Pseudobond parameters for QM/MM studies involving nucleosides, nucleotides, and their analogs

Robin Chaudret,¹ Jerry M. Parks,² and Weitao Yang^{1,3,a)}

¹*Department of Chemistry, Duke University, 124 Science Drive, 5301 French Science Center, Durham, North Carolina 27708-0346, USA*

²*UT/ORNL Center for Molecular Biophysics, Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*

³*Department of Physics, Faculty of Science, King Abdulaziz University, P.O. Box 80203, Jeddah 21589, Saudi Arabia*

(Received 20 September 2012; accepted 30 November 2012; published online 24 January 2013)

In biological systems involving nucleosides, nucleotides, or their respective analogs, the ribose sugar moiety is the most common reaction site, for example, during DNA replication and repair. However, nucleic bases, which comprise a sizable portion of nucleotide molecules, are usually unreactive during such processes. In quantum mechanical/molecular simulations of nucleic acid reactivity, it may therefore be advantageous to describe specific ribosyl or ribosyl phosphate groups quantum mechanically and their respective nucleic bases with a molecular mechanics potential function. Here, we have extended the pseudobond approach to enable quantum mechanical/molecular mechanical simulations involving nucleotides, nucleosides, and their analogs in which the interface between the two subsystems is located between the sugar and the base, namely, the $C(sp^3)-N(sp^2)$ bond. The pseudobond parameters were optimized on a training set of 10 molecules representing several nucleotide and nucleoside bases and analogs, and they were then tested on a larger test set of 20 diverse molecules. Particular emphasis was placed on providing accurate geometries and electrostatic properties, including electrostatic potential, natural bond orbital (NBO) and atoms in molecules (AIM) charges and AIM first moments. We also tested the optimized parameters on five nucleotide and nucleoside analogues of pharmaceutical relevance and a small polypeptide (triglycine). Accuracy was maintained for these systems, which highlights the generality and transferability of the pseudobond approach. © 2013 American Institute of Physics. [<http://dx.doi.org/10.1063/1.4772182>]

I. INTRODUCTION

In complex reactive systems, the influence of the macromolecular environment on chemical reactions is important. These effects are particularly relevant for biochemical systems, in which amino acid residues not only interact with ligands to facilitate specific binding but also participate directly in chemical reactions as catalyst. The computational cost of describing a complete biomolecular system quantum mechanically is, in general, prohibitive. Classical molecular mechanics approaches are then needed to model the biomolecular environment. The quantum mechanical/molecular mechanical (QM/MM) approach¹ enables chemical reactions to be modeled accurately while also including the influence of the macromolecular environment and solvent in simulations of complex molecular systems. The application of QM/MM methods to studies of biological systems has undergone sustained growth over the past two or more decades.^{2–7} Although the approach has achieved great success in numerous applications,^{8,9} several problems remain to be addressed. One such problem involves the interface between the QM and MM subsystems. Proper description of the junction between the two subsystems has been widely studied,^{8–13} and several approaches

have been developed to resolve it. The main methods used most often in QM/MM simulations are the link atom,^{14–16} frozen orbital,^{17–25} pseudobond,^{26–28} and quantum capping²⁹ approaches, each having its own advantages and disadvantages.

The pseudobond method was developed to enable a straightforward and accurate interface between QM and MM subsystems without introducing additional degrees of freedom.²⁶ Here, we provide a brief summary of the pseudobond method. A fluorine atom with a single free valence electron is introduced into the QM subsystem to replace the first MM atom at the QM/MM boundary and complete the valence of the QM subsystem. This atom is described using a special basis set and effective core potential (ECP), both of which are parameterized to mimic the original bond across the boundary while contributing minimal perturbations to the electronic structure of the QM subsystem. In the initial pseudobond formulation,²⁶ a 6-31G(d) basis set was used for all atoms (including the fluorine), and only $C_{ps}(sp^3)-C(sp^3)$ bonds were considered, enabling subsystem partitioning across aliphatic C–C bonds in, for example, the side chains of proteins.^{26,27} The testing set used in that work to assess parameter accuracy included a set of eight simple functionalized alkanes (i.e., ethanol, ethylamine, ethylthiol, and propionic acid in neutral and charged protonation states), and the molecular properties included various bond distances,

