigen Voxel repräsentieren, die über das Minimalabstandskriterium den jeweiligen Atomen des DFT-Fragments zugeordnet werden können. Über dieser Ebene $l = -1$ folgt dann diejenige der strukturellen Einheiten ($l = 0$), welche in (P)MM-MD Simulationen die unterste Hierarchieebene darstellt. Die Ausdehnungen der Cluster von Gitterpunkten, welche N

solcher Punkte i an den Orten $\mathbf{r}_i$ umfassen, werden hier analog zu den atomaren Clustern über sogenannte Gyrationsradien

$$R = \sqrt{1/N \sum_{i}^{N} |\mathbf{r}_i^2|} \tag{3.1}$$

berechnet.

Aufgrund der großen Anzahl N der Gitterpunkte wurde es notwendig die Berechnung der Gyrationsradien $\mathbf{R}$ auf höheren Ebenen $l > -2$ nicht durch eine Summation über alle Gitterpunkte i zu berechnen, sondern für die Evaluation analog zur hierarchischen Berechnung von Multipolen in SAMM auf die geometrischen Eigenschaften der Objekte der nächst niedrigeren Hierarchieebene $l - 1$ zurückzugreifen. Hier konnte ich die konkrete Form der hierarchischen Berechnung von Gyrationsradien $\mathbf{R}$ von Clustern von Gitterpunkten auf einer Ebene l durch die geometrischen Eigenschaften der Kindcluster von Gitterpunkten auf der nächst niedrigeren Ebene $l - 1$, wie in Kapitel 2 des *Supplementary Material* zu der Veröffentlichung [45] gegeben, herleiten.

Außerdem konnte ich die im Kontext der Verbindung der schnellen Multipolmethode SAMM mit der Kontinuumsmethode HADES vorgestellten Hamiltonschen SAMM-Kräfte, welche in [43] beschrieben sind, schon im Rahmen dieser Publikation zur Verfügung stellen. Dies ermöglichte es hier, die Korrektheit der Implementierung anhand des sogenannten "Shadow-Hamiltonian" [114] zu verifizieren und das algorithmische Rauschen des Algorithmus zu verringern, wobei der zusätzlich nötige Rechenaufwand hier nicht sonderlich ins Gewicht fällt, da dieser vornehmlich durch die notwendigen DFT-Berechnungen dominiert wird.

3.2 Parametrisierung polarisierbarer Wassermodelle

In einer Anwendung der beschriebenen DFT/PMM Kopplung hat Philipp Tröster verschiedene Wassermodelle parametrisiert [54, 55]. Diese Untersuchungen hatten das Ziel, wie in der Einleitung erwähnt, möglichst gute, d.h. den elektrostatischen Eigenschaften einer DFT Beschreibung eines Wassermoleküls ähnelnde, PMM Modelle zu kreieren. Diese Modelle sollen damit insbesondere als zuverlässiger Ersatz für die sehr rechenaufwändigen DFT-Beschreibung des Umgebungswasser bei der Berechnung von IR-Spektren dienen.

Die Abbildung 3.1 stellt solche Wassermodelle schematisch dar. Sie sind durch einen Gaußschen induzierbaren Dipol (die Breite der Verteilung ist durch die innere Kugel dargestellt) am Sauerstoffatom des Wassermoleküls, durch positive Partialladungen (rot) an Wasserstoffatomen und durch ein bis drei negative Partialladungen (blau) an masselosen Orten in der Nähe des Sauerstoffatoms charakterisiert.

In diesem Zusammenhang habe ich elementare Beiträge bei der Implementierung des verwendeten polarisierbaren Kraftfeldes in das Simulationsprogramm erbracht. Wichtiger Bestandteil dieser Beiträge war die korrekte Implementierung der Druckberechnung über das sogenannte Virial für die Kraftbeiträge aus den polarisierbaren Freiheitsgraden und die korrekte Implementierung der Wechselwirkungen zwischen zwei atomaren Ladungen bzw. Dipolen,

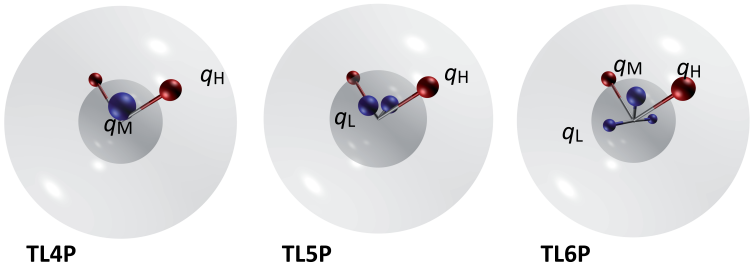


Abbildung 3.1: Schematische Darstellung der polarisierbaren Wassermodelle TL4P, TL5P und TL6P [54, 55]. Gezeigt ist die Geometrie der jeweiligen Modelle, in welchen den Wasserstoffatomen jeweils positive Partialladungen (q_H, rot) zugeordnet sind. Die masselosen Punkte in der Nähe des Sauerstoffatoms (Zentren der Kugeln) sind durch positive Ladungen (q_L und q_M, blau) charakterisiert. Die äußere Kugel beschriebt das effektive Volumen der Wassermodelle. Die innere Kugel beschreibt die Breite des Gaußschen induzierbaren Dipols. Abbildung adaptiert aus [116].

die als Punktobjekte oder über Gaußsche Dichten [vgl. (1.4)] der Breite σ beschrieben werden. Hierzu habe ich die Formeln verwendet, welche in der Doktorarbeit von B. Egwolf [115] gegeben sind.

Zusätzlich konnte ich eine Vorabversion der neuen SAMM Algorithmen bereitstellen, welche von der Einbettung der Lennard-Jones Dispersion in das FMM Schema profitierte. In dieser Version konnten so die Abstände l_0, ab denen die explizite Auswertung von atomaren Paarwechselwirkungen durch SAMM Wechselwirkungen von Wassermolekülen (strukturelle Einheiten) ersetzt werden, gegenüber der SAMM$_{2003}$-Version reduziert und so Effizienzgewinne erreicht werden. Dies geschah hier jedoch noch nicht über die intensive und gründliche Abschätzung der Restfehler der SAMM Entwicklungen, welche für verschiedene Typen von strukturellen Einheiten (unterste Hierarchiestufe) und von Clustern höherer Hierarchiestufen erst im Rahmen der Veröffentlichung [42] erfolgte und zu dem genauigkeitsbasierten Kriterium IAC führte. Da in den Simulationen in den Arbeiten [54, 55] jedoch ausschließlich homogene Wassersysteme verwendet wurden, fiel das Fehlen des IAC hier jedoch nicht ins Gewicht.

4 Résumé und Ausblick

In den vorangegangenen Kapiteln habe ich ein, in wesentlichen Teilen neues, FMM Verfahren für die linear skalierende Berechnung langreichweitiger Wechselwirkungen bei MD Simulationen bio-molekularer Systeme vorgestellt, welches die frühere "structure adapted multipole method" SAMM_{2003} [30, 32, 33, 51] unter großen Gewinnen an Genauigkeit und Effizienz ablöst. Große Vorteile bietet das neue SAMM Verfahren insbesondere für PMM Kraftfelder, DFT/PMM Hybridrechnungen [44, 45] und Kontinuumssimulationen mit HADES [48, 58].

