JCTC Journal of Chemical Theory and Computation

Letter

pubs.acs.org/JCTC

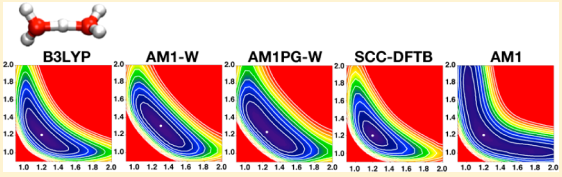
# Development of Semiempirical Models for Proton Transfer Reactions in Water

Shihao Wang,[†] Laurent MacKay,[†,‡] and Guillaume Lamoureux*[,†,‡]

[†]Department of Chemistry and Biochemistry and Centre for Research in Molecular Modeling (CERMM) and [‡]Department of Physics, Concordia University, Montréal, Canada

Supporting Information

**ABSTRACT:** This letter presents a method for the parametrization of semiempirical models for proton transfer reactions in water clusters. Two new models are developed: AM1-W, which is a reparameterization of the classic AM1 model, and AM1PG-W, which is a modified AM1-like model including a pairwise correction to the core repulsion function. Both models show good performance on hydrogen-bonding energies and on proton transfer energy profiles, which are of great importance for proton transfer reactions in large water clusters and in proteins. The parametrization method introduced is general and can be used to develop any other system-specific semiempirical models.




## 1. INTRODUCTION

Proton transfer is one of the most common reactions in chemistry and plays a fundamental role in many biological processes.[1,2] It is an essential step in numerous enzymatic reactions[3] and in proton-coupled membrane transport.[4]

Proton transfer in biological systems has been studied with a number of computational methods, including the commonly used hybrid quantum mechanical/molecular mechanical (QM/MM) approach.[5] In a large number of studies (see refs 6−8 for example) a QM/MM representation is used simply to obtain minimum-energy structures and energy profiles along the adiabatic reaction surface, while in other studies (see refs 9−11 for example) QM/MM molecular dynamics (MD) simulations, which incorporate thermal fluctuations, are used to estimate reaction free energies and to achieve large-scale relaxation of the protein and the solvent. In QM/MM MD simulations, the QM region is usually treated with Density Functional Theory (DFT) or with semiempirical (SE) quantum models. DFT methods can be highly accurate[12] but are computationally demanding and not easily amenable to large-scale simulations. On the other hand, SE models have much lower computational cost but are usually not as versatile.[13−16]

SE models are mostly reliable for molecular systems similar to those for which they are specifically parametrized. For instance, popular models such as AM1,[17] PM3,[18] and MNDO/d[19] were parametrized from a number of small organic molecules, but they may not perform well when applied to biological systems.[20] Notably, it has been found that some models, such as AM1, are unable to reproduce energies and conformations of several hydrogen-bonded clusters,[13−15,21−25] and display large errors in their prediction of reaction barriers.[16,26]

Despite their limited transferability, SE models can be reparameterized for specific systems or classes of reactions.[27−37] For the purpose of rapidly generating system-specific or reaction-specific SE models, it is therefore desirable to develop a general method that can calibrate any type of SE model to reproduce any desired properties of a set of molecular structures.

A number of SE models modified to better describe hydrogen bonding have been proposed over the years.[38−42] More recently, it has been shown that standard SE models could be reparameterized to improve their performance for proton transfer reactions in water.[31,36] In this letter, we report the development of semiempirical models that more accurately describe proton transfer reactions in water clusters. We examine the transferability of the models to establish, notably, whether they can be used in protein environments to describe water clusters involved in biological proton-transfer or proton-transport processes.[1,7−11,43−47]

## 2. METHODS

The SE models are parametrized to optimally reproduce a set of properties obtained from high-level DFT calculations on small water clusters (monomers, dimers, and trimers). This *training set* includes cluster geometries, proton affinities, hydrogen-bonding energies, reaction energies, activation energies ("reaction barriers"), and proton transfer energy profiles. The models are parametrized in rounds, one chemical element at a time, until no further improvement is observed. The final models are assessed using a second set of properties—the *testing set*—calculated from larger water clusters (tetramers, pentamers, and hexamers).

**2.1. Preparation of Training Set and Testing Set.** *2.1.1. Systems and Properties Selected.* The training set contains all stable conformations of the water monomer, dimer, and trimer in their neutral, protonated, and deprotonated

**Received:** February 25, 2014
**Published:** July 16, 2014

   dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

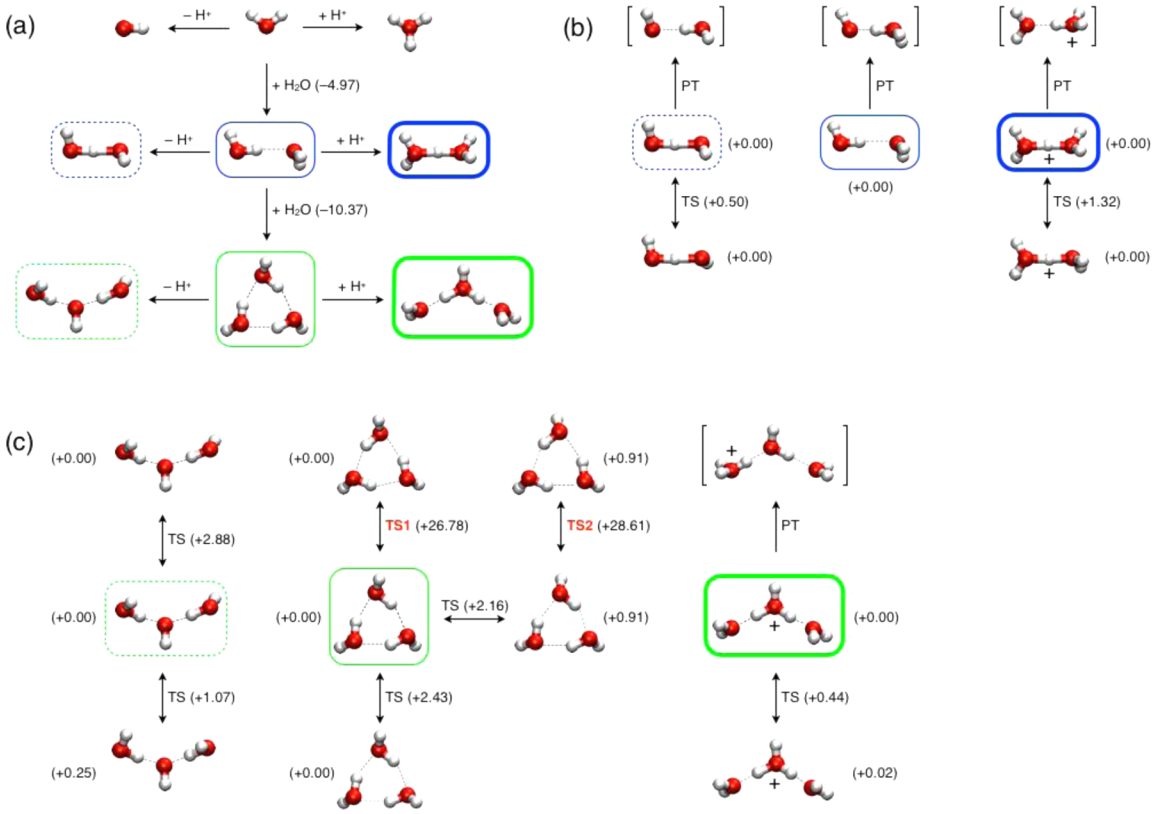



**Figure 1.** Water clusters in the training set. Panel a shows the lowest-energy conformer for each cluster, and panels b and c show how each structure from panel a (identified by a color frame) is related to the higher-energy conformers. The training set contains the following properties: (1) All nonequivalent structures in panels a−c, including transition states (labeled TS, structures not shown) and intermediate structures along proton transfer profiles (labeled PT, only final structure shown). (2) Deprotonation energies (panel a, labeled either $-H^+$ or $+H^+$). (3) Hydrogen bonding energies (panel a, labeled $+H_2O$). (4) Reaction energies between nonequivalent structures connected by a double arrow. (5) Activation energies between transition states (labeled TS) and both the reactant and product structures (if they are nonequivalent). (6) Proton transfer energy profiles (labeled PT).

states, as well as the transition structures (TS) connecting them (see Figure 1). Besides structures, the training set includes reaction energies and activation energies for all structural changes involving a single TS. It also includes proton affinities, hydrogen-bonding energies, and proton transfer energy profiles.

