JCTC Journal of Chemical Theory and Computation

pubs.acs.org/JCTC Article

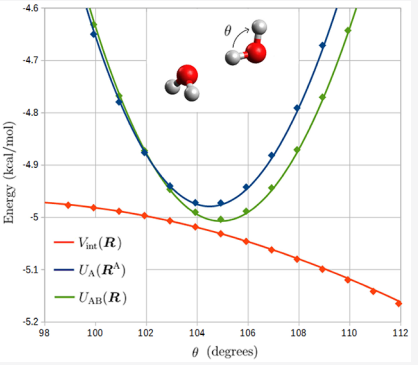
Automatic Generation of Flexible-Monomer Intermolecular Potential Energy Surfaces

Michael P. Metz and Krzysztof Szalewicz*

Cite This: J. Chem. Theory Comput. 2020, 16, 2317−2339 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: A method is developed for automatic generation of nonreactive intermolecular two-body potential energy surfaces (PESs) including intramonomer degrees of freedom. This method, called flex-autoPES, is an extension of the autoPES method developed earlier, which assumes rigid monomers. In both cases, the whole PES development proceeds without any human intervention. The functional form used is a sum of products of site−site functions (both atomic and off-atomic sites can be used). The leading terms with sites involving different monomers are of physically motivated form. The long-range part of a PES is computed from monomer properties without using any dimer information. The close-range part is fitted to dimer interaction energies computed using electronic structure methods. Virtually any method can be used in such calculations, but the use of symmetry-adapted perturbation theory provides a seamless connection to the long-range part of the PES. The performance of the flex-autoPES code was tested by developing a full-dimensional PES for the water dimer and PESs including only some soft intramonomer degrees of freedom for the ethylene glycol dimer and for the ethylene glycol−water dimer. In the case of the water dimer, the root-mean-square error (RMSE) of the PES from the data points with negative total energies is 0.03 kcal/mol, and we expect this PES to be more accurate than any previously published PES of this type. For the ethylene glycol dimer and the ethylene glycol−water dimers, the analogous RMSEs are 0.25 and 0.1 kcal/mol, respectively.



Downloaded via UNIV OF DELAWARE on April 22, 2020 at 15:59:38 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

I. INTRODUCTION

The (total) potential energy surface (PES) for a set of atoms or molecules (a cluster) is defined as the ground-state solution of the time-independent Schrödinger equation in the Born−Oppenheimer (BO) approximation, that is, the equation for the electron energy and the associated wave function in the field of nuclei clamped in space. A PES is a function of the coordinates of all atoms of a given cluster. It is convenient to separate the total PES into intermonomer and intramonomer parts. The latter is given by the sum of the total energies of all isolated monomers, and the former, called intermolecular interaction energy, is the difference between the total PES and the intramonomer PES (computed using monomer coordinates at the same values as within the cluster). Intermonomer PESs are often constructed in the rigid-monomer approximation. This gives a huge reduction in dimensionality: PESs are at the most 6-dimensional for dimers, 12-dimensional for trimers, and so on. This is to be compared with full-dimensional PESs, which have dimensions 12, 18, 24, and 30 for homogeneous dimers of 3-, 4-, 5-, and 6-atomic monomers. As defined so far, a PES is an assignment of energy to each possible configuration of a cluster. However, one usually applies this name to an analytic function fitted to a finite set of reference energies, referred to as grid points. PESs of this type are global, since they can be computed for any configuration of a cluster. The derivatives of PESs provide forces acting on atoms, hence one often uses the name force fields.

If a PES is available, one can solve the Schrödinger equation for the motions of nuclei on this surface, that is, the Schrödinger equation depending only on nuclear coordinates, with the PES being the potential energy term. For many problems, it suffices to solve classical Newton equations of motion with constraints imposed by statistical mechanics, a method referred to as molecular dynamics (MD). Such calculations give predictions of quantities that can be directly compared with experiment. In contrast, individual interaction energies cannot be compared with experiment except for very small clusters, and even then in an indirect way. Thus, PESs are of fundamental importance in science. Results of MD simulations can also be used to provide information to various kinetic and continuum approaches, leading to multiscale modeling protocols.[1,2] At this point, the majority of simulations of condensed phases and biomolecular aggregates

Received: December 14, 2019
Published: April 2, 2020

© 2020 American Chemical Society

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

are performed with empirical PESs, that is, PESs with parameters adjusted to get the best possible agreement between predictions of MD simulations and experimental data. PESs derived from first-principles, that is, by fitting to ab initio computed energies, are less often used, mainly because such PESs are much more difficult to develop. Nevertheless, first-principles or ab initio based PESs are steadily supplanting empirical PESs for larger and larger clusters. The former PESs have several advantages over the latter. In particular, while empirical PESs are applicable only to specific systems in some range of thermodynamic conditions, first-principles PESs are applicable to all systems containing given monomers. For example, the first-principles PES for water from refs 3−5 accurately predicts properties[6] from spectra of the water dimer to radial distribution functions of liquid water (provided reactive processes are absent). It should be clarified that we use, here, the phrase "first-principles" or "ab initio" mainly to distinguish PESs fitted to interaction energies obtained from solutions of Schrödinger's equation from empirical PESs. However, the autoPES methodology can be applied with interaction energies computed using electronic structure methods which are not completely ab initio, for example, some semiempirical density-functional theory (DFT) methods fitted to interaction energies from high-level ab initio calculations and supplemented by semiempirical dispersion functions (so-called DFT+D approaches) can predict interaction energies reasonably well.

One should mention here that many types of nuclear motion calculations can be performed without developing a PES. In MD, for example, the values of the potential energy and derivatives are needed only at the points in atomic coordinate space that are sampled by an MD trajectory, and therefore can be computed on-the-fly. This approach is generally referred to as ab initio MD (AIMD) (the term ab initio MD is somewhat misleading and a better name would be direct dynamics MD or on-the-fly MD). However, AIMD with a DFT method used for energy and gradient evaluations is about 6 orders of magnitude more expensive than MD using PESs. One should emphasize that dynamical calculations using PESs fitted to interaction energies from an electronic structure method are commensurate with on-the-fly calculations provided that such interaction energies are fitted with errors smaller than the uncertainties of these energies, that is, the fits are of high fidelity, which is often the case. In this way, ab initio PESs fitted to interaction energies computed using methods such as coupled clusters with single, double, and noniterative triple excitations [CCSD-(T)] reduce the cost of dynamical calculations by about nine orders of magnitude compared to on-the-fly calculations using CCSD(T).

The subject of the present work is flexible-monomer PESs of nonreactive systems developed from first principles. One should emphasize here that the rigid-monomer approximation works very well in most cases, at least for relatively small systems for which accurate PESs can be computed. This approximation was the foundation of most of the past work on spectra and scattering of atomic and molecular systems. In the case of PESs fitted to ab initio interaction energies, it is sometimes possible, paradoxically, for rigid-monomer PESs to provide more accurate results than flexible-monomer ones simply because uncertainties of the latter are much larger than those of the former due to more difficult fitting in a large number of dimensions and possibly the use of a lower level of theory to afford calculations of a large number of grid points. One example are recent PESs for H_2−CO.[7,8] Another example are energies of water clusters: a rigid-monomer two—plus three-body PES gave cluster energies much closer to accurate ab initio benchmarks than the best available flexible-monomer PESs.[9] Also, theoretical predictions for the water dimer rovibrational spectra from rigid- and flexible-monomer PESs have similar discrepancies compared to experiments.[10,11] Of course, as the accuracy of theoretical predictions increases, one has to take monomer-flexibility effects into account since their neglect becomes at some point the largest uncertainty contributor. At the present time, the water dimer is an example of a system where to get spectra to agree with experiment to better than one wavenumber, the monomer-flexibility effects have to be taken into account.[10,11] Perhaps even more importantly, there are several physical properties that require flexible-monomer PESs to make predictions, one example being shifts of monomer spectral lines due to intermolecular interactions. Another group of systems requiring flexible-monomer PESs are monomers that have "soft" degrees of freedom, such as rotations of functional groups around a chemical bond. The interaction energies may strongly depend on values of soft intramonomer coordinates. Knowledge of such dependence is particularly important in crystal structure predictions involving such monomers since the monomer conformations in a crystal can be completely different than in the gas phase.

From the discussion above, it is clear that the development of PESs that depend on monomers' degrees of freedom is an important current frontier in science. However, the difficulty of PES development increases exponentially with the number of degrees of freedom due to the fact that the number of grid points needed to sample a PES increases as $p^{3N_{at}-6}$, where p is the number of points per dimension and N_{at} is the number of atoms. This is sometimes called the dimensionality curse. Even if the grid-point generation is based on stochastic procedures in all dimensions considered, the increase in the number of such points with N_{at} is still steep. Full-dimensional PESs of spectroscopic accuracy can currently be developed for systems with up to a dozen or so degrees of freedom. For larger monomers with a few soft degrees of freedom, one can develop PESs dependent on these degrees, with the remaining internal ones frozen. The method presented here allows for both options.