^{a)} Author to whom correspondence should be addressed. Electronic mail: Weitao.Yang@duke.edu.

angles, bond dissociation energies, and Mulliken charges.³⁰ In subsequent work, $C_{ps}(sp^3)-C(sp^2)$ and $C_{ps}(sp^3)-N(sp^2)$ pseudobonds were also developed, which enabled QM/MM boundaries to be located along the protein backbone.²⁷ A smaller STO-2G basis and angular-momentum-independent ECP were used for increased efficiency. Later, $C_{ps}(sp^3)-C(sp^3)$, $C_{ps}(sp^3)-C(sp^2)$, and $C_{ps}(sp^3)-N(sp^2)$ pseudobond parameters were developed which provided an improved description of electrostatics and deprotonation energies, while maintaining accurate geometries.²⁸ In that work, the criteria used during parameter optimization and testing included accurate geometries (i.e., bonds and angles), bond dissociation energies of the parameterized bond, deprotonation energies for ionizable side chains, and electrostatic potential (ESP) fitted, rather than Mulliken, charges, the latter of which are known to suffer from significant basis set dependence effects.

Previous pseudobond development was primarily oriented toward studying enzymes. It is therefore desirable to extend the capabilities to include nucleosides, nucleotides, and their analogs for simulations of DNA polymerization or cleavage processes. Nucleosides can be divided into two distinct subunits: the nucleobase and the ribosyl or deoxyribosyl group. Likewise, nucleotides are composed of a nucleobase and a (deoxy)ribosyl phosphate. In chemical processes involving nucleosides or nucleotides, reactivity occurs mainly at the level of the sugar ring, whereas the bases are typically unreactive and only interact with the reactive center through non-covalent interactions. Partitioning a nucleotide base into a reactive (i.e., QM) sugar group and an unreactive (MM) base is desirable because it reduces the computational cost significantly relative to describing an entire nucleotide or nucleoside quantum mechanically.

Here, we develop pseudobond parameters specifically for QM/MM simulations of nucleic acid-containing systems. These parameters enable partitioning across the $N(sp^2)-C(sp^3)$ bond between the sugar moiety and the nucleobase, in which the sugar is contained in the QM subsystem and the base is included in the MM subsystem. In Secs. II–V, we first provide a brief summary of the pseudobond method. We then describe the approach used to carry out the pseudobond parameter optimizations and the training and testing sets used. Next, we assess the accuracy of various molecular properties using the resulting parameters. We then analyze the influence of the basis set of the QM subsystem on the accuracy of pseudobond-containing systems. Finally, we test the generality of our optimized parameters on a set of molecules that contain chemically diverse structures.

II. THEORY

As in previous pseudobond formulations,^{27,28} the atom described by the pseudopotential and STO-2G basis set is a fluorine atom, which has nine electrons. Two electrons are described by the ECP and the other seven are valence electrons, leaving one free valence for bonding to the rest of the QM subsystem. The nuclear charge is that of fluorine. The form of

the STO-2G basis set is

$$\Phi_s = g_s(\alpha_1, R) + d_1 g_s(\alpha_2, R),$$

$$\Phi_p = g_p(\alpha_1, R) + d_2 g_p(\alpha_2, R),$$

where g_s and g_p are normalized s - and p -type Gaussian functions, α_1 and α_2 are exponents, and d_1 and d_2 are coefficients. The functional form of the ECP is

$$V_{\text{eff}}(r) = a \exp(-br^2)/r,$$

where a and b are the coefficient and exponent, respectively. Thus, there are six adjustable parameters that are optimized: a , b , α_1 , α_2 , d_1 , and d_2 .

III. TRAINING AND TESTING SETS

Here, we adopt the following notation to describe the molecules used in the parameter optimization. A standard molecule (StdMol) refers to a reference molecule that contains a specific covalent linkage to be reproduced using a pseudobond (Figure 1(a)). A pseudobond-containing molecule (PsMol) is the analog of a StdMol in which a portion of the molecule has been replaced by a fluorine atom consisting of the parameterized ECP and basis set. The pseudobond parameters were optimized using a *training set* consisting of pairs of molecules (StdMol, PsMol). In this work, three different training sets were used:

- The *purine training set* included models of adenosine, deoxyadenosine, guanosine, and deoxyguanosine in which the CH_2OH group on C4' was replaced by CH_3 (Figure 1(b)).
- The *pyrimidine training set* included models of thymidine, deoxythymidine, cytidine, deoxycytidine, uridine, and deoxyuridine in which the $-\text{CH}_2\text{OH}$ group on C4' was replaced by CH_3 (Figure 1(b)).