Im Rahmen dieser Entwicklung von Rechenmethoden wurden zunächst die FMM Entwicklungen von der in SAMM_{2003} noch inkonsistent dargestellten Quadrupolordnung $p = 2$ auf die Hexadekapolordnung $p = 4$ erweitert [41]. Dazu wurden zweiseitige Taylorentwicklungen [52] p-ter Ordnung eingesetzt, so dass die Restfehler der Energie- und Kraftterme für ein Paar von wechselwirkenden Clustern in Abstand r proportional zu $r^{-(p+2)}$ skalieren. Das neue SAMM Verfahren erhält erfreulicherweise den Gesamtimpuls, da es das Newtonsche Reaktionsprinzip für alle wechselwirkenden Atompaare garantiert. Seine stark erhöhte Genauigkeit wurde anhand von Messungen absoluter und relativer Potential- und Kraftfehler sowie anhand der Untersuchung der algorithmischen Wärmeproduktion nachgewiesen.

Wegen des schnelleren Abfalls des Restfehlers durch die in Unterkapitel 2.1 dargestellten Neuerungen, der nun mit $\sim r^{-6}$ und nicht mehr lediglich mit $\sim r^{-4}$ skaliert, ergab sich die Frage, nach welchen Kriterien die Abstände r_{min}, ab denen die atomaren Wechselwirkungen über Multipolnäherungen behandelt werden, möglichst klein gewählt werden können.

Die Wahl von Abständen r_{min}, die im Vergleich zu dem festen Distanzklassenkriterium $l_0 = 10$ Å (vgl. Abbildung 1.7) klein sind, war vor allem deshalb von Interesse, weil die damit einhergehende Reduktion der Anzahl der zu berechnenden atomaren Paarwechselwirkungen auf Effizienzvorteile hoffen ließ. Insbesondere für komplexe polarisierbare Wassermodelle, bei denen der Zeitbedarf für die Auswertung der atomaren Interaktionen sehr viel größer ist als bei einfachen nicht-polarisierbaren Wassermodellen wie TIP3P [56], waren so relativ große Effizienzgewinne zu erwarten.

Zur systematischen Lösung dieser Frage wurde in der in Unterkapitel 2.2 nachgedruckten Veröffentlichung [42] ein neues, genauigkeits-korrigiertes Akzeptanzkriterium (interaction acceptance criterion, IAC) eingeführt, welches es ermöglicht, über einen zentralen Parameter Θ den absoluten Fehler, der bei der elektrostatischen Kraftberechnung für eine Cluster-Cluster Wechselwirkung maximal zulässig ist, einzustellen. Das Kriterium beruht auf empirisch bestimmten mittleren Fehlern für die in Protein-Wasser Systemen vorkommenden Atomcluster erster Stufe, den sogenannten "strukturellen Einheiten". Auf diese Weise berücksichtigt das IAC die typischen elektrostatischen Signaturen molekularer Bausteine des Systems und ermöglicht somit auch für heterogene Systeme die Wahl eines homogenen Kompromisses zwischen Genauigkeit und Effizienz.

Eine wichtige Voraussetzung für die effiziente Nutzung des neuen Wechselwirkungskriteriums war die Einbindung der $\sim 1/r^6$ Dispersion durch zweiseitige Taylor-Entwicklungen bis zur maximalen Ordnung $q = 3$ in das hierarchische Berechnungsszenario von SAMM. Es wurde gezeigt, dass damit Artefakte, die durch das Abschneiden der Dispersion bei kurzen Distanzen entstehen, eliminiert werden. (In Unterkapitel 2.3 wurde darüber hinaus gezeigt, wie auch die $\sim 1/r^{12}$ Lennard-Jones Repulsion in SAMM eingebettet werden kann, wobei hier Entwicklungen bis zur Ordnung $r = 1$ hinreichende Genauigkeit bieten.)

Der Einfluss des IAC-Parameters Θ wurde durch MM-MD Simulationen von Wasser- und Methanolsystemen anhand der resultierenden Energiedrift ermittelt. Aufbauend auf diese Analysen wurden drei Standard-IAC-Parameter Θ_χ mit $\chi =$ a,m,f [für: (a)ccurate, inter(m)ediate und (f)ast] in Kombination mit SAMM-Entwicklungen bis zu den Ordnungen $p = 4$ für die Elektrostatik und $q = 3$ für die Dispersion entwickelt. Damit liegt der minimale Abstand, ab dem die Wechselwirkungen über SAMM Näherungen ausgewertet werden können, für Wassermoleküle zwischen $r_\mathrm{f} = 5.42$ Å und $r_\mathrm{a} = 7.97$ Å.

Insbesondere stellte sich heraus, dass Simulationen von Wassersystemen, welche mit dem komplexen und polarisierbaren Sechspunktmodell TL6P [55] durchgeführt wurden, nur 4,6-fach langsamer waren als Vergleichssimulationen mit dem nicht-polarisierbaren Dreipunktmodell TIP3P. Im Vergleich dazu sind Simulationen, welche die Paarwechselwirkungen exakt auswerten, bei TL6P Wassersystemen um einen Faktor 13 langsamer als bei TIP3P-Systemen. Somit sind die weiterentwickelten SAMM Algorithmen effizient zur Simulation komplexer polarisierbarer Modelle einsetzbar.

Schließlich wurde in Unterkapitel 2.3 die Kombination von HADES-MD [48, 58] mit einer neuen, energieerhaltenden Version von SAMM dargestellt, die ein linear skalierendes HADES/SAMM-MD Verfahren ergab. Damit wurden auch für große Proteine MD Simulationen im dielektrischen Medium ermöglicht. Durch die energieerhaltenden SAMM Kräfte werden der Gesamtimpuls und der Gesamtdrehimpuls des simulierten Proteins numerisch genau und die Energie sehr gut erhalten.

4.1 Optimierung der Parallelisierung von SAMM

Da heutzutage die Erhöhung der Leistungsfähigkeit von Computern nicht mehr hauptsächlich, wie noch vor kurzem, durch die Erhöhung der Taktfrequenz der einzelnen Prozessoren erzielt werden kann, sondern durch eine Vergrößerung der Anzahl der parallel zu verwendenden Prozessorkerne realisiert werden muss, nimmt die Aufgabe der effizienten Parallelisierung von Programmen in der rechnergestützten Physik eine zentrale Rolle ein. Das MD Programm IPHIGENIE [53], in dem das von mir entwickelte neue SAMM Verfahren als Software implementiert ist, nutzt zur Parallelisierung zwei verschiedene, sich ergänzende Strategien.

Zu einen wird die Programmierschnittstelle OpenMP verwendet, die eine effiziente Shared-Memory-Programmierung ermöglicht. Dies ist insbesondere für lokale Rechnerarchitekturen von Interesse, in denen sich mehrere Prozessorkerne den selben physikalischen Arbeitsspeicher teilen. Zum anderen nutzt IPHIGENIE das Message Passing Interface (MPI), welches es

erlaubt, auszuführende Programme in mehrere Prozesse aufzuteilen. MPI bietet dabei mannigfaltige Möglichkeiten die jeweils nötige Kommunikation zwischen den einzelnen Prozessen zu gestalten. Dieser Ansatz ist insbesondere dazu geeignet, große Parallelrechner, die nicht über einen geteilten Speicher verfügen, da sie aus einzelnen, über ein Netzwerk verbundenen Rechnern bestehen, zur Beschleunigung von Rechnungen nutzbar zu machen. Ein aktuelles Beispiel für einen solchen Großrechner ist der SuperMUC am Leibniz-Rechenzentrum (LRZ) in Garching.

