

# Extension of the Self-Consistent-Charge Density-Functional Tight-Binding Method: Third-Order Expansion of the Density Functional Theory Total Energy and Introduction of a Modified Effective Coulomb Interaction

Yang Yang,<sup>†</sup> Haibo Yu,<sup>†</sup> Darrin York,<sup>‡</sup> Qiang Cui,<sup>†</sup> and Marcus Elstner<sup>\*,§</sup>

Department of Chemistry and Theoretical Chemistry Institute, University of Wisconsin—Madison, 1101 University Avenue, Madison, Wisconsin 53706, Department of Chemistry, University of Minnesota, 207 Pleasant Street Southeast, Minneapolis, Minnesota 55455, and Department of Physical and Theoretical Chemistry, Technische Universität Braunschweig, Hans-Sommer-Strasse 10, D-38106 Braunschweig, Germany

Received: May 30, 2007; In Final Form: July 24, 2007

The standard self-consistent-charge density-functional-tight-binding (SCC-DFTB) method (*Phys. Rev. B* **1998**, *58*, 7260) is derived by a second-order expansion of the density functional theory total energy expression, followed by an approximation of the charge density fluctuations by charge monopoles and an effective damped Coulomb interaction between the atomic net charges. The central assumptions behind this effective charge–charge interaction are the inverse relation of atomic size and chemical hardness and the use of a fixed chemical hardness parameter independent of the atomic charge state. While these approximations seem to be unproblematic for many covalently bound systems, they are quantitatively insufficient for hydrogen-bonding interactions and (anionic) molecules with localized net charges. Here, we present an extension of the SCC-DFTB method to incorporate third-order terms in the charge density fluctuations, leading to chemical hardness parameters that are dependent on the atomic charge state and a modification of the Coulomb scaling to improve the electrostatic treatment within the second-order terms. These modifications lead to a significant improvement in the description of hydrogen-bonding interactions and proton affinities of biologically relevant molecules.

## I. Introduction

The self-consistent-charge density-functional-tight-binding (SCC-DFTB) method<sup>1</sup> is an approximation to density functional theory (DFT), derived from a second-order expansion of the DFT total energy expression. In recent years, SCC-DFTB has been successfully applied to a wide range of problems involving structures and dynamics of biomolecules and biocatalysis in several enzymes; for comprehensive reviews see, for example, refs 2–6.

With respect to its computational efficiency, SCC-DFTB is comparable to the widely used semiempirical methods such as AM1 and PM3, i.e., being 2–3 orders of magnitude faster than DFT(HF) methods (with small to medium-sized basis sets). This increase in speed with respect to DFT is achieved without much loss of accuracy in the description of molecular geometries, while reaction energies and vibrational frequencies are usually less reliable.<sup>1,7</sup> This is confirmed by two recent thorough studies that evaluated SCC-DFTB for heats of formation, molecular structures, etc. on large sets of molecules.<sup>8,9</sup> While the most sophisticated neglect of diatomic differential overlap (NDDO) methods<sup>10</sup> are slightly superior to SCC-DFTB for heats of formation, the strength of SCC-DFTB is the overall excellent prediction of molecular structures, in particular for larger (bio)-molecular systems, where NDDO-type methods may have some limitations.<sup>11–13</sup>

There has been a resurgence of interest in developing improved fast quantum models such as SCC-DFTB and NDDO-based semiempirical methods for use in linear-scaling electronic structure<sup>14,15</sup> and combined quantum mechanical/molecular mechanical (QM/MM) simulations.<sup>6,16–22</sup> Some of the more recent efforts include the inclusion of orthogonalization corrections in the OMx model,<sup>23</sup> PDDG/PM3 model,<sup>24</sup> the PM3-MAIS and PM3-PIF models,<sup>25,26</sup> and the NO-MNDO model.<sup>27</sup> Other notable improvements include the PM3<sub>BP</sub> model<sup>28</sup> for accurate nucleic acid base-pairing interactions, the AM1/d-PhoT model<sup>29</sup> for phosphoryl transfer reactions based on a database of quantum calculations for RNA catalysis,<sup>30</sup> and a new method that greatly improves the modeling of charge-dependent response properties.<sup>31</sup>

Consequently, it is of considerable importance to identify systematic strengths and weaknesses of the different models to derive new functional forms and parametrizations that considerably advance the field. In the case of the SCC-DFTB method, several limitations have been identified over the years. The main problems of SCC-DFTB are twofold. First, because SCC-DFTB is implemented based on popular generalized gradient approximation (GGA) functionals, it inherits the problems associated with these approximate functionals. Examples include inaccuracies and failures in the description of electronically excited states involving long-range charge separation or dispersion interactions.<sup>4</sup> While there is no simple cure for the treatment of electronically excited states, an empirical correction for dispersion interactions has been proposed<sup>12</sup> and has been adopted for DFT calculations later on.<sup>32,33</sup> This correction was found to be crucial for predicting reliable nucleic acid base-stacking interactions<sup>12</sup> and polypeptide (protein) structures, as, for

\* Author to whom correspondence should be addressed. E-mail: m.elstner@tu-bs.de.

<sup>†</sup> University of Wisconsin—Madison.

<sup>‡</sup> University of Minnesota.

<sup>§</sup> Technische Universität Braunschweig.

example, exemplified in the relative stabilities of  $\alpha$ - and  $3_{10}$ -helices in proteins<sup>34</sup> (for a more detailed discussion, see refs 3 and 4).

However, it has also been recognized that the current SCC-DFTB model is not flexible enough to account for various chemical environments. In this work, we focus on the description of hydrogen-bonding interactions and proton affinities because these properties are of ultimate importance in the context of biological applications. Hydrogen-bonding interactions play a major role in maintaining the structural integrity and association of biomolecules as well as determining the specificity of substrate selection and chemical modification in enzymes.<sup>35</sup> Proton affinity is crucial because changes in the protonation state are involved in many catalytic processes as well as association events.<sup>36</sup> Quantitatively describing these two types of properties, however, is far from trivial especially if approximate or semiempirical QM methods are used. In this context, we emphasize that for biological problems sufficient conformational sampling is particularly important especially if the process spans a large spatial scale, such as in long-range proton pumping.<sup>6,37</sup> Therefore, there is a compelling reason for developing approximate QM methods that are semiquantitatively accurate and allow for at least nanosecond scale sampling in a QM/MM framework.<sup>6,17–20</sup>

To describe the interaction between molecules engaged in hydrogen bonds, a generally reliable method has to be able to treat a combination of components including electrostatics, charge-transfer effects, as well as dispersion interactions.<sup>38</sup> Ab initio methods<sup>39</sup> such as MP2 have been shown to generally give a reliable description for these interactions in complexes dominated by hydrogen bonding, but they are too demanding for realistic biological applications. Density functional theories such as B3LYP<sup>40–42</sup> give reasonable hydrogen-bonding structures and energies in most cases although the lack of dispersion and other deficiencies in the exchange–correlation functionals deserve further improvements. MNDO-based methods are computationally much more efficient, but the popular models underestimate weak hydrogen-bond strengths, and many attempts have been made to improve their performance in this regard (for a review, see ref 43). A major strategy has been to modify the core–core repulsion terms leading to the AM1 and PM3 models. Voityuk and Bliznyuk have added an additional set of Gaussian functions to the core–core repulsion terms within the MNDO framework, resulting in a major improvement for hydrogen-bonded compounds.<sup>44</sup> A different strategy has been applied by Jug and Geudtner<sup>45</sup> for the SINDO1 method by extending the minimal basis to include p-orbitals at the hydrogen atoms. This should allow a better description of the electron density in the interatomic regions. For the best results, they used a small Slater exponent and applied an additional empirical function that enhances bonding in the hydrogen-bonding region and damps the effect of the p-functions in the covalent region.

Because SCC-DFTB has rather systematic errors in the calculated hydrogen-bonding interactions, i.e., binding energies of weak hydrogen bonds are slightly underestimated by typically about 1–2 kcal/mol compared to high-level ab initio results,<sup>2,46,47</sup> there have been previous attempts to improve the description of hydrogen-bonding interactions in the SCC-DFTB framework. First, the DFTB repulsive potential contribution ( $\bar{E}_{\text{rep}}$ ) was modified to introduce a stronger attraction at hydrogen-bonding distances (such as the changes of core–core repulsion in MNDO). But this strategy is purely empirical and would imply that the energy of each X–H (X being heavy atoms) pair is increased by the same amount, irrespective of

the chemical environment. This approach works well for many small hydrogen-bonded complexes but fails to account properly for cooperative effects, such as the shrinking of hydrogen-bond lengths for larger water clusters. As an alternative, we followed the SINDO1 strategy within the framework of the SCC-DFTB method by including p-orbitals. The inclusion of the very diffuse 2p atomic orbital of the hydrogen atom would, in the case of covalent bonding of the hydrogen atom to oxygen, for example, lead very much to an extension of the basis functions located at the oxygen atom. This results in an erroneous description of atomic charges when applying the Mulliken charge analysis, on which the SCC-DFTB total energy expression is based. We therefore, similar to the procedure in SINDO1, convoluted the Hamiltonian and overlap matrix elements corresponding to the hydrogen 2p-orbitals with a Gaussian function centered at the hydrogen-bonding distance. This results in an effective damping of these matrix elements in the covalent bonding region, whereas in the hydrogen-bonding region they remain unaltered. We used the atomic 2p function of hydrogen and chose a Gaussian,  $\exp[-A(R_{\text{AB}} - r_0)^2]$ , with the width  $A = 2$  and  $r_0 = 1.9 \text{ \AA}$  for the convolution. These values are empirical and chosen to make the hydrogen 2p contributions vanish in the covalent bonding region.<sup>2</sup> This approach was also successful for small hydrogen-bonded complexes; however, it is not very elegant with respect to an unbalanced inclusion of polarization and leads to a significant slowdown of the calculations.

Calculating accurate proton affinities, especially for large molecules, is also not trivial. Sophisticated quantum chemical methods, such as coupled cluster methods, are not readily applicable to large systems, although remarkable progress has been made in recent years toward making such methods linear scaling. MNDO-based semiempirical methods<sup>10</sup> are much more practical but have limited accuracy. In a recent study,<sup>48</sup> for example, we have systematically studied the accuracy of several popular semiempirical methods including AM1 and PM3, using a series of phosphate-containing compounds. It was found that AM1, PM3, and SCC-DFTB have comparably large errors on the order of 14–19 kcal/mol (root-mean-square error, RMSE) compared to experimental data. If proton affinity or  $\text{p}K_{\text{a}}$  is the only property of interest, then one may choose to perform systematic corrections based on either empirical correlation<sup>48</sup> or single-point energy calculations at higher levels. In fact, the latter approach was found to be quite effective in a recent QM/MM study of  $\text{p}K_{\text{a}}$  in solution and protein environments.<sup>49</sup> In many cases, however, exchange of proton(s) is an important part of the chemical reaction under study;<sup>36</sup> a poor description of relative proton affinities of the participating groups may cause qualitative errors in the chemical nature of intermediate state(s) and configuration sampling.

In light of all previous attempts to improve the performance of NDDO-based semiempirical methods and SCC-DFTB, it seems that ad hoc modifications and re-parametrizations of existing methods may lead to only a partial improvement for some systems, causing problems for others. Therefore, a systematic improvement requires an extension of the existing formalisms, probably for both SCC-DFTB and traditional NDDO methods.

