An empirical extrapolation scheme for efficient treatment of induced dipoles

Andrew C. Simmonett,[1,a)] Frank C. Pickard IV,[1] Jay W. Ponder,[2] and Bernard R. Brooks[1]
[1]*Laboratory of Computational Biology, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, Maryland 20892, USA*
[2]*Department of Chemistry, Washington University in St. Louis, St. Louis, Missouri 63130, USA*

(Received 19 July 2016; accepted 4 October 2016; published online 24 October 2016)

Many cutting edge force fields include polarization, to enhance their accuracy and range of applicability. In this work, we develop efficient strategies for the induced dipole polarization method. By fitting various orders of perturbation theory (PT) dipoles to a diverse training set, we arrive at a family of fully analytic methods — whose *n*th order is referred to OPTn — that span the full spectrum of polarization methods from the fast zeroth-order approach that neglects mutual dipole coupling, approaching the fully variational approach at high order. Our training set contains many difficult cases where the PT series diverges, and we demonstrate that our OPTn methods still deliver excellent results in these cases. Our tests show that the OPTn methods exhibit rapid convergence towards the exact answer with each increasing PT order. The fourth order OPT4 method, whose costs are commensurate with three iterations of the leading conjugate gradient method, is a particularly promising candidate to be used as a drop-in replacement for existing solvers without further parameterization. [http://dx.doi.org/10.1063/1.4964866]

I. INTRODUCTION

Modern classical simulation methods use increasingly elaborate physics, such as multipole moments[1–11] and polarization,[12–20] to describe molecular interactions. The extra flexibility of these next-generation models affords accurate descriptions of interactions over a range of chemical environments, but introduces a computational penalty; it is important to devise algorithms that minimize this penalty in order to effectively sample configurational space. Polarization allows modeling of electron cloud distortions in response to the local electric field and is commonly effected by Drude oscillators[12–18,21,22] or induced dipoles.[3,23–40]

A Drude oscillator is simply a pair of particles with equal and opposite charges, connected by a harmonic spring. One of the particles is tethered to the atom whose polarizability is to be simulated, while the other moves in response to the field; thus providing the desired redistribution of the charge density. Because the Drude particles experience the field due to both the nuclei and the other Drude particles, minimization of the energy with respect to the Drude particle positions is formally an iterative self-consistent field (SCF) procedure. The simple charge-on-a-spring nature of Drude oscillators means that their treatment closely mirrors the classical treatment of nuclei, which makes the adaptation of existing codebases to include Drude polarization quite a straightforward, popular approach.

Induced dipoles are very closely related to Drude oscillators; instead of forming a "finite difference" dipole at each polarizable center using point charges, an analytic dipole is created. The ensuing need to evaluate dipole-dipole interactions leads to more complicated mathematical expressions than are required for the Drude model, which makes implementation into existing codes difficult. However, with fewer pairwise interactions to evaluate than for the Drude model, and with multipole evaluation algorithms being actively developed, the induced dipole strategy is widely used. The induced dipoles are defined as a response to the field resulting from the fixed charge distributions as well as the induced dipoles on other centers, so obtaining them is formally a self-consistent procedure, as for the Drude oscillators.

To avoid the computationally expensive process of self-consistently evaluating polarization, extended Lagrangian (EL) techniques have been developed for Drude oscillators,[14,15,21] induced dipoles,[3,41] and fluctuating charges.[42] The EL approach propagates the electronic and nuclear degrees of freedom simultaneously and is therefore a classical analog to the Car-Parinello *ab initio* dynamics method. The choice of mass of the Drude particle can be important in the EL scheme; too light a mass will necessitate short timesteps, while too heavy a mass will cause difficulty in maintaining separation of nuclear and electronic degrees of freedom. Introducing dual thermostats for the nuclear and electronic degrees of freedom helps to maintain adiabatic separation and permits the use of timesteps commensurate with conventional dynamics simulations.[14] Moreover, a potentially promising hybrid (EL/SCF) method has recently been developed that permits loose SCF convergence to be used, starting from a propagated EL guess; the resulting method offers very good energy conservation with conventional MD timesteps.[43]

While the EL and hybrid EL methods offer much promise for efficient, scalable simulations when run under suitable

a)Electronic mail: andrew.simmonett@nih.gov

conditions, we will focus on the problem of obtaining induced dipoles using methodology that is compatible with conventional integration techniques. Many numerical solvers exist for self-consistently obtaining induced dipoles but, for reasons that will be outlined in Section II, care must be exercised when using them as instabilities may result from overly permissive convergence criteria. To address this instability, while maintaining efficiency, we recently developed a strategy[19] to obtain an analytic representation of induced dipoles; the analytic nature of this polarization method makes it a close relative of conventional, fixed point charge methods even though it encompasses the many-body character of induced dipoles. Based on perturbation theory (PT), our approach is equivalent to the variational, self-consistent approach at infinite order, when in the convergent regime. By analyzing the properties of the PT series, we developed an extrapolation procedure to approximate the infinite-order solution using only low order solutions. In this work, we identify deficiencies in the previous extrapolation procedure and consider a more pragmatic fitting approach to combine the lowest orders of the PT series, resulting in a series of accurate, efficient, and analytic expressions for induced dipoles.

II. THEORY

The polarization energy for a system of N induced dipoles $\boldsymbol{\mu}$ can be obtained from a $3N \times 3N$ coupling tensor $\mathbf{T}$ and the electric field due to permanent multipole moments, $\mathbf{E}$, at the polarizable centers

$$U = \frac{1}{2}\boldsymbol{\mu}^T\mathbf{T}\boldsymbol{\mu} - \mathbf{E}^T\boldsymbol{\mu}. \tag{1}$$

The tensor $\mathbf{T}$ comprises two components,

$$\mathbf{T} = \boldsymbol{\alpha}^{-1} - \boldsymbol{\mathcal{T}}, \tag{2}$$

whose diagonal blocks are the 3×3 inverse atomic polarizabilities, and the off-diagonal terms $\boldsymbol{\mathcal{T}}$ are the 3×3 coupling terms that describe the damped[44] interactions between induced dipoles on different centers, whose exact formulation is not important for the present discussion. For brevity, we will focus on the case where $\boldsymbol{\alpha}$ is isotropic; the extension to anisotropic tensors is straightforward and has been discussed in Ref. 34 as have the details of the coupling terms $\boldsymbol{\mathcal{T}}$.