The testing set contains a selection of tetramer, pentamer, and hexamer structures (see Figure 2). Similarly to the training set, each cluster has neutral, protonated, and deprotonated states. The lowest-energy conformation of each cluster is shown in Figure 2a and related, higher-energy structures are shown in Figure 2b−d.

*2.1.2. Training Set and Testing Set Generation.* All properties of the training and testing sets are calculated from fully optimized cluster geometries.

In Figure 1a, each horizontal arrow represents either a deprotonation ($-H^+$) or a protonation ($+H^+$) and corresponds to a deprotonation energy $E_D$ in the training set. $E_D$ is equivalent to a proton affinity and is calculated for the protonated and neutral water clusters as follows:[48]

$$E_D = -\Delta E - \Delta\mathrm{ZPE} + \frac{5}{2}RT \tag{1}$$

where $E$ is the electronic energy and ZPE is the zero-point energy. $\Delta E$ is the energy difference between the states before and after the deprotonation. When calculating proton affinities using semiempirical models, the experimental heat of formation of proton (367.2 kcal/mol)[49] is used. Each vertical arrow represents the addition of one water molecule ($+H_2O$), and corresponds to a hydrogen-bonding energy $E_{HB}$ calculated as follows:

$$E_{HB} = E(AW) - E(A) - E(W) \tag{2}$$

where $E(A)$, $E(W)$, and $E(AW)$ are energies of complex $A$, water molecule $W$, and hydrogen-bonded complex $AW$, respectively. Because the addition of a water molecule creates two new hydrogen bonds in most clusters considered, $E_{HB}$ typically amounts to twice the energy of a single hydrogen bond.

In Figure 1b and c, each double-headed arrow (labeled TS) represents an isomerization between two stable conformers, and corresponds to *one* reaction energy $E_R$ and *two* activation energies $E_A$ (from reactant to TS and from product to TS). This places an additional emphasis on activation energies compared to reactions energies. When the reactant and product are structurally equivalent—either by being superimposable or by being mirror images of one another—the reaction energy (which is zero by definition) is ignored and the activation energy is considered only once. Each single-headed arrow (labeled PT) represents a proton transfer reaction, corresponding to a proton transfer energy profile $E_{PT}$. Unlike other reactions from Figures 1 and 2, the PT reactions do not lead to stable products in gas phase. They are nevertheless essential for the development of the models because they represent processes that may have a stable product in solution or in a


dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

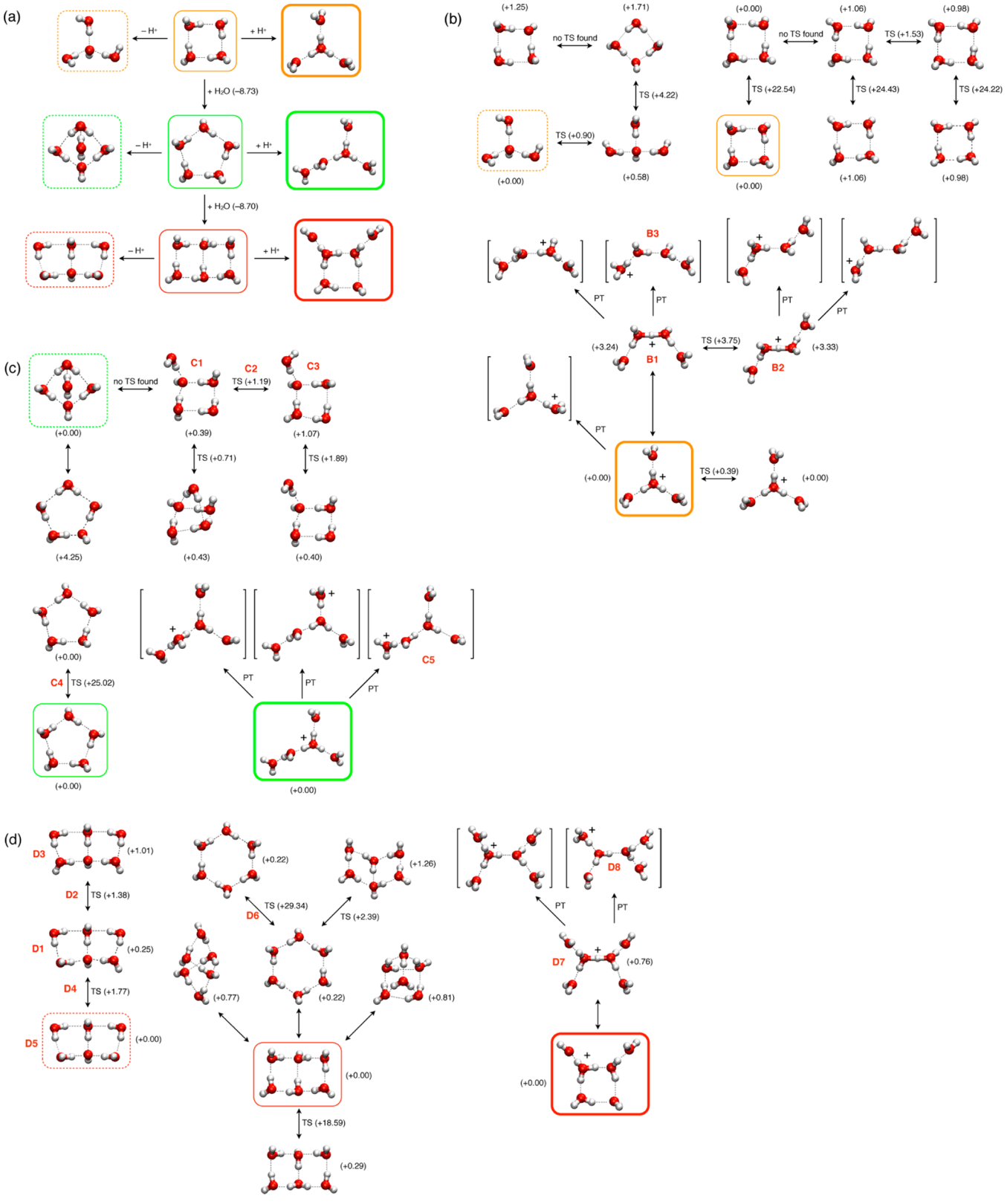



**Figure 2.** Water clusters in the testing set. Panel a shows the lowest-energy conformer for each cluster and panels b to d show how each structure from panel a (identified by a color frame) is related to the higher-energy conformers. The properties in the testing set are indicated as in Figure 1 (see Figure 1 caption for details).

protein environment, under large transient or permanent electric fields. In order to get proton transfer energy profiles in the gas phase, constrained energy scans are used. Specifically, the O···H···O angle is kept at 180° and the O(acceptor)···H distance is scanned from 1.7 to 0.9 Å in steps of −0.2 Å. The energy of each of the five constrained structures is calculated as

$$E_{PT}(k) = E(k) - E(0) \quad (3)$$


dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

where $E(k)$ is the energy of the $k$th constrained structure and $E(0)$ is the energy of the stable reactant structure.