FIG. 1. (a) Example of a truncated nucleoside (deoxycytidine) analog used in the training set during pseudobond parameter optimization. Atoms in the QM subsystem are shown in black, the N_{ps} boundary atom in blue, and MM atoms in red. (b) Representation of the different training and test sets used in this work.

- The *complete training set* included all molecules from both the purine and pyrimidine sets.

In the following study, five different sets of parameters are discussed. The Pur1 and Pur2 (or Pyr1 and Pyr2) sets were optimized using only the purine (or pyrimidine) training set, and were optimized using two different sets of initial parameters. Finally, the Total set was optimized using the Total training set, which is simply the union of the purine and pyrimidine training sets. Only one initial guess set of parameters was used in that case.

After parameter optimization, the performance of the resulting parameters was tested using a larger test set of molecules and properties. In addition to the molecules present in the total training set, the testing set also included the monophosphorylated (phosphate dianion) counterparts of each nucleoside (i.e., nucleotides) (Figure 1(b)). Additional properties included in the assessment were geometries (bond lengths, angles, and dihedrals), ESP and NBO^{31,32} charges, AIM^{33,34} charges and AIM first moments³⁵ (M1). The AIM M1 is the first electrostatic moment of a molecular space (an atom in AIM) and can be compared to the dipole of an atom in a molecule. Unlike the parameter optimizations, for which the geometry of each PsMol was kept fixed at its corresponding StdMol geometry, full geometry optimizations were carried out for all molecules in the testing set. As in previous studies,^{27,28} pure QM optimizations were performed instead of full QM/MM optimizations.

The various properties for each PsMol were compared to the corresponding StdMols. All atoms from the QM subsystem were included in the comparisons except for the phosphate group and the C5' hydrogens (for charges and dipoles) and also any bond, angle, and dihedral that included those atoms. The phosphate group was not included in the assessments because it was quite mobile and therefore significant artificial variations occurred. As in the training set, three different test sets were used to investigate the influence of the training set on the accuracy of the pseudobond, and purine, pyrimidine, and total *testing* sets were defined. The purine testing set contained all molecules from the purine training set and their monophosphate equivalents. Likewise, the pyrimidine testing set included the pyrimidine training set molecules and their monophosphate equivalents. Finally, the total testing set combined the purine and pyrimidine testing sets.

IV. PARAMETER OPTIMIZATION

The various sets of pseudobond parameters were optimized by minimizing an objective error function, E_{total} , using the gradient-free Powell optimization procedure.³⁶ The objective function consisted of a weighted sum of error terms:

$$E_{\text{total}} = \sum_{\text{promp } i} \omega_i E_i(\lambda), \quad \lambda = (\alpha, b, \alpha_1, \alpha_2, d_1, d_2),$$

where $E_i(\lambda)$ represents the error for property i computed with the parameter λ and ω_i is the weight assigned to that property.

The weight factors were taken from our previous work,²⁸ and were chosen such that each term was treated approximately equally by taking into account the relative magnitudes of each energy or property. However, geometries and ESP charges were given greater weights than bond dissociation energies to reflect their relative importance. The error for a given property i for a standard reference molecule (StdMol) and for the same property computed for the truncated molecule containing a pseudobond (PsMol). The three different properties of interest in this optimization are:

- The **gradient norm error** was calculated as the norm of a single vector containing all the individual geometric gradients of the pseudobond-containing molecules fixed at their fully optimized StdMol geometries.
- The **ESP charge error** was calculated as the rms deviation in ESP charges between all non-pseudobond atoms in PsMol and its corresponding StdMol.
- The **bond dissociation energy error** was calculated as the rms deviation between the bond dissociation energies of the parameterized bonds (i.e., $N_{\text{ps}}(sp^2)-C(sp^3)$ bonds) of molecules containing a pseudobond (PsMol) and their corresponding StdMols.