Here, considering the promise of the SCC-DFTB approach in biophysical studies,<sup>3,6</sup> we make two physically motivated improvements to the method. First, we propose an extension of the DFTB formalism by a third-order expansion of the DFT total energy. This leads to charge dependence of the chemical hardness (Hubbard) parameter, which has a major impact on the predicted proton affinities. Second, we reexamine the

assumptions underlying the Coulomb interactions in the second-order terms; the proposed revision has a significant effect on the calculated hydrogen-bonding interactions. The small number of parameters associated with the improvements have been determined based on a set of biologically relevant molecules and are expected to be transferable. For specific applications that demand an even higher accuracy than achieved here, the current developments also offer the framework for developing specific reaction parameters.

In the following sections, we first describe the relevant theoretical developments and computational algorithms used to determine the relevant parameters; this is followed by test calculations that illustrate the improvements in the performance of the SCC-DFTB method for a series of molecules of general biological interest. Finally, we draw a few conclusions.

## II. Theory and Computational Methods

To facilitate the discussion of the new developments, we first briefly review the current formulation of the SCC-DFTB approach. Extensions of the formalism are discussed in section II.B.

**A. The SCC-DFTB Method.** The first step in the derivation of the SCC-DFTB model<sup>1</sup> is a second-order expansion of the DFT total energy functional with respect to the charge density fluctuations  $\delta\rho$  around a given reference density  $\rho_0$  ( $\rho'_0 = \rho_0(\vec{r}')$ ,  $f' = f d\vec{r}'$ )

$$E = \sum_i^{\text{occ}} \langle \Psi_i | \hat{H}^0 | \Psi_i \rangle + \frac{1}{2} \int \int' \left( \frac{1}{|\vec{r} - \vec{r}'|} + \frac{\delta^2 E_{\text{xc}}}{\delta\rho\delta\rho'} \Big|_{\rho_0} \right) \delta\rho\delta\rho' - \frac{1}{2} \int \int' \frac{\rho'_0\rho_0}{|\vec{r} - \vec{r}'|} + E_{\text{xc}}[\rho_0] - \int V_{\text{xc}}[\rho_0]\rho_0 + E_{\text{cc}} \quad (1)$$

$\rho_0$  is usually taken as the superposition of the electron densities  $\rho_0^\alpha$  of the neutral atoms  $\alpha$  constituting the molecular system of interest.  $\hat{H}^0 = \hat{H}[\rho_0]$  is the effective Kohn–Sham Hamiltonian evaluated at the reference density  $\rho_0$ , and the  $\Psi_i$ 's are Kohn–Sham orbitals.  $E_{\text{xc}}$  and  $V_{\text{xc}}$  are the exchange–correlation energy and potential, respectively, and  $E_{\text{cc}}$  is the core–core repulsion energy.

In a second step, the energy contributions in eq 1 are subjected to several approximations described below.

**1. Determination of the Hamiltonian Matrix Elements.** The Hamiltonian matrix elements  $\langle \Psi_i | \hat{H}^0 | \Psi_i \rangle$  in the first term of eq 1 are represented in a minimal basis set of confined, pseudo-atomic orbitals  $\phi_\mu$  (see refs 1 and 50 for more details)

$$\Psi_i = \sum_\mu c_\mu^i \phi_\mu \quad (2)$$

The basis functions  $\phi_\mu$  are determined by solving the atomic Kohn–Sham equations in the presence of an additional harmonic potential,<sup>50</sup> which leads to a confinement of the basis functions. The Hamiltonian matrix elements in this linear combination of atomic orbital (LCAO) basis,  $H_{\mu\nu}^0$ , are then calculated as follows. The diagonal elements  $H_{\mu\mu}^0$  are taken to be the Kohn–Sham energies of the atomic orbitals  $\phi_\mu$ , and the nondiagonal elements  $H_{\mu\nu}^0$  are calculated in a two-center approximation

$$H_{\mu\nu}^0 = \langle \phi_\mu | \hat{T} + v_{\text{eff}}[\rho_\alpha^0 + \rho_\beta^0] | \phi_\nu \rangle \quad \mu \in \alpha, \nu \in \beta \quad (3)$$

$H_{\mu\nu}^0$  and the overlap matrix elements  $S_{\mu\nu} = \langle \phi_\mu | \phi_\nu \rangle$  are tabulated as a function of the interatomic distance  $R_{\alpha\beta}$ .  $v_{\text{eff}}$  is the effective Kohn–Sham potential according to the superposition of the densities of neutral atoms  $\alpha$  and  $\beta$ . The exchange–correlation functional applied is that suggested by Perdew, Burke, and Ernzerhof.<sup>51</sup>

**2. The Second-Order Term.** The second-order term in the charge density fluctuations  $\delta\rho$  (second term in eq 1) is approximated by writing  $\delta\rho$  as a superposition of atomic contributions  $\delta\rho = \sum_\alpha \Delta\rho_\alpha$ , which decay quickly with increasing distance from the corresponding center

$$E^{2\text{nd}} = \frac{1}{2} \sum_{\alpha\beta} \int \int' \Gamma[\vec{r}, \vec{r}', \rho_0] \Delta\rho_\alpha \Delta\rho_\beta \quad (4)$$

where  $\Gamma[\vec{r}, \vec{r}', \rho_0]$  denotes the second derivative of the Hartree and exchange–correlation contributions with respect to the atomic-like charge densities.

To further simplify  $E^{2\text{nd}}$ , we apply a monopole approximation<sup>1</sup>

$$\Delta\rho_\alpha \approx \Delta q_\alpha F_{00}^\alpha Y_{00} \quad (5)$$

$F_{00}^\alpha$  denotes the normalized radial dependence of the density fluctuation on atom  $\alpha$ , which is constrained (approximated) to be spherical ( $Y_{00}$  is the zeroth-order spherical harmonics); i.e., the angular deformation of the charge density change in second order is neglected. After integration,  $E^{2\text{nd}}$  becomes a simple two-body expression depending on atomic-like charges

$$E^{2\text{nd}} = \frac{1}{2} \sum_{\alpha\beta} \Delta q_\alpha \Delta q_\beta \gamma_{\alpha\beta} \quad (6)$$

and a function

$$\gamma_{\alpha\beta} = \int \int' \Gamma[\vec{r}, \vec{r}', \rho_0] F_{00}^\alpha F_{00}^\beta Y_{00}^2 \quad (7)$$

The diagonal terms  $\gamma_{\alpha\alpha}$  model the dependence of the total energy on charge density fluctuations (decomposed into atomic contributions) in the second order. The monopole approximation restricts the change of the electron density considered, and no spatial deformations are included; only the change of energy with respect to the change of charge on the atom  $\alpha$  is considered. By neglecting the effect of the chemical environment on atom  $\alpha$ , the diagonal part of  $\gamma$  can be approximated by the chemical hardness  $\eta$  of the atom

$$\gamma_{\alpha\alpha} = U_\alpha = 2\eta_\alpha = \frac{\partial^2 E_\alpha}{\partial^2 q_\alpha} \quad (8)$$

$E_\alpha$  is the energy of the isolated atom  $\alpha$ .  $U_\alpha$  is known as the Hubbard parameter and is twice the chemical hardness of atom  $\alpha$ , which can be estimated from the difference of the ionization potential and the electron affinity of atom  $\alpha$ . For SCC-DFTB, it is calculated using Janak's theorem<sup>52</sup> by taking the first derivative of the energy of the highest occupied orbital with respect to occupation number.<sup>1</sup>

For  $\alpha \neq \beta$ ,  $\gamma_{\alpha\beta}$  is determined analytically by considering, for the moment, only the Hartree contribution and the exchange–correlation contributions will be included implicitly later on. By approximating the charge density fluctuations with spherical charge densities, Slater-like distributions

$$\rho_\alpha(r) = \frac{\tau_\alpha^3}{8\pi} \exp(-\tau_\alpha |\vec{r} - \vec{R}_\alpha|) \quad (9)$$



located at  $\bar{\mathbf{R}}_\alpha$  allow for an analytical evaluation of the Hartree contribution. This leads to a function for  $\gamma_{\alpha\beta}$ , which depends on the parameters  $\tau_\alpha$  and  $\tau_\beta$  that determine the extension of the charge densities of atoms  $\alpha$  and  $\beta$ . This function has a  $1/R_{\alpha\beta}$  dependence for large  $R_{\alpha\beta}$  and approaches a finite value for  $R_{\alpha\beta} \rightarrow 0$ . The neglect of exchange–correlation contributions is a good approximation for large interatomic distances because the exchange–correlation energy decays in a manner proportional to the density overlap for standard GGA functionals. For zero interatomic distances, i.e.,  $\alpha = \beta$ , one finds<sup>1</sup> that

$$\tau_\alpha = \frac{16}{5} \gamma_{\alpha\alpha} \quad (10)$$

A consistent approximation at the Hartree level would consider only the Hartree contributions in  $\gamma_{\alpha\alpha} = U_\alpha$ . Because our calculated Hubbard parameters include exchange–correlation contributions for  $R_{\alpha\beta} \rightarrow 0$ , they are also extrapolated into the binding region due to the curve shape of  $\gamma_{\alpha\beta}$ .

In summary, the standard approximations of the second-order terms in SCC-DFTB<sup>1</sup> contain three major items:

- The charge monopole approximation: This approximation does not imply that higher multipole moments in the electron–electron interaction are completely neglected in DFTB. They are included to a large degree in the  $H_{\mu\nu}^0$  terms. Therefore, the higher multipole terms are neglected only for electron–electron interactions arising from the charge density fluctuations  $\delta\rho$ . Therefore, this approximation is probably uncritical for small charge transfer, i.e., within the limits of the expansion underlying the SCC-DFTB formalism. This is the main difference with respect to CNDO-like methods in semiempirical theory; see also ref 53.

- The Hubbard parameters, evaluated for neutral atoms, are independent of the charge state of the atom. More realistically, the atomic hardness changes with the charge state of the atom and this effect can be captured by including higher-order terms as discussed below.

- Equation 10 makes an interesting statement. It implies that the extension of the charge distribution is inversely proportional to the chemical hardness of the respective atom; i.e. the size of an atom is inversely related to its chemical hardness. It should be emphasized that SCC-DFTB is based on this relation irrespective of its empirical validity. DFTB makes use of this relation, requiring the Hubbard parameter to represent the inverse of the atomic size in  $\gamma_{\alpha\beta}$ ; i.e., for large atoms the onset of the overlap occurs already at large interatomic distances and leads to a deviation from the  $1/R_{\alpha\beta}$  behavior. This deviation effectively decreases the electron–electron interaction in the binding region where the atomic densities overlap. That this relation is not empirically valid throughout the periodic table is the basis of modifications, which will be discussed below.

The second and third approximations are the subject of developments in this work and will be discussed in detail below. To complete the description of SCC-DFTB, we now discuss the last term in the total energy expression.