Stationarity of Eq. (1) defines the variational condition for the induced dipoles

$$\boldsymbol{\mathcal{R}} \equiv \left(\frac{\partial U}{\partial \boldsymbol{\mu}}\right)^T = \mathbf{T}\boldsymbol{\mu} - \mathbf{E} = \mathbf{0} \tag{3}$$

and the polarization energy gradient, with respect to nuclear positions $\mathbf{r}$, is

$$\frac{\mathrm{d}U}{\mathrm{d}\mathbf{r}} = \frac{\partial U}{\partial \mathbf{r}} + \frac{\partial U}{\partial \boldsymbol{\mu}}\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{r}}. \tag{4}$$

The final term in Eq. (4) which is due to dipole response, contains derivatives of the induced dipoles $\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{r}}$. As noted in Sec. III, the $\boldsymbol{\mu}$ derivatives are problematic to evaluate because $\boldsymbol{\mu}$ itself is obtained as a numerical solution to Eq. (3). Because the residual $\boldsymbol{\mathcal{R}}$ appears as a factor of the dipole response terms, they are usually neglected. The extent to which the residual $\boldsymbol{\mathcal{R}}$ can be assumed to be zero depends on how tightly the numerical solution for $\boldsymbol{\mu}$ is obtained; if loose convergence criteria are employed, the dipole response terms could be too large to safely neglect, leading to unstable trajectories.

The PT approach can be derived by introducing an ordering parameter, λ, into the $\mathbf{T}$ coupling tensor

$$\mathbf{T} = \boldsymbol{\alpha}^{-1} - \lambda\boldsymbol{\mathcal{T}} \tag{5}$$

and expressing the resulting dipoles using a power series in λ,

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_{(0)} + \lambda\boldsymbol{\mu}_{(1)} + \lambda^2\boldsymbol{\mu}_{(2)} + \cdots + \lambda^n\boldsymbol{\mu}_{(n)}. \tag{6}$$

The nth order dipole comprises $n+1$ components, which are labeled with their order in parentheses. Substituting the expanded quantities, Eqs. (5) and (6), into Eq. (3) and collecting by powers of λ yields analytic expressions for the induced dipoles at each order

$$\begin{aligned} \boldsymbol{\mu}_{(0)} &= \boldsymbol{\alpha}\mathbf{E}, \\ \boldsymbol{\mu}_{(1)} &= \boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\boldsymbol{\alpha}\mathbf{E}, \\ \boldsymbol{\mu}_{(2)} &= \boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\boldsymbol{\alpha}\boldsymbol{\mathcal{T}}\boldsymbol{\alpha}\mathbf{E}, \\ &\vdots \qquad\qquad \vdots \\ \boldsymbol{\mu}_{(n)} &= \boldsymbol{\alpha}(\boldsymbol{\mathcal{T}}\boldsymbol{\alpha})^n\mathbf{E}. \end{aligned} \tag{7}$$

The nth order energy is simply

$$U_{\mathrm{n}} = -\frac{1}{2}\mathbf{E}^T\boldsymbol{\mu}_n. \tag{8}$$

The spectrum of PT methods represents a family of methods that each offer a different level of compromise between accuracy and efficiency. One crucial difference between a loosely converged variational solution and a PT approach is that the former requires additional linear equations to be solved to properly obtain the nuclear energy gradient, while the latter is analytically differentiable.

In our original development of PT,[19] we focused on the odd terms of the U_{n} series, which we denoted U_{PTn}. Because each dipole component in the U_{PTn} series converges exponentially, a three-point exponential fit $U_{\mathrm{PT}\infty} = U_{\mathrm{PTn}} - b\exp(-cn)$ is an effective way to reach the infinite order limit, under the assumption that all components converge at the same rate. To obtain the three unknowns, $U_{\mathrm{PT}\infty}$, b, and c, would require $\{U_{\mathrm{PT0}}, U_{\mathrm{PT1}}, U_{\mathrm{PT2}}\}$ or, equivalently, $\{U_1, U_3, U_5\}$. Our extrapolated PT (ExPT) method reduced these requirements by additionally assuming that the exponent c is fixed for all systems, reducing the fit to just two points

$$\boldsymbol{\mu}_{\mathrm{ExPT}} = c_0\boldsymbol{\mu}_1 + c_1\boldsymbol{\mu}_3, \tag{9}$$

which has a single empirical parameter due to the constraint $c_0 + c_1 = 1$.

Inspection of Eq. (7) reveals that generation of $\boldsymbol{\mu}_3$ also yields $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ of which only $\boldsymbol{\mu}_1$ is utilized in the ExPT approach. In this work, we adopt a more empirical approach to determine the coefficients of the PT dipoles, which remedies the wasted information in ExPT by using a more general ansatz

$$\boldsymbol{\mu}_{\mathrm{OPTn}} = M_0\boldsymbol{\mu}_0 + M_1\boldsymbol{\mu}_1 + M_2\boldsymbol{\mu}_2 + \cdots + M_n\boldsymbol{\mu}_n, \tag{10}$$

where the coefficients $\{M_0, M_1, \ldots, M_n\}$ are to be determined by a fitting procedure. The resulting nth-order optimized PT method is denoted OPTn. With this notation, ExPT is a special case of OPT3, with coefficients $\{0, c_0, 0, c_1\}$. To parameterize the functional form, we will consider a diverse collection of 50 systems, described in Section III and the supplementary material.