In all cases, a relative weight of 1.0 was used for the gradient norm term. A weight factor of 0.46 was used for ESP charges, and bond dissociation energies were given a weight of 0.0006. No further manual adjustment of weights was performed. The units of each weight are the inverse of the corresponding property. All calculations were performed at the B3LYP^{37,38}/6-31G*³⁹ level of theory using GAUSSIAN09⁴⁰ unless otherwise noted.

V. RESULTS AND DISCUSSION

Five sets of optimized pseudobond parameters are listed in Table I. Each set was obtained by optimizing a different initial guess set of parameters against a particular training set. The initial guesses came from previously optimized parameters or pseudorandom numbers within a specified numerical range. In general, the different parameter sets provide similar accuracies (see Table II) suggesting that, for a given bond, the nature of the molecule used in the training set has only a minor influence on the accuracy of the parameters, provided that somewhat similar and chemically reasonable training sets are used. For example, the Pur2 parameter set yields errors comparable to the Pyr2 set when assessed using the Total test set.

TABLE I. Five sets of optimized pseudobond parameters from the total test set (Tot), the purine test sets (Pur1 and Pur2), and the pyrimidine test sets (Pyr1 and Pyr2).

Set	a	b	α_1	α_2	d_1	d_2
Tot	24.92677	18.70758	1.82997	0.28389	3.95334	1.19779
Pur1	24.88664	17.70140	1.87873	0.28288	4.53586	1.27451
Pur2	33.87962	25.08505	2.13980	0.30586	6.73700	1.33034
Pyr1	33.29602	26.98681	2.04912	0.31768	5.00416	1.24585
Pyr2	22.92485	15.27730	1.68929	0.27526	4.18348	1.20014

Thus, the various pseudobond parameter sets exhibit some degree of transferability and are therefore able to model the desired bond in different molecular systems. In the remainder of this work, only the Tot parameters optimized on the Total training set are discussed further as they were found to be slightly more accurate than the other sets. Bond dissociation errors computed for the total training set molecules are provided in the supplementary material (Table S2).⁴² All errors are less than 5% of the total bond dissociation energy.

A. Geometric properties

The bond and angle deviations for the PsMol set relative to the StdMol equivalents are shown in Figure 2 and Table II. The rmse is 0.01 Å for the bonds and 1.1° for the angles, and the maximum signed errors are +0.06 Å (O3'–H3' bond of deoxycytidine monophosphate), −0.09 (O3'–H3' bond of thymidine monophosphate) for the bonds and +5.1° (C3'–C4'–C5' angle of deoxyadenosine) and −5.0° (C3'–C4'–H4' angle of deoxyadenosine) for the angles. The rmse values are only slightly larger than those for proteins in the previous study.²⁸ The presence of hydrogen bonds of varying strength (depending on the presence or absence of the pseudobond) between a phosphoryl oxygen and H3' of the sugar ring is responsible for the majority of the deviations for the bonds

FIG. 2. Correlations of (a) bond distances (in Å) and (b) angles (in °) between pseudobond-containing molecules (PsMol) and corresponding standard molecules (StdMol) for the total test set using the Tot parameters.

TABLE II. RMS error for ESP, NBO, and AIM charges, AIM M1, bond distances, and bond angles for the different optimized parameters. The tables summarize the rmse values obtained for the purine (a), pyrimidine (b), and total (c) test sets.