*2.1.3. DFT and Semiempirical Calculations.* Benchmark DFT calculations are performed using Gaussian 09[50] at the B3LYP/6-311++G(2d,2p) level of theory, which has been shown to perform well for water clusters[51] and is much less time-consuming than accurate *ab initio* methods such as MP2, especially for water hexamers. All stable structures and transition states are confirmed with frequency calculations. For transition states, the negative-frequency mode is followed to confirm that they are connecting the correct reactant and product states. The energies are not corrected for BSSE.

Semiempirical calculations are performed using an in-house version of MOPAC7.[52] Geometry optimization is performed with a modified[52] BFGS method[53−56] without any constraints, using the DFT-optimized structures as initial guesses. Transition state searching is based on the dimer method,[57] using the DFT-optimized transition state structures as initial guesses. For MOPAC7 geometry optimizations, the termination criterion GNORM is set to 1.0 kcal/mol·Å during the parametrization and to 0.01 kcal/mol·Å for the final assessment of the models. When GNORM is 0.01 kcal/mol·Å, the optimal structure closest to the reference DFT geometry is found by restricting the search space to stable configurations for which no atom deviates from its reference position by a distance greater than $d$, and by increasing $d$ until a stable configuration is found. The following penalty energy is added on every atom $a$:

$$E_p = W_p(|\mathbf{r}_a^{model} - \mathbf{r}_a^{ref}|^2 - d^2) \quad \text{if} \quad |\mathbf{r}_a^{model} - \mathbf{r}_a^{ref}| > d$$
$$= 0 \quad \text{if} \quad |\mathbf{r}_a^{model} - \mathbf{r}_a^{ref}| \le d \tag{4}$$

where $W_p$ is a force constant set to 50 kcal/mol·Å$^2$, $|\mathbf{r}_a^{model} - \mathbf{r}_a^{ref}|$ is the distance between the model and reference positions of atom $a$, and $d$ is the maximum tolerated deviation. The penalty term $E_p$ is only applied when its value is positive. A first geometry optimization is performed with $d$ set to 0.5 Å. If $E_p$ is nonzero for the optimal structure found, or if the structural difference is too close to $d$ (greater than $d$ − 0.1 Å), parameter $d$ is increased to 1 Å, then to 2 Å, then to 3 Å. If still no optimal structure is found, the penalty term is removed and a free optimization is performed. This process is used for all geometry optimizations except for those in proton transfer energy scans. The validity of the optimal geometries obtained using the MOPAC7 BFGS algorithm is confirmed by repeating the geometry optimizations of the final SE models with CP2K,[58] which also uses the BFGS algorithm.

For SE calculations, all "$E$" values mentioned in section 2.1.2 correspond to the total energy of the optimized system (called "$E_{eq}$" in MOPAC), and ZPE values are calculated from the sum of vibration frequencies.

*2.1.4. Error Function.* The overall error function used in the parametrization is as follows:

$$\chi = W_s \sum_i \mathrm{MSD}(i) + \sum_i |E_D^{model}(i) - E_D^{ref}(i)| + \sum_i |E_{HB}^{model}(i) - E_{HB}^{ref}(i)| + \sum_i |E_R^{model}(i) - E_R^{ref}(i)| + \sum_i |E_A^{model}(i) - E_A^{ref}(i)| + W_{PT} \sum_i \left[ \frac{1}{n_{PT}} \sum_{k=1}^{n_{PT}} |E_{PT}^{model}(i,k) - E_R^{ref}(i,k)| \right] \tag{5}$$

Superscript "model" represents SE calculations, and superscript "ref" represents DFT calculations. MSD is the mean square deviation of the model (SE) structure relative to the reference (DFT) structure. Each model cluster is aligned with the reference structure and the MSD is calculated as follows:

$$\mathrm{MSD} = \frac{1}{n} \sum_{a=1}^{n} |\mathbf{r}_a^{model} - \mathbf{r}_a^{ref}|^2 \tag{6}$$

where $n$ is the number of atoms in the cluster and $|\mathbf{r}_a^{model} - \mathbf{r}_a^{ref}|$ is the difference between the model and reference structures on position of atom $a$. MSD values are calculated for all nonequivalent structures, with the exception of $OH^-$, considered too small to be meaningfully compared. For each "PT" reaction, a single MSD value is used, corresponding to the average of the individual MSD values of the five constrained "PT" structures. $W_s$ is a weighting factor set to 33 kcal/mol·Å$^2$. This value is chosen to achieve a balanced compromise between structures and energies. The remaining terms in eq 5 correspond to deprotonation energies ($E_D$), hydrogen-bond energies ($E_{HB}$), reaction energies ($E_R$), activation energies ($E_A$), and proton transfer profile energies ($E_{PT}$). $n_{PT}$ is the number of points in each proton transfer profile, which is set to 5 in our parametrization scheme. $W_{PT}$ is a weighting factor for proton transfer, set to 2 to give about as much importance to a proton transfer profile as to any other reaction. Error function $\chi$ depends on the parameters of the SE model through the "model" quantities. For the training set illustrated in Figure 1, function $\chi$ is composed of 24 MSD terms, 6 $E_D$ terms, 2 $E_{HB}$ terms, 3 $E_R$ terms (6 are ignored), 12 $E_A$ terms, and 20 $E_{PT}$ terms (5 for each of the 4 PT energy profiles).

**2.2. Parameterization of Semiempirical Models.**

*2.2.1. Reparameterization of the AM1 Model.* The original AM1 model for oxygen and hydrogen is reparameterized to improve its performance for hydrogen-bonded water clusters. The new model, called AM1-W ("AM1 for water clusters"), has exactly the same formalism as AM1.

*2.2.2. Extension of the AM1 Model.* In order to improve the model's performance on proton transfer profiles, and inspired by the MNDO/HB[38] and PDDG[59] methods, we modify the AM1 model by adding a pairwise Gaussian function to the core repulsion function (CRF):

$$\mathrm{CRF}(a,b) = \mathrm{CRF}_{AM1}(a,b) + P_{ab} e^{-\beta_{ab}(R_{ab} - C_{ab})^2} \tag{7}$$

where $P_{ab}$, $\beta_{ab}$, and $C_{ab}$ are parameters for all pairs of elements that form covalent bonds, and $R_{ab}$ is the distance between atoms $a$ and $b$. For water clusters, the only such combination is OH. (The other two combinations, OO and HH, do not form covalent bonds.) Therefore, three new parameters are introduced: $P_{OH}$, $\beta_{OH}$, and $C_{OH}$. This new model is called AM1PG-W ("AM1 with pairwise Gaussians for water clusters").

*2.2.3. Parameterization Procedure.* Parameterization is performed using a genetic algorithm approach, which has been used for model parametrization in many previous studies.[42,60−72] A parallel version of the PIKAIA program[73] is used in the present work.

The AM1-W and AM1PG-W model parameters are allowed to change by up to ±50% of the original AM1 values. Since the pairwise Gaussian parameters of the AM1PG-W model have no equivalent in the AM1 model, their starting values are set by minimizing the error function while keeping all other parameters at their original AM1 values. The resulting starting