3. *The Repulsive Potential.* The “double counting” contributions and the core–core repulsion energy (the last four terms in eq 1) are represented as  $E_{\text{rep}}$

$$E_{\text{rep}}[\rho_0] = -\frac{1}{2} \int \int' \frac{\rho'_0 \rho_0}{|\bar{\mathbf{r}} - \bar{\mathbf{r}}'|} + E_{\text{xc}}[\rho_0] - \int V_{\text{xc}}[\rho_0] \rho_0 + E_{\text{cc}} \quad (11)$$

Writing the initial charge density as a superposition of atomic-like neutral charge densities

$$\rho_0 = \sum_{\alpha} \rho_0^{\alpha} \quad (12)$$

centered at the atoms  $\alpha$ , the repulsive energy  $E_{\text{rep}}$  does not depend on the charge density fluctuations and contains no long-range Coulombic interactions due to the neutrality of the atomic-like densities  $\rho_0^{\alpha}$ . However, the repulsive energy as defined above does not go to zero for large interatomic distances  $R_{\alpha\beta}$  but to a constant given by the atomic contributions

$$E_{\text{rep}}[\rho_0] = \sum_{\alpha} E_{\text{rep}}[\rho_0^{\alpha}] \quad R_{\alpha\beta} \rightarrow \infty \quad (13)$$

Therefore, by neglecting the atomic contributions  $E_{\text{rep}}$  can be approximated as a sum of short-ranged two-center terms with respect to the energies  $E_{\text{rep}}[\rho_0^{\alpha}]$  of neutral atomic fragments

$$\tilde{E}_{\text{rep}}[\rho_0] = E_{\text{rep}}[\rho_0] - \sum_{\alpha} E_{\text{rep}}[\rho_0^{\alpha}] = \frac{1}{2} \sum_{\alpha\beta} V[\rho_0^{\alpha}, \rho_0^{\beta}; R_{\alpha\beta}] \quad (14)$$

For given densities  $\rho_0^{\alpha}$ ,  $E_{\text{rep}}$  could be calculated in principle. However, it is convenient to fit this expression to ab initio calculations, as have been done in current implementations.

4. *The SCC-DFTB Total Energy.* With these definitions and approximations, the SCC-DFTB energy finally reads

$$E^{\text{SCC}} = \sum_{\mu\nu} c_{\mu}^i c_{\nu}^j H_{\mu\nu}^0 + \frac{1}{2} \sum_{\alpha\beta} \gamma_{\alpha\beta} \Delta q_{\alpha} \Delta q_{\beta} + \frac{1}{2} \sum_{\alpha\beta} V[\rho_0^{\alpha}, \rho_0^{\beta}; R_{\alpha\beta}] \quad (15)$$

The variational principle leads to approximate Kohn–Sham equations, which have to be solved iteratively for the wavefunction expansion coefficients  $c_{\mu}^i$ , because the Hamiltonian matrix elements depend on the  $c_{\mu}^i$ 's due to the Mulliken charges. The two-body contributions  $V[R_{\alpha\beta}]$  are determined by comparison of the energy according to eq 15 with that from full DFT calculations with respect to the interatomic distance  $R_{\alpha\beta}$ . The resulting energy curve  $V[R_{\alpha\beta}]$  is then analytically represented by splines; for more details, see ref 4.

The neglect of the atomic contributions  $E_{\text{rep}}[\rho_0^{\alpha}]$  has consequences for the calculation of proton affinities and deprotonation energies: The proton has a finite energy of  $0.5U_{\text{H}}$ , because the total energy in eq 15 is not zero due to the use of a neutral hydrogen atom as the reference

$$E^{\text{SCC}} = \frac{1}{2} U_{\text{H}} \quad (16)$$

Clearly, this should be compensated by  $E_{\text{rep}}[\rho_0^{\text{H}}]$ ; therefore we obtain

$$E_{\text{rep}}[\rho_0^{\text{H}}] = -\frac{1}{2} U_{\text{H}} \quad (17)$$

As discussed previously,<sup>54,55</sup> when calculating proton affinities with SCC-DFTB one simply could take the energy of the proton into account. However, this is based on the chemical hardness parameter, which is evaluated for the neutral atom and therefore gives only a rough estimate for the ionized system of  $E_{\text{rep}}[\rho_0^{\text{H}}] = -131.6$  kcal/mol. Calculating  $E_{\text{rep}}[\rho_0^{\text{H}}]$  directly<sup>54</sup> leads to a value of  $E_{\text{rep}}[\rho_0^{\text{H}}] = -141.8$  kcal/mol, which has been used to calculate the proton affinity values in previous studies.

**B. New Developments.** As pointed out above, as a consequence of the monopole approximation, the “shape” of the charge density in SCC-DFTB is not iteratively updated but only

**TABLE 1: Covalent Radii  $r_c$  (Å) Estimated by Politzer et al.,<sup>58</sup> Calculated ( $U_H$ ) and Experimental ( $U_H^{\text{exp}}$ )<sup>59</sup> Hubbard Parameters (in bohr<sup>-1</sup>), and the Effective Radii  $r_c = 5/(16U_H)$  (Å) Estimated Using the Calculated Hubbard Parameters**

	H	C	N	O	F	Si	P	S	Cl
$r_{\text{cov}}$									
with H		0.70	0.65	0.62	0.59	1.00	0.96	0.91	0.59
with first row	0.37	0.74	0.72	0.70	0.69	0.96	0.97	0.98	0.97
with second row	0.46	0.82	0.77	0.74	0.69	1.09	1.08	1.03	1.00
$U_H$	0.42	0.36	0.43	0.50	0.59	0.25	0.29	0.33	0.37
$U_H^{\text{exp}}$	0.47	0.37	0.53	0.45	0.52	0.25	0.36	0.30	0.34
$r_c = 5/(16U_H)$	0.39	0.44	0.39	0.33	0.28	0.66	0.57	0.50	0.44

the distribution of the net atomic (Mulliken) charges. The interaction of the charge density fluctuations in the monopole approximation is governed by the analytic function  $\gamma$ , which assumes the chemical hardness (Hubbard parameter) to be inversely proportional to the atomic size (eq 10). A second approximation is that the Hubbard parameter is independent of the charge state of the atom. In the following, we will discuss these approximations in further detail and suggest corresponding extensions of the SCC-DFTB formalism.

*1. Improving the Interatomic Electrostatic Description.* As described above,  $\gamma_{\alpha\beta}$  is derived from the assumption that the electron–electron interaction in the second-order terms of the DFTB total energy can be evaluated from the interaction of two exponentially decaying charge densities (eq 9), in which the exponent  $\tau_\alpha$  is a measure for the extension of the atomic charge density, or inverse of the atomic “size”. Further, the on-site interaction  $\gamma_{\alpha\alpha}$  should correspond to the electron self-interaction on the atom; i.e., it can be expressed via the Hubbard parameters  $U_\alpha$ , which are equal to twice the chemical hardness  $\eta_\alpha$

$$\gamma_{\alpha\alpha} = U_\alpha = 2\eta_\alpha \quad (18)$$

This immediately leads to the relation between  $\tau_\alpha$  and  $U_\alpha$  (eq 10). In other words, the function  $\gamma_{\alpha\beta}$ , as used in the SCC-DFTB method, assumes that there is an inverse correspondence between the size of an atom,  $1/\tau_\alpha$ , and its chemical hardness parameter,  $U_\alpha$ .<sup>1</sup> For  $R_{\alpha\beta} = 0$ ,  $\gamma_{\alpha\beta}$  assumes a finite value of  $U_\alpha$  and the deviation from  $1/R$  in the region of covalent bonding (1–3 Å) is largely dependent on the size of the respective atoms modeled by  $1/U_\alpha$ . In fact, a very similar approximation is used in semiempirical quantum chemical methods such as MNDO, AM1, or PM3, where  $\gamma$  has a simpler form, as given, for example, by the Klopman–Ohno approximation<sup>56,57</sup>

$$\gamma_{\alpha\beta} = \frac{1}{\sqrt{R_{\alpha\beta}^2 + 0.25(1/U_\alpha + 1/U_\beta)^2}} \quad (19)$$

which also assumes that the size of an atom, which is crucial for determining the deviation of  $\gamma$  from the  $1/R$  behavior, can be estimated based on the chemical hardness of this atom.

To check the validity of this crucial assumption, i.e., to assess how well  $1/U_\alpha$  can be used as a measure of the size of an atom, one can compare covalent radii with the respective chemical hardness values. In a recent work of Politzer and co-workers,<sup>58</sup> various sets of covalent radii have been examined and an overall reasonable agreement between the different concepts has been found. Large deviation, however, has been found in particular for the hydrogen atom.

In Table 1, we summarize the covalent radii from Politzer et al.<sup>58</sup> and the calculated (as described above in section II.A.2) and experimental chemical hardness values (taken from ref 59). The covalent radii of the atoms depend on whether they are

bonded to hydrogen, first row, or second row atoms. In addition, Table 1 shows the atomic radii as calculated from the chemical hardness values with the relation  $r_c = 5/(16U_H)$  (eq 10). The calculated  $r_c$  values are systematically smaller than the covalent radii because they only reflect the half-widths of the Slater-like distribution (eq 9) and not the true covalent radius. Close inspection of Table 1 suggests that the inverse relationship of chemical hardness and atomic size, as suggested by eq 10, only holds well for group II–IV elements; thus application of the  $\gamma_{\alpha\beta}$  expression derived in ref 1 is justified for these elements. A major exception is the hydrogen atom. It has a chemical hardness comparable to nitrogen but has only half of the size.

Because  $\gamma_{\alpha\beta}$  approaches the value  $\gamma_{\alpha\alpha} = U_\alpha$  at short distances, the poor relation between its size and the chemical hardness for H means that modifications have to be made for  $\gamma_{\alpha\beta}$  for all X–H (X being heavy atoms) pairs. In principle, this could be done by modifying the value of  $U_H$  for hydrogen according to its atomic size, which would, however, make the on-site interaction on H,  $\gamma_{H-H}$ , inconsistent with its chemical hardness.

We propose to modify  $\gamma_{\alpha\beta}$  in the intermediate region only, leaving the limiting cases at short and long interatomic distances unchanged. Specifically,  $\gamma_{\alpha\beta}$  has the following form in the standard implementation of the SCC-DFTB method<sup>1</sup>

$$\gamma_{\alpha\beta} = \frac{1}{R_{\alpha\beta}} - S \quad (20)$$

with  $S$  being a short-range function that leads to the desired limit for small interatomic distances. Because the hydrogen atom size according to  $r_c = 5/(16U_H)$  is too large, the density overlap is overestimated; i.e., the electronic interaction starts to deviate from  $1/R_{\alpha\beta}$  too early. To correct for this, an additional damping term is added for the X–H pairs

$$\gamma_{\alpha H} = \frac{1}{R_{\alpha H}} - S \exp\left[-\left(\frac{U_\alpha + U_H}{2}\right)^\zeta R_{\alpha H}^2\right] \quad (21)$$

This leads to a faster decay for the influence of  $U_H$  on the shape of  $\gamma_{\alpha H}$ , thereby reducing the effect of the overlap. This modification contains a single parameter, the exponent  $\zeta$ , which can be fitted to appropriate reference systems as described below.

*2. Third-Order Contributions.* The formal second-order expansion of the DFT total energy leads to the SCC-DFTB formalism,<sup>1</sup> where the second-order one-center integrals are approximated using the Hubbard (chemical hardness) parameters. However, the chemical hardness calculated from small variations around the reference density may be different from chemical hardness parameters calculated from the difference of ionization potential and electron affinity values. More importantly, the Hubbard value may not be a constant for different atomic charge states, as assumed in the second-order SCC-DFTB method.

The change of the (neutral atomic) chemical hardness parameters due to environmental factors can be estimated by their derivatives with respect to the atomic charge. These chemical hardness derivatives can be determined by calculating the chemical hardness values as described above but for charged atoms. Taking the numerical derivative leads to the derivatives of the Hubbard parameters, i.e., to third-order derivatives of the energy of an atom; we obtain  $-0.16$  a.u. for H, C, and N and  $-0.17$  a.u. for O. Interestingly, these values are very similar although the chemical hardness values are quite different (Table 1).