a					
RMSE	Tot	Pur1	Pur2	Pyr1	Pyr2
ESP charge (a.u.)	0.070	0.075	0.067	0.067	0.069
NBO charge (a.u.)	0.026	0.031	0.024	0.021	0.033
AIM charge (a.u.)	0.081	0.084	0.087	0.071	0.075
AIM M1 (a.u.)	0.042	0.043	0.042	0.040	0.041
Distances (Å)	0.007	0.010	0.007	0.008	0.007
Angles (°)	0.92	1.59	0.92	1.24	0.92
b					
RMSE	Tot	Pur1	Pur2	Pyr1	Pyr2
ESP charge (a.u.)	0.071	0.074	0.074	0.071	0.071
NBO charge (a.u.)	0.021	0.024	0.020	0.016	0.028
AIM charge (a.u.)	0.067	0.072	0.066	0.058	0.068
AIM M1 (a.u.)	0.035	0.037	0.038	0.035	0.037
Distances (Å)	0.017	0.018	0.018	0.017	0.017
Angles (°)	1.19	1.44	1.26	1.40	1.39
c					
RMSE	Tot	Pur1	Pur2	Pyr1	Pyr2
ESP charge (a.u.)	0.071	0.074	0.071	0.070	0.071
NBO charge (a.u.)	0.023	0.027	0.022	0.018	0.030
AIM charge (a.u.)	0.079	0.077	0.080	0.070	0.077
AIM M1 (a.u.)	0.038	0.040	0.040	0.037	0.039
Distances (Å)	0.014	0.015	0.015	0.014	0.014
Angles (°)	1.09	1.50	1.13	1.34	1.22

and angles. This hydrogen bond is artificial and would not be present in a real QM/MM simulation where steric effects from the MM environment would preclude such an interaction. Figure 2 shows that the deviations for the geometric parameters are in general minor, although they are more significant for the angles than for the bonds. Deviations in dihedral angles tend to be more significant than for angles, mainly because there exists the possibility of phosphoryl group rotation in the nucleotides that is absent in nucleosides. Further analysis of dihedrals is provided in the supplementary material.⁴²

B. Atomic partial charges

In this study, three different atomic partial charge schemes have been studied as well as, for the first time in pseudobond studies, the AIM atomic M1, which is associated with the atomic dipoles (Table II). The ESP charge rmse value is 0.07 a.u. and the maximum deviations are +0.35 a.u. (C4' atom of adenosine monophosphate) and −0.32 a.u. (C2' of adenosine monophosphate), which are of the same order of magnitude as errors in previous studies. This deviation is similar to the AIM results (rmse = 0.07 a.u. and maximum deviations = +0.19 a.u. and −0.32 a.u., see also Figure 3(c)) but is worse than the NBO one (rmse = 0.02 and maximum deviations = +0.07 a.u. and −0.11 a.u., see also Figure 3(b)). These results show that the pseudobond method is not only able to reproduce ESP charges, which are important for obtaining accurate QM/MM electrostatic energies, but

TABLE III. RMS error for charges, first moments M1 (a.u.), bond distances ($\text{\AA}$), and angles (degrees) and the dihedrals for adenosine monophosphate computed using various basis sets.

RMSE	ESP charges	NBO charges	AIM charges	AIM M1	Bond	Angle	Dihedral
LANL2DZ	0.057	0.017	0.056	0.016	0.017	1.71	31.0
3-21G	0.034	0.027	0.066	0.014	0.014	1.25	13.1
6-31G(d)	0.020	0.022	0.074	0.040	0.005	1.69	2.3
6-31+G(d)	0.055	0.026	0.069	0.026	0.013	1.55	25.9
6-31++G(d,p)	0.055	0.024	0.088	0.018	0.013	1.57	26.2
6-311++G(d,p)	0.056	0.030	0.033	0.022	0.010	2.40	27.7

also AIM and NBO charges, which are commonly used to analyze atomic charges in quantum chemical studies.

The AIM M1 values are also well reproduced in the pseudobond systems (rmse = 0.04 a.u., positive and negative maximum deviations +0.18 a.u. and -0.18 a.u.; see also Figure 3(d)), which indicates that the polarization properties of the molecules with or without pseudobonds are similar, and therefore the polarization of the QM subsystem will be similar at the pure QM or QM/MM level. Future applications of polarizable QM/MM schemes will likely benefit from such properties.