Table 1. Parameters of Original AM1 Model and of AM1-W and AM1PG-W Models (This Work)[a]

| | AM1 | | AM1-W | | AM1PG-W | |
|---|---|---|---|---|---|---|
| param. | H | O | H | O | H | O |
| $U_{ss}$ (ev) | −11.396427 | −97.83000 | −12.1942 | −128.1573 | −10.9406 | −115.4394 |
| $U_{pp}$ (ev) | | −78.26238 | | −79.8276 | | −85.3060 |
| $Z_s$ (au) | 1.188078 | 3.108032 | 1.0574 | 3.4188 | 0.9980 | 3.1702 |
| $Z_p$ (au) | | 2.524039 | | 2.5493 | | 2.9279 |
| $B_s$ (eV) | −6.173787 | −29.27277 | −5.0625 | −36.2982 | −4.7538 | −35.4201 |
| $B_p$ (eV) | | −29.27277 | | −38.9328 | | −38.3473 |
| $G_{ss}$ (eV) | 12.848 | 15.42 | 14.7752 | 21.4338 | 14.9037 | 15.5742 |
| $G_{sp}$ (eV) | | 14.48 | | 18.6792 | | 15.7832 |
| $G_{pp}$ (eV) | | 14.52 | | 17.5692 | | 15.1008 |
| $G_{p2}$ (eV) | | 12.98 | | 9.8648 | | 13.7588 |
| $H_{sp}$ (eV) | | 3.94 | | 2.8368 | | 4.3734 |
| $\alpha$ (Å$^{-1}$) | 2.882324 | 4.455371 | 2.7670 | 5.7029 | 2.8535 | 5.4356 |
| $K_1$ (eV) | 0.122796 | 0.280962 | 0.0847 | 0.3484 | 0.1621 | 0.1601 |
| $L_1$ (Å$^{-1}$) | 5.0000 | 5.0000 | 6.6995 | 3.0000 | 4.5000 | 3.4100 |
| $M_1$ (Å) | 1.2000 | 0.847918 | 1.0320 | 1.0005 | 1.0080 | 1.1193 |
| $K_2$ (eV) | 0.00509 | 0.08143 | 0.0038 | 0.0497 | 0.0054 | 0.0489 |
| $L_2$ (Å$^{-1}$) | 5.0000 | 7.0000 | 6.4500 | 6.0900 | 3.2500 | 9.0300 |
| $M_2$ (Å) | 1.8000 | 1.445071 | 1.6020 | 1.9364 | 1.0260 | 1.9075 |
| $K_3$ (eV) | −0.018336 | | −0.0260 | | −0.0193 | |
| $L_3$ (Å$^{-1}$) | 2.0000 | | 2.0400 | | 2.5600 | |
| $M_3$ (Å) | 2.1000 | | 1.1760 | | 1.1550 | |
| $P_{OH}$ (eV) | | | | | 0.0147 | |
| $\beta_{OH}$ (Å$^{-2}$) | | | | | 13.1790 | |
| $C_{OH}$ (Å) | | | | | 1.0282 | |

[a]The last three rows are pairwise Gaussian parameters as defined in eq 7 for model AM1PG-W.

values are 0.00905 eV for $P_{OH}$, 9.55 Å$^{-2}$ for $\beta_{OH}$, and 0.970 Å for $C_{OH}$.

A single PIKAIA run simulates the evolution of 100 individuals for 300 generations (see Supporting Information Table S1 for details). Each individual represents a set of SE model parameters, which fitness is evaluated as $1/(\chi+1)$ (using eq 5). PIKAIA initially generates 100 individuals randomly distributed inside the search space, such that none of the original AM1 parameters are preserved. Each new generation is obtained by genetic recombination (crossover) of pairs of individuals selected from the previous generation, followed by random mutation. The individual that has yielded the highest fitness value during any of the 300 generations is chosen as the final parameter set for the run. (See Supporting Information Figure S1 for a graphical representation of the workflow.)

In order to reduce the dimensionality of the parameter space for PIKAIA and speed up the convergence, parameters are optimized in groups instead of all at once. One full round of optimization consists of a PIKAIA run to optimize the 18 oxygen parameters, followed by a run to optimize the 14 hydrogen parameters (followed, for the AM1PG-W model, by a run to optimize the 3 pairwise Gaussian parameters). In the parametrization of both AM1-W and AM1PG-W models, three rounds are required to converge the error value. A fourth round does not result in any improvement.

## 3. RESULTS AND DISCUSSION

**3.1. Performance in the Training Set.** The parameters for the new models, AM1-W and AM1PG-W, are reported in Table 1. They deviate from the original AM1 parameters by 21% on average, with a maximum deviation of 45% (within the ±50% margins set initially). In Figure 3, the performance of the AM1-W and AM1PG-W models is compared with that of the AM1,[17] RM1,[74] and PM3[18] models. As expected, the new models perform better for all properties in the training set, especially for structures, proton affinities ($E_D$), and activation energies ($E_A$). After a reoptimization of all structures using GNORM = 0.01 kcal/mol·Å (instead of 1 kcal/mol·Å)—following the optimization scheme described in section 2.1.3—the error function values are (in kcal/mol) 5.9 for AM1, 3.0 for RM1, 4.5 for PM3, 1.2 for AM1-W, and 1.0 for AM1PG-W.

As shown in Figure 3a, the original AM1 model yields large errors in structures, proton affinities, and activation energies. As previously documented for neutral water clusters,[75−77] many dimer and trimer structures form incorrect, bifurcated hydrogen bonds. The two most poorly reproduced structures, with an RMSD greater than 0.67 Å (that is, with a "structure" error higher than 15 kcal/mol), are the protonated trimer $H_2O{\cdot}H_3O^+{\cdot}H_2O$ and its corresponding "PT" structure $H_3O^+{\cdot}H_2O{\cdot}H_2O$ (see Figure 1c), which for AM1 adopt conformations in which the water molecules form bifurcated hydrogen bonds with the hydronium fragment. The largest errors on the activation energy are above 30 kcal/mol and correspond to proton transfer reactions in the neutral trimer cluster (TS1 and TS2 in Figure 1c). For instance, the activation energy associated with TS1 is 26.78 kcal/mol according to the benchmark DFT calculations but is 65.61 kcal/mol for the AM1 model—an overestimation of 38.83 kcal/mol. The error on the activation energy associated with TS2 is even larger (41.42 kcal/mol).

The RM1 and PM3 models perform better than AM1 on most properties (see Figures 3b and c) but, similar to AM1, cannot predict cluster properties with consistent accuracy. For instance, while RM1 reproduces the deprotonation energies of $H_3O^+$, $H_5O_2^+$, and $H_7O_3^+$ within 5 kcal/mol or less, it


dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

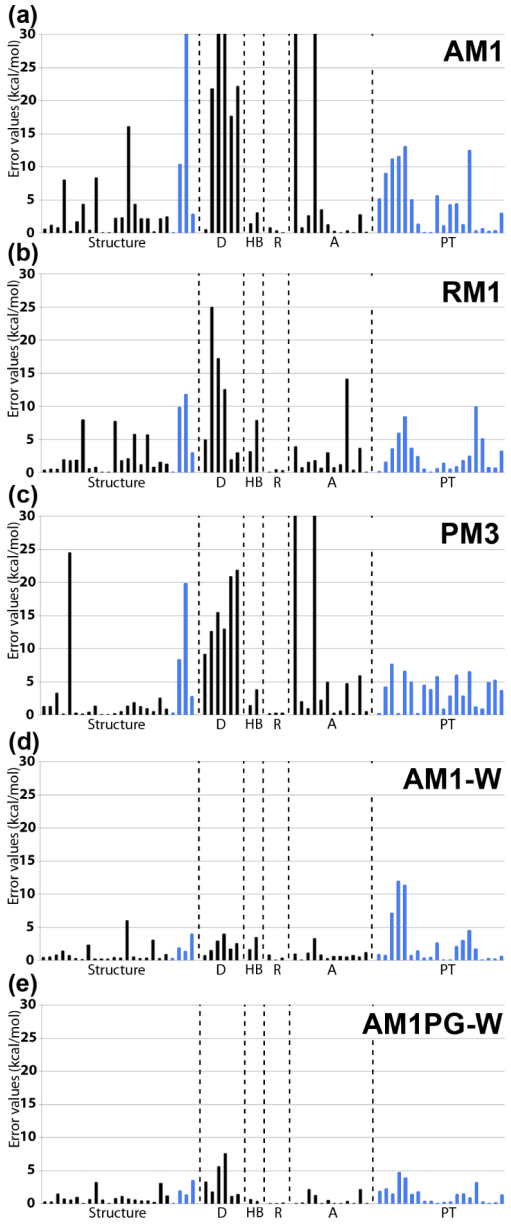



**Figure 3.** Performance of different SE models for each property in the training set. Panels a, b, c, d, and e are showing the error values of AM1, RM1, PM3, AM1-W, and AM1PG-W models, respectively. Structure errors are defined in eq 6. (A structure error of 5 kcal/mol corresponds to an RMSD of 0.39 Å.) Labels D, HB, R, A, and PT represent errors on deprotonation energies (or proton affinities), hydrogen-bonding energies, reaction energies, activation energies, and proton transfer energy profiles, respectively. The error bars associated with proton transfer profiles are marked in blue.