Until recently, the generation of each PES, even assuming rigid monomers, required significant amounts of human effort and of computer resources. The latter was due both to the necessity of performing calculations for a large number of grid points (10^5 in some cases) and the significant costs of a single ab initio calculation if an accurate electronic structure method, such as CCSD(T), was used. Furthermore, accuracy of the fits was often a problem, in particular at long-range. This situation has changed with the development of the autoPES method and the corresponding system of computer codes.[12,13] The human effort was reduced almost to zero since the autoPES program requires only monomer coordinates on input and generates a PES automatically. The huge computational effort of a single ab initio calculation with the CCSD(T) method was reduced by using SAPT based on DFT description of monomers, SAPT(DFT).[14−25] While the cost of CCSD(T) scales as the seventh power of the system size, SAPT(DFT) scales only as the fifth power, and, for dimers with about 60 atoms, takes about the same time as a supermolecular DFT calculation of interaction energy (since a DFT calculation for a dimer is

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

about as expensive as a SAPT postmonomer calculation). At the same time, SAPT(DFT) gives interaction energies which are only a couple of percent different from the CCSD(T) ones in the same basis set.[26] However, autoPES can use CCSD(T) or any electronic structure method for close-range calculations, depending on the accuracy requirements of the application. In particular, for use in predictions of spectra of small van der Waals complexes, levels as high as CCSDT(Q) have been used.[27,28] Perhaps the most important reduction in computer costs was due to the reduction of the number of grid points. This has been achieved in two ways. First, while many published PESs were fitted to products of one-dimensional grids (which leads to the $p^{3N_{at}-6}$ scaling), autoPES uses a Monte Carlo grid generation method sampling directly all-dimensional space and using importance sampling guided by the strength of interactions. Second, for a large part of the coordinate space, electronic structure calculations are not needed since this region is accurately described by the asymptotic expansion, discussed below. In effect, autoPES needs only on the order of 10^3 grid points. The accuracy problems of some published PESs result from either overly simple forms of fitting functions or from incorrect asymptotic behavior. Fits being a sum of atom−atom pair functions of Lennard-Jones or Buckingham form are not sufficiently malleable to represent accurate interaction energies as functions of atomic coordinates. In autoPES, this problem has been solved by using off-atomic sites in addition to atomic ones and by using a more sophisticated form of pair functions, including terms such polynomials multiplied by exponentials, inverse powers of site−site distances r_{ab} up to r_{ab}^{-10}, and damping functions. It would be easy to program still more sophisticated fit functions in autoPES, but the advantage of the chosen one is that it can be used without changes in virtually all MD programs. The problem of accuracy of PESs at large separations plagues in particular the approaches that fit the PES locally or as the sum of products (SOP) of a simple pair functions. While such approaches are capable of achieving high accuracy in the close-range region, the PESs decay in an unphysical way. In autoPES, the asymptotic expansion of interaction energy is used with coefficients computed from monomers' multipole moments and static and dynamic polarizabilities. Interaction energies from the asymptotic expansion are seamlessly connected with SAPT(DFT) interaction energies in the sense that the differences between the two quantities are arbitrarily small if the separation between monomers is sufficiently large. In addition to dimer (two-body) PESs, autoPES can generate a polarization component that partly accounts for pairwise nonadditive effects if the PES is applied in a many-body context. For bulk systems of polar monomers, polarization effects are a very good approximation to the complete many-body effects.[29]

The main limitation of autoPES is that it assumes the rigid-monomer approximation. In the present work, this limitation is removed. The reason an extension to flexible monomers is nontrivial is that autoPES fits interaction energies in a physically motivated way. In contrast, methods based on local fits, such as interpolating moving least-squares (IMLS) method,[30−32] SOPs of simple pair functions, such as the permutationally invariant polynomials (PIP) approach,[33−36] or many machine learning (ML) methods,[37−40] do not in general distinguish inter- and intramonomer degrees of freedom and, therefore, can generate flexible-monomer PESs as readily as rigid-monomer ones. Published flexible-monomer PESs from our group[8,27,41−45] used expansions into a sum of intermonomer site−site pair functions similar to those used in autoPES and then expanded parameters of this expansion as polynomials of intramonomer coordinates. [Fits for smaller systems used functions based on expansion in the distance R between centers of mass (COMs) of monomers and Euler angles defining the relative orientation of monomers.]

Our present approach to the design of the functional form for flexible-monomer PESs takes an intermediate route between the expansion into a sum of intermonomer site−site functions with parameters of this expansion dependent on intramonomer distances and the SOP method. The latter approach expands the intermolecular PES into the sum of all possible products of atom−atom pair functions, where each term has to contain at least one intermonomer pair, with up to 6-fold products used in some fits.[33−35] The atom−atom functions in this approach are very simple, for example, $e^{-\alpha r_{ab}}$, and sometimes identical for all pairs. In the present approach, the atom−atom functions are much more complex and reflect the physical nature of atom−atom interactions (in the sense of an atom-distributed description of interaction energy). We also use off-atomic sites. However, the products are limited to double products only, and the intramolecular pair function in such products is a polynomial of site−site distances. Furthermore, we use this form only for the exponential component of the potential, the asymptotic component is treated differently. The functional form of the former component consists of a sum of exponentials of (only) intermonomer site−site distances r_{ab} multiplied by (a) a polynomial of r_{ab}, (b) a sum of polynomials of $r_{ab'}$ and $r_{a'b}$, and (c) a sum of polynomials of $r_{a'a''}$ and of $r_{b'b''}$, where all a, a', a'' (b, b', b'') indices denote sites of monomer A (B). The a and b terms describe short-range intermolecular interactions of rigid monomers. The term a is of the form used in most published rigid-monomer PESs from our group. It does depend on monomer deformations but is not expected to represent them well.[46,47] The term b can be considered an extension of the sum of site−site functions approach in the SOP spirit and can be used also for rigid-monomer PESs. In the flexible-monomer context, it should provide a partial account of the monomer-flexibility effects. The main coupling of the intramonomer and intermonomer effects originates from term c, which is the leading term for such coupling in the SOP approach. The asymptotic component uses an expansion of charges and of van der Waals coefficients into polynomials of $r_{aa'}$ and $r_{bb'}$.

We test the flex-autoPES methodology on three dimers: (i) the water dimer in full dimensionality, 12D; (ii) the ethylene glycol dimer with partial monomer flexibility: the three soft rotational degrees of freedom are modeled, with the remaining internal degrees of freedom frozen, so that the dimer PES is 12-dimensional; and (iii) the interaction of water with ethylene glycol, with the same selection of internal degrees of freedom (IDOFs) as are used in the two homogeneous cases. The existing 12D water dimer PESs[33,43,45,48−51] have been recently compared in refs 11, 45, and 52. We will evaluate the quality of the new PES by comparing fitting errors and stationary points on the surface with those of literature potentials. For $(HO(CH_2)_2OH)_2$ and $HO(CH_2)_2OH{\cdot}H_2O$, for which no system-specific literature PESs exist, we instead compare to the OPLS-AA[53,54] force field and the recently developed interaction potentials from machine learning (IPML) method.[55]

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

II. GEOMETRY GRID

In the rigid monomer case, the dimer interaction energy depends only on the relative configuration of the monomers, which is specified by a point on a surface of dimension D_{rigid} = 6 or less (from D_{rigid} = 2 for atom−diatom interactions). To describe a given dimer configuration i in the case of D_{rigid} = 6, we use the coordinates

$$\boldsymbol{R}_i^{\text{rigid}} = (R_i, \boldsymbol{\Omega}_i) = (R_i, \beta_i^{\text{A}}, \gamma_i^{\text{A}}, \alpha_i^{\text{B}}, \beta_i^{\text{B}}, \gamma_i^{\text{B}}) \tag{1}$$

where R_i is the distance between the COMs of the monomers, β_i^{A} and γ_i^{A} are Euler angles specifying the orientation of monomer A relative to some space-fixed reference frame, and α_i^{B}, β_i^{B}, and γ_i^{B} specify the orientation of monomer B.

The most general description of a flexible-monomer PES places no restriction on the relative positions of nuclei. Such a surface has dimensionality $D = 3N_{\text{at}} - 6$ for $N_{\text{at}} \geq 3$. To address the widest possible range of applications, we have designed the software to include explicitly any number of IDOFs, and so the dimensionality is in the range $6 \leq D \leq 3N_{\text{at}} - 6$ for D_{rigid} = 6. The dimer coordinates in the general case are given by

$$\begin{aligned}\boldsymbol{R}_i &= (\boldsymbol{R}_i^{\text{rigid}}, \boldsymbol{R}_i^{\text{A}}, \boldsymbol{R}_i^{\text{B}}) \\ &= (R_i, \boldsymbol{\Omega}_i, X_{\text{A},i}^1, ..., X_{\text{A},i}^{N_{\text{A}}^{\text{IDOF}}}, \chi_{\text{A},i}, X_{\text{B},i}^1, ..., X_{\text{B},i}^{N_{\text{B}}^{\text{IDOF}}}, \chi_{\text{B},i})\end{aligned} \tag{2}$$

where $X_{\text{B},i}^j$ and $X_{\text{B},i}^j$ are real numbers specifying the amount of the internal deformations of the jth IDOF of monomer A and B, respectively. The boolean values $\chi_{\text{A},i}$ and $\chi_{\text{B},i}$ indicate whether the monomers should be reflected about the origin prior to the other transformations. In the case that monomer A(B) is achiral, this reflection operation is unnecessary, and $\chi_{\text{C},i}$, C = A or B, is always set to false. Note that in the chiral case, it is possible to use a more elaborate fitting function which uses separate sets of fit parameters for same-handed dimers and opposite-handed dimers, in which case the reflection operation is again unnecessary. We choose the convention such that for the undeformed reference monomer geometry, each of the $X_{\text{C},i}^j$ takes the value zero. We denote the reference geometry of monomer A as $\boldsymbol{R}_0^{\text{A}}$ = (0, 0, ..., 0) and similarly for B.

Each IDOF of monomer A is specified by a 4-tuple

$$\boldsymbol{D}_{\text{A}}^j = (T_{\text{A}}^j, v_{\text{A}}^j, w_{\text{A}}^j, S_{\text{A}}^j) \tag{3}$$

The boolean value T_{A}^j specifies the type of the IDOF: rotational or translational. The values v_{A}^j and w_{A}^j identify two sites in the monomer that define the axis of rotation in the case of a rotational IDOF (such rotations change the dihedral angles in a molecule) or a direction of translation for a translational one. The maximum and minimum limits of rotation or translation are either set by the user or determined by automatic default procedures for different parts of the PES generation (see below). Finally, S_{A}^j is the set of atoms that are translated or rotated according to the corresponding IDOF coordinate values $X_{\text{A},i}^j$. The sites v_{A}^j and w_{A}^j are usually atoms or off-atomic fitting sites in the monomer (the positions of the off-atomic sites are determined by the positions of the atoms, as described in section V). This means that the axis from v_{A}^j to w_{A}^j is always well-defined as the monomer deforms. Another type of rotation is the change of angle between two chemical bonds. In this case, v_{A}^j is some atom, and the axis of rotation goes through v_{A}^j in the direction normal to the plane formed by v_{A}^j and two other atoms bound to it and jointly denoted w_{A}^j (this is used, for example, to modify the angle between bonds of the water molecule). If necessary, off-atomic sites can be included, which are only used as v_{A}^j or w_{A}^j to define IDOFs, and are not used in the functional form of the fit. Such sites can be introduced, for example, to perform changes of angles between two adjacent bonds. In the most general case of a fully flexible large molecule, one could simply use as the $\boldsymbol{D}_{\text{A}}^j$ translations of each atom in each Cartesian coordinate, although we have not explored such cases here.

During each translation or rotation j, all atoms not in the set S_{A}^j are stationary. This means that, in cases where the internal deformations $X_{\text{A},i}^j$ are nonzero, the locations of the monomers' COMs will generally differ from those of the reference geometries. In our convention, the COM of undeformed monomer A remains at the origin of the coordinate system, whereas the COM of undeformed monomer B is at R_i along the positive z-axis. Therefore, the true COM-COM distance will generally only equal R_i in the case that the value of all IDOF coordinates are zero.