For each system, we minimize the objective function

$$S = \sum_{\alpha}^{3N} (\mu_\alpha - \mu_{\mathrm{OPTn},\alpha})^6 \tag{11}$$

using standard minimization techniques,[45] where the summation runs over all $3N$ dipole components for each system. The residual is normally raised to the second power in conventional least squares fitting, but we use a power of 6 here to favor reduction of outliers, thus creating a more uniform distribution of errors; we deem this uniformity of errors more important than obtaining a lower RMS error.

Analogous to Eq. (8), the OPTn energy is

$$\begin{aligned} U_{\mathrm{OPTn}} &= -\frac{1}{2}\mathbf{E}^T \boldsymbol{\mu}_{\mathrm{OPTn}} \\ &= -\frac{1}{2}\mathbf{E}^T \left(M_0\boldsymbol{\mu}_0 + M_1\boldsymbol{\mu}_1 + \cdots + M_n\boldsymbol{\mu}_n\right) \\ &= -\frac{1}{2}\mathbf{E}^T \left(m_0\boldsymbol{\mu}_{(0)} + m_1\boldsymbol{\mu}_{(1)} + \cdots + m_n\boldsymbol{\mu}_{(n)}\right), \end{aligned} \tag{12}$$

where, for convenience, we have expressed the induced dipoles in terms of their components using the relationship

$$m_i = \sum_{j=i}^{n} M_j. \tag{13}$$

Because the components of $\boldsymbol{\mu}_{\mathrm{OPTn}}$ have an analytic form, computing polarization energy derivatives are very straightforward

$$\begin{aligned} \frac{\partial U_{\mathrm{OPTn}}}{\partial \mathbf{r}} &= -\boldsymbol{\mu}_{\mathrm{OPTn}}^T \frac{\partial \mathbf{E}}{\partial \mathbf{r}} \\ &\quad + \frac{1}{2}\sum_{l}^{n} m_l \sum_{m=0}^{n-1} \boldsymbol{\mu}_{(m)} \frac{\partial \boldsymbol{\mathcal{T}}}{\partial \mathbf{r}} \boldsymbol{\mu}_{(l-m-1)}. \end{aligned} \tag{14}$$

The leading term of Eq. (14) is present in both the variational and "direct" (where $\boldsymbol{\mathcal{T}}$ is neglected) polarization algorithms. The additional $n(n+1)$ terms arising from dipole response can be efficiently evaluated by caching the field and field gradient due to $\{\boldsymbol{\mu}_{(0)}, \boldsymbol{\mu}_{(1)}, \ldots, \boldsymbol{\mu}_{(n-1)}\}$ during the formation of the induced dipoles, Eq. (7).

The terms required to implement the energies and forces for OPT are similar to those needed in the variational algorithm, and we have implemented the OPT method in development versions of CHARMM,[46,47] TINKER,[48] and OpenMM.[49] All computations described hereafter were performed using the TINKER[48] simulation package.

Constructing a dataset that spans the complete chemical and configurational space is not possible, so our choice was motivated by the following considerations. Inspection of the AMOEBA parameters reveals that there is little variety in the polarization parameters for many main group atoms, with the exception of some highly polarizable species such as sulfur and chloride ions. By including some archetypal protein, RNA, and DNA systems, we have many representatives of the "typical" polarizabilities and we must be sure to include systems with sulfur and chlorine atoms to also consider the outliers. We also include some liquids with a range of dipole moments to probe homogeneous systems and solvated ions to represent the more problematic inhomogeneous liquid systems.

The resulting set of 50 training systems is detailed in the supplementary material and, for the following discussion, are loosely categorized into three groups: homogeneous liquids, solvated ions, and biological systems. The homogeneous liquids are 13 small, organic molecules including benzene, acetonitrile, dimethyl sulfoxide, ammonia, and methanol. These systems were chosen to represent a range of polarities. Because doubly charged cations are known to be tough to describe, our seven solvated ion systems comprise a 34.14 Å cubic box containing 1331 water molecules, as well as the same water box containing 1, 2, 3, 4, 5, and 6 $MgCl_2$ molecules. The biological systems include a range of proteins, RNA and DNA systems, harvested from the protein databank (PDB). The liquid and ion systems were equilibrated at 300 K, while the biological systems were subjected to a crude energy minimization, followed by 100 steps of dynamics to eliminate any bad contacts.

III. RESULTS

We parameterized the OPTn (n = 0–4) family of methods for each of the 50 test molecules using Equation (11) with tightly converged variational dipoles as the reference. To test the sensitivity of the optimization solutions to the initial guess coefficients, a number of starting conditions were tried. First, we used the guess coefficients $\{0, 0, \ldots, 0, 1\}$, which correspond to the nth order perturbation theory method U_n. Second, we tried the uniform guess $\{\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\}$, which weights all PT components equally. Finally, after observing the oscillatory convergence of the PT series, discussed below, we tried the guess $\{0, 0, \ldots, 0, \frac{1}{2}, \frac{1}{2}\}$ which is the mean of the two highest orders of PT, $\frac{U_n + U_{n-1}}{2}$. Although all of these guesses have coefficients that sum to one, consistent with the ExPT treatment, Eq. (9), no restrictions were placed on the sum of the coefficients during the optimization. For all systems in the training set and all orders of OPTn optimization, the final parameters were identical for all starting guesses.

The resulting parameters are shown alongside the resulting RMS induced dipole errors in Table I. We included OPT0 in our parameterization because this uses the same "direct" polarization algorithm as the iAMOEBA method. Direct polarization is obtained by completely neglecting coupling between induced dipoles, averting the need for any expensive matrix-vector products. In iAMOEBA, the entire set of bonded and noncovalent parameters were re-optimized to compensate for the lack of mutual dipole coupling; our PT0 parameterization simply introduces a polarization scale factor, which is equivalent to uniformly scaling all polarizabilities.