Im Rahmen der KONWIHR-Projekte des LRZ wurde Magnus Schwörer und mir unter Leitung von Gerald Mathias dankenswerterweise die Möglichkeit gegeben, die Parallelisierung von IPHIGENIE für die Nutzung auf dem SuperMUC zu optimieren. Während Magnus Schwörer hier in erster Linie die OpenMP-seitige Parallelisierung überarbeitet hat, bestand meine Aufgabe in der Revision der Einbindung der MPI Schnittstelle in IPHIGENIE.

Die vorherige MPI Parallelisierungsstrategie, welche in Unterkapitel 2.2 dargestellt ist, zielt auf kleinere Computercluster. Hier wird die Rechenlast anhand der obersten Clusterebene λ des Systems auf die einzelnen MPI Prozesse verteilt. Dabei wird jeweils mindestens ein Cluster C der Ebene λ einem MPI Prozess zugeordnet, welcher dann für die Berechnung der Kräfte auf die Atome $i \in C$ zuständig ist.

Dazu müssen für die Nutzung von N_c Prozessorkernen, welche jeweils einen MPI Prozess ausführen, mindestens N_λ Cluster der höchsten Hierarchieebene λ in dem gegebenen System vorhanden sein. Diese Anzahl N_λ liegt in dem alten Parallelisierungsansatz, z.B. für Wassersysteme, üblicherweise zwischen 60 und 250, was die effektive Ausnutzung einer Rechnerarchitektur mit 10^5 bis 10^6 Rechenkernen (wie sie vom SuperMUC geboten wird) auch bei hybrider Nutzung der OpenMP und MPI Parallelisierungen unmöglich macht.

Im Rahmen des KONWIHR Projekts habe ich die MPI Implementierung derart neu gestaltet, dass nunmehr nicht die oberste Hierarchieebene $l = \lambda$ sondern die unterste Cluster-Ebene $l = 0$ der strukturellen Einheiten zur Verteilung des Systems auf die Rechenkerne genutzt wird. Hierzu mussten insbesondere die Erstellung der Wechselwirkungslisten überarbeitet und neue Kommunikationsfunktionen eingeführt werden, welche es ermöglichen, die von der Berechnung der Multipolmomente aufgeworfene Rechenlast nunmehr auf alle Prozesse zu verteilen. Die Optimierung dieser neuen Parallelisierungsstrategie ist zur Zeit noch Gegenstand weiterer laufender Arbeiten und es ist geplant, die endgültige Version im Rahmen einer Veröffentlichung, welche die unter der Leitung von Gerald Mathias neu entwickelten Features von IPHIGENIE zusammenfassen soll, zu publizieren. Dennoch möchte ich schon hier den gegenwärtigen Entwicklungsstand anhand von Beispiel-Simulationen andeuten.

Abbildung 4.1 zeigt das Skalierungsverhalten des gegenwärtigen Programms anhand der Beschleunigung der SAMM/RF Simulationen eines kubischen TIP3P [56] Systems mit $N_A = 43890$ Atomen relativ zur Simulationsgeschwindigkeit der sequentiellen Version. Dieser Speedup ist als Funktion der Anzahl N_MPI der MPI Prozesse gegeben.

Erfreulicherweise ist über den gesamten gezeigten Bereich (selbst bei $N_\mathrm{MPI} = 512$ und damit für nur 86 Atome pro MPI Prozess) noch ein zusätzlicher Speedup durch die Nutzung von weiteren MPI Prozessen gegeben. Im Ergebnis ist nun der Nutzung sehr großer Anzahlen von MPI Prozessen N_MPI keine prinzipielle Grenze gesetzt, so dass in der neuen Version die Güte

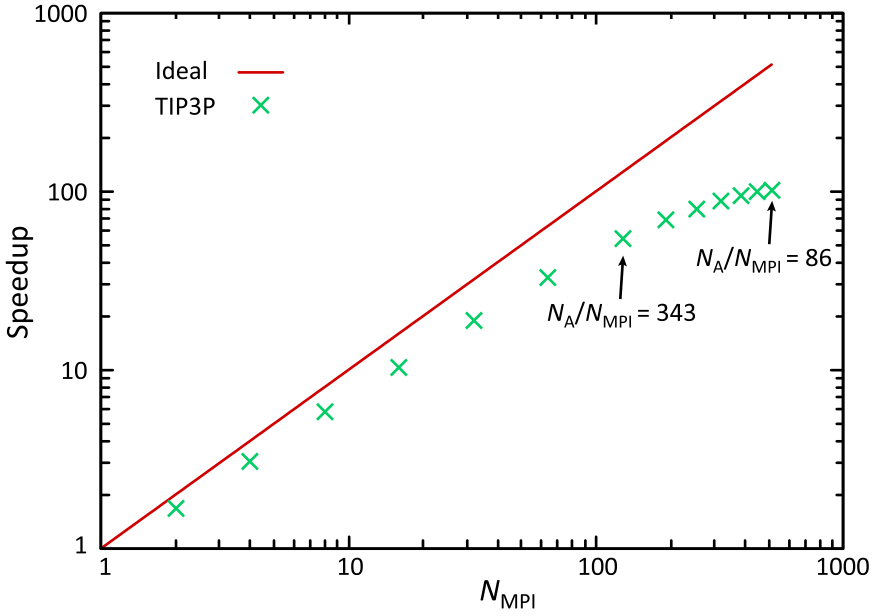


Abbildung 4.1: Die Abbildung zeigt das Skalierungsverhalten der neuen Parallelisierung anhand des Speedups, d.h. der Beschleunigung der Simulationsgeschwindigkeit der parallelen Version relativ zur sequentiellen Version, als Funktion der Anzahl N_{MPI} der genutzten MPI Prozesse. Die Simulationen wurden an einem TIP3P [56] Wassersystem mit $N_A = 43890$ Atomen durchgeführt. Für $N_{\text{MPI}} = 128$ und $N_{\text{MPI}} = 512$ sind zusätzlich die Anzahlen der Atome pro MPI Prozesse angegeben.

der Parallelisierung der einzelnen Programmteile die sinnvollerweise zu wählende Anzahl N_{MPI} bestimmt.

4.2 Modifikation von SAMM für periodische Randbedingungen

Ein Nachteil der SAMM/RF Methode [51] für MD-Simulationen von Protein-Lösungsmittel-Systemen ist es, dass die ständige Drift von Wassermolekülen aus dem atomar dargestellten Bereich innerhalb der MIC-Kugeln, welche jeden atomaren Cluster der höchsten SAMM Ebene umgeben (vgl. Abb. 1.5), in das umgebende Kontinuum (und *vice versa*) zu algorithmischem Rauschen führt. Die resultierende Dynamik, weist daher eine kleine, mit der Systemgröße abnehmende, aber nicht zu vernachlässigende systematische Energiezunahme auf. Es stellt sich nun, angesichts der großen Genauigkeit des neuen SAMM Verfahrens, die Frage, ob und wie diese verbleibende Rauschquelle beseitigt werden kann.