The set of IDOFs $\boldsymbol{D}_{\text{A}}^j$ and $\boldsymbol{D}_{\text{B}}^j$ determine the explicit dimensionality D of the intermolecular interaction PES. When the monomer geometries are constrained, that is, when $D < 3N_{\text{at}} - 6$, the PES remains well-defined on the entire $3N_{\text{at}} - 6$ physical input space. In these cases, the model is generally able to make reasonable predictions of interaction energies for geometries which depart from the D dimensional fitted input space, so long as these departures are small.

II.A. Intramolecular Grid Generation Procedure.

To fit the intramolecular portion of the PES, a separate set of electronic structure calculations is needed for each distinct monomer. One would think that monomer calculations performed during the close-range fitting process can be used to fit monomer PESs, but this is not the case. The reason is that both for SAPT and for supermolecular methods, one uses bases that have functions on the interacting partner [in the SAPT case to speed up convergence, in the supermolecular case to remove basis set superposition error (BSSE)]. These functions alter monomer energies significantly from their values in a pure monomer basis set, and since the alternation depends on the position of the partner, separate calculations for monomers are needed for each dimer's geometry. The separate calculations for monomers could be avoided if large basis sets and extrapolation to complete basis set (CBS) limits are used, but we have not included such an option. Also, in some cases one could take monomer PESs from literature and avoid monomer calculations in this way. While we could have used monomer grid points generated in other stages of PES development, it was more convenient to use a separate grid. Because the monomer calculations require much less time than the dimer calculations for a given accuracy, there is little need for fine-tuning the grid selection in the former case. Rather, we simply use an oversized fitting grid, with 40 grid points per free parameter of the fit, generated as described below.

In the case that the IDOF limits for monomer A are not specified by the user, they are set based on the geometric parameter $\delta_{\text{A}}(\boldsymbol{R}^{\text{A}})$, defined as

$$\delta_{\text{A}}(\boldsymbol{R}^{\text{A}}) = \max_{\{aa'\}}\left(\frac{|r_{aa'} - r_{aa'}^0|}{\min(r_{aa'}, r_{aa'}^0) - 0.5(r_a^{\text{cov}} + r_{a'}^{\text{cov}})}\right) \tag{4}$$

where $r_{aa'}$ is the distance between atoms a and a' in the monomer geometry $\boldsymbol{R}^{\text{A}}$, and $r_{aa'}^0$ is the corresponding distance

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

in $\boldsymbol{R}_0^A$. The term $0.5(r_a^{cov} + r_{a'}^{cov})$, which depends on the covalent radii of the atoms,[56] has the effect of restricting the relative motion of covalently bonded atom pairs more, while having a comparatively small effect on the range of motion of nonbonded atom pairs. For each IDOF j, we set the IDOF limits as the minimum and maximum values of X_A^j such that $\delta_A(\boldsymbol{R}_i^A)$ does not exceed δ_{max}, with all other IDOFs fixed at zero. In the present work, we use the value $\delta_{max} = 0.5$. All the corresponding quantities for monomer B are, of course, defined in the same way. In the case of water, this gives the range of X_A^j equal to (− 0.16 Å, 0.23 Å) where j is either of the O−H bond length deformations. For the water bond angle deformation, the minimum value of X_A^j is −33° (giving a minimum bond angle of 71.3°), and the maximum value permits the bond angle to go all the way to the minimum value in the opposite direction (which has the same effect as making the maximum bond angle give the linear molecule).

The values of X_A^j in the intramolecular fitting grid are generated using a uniform N_A^{IDOF}-dimensional Sobol sequence[57] in the ranges determined by the procedure described above (and similarly for monomer B). No energy-dependent weights are used in the intramolecular grid generation process.

II.B. Close-Range Intermolecular Grid Generation Procedure. Because the calculation of interaction energies in the nonasymptotic region dominates the overall computational cost of PES generation, points in the close-range intermolecular grid must be carefully chosen. Two competing objectives influence the choice of fitting data: that every relevant region of the configuration space must be adequately sampled and that the number of data points must be as small as possible. These objectives are intimately linked to the functional form of the model used. Adding free parameters to the model has the advantage of decreasing its fitting error on a given training set. Conversely, free parameters introduce statistical variance to certain regions of the configuration space, which, for a fixed set of input data, may result in decreased accuracy (see ref 58 for a detailed exploration of this relationship). The quality of the model, for a given amount of computer effort, therefore, depends critically on the choice of both fitting data and free parameters, in an interdependent way.

Various statistical methods have been devised wherein model fitting data is chosen according to some optimality criteria.[59] We describe one such approach in the context of intermolecular PESs in ref 60. The procedure of grid generation used here is a straightforward extension of that used previously in the rigid-monomer case, described in ref 12. This procedure uses a combination of heuristics that are tuned to the class of models investigated here. While not statistically well-founded, we have chosen this approach in the present work to isolate complications due to monomer flexibility from other factors. The intermolecular fitting grid is iteratively refined, with the close-range intermolecular component of the PES being refitted after each iteration of grid generation.

During a given iteration of grid generation, each coordinate of a grid point $\boldsymbol{R}_i$ is sampled from a separate, appropriately distributed random variable. For the angular coordinates β_i^A, γ_i^A, α_i^B, β_i^B, and γ_i^B, we use the same distribution that was used in the previous work, giving a uniform sampling of the unit sphere for each monomer. The probability density F for the radial variable R_i is modified from that of ref 12 and is given by

$$F(\rho) = \frac{A}{\rho}\exp\left(-\frac{[\ln \rho]^2}{2}\right)\Theta(\rho)\Theta(1-\rho) \tag{5}$$

where A is a normalization constant, Θ denotes the Heaviside step function, and $\rho = [R - R_{min}]/[R_{max} - R_{min}]$, where R_{min} and R_{max} are dependent on the dimer orientation. R_{min} is set to the COM distance such that the closest-contact atoms between the two monomers are separated by 1.3 times the sum of their covalent radii, and R_{max} to the distance such that the closest-contact atoms are separated by 1.3 times the sum of their van der Waals radii.[61] This distribution results in a large proportion of the grid points placed between the potential minimum at a given orientation and low on the repulsive wall, which was found to be optimal for overall convergence of the fit. This modification to the radial distribution is not related to the introduction of monomer flexibility—it was an improvement of the rigid-monomer approach that was introduced after the publication of the original method.

Rather than using the geometric parameter $\delta_C(\boldsymbol{R}^C)$, C = A or B, to determine the ranges of the IDOFs as is done for the intramolecular grids, here, we set the ranges of the IDOFs, denoted $(X_{C,min}^j, X_{C,max}^j)$, to the minimum and maximum values such that the energy of the intramolecular PES does not exceed 15 kcal/mol with all other IDOFs kept fixed. This avoids the possibility of generating points in physically irrelevant configurations, which may happen using the geometric criterion. For water, this gives $(X_{C,min}^j, X_{C,max}^j)$ equal to (−0.16 Å, 0.20 Å), where j is one of the O−H bond length deformations, and $(X_{C,min}^j, X_{C,max}^j)$ equal to (−22°, 39°), where j is the H−O−H angle deformation. For the internal coordinates $X_{C,j}^i$ of monomer C we use a truncated normal distribution $\mathcal{N}(0, \sigma^2)$ with mean $\mu = 0$ (corresponding to the undeformed reference monomer geometry) and standard deviation $\sigma = 0.4[X_{C,max}^j - X_{C,min}^j]$, with the probability set to zero outside the range $(X_{C,min}^j, X_{C,max}^j)$. In the case of a rotational IDOF for which the entire range of motion is included, a uniform distribution is used instead. No points are generated with the monomers fixed at their $\boldsymbol{R}_0^C$ geometries.

The grid point distributions defined above are additionally modified by an energy dependent factor. These energies are computed from the intermolecular PES $V_{int}(\boldsymbol{R})$ and intramolecular PES $U_A(\boldsymbol{R}^A) + U_B(\boldsymbol{R}^B)$, with their sum $U_{AB}(\boldsymbol{R})$ being the total energy of the dimer, defined precisely in section III. For $V_{int}(\boldsymbol{R})$, we use the PES from the previous fitting iteration. For the first grid generation iteration, we instead use the OPLS-AA[53] force field. The intramolecular PES is fitted before the first iteration of close-range grid generation, so no special considerations are needed in that case. Using these guiding potentials, each generated grid point $\boldsymbol{R}_i$ is rejected with probability P_i given by

$$P_i = \max[0,\ 1 - w(V_{int}(\boldsymbol{R}_i);\ V_{int}(\boldsymbol{R}_{min}),\ \eta_{int}) \times w(U_A(\boldsymbol{R}_i^A) + U_B(\boldsymbol{R}_i^B);\ 0,\ \eta_{mon})] \tag{6}$$

The weighting function $w(E; E_0, \eta)$ is to be thought of as a function of energy with two tuning parameters and is given by

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

$$w(E;\, E_0,\, \eta) = \exp(-\eta)\begin{cases}\exp(-w') & w' \le 0\\ [1 + 0.2w']^{-3} & w' \ge 0\end{cases}$$
$$w' = \eta\left[\frac{E - E_0}{4 + |E_0|} - 1\right] \quad (7)$$

The weight strength parameters η_{int} and η_{mon} have default values of 2 and 4, respectively, and $\boldsymbol{R}_{min}$ is the geometry of the minimum of $U_{AB}(\boldsymbol{R})$. The value of $w(E;\, E_0,\, \eta)$ is equal to 1 when $E = E_0$, and is between 0 and 1 when $E > E_0$. Thus, all points with $V_{int}(\boldsymbol{R}_i) < V_{int}(\boldsymbol{R}_{min})$ and $\boldsymbol{R}_i^C$ close to the $\boldsymbol{R}_0^C$ values are accepted. Recall that we use the convention that the monomer energies at the reference configurations are zero; hence, the value zero of the parameter E_0 for the factor dependent on the intramolecular PES in eq 6. The intermolecular part of eq 6 is very similar to that used in ref 12. For the intramolecular part, a stronger energy dependence was required to get the desired density of grid points in the low-energy region of the PES. This is because of the overall shape of the intermolecular PES as compared to the intramolecular one. In the intermolecular case, there is a relatively small (although important) region of configuration space low on the repulsive wall, with $V_{int}(\boldsymbol{R})$ between roughly $V_{int}(\boldsymbol{R}_{min})$ and 10 kcal/mol. In the intramolecular case, there is typically a large volume of configuration space in this energy range, and so, stronger weighting is required to put a reasonable proportion of points in the most physically important region with negative total energy $U_{AB}(\boldsymbol{R})$. This fact is demonstrated in section VIII. For the fitting weights, discussed in section III, we use a weight function which depends only on the total energy.