Dipole response force errors notwithstanding, a target RMS change in the dipoles of 0.01 D has been considered

TABLE I. Results of the OPTn fitting procedures.[a]

Method	M_0	M_1	M_2	M_3	M_4	$\sum_j M_j$	Dipole error (D)[b]
OPT0	1.044 (0.142)					1.044	0.084 (0.036)
OPT1	0.412 (0.103)	0.784 (0.096)				1.197	0.029 (0.012)
OPT2	−0.115 (0.081)	0.568 (0.079)	0.608 (0.126)			1.062	0.012 (0.006)
OPT3	−0.154 (0.036)	0.017 (0.120)	0.657 (0.050)	0.475 (0.125)		0.995	0.006 (0.003)
OPT4	−0.041 (0.032)	−0.176 (0.026)	0.169 (0.154)	0.663 (0.027)	0.374 (0.124)	0.987	0.004 (0.003)

[a]The M_j coefficients shown are those defined in Eq. (10). Standard deviations across the data set are shown in parentheses.
[b]The mean RMS error in the OPTn induced dipoles across the training set of 50 molecules with respect to the tightly-converged, variational reference values.

a sufficient stopping criterion for iterative dipole solvers in previous works.[29,50] Although modern protocols commonly specify much tighter convergence of at least 10^{-5} D, we will consider 0.01 D as a desirable threshold error for our methods to deliver, keeping in mind that the forces are always evaluated exactly in these approximations, unlike the loosely converged variational solutions. Inspection of Table I reveals that the mean error in the OPT2 dipoles is 0.012 D across the data set, while for OPT3 this drops to 0.006 D; we will therefore focus much of our discussion on the OPT3 method. To visualize the spread in ideal third order coefficients (i.e., those that are optimal for each system) for each system in the training set, whose mean defines the consensus OPT3 coefficients, those ideal coefficients are plotted in Figure 1, alongside the OPT3 coefficients. The homogeneous liquids and ionic liquids adopt similar coefficients, while the biological systems generally adopt more positive M_1 and more negative M_3 coefficients than the ions and liquids. Among the ionic liquids, the single outlier possessing a large M_1 and corresponding low M_3 is benzene, which, like the biological systems, has a relatively low dielectric. For the even terms in the series, the coefficients for all systems are more closely clustered.

The OPT3 coefficients closely resemble an average of the two highest orders of perturbation theory, which is akin to the quantum mechanical Møller-Plesset MP2.5 method[51] that is an average of the MP2 and MP3 methods, consistent with the oscillatory convergence patterns. On the other hand, the ExPT coefficients have a rather different structure, while the $\boldsymbol{\mu}_2$ coefficient in the OPT3 method is the largest for any of the four $\boldsymbol{\mu}_n$ components, that same coefficient is zero in the ExPT approach as a direct consequence of the assumption that all dipole components converge exponentially, with the same exponent. Those same assumptions lead to the constraint that the ExPT coefficients must sum to one; although no such constraint was applied in the OPT3 fit, the coefficients sum to 0.995.

Figure 2 shows the RMS atomic induced dipole errors for each system in the training set, for a range of induced dipole algorithms. The ExPT method offers a significant reduction in the errors for the liquids and ionic systems upon which it was initially tested, but performs very poorly for the biological systems, for which even the direct algorithm offers better performance. The poor performance of ExPT for highly inhomogeneous systems can be explained by the plots in Figure 3, which depict the convergence behavior of the PT series for some representative cases for each of the three system types in our training set. The odd terms in the series are convergent for the $MgCl_2$ and acetic acid cases but divergent for the protein test. Our preliminary development of ExPT included a test that was divergent but exhibited convergence in the lower orders of PT before diverging; for cases such as the dry protein crystals examined herein, the lower odd orders of PT are often divergent, causing ExPT to fail. The PT0 method offers little improvement over the direct algorithm, while PT1 introduces massive improvements, especially for proteins, with all systems possessing an RMS induced dipole error below 0.05 D. The OPT4 method offers only a marginal

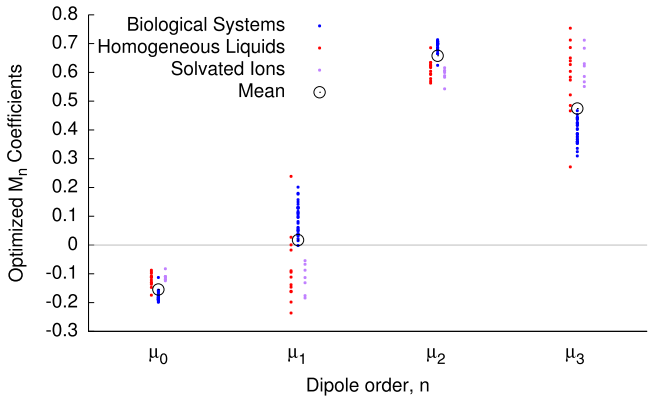

FIG. 1. Ideal third-order coefficients for each system in the training set, classified into three broad categories, described in the text. The mean of the set of 50 values for each coefficients constitute the OPT3 coefficients, which are depicted as hollow, black circles.

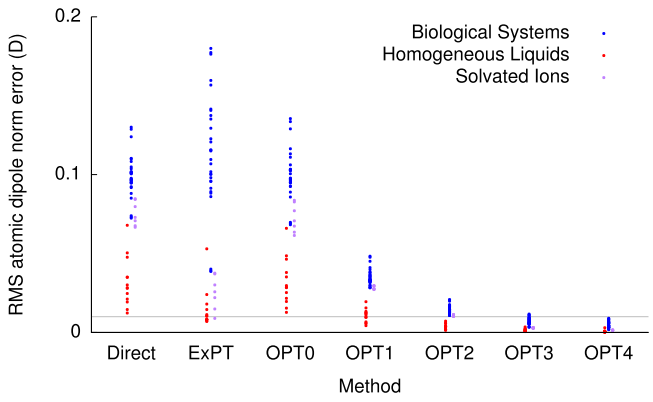

FIG. 2. The RMS atomic induced dipole errors for each system in the training set, broken down by system type, for a range of induced dipole algorithms. The horizontal gray line depicts an error of 0.01 D.