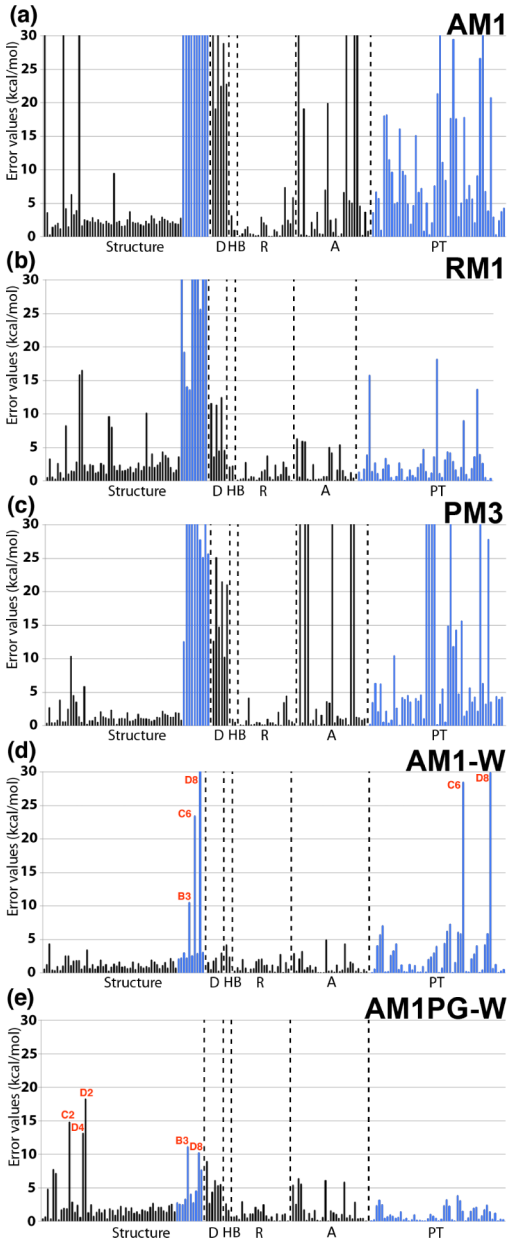



**Figure 4.** Performance of different SE models for each property in the testing set (see Figure 3 caption for details). For RM1 and PM3 models, some of the structures cannot be optimized by MOPAC and are therefore not included in panels b and c.

overestimates the deprotonation energies of $H_2O$, $(H_2O)_2$, and $(H_2O)_3$ by 12−25 kcal/mol.

As expected, the AM1-W and AM1PG-W models show good performance on small clusters. Errors are consistently small, and almost all errors are below 5 kcal/mol, as shown in Figure 3d and e. The AM1PG-W model performs noticeably better than AM1-W for proton transfer energies.

**3.2. Performance in the Testing Set.** Since the training set only contains water clusters up to trimers, it is essential to check that the optimized models are transferable to larger water clusters.

The performance of the models in the testing set is compared in Figure 4. The average values of the individual error contributions from Figure 4 are (in kcal/mol) 13.4 for AM1, 4.6 for RM1, 32.5 for PM3, 1.9 for AM1-W, and 2.0 AM1PG-W. As shown in Figure 4a, the AM1 model performs relatively poorly for structures, activation energies, and proton transfer profiles. An error of 30 kcal/mol on a structure represents a 0.91-Å root-mean-square displacement in atomic positions, which indicates incorrect overall shape and hydrogen-bonding structure. The hydrogen-bonding energies and reaction energies are well predicted for all models, likely because they remain small (around 9 kcal/mol for $E_{HB}$ and at most 4.2 kcal/mol for $E_R$).

The RM1 model displays good performance for hydrogen-bonding energies, reaction energies, activation energies, and for


dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

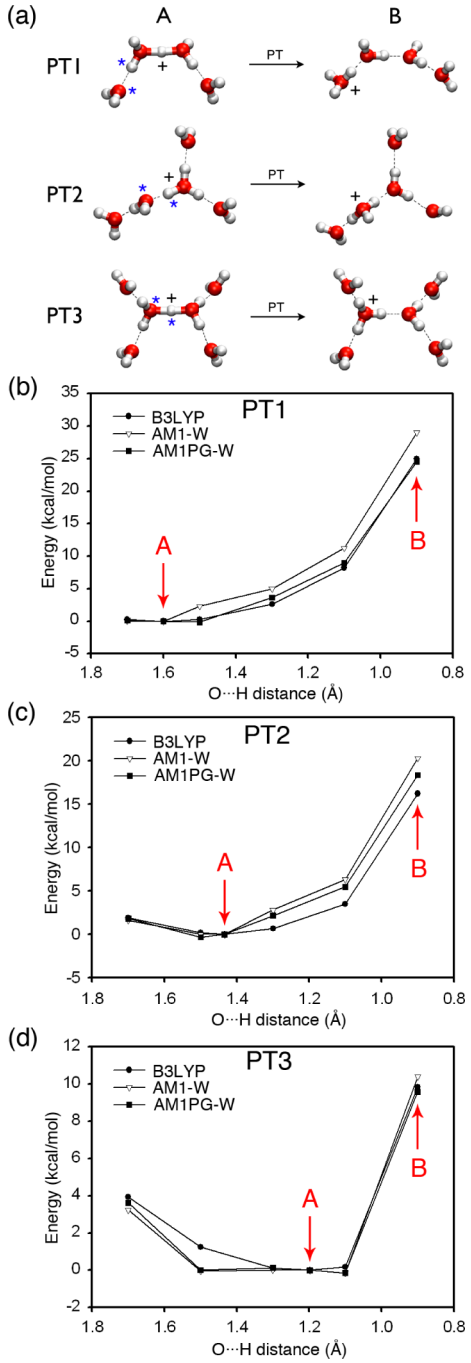



**Figure 5.** Proton transfer energy profiles calculated using the AM1-W and AM1PG-W models, compared to B3LYP benchmark values. Panel a shows the three proton transfer reactions presented, where structure A is the benchmark minimum-energy conformation (before proton transfer is induced) and structure B is the benchmark conformation after the transfer (once the proton is at a distance of 0.9 Å from the acceptor oxygen atom). Panels b, c, and d show the energies as a function of the distance between O and H atoms (identified with blue stars in panel a). Note that the O−H distance decreases from left to right.

most of the proton transfer profiles, but is inaccurate at predicting some of the structures, especially the constrained structures obtained from the proton transfer energy scans. In Figure 4b, errors above 20 kcal/mol (that is, RMSD above 0.78 Å) all correspond to structures in proton transfer profiles.

The PM3 model has better performance in structure prediction (Figure 4c), but shows large errors on some constrained structures in proton transfer energy scans, and in the proton transfer energies and barriers. In Figure 4c, all the large errors (above 30 kcal/mol) in the "A" section correspond to overestimated proton transfer energy barriers.

By comparison, the AM1-W model gives small errors in all properties, except for five outliers marked B3, C6 (twice), and D8 (twice) (see Figure 4d). Label B3 corresponds to errors in the structures of B1 → B3 proton transfer profile (see Figure 2b). Because of the small energy difference and low energy barrier between the *cis* (B1) and *trans* (B2) states, the conformer undergoes a *cis*-to-*trans* transition during the proton transfer when AM1-W model is used. Label D8 corresponds to errors in the structures of D7 → D8 proton transfer profile (see Figure 2d). Conformer D8 is highly unstable and, for the AM1-W model, collapses into a structure in which the hydronium fragment is coordinated by three water molecules instead of one. This structural change results in the large error in proton transfer energy marked D8 in Figure 4d. C6 corresponds to errors in the C5 → C6 proton transfer profile. The two largest errors in activation energies are from proton transfer reactions (C4 and D6 in Figure 2d): 25.00 and 29.12 kcal/mol in benchmark, predicted to be respectively 29.95 and 33.51 kcal/mol by AM1-W.