In addition to the geometry-dependent initial distribution and energy-dependent refinement, a final, deterministic refinement is used to eliminate grid points that are redundant with previously generated points. The similarity of each grid point $\boldsymbol{R}_i$, generated according to the above distribution, to each of the previously generated grid points $\boldsymbol{R}_j$ is evaluated using the ϵ_{ij} metric defined in ref 12. This metric is extended in the natural way to account for intramonomer deformations, by including in the distance vectors the intramonomer atom−atom distances in addition to the intermolecular ones. If the new point is found to be too similar, it is rejected. The ϵ_{ij} threshold below which points are considered too similar is adjusted such that three out of every four randomly generated points are rejected, see section IIA of ref 12 for details. A typical value of the ϵ_{ij} threshold after several thousand points are added is about 0.3 Å for the examples reported here.

The number of grid points in the initial iteration of the close-range intermolecular grid N_{grid} is given by $N_{grid} = 5N_{atom} + 5N_{FP}$, where N_{atom} is the total number of atoms in the dimer and N_{FP} is the total number of free parameters in the close-range intermolecular part of the fit (see section III.B). The term dependent on the number of atoms is absent from the formula used in ref 12 and was added (also to the rigid autoPES case) because some systems with very high symmetry, such as benzene, otherwise use too few grid points.

The parameters of the OPLS-AA force field have to be provided by the user by specifying OPLS-AA atom type. If no parameters are given, the program does not use the energy-dependent rejection criterion for the first iteration. In flexAutoPES, only the noncovalent part of OPLS-AA is needed (i.e., the same part as in the rigid case) since the intramonomer PESs are available at the start of intermolecular stage. The intramonomer component of OPLS-AA is not used at any point.

As in ref 12, a fraction of the grid points are placed near minima of the total PESs, rather than using the procedure described above. In the initial iteration, up to 10% of the total number of points is reserved for this purpose, whereas in subsequent iterations, where the locations of the minima are known more accurately, up to 25% are used; in both cases, the total number of points placed near minima in a given iteration is limited to no more than 10 times the number of minima. One grid point is placed at the exact location of each minimum, with the remaining points uniformly distributed near each minimum in a range of ±0.5 Å in the radial coordinate and ±12° in each of the angular dimer coordinates, with the monomer coordinates fixed at their values for a given minimum configuration. The number of points assigned to each minimum is proportional to the magnitude of the PES energy. The procedure for locating the minima is described in section IV.

II.C. Asymptotic Grid Generation Procedure. Interaction energies at large monomer separation are computed from charge distributions and frequency-dependent density susceptibility functions (FDDSs, also called density−density response functions) of each monomer, as described in detail in ref 12. In the rigid case, these quantities only need to be computed once for each monomer, and interaction energies can then be computed in negligible time for arbitrary dimer configurations. In the flexible monomer case, it is instead necessary to generate a grid of monomer coordinates $\boldsymbol{R}_i^A$ for monomer A, denoted G_A. In the case of a heterogeneous dimer, a separate grid G_B is created for monomer B, otherwise we set $G_B = G_A$. The charge distributions and FDDSs are then computed for each element of G_A and G_B. Each element of the monomer product grid $G_A \times G_B$ can be treated as a pair of rigid monomers, which is then combined with a separate grid of dimer coordinates $\boldsymbol{R}_i^{rigid}$. The union of all such sets gives the full asymptotic grid. The use of the product form $G_A \times G_B$ for the monomer coordinates allows many rigid-monomer configurations to be generated from a much smaller number of monomer properties, which is desirable because the overall cost of computing the asymptotic interaction energies is dominated by the monomer calculations. For the dimer coordinates $\boldsymbol{R}_i^{rigid}$, where there is no such constraint, we randomly generate a separate grid for each element of $G_A \times G_B$.

The monomer grids G_A and G_B are created using the same normal distribution $\mathcal{N}(0, \sigma^2)$ described in section II.B but without the refinement stages. The dimer grids are generated as in the rigid autoPES method, described in section II D of ref 12. The same ranges for the IDOFs are used here as in section II.B. One of the elements of G_A is fixed at the undeformed monomer $\boldsymbol{R}_0^A$. We set the number of grid points in the monomer grids to $|G_C| = 5N_{IDOF}^C$, C = A or B, including the point $\boldsymbol{R}_0^C$. The total number of points N_{grid}^{asym} in the dimer grid is set to $300|G_A \times G_B|$, that is, 300 dimers are associated with each element of $G_A \times G_B$.

III. ANALYTIC FIT

The total fitted electronic energy U_{AB} of a dimer at a grid point i can be written as the sum of the interaction energy (potential) V_{int} and of monomer energies U_C, C = A and B

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

$$U_{AB}(\boldsymbol{R}_i) = V_{int}(\boldsymbol{R}_i) + U_A(\boldsymbol{R}_i^A) + U_B(\boldsymbol{R}_i^B) \quad (8)$$

where $\boldsymbol{R}_i$ is the multidimensional variable defined in eq 2. We denote the corresponding energies obtained directly from electronic structure calculations using E rather than U or V, so we also have

$$E_{AB}(\boldsymbol{R}_i) = E_{int}(\boldsymbol{R}_i) + E_A(\boldsymbol{R}_i^A) + E_B(\boldsymbol{R}_i^B) \quad (9)$$

with E_{AB}, E_A, and E_B defined as the total energies of the dimer, monomer A, and monomer B, respectively. Notice that this definition implies that all the energies are calculated with a given monomer always at the same geometry, so all interaction potentials are "vertical". As already mentioned, we use the convention that the monomer energies have zero values at the monomer reference geometries. The reference monomer geometries do not necessarily need to be their equilibrium geometries.

The dimer interaction potentials are the leading component of the interaction energy of a many-body system, that is, a cluster with more than two monomers. The sum of dimer interaction potentials for all pairs of molecules within a cluster gives so-called two-body (or pairwise additive) contribution to this cluster energy. The monomer terms in eq 8 give the so-called one-body contributions. Thus, the dimer PESs such as developed here allow building one- plus two-body parts of the energy of an arbitrary cluster. The remaining part of a cluster energy is called pairwise nonadditive many-body energy within the many-body expansion formalism.[62] The total interaction energy in molecular systems is dominated by the pairwise interaction terms. The pairwise nonadditive terms contribute up to about 20%, the upper limit reached for highly polar systems such as water.[63] For such systems in bulk, pairwise nonadditive energies are very well represented by the polarization model.[29] Therefore, the PESs developed here include such a model. Once this model is fitted in dimer calculations, it can be trivially extended to any many-body case.

While our training data consists exclusively of energies computed at each grid point, it is also possible to make use of analytically computed energy gradients in fitting PESs.[64,65] We have not done so for the following reasons. First, the SAPT codes that we use as the main electronic structure tool do not have gradients. The point of view of SAPT developers has been that SAPT data should be fitted by analytic surfaces and then calculations of analytic gradients is trivial. The second reason is that calculations of analytic gradients are expensive at the levels of theory adequate for developments of intermolecular PESs, such as CCSD(T). Time spent on computing gradients can be used instead to compute more energy grid points. The latter strategy also allows a better coverage of the configuration space, which is complicated in the case of intermolecular interactions (for larger monomers, surfaces have dozens of nonequivalent minima).

When doing regression, it is common practice to first map the input space into a higher-dimensional feature space, which is then used to express the functional dependence. For molecular systems, a natural choice of feature space is the set of site−site distances. Thus, while we defined grid points in terms of $\boldsymbol{R}_i$ coordinates, the fits are expressed entirely in terms of site−site distances. This representation also allows to impose the correct symmetry on PESs by applying appropriate constraints to parameters associated with specific fitting sites or site-pairs, as described below.

Any two atoms of the same charge, which share identical environments within the monomer or which can be made to share identical environments by some monomer deformations which do not break or form covalent bonds, must be treated as symmetry-equivalent. Such an assignment of site types is shown for the example of ethylene glycol in Figure 1. Note that

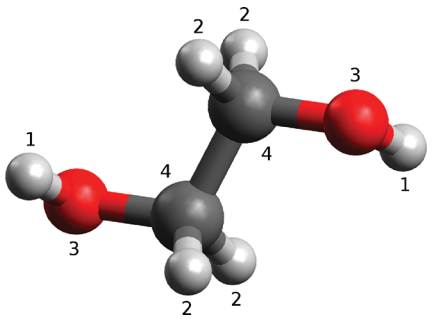


Figure 1. Most general possible assignment of atom types in the ethylene glycol molecule, which preserves the correct permutational symmetry under the constraint that covalent bond breaking is prohibited.

we use separate parameters for the hydrogens bonded to the carbons and the ones bonded to the oxygens; in this way, the model is able to account for their distinct electronic environments. Given the assignment of site types, we define symmetry-distinct site-pair types as any unique combination of two (possibly the same) site types and, similarly, site-pair-pair types as any unique combination of two (possibly the same) site-pair types. This could be continued, of course, but we do not use terms beyond those dependent on site-pair-pairs. Our functional form treats intramonomer site-pair types as distinct from intermonomer site-pair types, which means that we do not impose symmetry with respect to exchange of atoms between distinct monomers (except for complete exchange of the two monomers in the homogeneous case). For example, in the water dimer, the PES will give identical results upon exchanging the positions of the two hydrogens within a given monomer but exchanging the positions of two hydrogens from distinct monomers will give incorrect results. This approach is appropriate because we only seek to model noncovalent interactions and do not allow bond breaking. While it would be possible to automate the assignment of site types to each of the atoms or off-atomic sites of the monomers, this task is straightforward and in the current version of autoPES is left to the user.