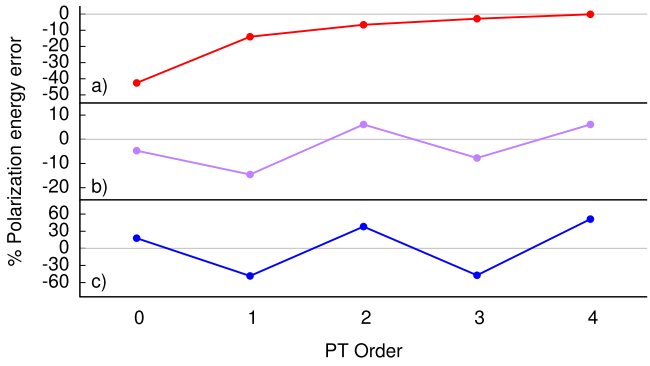

FIG. 3. Observed behavior of polarization energy errors for the perturbation series as a function of series order for (a) a monotonically convergent case (acetic acid) (b) oscillatory convergence (6 $MgCl_2$ in 1331 waters), and (c) a divergent case (albumin binding protein, 1PRB).

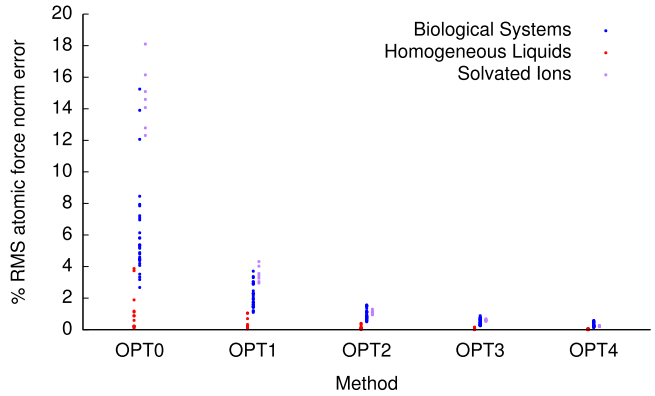

FIG. 4. Mean absolute errors in the norm of the atomic forces, for each system in the training set for the OPTn methods developed in this work.

improvement over OPT3, reducing the mean RMS induced dipole error by just 0.002 D.

Despite the very disparate convergence patterns observed in our training set, optimized perturbation theory performs very well across the board. Remembering that the OPT3 energy is mostly an average of U_2 and U_3, with a slight emphasis on the former, it is evident from Figure 3 that such a strategy should work. In the monotonically convergent case, both values fall close to the exact result and are pushed closer by the small amount of U_0 that is subtracted. When the series is oscillatory, whether convergent or not, the averaging of contiguous PT methods yields a result with close to zero error by virtue of the odd and even terms bounding the exact result. A detailed breakdown of the induced dipole errors, OPT3 force errors, and convergence analysis of the PT series, for each system in the training set is provided in the supplementary material.

Figure 4 shows the mean error in the norm of the force on each atom, plotted for each of the OPTn methods, for all systems in the training set. The homogeneous liquids clearly represent a far less challenging system than the other cohorts; the OPT0 method produces errors in the atomic force norms within 4% of the exact value, and this drops down to 0.4% for OPT2. The solvated ions have force errors as large as 18%, which reduces below 2% for OPT2 and just 0.7% for OPT3; a similar trend is observed for the protein systems, with OPT3 delivering errors within 0.9%.

To gauge the quality of condensed phase properties, Table II shows some computed properties of water for OPTn (n = 0-4), ExPT and the variational reference method computed from a 1 ns NPT simulation of 729 AMOEBA03 water molecules.[29] The density appears to be well modeled for all methods, with the sequence OPT1 to OPT4 offering systematically decreasing errors from −0.5% for OPT1, falling to just −0.2% for OPT3 and culminating in agreement between OPT4 and the reference; OPT0 is fortuitously close for this property. The self-diffusion is harder to model correctly, with a very large deviation observed for OPT0, and even OPT3 overestimates the water self-diffusion by 10%. As is the case for the density, complete agreement is obtained between OPT4 and the reference calculation. The error in the mean potential energy is −0.19 kcal mol^{-1} per molecule for OPT3, dropping to just 0.01 kcal mol^{-1} per molecule for OPT4, with similar fluctuations observed for both. The formulation of ExPT considered only energies, so its deviation of just 0.08 kcal mol^{-1} per molecule is unsurprising. This shows that accurate energies can be captured by third order methods. However, the fact that ExPT provides the poorest description of the density for all methods tested here shows that properties beyond the energy should be considered in the model development. Our use of the dipoles as a target for parameterization in OPTn has yielded a series of methods that offer systematically improving performance for describing water.

One potential pitfall of a perturbative scheme is the singularity at zero bond length, which could lead to far more extreme divergence of the series than we observe in our training set, in the presence of anomalously short contacts. The use of Thole damping — effectively blurring the point induced dipole — mitigates this, as is evident from Figure 5, which compares the OPTn (n = 1-3) methods to the iterative method in describing the potential energy curve, and derivative thereof, for $H_2O\cdots Mg^{2+}$ dissociation along the C_{2v} axis. At large separations, the polarization effect is small, and at short distances the Thole damping greatly diminishes the magnitude; in these extremes, the agreement for OPT2 and OPT3 with the variational reference curve is excellent. Around the equilibrium region, the OPT2 and OPT3

TABLE II. Properties for AMOEBA water, computed from 1 ns NPT simulations, using a range of polarization algorithms.[a]

Method	ρ[b]	D[c]	V[d]
ExPT	0.992 (0.005)	1.63 (0.07)	−9.100 (0.072)
OPT0	0.999 (0.005)	4.52 (0.15)	−7.984 (0.063)
OPT1	0.995 (0.007)	1.41 (0.05)	−9.510 (0.073)
OPT2	0.997 (0.007)	1.49 (0.08)	−9.313 (0.070)
OPT3	0.998 (0.006)	2.04 (0.06)	−8.831 (0.072)
OPT4	1.000 (0.006)	1.85 (0.06)	−9.015 (0.073)
SCF	1.000 (0.006)	1.85 (0.06)	−9.025 (0.068)

[a]The SCF entry corresponds to tightly converged, variational reference values.
[b]The density (g cm^{-3}) with standard deviation in parentheses.
[c]The self-diffusion constant (10^5 cm^2 s^{-1}) with standard deviation in parentheses.
[d]The mean potential energy per molecule (kcal mol^{-1}) with standard deviation in parentheses.