Formally, the charge dependence of the Hubbard parameter can be accounted for by expanding the DFT total energy up to third order in the density fluctuations ( $\int d\mathbf{r}' = f'$ )

$$E[\rho] = E[\rho_0] + \int \left[ \frac{\delta E[\rho]}{\delta \rho} \right]_{\rho_0} \delta \rho + \frac{1}{2} \int \int' \left[ \frac{\delta^2 E[\rho]}{\delta \rho \delta \rho'} \right]_{\rho_0} \delta \rho \delta \rho' + \frac{1}{6} \int \int' \int'' \left[ \frac{\delta^3 E[\rho]}{\delta \rho \delta \rho' \delta \rho''} \right]_{\rho_0} \delta \rho \delta \rho' \delta \rho'' \quad (22)$$

In the following we will introduce similar approximations to the third-order term as already used in the second-order formalism. As for the second-order formalism, we introduce atomic density fluctuations,  $\delta \rho = \sum_{\alpha} \Delta \rho_{\alpha}$  in the monopole approximation (eq 5) and a functional  $\Omega[\mathbf{r}, \mathbf{r}', \mathbf{r}'', \rho_0]$ , which represents the third-order derivative of the total energy with respect to the atomic-like densities (at the reference density  $\rho_0$ ):

$$E^{3\text{rd}} \approx \frac{1}{6} \sum_{\alpha} \sum_{\beta} \sum_{\lambda} \times \Delta q_{\alpha} \Delta q_{\beta} \Delta q_{\lambda} \int \int' \int'' \Omega[\mathbf{r}, \mathbf{r}', \mathbf{r}'', \rho_0] F_{00}^{\alpha} F_{00}^{\beta} F_{00}^{\lambda} Y_{00}^3 \quad (23)$$

In analogy to the second-order formalism, we have to evaluate in particular the integral

$$\omega_{\alpha\beta\lambda} = \int \int' \int'' \Omega[\mathbf{r}, \mathbf{r}', \mathbf{r}'', \rho_0] F_{00}^{\alpha} F_{00}^{\beta} F_{00}^{\lambda} Y_{00}^3 \quad (24)$$

As the simplest approximation, we consider only the one-center (on-site) terms; i.e., we consider the case  $\alpha = \beta = \lambda$ , for which we have to evaluate the third derivative of energy with respect to the density fluctuation on atom  $\alpha$ . In the spirit of evaluating the Hubbard parameters, we approximate this as the third derivative of the energy of an atom ( $E_{\text{at}}$ ) with respect to the atomic charge  $q_{\alpha}$

$$E_{\alpha\alpha\alpha}^{3\text{rd}} = \frac{1}{6} \omega_{\alpha\alpha\alpha} \Delta q_{\alpha}^3 \approx \frac{1}{6} \frac{\partial^3 E_{\text{at}}}{\partial q_{\alpha}^3} \Delta q_{\alpha}^3 = \frac{1}{6} U_{\alpha}^{\text{d}} \Delta q_{\alpha}^3 \quad (25)$$

which contains the derivative of the Hubbard parameter  $U_{\alpha}$  with respect to the atomic charge, denoted by  $U_{\alpha}^{\text{d}}$ . Finally, we arrive at the total energy expression with on-site third-order contributions

$$E = \sum_{i\mu\nu} c_{\mu}^i c_{\nu}^i H_{\mu\nu}^0 + \frac{1}{2} \sum_{\alpha\beta} \gamma_{\alpha\beta} \Delta q_{\alpha} \Delta q_{\beta} + \frac{1}{2} \sum_{\alpha\beta} V[\rho_0^{\alpha}, \rho_0^{\beta}; R_{\alpha\beta}] + \frac{1}{6} \sum_{\alpha} U_{\alpha}^{\text{d}} \Delta q_{\alpha}^3 \quad (26)$$

The extension of this scheme including the off-center contributions is straightforward but requires, within the SCC-DFTB framework, calculating the derivative of  $\gamma_{\alpha\beta}$  with respect to the charge on atom  $\lambda$ ,  $q_{\lambda}$ . This will be explored in future work.

**C. Parameter Fitting and Benchmark Calculations.** For the modified Coulomb interaction, we introduce a single parameter  $\zeta$ , in the damping function associated with  $\gamma_{\alpha\text{H}}$  in eq 21. For the chemical hardness (Hubbard) derivative, one new parameter is required per element.

The modified  $\gamma_{\alpha\text{H}}$  function has a significant impact on hydrogen bonding. For example, the standard SCC-DFTB method yields a binding energy of 3.3 kcal/mol for the water dimer. Choosing  $\zeta = 3.6$  in eq 21 increases this binding energy to 4.6 kcal/mol, which is close to the expected value of 5.0 kcal/mol.<sup>60</sup> The third-order contribution, however, improves the predicted proton affinity substantially. For example, with the estimated  $U_{\alpha}^{\text{d}}$  based on atomic calculations mentioned above, the error in the calculated proton affinity of water is reduced from 26.5 to  $-5.4$  kcal/mol.

These results encouraged us to systematically optimize the parameters in a second step by fitting based on the binding energies and proton affinities of a set of gas-phase compounds that are of general biological interest. For testing, an additional set of small molecules are studied. All reference calculations are carried out using the Gaussian 03<sup>61</sup> program, and all SCC-DFTB calculations are carried out using a locally modified version of CHARMM.<sup>62</sup>

*1. Protocols for Parameter Fitting.* The general fitting set includes a series of biologically relevant molecules (e.g., water clusters and amino acid side chains), and the corresponding properties of interest include 32 proton affinities and 22 binding energies in the gas phase. A genetic algorithm (GA)<sup>63</sup> is used to optimize the Hubbard derivatives and the damping exponent in  $\gamma_{\alpha\text{H}}$  to minimize the penalty function defined as

$$\chi = \frac{\sum_i w_i (Y_i^{\text{ref}} - Y_i^{\text{SCC}})^2}{\sum_i w_i} \quad (27)$$

where the summation is over all properties of interest in a particular set of optimizations (see below),  $w_i$  is the weight of a specific property, and  $Y_i^{\text{ref}}/Y_i^{\text{SCC}}$  are the values of the  $i$ th property from a reference calculation (see below) and a SCC-DFTB calculation with a specific set of  $\{U_{\alpha}^{\text{d}}, \zeta\}$ , respectively. During the GA optimization, the properties of interest include proton affinities, binding energies, and the root-mean-square gradient (GRMS) of the molecule at the reference geometry, addressing both energetic and structural information; the corresponding weights in  $\chi$  are 10, 10, and 1, respectively. The micro-GA technique<sup>63</sup> is applied with a population of 10 chromosomes for 100 generations with uniform crossovers.

Rigorously speaking, the proton affinity of molecule  $A^-$  is the negative of the enthalpy change for the gas-phase reaction  $A^-(g) + H^+(g) \rightarrow AH(g)$  at a given (room) temperature, which involves the thermal vibrational contribution. To avoid a large number of vibrational calculations, we consistently consider only the potential energy contribution in both the reference calculations and the SCC-DFTB calculations during the GA optimization for both proton affinities and binding energies. Another subtle point is, as discussed above, that the energy (eq 15) of a proton in the SCC-DFTB method is not zero; however, once a



**TABLE 2: Different Sets of Parameters Optimized for Improving the SCC-DFTB Approach for Proton Affinity (PA) and Hydrogen-Bonding Binding Energy (BE) Calculations<sup>a</sup>**

set	parameters	NH $\tilde{E}_{\text{rep}}^c$	reference data <sup>d</sup>	$\zeta$	$U_{\text{O,N,C,H}}^d$
0 <sup>b</sup>					-0.17, -0.16, -0.16, -0.16
1	$\zeta$	NHorg	22 BEs	4.50	
2	$U_{\alpha}^d$	NHorg	32 PAs		-0.14, -0.09, -0.08, -0.08
3	$U_{\alpha}^d$	NHmod	32 PAs		-0.14, -0.13, -0.08, -0.14
4	$U_{\alpha}^d$	NHmix	32 PAs		-0.15, -0.13, -0.08, -0.08
5	$\zeta, U_{\alpha}^d$	NHorg	all	4.95	-0.14, -0.08, -0.04, -0.07
6	$\zeta, U_{\alpha}^d$	NHmod	all	4.88	-0.14, -0.13, -0.04, -0.05
7	$\zeta, U_{\alpha}^d$	NHmix	all	4.85	-0.14, -0.12, -0.08, -0.08

<sup>a</sup>  $U_{\alpha}^d$  is the Hubbard derivative (in bohr<sup>-1</sup>) defined in eq 25;  $\zeta$  is the exponent (unitless) in the damping function in  $\gamma_{\text{aH}}$  defined in eq 21. <sup>b</sup> The Hubbard derivatives are calculated based on atoms. <sup>c</sup> “NHorg” is the standard NH repulsive potential; “NHmod” is the shifted NH repulsive potential developed in ref 68; “NHmix” means applying “NHmod” for sp<sup>3</sup>-hybridized acidic nitrogen and “NHorg” for the rest. <sup>d</sup> “All” means that all 32 PAs and 22 BEs are considered in the optimization.

value for  $E_{\text{rep}}[\rho_0^{\text{H}}] = -141.8$  kcal/mol is selected,<sup>54</sup> the results are consistent among all SCC-DFTB calculations.

Regarding the level of reference calculations, except for neutral water hexamer clusters and methylimidazole water clusters, the reference data (energy and geometry) are obtained at the G3B3 level.<sup>64,65</sup> Previous benchmark calculations showed that the G3B3 method predicts the proton affinity for small molecules very well compared to experiments; for 16 species studied, RMSE was 1.2 kcal/mol compared to available experimental data, which makes G3B3 one of the best methods available for proton affinity calculations.<sup>48</sup> For the four neutral water hexamers, complete basis set (CBS) results from Xantheas et al.<sup>66</sup> are used. For the relatively large methylimidazole water clusters, B3LYP/6-31G(d) geometries and single-point energies at the level of MP2 with the G3Large basis set are used; G3Large is a modified version of the 6-311+G(3df,2p) basis set applied in the G2 theory. For the systems in our training set that can be studied by the G3B3 approach, MP2/G3Large shows a strictly systematic negative deviation on the order of 1.0 kcal/mol and thus supports its role as a reliable reference for the methylimidazole–water clusters.

Several sets of optimizations have been carried out as summarized in Table 2. First, the Hubbard derivatives and the damping exponent are optimized separately based on proton affinity and binding energy data, respectively, and their impact on the corresponding properties is made clear by comparison to the standard second-order SCC-DFTB approach. Next, both sets of parameters are optimized simultaneously based on all of the reference systems to establish an improved SCC-DFTB approach for both proton affinity and hydrogen-bonding interactions. Additional complication arises because it is found that nitrogen-containing compounds behave rather differently in terms of proton affinity; thus two additional sets of optimizations are done with slightly adjusted repulsive potentials for the N–H pair (see Table 2 and below for details).

**2. Additional Benchmark Calculations.** To test the transferability of the fitted parameters and modifications to the SCC-DFTB approach, additional benchmark calculations are carried out. For hydrogen-bonding interactions, systems chosen include DNA base pairs and a set of clusters involving small molecules; a set of different conformers of the water dimer studied by Quack and co-workers<sup>67</sup> is also included to probe different regions of the water-dimer potential surface. For proton affinities, tautomerization energies in DNA/RNA bases and the proton affinities of a set of small molecules that mimic commonly found biological cofactors are selected. Most structures involved in the benchmarks are optimized at the B3LYP/6-311++G(d,p) level while higher-level calculations are done for the

energetics wherever possible (see Tables 8–11 for details). For the smaller hydrogen-bonding clusters, G3B3 calculations<sup>64,65</sup> are done to generate the structure and energetics. In the SCC-DFTB calculations, the structures are reoptimized at the respective level.