Consider a function F representing the dependence of a PES on site-pair-pairs in the general form

$$F = \sum_{s_1,s_2 \in S} \sum_{s_3,s_4 \in S} f_{t(s_1,s_2,s_3,s_4)}(r_{12}, r_{34}) \quad (10)$$

where S denotes the set of sites in the dimer, $t(s_1, s_2, s_3, s_4)$ denotes the site-pair-pair type of the site-pairs (s_1, s_2) and (s_3, s_4), r_{ij} is the distance between sites i and j, and $f_{t(s_1,s_2,s_3,s_4)}$ is an arbitrary function of two real numbers. The exchange of any two symmetry-equivalent sites only changes the order of terms in eq 10. Thus, the problem of enforcing the correct permutational symmetry is reduced to constraining parameters associated with all site-pair-pairs of the same type to have the same value, which ensures that the functions $f_{t(s_1,s_2,s_3,s_4)}$ are indeed only dependent on the type $t(s_1, s_2, s_3, s_4)$ and not on

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

the specific sites. In addition to terms of the form shown in eq 10, the full PES U_{AB} includes terms that depend only on site-pairs or, in the case of the polarization model, on the relative positions of neighboring atoms rather than just their distances. The permutational symmetry follows from similar arguments in these cases.

III.A. Isolated Monomer Fits. At a given level of accuracy, generation of the intramolecular component of the PES is generally easier than generation of the intermolecular component due to the fact that the intramolecular component is only required to model a relatively small region of the configuration space near a single reference configuration. An exception are soft degrees of freedom, but we allow only for a relatively small number of such variables. In some applications, the use of a standard biomolecular force field, such as OPLS-AA,[53] to model the intramonomer component in conjunction with a system-specific intermonomer component may suffice. We include, here, a straightforward automated fitting system for the intramolecular component.

For simplicity, the intramonomer component $U_A(\boldsymbol{R}^A)$ is fitted using the same feature space of atom−atom distances (we do not use off-atomic sites in the monomer fits), restricted to those atoms within monomer A.

$$U_A(\boldsymbol{R}^A) = P_0 + \sum_{a_1\in A}\sum_{a_2\in A} P_l^{a_1a_2}(r_{a_1a_2} - \overline{r}_{a_1a_2}) + \sum_{(a_1,a_2,a_3,a_4)\in A^4} P_{l'}^{a_1a_2a_3a_4}(r_{a_1a_2} - \overline{r}_{a_1a_2}, r_{a_3a_4} - \overline{r}_{a_3a_4}) \tag{11}$$

where the symbol A^4 denotes the set of all atom-pair-pairs in the monomer (and similarly for monomer B). Some terms will be set to zero in cases of partial monomer flexibility, see below. The constant P_0 is treated as an additional fitting parameter and is needed since in general $r_{a_1a_2} - \overline{r}_{a_1a_2} \neq 0$ at $\boldsymbol{R}_0^A$. The polynomials P are defined as

$$P_l^{a_1a_2}(r_{a_1a_2} - \overline{r}_{a_1a_2}) = \sum_{i=1}^{l} c_i^{a_1a_2}[r_{a_1a_2} - \overline{r}_{a_1a_2}]^i \tag{12}$$

and

$$P_{l'}^{a_1a_2,a_3a_4}(r_{a_1a_2} - \overline{r}_{a_1a_2}, r_{a_3a_4} - \overline{r}_{a_3a_4}) = \sum_{i=1}^{l'}\sum_{j=1}^{l'} c_{ij}^{a_1a_2,a_3a_4}[r_{a_1a_2} - \overline{r}_{a_1a_2}]^i[r_{a_3a_4} - \overline{r}_{a_3a_4}]^j \tag{13}$$

The distance $\overline{r}_{a_1a_2}$ denotes the arithmetic mean of all site−site distances $r^0_{aa'}$ in the reference monomer, where aa' is the same site-pair type as a_1a_2. This averaging is necessary to preserve the correct permutational symmetry of the PES (using $r^0_{a_1a_2}$ in place of $\overline{r}_{a_1a_2}$ in eq 11 would result in nonidentical functions between site-pairs of the same type). Note that the value of $\overline{r}_{a_1a_2}$ may be very different from an individual $r^0_{a_1a_2}$. For example, $\overline{r}_{23}$ in Figure 1 is an average of the distances from one of the O atoms to all four of the carbon-bonded H atoms. This is a consequence of our assumption that site-pairs involving equivalent sites are equivalent. We could have relaxed this assumption and made site-pairs involving an oxygen and the hydrogen atoms bound to the same carbon that the oxygen is bound to different from pairs involving hydrogens bound to the other carbon, but the more restricted form proved to be adequate for our purposes.

We have opted to express the monomer PESs in terms of polynomials of site−site distances rather than as functions of bond angles and dihedral angles primarily because it is more difficult to automatically enforce the correct permutational symmetry for arbitrary systems and arbitrary deformations in the latter case. We show in section VIII that this functional form is able to reproduce the correct angular dependence.

While we use the convention that $E_C(\boldsymbol{R}_0^C) = 0$, the value $U_C(\boldsymbol{R}_0^C)$ will generally differ slightly from zero due to fitting error. Note that $E_C(\boldsymbol{R}^C)$ can be less than zero if $\boldsymbol{R}_0^C$ is chosen as a minimum other than the global one.

III.B. Intermolecular Fit. In the rigid-monomer case, the dimer interaction energy PES has the functional form (see ref 12, section IV)

$$V_{int}(\boldsymbol{R}^{rigid}) = V_{ind}(\boldsymbol{R}^{rigid}) + \sum_{a\in A}\sum_{b\in B} u_{ab}(r_{ab}) \tag{14}$$

The first term is a two-body polarization model (which can be extended to many bodies in a straightforward way), and the second term is a sum of functions of site−site distances, where the first and second sites belong to monomers A and B, respectively. In addition to atoms, the set of sites may include off-atomic sites, which serve to reduce uncertainty of the model. The site−site function $u_{ab}(r_{ab})$ is given by

$$u_{ab}(r_{ab}) = A_u P_k^{ab}(r_{ab}) e^{\alpha^{ab} - \beta^{ab} r_{ab}} + \frac{A_{12}^{ab}}{(r_{ab})^{12}} + f_1(\delta_1^{ab}, r_{ab})\frac{q_a q_b}{r_{ab}} - \sum_{n=6,8,10} f_n(\delta_n^{ab}, r_{ab})\frac{C_n^{ab}}{(r_{ab})^n} \tag{15}$$

where q_x are partial charges on each site, C_n^{ab} are distributed dispersion coefficients, f_n are Tang−Toennies damping functions,[66] and $P_k^{ab}(r) = 1 + \sum_{i=1}^{k} a_i^{ab} r^i$. The coefficients A_{12}^{ab} are constrained to be positive and are included to ensure the correct repulsive behavior of the potential at very close range (see ref 12 for a discussion of the role of this term). The factor A_u is needed for consistency of energy units and is equal to 1 kcal/mol. Any of the sites in the fit may optionally exclude electrostatic, polarization, dispersion, or exponential components. In the case that one monomer has nonzero charge (or both monomers are charged), the last summation in eq 15 starts with $n = 4$ rather than $n = 6$ to reproduce the correct asymptotic behavior of the induction energy. If the polarization model is not used, the coefficients C_n^{ab} represent both the dispersion and induction energies. If the polarization model is included, C_n^{ab} still account for a small fraction of induction energy that is missed by this model. The first two terms in eq 15 summed over a and b give the exponential component of the PES and will be denoted as V_{exp}. Similarly, the third term will give V_{elst}, and the last term $V^{(2)}$, denoting the electrostatic and second-order contributions, respectively.

The polarization model defining V_{ind} consists of a set of induced point-dipoles $\boldsymbol{\mu}_a^{ind}$ on sites $a \in A$ (and analogously for $b \in B$), defined as

$$\boldsymbol{\mu}_a^{ind} = \alpha_a\left[\mathcal{E}_a + \sum_{b\in B} \boldsymbol{T}_{ab}\boldsymbol{\mu}_b^{ind}\right] \tag{16}$$

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

where α_a is an isotropic polarizability assigned to a site a, $\boldsymbol{\mathcal{E}}_a$ is the damped electric field at point a due to all permanent point charges q_b of monomer B

$$\boldsymbol{\mathcal{E}}_a = \sum_{b\in B} f_1(\delta_1^{ab}, r_{ab}) \frac{q_b \boldsymbol{r}_{ab}}{r_{ab}^3} \tag{17}$$

and $\boldsymbol{T}_{ab}$ is the damped dipole−dipole 3 × 3 interaction tensor

$$\boldsymbol{T}_{ab} = f_3(\delta_p^{ab}, r_{ab}) \left[3\frac{\boldsymbol{r}_{ab} \otimes \boldsymbol{r}_{ab}}{r_{ab}^5} - \frac{\mathbf{1}}{r_{ab}^3} \right] \tag{18}$$

with $\boldsymbol{r}_{ab}$ denoting the displacement vector from site a to site b, $\otimes$ denotes the Kronecker product, and $\mathbf{1}$ is the 3 × 3 unit matrix. After the set of equations is solved iteratively, the energy contribution from the polarization model is given by

$$V_{\text{ind}} = -\frac{1}{2} \sum_c \boldsymbol{\mathcal{E}}_c \cdot \boldsymbol{\mu}_c^{\text{ind}} \tag{19}$$

where the sum is over all atoms c in both monomers. For small molecules, it may suffice to place the induced dipoles at a single site, close to the molecule's COM. For larger molecules, placement on heavy atoms only may be sufficient.

The two-body implementation of the polarization model can easily be extended to more bodies by making the summations in eqs 16 and 17 over atoms in all monomers other than the one in which the atom a resides and summing eq 19 over atoms in all monomers.

The rigid-monomer model described above can be generalized to the flexible-monomer case in the following way

$$V_{\text{int}}(\boldsymbol{R}) = V_{\text{ind}}(\boldsymbol{R}) + \sum_{a\in A} \sum_{b\in B} u_{ab}(r_{ab}, (r_{ab_1}), (r_{a_1b}), (r_{a_1a_2}), (r_{b_1b_2})) \tag{20}$$

$V_{\text{ind}}(\boldsymbol{R})$ is, here, the same as in the rigid case except for being evaluated, in general, for deformed monomers, and using the field of partial charges that are dependent on monomer deformation. The distributed polarizabilities α_a in eq 16 are fitted to the asymptotic induction energies. The site−site functions u_{ab} have been generalized to include dependence on site−site distances in the dimer other than r_{ab}. The tuple (r_{ab_1}) denotes the site−site distances, including off-atomic sites, between site a and any site b_1 in monomer B, and similarly for (r_{a_1b}). The tuples $(r_{a_1a_2})$ and $(r_{b_1b_2})$ denote all the site−site distances in the respective monomer.