Die Kombination von SAMM mit Gittersummenmethoden ist dazu ein vielversprechender Ansatz, da die Zusammensetzung der zugehörigen, periodisch fortgesetzten Systeme nicht fluktuiert. Die Einbettung von dreidimensionalen FMM Algorithmen in Gittersummenmethoden ist in der Literatur bekannt. So haben schon Schmidt und Lee 1991 eine solche Implementierung für sphärische Multipolmomente beschrieben [117]. Ausführlichere Herleitungen

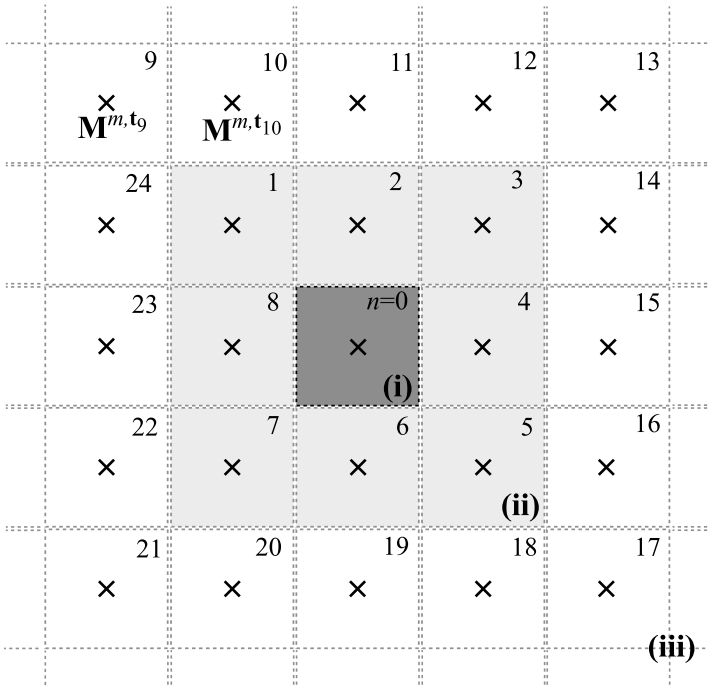


Abbildung 4.2: Kombination von FMM Verfahren mit periodischen Randbedingungen als Alternative zu den toroidalen Randbedingungen, die im Reaktionsfeldansatz SAMM/RF [51] eingesetzt werden. Für die Atome im zentralen System (i) werden alle Kräfte ausgewertet, welche durch die Atome in diesem System (i), in den direkt benachbarten Bildern (ii) und in den weit entfernten Bildern (iii) hervorgerufen werden. Hierzu wird die Wirkung der Regionen (i) und (ii) durch Multipolentwicklungen und die der Region (iii) über kombinierte Ewald-Multipol-Methoden ausgewertet.

finden sich in [101, 103] und eine aktuelle Implementierung dieser Methode in [118]. Auch für Multipolentwicklungen in kartesischen Koordinaten finden sich in der Literatur Vorbilder [106, 119] für die Verwendung von periodischen Randbedingungen in Verbindung mit Multipolmethoden, welche für den SAMM Algorithmus adaptiert werden können.

Abbildung 4.2 zeigt das typische Berechnungsschema der FMM-Gittersummenmethoden, in welchem drei räumliche Bereiche voneinander zu unterscheiden sind. Diese Bereiche sind (i) das zentrale System, (ii) eine Schicht aus periodisch angelagerten Abbildern (hier konkret am Beispiel unmittelbarer Nachbarn, die mindestens an einem Punkt direkt an das zentrale System angrenzen) und (iii) alle weiteren und somit weiter entfernten periodischen Bilder. Das zentrale System (i) ist hier insofern ausgezeichnet, da nur für dieses System die auf die Atome wirkenden Kräfte ausgerechnet werden. Dazu werden in [101, 103, 106, 117] über Ewald-Methoden Ausdrücke entwickelt, welche den Einfluss der nicht-benachbarten periodischen Bilder (iii) auf die Atome des zentralen Systems (i) über schnelle Multipolmethoden berechenbar machen. Hierfür werden jeweils die Ladungsverteilungen der einzelnen periodischen Systeme über Multipolmomente beschrieben. Die Wirkung dieser Momente $\mathbf{M}^{m,\mathbf{t}_n}$ m-ter Ordnung der Bilder aus dem Bereich (iii), welche um den Vektor $\mathbf{t}_n$ gegenüber dem zentralen System verschoben sind, wird dann durch die Berechnung von lokalen Taylorent-

wicklungskoeffizienten (also der Multipolpotentiale und ihrer Ableitungen) im Zentrum $\mathbf{t}_0$ des zentralen Systems beschrieben. Durch Taylorentwicklungen können dann im Rahmen der FMM-Methodik Kräfte und Felder an den Atomorten ausgewertet werden.

Diese Berechnung nutzt den Vorteil, dass die Multipolmomente $\mathbf{M}^{m,\mathbf{t}_n}$ aller n periodischen Bilder in (iii) bzgl. ihres jeweiligen Referenzpunktes $\mathbf{t}_n$ identisch sind und deshalb, z.B. in der Berechnung des periodischen Potentials

$$\Phi^{\mathrm{GS}}(\mathbf{t}_0) = \sum_{m=0}^{p} \mathbf{M}^{m,\mathbf{t}_0} \odot \sum_{n\in(iii)} \frac{1}{m!} \frac{(\mathbf{t}_n - \mathbf{t}_0)^{(m)}}{|(\mathbf{t}_n - \mathbf{t}_0)|^{2m+1}}, \tag{4.1}$$

welches hier bis zur maximalen Multipolordnung p ausgeführt wird, ausgeklammert werden können. Für eine unveränderte Geometrie, d.h. für Simulationen in denen die Verschiebevektoren $\mathbf{t}_n$ konstant bleiben, müssen die Kontraktionsmatrizen

$$K^m = \sum_{n\in(iii)} (1/m!) \left((\mathbf{t}_n - \mathbf{t}_0)^{(m)}/|(\mathbf{t}_n - \mathbf{t}_0)|^{2m+1}\right)$$

somit nur einmal zu Beginn einer Simulation über Ewald-Methoden berechnet werden, da einzig die Werte der Multipolmomente $\mathbf{M}^{m,\mathbf{t}_0}$ dynamischen Änderungen unterliegen.

Die Wirkung der Ladungsverteilung in den direkt benachbarten Bildern (ii) und im zentralen System (i) auf die Ladungen in ebendiesem System (i) wird im Gegensatz dazu direkt über explizite FMM Ausdrücke berechnet. Diese explizite Berechnung entspricht weitestgehend dem bisherigen Ansatz des SAMM/RF Algorithmus mit dem einzigen Unterschied, dass in SAMM/RF die jeweiligen Wechselwirkungen der atomaren Cluster des zentralen Systems mit den Wechselwirkungspartnern aus den Bereichen (i) und (ii) nur für Distanzen $r \leq d_{\mathrm{MIC}}$ ausgewertet wurden (vgl. Abb. 1.5). Es ist somit für die oben beschriebene Implementierung periodischer Randbedingungen mit einem gewissen Mehraufwand an Rechenzeit gegenüber der Nutzung von toroidalen Randbedingungen in SAMM/RF Algorithmus zu rechnen. Da die zusätzlich auszuwertenden Wechselwirkungen jedoch durch relativ große Abstände $r > d_{\mathrm{MIC}}$ gekennzeichnet sind, sollten diese Interaktionen, gesteuert durch das IAC, über Clusterpaare mit sehr großen Radien berechnet werden und damit der Mehraufwand überschaubar bleiben.