### III. Results and Discussions

In this section, we first discuss how different modifications of the SCC-DFTB approach impact the calculation of hydrogen-bonding interactions. Next, we present the corresponding discussions regarding proton affinity calculations. Finally, we show results for molecules not included in the fitting set.

**A. Hydrogen-Bonding Interactions.** As shown in Table 3, the standard SCC-DFTB method in almost all cases underestimates the strength of hydrogen-bonding interactions. The magnitude of error is on the order of 2–3 kcal/mol per hydrogen bond and increases slightly as the number of hydrogen bonds increases. For example, the binding energy of the water dimer is underestimated by 1.6 kcal/mol, while that of the water hexamer is underestimated by ~18 kcal/mol, which amounts to about 3 kcal/mol per hydrogen bond. The errors are larger in magnitude for protonated water clusters and protonated imidazole–water complexes. We note that the binding energy of water and hydroxide is overestimated by the standard SCC-DFTB approach by 5.1 kcal/mol, and error cancellation makes the description of multiple-water–hydroxide clusters fortuitously good. Overall, RMSE of 10.5 kcal/mol (3.1 kcal/mol per hydrogen bond) is rather large.

With the damped  $\gamma_{\text{XH}}$  modification (eq 21), the situation improves substantially, especially for neutral and protonated complexes. The largest error is reduced from 20.0 kcal/mol for the standard SCC-DFTB method to 10.9 kcal/mol, and the RMSE is reduced from 10.5 to 6.6 kcal/mol. For the water hexamer, for example, the error per hydrogen bond is reduced to ~1.5 kcal/mol. Unfortunately, because electrostatic interactions are generally enhanced with this modification the overestimated hydrogen bonding for hydroxide–water clusters becomes even worse. For water–hydroxide, for example, the error increases from -5.1 kcal/mol for the standard SCC-DFTB method to -9.3 kcal/mol. As a result, the RMSE per hydrogen bond for all 22 cases studied is only reduced modestly from 3.1 to 2.8 kcal/mol.

With the third-order extension of SCC-DFTB, for which the Hubbard derivatives are either computed for atoms by calculating the third derivative of the energy or optimized based on proton affinity only, the performance for hydrogen-bonding interactions is similar to that of the standard SCC-DFTB method, as expected. The RMSEs are 9.2 and 9.5 kcal/mol with the computed and optimized Hubbard derivatives, respectively, as

**TABLE 3: Binding Energy (in kcal/mol) Comparison between SCC-DFTB and High-Level Ab Initio Methods<sup>a</sup>**

molecules <sup>b</sup>	high level <sup>c</sup>	SCC-DFTB <sup>d</sup>			
		standard	HBond <sup>e</sup>	third order <sup>f</sup>	third order and HBond <sup>g</sup>
2H <sub>2</sub> O	-4.9/-0.1	1.6	0.7	1.1/1.2	0.2
3H <sub>2</sub> O	-15.1/-0.6	5.5	2.3	3.7/4.2	0.3
4H <sub>2</sub> O	-27.4/-1.0	10.3	5.5	7.6/8.2	2.7
5H <sub>2</sub> O	-36.3/-1.3	14.0	7.9	10.6/11.4	4.2
2H <sub>2</sub> O(H <sup>+</sup> )	-33.9/-0.8	4.6	1.2	6.1/5.7	2.5
3H <sub>2</sub> O(H <sup>+</sup> )	-57.3/-1.0	11.1	5.4	11.8/11.7	6.3
4H <sub>2</sub> O(H <sup>+</sup> )	-77.2/-1.0	15.6	8.6	15.2/15.3	8.7
5H <sub>2</sub> O(H <sup>+</sup> )	-91.9/-1.2	20.0	10.9	18.6/19.0	10.0
2H <sub>2</sub> O(-H <sup>+</sup> )	-27.4/-1.2	-5.1	-9.3	2.8/1.3	-3.4
3H <sub>2</sub> O(-H <sup>+</sup> )	-48.6/-1.3	-2.4	-10.4	4.6/3.3	-5.7
4H <sub>2</sub> O(-H <sup>+</sup> )	-66.7/-1.7	1.1	-9.0	8.8/7.4	-3.5
5H <sub>2</sub> O(-H <sup>+</sup> )	-86.3/-1.8	7.2	-6.5	11.1/11.0	-4.5
NH <sub>3</sub> (H <sub>2</sub> O)	-6.6/-0.2	2.6	2.0	2.1/2.2	1.4
NH <sub>4</sub> <sup>+</sup> (H <sub>2</sub> O)	-20.4/-0.4	1.7	-0.4	1.7/1.8	-0.2
6H <sub>2</sub> O_book	-45.6 (CBS)/-	17.3	9.3	12.9/14.0	4.4
6H <sub>2</sub> O_cage	-45.8 (CBS)/-	17.1	8.1	12.3/13.4	2.3
6H <sub>2</sub> O_prism	-45.9 (CBS)/-	16.9	7.3	11.9/13.1	1.6
6H <sub>2</sub> O_ring	-44.9 (CBS)/-	17.6	10.1	13.3/14.3	5.6
methylimidazole(-H <sup>+</sup> )(H <sub>2</sub> O)	-16.2 (MP2)/-	0.1	-1.3	-6.1/-3.9	-6.0
methylimidazole(H <sub>2</sub> O)_1	-6.4 (MP2)/-	2.7	2.1	2.6/2.6	2.1
methylimidazole(H <sub>2</sub> O)_2	-8.3 (MP2)/-	2.7	2.0	1.2/1.5	0.8
methylimidazoleH <sup>+</sup> (H <sub>2</sub> O)	-16.4 (MP2)/-	4.4	3.2	4.2/4.2	3.0
Error Analysis <sup>h</sup>					
MAXE	-1.7 <sup>i</sup>	20.0	10.9	18.6/19.0	10.0
RMSE	1.1 <sup>i</sup>	10.5	6.6	9.2/9.5	4.4
MUE	1.0 <sup>i</sup>	8.2	5.6	7.7/7.8	3.6
MSE	-1.0 <sup>i</sup>	7.6	2.3	7.2/7.4	1.5
Error Analysis (per Hydrogen Bond)					
MAXE	-1.2	5.6	-9.3	6.1/5.9	-6.0
RMSE	0.5	3.1	2.8	3.2/3.0	2.2
MUE	0.4	2.7	2.1	2.8/2.6	1.6
MSE	-0.4	2.1	0.4	2.2/2.3	0.4

<sup>a</sup> The binding energy (BE) is computed as the energy difference between the complex and the isolated molecules at 0 K in the gas phase. No zero-point energy correction has been included. <sup>b</sup> Examples of notation: "2H<sub>2</sub>O", neutral water dimer; "2H<sub>2</sub>O(H<sup>+</sup>)", protonated water dimer; "2H<sub>2</sub>O(-H<sup>+</sup>)", deprotonated water dimer; "6H<sub>2</sub>O\_book", neutral water hexamer in the book configuration; "methylimidazole(-H<sup>+</sup>)(H<sub>2</sub>O)", deprotonated methylimidazole complexed with water; "methylimidazole(H<sub>2</sub>O)\_1", neutral methylimidazole complexed with water as the hydrogen-bond donor; "methylimidazole(H<sub>2</sub>O)\_2", neutral methylimidazole complexed with water as the hydrogen-bond acceptor; "methylimidazoleH<sup>+</sup>(H<sub>2</sub>O)", protonated methylimidazole complexed with water. <sup>c</sup> The number before the slash is the high-level reference data, which is based on G3B3 calculations for the first 14 molecules, CBS results of ref 50 for the water hexamers, and MP2/G3Large for the rest. The number after the slash is the deviation of the MP2/G3Large result from the G3B3 value (i.e., BE<sub>MP2</sub> - BE<sub>G3B3</sub>). For the G3Large basis set, please refer to <http://chemistry.anl.gov/compmat/g3theory.htm>. <sup>d</sup> The numbers are the BE differences between various SCC-DFTB models and the corresponding high-level result (i.e., BE<sub>SCC-DFTB</sub> - BE<sub>high-level</sub>). Unless indicated otherwise, the "NHmod" repulsive potential is used. <sup>e</sup> Obtained with parameter set 1 in Table 2. <sup>f</sup> The numbers before and after the slash are obtained with parameter sets 0 and 3 (Table 2), respectively. <sup>g</sup> The values are obtained with parameter set 6 (Table 2). <sup>h</sup> The notation for errors is the same throughout the work: MAXE, error with the largest magnitude, defined as sign(err) max(|err|); RMSE, root-mean-square error, defined as  $\langle(\text{err})^2\rangle^{1/2}$ ; MUE, mean unsigned error, defined as  $\langle|\text{err}|\rangle$ ; MSE, mean signed error  $\langle\text{err}\rangle$ . <sup>i</sup> The error analysis for MP2/G3Large is based on the first 14 molecules only.

compared to the value of 10.5 kcal/mol for the standard SCC-DFTB method. The major difference is that the water-hydroxide interactions are no longer overestimated.

When both the third-order extension and the damped  $\gamma_{\text{XH}}$  modification are introduced, with the Hubbard derivatives and the damping exponent optimized based on all reference systems, the performance improves substantially over using the damped  $\gamma_{\text{XH}}$  alone. The RMSE, for example, is reduced to 4.4 kcal/mol for total binding energies (2.2 kcal/mol per hydrogen bond). For neutral water clusters, the error per hydrogen bond is about 1.0 kcal/mol; the error is slightly larger for protonated water and about half of that of the standard SCC-DFTB method. For the hydroxide-water clusters, the strength of interaction is still overestimated although the magnitude is substantially reduced from using the damped  $\gamma_{\text{XH}}$  alone; e.g., the error for the water-hydroxide interaction is reduced from -9.3 to -3.4 kcal/mol.

**B. Proton Affinities.** 1. *Impact of the Third-Order Contribution with Calculated Hubbard Derivatives.* Consistent with the previous study of Range et al.,<sup>48</sup> the standard SCC-DFTB approach has rather large errors for the proton affinities (Table

4) that are comparable to the AM1 and PM3 approaches. The RMSE for the 32 reference systems is 11.6 kcal/mol, and the largest error is 26.5 kcal/mol (for water deprotonation). This magnitude of error is unacceptable for most applications. It can be observed that the error is the largest for small molecules in which the excess charge upon deprotonation is strongly localized. For these systems, we expect the charge-dependent Hubbard parameters to have a significant impact, which is indeed the case as shown in Table 4. For these deprotonation processes involving oxygen, the largest error is reduced to -6.3 kcal/mol and the RMSE is only 3.6 kcal/mol! Considering the simplicity of the approach (only one extra parameter is introduced per element), the performance is remarkable.