The V_{exp} contribution is generalized to

$$V_{\text{exp}} = A_u \sum_{a\in A} \sum_{b\in B} e^{\alpha^{ab} - \beta^{ab} r_{ab}} [P_k^{ab}(r_{ab}) + \sum_{b_1 \neq b} P_{k'}^{ab,ab_1}(r_{ab_1}) + \sum_{a_1 \neq a} P_{k'}^{ab,a_1b}(r_{a_1b}) + \sum_{a_1a_2 \in A^2} P_{k''}^{ab,a_1a_2}(r_{a_1a_2}) + \sum_{b_1b_2 \in B^2} P_{k''}^{ab,b_1b_2}(r_{b_1b_2})] \tag{21}$$

where A^2 and B^2 refer to the sets of all pairs of sites in the respective monomers and the polynomials are defined as

$$P_{k'}^{ab,ab_1}(r_{ab_1}) = \sum_{i=1}^{k'} b_i^{ab,ab_1} [r_{ab_1}]^i \tag{22}$$

and analogously for $P_{k'}^{ab,a_1b}(r_{a_1b})$ and

$$P_{k''}^{ab,a_1a_2}(r_{a_1a_2}) = \sum_{i=1}^{k''} b_i^{ab,a_1a_2} [r_{a_1a_2} - \overline{r}_{a_1a_2}]^i \tag{23}$$

where $\overline{r}_{a_1a_2}$ is defined in the same way as in eq 12, and analogously for $P_{k''}^{ab,b_1b_2}(r_{b_1b_2})$. Note that the second and third terms of eq 21 do not introduce any dependence on the IDOFs, but are included to generalize the intermolecular part of the fit. Used in the flexible-monomer context, it provides additional pliability of the PES to reproduce points with changing intramonomer geometries. The last two terms are expected to provide the major representation of the monomer-flexibility effects on interaction energies. It is the second leading term in the SOP approach. We have not included further terms since eq 21 and the other monomer-flexibility adjustment described below gave satisfactory results.

The remaining terms in eq 15 were generalized as follows. The terms dependent on $1/r_{ab}^{12}$, whose role is to prevent unphysical behavior at very short intermonomer separations, remain in the rigid-monomer form. The generalized partial charges are given by

$$q_a((r_{a_1a_2})) = q_a^0 + \sum_{(a_1,a_2)\in A^2} \sum_{i=1}^{k_q} q_i^{a,a_1a_2} [r_{a_1a_2} - \overline{r}_{a_1a_2}]^i \tag{24}$$

and similarly for q_b. The generalized induction plus dispersion coefficients C_n^{ab} always make use of the combination rule $C_n^{ab} = \sqrt{C_n^a C_n^b}$, where

$$C_n^a((r_{a_1a_2})) = C_n^{a0} + \sum_{(a_1,a_2)\in A^2} C_n^{a,a_1a_2} [r_{a_1a_2} - \overline{r}_{a_1a_2}] \tag{25}$$

and similarly for C_n^b. We have not implemented an expansion of the C_n^a coefficients beyond the linear terms because higher terms were found to have a negligible benefit to the model's accuracy.

IV. FITTING PROCEDURE

Electronic structure calculations on grids generated as described in section II result in a set of asymptotic interaction energies on the asymptotic grid, a set of close-range interaction energies on the close-range grid, and a set of intramonomer energies on the isolated monomer grid (or two intramonomer sets in the case of a heterogeneous dimer). We will sometimes refer jointly to the energy plus geometry data as a grid point. As in the rigid-monomer case, the asymptotic energies are fitted first, and the resulting parameters are kept fixed during the fitting of the close-range parameters. All nonlinear optimizations use Powell's method[67] as implemented in the Praxis Fortran code,[68] and all linear optimizations are done using singular value decomposition as implemented in standard linear algebra libraries.

The monomer fits are performed separately using linear least-squares optimization. The fitting weights for the monomer fits are given by the weight function $w(E; 0, \eta_{\text{fit}})$, which is of the same form as the one used in the grid generation, given by eq 7. The weight strength parameter η_{fit} has the default value of 5.

In general, some of the intramonomer distances $r_{a_1a_2}$ will not vary with any of the IDOF coordinates $X_{A,i}^j$, because the PES may be constrained to some subspace of the full-dimensional configuration space, depending on which IDOFs $\boldsymbol{D}_A^j$ are included. For example, in the ethylene glycol dimer PES, where

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

we have modeled only rotational degrees of freedom, the distances between bonded atoms are fixed. Each site-pair and each site-pair-pair is, therefore, categorized according to whether it is variable within the user-specified constraints. Any site-pair (a_1, a_2) such that $|r_{a_1a_2} - r^0_{a_1a_2}|/r^0_{a_1a_2} < 0.02$ for all values of the IDOF coordinates within the ranges $(X^j_{A,min}, X^j_{A,max})$ is considered nonvariable (and similarly for monomer B). The corresponding parameters b_i^{ab,a_1a_2}, q_i^{a,a_1a_2}, and C_n^{a,a_1a_2} are all set to zero, as are the parameters $c_{ij}^{a_1a_2,a_3a_4}$ in the case that either of the pairs (a_1, a_2) or (a_3, a_4) are nonvariable. For the isolated monomer potential, this means that the PES will not necessarily have reasonable behavior if the fit is evaluated for atomic positions which do not correspond to any possible choice of the IDOF coordinates. Therefore, the PES is only useful for applications in which the monomer geometries are properly constrained. This is the case in many crystal structure prediction (CSP) programs, so that CSPs are the main intended application of PESs with a partial account of monomer flexibility. In applications that cannot recognize rigid fragments of a monomer, one can simulate rigidity by adding to eq 11 harmonic potentials with a large force constant for those coordinates that are not included in the fit, so that motions in these coordinates are strongly restrained. An alternative approach would be to use a generic force field to describe the remaining coordinates of isolated monomers and possibly reparameterize it if higher accuracy is required.

IV.A. Asymptotic Fitting. After charge densities and FDDSs are computed for each of the monomer geometries in G_C (defined in Sec. II.C), C = A or B, COM-based or distributed[69] multipole moments and static and dynamic polarizabilities are obtained from them. This asymptotic expansion is used to find the values for the electrostatic, induction, and dispersion energy components, denoted $E_{elst}(\boldsymbol{R}_i)$, $E_{ind}(\boldsymbol{R}_i)$, and $E_{disp}(\boldsymbol{R}_i)$, at each grid point $\boldsymbol{R}_i$ in the asymptotic fitting grid, described in section II.C. Because the asymptotic form of CCSD(T) theory is unknown, asymptotic interaction energies are computed at the same level of theory regardless of what electronic structure method is used for the close-range grid. While this unfortunately means that the connection to asymptotics cannot be seamless when CCSD(T) is used, the difference between the two theories is less than one percent for water at asymptotic separations, see Table I of ref 11 for a comparison.

Fitting of the asymptotic model proceeds in three stages: electrostatic, polarization, and induction plus dispersion. The fitting weights are identical to those used in ref 12. As we did not introduce any explicit dependence of the polarization model V_{ind} on the IDOFs except for the partial charges, that part of the fitting remains unchanged from the earlier work. The atomic polarizabilities α_a and α_b are simply fitted to the induction energy component of the asymptotic grid, which in the flexible-monomer case includes dimers with deformed monomer geometries and partial charges that depend on monomer deformations. Thus, α_c are averages of some monomer-deformation-dependent distributed polarizabilities. Because the asymptotic induction energies $E_{ind}(\boldsymbol{R}_i)$ are computed at the level of second-order perturbation theory, the polarization model is iterated only once when fitting the α_c parameters, rather than iterating to convergence as is done when evaluating the PES.

The partial charges q_a^0 of monomer A are fitted under the constraint that the total monomer charge is equal to $\sum_{a\in A} q_a^0$. The total monomer charge is kept independent of monomer deformations by the constraint $\sum_{a\in A} q_i^{a,a_1a_2} = 0$ for all $(a_1, a_2) \in A^2$ and all i from 1 to k_q. The electrostatic fitting is done in two stages. The partial charges q_a^0 and q_b^0 are fitted first (on the flexible-monomer grid), with the $q_a^{a_1a_2}$ and $q_b^{b_1b_2}$ fixed at zero. The $q_a^{a_1a_2}$ and $q_b^{b_1b_2}$ parameters are then fitted together with the q_a^0 and q_b^0 parameters, with the q_c^0 values from the previous stage used as starting points for optimization. As in other cases where multiple fitting stages are used, the purpose of first fitting the simplified model is to aid the nonlinear optimization toward the global minimum of the objective function. Note that one consequence of this fitting procedure is that our fit is not the optimal one for the $\boldsymbol{R}_0^A$ and $\boldsymbol{R}_0^B$ monomer configurations.

A simple weighted least-squares fit of the partial charges to asymptotic electrostatic energies sometimes results in charges that are unphysically large in magnitude. We have, therefore, used a constraint based on partial charges obtained using the electrostatic potentials on grid (CHELPG) method.[70] The constrained fitting minimizes the error functional

$$\chi^2 = \frac{1}{N_{grid}^{asym}} \sum_{i=1}^{N_{grid}^{asym}} w_i^{elst} [V_{elst}(\boldsymbol{R}_i) - E_{elst}(\boldsymbol{R}_i)]^2 + \frac{C_{els}}{N_{ST}} \sum_{c=1}^{N_{ST}} [q_c^0 - q_c^{CHELPG}]^2 \quad (26)$$

where N_{ST} is the number of symmetry-distinct atom types in the dimer. The CHELPG partial charges q_c^{CHELPG} for each atom type c are computed using ORCA[71] at the reference geometry. The value of the parameter C_{els} is set equal to the value of the first summation in eq 26, evaluated using the CHELPG values as the partial charges. We use weights w_i^{elst} from ref 12, multiplied by an additional factor $w(U_A(\boldsymbol{R}_i^A) + U_B(\boldsymbol{R}_i^B); 0, \eta_{fit})$, with $\eta_{fit} = 5$ as for the intramolecular fits.

The correct permutational symmetry of the PES is enforced by setting q_i^{a,a_1a_2} in eq 24 equal for all sites a of a given symmetry-equivalent site type and all (a_1, a_2) of a given symmetry-equivalent site-pair type. In the interest of making the PESs physically reasonable, we want to enforce some degree of locality to the transfer of charge within the monomers. For this reason, all parameter values q_i^{a,a_1a_2} in eq 24 for which $\min(r^0_{aa_1}, r^0_{aa_2}) > 3$ Å are set to zero. If there are multiple triples of sites sharing the same site types, then the q_i^{a,a_1a_2} values are only set to zero in the case that the site−site distances of all such triples exceed the 3 Å cutoff.