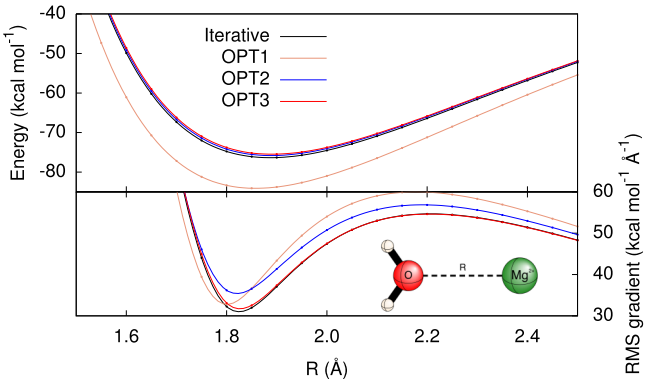

FIG. 5. Constrained potential energy scans for the $H_2O\cdots Mg^{2+}$ dimer, using the variational, OPT1, OPT2, and OPT3 methods. The top plot depicts the potential energy while the bottom shows the RMS atomic force.

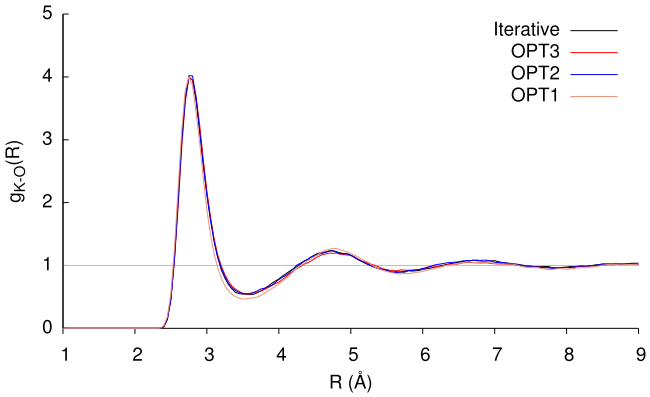

FIG. 7. Radial distribution function for the K^+–O pair, derived from 500 ps simulation of KCl in 1331 water molecules at 300 K.

methods yield a slight under-binding of *ca*. 1 kcal mol^{-1}, but both greatly outperform the simpler OPT1 method, which greatly overestimates the polarization stabilization across the entire potential curve. The plot of the RMS force on the system reveals that, although the OPT2 energies very closely track the reference values, the forces are quite systematically overestimated along the dissociation coordinate, with a minimum that occurs at a slightly shorter bond length. To investigate the effects of these errors, we simulated a periodic system comprising a single $MgCl_2$ molecule in 1331 waters for 500 ps, in an NVT ensemble at 300 K; the resulting radial distribution functions are shown in Figure 6. The OPT2 and OPT3 distribution functions are almost indistinguishable from the variational reference. Although OPT1 correctly predicts a sharp peak at 2.1 Å, corresponding to the first solvation shell, the second solvation shell is erroneously placed at 4.4 Å instead of 4.2 Å. As a test of performance for monovalent ion solvation, Figure 7 shows the analogous radial distribution function for KCl. As for $MgCl_2$, all three OPT methods tested provide a very accurate description of the first solvation shell. The weakly-structured second and third solvation shells are described very well by OPT2 and OPT3, with OPT1 offering a very slightly over structured description in the vicinity of the second shell.

To further probe the effect of simulation conditions on the parameterization, we studied ubiquitin with the 58 water molecules found in the PDB file, the same system with no water molecules and the same system with a total of 3071 water molecules to fill the unit cell; the resulting ideal third order coefficients are $\{-0.18, 0.06, 0.70, 0.42\}$, $\{-0.18, 0.10, 0.70, 0.39\}$, and $\{-0.19, 0.08, 0.70, 0.40\}$, respectively. The same molecule with no waters and no periodic images has ideal third order coefficients of $\{-0.18, 0.05, 0.70, 0.42\}$. Such insensitivity to periodicity and solvation effects supports the idea that a universal set of coefficients may be employed for all molecular systems.

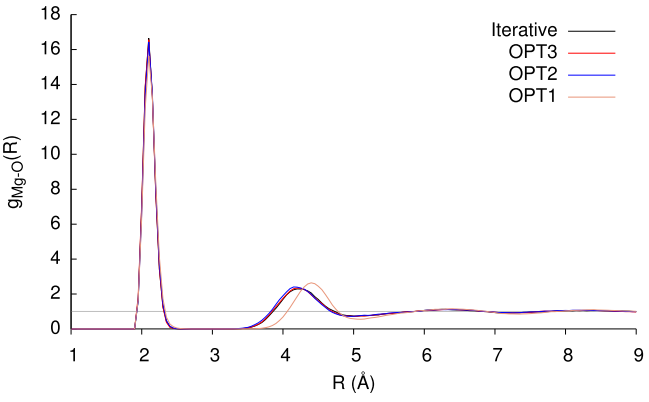

FIG. 6. Radial distribution function for the Mg^{2+}–O pair, derived from 500 ps simulation of $MgCl_2$ in 1331 water molecules at 300 K.

Our preliminary implementation provides functionality for users to determine ideal expansion coefficients for any system of interest. This offers a remedy for any difficult cases that may be encountered, for which the consensus OPTn coefficients may not yield satisfactory agreement with reference values. However, any coefficient tailored this way must be explicitly reported in the interests of reproducibility.