As shown in Table 5, the SCC-DFTB proton affinities (PAs) involving nitrogen show a peculiarity: The error seems to correlate with the hybridization state of the nitrogen. For sp<sup>3</sup> cases (e.g., NH<sub>4</sub><sup>+</sup> or lysine side chain), the errors tend to be substantially larger than those for sp<sup>2</sup> cases (e.g., protonated methylimidazole) by approximately 10 kcal/mol. This trend holds when the third-order on-site terms are included (first



**TABLE 4: Proton Affinity (in kcal/mol) Comparison between SCC-DFTB (Standard and Set 0 in Table 2) and High-Level Ab Initio Methods for Molecules with Acidic Oxygen<sup>a</sup>**

molecules <sup>b</sup>	high level <sup>c</sup>	SCC-DFTB <sup>d</sup>	
		standard	third order
H <sub>2</sub> O	398.4/−1.1	26.5	−5.4
2H <sub>2</sub> O	375.9/−2.2	19.8	−3.6
3H <sub>2</sub> O	365.0/−2.0	18.5	−4.6
4H <sub>2</sub> O	359.1/−1.9	17.4	−4.2
5H <sub>2</sub> O	348.4/−1.7	19.7	−4.9
CH <sub>3</sub> OH	392.6/−1.5	4.5	−6.3
CH <sub>3</sub> CH <sub>2</sub> OH	388.3/−1.2	8.7	−2.8
CH <sub>3</sub> CH <sub>2</sub> CH <sub>2</sub> OH	387.6/−1.3	7.9	−3.5
CH <sub>3</sub> −CH(OH)−CH <sub>3</sub>	385.6/−1.1	11.5	−0.5
HCOOH	351.2/−1.7	11.9	3.1
CH <sub>3</sub> COOH	355.1/−1.6	11.3	1.6
CH <sub>3</sub> CH <sub>2</sub> COOH	354.5/−1.5	11.2	1.9
C <sub>6</sub> H <sub>5</sub> OH	356.7/−1.9	5.5	0.2
<i>p</i> -CH <sub>3</sub> −C <sub>6</sub> H <sub>5</sub> OH	357.9/−1.8	4.6	−0.4
<i>p</i> -NO <sub>2</sub> −C <sub>6</sub> H <sub>5</sub> OH	334.6/−1.1	0.9	−5.4
H <sub>3</sub> O <sup>+</sup>	171.2/−0.8	9.8	3.4
2H <sub>2</sub> O(H <sup>+</sup> )	200.2/−0.2	6.8	−1.6
3H <sub>2</sub> O(H <sup>+</sup> )	213.4/−0.4	4.2	−4.6
4H <sub>2</sub> O(H <sup>+</sup> )	221.1/−0.9	4.3	−4.3
5H <sub>2</sub> O(H <sup>+</sup> )	226.7/−0.7	3.9	−4.5
CH <sub>3</sub> OH <sub>2</sub> <sup>+</sup>	186.8/−1.2	1.9	−1.5
H <sub>2</sub> COH <sup>+</sup>	177.1/−2.3	−1.6	−3.7
CH <sub>3</sub> CHOH <sup>+</sup>	190.2/−2.2	0.1	−2.1
Error Analysis			
MAXE	−2.3	26.5	−6.3
RMSE	1.5	11.5	3.6
MUE	1.4	9.2	3.2
MSE	−1.4	9.1	−2.3

<sup>a</sup> The proton affinity (PA) is calculated with the potential energies at 0 K without any vibrational contribution. <sup>b</sup> The molecules are given in the protonated form in PA calculations. <sup>c</sup> The number before the slash is the PA at the G3B3 level; the number after the slash is the MP2/G3Large PA difference from the G3B3 result (i.e., PA<sub>MP2</sub> − PA<sub>G3B3</sub>). <sup>d</sup> The numbers are the differences between the calculated PA with various SCC-DFTB models and the G3B3 results (i.e., PA<sub>SCCDFTB</sub> − PA<sub>G3B3</sub>). The third-order results are based on the calculated (not optimized) Hubbard derivatives, i.e., parameter set 0 in Table 2.

**TABLE 5: Proton Affinity (in kcal/mol) Comparison between SCC-DFTB (Standard and Set 0 in Table 2) and High-Level Ab Initio Methods for Molecules with Acidic Nitrogen<sup>a</sup>**

molecules	high level	SCC-DFTB <sup>b</sup>	
		standard	third order
HCNH <sup>+</sup>	176.0/−1.7	−2.2/9.4/−2.2	−3.7/8.0/−3.7
CH <sub>3</sub> CNH <sup>+</sup>	192.3/−1.8	−4.1/7.6/−4.1	−5.9/5.9/−5.9
C <sub>5</sub> H <sub>5</sub> NH <sup>+</sup>	229.5/−2.0	−6.9/4.7/−6.9	−7.6/4.0/−7.6
methylimidazoleH <sup>+</sup>	237.3/−2.2	−2.5/9.1/−2.5	−3.6/8.0/−3.6
arginineH <sup>+</sup>	249.3/−1.2	−1.8/10.0/−1.8	−7.7/4.1/−7.7
NH <sub>3</sub>	413.9/−1.4	20.6/32.8/32.8	−22.2/−10.8/−10.8
NH <sub>4</sub> <sup>+</sup>	212.3/−1.1	−14.2/−2.9/−2.9	−18.1/−6.9/−6.9
CH <sub>3</sub> NH <sub>3</sub> <sup>+</sup>	223.3/−1.2	−16.6/−5.2/−5.2	−18.5/−7.1/−7.1
lysineH <sup>+</sup>	228.2/−1.2	−16.5/−5.1/−5.1	−18.5/−7.1/−7.1
Error Analysis			
MAXE	−2.2	20.6/32.8/32.8	−22.2/−10.8/−10.8
RMSE	1.6	11.8/12.9/11.6	13.7/7.2/7.0
MUE	1.5	9.5/9.6/7.1	11.8/6.9/6.7
MSE	−1.5	−4.9/6.7/0.2	−11.8/−0.2/−6.7

<sup>a</sup> See the footnotes of Table 4 for the notation of molecules and details of the high-level results. <sup>b</sup> For both the standard and the third-order SCC-DFTB results, the three numbers for each molecule are calculated with the “NHorg”, “NHmod”, and “NHmix” sets of the NH repulsive potential (see Table 2 footnotes), respectively.

column under “third-order” in Table 5). In a previous study,<sup>68</sup> where we had to describe a proton transfer from nitrogen to

**TABLE 6: Proton Affinity (in kcal/mol) Errors for the Optimized SCC-DFTB Models as Compared to High-Level Ab Initio Methods for Molecules with Acidic Oxygen<sup>a</sup>**

error analysis	SCC-DFTB	
	third order <sup>b</sup>	third order and HBond <sup>c</sup>
MAXE	−4.4/4.5/5.1	−6.1/−7.5/−6.8
RMSE	2.4/2.5/2.4	3.1/3.3/3.4
MUE	1.9/2.1/1.9	2.8/2.9/3.0
MSE	0.0/−0.1/−0.1	−0.2/−0.5/−0.4

<sup>a</sup> See the footnotes of Table 4 for the notation of molecules and details of the high-level results. For the specific PA results, see the Supporting Information. <sup>b</sup> The three numbers for each molecule are obtained with parameter sets 2, 3, and 4 (Table 2), respectively. <sup>c</sup> The three numbers for each molecule are obtained with parameter sets 5, 6, and 7 (Table 2), respectively.

oxygen, a special parametrization of the repulsive potential for the N−H pair, termed “NHmod”, has been introduced, which was introduced to correct for proton affinity errors for sp<sup>3</sup>-hybridized nitrogens. Specific problems with nitrogen in certain chemical environments, mostly for sp<sup>3</sup> chemical environments, could not be resolved by the third-order terms: Here, we found errors of about 10 kcal/mol. We first recognized this problem when investigating intramolecular proton-transfer reactions in the DNA bases guanine and uracil, where proton acceptors can be either oxygen or nitrogen. To correct this error, we developed a special parametrization for nitrogen by modifying the N−H repulsive potential to correct for the wrong energetics.<sup>4</sup> Technically, this is done by adding a constant shift of 10 kcal/mol to the N−H repulsive energy pair potential. Of course, this is a severe limitation because this shift should be only applied to N−H bonds with sp<sup>3</sup> nitrogen.

Because the major effect of “NHmod” is to uniformly shift the nitrogen proton affinity by approximately 10 kcal/mol, using this set of repulsive potential (second columns for both “standard” and “third-order” in Table 5) tends to produce errors of comparable absolute values for different nitrogen-containing species although the sign of error varies depending on the hybridization state of nitrogen, regardless of whether third-order terms are included or not.

Clearly, the introduction of “NHmod” is not a generally satisfying solution because it attempts to account for deficiencies in the electronic part of the SCC-DFTB method, which is obviously not remedied by the current third-order formalism. The problems seem to be rooted in the Hamiltonian matrix elements, and the precise reasons for such dependence on the nitrogen hybridization state are not clear and currently under investigation. As a practical solution at this stage, we recommend to use the “NHmod” repulsive potential when treating proton-transfer reactions for sp<sup>3</sup> nitrogen species and the standard parametrization for the rest, whenever this is possible.

Applying “NHmod” only to the sp<sup>3</sup> nitrogen species (last four molecules in Table 5 and the standard repulsive potential for the rest), we find a mean deviation of 6.7 kcal/mol for the nine molecules and a RMSE of 7.0 kcal/mol.

**2. Results with Optimized Parameters.** As an attempt to further improve the calculated PAs, the Hubbard derivatives are treated as free parameters to be optimized using a genetic algorithm. As summarized in Table 2, six sets of parameters have been developed, three sets using the third-order formalism alone (with different NH repulsive potentials) and three sets combining the third-order and the modified Coulomb interaction.

For the PAs of the oxygen species, the performance of the three sets is very similar (Table 6). Basically, the systematic error in the PAs is removed with a MSE close to be zero; the

**TABLE 7: Proton Affinity (in kcal/mol) Comparison between Optimized SCC-DFTB Models and High-Level Ab Initio Methods for Molecules with Acidic Nitrogen<sup>a</sup>**

molecules	high level	SCC-DFTB	
		third order <sup>b</sup>	third order and HBond <sup>c</sup>
HCNH <sup>+</sup>	176.0/−1.7	−3.0/8.3/−3.0	−3.6/8.1/−3.9
CH <sub>3</sub> CNH <sup>+</sup>	192.3/−1.8	−5.0/6.3/−5.1	−5.3/6.4/−5.6
C <sub>3</sub> H <sub>5</sub> NH <sup>+</sup>	229.5/−2.0	−7.3/4.1/−7.4	−7.7/3.8/−7.8
methylimidazoleH <sup>+</sup>	237.3/−2.2	−3.0/8.2/−3.2	−3.3/8.2/−3.5
arginineH <sup>+</sup>	249.3/−1.2	−4.8/5.2/−6.5	−5.1/4.6/−6.7
NH <sub>3</sub>	413.9/−1.4	3.3/2.5/1.2	5.1/0.2/4.0
NH <sub>4</sub> <sup>+</sup>	212.3/−1.1	−16.1/−6.1/−5.9	−18.4/−8.7/−8.6
CH <sub>3</sub> NH <sub>3</sub> <sup>+</sup>	223.3/−1.2	−17.6/−6.8/−6.6	−19.1/−8.2/−8.3
lysineH <sup>+</sup>	228.2/−1.2	−17.6/−6.8/−6.6	−18.8/−8.0/−8.1
Error Analysis			
MAXE	−2.2	−17.6/8.3/−7.4	−19.1/−8.7/−8.6
RMSE	1.6	10.6/6.3/5.4	11.6/6.8/6.6
MUE	1.5	8.6/6.0/5.1	9.6/6.2/6.3
MSE	−1.5	−7.9/1.7/−4.8	−8.5/0.7/−5.4

<sup>a</sup> See the footnotes of Table 4 for the notation of molecules and details of the high-level results. <sup>b</sup> The three numbers for each molecule are obtained with parameter sets 2, 3, and 4 (Table 2), respectively. <sup>c</sup> The three numbers for each molecule are obtained with parameter sets 5, 6, and 7 (Table 2), respectively.