The parameters of the induction plus dispersion model are fitted similarly to those of the electrostatic model, except that there is no charge neutrality constraint or CHELPG-like penalty function. We have used the constraint that $C_n^{a,a_1a_2} = C_6^{a,a_1a_2} C_n^{a0} / C_6^{a0}$ for $n > 6$ to reduce number of free parameters in the $V^{(2)}$ part of the model.

IV.B. Complete Intermolecular PES Fitting. Optimization of the remaining free parameters of the PES uses the close-range dimer grid, while the asymptotic parameters are kept fixed. The damping parameters δ_1^{ab}, δ_n^{ab}, and δ_P^{ab} of eqs 15 and 18 are assumed to be independent of monomer geometry but are optimized using the electrostatic, induction, exchange-induction, dispersion, and exchange-dispersion components of the SAPT interaction energies on the flexible-monomer grid. See section VI of ref 12 for a discussion of this fitting stage. If close-range energies are computed using the supermolecular

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

method rather than SAPT, the damping parameters are instead optimized together with other parameters to the total interaction energies. The A_{12}^{ab} parameters are set using a combination of the interaction energies and the covalent radii of the atoms (see below). The remaining nonlinear parameters, α^{ab} and β^{ab}, are optimized using the interaction energies. The error functional is given by

$$\chi^2 = \sum_{i=1}^{N_{\text{grid}}} [V_{\text{int}}(\boldsymbol{R}_i) - E_{\text{int}}(\boldsymbol{R}_i)]^2 w_i + P_A \tag{27}$$

where the fitting weights w_i are given by

$$w_i = w(E_{\text{int}}(\boldsymbol{R}_i) + U_{\text{A}}(\boldsymbol{R}_i^{\text{A}}) + U_{\text{B}}(\boldsymbol{R}_i^{\text{B}}); U_{\text{AB}}(\boldsymbol{R}_{\text{min}}), \eta_{\text{fit}}) \tag{28}$$

where $\eta_{\text{fit}} = 5$ as before and $\boldsymbol{R}_{\text{min}}$ is the geometry of the global minimum of U_{AB}. We compute the total energy used for the weights by combining $E_{\text{int}}(\boldsymbol{R}_i)$ with intramolecular energy components computed using the PESs U_{A} and U_{B} because the values $E_{\text{A}}(\boldsymbol{R}_i^{\text{A}})$ and $E_{\text{B}}(\boldsymbol{R}_i^{\text{B}})$ are not available (although monomer energies computed in the dimer basis are available, see section II.A). The penalty term P_A is described below. At each evaluation of χ^2 during the nonlinear optimization, the linear parameters a_i^{ab}, b_i^{ab,a_1a_2}, b_i^{ab,b_1b_2}, b_i^{ab,ab_2}, and b_i^{ab,a_2b} of eq 21 are fitted using the linear least-squares method.

While the PESs of the type developed here are accurate at arbitrarily large intermonomer separations by construction, they break down at energies above some threshold on the repulsive wall. This threshold is determined by the weighting parameter η_{fit} and by the parameter E_{wall}, which is discussed in ref 12. and later in section VII here. The PESs described in section VIII are accurate up to at least 20 kcal/mol above the dissociation limit (although the repulsive wall extends beyond this value in most configurations), which is adequate for most molecular dynamics applications. For applications that require accuracy at higher energies, the height of the repulsive wall can be increased at the possible cost of decreasing accuracy in the low-energy region.

Two modifications were made to the fitting procedure of ref 12, which are not related to the introduction of flexible monomers. First, the procedure for setting the A_{12}^{ab} parameter values was improved. As in ref 12, each A_{12}^{ab} parameter is dependent on the sum of the covalent radii of atoms a and b. Unlike before, they are given by

$$A_{12}^{ab} = A_{\text{s}}(r_a^{\text{cov}} + r_b^{\text{cov}})^{12} \frac{\text{kcal}}{\text{mol}} \tag{29}$$

where A_{s} is a dimensionless overall scale parameter, which is optimized simultaneously with the α^{ab} and β^{ab} parameters (as a nonlinear parameter). The goal is to make the $A_{12}^{ab} r_{ab}^{-12}$ terms as large as possible without significantly increasing the fitting error to maximize the tendency of the repulsive wall to have the correct monotonically increasing behavior with decreasing intermonomer separation. This goal is reinforced by the P_A penalty term of eq 27, which is given by

$$P_A = -10^{-5} \log(A_{\text{s}}) \sum_{i=1}^{N_{\text{grid}}} [E_{\text{int}}(\boldsymbol{R}_i)]^2 w_i \tag{30}$$

The logarithmic dependence in eq 30 is used to give reasonable weight strength over a large range of values of A_{s}. This is important because the A_{12}^{ab} parameters can vary by several orders of magnitude between different systems, corresponding to only small shifts in the location of the repulsive wall. The second change relative to ref 12 was to remove the dependence of the fitting weights $w(E; E_0, \eta)$ on the values of the (partially optimized) PES, described in eq 20 of ref 12. This change was due to instability of the fitting convergence in rare cases.

V. OFF-ATOMIC FITTING SITES

One effective way to improve the accuracy of a PES is to introduce off-atomic fitting sites. Such sites act as centers for the site−site functions u_{ab}, in addition to the positions of the atomic nuclei. The locations of off-atomic sites within the monomers are usually determined by a combination of physical intuition and optimization by minimization of the fitting error. For rigid monomers, once the positions of the off-atomic sites are determined, they are simply rotated and translated with the monomer. For flexible monomers, the situation is more complex, as some choice must be made regarding how the off-atomic sites are to be positioned in an arbitrarily deformed monomer. While many such choices are possible,[46,47] we have implemented two options which address most of the common use cases.

The first option is appropriate for off-atomic sites which are located near a noncolinear triple of atomic nuclei. The projection of the off-atomic site location onto the plane of the three atoms, denoted $\boldsymbol{p}_{\text{oa}}$, can be written as a linear combination of the positions of the three nuclei, $\boldsymbol{p}_{\text{oa}} = c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2 + c_3\boldsymbol{r}_3$, with $c_1 + c_2 + c_3 = 1$. The coefficients c_1, c_2, and c_3, together with the distance from the off-atomic site to the plane, are determined using the undeformed monomer geometry. These four numbers are then used to position the off-atomic site relative to the three atoms for any deformed geometry by using the same linear combination of the new nuclear positions, and the same distance to the plane. So long as the nuclei do not become approximately colinear as a result of the monomer deformations, this procedure gives a stable and physically reasonable way to position the off-atomic sites. In cases where the nuclei may become colinear, this option should only be used when the off-atomic site is positioned exactly on the plane.

The second option is primarily useful for off-atomic sites that are placed on atomic bonds, or for linear molecules. The off-atomic site location $\boldsymbol{r}_{\text{oa}}$ is assumed to be colinear with two nuclei and is expressed as $\boldsymbol{r}_{\text{oa}} = c_1\boldsymbol{r}_1 + c_2\boldsymbol{r}_2$, with $c_1 + c_2 = 1$. As before, the reference monomer geometry is used to determine c_1 and c_2, which are then kept fixed for arbitrary monomer deformations. Often, the site is placed halfway between the atoms, with $c_1 = c_2 = 0.5$.

VI. MINIMA SEARCH

The fitted PESs are searched for all symmetry-distinct local minima. These configurations are of direct physical interest in many contexts, and are used in part of the iterative grid enlargement procedure (see section VII). To reliably locate all local minima on a high-dimensionality surface is not trivial, and we use an evolutionary algorithm (EA) for the purpose. Such algorithms have been shown to be able to reliably locate minima on intermolecular PESs.[72]

As in standard EAs, new population members, in this case dimer geometries, are generated by combining two existing members. Different schemes may be used for this combination procedure.[72] We use the straightforward approach of random

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

bitwise splicing of the 8 most significant digits of each of the coordinates (both inter- and intramonomer) described in Sec. II. In a typical EA, the least-fit existing population member (i.e., the member with the greatest interaction energy) would be replaced by the newly generated member in each iteration. We follow instead Algorithm 1 of ref 73. In addition to the usual fitness function, this algorithm requires the specification of a separate diversity metric (defined below) to quantify the difference between two dimer configurations. After each new member is generated, a local fitness maximization (i.e., energy minimization) is performed on it, and the diversity metric is then evaluated between the resulting locally optimized dimer configuration and each of the other existing population members. A replacement of the least-fit existing member only occurs if the new member is sufficiently distinct from any of the others. The effect of this procedure is to maintain diversity of the population, which is needed to generate a comprehensive list of local minima; in more standard EAs, the entire population would tend to converge to the global minimum, which is insufficient for this purpose.

The EA is implemented using a modified version of the Pikaia Fortran routines.[74,75] The local fitness maximization stage was performed using Powell's algorithm, in the coordinates described in section II. For full-dimensional PESs, it would have been simpler to instead use Cartesian coordinates of all atoms, but our procedure is also applicable to partially flexible monomers. For the diversity metric, we have used the epsilon metric described in ref 12. The procedure described here is stochastic and so cannot guarantee convergence to a complete list of minima but has been found to perform reliably in all tested cases. A fraction of the grid points are placed at or near the minima, as described in section II.B.

VII. ITERATIVE IMPROVEMENT

The overall flowchart for flex-autoPES is very similar to that presented in Figure 1 of ref 12. The main difference is the additional initial step of fitting the monomer PESs. Following the PES fitting and minimum search stages, the quality of the PES is evaluated. If necessary, additional intermolecular grid points are computed, and the PES is then refitted until it converges. The number of grid points in the additional iterations is equal to 30% of the number in the initial iteration. We use two convergence criteria. First, the total close-range intermolecular fitting grid is randomly split into a fitting set and a test set, with 30% of the points in the test set. The RMSE evaluated on the test set must be no more than 20% larger than the RMSE on the fitting set. Second, a hole searching procedure, described below, must not locate any holes in the repulsive wall of the PES. After convergence is achieved, a final fit is performed to the entire data set, and additional iterations are only performed in the (rare) case that new holes appear in this final fit.