Recent efforts to improve the efficiency of induced dipole treatments have yielded some promising methods, including conjugate gradient (CG) SCF solvers[34] and SCF using a propagated EL guess.[43] Although the number of iterations needed to converge the conventional SCF equations to a given tolerance depends on the algorithm used, the nature of the system under study and the desired convergence level, we will briefly compare the computational cost of these methods to OPTn. Reference 43 reports that the leading CG solver with a predictor guess[34] achieves convergence of the dipoles to 10^{-6} D in 5 SCF cycles, which generates a drift of 4.63×10^{-6} kcal mol^{-1} ps^{-1} for a water box. Because CG requires a matrix-vector product (MVP) in the setup and another in each iteration, this corresponds to 6 MVPs, which provides a good measure of the overall polarization cost. By introducing a thermostat for the auxiliary degrees of freedom used to obtain the dipole guess in the hybrid EL/SCF approach, similar energy conservation can be achieved by converging the SCF equations to just 10^{-2} D, which requires 4 iterations of CG (5 MVPs);[43] good energy conservation ($\sim 3 \times 10^{-5}$ kcal mol^{-1} ps^{-1}) is also realized when a criterion of 10^{-1} D is used, at a cost of 3 iterations (4 MVPs).

The ExPT method conserves energy[19] due to the forces being calculated analytically, but the water properties

computed herein show some significant deviations with respect to SCF reference values. The OPT3 method, like ExPT, has analytic forces and costs 3 MVPs but provides much better properties for water. Moreover the OPT3 method produces accurate dipoles over a diverse range of compounds, where ExPT fails; this is because the former uses a less constrained parameterization scheme and makes no *a priori* assumptions about the convergence behavior of the PT series. Adaptation of an ExPT code to use OPT3 coefficients is trivial due to the similarity of their formulation. Similarly, generalizing the implementation for arbitrary-order OPTn is also straightforward.

IV. CONCLUSIONS

Building on our previous work, we have developed a new series of perturbation theory techniques for induced dipoles; the current work proposes new extrapolation techniques to accurately approximate the exact solution with only a few low order terms. By considering a diverse set of molecules and simply optimizing the coefficients for each term in the PT series up to *n*th order, we have developed the OPTn family of methods. The resulting methods form a hierarchy of approaches that span the spectrum from the "direct" algorithm, where mutual coupling is completely neglected, approaching the exact solution. One key feature of the OPTn methods is that they are fully analytic at all levels of approximation. While forces from the more approximate, lower order OPTn methods are less accurate than their higher order analogs, these forces are just as precise as the energies — leading to energy conservation. In contrast, attempting to accelerate iterative solvers by loosening convergence criteria for the variational induced dipole method yields a family of methods that are accurate but lack precision; this reduced precision can manifest itself in catastrophically erroneous forces, so caution must be exercised when attempting to tune such approaches for dynamics.

The OPT3 method offers excellent computational efficiency, requiring just three of the rate-limiting matrix-vector products and is able to deliver dipoles with an RMS error of just 0.006 D across a diverse set of 50 molecules. More extensive testing is being performed to understand how these small errors manifest themselves in various chemical properties. The data presented herein suggest that OPT3 is the minimal possible OPTn method that may be considered for use as a drop-in replacement for iterative algorithms, without any reparameterization of the force field. The more approximate OPT2 method may not be accurate enough to be considered as a drop-in replacement for CG solvers, and its use may require reparameterization of the underlying force field. We note that an algorithm equivalent to OPT1 is already used as the polarization algorithm for the POSSIM force field,[52] with the parameters defined accordingly. The OPT4 method requires one more matrix-vector product than OPT3, but appears to be very robust with respect to accurately describing a range of compounds with a universal set of coefficients. We do not recommend pursuing higher order methods (n > 4), as they would not be competitive with currently used SCF methods.[34,43,53]

SUPPLEMENTARY MATERIAL

See supplementary material for details of the training set, and a listing of the ideal coefficients, induced dipole errors, PT series convergence patterns, and OPT3 atomic force error distributions for each member of this set.

ACKNOWLEDGMENTS

This work was supported by the intramural research program of the National Heart, Lung and Blood Institute. J.W.P. wishes to acknowledge support for development of the AMOEBA force field from Nos. NIH GM106137 and NIH GM114237. A.C.S. is thankful to Dr. P. K. Eastman for developing the CUDA implementation of OPT*n* in OpenMM.