C. Additional tests

Three additional tests were performed to study the generality of the pseudobond approach for DNA-like systems. First, the basis set dependence for the accuracy of the pseudobond parameters was investigated. Using the optimized Tot

parameter set, the basis set for the rest of the QM subsystem was varied and LANL2DZ, 3-21G, 6-31G, 6-31+G(d), 6-31++G(d,p), 6-311++G(d,p) basis sets were selected. They were chosen to be representative of the “typical” basis sets used in QM/MM simulations in order to give an idea of the expected accuracy of the pseudobonds in biomolecular simulations. As a second test, we assessed the transferability of the optimized parameters to chemically diverse, pharmaceutically relevant systems. Four drug molecules, each containing an $N(sp^2)-C(sp^3)$ bond, were chosen. The chemical structures of acyclovir, clofarabine, emtricitabine, and tenofovir are shown in Figure 2. Clofarabine (Figure 4(b)) is a purine nucleoside antimetabolite used in leukemia treatment. Emtricitabine (Figure 4(c)) and tenofovir (Figure 4(d)) are nucleoside reverse transcriptase inhibitors that are used to treat HIV. The electronic structures of these compounds are significantly different from standard nucleosides and nucleotides. Acyclovir (Figure 4(a)), or acycloguanosine, is a guanosine analogue that is used to treat viral infections. As an additional

FIG. 3. Correlation between pseudobond containing molecules (PsMol) and standard molecules (StdMol) for (a) ESP charges, (b) NBO charges, (c) AIM charges, and (d) AIM M1 computed using the Total testing set.

FIG. 4. Representations of the molecules used for pseudobond testing: (a) acyclovir, (b) clofarabine, (c) emtricitabine, (d) tenofovir, and (e) GLY₃ (triglycine) polypeptide. Atoms in the QM subsystem are shown in black, the N_{ps} boundary atom in blue, and MM atoms in red.

test, we also included a polypeptide, triglycine (Figure 4(e)), to show that the pseudobonds can be used in a new way to cut along the peptide backbone in protein systems.

1. Basis set effects

During the parameter optimizations, all calculations were performed at the B3LYP/6-31G(d) level of theory. However, it is important to assess the performance of the resulting parameters when other basis sets are used to describe the QM subsystem. A total of six different basis sets were tested, and the results are summarized in Table III. As expected, the 6-31G(d) basis set gives, on average, better results than the others, although other basis sets also provide reasonable results. The most pronounced differences are evident in the dihedral terms. The dihedral rmse values can be explained by the significant differences between the PsMols and StdMols in the ring conformations that occur for all basis sets except 6-31G(d). It is also interesting to note that larger or smaller basis sets yield practically the same levels of accuracy relative to their respective StdMol reference molecules. Indeed, the pseudobonds were optimized for use with the 6-31G(d) basis set. Therefore, although an oversimplification, the trend is that the more different the basis set is from 6-31G(d), the larger the deviation. However, on average, the error remains of the same order of magnitude as for the training basis set, 6-31G(d), demonstrating that the pseudobond approach can be used with any basis set as long as the accuracy for important properties, e.g., dihedral angles, is carefully checked.

The dependence of the accuracy of the pseudobond method on the particular quantum chemical method was not studied here. However, it has been shown previously^{26,28} that

TABLE IV. RMS errors for partial charges (a.u.), M1 (a.u.), bond distances (Å), and angles (degrees) for acyclovir, clofarabine, emtricitabine, tenofovir, and triglycine polypeptide.

RMSE	ESP	NBO	AIM charge	AIM M1	Bond	Angle
Acyclovir	0.044	0.026	0.090	0.035	0.004	1.15
Clofarabine	0.038	0.020	0.063	0.050	0.004	0.44
Emtricitabine	0.080	0.023	0.129	0.029	0.006	0.73
Tenofovir	0.064	0.013	0.104	0.081	0.004	0.51
Triglycine	0.023	0.018	0.006	0.016	0.003	0.58

the accuracy was quite insensitive to the particular method used, including HF, MP2, and various DFT methods.

2. Generality of the pseudobond parameters

Finally, to assess the generality of the parameterized pseudobonds, four diverse molecules of pharmaceutical relevance and one polypeptide were selected for further testing. These molecules were chosen to have QM (and/or MM) subsystems that are quite different from standard nucleobases. The results are summarized in Table IV and S1.