Abbildungsverzeichnis

Literaturverzeichnis

[1] Großjean, M. F., P. Tavan und K. Schulten. Quantumchemical Vibrational Analysis of the Retinal Chromophore of Bacteriorhodopsin. *J. Phys. Chem.* **94**, 8059–8069 (1990).

[2] Babitzki, G., G. Mathias und P. Tavan. The Infrared Spectra of the Retinal Chromophore in Bacteriorhodopsin Calculated by a DFT/MM Approach. *J. Phys. Chem. B* **113**, 10496–10508 (2009).

[3] Nonella, M. und P. Tavan. An Unscaled Quantum Mechanical Harmonic Force Field for p-Benzoquinone. *Chem. Phys.* **199**, 19 – 32 (1995).

[4] Nonella, M., G. Mathias und P. Tavan. Infrared Spectrum of p-Benzoquinone in Water Obtained from a QM/MM Hybrid Molecular Dynamics Simulation. *J. Phys. Chem. A* **107**, 8638–8647 (2003).

[5] Rieff, B., G. Mathias, S. Bauer und P. Tavan. Density Functional Theory Combined with Molecular Mechanics: The Infrared Spectra of Flavin in Solution. *Photochem. Photobiol.* **87**, 511–523 (2011).

[6] Rieff, B., S. Bauer, G. Mathias und P. Tavan. IR Spectra of Flavins in Solution: DFT/MM Description of Redox Effects. *J. Phys. Chem. B* **115**, 2117–2123 (2011).

[7] Rieff, B., S. Bauer, G. Mathias und P. Tavan. DFT/MM Description of Flavin IR Spectra in BLUF Domains. *J. Phys. Chem. B* **115**, 11239–11253 (2011).

[8] Oesterhelt, D. und W. Stoeckenius. Rhodopsin-Like Protein from the Purple Membrane of Halobacterium halobium. *Nature New Biol.* **233**, 149–152 (1971).

[9] Lanyi, J. K. Bacteriorhodopsin. *Annu. Rev. Physiol.* **66**, 665–688 (2004).

[10] Deisenhofer, J., O. Epp, K. Miki, R. Huber und H. Michel. Structure of the Protein Subunits in the Photosynthetic Reaction Centre of Rhodopseudomonas viridis at 3ÅResolution. *Nature* **318**, 618–624 (1985).

[11] Palczewski, K. G Protein-Coupled Receptor Rhodopsin. *Annu. Rev. Biochem.* **75**, 743–767 (2006).

[12] Gomelsky, M. und G. Klug. BLUF: a Novel FAD-Binding Domain Involved in Sensory Transduction in Microorganisms. *Trends Biochem. Sci.* **27**, 497 – 500 (2002).

[13] Koller, F. O., W. J. Schreier, T. E. Schrader, A. Sieg, S. Malkmus, C. Schulz, S. Dietrich, K. Rück-Braun, W. Zinth und M. Braun. Ultrafast Structural Dynamics of Photochromic Indolylfulgimides Studied by Vibrational Spectroscopy and DFT Calculations. *J. Phys. Chem. A* **110**, 12769–12776 (2006).

[14] Zinth, W., T. Schrader, W. Schreier, F. Koller, T. Cordes, G. Babitzki, R. Denschlag, P. Tavan, M. Löweneck, S.-L. Dong, L. Moroder und C. Renner. Ultrafast Unzipping of a Beta-Hairpin Peptide. In P. Corkum, D. Jonas, R. Miller und A. Weiner (Herausgeber), *Ultrafast Phenomena XV*, Band 88 von *Springer Series in Chemical Physics*, Seiten 498–500. Springer Berlin Heidelberg (2007).

[15] Schrader, T. E., W. J. Schreier, T. Cordes, F. O. Koller, G. Babitzki, R. Denschlag, C. Renner, M. Löweneck, S.-L. Dong, L. Moroder, P. Tavan und W. Zinth. Light-Triggered β-Hairpin Folding and Unfolding. *Proc. Natl. Acad. Sci.* **104**, 15729–15734 (2007).

[16] Denschlag, R., W. J. Schreier, B. Rieff, T. E. Schrader, F. O. Koller, L. Moroder, W. Zinth und P. Tavan. Relaxation Time Prediction for a Light Switchable Peptide by Molecular Dynamics. *Phys. Chem. Chem. Phys.* **12**, 6204–6218 (2010).

[17] Schrader, T., T. Cordes, W. Schreier, F. Koller, S.-L. Dong, L. Moroder und W. Zinth. Folding and Unfolding of Light-Triggered beta-Hairpin Model Peptides. *J. Phys. Chem. B* **115**, 5219–5226 (2011).

[18] Lifson, S. und A. Warshel. Consistent Force Field for Calculations of Conformations, Vibrational Spectra, and Enthalpies of Cycloalkane and n-Alkane Molecules. *J. Chem. Phys.* **49**, 5116–5129 (1968).

[19] Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez und J. A. Pople. Gaussian 03, Revision C.02. Gaussian, Inc., Wallingford, CT, 2004.

[20] Hohenberg, P. und W. Kohn. Inhomogeneous Electron Gas. *Phys. Rev.* **136**, B864–B871 (1964).

[21] Kohn, W. und L. J. Sham. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133–A1138 (1965).

[22] Parr, R. G. Density Functional Theory. *Ann. Rev. Phys. Chem.* **34**, 631–656 (1983).

[23] Neugebauer, J. und B. A. Hess. Fundamental Vibrational Frequencies of Small Polyatomic Molecules from Density-Functional Calculations and Vibrational Perturbation Theory. *J. Chem. Phys.* **118**, 7215–7225 (2003).

[24] MacKerell, A. D., D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin und M. Karplus. All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).

[25] Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell und P. A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **117**, 5179–5197 (1995).

[26] Oostenbrink, C., A. Villa, A. Mark und W. Van Gunsteren. A Biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: The GROMOS Force Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem. B* **25**, 1656–1676 (2004).

[27] Warshel, A. und M. Levitt. Theoretical Studies of Enzymic Reactions: Dielectric, Electrostatic and Steric Stabilization of the Carbonium Ion in the Reaction of Lysozyme. *J. Mol. Biol* **103**, 227 – 249 (1976).

[28] Eichinger, M., P. Tavan, J. Hutter und M. Parrinello. A Hybrid Method for Solutes in Complex Solvents: Density Functional Theory Combined with Empirical Force Fields. *J. Chem. Phys.* **110**, 10452–10467 (1999).

[29] Hutter, J., A. Alavi, T. Deutsch, M. Bernasconi, S. Goedecker, D. Marx, M. Tuckerman und M. Parrinello. *CPMD: Car–Parinello Molecular Dynamics, version 3.10.* © IBM Corp 1990–2008 and MPI für Festkörperforschung Stuttgart, `www.cpmd.org` (1997-2001).

[30] Eichinger, M., H. Grubmüller, H. Heller und P. Tavan. FAMUSAMM: An Algorithm for Rapid Evaluation of Electrostatic Interactions in Molecular Dynamics Simulations. *J. Comput. Chem.* **18**, 1729–1749 (1997).

[31] Greengard, L. und V. Rokhlin. A Fast Algorithm for Particle Simulations. *J. Comput. Phys.* **73**, 325 – 348 (1987).