**TABLE 8: Benchmark Calculations of SCC-DFTB for Hydrogen-Bond Interactions (in kcal/mol) in DNA Base Pairs<sup>a</sup>**

error analysis	SCC-DFTB		
	standard	third order <sup>b</sup>	third order and HBond <sup>c</sup>
MAXE	3.0	1.7/1.7	−1.2
RMSE	1.6	0.9/0.9	0.8
MUE	1.4	0.7/0.7	0.7
MSE	1.4	0.6/0.7	−0.3

<sup>a</sup> The reference data are MP2 calculations with large basis sets by Hobza et al.<sup>53</sup> For the specific data, see the Supporting Information. <sup>b</sup> The numbers before and after the slash are obtained with parameter sets 0 and 3 (Table 2), respectively. <sup>c</sup> The numbers are obtained with parameter set 6 (Table 2).

RMSE is about 2.4 kcal/mol, which is very encouraging for a semiempirical method. When both the third-order and damped  $\gamma_{\text{XH}}$  modification are considered, the errors in the calculated PAs are somewhat increased because PA and binding energy need to be balanced. The RMSE is about 3.0 kcal/mol, and the largest error is −6.1 kcal/mol, which are quite comparable for the third-order model optimized based on PA alone.

The situation is different for the nitrogen species (Table 7); using parameters optimized either with the standard N–H repulsive potential or with the “NHmod” alone leads to very large errors. A consistent trend is observed only if different NH repulsive potentials are used for different species based on the hybridization state of the acidic nitrogen (last column for “third-order” in Table 7). Even with this “NHmix” optimization set, the NH<sub>3</sub> molecule shows up as an exception, for which the PA is overestimated while for all other species the PAs are underestimated (i.e., negative error). The “NHmix” optimization set has a MSE of −4.8 kcal/mol and a RMSE of 5.4 kcal/mol, which are only slightly smaller than the results obtained with the calculated Hubbard derivatives (last column in Table 5). Including the damped  $\gamma_{\text{XH}}$  modification does not change the trend and slightly increases the errors in the PAs.

In the context of realistic applications, the quantities of primary interest are the relative PAs between O and N species, which often act as proton donors and acceptors in biological processes. Because the optimized PAs show a mean error of about 0.0 kcal/mol for oxygen species and approximately −5

**TABLE 9: Benchmark Calculations for SCC-DFTB for the Binding Energy of Small Molecule Clusters<sup>a</sup>**

error analysis	SCC-DFTB		
	standard	third order <sup>b</sup>	third order and HBond <sup>c</sup>
MAXE	4.5	4.2/4.2	3.0
RMSE	2.5	2.3/2.3	1.4
MUE	2.1	1.9/1.9	1.1
MSE	1.9	1.5/1.5	0.6

<sup>a</sup> The reference data are based on G3B3 calculations; for their structures and specific data, see the Supporting Information. <sup>b</sup> The numbers before and after the slash are obtained with parameter sets 0 and 3 (Table 2), respectively. <sup>c</sup> The number is obtained with parameter set 6 (Table 2).

**TABLE 10: Benchmark Calculations for SCC-DFTB for the Tautomerization Energy (in kcal/mol) of Neutral DNA and RNA Bases**

molecules	B3LYP 6-311++G**	SCC-DFTB <sup>a</sup>		
		standard	third order <sup>b</sup>	third order and HBond <sup>c</sup>
adenine	11.9	0.1	−1.6/−1.1	−1.1/−0.8
cytosine	1.6	3.2	1.7/2.3	2.2/2.3
guanine	1.3	4.5	4.4/4.3	3.1/−7.6
thymine	13.2	0.9	1.3/1.1	−0.1/−11.1
uracil	12.4	1.4	1.7/1.5	0.4/−10.6
Error Analysis				
MAXE		4.5	4.4/4.3	3.1/−11.1
RMSE		2.6	2.4/2.4	1.8/7.7
MUE		2.0	2.1/2.1	1.4/6.5
MSE		2.0	1.5/1.6	0.9/−5.6

<sup>a</sup> The modified set of NH repulsive potentials (NHmod) is used unless stated otherwise. <sup>b</sup> The numbers before and after the slash are obtained with parameter sets 0 and 3 (Table 2), respectively. <sup>c</sup> The numbers before and after the slash are obtained with parameter sets 6 and 5 (Table 2), respectively.

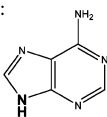
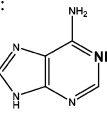
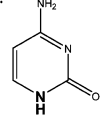
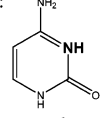
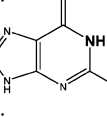
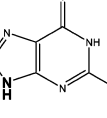
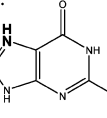
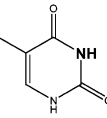
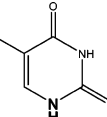
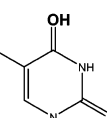
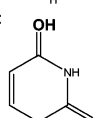
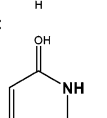
kcal/mol for nitrogen species, we expect average errors in this range for many applications. Interestingly, this average error is comparable, in fact even slightly larger, than that obtained with the calculated (not optimized) Hubbard derivatives, which is −4.4 kcal/mol (without the damped  $\gamma_{\text{XH}}$ ). This consistency suggests that the Hubbard derivative parameters should be rather transferable to many molecular systems.

Benchmark calculations using AM1 and PM3 on the same molecule set indicate (see Supporting Information for more details) that these methods have similar problems to describe the PAs consistently, comparable to the standard SCC-DFTB method. This may be a general problem of minimal basis set methods, because DFT calculations without diffuse functions face a similar problem, yielding a very inhomogeneous description of the PAs in the molecule set finding errors of a similar size (see Supporting Information).

**C. Additional Benchmarks. 1. Hydrogen-Bonding Systems.** As shown in Table 8, the standard SCC-DFTB method does a rather good job for the hydrogen-bonding energies for all 24 base pairs, with a RMSE of 1.6 kcal/mol and a maximal error of 3.0 kcal/mol, as compared to MP2 calculations of Hobza and co-workers.<sup>69</sup> The third-order extension, as expected, does not change the result dramatically although the errors tend to systematically decrease; the RMSE is 0.9 kcal/mol, and the maximum error is 1.7 kcal/mol. With both third-order and damped  $\gamma_{\text{XH}}$ , the RMSE is further reduced slightly to 0.8 kcal/mol and the maximal error is −1.2 kcal/mol.

For the set of hydrogen-bonding complexes studied in Table 9, which includes both neutral and charged species, a small but systematic decrease in error is also observed when modified

TABLE 11: Benchmark Calculations for SCC-DFTB for the Proton Affinity (in kcal/mol) of DNA and RNA Bases<sup>a</sup>

Molecules	B3LYP 6-311++G**	SCCDFTB <sup>b</sup>		
		Standard	3 <sup>rd</sup> Order <sup>c</sup>	3 <sup>rd</sup> Order and HBond <sup>d</sup>
A: 	344.0	26.1	21.3/22.3	22.5/12.5
A: 	232.7	14.2	11.6/11.9	11.6/1.1
C: 	354.8	15.3	9.4/10.5	10.5/0.3
C: 	235.8	8.0	4.3/4.7	4.3/-6.3
G: 	346.1	21.3	14.9/15.7	15.4/6.1
G: 	343.7	27.2	21.1/22.5	22.7/12.2
G: 	236.3	15.0	11.8/12.5	12.4/1.3
T: 	354.8	13.2	5.7/6.6	6.5/-3.7
T: 	342.4	12.5	5.1/6.3	6.5/-4.1
T: 	210.6	9.5	5.2/5.8	6.2/6.5
U: 	211.6	9.8	5.5/6.0	6.5/6.8
U: 	354.1	13.8	6.1/7.0	6.9/-3.3
		Error Analysis		
MAXE		27.2	21.3/22.5	22.7/12.5
RMSE		16.6	11.7/12.5	12.5/6.6
MUE		15.5	10.2/11.0	11.0/5.35
MSE		15.5	10.2/11.0	11.0/2.5

<sup>a</sup> A, C, G, T, and U represent the bases for purine adenine, pyrimidine cytosine, purine guanine, pyrimidine thymine and pyrimidine uracil, respectively. The deprotonation position is in bold. <sup>b</sup> The modified set of NH repulsive potentials (NHmod) is used unless stated otherwise. <sup>c</sup> The numbers before and after the slash are obtained with parameter sets 0 and 3 (Table 2), respectively. <sup>d</sup> The numbers before and after the slash are obtained with parameter sets 6 and 5 (Table 2), respectively.



**TABLE 12: Benchmark Calculations for SCC-DFTB for the Proton Affinities (in kcal/mol) of Several Model Systems of Biological Cofactors**

molecules <sup>a</sup>	B3LYP <sup>b</sup>	SCC-DFTB		
		standard	third order <sup>c</sup>	third order and HBond <sup>d</sup>
GFPH	335.8	8.4	3.7/4.5	5.7
UBQH	437.7	15.6	3.8/6.5	7.9
UBQH <sub>2</sub>	346.1	10.6	3.9/5.4	6.2
FADH	333.6	11.6	7.1/8.0	8.2
FADH <sub>2</sub>	332.7	10.3	3.5/4.6	4.6
Error Analysis				
MAXE		15.6	7.1/8.0	8.2
RMSE		11.5	4.6/5.9	6.7
MUE		11.3	4.4/5.8	6.5
MSE		11.3	4.4/5.8	6.5

<sup>a</sup> Protonated states are listed: "GFPH", the chromophore in the green fluorescent protein; UBQH/UBQH<sub>2</sub>, ubiquinone; FADH/FADH<sub>2</sub>, flavin adenine dinucleotide. <sup>b</sup> The basis set is 6-311++G\*\*. <sup>c</sup> The numbers before and after the slash are obtained with parameter sets 0 and 3 (Table 2), respectively. <sup>d</sup> The number is obtained with parameter set 6 (Table 2).

$\gamma_{\text{XH}}$  is used; for example, the RMSE for the standard SCC-DFTB method is 2.5 kcal/mol, while that for the optimized parameter set 3 (third-order plus damped  $\gamma_{\text{XH}}$ ) is 1.3 kcal/mol.

2. *Proton Affinities.* For the tautomerization energies of DNA and RNA bases, the standard SCC-DFTB method with the NHmod repulsive potential gives rather good results due to error cancellations for the PAs associated with the two tautomers (see below); the RMSE is only 2.6 kcal/mol (Table 10). With the third-order extension, the result is essentially the same with a RMSE of 2.4 kcal/mol. With both third-order and damped  $\gamma_{\text{XH}}$ , the result is further improved slightly with a RMSE of 1.8 kcal/mol. It should be noted that the NH repulsive potential makes a notable difference here; with the original NH repulsive potential with the third-order and damped  $\gamma_{\text{XH}}$ , for example, the RMSE is as large as 7.7 kcal/mol.

The absolute PAs of the DNA and RNA bases, by contrast, still have sizable errors (Table 11). While set 6 leads to quite reasonable tautomerization energies, it is not acceptable for the calculation of absolute PAs. This clearly shows the limits of the current DFTB version. The errors may be associated with the use of a minimal basis set; an extension of the basis set may remedy the situation. However, the current version of DFTB should therefore be carefully tested before application to new chemical species.

Finally, for the model biological "cofactors" shown in Table 12, the standard SCC-DFTB has large errors with a RMSE of 11.5 kcal/mol. Including the third-order extension substantially reduces the error to a RMSE of about 5 kcal/mol. With both third-order and damped  $\gamma_{\text{XH}}$ , however, the errors become larger in magnitude; the RMSE increases to 6.7 kcal/mol, which is quite significant although still a major improvement over the standard SCC-DFTB method.