Overall, the accuracy for the different molecules is, on average, quite reasonable. Acyclovir presents the largest deviations in the angles, with a rms deviation of slightly more than 1°. Indeed, during the QM optimization of the full molecule (StdMol) a hydrogen bond is formed between the terminal hydroxyethoxy group in the QM subsystem and N7 of the guanosyl group. This hydrogen bond induces the formation of a 9-membered ring and therefore constrains the conformation slightly. N7 is not included into the QM subsystem in the PsMol. Therefore, the QM geometry optimization of the PsMol destroys the ring structure and allows the QM subsystem to relax completely. Interestingly, the pseudobond also showed good accuracy, using any of the parameter sets, for the polypeptide triglycine. These results show that our new pseudobond parameters can also be used to cut along the peptide backbone in protein systems. The present pseudobond parameters can therefore be considered as general to any $C(sp^3)$ – $N(sp^2)$ bond treated with any method and basis set. The pseudobond parameters provided here have been optimized to reproduce ground state properties. It is, however, possible to generalize the approach, in the same spirit as the work of Slavíček and Martínez,⁴¹ by introducing properties computed for excited states or other properties.

VI. CONCLUSIONS

Here, we have developed pseudobond parameters for QM/MM simulations involving nucleosides, nucleotides, and their analogs such that the boundary is located across the bond between the ribosyl group and the nucleobase. Because reactions of biological interest commonly involve the (deoxy)ribose sugar (e.g., DNA replication, repair, phosphoryl transfer) the nucleobase can be described using molecular mechanics, significantly improving the efficiency of the simulation. Using a previously described protocol, we have carried out pseudobond parameter optimizations with three different

training sets and tested their performance on seven molecular properties (bond distances, angles, dihedrals, ESP, NBO and AIM charges, and AIM first moment (M1)) using a test set of 20 molecules. No large differences in accuracy were observed when the different optimized pseudobonds sets were compared, with each training set yielding essentially equal performance. The size of the training set is not, therefore, crucial for pseudobond generality.

Pseudobonds are able to provide not only accurate ESP charges, which are useful for electrostatic potential fitting, but also AIM and NBO charges, which may provide more physically and chemically meaningful quantities. Perhaps more importantly, the approach also reproduces the AIM first moment M1, which is related to the atomic dipole and then shows the generality of the approach toward polarizable simulations. This approach is therefore a valuable tool for linking QM and MM subsystems in QM/MM simulations, including future applications involving polarizable force fields.

The pseudobond parameters developed here should enable accurate and efficient QM/MM calculations to be carried out on biomolecular systems containing DNA-like molecules. They are also general enough to be used at different levels of theory, with different basis sets and for any system containing a covalent $C(sp^3)$ – $N(sp^2)$ bond at the QM/MM interface (from sugar/base to protein backbone junctions). We have also shown that the pseudobond approach is robust enough to preserve dipolar properties and may therefore be applicable in possible future polarizable QM/MM schemes.

The optimization program, scripts, and instructions for use are available upon request from the authors.

ACKNOWLEDGMENTS

Support from the National Institute of Health (NIH R01-GM061870) is greatly appreciated.

¹A. Warshel and M. Levitt, *J. Mol. Biol.* **103**, 227–249 (1976).

²P. Wu, G. A. Cisneros, H. Hu, R. Chaudret, X. Q. Hu, and W. T. Yang, *J. Phys. Chem. B* **116**, 6889–6897 (2012).

³M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar, *Science* **303**, 186–195 (2004).

⁴M. J. Field, *J. Comput. Chem.* **23**, 48–58 (2002).

⁵A. Shurki and A. Warshel, *Protein Simulations* (Academic, San Diego, 2003), pp. 249–313.

⁶A. J. Mulholland, P. D. Lyne, and M. Karplus, *J. Am. Chem. Soc.* **122**, 534–535 (2000).

⁷H. Hu and W. T. Yang, *Annu. Rev. Phys. Chem.* **59**, 573–601 (2008).

⁸H. M. Senn and W. Thiel, *Angew. Chem., Int. Ed.* **48**, 1198–1229 (2009).

⁹H. Hu and W. T. Yang, *J. Mol. Struct. (THEOCHEM)* **898**, 17–30 (2009).

¹⁰H. Lin and D. G. Truhlar, *Theor. Chem. Acc.* **117**, 185–199 (2007).

¹¹Y. Hagiwara and M. Tateno, *J. Phys.-Condens. Matter* **22**, 413101 (2010).

¹²K. E. Ranaghan and A. J. Mulholland, *Int. Rev. Phys. Chem.* **29**, 65–133 (2010).