[32] Niedermeier, C. und P. Tavan. A Structure Adapted Multipole Method for Electrostatic Interactions in Protein Dynamics. *J. Chem. Phys.* **101**, 734–748 (1994).

[33] Niedermeier, C. und P. Tavan. Fast Version of the Structure Adapted Multipole Method - Efficient Calculation of Electrostatic Forces in Protein Dynamics. *Mol. Simul.* **17**, 57–66 (1996).

[34] Eichinger, M. *Berechnung molekularer Eigenschaften in komplexer Lösungsmittelumgebung: Dichtefunktionaltheorie kombiniert mit einem Molekularmechanik-Kraftfeld*. Doktorarbeit, Ludwig-Maximilians Universität München, Fakultät für Physik, Germany (1999).

[35] Nonella, M., G. Mathias, M. Eichinger und P. Tavan. Structures and Vibrational Frequencies of the Quinones in Rb. sphaeroides Derived by a Combined Density Functional/Molecular Mechanics Approach. *J. Phys. Chem. B* **107**, 316–322 (2003).

[36] Klähn, M., G. Mathias, C. Kötting, M. Nonella, J. Schlitter, K. Gerwert und P. Tavan. IR Spectra of Phosphate Ions in Aqueous Solution: Predictions of a DFT/MM Approach Compared with Observations. *The Journal of Physical Chemistry A* **108**, 6186–6194 (2004).

[37] VandeVondele, J., P. Tröster, P. Tavan und G. Mathias. Vibrational Spectra of Phosphate Ions in Aqueous Solution Probed by First-Principles Molecular Dynamics. *J. Phys. Chem. A* **116**, 2466–2474 (2012).

[38] Schultheis, V., R. Reichold, B. Schropp und P. Tavan. A Polarizable Force Field for Computing the Infrared Spectra of the Polypeptide Backbone. *J. Phys. Chem. B* **112**, 12217–12230 (2008).

[39] Schmitz, M. und P. Tavan. On the Art of Computing the IR Spectra of Molecules in Condensed Phase. In S. Tanaka und J. Lewis (Herausgeber), *Modern Methods for Theoretical Physical Chemistry of Biopolymers*, Kapitel 8, Seiten 157–177. Elsevier, Amsterdam (2006).

[40] Breitenfeld, B. *Kopplung des Dichtefunktionaltheorie-Programms CPMD an ein polarisierbares Kraftfeld: Implementierung und Tests*. Diplomarbeit, Ludwig-Maximilians Universität München (2010).

[41] Lorenzen, K., M. Schwörer, P. Tröster, S. Mates und P. Tavan. Optimizing the Accuracy and Efficiency of Fast Hierarchical Multipole Expansions for MD Simulations. *J. Chem. Theory Comput.* **8**, 3628–3636 (2012).

[42] Lorenzen, K., C. Wichmann und P. Tavan. Including the Dispersion Attraction into Structure-Adapted Fast Multipole Expansions for MD Simulations. *J. Chem. Theory Comput.* **10**, 3244–3259 (2014).

[43] Lorenzen, K., G. Mathias und P. Tavan. Linearly Scaling and Almost Hamiltonian Dielectric Continuum Molecular Dynamics Simulations through Fast Multipole Expansions. *J. Chem. Phys.* **143**, 184114 (2015).

[44] Schwörer, M., B. Breitenfeld, P. Tröster, K. Lorenzen, P. Tavan und G. Mathias. Coupling DFT to Polarizable Force Fields for Efficient and Accurate Hamiltonian Molecular Dynamics Simulations. *J. Chem. Phys.* **138**, 244103 (2013).

[45] Schwörer, M., K. Lorenzen, G. Mathias und P. Tavan. Utilizing Fast Multipole Expansions for Efficient and Accurate Quantum-Classical Molecular Dynamics Simulations. *J. Chem. Phys.* Seite submitted (2015).

[46] Babitzki, G. *Analysen der Schwingungsspektren von Biomolekülen mit Hybridmethoden*. Doktorarbeit, Ludwig-Maximilians Universität München, Fakultät für Physik, Germany (2009).

[47] Babitzki, G., R. Denschlag und P. Tavan. Polarization Effects Stabilize Bacteriorhodopsin's Chromophore Binding Pocket: A Molecular Dynamics Study. *J. Phys. Chem. B* **113**, 10483–10495 (2009).

[48] Bauer, S., P. Tavan und G. Mathias. Electrostatics of Proteins in Dielectric Solvent Continua. II. Hamiltonian Reaction Field Dynamics. *J. Chem. Phys.* **140**, 104103 (2014).

[49] Pulay, P. Convergence acceleration of iterative sequences - The case of SCF iteration. *Chem. Phys. Lett.* **73**, 393–398 (1980).

[50] Császár, C. und P. Pulay. Geometry optimization by direct inversion in the iterative subspace. *J. Mol. Struct.* **114**, 31 (1984).

[51] Mathias, G., B. Egwolf, M. Nonella und P. Tavan. A Fast Multipole Method Combined with a Reaction Field for Long-Range Electrostatics in Molecular Dynamics Simulations: The Effects of Truncation on the Properties of Water. *J. Chem. Phys.* **118**, 10847–10860 (2003).

[52] Dehnen, W. A Hierarchical $O(N)$ Force Calculation Algorithm. *J. Comput. Phys.* **179**, 27–42 (2002).

[53] IPHIGENIE is available for download free of charge under the GPL licence at `http://sourceforge.net/projects/iphigenie`.

[54] Tröster, P., K. Lorenzen, M. Schwörer und P. Tavan. Polarizable Water Models from Mixed Computational and Empirical Optimization. *J. Phys. Chem. B* **117**, 9486–9500 (2013).

[55] Tröster, P., K. Lorenzen und P. Tavan. Polarizable Six-Point Water Models from Computational and Empirical Optimization. *J. Phys. Chem. B* **118**, 1589–1602 (2014).

[56] Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey und M. L. Klein. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **79**, 926–935 (1983).

[57] Schropp, B., C. Wichmann und P. Tavan. Spectroscopic Polarizable Force Field for Amide Groups in Polypeptides. *J. Phys. Chem. B* **114**, 6740–6750 (2010).

[58] Bauer, S., G. Mathias und P. Tavan. Electrostatics of Proteins in Dielectric Solvent Continua. I. An Accurate and Efficient Reaction Field Description. *J. Chem. Phys.* **140**, 104102 (2014).

[59] Lehninger, A. L. *Biochemistry*. Worth publishers, New York (1975).

[60] Creighton, T. E. *Proteins: Structures and Molecular Properties*. W. H. Freeman, New York (1993).

[61] ConsortiumInternational, H. G. S. Finishing the Euchromatic Sequence of the Human Genome. *Nature* **431**, 931 – 945 (2004).

[62] Tavan, P., H. Carstens und G. Mathias. Molecular Dynamics Simulations of Proteins and Peptides: Problems, Achievements, and Perspectives. In J. Buchner und T. Kiefhaber (Herausgeber), *Protein Folding Handbook*, Band 1, Seiten 1170–1195. Wiley-VCH, Weinheim (2005).

[63] Bauer, S., P. Tavan und G. Mathias. Exploring Hamiltonian Dielectric Solvent Molecular Dynamics. *Chem. Phys. Lett.* **612**, 20–24 (2014).