As shown in Figure 4e, the AM1PG-W model yields a few larger errors in structure compared to AM1-W but smaller errors in proton transfer profiles. The error bar labeled C2 corresponds to the transition state between C1 and C3 (see Figure 2c). When AM1PG-W is used, the out-of-plane water molecule moves toward the right, and acts as a hydrogen-bond acceptor to the water molecule on the right. Error bar D2 corresponds to the transition state between D1 and D3 (see Figure 2d). When AM1PG-W is used, one of the two rings is distorted and more open (although no hydrogen bond is broken). The RMSD of structure D2 is 0.74 Å, corresponding to an error of 18.3 kcal/mol. Error bar D4 corresponds to the transition state between D1 and D5, which shows ring distortion similar to that for structure D2. Error bars B3 and D8 from Figure 4e have already been discussed in the context of Figure 4d. The five largest errors in activation energies (larger than 5 kcal/mol) correspond to proton transfer reactions, for which energy barriers are in the 22−29 kcal/mol range for the benchmark data but in the 28−35 kcal/mol range for the AM1PG-W model.

With its additional pairwise Gaussian function centered at $C_{OH}$ = 1.0282 Å (with a standard deviation $\sigma$ = 0.195 Å), the AM1PG-W model gives more accurate intramolecular O−H distances than AM1-W. Over all clusters from the testing set, the average O−H bond lengths are 0.977 Å for DFT, 0.956 Å for AM1-W, and 0.972 Å for AM1PG-W. The average intermolecular O···H distances (distributed between 1.5 and 2.1 Å) are 1.768, 1.733, and 1.755 Å, respectively. The average O···O distances between hydrogen-bonding water molecules (between oxygen atoms forming an O−H···O motif) are 2.709, 2.680, and 2.677 Å, respectively, and the average H···H distances (between hydrogen atoms forming an H—O···H motif) are 2.298, 2.249, and 2.293 Å, respectively. See Supporting Information Figure S2 for the distance distributions.

While both models can accurately reproduce most properties, the AM1PG-W model gives smaller errors in proton transfer energy profiles. See Supporting Information Figure S3 for


dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

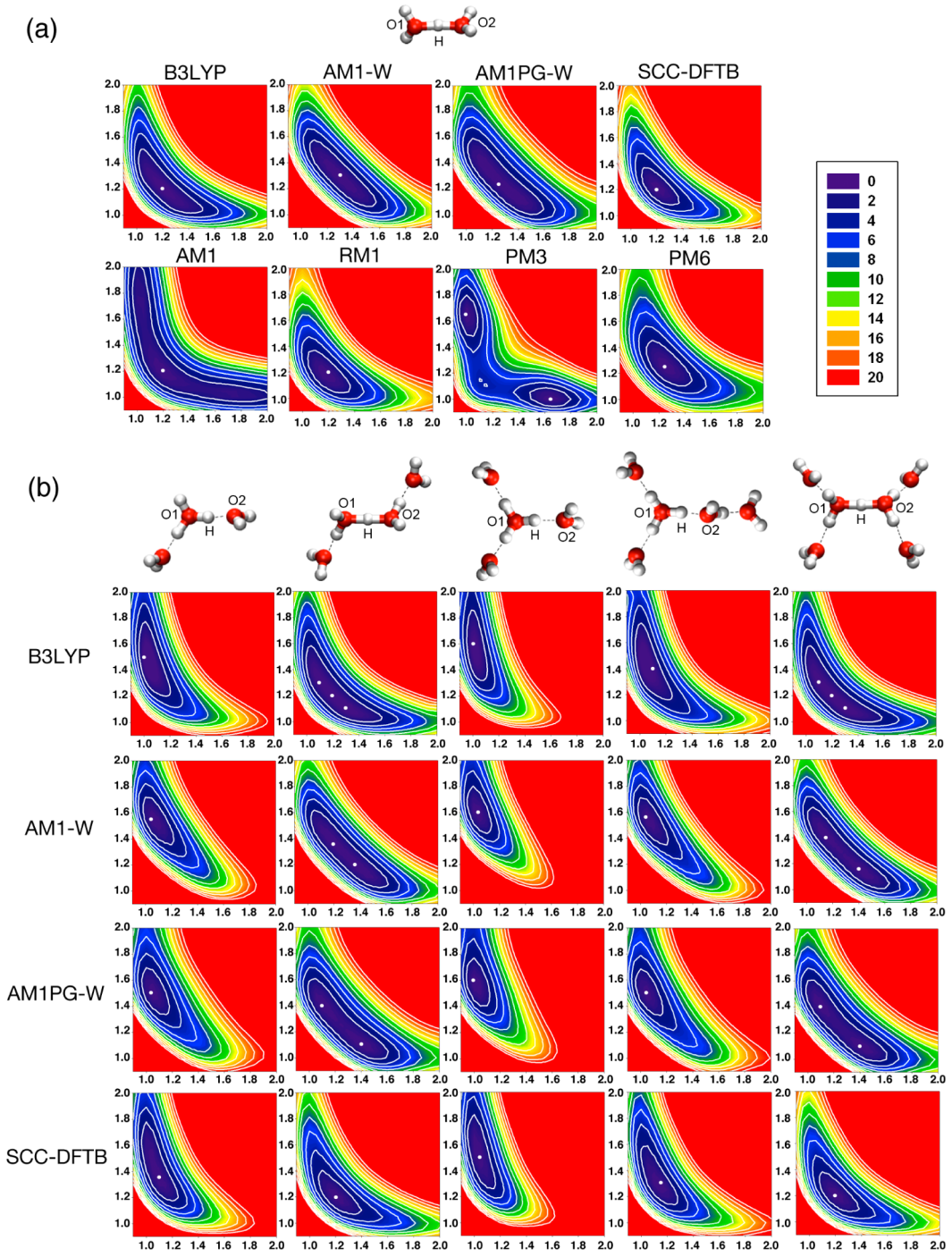



**Figure 6.** Proton transfer energy surfaces calculated using the AM1-W and AM1PG-W models, compared to the B3LYP benchmark and to other SE models from the literature (SCC-DFTB, AM1, RM1, PM3, and PM6). For each graph, the *x* and *y*-axis represent the O1−H and H−O2 distances (in Å), respectively, as illustrated in the corresponding molecular structure. Energies are in kcal/mol, color-coded according to the legend in the upper right corner. Panel a shows proton transfer energies for the $H_2O\cdots H^+\cdots OH_2$ motif and panel b shows the influence of additional hydrogen-bonded water molecules on proton transfer within a central $H_5O_2^+$ motif. In panel b, energy surfaces are not reported for AM1, RM1, PM3, and PM6 models because the hydrogen-bonded structures are unstable and cannot be optimized for all O1−H and H−O2 distances of interest. White contours represent energy levels from 0 to 20 kcal/mol, in 2-kcal/mol steps.

"model versus reference" correlation plots of all properties in the training and testing sets.

Figure 5 shows three proton transfer energy profiles from the testing set. Energies are defined as in eq 3 and plotted relative to the energy of the stable reactant structures (marked with label A). In accordance with the results of Figures 3 and 4, both AM1-W and AM1PG-W models accurately reproduce the proton transfer energy profiles, with AM1PG-W results closer to the DFT benchmark data. The equilibrium positions of the proton in AM1-W and AM1PG-W are different from the DFT positions but, considering that the energy difference between the stable reactant structure (A) and the minimum-energy structure is within 0.6 kcal/mol in all profiles, this structural discrepancy is likely to be attenuated in clusters at room


dx.doi.org/10.1021/ct500164h | J. Chem. Theory Comput. 2014, 10, 2881−2890

temperature. The models are performing well even in the higher-energy sections of the profiles, with a maximum error of 4.0 kcal/mol. This suggests that the models will be useful in studying proton transfer reactions with high energy barriers.