¹³O. Acevedo and W. L. Jorgensen, *Acc. Chem. Res.* **43**, 142–151 (2010).

¹⁴M. J. Field, P. A. Bash, and M. Karplus, *J. Comput. Chem.* **11**, 700–733 (1990).

¹⁵U. C. Singh and P. A. Kollman, *J. Comput. Chem.* **7**, 718–730 (1986).

¹⁶P. Amara and M. J. Field, *Theor. Chem. Acc.* **109**, 43–52 (2003).

¹⁷V. Thery, D. Rinaldi, J. L. Rivail, B. Maigret, and G. G. Ferenczy, *J. Comput. Chem.* **15**, 269–282 (1994).

¹⁸G. Monard, M. Loos, V. Thery, K. Baka, and J. L. Rivail, *Int. J. Quantum Chem.* **58**, 153–159 (1996).

¹⁹X. Assfeld and J. L. Rivail, *Chem. Phys. Lett.* **263**, 100–106 (1996).

²⁰N. Ferre, X. Assfeld, and J. L. Rivail, *J. Comput. Chem.* **23**, 610–624 (2002).

²¹J. L. Gao, P. Amara, C. Alhambra, and M. J. Field, *J. Phys. Chem. A* **102**, 4714–4721 (1998).

²²P. Amara, M. J. Field, C. Alhambra, and J. L. Gao, *Theor. Chem. Acc.* **104**, 336–343 (2000).

²³M. Garcia-Viloca and J. L. Gao, *Theor. Chem. Acc.* **111**, 280–286 (2004).

²⁴J. Z. Pu, J. L. Gao, and D. G. Truhlar, *J. Phys. Chem. A* **108**, 632–650 (2004).

²⁵J. Z. Pu, J. L. Gao, and D. G. Truhlar, *ChemPhysChem* **6**, 1853–1865 (2005).

²⁶Y. Zhang, T.-S. Lee, and W. Yang, *J. Chem. Phys.* **110**, 46–54 (1999).

²⁷Y. K. Zhang, *J. Chem. Phys.* **122**, 024114 (2005).

²⁸J. M. Parks, H. Hu, A. J. Cohen, and W. T. Yang, *J. Chem. Phys.* **129**, 154106 (2008).

²⁹G. A. DiLabio, M. M. Hurley, and P. A. Christiansen, *J. Chem. Phys.* **116**, 9578 (2002).

³⁰R. S. Mulliken, *J. Chem. Phys.* **23**, 1833 (1955).

³¹A. E. Reed and F. Weinhold, *J. Chem. Phys.* **78**, 4066 (1983).

³²A. E. Reed, R. B. Weinstock, and F. Weinhold, *J. Chem. Phys.* **83**, 735 (1985).

³³R. F. W. Bader, *Atoms in Molecules: A Quantum Theory* (Oxford University Press, Oxford, 1990).

³⁴C. F. Matta and R. J. Boyd, *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design* (Wiley-VCH, Weinheim, 2007).

³⁵J. Pilme and J. P. Piquemal, *J. Comput. Chem.* **29**, 1440–1449 (2008).

³⁶W. H. Press, S. Teukovsky, W. Vetterling, and B. Flannery, *Numerical Recipes in Fortran 77. The Art of Scientific Computing*, Numerical Recipes, 2nd ed. (Cambridge University Press, 1992).

³⁷A. D. Becke, *J. Chem. Phys.* **98**, 5648–5652 (1993).

³⁸C. Lee, Y. Yang, and R. G. Parr, *Phys. Rev. B* **37**, 785–789 (1988).

³⁹P. Hariharas and J. A. Pople, *Theor. Chim. Acta* **28**, 213–222 (1973).

⁴⁰M. J. Frisch, G. W. Trucks, H. B. Schlegel *et al.*, GAUSSIAN 09, Gaussian, Inc., Wallingford, CT, 2009.

⁴¹P. Slavicek and T. J. Martinez, *J. Chem. Phys.* **124**, 084107 (2006).

⁴²See supplementary material at <http://dx.doi.org/10.1063/1.4772182> for a discussion of the accuracy of pseudobonds regarding dihedral angles and the bond dissociation errors.