[1]A. J. Stone, *The Theory of Intermolecular Forces* (Oxford University Press, 2013).
[2]W. Smith, CCP5 Quarterly **4**, 13 (1982).
[3]A. Toukmaji, C. Sagui, J. Board, and T. A. Darden, J. Chem. Phys. **113**, 10913 (2000).
[4]C. Sagui, T. A. Darden, and L. G. Pedersen, J. Chem. Phys. **120**, 73 (2004).
[5]T. J. Giese and D. York, J. Chem. Phys. **128**, 064104 (2008).
[6]A. C. Simmonett, F. C. Pickard, H. F. Schaefer, and B. R. Brooks, J. Chem. Phys. **140**, 184101 (2014).
[7]M. Devereux, S. Raghunathan, D. G. Fedorov, and M. Meuwly, J. Chem. Theory Comput. **10**, 4229 (2014).
[8]H. A. Boateng and I. T. Todorov, J. Chem. Phys. **142**, 034117 (2015).
[9]D. M. Rogers, J. Chem. Phys. **142**, 074101 (2015).
[10]T. J. Giese, M. T. Panteva, H. Chen, and D. York, J. Chem. Theory Comput. **11**, 436 (2015).
[11]D. Lin, J. Chem. Phys. **143**, 114115 (2015).
[12]P. Drude, Ann. Phys. **308**, 369 (1900).
[13]P. Drude, Ann. Phys. **306**, 566 (1900).
[14]G. Lamoureux and B. Roux, J. Chem. Phys. **119**, 3025 (2003).
[15]W. Jiang, D. J. Hardy, J. C. Phillips, A. D. MacKerell, Jr., K. Schulten, and B. Roux, J. Phys. Chem. Lett. **2**, 87 (2011).
[16]J. Huang, P. E. M. Lopes, B. Roux, and A. D. MacKerell, Jr., J. Phys. Chem. Lett. **5**, 3144 (2014).
[17]K. Vanommeslaeghe and A. D. MacKerell, Jr., Biochim. Biophys. Acta **1850**, 861 (2015).
[18]J. A. Lemkul, B. Roux, D. van der Spoel, and A. D. MacKerell, J. Comput. Chem. **36**, 1473 (2015).
[19]A. C. Simmonett, F. C. Pickard IV, Y. Shao, T. E. Cheatham III, and B. R. Brooks, J. Chem. Phys. **143**, 074115 (2015).
[20]Y. Shi, P. Y. Ren, M. Schnieders, and J.-P. Piquemal, Rev. Comput. Chem. **28**, 51 (2015).
[21]P. E. M. Lopes, B. Roux, and A. D. MacKerell, Jr., Theor. Chem. Acc. **124**, 11 (2009).
[22]J. A. Lemkul, J. Huang, B. Roux, and A. D. MacKerell, Chem. Rev. **116**, 4983 (2016).
[23]L. Silberstein, Philos. Mag. Ser. **33**, 92 (1917).
[24]L. Silberstein, Philos. Mag. Ser. **33**, 521 (1917).
[25]J. Applequist, J. R. Carl, and K. K. Fung, J. Am. Chem. Soc. **94**, 2952 (1972).
[26]T. P. Straatsma and J. A. McCammon, Chem. Phys. Lett. **167**, 252 (1990).
[27]T. P. Straatsma and J. A. McCammon, Chem. Phys. Lett. **177**, 433 (1991).
[28]B. Roux, Chem. Phys. Lett. **212**, 231 (1993).
[29]P. Y. Ren and J. W. Ponder, J. Phys. Chem. B **107**, 5933 (2003).
[30]G. A. Kaminski, R. A. Friesner, and R. Zhou, J. Comput. Chem. **24**, 267 (2003).
[31]J. W. Ponder, C. Wu, P. Y. Ren, V. S. Pande, J. D. Chodera, M. J. Schnieders, I. Haque, D. L. Mobley, D. S. Lambrecht, and R. A. DiStasio, J. Phys. Chem. B **114**, 2549 (2010).
[32]P. Y. Ren, C. Wu, and J. W. Ponder, J. Chem. Theory Comput. **7**, 3143 (2011).
[33]L.-P. Wang, T. Head-Gordon, J. W. Ponder, P. Y. Ren, J. D. Chodera, P. K. Eastman, T. J. Martinez, and V. S. Pande, J. Phys. Chem. B **117**, 9956 (2013).
[34]F. Lipparini, L. Lagardère, B. Stamm, E. Cancès, M. Schnieders, P. Y. Ren, Y. Maday, and J.-P. Piquemal, J. Chem. Theory Comput. **10**, 1638 (2014).
[35]M. L. Laury, L.-P. Wang, V. S. Pande, T. Head-Gordon, and J. W. Ponder, J. Phys. Chem. B **119**, 9423 (2015).
[36]R. Qi, L.-P. Wang, Q. Wang, V. S. Pande, and P. Y. Ren, J. Chem. Phys. **143**, 014504 (2015).

[37]M. S. Gordon, D. G. Fedorov, S. R. Pruitt, and L. V. Slipchenko, Chem. Rev. **112**, 632 (2012).
[38]G. A. Cisneros, J. Chem. Theory Comput. **8**, 5072 (2012).
[39]R. E. Duke, O. N. Starovoytov, J.-P. Piquemal, and G. A. Cisneros, J. Chem. Theory Comput. **10**, 1361 (2014).
[40]O. Engkvist, P.-O. Åstrand, and G. Karlström, Chem. Rev. **100**, 4087 (2000).
[41]M. Souaille, H. Loirat, D. Borgis, and M. P. Gaigeot, Comput. Phys. Commun. **180**, 276 (2009).
[42]S. W. Rick, S. J. Stuart, and B. J. Berne, J. Chem. Phys. **101**, 6141 (1994).
[43]A. Albaugh, O. Demerdash, and T. Head-Gordon, J. Chem. Phys. **143**, 174104 (2015).
[44]B. Thole, Chem. Phys. **59**, 341 (1981).
[45]J. W. Ponder and F. M. Richards, J. Comput. Chem. **8**, 1016 (1987).
[46]B. R. Brooks, R. E. Bruccoleri, D. J. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[47]B. R. Brooks, C. L. Brooks III, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**, 1545 (2009).
[48]J. W. Ponder, *TINKER: Software Tools for Molecular Design, 7.0* (Washington University School of Medicine, Saint Louis, MO, 2015).
[49]P. K. Eastman, M. S. Friedrichs, J. D. Chodera, R. J. Radmer, C. M. Bruns, J. P. Ku, K. A. Beauchamp, T. J. Lane, L.-P. Wang, D. Shukla, T. Tye, M. Houston, T. Stich, C. Klein, M. R. Shirts, and V. S. Pande, J. Chem. Theory Comput. **9**, 461 (2013).
[50]T. A. Darden, P. Y. Ren, D. Jiao, C. King, and A. Grossfield, J. Phys. Chem. B **110**, 18553 (2006).
[51]M. Pitoňák, P. Neogrády, J. Černý, S. Grimme, and P. Hobza, ChemPhysChem **10**, 282 (2009).
[52]X. Li, S. Y. Ponomarev, D. L. Sigalovsky, J. P. Cvitkovic, and G. A. Kaminski, J. Chem. Theory Comput. **10**, 4896 (2014).
[53]L. Lagardère, F. Lipparini, É. Polack, B. Stamm, E. Cancès, M. Schnieders, P. Y. Ren, Y. Maday, and J.-P. Piquemal, J. Chem. Theory Comput. **11**, 2589 (2015).