[64] Karplus, M. und J. A. McCammon. Molecular Dynamics Simulations of Biomolecules. *Nature* **9**, 646 – 652 (2002).

[65] van Gunsteren, W. F., D. Bakowies, R. Baron, I. Chandrasekhar, M. Christen, X. Daura, P. Gee, D. P. Geerke, A. Glättli, P. H. Hünenberger, M. A. Kastenholz, C. Oostenbrink, M. Schenk, D. Trzesniak, N. F. A. van der Vegt und H. B. Yu. Biomolecular Modeling: Goals, Problems, Perspectives. *Angew. Chem. Int. Ed.* **45**, 4064–4092 (2006).

[66] Marx, D. und J. Hutter. *Ab initio Molecular Dynamics: Basic Theory and Advanced Methods*. Cambridge University Press, Cambridge (2009).

[67] Ponder, J. und D. Case. Force Fields for Protein Simulations. *Adv. Prot. Chem.* **66**, 27–85 (2003).

[68] Kaminski, G. A., H. A. Stern, B. J. Berne, R. A. Friesner, Y. X. Cao, R. B. Murphy, R. Zhou und T. A. Halgren. Development of a Polarizable Force Field for Proteins via Ab Initio Quantum Chemistry: First Generation Model and Gas Phase Tests. *J. Comput. Chem.* **23**, 1515–1531 (2002).

[69] Harder, E., B. Kim, R. A. Friesner und B. J. Berne. Efficient Simulation Method for Polarizable Protein Force Fields: Application to the Simulation of BPTI in Liquid Water. *J. Chem. Theory Comput.* **1**, 169–180 (2005).

[70] Wang, Z.-X., W. Zhang, C. Wu, H. Lei, P. Cieplak und Y. Duan. Strike a Balance: Optimization of Backbone Torsion Parameters of AMBER Polarizable Force Field for Simulations of Proteins and Peptides. *J. Comput. Chem.* **27**, 781–790 (2006).

[71] Swope, W. C., H. C. Andersen, P. H. Berens und K. R. Wilson. A Computer Simulation Method for the Calculation of Equilibrium Constants for the Formation of Physical Clusters of Molecules: Application to Small Water Clusters. *J. Chem. Phys.* **76**, 637–649 (1982).

[72] Alder, B. J. und T. E. Wainwright. Phase Transition for a Hard Sphere System. *J. Chem. Phys.* **27**, 1208–1209 (1957).

[73] Rahman, A. Correlations in the Motion of Atoms in Liquid Argon. *Phys. Rev.* **136**, A405–A411 (1964).

[74] McCammon, J. A., B. R. Gelin und M. Karplus. Dynamics of Folded Proteins. *Nature* **267**, 585–590 (1977).

[75] Palmö, K., B. Mannfors, N. G. Mirkin und S. Krimm. Potential Energy Functions: From Consistent Force Fields to Spectroscopically Determined Polarizable Force Fields. *Biopolymers* **68**, 383–394 (2003).

[76] Mahoney, M. W. und W. L. Jorgensen. A Five-Site Model for Liquid Water and the Reproduction of the Density Anomaly by Rigid, Nonpolarizable Potential functions. *J. Chem. Phys.* **112**, 8910–8922 (2000).

[77] Lopes, P. E. M., B. Roux und A. D. MacKerell. Molecular Modeling and Dynamics Studies with Explicit Inclusion of Electronic Polarizability. Theory and Applications. *Theor. Chem. Acc.* **124**, 11–28 (2009).

[78] Schropp, B. und P. Tavan. The Polarizability of Point-Polarizable Water Models: Density Functional Theory/Molecular Mechanics Results. *J. Phys. Chem. B* **112**, 6233–6240 (2008).

[79] Bashford, D. und A. Case. Generalized Born Models for Macromolecular Solvation Effects. *Ann. Rev. Phys. Chem.* **51**, 129–152 (2000).

[80] Grycuk, T. Deficiency of the Coulomb-Field Approximation in the Generalized Born Model: An Improved Formula for Born Radii Evaluation. *J. Chem. Phys.* **119**, 4817–4826 (2003).

[81] Fenley, A., J. C. Gordon und A. Onufriev. An Analytical Approach to Computing Biomolecular Electrostatic Potential. I. Derivation and Analysis. *J. Chem. Phys.* **129**, 075101 (2008).

[82] Im, W., D. Beglov und B. Roux. Continuum Solvation Model: Computation of Electrostatic Forces from Numerical Solutions to the Poisson-Boltzmann Equation. *Comput. Phys. Commun.* **111**, 59–75 (1997).

[83] Geng, W. und G. W. Wei. Multiscale Molecular Dynamics Using the Matched Interface and Boundary Method. *J. Comput. Phys.* **230**, 435–457 (2011).

[84] Egwolf, B. und P. Tavan. Continuum Description of Solvent Dielectrics in Molecular-Dynamics Simulations of Proteins. *J. Chem. Phys.* **118**, 2039–2056 (2003).

[85] Stork, M. und P. Tavan. Electrostatics of Proteins in Dielectric Solvent Continua. I. Newton's Third Law Marries $q\mathbf{E}$ forces. *J. Chem. Phys.* **126**, 165105 (2007).

[86] Zhou, R. und B. J. Berne. Can a Continuum Solvent Model Reproduce the Free Energy Landscape of a β-Hairpin Folding in Water? *Proc. Natl. Acad. Sci. USA* **99**, 12777–12782 (2002).

[87] Nymeyer, H. und A. E. García. Simulation of the Folding Equilibrium of α-Helical Peptides: A Comparison of the Generalized Born Approximation with Explicit Solvent. *Proc. Natl. Acad. Sci. USA* **100**, 13934–13939 (2003).

[88] Feig, M., A. Onufriev, M. S. Lee, W. Im, D. A. Case und C. L. B. III. Performance Comparison of Generalized Born and Poisson Methods in the Calculation of Electrostatic Solvation Energies for Protein Structures. *J. Comput. Chem.* **25**, 265–284 (2003).

[89] Levy, N., D. Borgis und M. Marchi. A Dielectric Continuum Model of Solvation for Complex Solutes. *Comput. Phys. Commun.* **169**, 69–74 (2005).

[90] Sklenar, H., F. Eisenhaber, M. Poncin und R. Lavery. Including Solvent and Counterion Effects in the Force Fields of Macromolecular Mechanics: The Field Integrated Electrostatic Approach (FIESTA). In D. L. Beveridge und R. Lavery (Herausgeber), *Theoretical Biochemistry & Molecular Biophysics, 2. Proteins*, Seiten 317–335. Adenine Press, New York (1991).

[91] Tironi, I. G., R. Sperb, P. E. Smith und W. F. van Gunsteren. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **102**, 5451–5459 (1995).

[92] Hünenberger, P. H. und W. F. van Gunsteren. Alternative Schemes for the Inclusion of a Reaction-Field Correction into Molecular Dynamics Simulations: Influence on the Simulated Energetic, Structural, and Dielectric Properties of Liquid Water. *J. Chem. Phys.* **108**, 6117–6134 (1998).

[93] Ewald, P. P. Die Berechnung Optischer und Elektrostatischer Gitterpotentiale. *Ann. Phys.* **369**, 253–287 (1921).

[94] Allen, M. P. und D. Tildesley. *Computer Simulations of Liquids*. Clarendon, Oxford (1987).