As a final assessment of the AM1-W and AM1PG-W models, Figure 6 presents the energies of Zundel-type systems, calculated as a function of O1−H and H−O2 distances, while fixing the O1−H−O2 angle to 180°. In this figure, the results from AM1-W and AM1PG-W models are compared to those from B3LYP and from the SCC-DFTB,[78,79] AM1, RM1, PM3, and PM6 models. (Results were also produced for the recently reported AM1n model[31] but are not included in this figure because the energy profiles are undistinguishable from those from AM1.) For the simple $H_5O_2^+$ dimer, models AM1-W and AM1PG-W perform clearly better than AM1 and PM3 models, which significantly underestimate hydrogen bonding and fail to reproduce the low barrier for proton transfer. Compared to SCC-DFTB, RM1, and PM6 models, models AM1-W and AM1PG-W better describe the facile movement of the excess proton from one $H_2O$ fragment to the other, as reflected by the elongated shape of the low-energy region of the surface. For larger systems, the AM1-W and AM1PG-W models correctly describe the polarization effect created by additional water molecules hydrogen-bonded to the central $H_5O_2^+$ group (see Figure 6b). The SCC-DFTB model performs comparably well but lacks for the $H_9O_4^+$ tetramer and $H_{13}O_6^+$ hexamer the same facile proton movement as for the $H_5O_2^+$ dimer. By contrast, the AM1, RM1, PM3, and PM6 models fail to stabilize the hydrogen bond network and favor structures in which the excess proton is coordinated by three or more water molecules (data not shown in Figure 6b). The intermolecular O···H distances predicted by AM1-W and AM1PG-W models are slightly too long, with an average error of 0.079 Å for the six structures of Figure 6. However, this average deviation in the O···H distances does not seem to be systematic, as it is actually significantly smaller for both models when considering the entire set of structures calculated (Figures 1 and 2). It should be noted that the training set contained information about proton transfer processes in the $H_5O_2^+$ dimer and the $H_7O_3^+$ trimer only (the first two structures of Figure 6).

## 4. CONCLUSION

In summary, a new method is introduced to develop reaction-specific semiempirical models for proton transfer reactions in molecular clusters. It is applied to small water clusters, used as a model system for proton transfer reactions in confined environments such as active sites of proteins. The training set for the model is generated using a series of standard transformations (protonation/deprotonation, addition of a water molecule, reaction involving a transition state, and proton transfer). It enumerates all stable structures and all transition states of water monomer, dimer, and trimer.

The approach is general and was implemented as such. The benchmark data for the training and testing sets are prepared semiautomatically using in-house scripts, and stored as a highly structured database that can be easily queried to calculate the error function. The parametrization pipeline was designed to provide a flexible definition of the error function and, provided the relevant molecular structures have been inserted into the database, can be used to generate any other reaction-specific or system-specific SE models.

The final parameter sets, models AM1-W and AM1PG-W, represent a significant improvement over the AM1, RM1, and PM3 models for the description of water clusters and proton transfer reactions. Although the models were trained on small water clusters, they perform well for water tetramers, pentamers, and hexamers. In particular, the proton transfer energy profiles generated by the new models match the high-level DFT results very well. Excluding the error from the high-energy structures C6 and D8 (Figure 2d), the maximum errors in proton transfer energy profiles are 7.3 kcal/mol for AM1-W and 3.8 kcal/mol for AM1PG-W, and the average errors are 1.9 and 1.0 kcal/mol, respectively. Furthermore, as shown in Figure 6, the proton transfer energy surfaces are close to the B3LYP results and the models correctly capture the influence of the environment's polarity on the proton transfer profile of the $H_2O\cdots H^+\cdots OH_2$ motif. The good performance of the models suggests that they are transferable to polarized environments and may be applied to more extended water molecule networks in proteins (within a QM/MM simulation scheme). Future work will be directed toward models describing other proton-donating and accepting groups such as imidazole and carboxylate, as well as metal-bound ligands.

## ■ ASSOCIATED CONTENT

### Supporting Information

Parameterization workflow, detailed analysis of the performance of AM1-W and AM1PG-W models in the training and testing sets, and input parameters for the PIKAIA program. This material is available free of charge via the Internet at http://pubs.acs.org.

## ■ AUTHOR INFORMATION

### Corresponding Author

*Tel.: 514-848-2424 ext 5314. Fax: 514-848-2868. Email: guillaume.lamoureux@concordia.ca.

### Notes

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

This work has been supported by an NSERC Discovery Grant to G.L. and an NSERC Undergraduate Student Research Award to L.M. Additional computational resources were provided by Calcul Québec.

## ■ REFERENCES

(1) Marx, D. *ChemPhysChem* **2006**, *7*, 1848−1870.
(2) Jeffrey, G. A.; Saenger, W. In *Hydrogen Bonding in Biological Structures*; Springer-Verlag: Berlin, 1994.
(3) Fersht, A. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding*; W.H. Freeman & Company: New York, 1998.
(4) Maffeo, C.; Bhattacharya, S.; Yoo, J.; Wells, D.; Aksimentiev, A. *Chem. Rev.* **2012**, *112*, 6250−6284.
(5) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227−249.
(6) Ribeiro, A. J. M.; Ramos, M. J.; Fernandes, P. A. *J. Am. Chem. Soc.* **2012**, *134*, 13436−13447.
(7) Phoon, L.; Burton, N. A. *J. Mol. Graph. Model.* **2005**, *24*, 94−101.
(8) Szeto, M. W. Y.; Mujika, J. I.; Zurek, J.; Mulholland, A. J.; Harvey, J. N. *J. Mol. Struct. THEOCHEM* **2009**, *898*, 106−114.
(9) Xu, D.; Guo, H. *J. Am. Chem. Soc.* **2009**, *131*, 9780−9788.
(10) Blumberger, J.; Lamoureux, G.; Klein, M. L. *J. Chem. Theory Comput.* **2007**, *3*, 1837−1850.
(11) Riccardi, D.; Yang, S.; Cui, Q. *Biochim. Biophys. Acta* **2010**, *1804*, 342−351.
(12) Nachimuthu, S.; Gao, J.; Truhlar, D. G. *Chem. Phys.* **2012**, *400*, 8−12.