#### IV. Conclusions

Many biological applications of QM/MM simulations require that the method is capable of accurately describing proton affinities and hydrogen-bonding interactions. This is a significant challenge especially for semiempirical QM methods, which currently are the most practical for carrying out QM/MM simulations with a sufficient amount of sampling. Motivated by such considerations, we have improved the formulation of the SCC-DFTB approach by including third-order terms in density expansion and modifying the short-range behavior of

the  $\gamma$  function for X–H pairs. Both improvements have been proposed based on physical considerations rather than ad hoc parametrizations.

These modifications are shown to significantly improve the reliability of the SCC-DFTB approach. In particular, the third-order terms, even if only the on-site terms are considered, improve proton affinities dramatically. The damped  $\gamma_{\text{XH}}$ , however, improves the description of hydrogen-bonding interactions (by approximately 1–2 kcal/mol per hydrogen bond). Using a set of small molecules of biological interest, several sets of parameters have been fitted. Considering the small number of parameters needed (one Hubbard derivative for each element type, one parameter that describes the damping of  $\gamma_{\text{XH}}$ ), the results are expected to be rather transferable to systems beyond the fitting set, which is largely supported by additional benchmark systems.

Although satisfying progress has been made, there are still major limitations in the improved SCC-DFTB approach. For example, although the RMSE of proton affinities for oxygen species is fairly small, typically on the order of 2–3 kcal/mol, the errors in the proton affinities for nitrogen species appear to be dependent on the hybridization state of the nitrogen. The origin of this is not well understood and requires further study. Although this limitation can be somewhat alleviated by adopting different repulsive potentials for the N–H pair, such a "remedy" is clearly only useful for proton affinity calculations but less suitable for studying reactions. For the hydrogen-bonding interactions, the errors in the neutral/positively charged species and negatively charged species tend to be of different signs; this systematic behavior also requires further studies to improve.

In short, our improvements in the SCC-DFTB method are expected to enhance the applicability of this approximate density functional method, especially in biological applications. On the basis of the current set of benchmark calculations, it appears that set 7 in Table 2 is the most useful in many applications. It leads to a quite reasonable overall performance for hydrogen-bonded systems, although there are still problems for some systems, as has been shown for the absolute PAs of the DNA bases. Whether the accuracy of the method is sufficient for the question of interest depends on the system of interest and needs to be established with careful benchmark calculations.

**Acknowledgment.** The research discussed here has been partially supported by the National Institutes of Health (Grant Nos. R01-GM071428-01 to Q.C. and 1R01-GM62248-06 to D.Y.). Q.C. also acknowledges a Research Fellowship from the Alfred P. Sloan Foundation. M.E. acknowledges funding from the German Science Foundation. Computational resources from the National Center for Supercomputing Applications at the University of Illinois are greatly appreciated.

**Supporting Information Available:** Detailed data for specific molecules associated with Tables 6, 8, and 9, comparison of DFT, AM1, and PM3 proton affinities with ab initio data for the test set described in the main text, and structures for the hydrogen-bonded complexes in Table 9. This material is available free of charge via the Internet at <http://pubs.acs.org>.

#### References and Notes

- (1) Elstner, M.; Porezag, D.; Jungnickel, G.; Elstner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.
- (2) Elstner, M.; Frauenheim, T.; Kaxiras, E.; Seifert, G.; Suhai, S. *Phys. Status Solid B* **2000**, *217*, 357.
- (3) Elstner, M.; Frauenheim, T.; Suhai, S. *J. Mol. Struct.: THEOCHEM* **2003**, *632*, 29.
- (4) Elstner, M. *Theor. Chem. Acc.* **2006**, *116*, 316.

- (5) Cui, Q. *Theor. Chem. Acc.* **2006**, *116*, 51.
- (6) Riccardi, D.; Schaefer, P.; Yang, Y.; Yu, H.; Ghosh, N.; Prat-Resina, X.; Konig, P.; Li, G.; Xu, D.; Guo, H.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2006**, *110*, 6458.
- (7) Kruger, T.; Elstner, M.; Schifffels, P.; Frauenheim, T. *J. Chem. Phys.* **2005**, *122*, 114110.
- (8) Sattelmeyer, K. W.; Tirado-Rives, J.; Jorgensen, W. *J. Phys. Chem. A* **2006**, *110*, 13551.
- (9) Otte, N.; Scholten, M.; Thiel, W. *J. Phys. Chem. A* **2007**, *111*, 5751.
- (10) Thiel, W. *Adv. Chem. Phys.* **1996**, *93*, 703.
- (11) Elstner, M.; Jalkanen, K. J.; Knapp-Mohammady, M.; Frauenheim, T.; Suhai, S. *Chem. Phys.* **2001**, *263*, 203.
- (12) Elstner, M.; Hobza, P.; Frauenheim, T.; Suhai, S.; Kaxiras, E. *J. Chem. Phys.* **2001**, *114*, 5149.
- (13) Mohle, K.; Hofmann, H. J.; Thiel, W. *J. Comput. Chem.* **2001**, *22*, 509.
- (14) Khandogin, J.; Musier-Forsyth, K.; York, D. M. *J. Mol. Biol.* **2003**, *330*, 993.
- (15) Khandogin, J.; York, D. M. *Proteins* **2004**, *56*, 724.
- (16) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227.
- (17) Gao, J.; Truhlar, D. G. *Annu. Rev. Phys. Chem.* **2002**, *53*, 467.
- (18) Shurki, A.; Warshel, A. *Adv. Protein Chem.* **2003**, *66*, 249.
- (19) Friesner, R. A.; Guallar, V. *Annu. Rev. Phys. Chem.* **2005**, *56*, 389.
- (20) Shaik, S.; Kumar, D.; de Visser, S. P.; Altun, A.; Thiel, W. *Chem. Rev.* **2005**, *105*, 2279.
- (21) Gregersen, B. A.; Lopez, X.; York, D. M. *J. Am. Chem. Soc.* **2004**, *126*, 7504.
- (22) Nam, K.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 2.
- (23) Weber, W.; Thiel, W. *Theor. Chem. Acc.* **2000**, *103*, 495.
- (24) Tubert-Brohman, I.; Guimaraes, C. R. W.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2005**, *1*, 817.
- (25) Bernal-Uruchurtu, M.; Ruiz-Lopez, M. *Chem. Phys. Lett.* **2000**, *330*, 118.
- (26) Bernal-Uruchurtu, M. I.; Martins-Costa, M. T. C.; Millot, M. F. R.-L. *C. J. Comput. Chem.* **2000**, *21*, 572.
- (27) Sattelmeyer, K. W.; Tubert-Brohman, I.; Jorgensen, W. L. *J. Chem. Theory Comput.* **2006**, *2*, 413.
- (28) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 1275.
- (29) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486.
- (30) Giese, T. J.; Gregersen, B. A.; Liu, Y.; Nam, K.; Mayaan, E.; Moser, A.; Range, K.; Nieto Faza, O.; Silva Lopez, C.; Rodriguez de Lera, A., et al. *J. Mol. Graphics Modell.* **2006**, *25*, 423.
- (31) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2005**, *123*, 164108.
- (32) Wu, Q.; Yang, W. T. *J. Chem. Phys.* **2002**, *116*, 515.
- (33) Grimme, S. *J. Comput. Chem.* **2004**, *25*, 1463.
- (34) Liu, H. Y.; Elstner, M.; Kaxiras, E.; Frauenheim, T.; Hermans, J.; Yang, W. T. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 484.
- (35) Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell*, 3rd ed.; Garland Publishing: New York, 1994.
- (36) Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W. H. Freeman: New York, 1999.
- (37) Koenig, P.; Ghosh, N.; Hoffman, M.; Elstner, M.; Tajkhorshid, E.; Frauenheim, T.; Cui, Q. *J. Phys. Chem. A* **2006**, *110*, 548.
- (38) Reed, A. E.; Curtiss, L. A.; Weinhold, F. *Chem. Rev.* **1988**, *88*, 899.
- (39) Head-Gordon, M. *J. Phys. Chem.* **1996**, *100*, 13213.
- (40) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785.
- (41) Becke, A. D. *Phys. Rev. A* **1988**, *38*, 3098.
- (42) Becke, A. D. *J. Chem. Phys.* **1993**, *98*, 5648.
- (43) Hadzi, D.; Koller, J. *Theoretical Treatment of Hydrogen Bonding*; Wiley: New York, 1997.
- (44) Voityuk, A. A.; Bliznyuk, A. A. *Theor. Chim. Acta* **1987**, *72*, 223.
- (45) Jug, K.; Geudtner, G. *J. Comput. Chem.* **1993**, *14*, 639.
- (46) Han, W. G.; Elstner, M.; Jalkanen, K. J.; Frauenheim, T.; Suhai, S. *Int. J. Quantum Chem.* **2000**, *78*, 459.
- (47) Riccardi, D.; Li, G.; Cui, Q. *J. Phys. Chem. B* **2004**, *108*, 6467.
- (48) Range, K.; Riccardi, D.; Elstner, M.; Cui, Q.; York, D. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3070.
- (49) Riccardi, D.; Schaefer, P.; Cui, Q. *J. Phys. Chem. B* **2005**, *109*, 17715.
- (50) Porezag, D.; Frauenheim, T.; Kohler, T.; Seifert, G.; Kaschner, R. *Phys. Rev. B* **1995**, *51*, 12947.
- (51) Perdew, J. P.; Burke, K.; Ernzerhof, M. *Phys. Rev. Lett.* **1996**, *77*, 3865.
- (52) Janak, J. F. *Phys. Rev. B* **1978**, *18*, 7165.
- (53) Elstner, M. *J. Phys. Chem. A* **2007**, *111*, 5614.
- (54) Zhou, H. Y.; Tajkhorshid, E.; Frauenheim, T.; Suhai, S.; Elstner, M. *Chem. Phys.* **2002**, *277*, 91.
- (55) Elstner, M.; Cui, Q.; Munih, P.; Kaxiras, E.; Frauenheim, T.; Karplus, M. *J. Comput. Chem.* **2003**, *24*, 565.
- (56) Klopman, G. *J. Am. Chem. Soc.* **1964**, *86*, 4550.
- (57) Ohno, K. *Theor. Chim. Acta.* **1964**, *2*, 219.
- (58) Politzer, P.; Murray, J. S.; Lane, P. *J. Comput. Chem.* **2003**, *24*, 505.
- (59) Parr, R. G.; Yang, W. T. *Density-Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
- (60) Klopper, W.; van Duijneveldt-van de Rijdt, J. G. C. M.; van Duijneveldt F. B. *Phys. Chem. Chem. Phys.* **2000**, *2*, 2227.
- (61) 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 B.05; Gaussian, Inc.: Wallingford, CT, 2004.
- (62) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. *J. Comput. Chem.* **1983**, *4*, 187.
- (63) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.
- (64) Curtiss, L. A.; Raghavachari, K.; Redfern, P. C.; Rassolov, V.; Pople, J. A. *J. Chem. Phys.* **1998**, *109*, 7764.
- (65) Baboul, A. G.; Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. *J. Chem. Phys.* **1999**, *110*, 7650.
- (66) Xantheas, S.; Burnham, C. J.; Harrison, R. J. *J. Chem. Phys.* **2002**, *116*, 1493.
- (67) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer, H. F., III; Quack, M. *J. Chem. Phys.* **2002**, *116*, 690.
- (68) Bondar, A. N.; Fischer, S.; Smith, J. C.; Elstner, M.; Suhai, S. *J. Am. Chem. Soc.* **2004**, *126*, 14668.
- (69) Hobza, P.; Kabelac, M.; Sponer, J.; Mejzlik, P.; Vondrasek, J. *J. Comput. Chem.* **1997**, *18*, 1136.