[95] Sagui, C., L. G. Pedersen und T. A. Darden. Towards an Accurate Representation of Electrostatics in Classical Force Fields: Efficient Implementation of Multipolar Interactions in Biomolecular Simulations. *J. Chem. Phys.* **120**, 73–87 (2004).

[96] Perram, J. W., H. G. Petersen und S. W. D. Leeuw. An Algorithm for the Simulation of Condensed Matter which Grows as the 3/2 Power of the Number of Particles. *Mol. Phys.* **65**, 875–893 (1988).

[97] Darden, T. A., D. York und L. Pedersen. Particle Mesh Ewald: An N log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **98**, 10089–10092 (1993).

[98] Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee und L. G. Pedersen. A Smooth Particle Mesh Ewald Method. *J. Chem. Phys.* **103**, 8577–8593 (1995).

[99] Luty, B. A., I. G. Tironi und W. F. van Gunsteren. Lattice-Sum Methods for Calculating Electrostatic Interactions in Molecular Simulations. *J. Chem. Phys.* **103**, 3014–3021 (1995).

[100] Hünenberger, P. H. und J. A. McCammon. Effect of Artificial Periodicity in Simulations of Biomolecules under Ewald Boundary Conditions: a Continuum Electrostatics Study. *Biophys. Chem.* **78**, 69–88 (1999).

[101] Figueirido, F., R. M. Levy, R. Zhuo und B. J. Berne. Large Scale Simulation of Macromolecules in Solution: Combining the Periodic Fast Multipole Method with Multiple Time Step Integrators. *J. Chem. Phys.* **106**, 9835–9849 (1997).

[102] Challacombe, M., C. White und M. Head-Gordon. Periodic Boundary Conditions and the Fast Multipole Method. *J. Chem. Phys.* **107**, 10131–10139 (1997).

[103] Amisaki, T. Precise and Efficient Ewald Summation for Periodic Fast Multipole Method. *J. Comput. Chem.* **21**, 1075–1087 (2000).

[104] Ding, H.-Q., N. Karasawa und W. A. Goddard III. Atomic Level Simulations on a Million Particles: The Cell Multipole Method for Coulomb and London Nonbond Interactions. *J. Chem. Phys.* **97**, 4309–4315 (1992).

[105] Ding, H.-Q., N. Karasawa und W. A. G. III. The Reduced Cell Multipole Method for Coulomb Interactions in Periodic Systems with Million-Atom Unit Cells. *Chem. Phys. Lett.* **196**, 6–10 (1992).

[106] Shimada, J., H. Kaneko und T. Takada. Performance of Fast Multipole Methods for Calculating Electrostatic Interactions in Biomacromolecular Simulations. *J. Comput. Chem.* **15**, 28–43 (1994).

[107] Warren, M. S. und J. K. Salmon. A Portable Parallel Particle Program. *Comput. Phys. Commun.* **87**, 266–290 (1995).

[108] Martinetz, T., S. Berkovich und K. Schulten. 'Neural-Gas" Network for Vector Quantization and its Application to Time-Series Prediction. *IEE Trans. Neur. Networks* **4**, 558–569 (1993).

[109] Prakash, B., G. J. K. Praefcke, L. Renault, A. Wittinghofer und C. Herrmann. Structure of Human Guanylate-Binding Protein 1 Representing a Unique Class of GTP-Binding Proteins. *Nature* **403**, 567 – 571 (2000).

[110] Neumann, M. Dipole Moment Fluctuation Formulas in Computer Simulations of Polar Systems. *Mol. Phys.* **50**, 841–858 (1983).

[111] Dehnen, W. A Very Fast and Momentum-Conserving Tree Code. *Astrophys. J.* **536**, L39–L42 (2000).

[112] Shanker, B. und H. Huang. Accelerated Cartesian Expansions: A Fast Method for Computing of Potentials of the Form R^{-v} for All Real v. *J. Computat. Phys.* **226**, 732–753 (2007).

[113] Barnes, J. und P. Hut. A Hierarchical $O(N \log N)$ Force-Calculation Algorithm. *Nature* **324**, 446–449 (1986).

[114] Tuckerman, M. E. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford University Press USA, New York, 1 Auflage (2010).

[115] Egwolf, B. *Proteine in wässriger Umgebung: Kontinuumstheorie der Lösungsmittelelektrostatik und ihre effiziente Berechnung*. Doktorarbeit, Ludwig-Maximilians Universität München, Fakultät für Physik, Germany (2004).

[116] Tröster, P. und P. Tavan. The Microscopic Physical Cause for the Density Maximum of Liquid Water. *J. Phys. Chem. Lett.* **5**, 138–142 (2014).

[117] Schmidt, K. und M. Lee. Implementing the Fast Multipole Method in Three Dimensions. *J. Stat. Phys.* **63**, 1223–1235 (1991).

[118] Andoh, Y., N. Yoshii, K. Fujimoto, K. Mizutani, H. Kojima, A. Yamada, S. Okazaki, K. Kawaguchi, H. Nagao, K. Iwahashi, F. Mizutani, K. Minami, S.-i. Ichikawa, H. Komatsu, S. Ishizuki, Y. Takeda und M. Fukushima. MODYLAS: A Highly Parallelized General-Purpose Molecular Dynamics Simulation Program for Large-Scale Systems with Long-Range Forces Calculated by Fast Multipole Method (FMM) and Highly Scalable Fine-Grained New Parallel Processing Algorithms. *J. Chem. Theory Comput.* **9**, 3201–3209 (2013).

[119] Shimada, J., H. Kaneko und T. Takada. Efficient Calculations of Coulombic Interactions in Biomolecular Simulations with Periodic Boundary Conditions. *J. Comput. Chem.* **14**, 867–878 (1993).

Danksagung

Mein besonderer Dank am Gelingen dieser Arbeit gilt hier meinem Doktorvater Prof. Paul Tavan, der mir bei der Bearbeitung meines Themas immer geduldig mit Rat und Tat zu Seite stand. Insbesondere möchte ich mich hier für die Unterstützung beim Verfassen ansprechender wissenschaftlicher Texte und für das entgegengebrachte Vertrauen und die großen Freiräume in der täglichen Forschungsarbeit bedanken.

Vielen Dank auch an Gerald Mathias, der immer wertvolle Ratschläge zu erteilen wusste und so an den Lösungen vieler Fragestellungen großen Anteil hat. Insbesondere danke ich Gerald auch dafür, dass er mir mit der Beteiligung an einem KONWIHR Projekt am Ende meiner Doktorarbeit auch in diesem Zeitraum eine Finanzierung ermöglicht hat.

An dieser Stelle sollen auch Philipp Tröster, Christoph Wichmann, Magnus Schwörer und Sebastian Bauer nicht unerwähnt bleiben. Ich bin für die angenehme Gesellschaft dieser Kollegen sehr dankbar, die mir sicherlich half die eine oder andere Frustration im Forschungsalltag zu überwinden. Mit dem wertvollen Austausch über wissenschaftliche Probleme und auch der einen oder anderen fachfremden Freizeitgestaltung haben diese Kollegen großen Anteil daran, dass mir die Arbeit am BMO als gute Zeit in Erinnerung bleiben wird.

Last but not least danke ich meiner Familie für ihre Hilfe. Ohne die große Unterstützung von Monika, Klaus und Katharina über die Jahre des Studiums und der Doktorarbeit wäre ich sicher nie bis zu diesem Punkt gekommen. Danke!