(13) McNamara, J. P.; Hillier, I. H. *Phys. Chem. Chem. Phys.* **2007**, *9*, 2362−2370.
(14) Korth, M. *J. Chem. Theory Comput.* **2010**, *6*, 3808−3816.
(15) S. Rzepa, H.; Yi, M. *J. Chem. Soc. Perkin Trans. 2* **1990**, 943−951.
(16) Arillo-Flores, O. I.; Ruiz-López, M. F.; Bernal-Uruchurtu, M. I. *Theor. Chem. Acc.* **2007**, *118*, 425−435.
(17) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(18) Stewart, J. J. P. *J. Comput. Chem.* **1989**, *10*, 209−220.
(19) Thiel, W.; Voityuk, A. A. *Theor. Chim. Acta* **1992**, *81*, 391−404.
(20) Bräuer, M.; Kunert, M.; Dinjus, E.; Klußmann, M.; Döring, M.; Görls, H.; Anders, E. *J. Mol. Struct. THEOCHEM* **2000**, *505*, 289−301.
(21) Harb, W.; Bernal-Uruchurtu, M. I.; Ruiz-López, M. F. *Theor. Chem. Acc.* **2004**, *112*, 204−216.
(22) Bernal-Uruchurtu, M. I.; Martins-Costa, M. T. C.; Millot, C.; Ruiz-López, M. F. *J. Comput. Chem.* **2000**, *21*, 572−581.
(23) Bernal-Uruchurtu, M.; Ruiz-López, M. *Chem. Phys. Lett.* **2000**, *330*, 118−124.
(24) Korth, M.; Pitoňák, M.; Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2010**, *6*, 344−352.
(25) Csonka, G. I.; Ángyán, J. G. *J. Mol. Struct. THEOCHEM* **1997**, *393*, 31−38.
(26) Buesnel, R.; Hillier, I. H.; Masters, A. J. *Chem. Phys. Lett.* **1995**, *247*, 391−394.
(27) González-Lafont, A.; Truong, T. N.; Truhlar, D. G. *J. Phys. Chem.* **1991**, *95*, 4618−4627.
(28) Layfield, J. P.; Owens, M. D.; Troya, D. *J. Chem. Phys.* **2008**, *128*, 194302.
(29) Nam, K.; Cui, Q.; Gao, J.; York, D. M. *J. Chem. Theory Comput.* **2007**, *3*, 486−504.
(30) Troya, D.; García-Molina, E. *J. Phys. Chem. A* **2005**, *109*, 3015−3023.
(31) Wu, X.; Thiel, W.; Pezeshki, S.; Lin, H. *J. Chem. Theory Comput.* **2013**, *9*, 2672−2686.
(32) Liang, S.; Roitberg, A. E. *J. Chem. Theory Comput.* **2013**, *9*, 4470−4480.
(33) Korth, M.; Thiel, W. *J. Chem. Theory Comput.* **2011**, *7*, 2929−2936.
(34) Choi, T. H.; Liang, R.; Maupin, C. M.; Voth, G. A. *J. Phys. Chem. B* **2013**, *117*, 5165−5179.
(35) Chang, D. T.; Schenter, G. K.; Garrett, B. C. *J. Chem. Phys.* **2008**, *128*, 164111.
(36) Lin, Y.; Wynveen, A.; Halley, J. W.; Curtiss, L. A.; Redfern, P. C. *J. Chem. Phys.* **2012**, *136*, 174507.
(37) Řezáč, J.; Hobza, P. *J. Chem. Theory Comput.* **2012**, *8*, 141−151.
(38) Burshtein, K. Y.; Isaev, A. N. *J. Struct. Chem.* **1986**, *27*, 347−351.
(39) Goyal, P.; Elstner, M.; Cui, Q. *J. Phys. Chem. B* **2011**, *115*, 6790−6805.
(40) Maupin, C. M.; Aradi, B.; Voth, G. A. *J. Phys. Chem. B* **2010**, *114*, 6922−6931.
(41) Rodríguez, J. *J. Comput. Chem.* **1994**, *15*, 183−189.
(42) Zhang, P.; Fiedler, L.; Leverentz, H. R.; Truhlar, D. G.; Gao, J. *J. Chem. Theory Comput.* **2011**, *7*, 857−867.
(43) Maupin, C. M.; Voth, G. A. *Biochim. Biophys. Acta* **2010**, *1804*, 332−341.
(44) Freier, E.; Wolf, S.; Gerwert, K. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 11435−11439.
(45) Garczarek, F.; Gerwert, K. *Nature* **2006**, *439*, 109−112.
(46) Vardi-Kilshtain, A.; Major, D. T.; Kohen, A.; Engel, H.; Doron, D. *J. Chem. Theory Comput.* **2012**, *8*, 4786−4796.
(47) Wang, S.; Orabi, E. A.; Baday, S.; Bernèche, S.; Lamoureux, G. *J. Am. Chem. Soc.* **2012**, *134*, 10419−10427.
(48) Lohr, L. L.; Schlegel, H. B.; Morokuma, K. *J. Phys. Chem.* **1984**, *88*, 1981−1987.
(49) Stull, D. R.; Prophet, H. *Janaf Thermochemical Tables*; 2nd ed.; National Bureau of Standards: Washington, DC, 1971.
(50) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; 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.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09, Revision B.01, 2009.
(51) Bryantsev, V. S.; Diallo, M. S.; van Duin, A. C. T.; Goddard, W. A. *J. Chem. Theory Comput.* **2009**, *5*, 1016−1026.
(52) Stewart, J. J. P. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1−103.
(53) Shanno, D. F. *Math. Comput.* **1970**, *24*, 647−656.
(54) Goldfarb, D. *Math. Comput.* **1970**, *24*, 23−26.
(55) Fletcher, R. *Comput. J.* **1970**, *13*, 317−322.
(56) Broyden, C. G. *J. Inst. Math. Appl.* **1970**, *6*, 222−231.
(57) Henkelman, G.; Jònsson, H.; Jonsson, H. *J. Chem. Phys.* **1999**, *111*, 7010−7022.
(58) Laino, T.; Mohamed, F.; Laio, A.; Parrinello, M. *J. Chem. Theory Comput.* **2005**, *1*, 1176−1184.
(59) Repasky, M. P.; Chandrasekhar, J.; Jorgensen, W. L. *J. Comput. Chem.* **2002**, *23*, 1601−1622.
(60) Giese, T. J.; Sherer, E. C.; Cramer, C. J.; York, D. M. *J. Chem. Theory Comput.* **2005**, *1*, 1275−1285.
(61) McNamara, J. P.; Sundararajan, M.; Hillier, I. H. *J. Mol. Graph. Model.* **2005**, *24*, 128−137.
(62) Tafipolsky, M.; Schmid, R. *J. Phys. Chem. B* **2009**, *113*, 1341−1352.
(63) Isegawa, M.; Fiedler, L.; Leverentz, H. R.; Wang, Y.; Nachimuthu, S.; Gao, J.; Truhlar, D. G. *J. Chem. Theory Comput.* **2013**, *9*, 33−45.
(64) Hutter, M. C.; Reimers, J. R.; Hush, N. S. *J. Phys. Chem. B* **1998**, *102*, 8080−8090.
(65) Brothers, E. N.; Merz, K. M., Jr.; Merz, K. M. *J. Phys. Chem. B* **2002**, *106*, 2779−2785.
(66) Giese, T. J.; York, D. M. *J. Chem. Phys.* **2005**, *123*, 164108.
(67) Williams, D. E.; Peters, M. B.; Wang, B.; Merz, K. M. *J. Phys. Chem. A* **2008**, *112*, 8829−8838.
(68) Mane, J. Y.; Klobukowski, M. *Chem. Phys. Lett.* **2010**, *500*, 140−143.
(69) Tejero, I.; González-Lafont, À.; Lluch, J. M. *J. Comput. Chem.* **2007**, *28*, 997−1005.
(70) Rossi, I.; Truhlar, D. G. *Chem. Phys. Lett.* **1995**, *233*, 231−236.
(71) Cundari, Thomas R.; Jun Deng, W. F. *Int. J. Quantum Chem.* **2000**, *77*, 421−432.
(72) Wang, J.; Cieplak, P.; Li, J.; Hou, T.; Luo, R.; Duan, Y. *J. Phys. Chem. B* **2011**, *115*, 3091−3099.
(73) Charbonneau, P. *Astrophys. J. Suppl. Ser.* **1995**, *101*, 309−334.
(74) Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P. *J. Comput. Chem.* **2006**, *27*, 1101−1111.
(75) Herndon, W.; Radhakrishnan, T. *Chem. Phys. Lett.* **1988**, *148*, 492−496.
(76) Dannenberg, J. J. *J. Phys. Chem.* **1988**, *92*, 6869−6871.
(77) Rzepa, H. S.; Yi, M. Y. *J. Chem. Soc. Perkin Trans. 2* **1990**, 943−951.
(78) Zhechkov, L.; Heine, T.; Patchkovskii, S.; Seifert, G.; Duarte, H. A. *J. Chem. Theory Comput.* **2005**, *1*, 841−847.
(79) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. *Phys. Rev. B* **1998**, *58*, 7260.