Hole scanning is similar to the procedure described in ref 12. Each of the five Euler angles defining the orientation of the dimer is iterated through, and for each orientation, a radial scan is done to check for the existence of the repulsive wall. The required height of repulsive wall is determined by a parameter E_{wall}, which can be adjusted depending on the intended application of the PES. The default value is E_{wall} = 20 kcal/mol, and we use this value for the test cases in section VIII. If the wall is found, a secondary search stage is used, starting from the dimer configuration corresponding to the start of the wall. In this secondary stage, a weak attractive term is added to the PES, and a local energy minimization is performed on the modified surface, in an attempt to move along the wall to find possible holes. See section VII of ref 12 for a more detailed description. To generalize this procedure to the flexible case, starting from each dimer orientation, several radial scans are performed, each using randomized values for the monomer IDOFs (fixed in each radial scan). The number of scans per orientation is equal to $1 + N_A^{IDOF} + N_B^{IDOF}$. The energy minimization stage also keeps the randomized values of the IDOFs fixed while searching only in the dimer coordinates.

VIII. RESULTS

We have generated PESs for three systems: the water dimer, the ethylene glycol [$HO(CH_2)_2OH$] dimer, and ethylene glycol interacting with water. The water dimer is a natural choice of test system because multiple full-dimensional ab initio PESs have been published,[45,48,51] which we use for comparison with the present work. Ethylene glycol is convenient as a larger system with multiple soft IDOFs, and $HO(CH_2)_2OH{\cdot}H_2O$ was chosen as a heterogeneous test case. The systems involving $HO(CH_2)_2OH$ are particularly challenging due to the large range of conformations that are accessible to that monomer and due to the many possible hydrogen bonding configurations. In fact, we have identified about 700 distinct local minima of the $(HO(CH_2)_2OH)_2$ dimer, which are listed in the Supporting Information (SI). There do not appear to be any prior publications of ab initio PESs for $(HO(CH_2)_2OH)_2$ or $HO(CH_2)_2OH{\cdot}H_2O$. In fact, we could not identify any literature work developing PESs with partially flexible monomers. For the water dimer, the surface is full-dimensional, with the H−O−H bond angle and both O−H bond lengths of each monomer able to vary, for a total of 12 dimensions in the dimer. For $HO(CH_2)_2OH$, we model only three IDOFs per monomer, indicated in Figure 2, so the

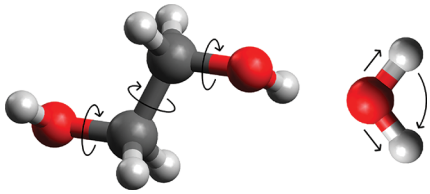


Figure 2. Ethylene glycol and water molecules in their $\boldsymbol{R}_0^A$ geometries, with modeled IDOFs indicated by arrows.

$HO(CH_2)_2OH{\cdot}H_2O$ and $(HO(CH_2)_2OH)_2$ surfaces are also 12-dimensional. The software is capable of modeling all degrees of freedom of the ethylene glycol dimer, but the additional free parameters would require many more grid points to fit. The rotational IDOFs of $HO(CH_2)_2OH$ accommodate for the full 360° range of motion, which includes the regions of both the local and global minima of the monomer surface. Full specifications of fitting parameters and data are given in the SI.

For the water dimer, we use the all-electron CCSD(T) method as implemented in the MOLPRO package.[77] Energies were extrapolated from calculations using the aug-cc-pVTZ (aTZ) and aug-cc-pVQZ (aQZ) basis sets of Dunning and co-workers[78] plus midbond functions, using the extrapolation scheme described in ref 11. Midbond function exponents and placement are as described in ref 79, with the same midbond functions used for the aTZ and aQZ basis sets. For the

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

interaction energies needed in the close-range grid, we use the counterpoise-corrected[80] supermolecular CCSD(T) method, whereas for the intramolecular grid, total CCSD(T) energies are computed in the monomer basis. The $\boldsymbol{R}_0^A$ reference geometry for the water monomer was obtained by optimization at the same level of theory as applied for the intramolecular grid using MOLPRO. We have utilized the five off-atomic site locations from the fit published in ref 81, referred to as SAPT-5s. The positions of these sites are adjusted with the deformations of the monomers using the planar atom-following mode defined in section V. Table I compares the present value of the ab initio energy to the most accurate literature calculations of Lane.[76] This table shows that the basis set incompleteness error of our surface is 0.04 kcal/mol. The remaining corrections amount to only 0.01 kcal/mol. Thus, the overall uncertainty of our ab initio interaction energies is 0.05 kcal/mol or 1% in the region of the van der Waals (vdW) minimum. This is slightly larger than the uncertainty of the fit, discussed below. Compared to the best published full-dimensional PESs for the water dimer, our basis set and level of theory is essentially the same as that used to develop the MB-pol PES,[51] although the authors of MB-pol have not specified if they correlated all electrons or used the frozen-core (FC) approximation. This approximation introduces an error of about 0.03 kcal/mol near the vdW minimum. The HBB2 PES[48] was developed using only the aTZ basis set which, with the use of the counterpoise (CP) correction, underestimates the magnitude of the interaction energy by at least 0.2 kcal/mol. However, Shank et al.[48] computed also interaction energies without the CP correction (in this case getting an overestimation of the magnitude) and then linearly interpolated between these two values with a factor that recovers the benchmark interaction energy at the minimum computed in ref 82.

Table I. Convergence of Interaction Energy (in kcal mol^{-1}) in Basis Set Size and Theory Level at the Minimum Configuration of Water Dimer[a]

theory level/basis set	U_{AB}^{min}
CCSD(T) CBS(TQ+mb)[b]	−4.9710
CCSD(T)-F12b CBS(Q5)	−5.0103
plus CCSDTQ-CCSD(T)	−5.0184
plus relativistic and DBOC	−5.0208

[a]The interaction energies (U_{AB}^{min}) are computed with respect to the monomer at its equilibrium geometry. All results are from the work of Lane,[76] unless noted otherwise. DBOC denotes the diagonal BO correction. [b]Computed by us using the theory level and basis set of the present work at the geometry in the first column "CCSD(T)-F12b/CBS" of Table 8 in ref 76. All energies are relative to monomers at their isolated-monomer equilibrium geometries taken from the line CCSD(T) and the last two columns of Table 2 in ref 76.

The CCpol-8sfIR PES of ref 45 was fitted to ab initio data in the rigid-monomer sector of similar accuracy to our values, but the dependence on monomer deformations used a lower level of theory. Compared to our recent rigid-monomer PES of ref 11, we have removed the FC approximation. Thus, our ab initio energies are on par or more accurate than those used to develop published full-dimensional PESs for the water dimer. See ref 11 for a more detailed discussion of ab initio methods used to generate literature PESs for the water dimer.

Interaction energies for the asymptotic grid were computed using the distributed asymptotic expansion method of ref 69. Because of software limitations, the asymptotic calculations could only be performed in the aTZ basis set. However, comparisons of energies computed using SAPT(DFT) in the aTZ and aQZ basis sets at several configurations used in the asymptotic grid showed that the electrostatic energy computed in the aTZ basis was consistently within 1% of the aQZ value, and the induction and dispersion energies were consistently between 2% and 3% more negative in the aQZ case. Therefore, for the $(H_2O)_2$ system, we simply used the aTZ values for the electrostatic energy, and scaled the induction and dispersion energies of points in the asymptotic grid by 1.025.

For the $(HO(CH_2)_2OH)_2$ and $HO(CH_2)_2OH{\cdot}H_2O$ systems, we computed interaction energies using SAPT(DFT) in the density fitting version[18,23,24] with the PBE0 functional.[83,84] The $\delta E_{int,resp}^{HF}$ correction[85,86] is included for both systems. The aTZ basis set is used in both cases, with the same midbond functions as for the water dimer. Energies for the intramonomer grid were computed using PBE0 (so the isolated monomer PESs are different in the $(H_2O)_2$ and $HO(CH_2)_2OH{\cdot}H_2O$ fits). We have used two (symmetry-equivalent) off-atomic sites, with the first (second) placed at the midpoint of the triple formed by the first (second) oxygen and the two carbons. These use the planar atom-following mode (which in this case means that they are always located at the midpoint of their respective sets of three atoms). For $HO(CH_2)_2OH{\cdot}H_2O$, the same off-atomic sites are used for each monomer as in the heterogeneous cases. We use the site type assignment from Figure 1, with one additional type for the two off-atomic sites. The reference geometry $\boldsymbol{R}_0^A$ in the case of $HO(CH_2)_2OH$ is not the global minimum of the monomer, but a symmetrized version of a secondary minimum with energy 2.3 kcal/mol higher than the global minimum. The (unsymmetrized) secondary minimum was obtained by geometry optimization using PBE0. In the symmetrized version, the four heavy atoms are coplanar, and the energy of the monomer is invariant when the hydrogens of the hydroxy groups are both reflected about this plane. This is important due to the partial rigidity of the model, which would

Table II. Comparison of Generated Potentials[a]

system	N_{atom}	N_{site}	N_{grid}	N_{iter}	N_{FP}^{lin}	N_{FP}^{nonlin}	$U_{AB}(\boldsymbol{R}_{min})$	$V_{int}(\boldsymbol{R}_{min})$
$(H_2O)_2$	6	18	4758	5	443	41	−5.003	−5.031
$HO(CH_2)_2OH{\cdot}H_2O$	13	20	11311	10	915	50	−10.065	−8.246
$(HO(CH_2)_2OH)_2$	20	24	5507	7	441	30	−16.269	−12.976

[a]The number of linear and nonlinear free parameters in the intermolecular fit, excluding those parameters fitted to the asymptotic grid, are given by N_{FP}^{lin} and N_{FP}^{nonlin}, respectively. The numbers of grid points in the close-range intermolecular fitting grids of the final fits, which are fitted to all points (the union of fitting and test sets), are is given by N_{grid}. The numbers of fitting iterations for the intermolecular PES N_{grid} include hole-fixing iterations but do not include the final fits to all grid points. The geometry of the minimum of $U_{AB}(\boldsymbol{R})$ is denoted $\boldsymbol{R}_{min}$. Energies are in kcal/mol.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

Table III. Comparison of the Uncertainties of Generated Potentials on Fitting and Testing Subsets of the Intermolecular Grid for the Last Iteration for Each System and for the Final Fits[a]

	N_{grid} fit	N_{grid} test	V_{int} fit RMSE	V_{int} test RMSE	$U_A + U_B$ RMSE	U_{AB} fit RMSE	U_{AB} test RMSE	V_{int} final RMSE	U_{AB} final RMSE
					$(H_2O)_2$				
$E_{AB}(R_i) < 0$	198	88	0.015	0.023	0.021	0.028	0.029	0.015	0.026
$E_{AB}(R_i) < 10$	2407	1093	0.038	0.043	0.084	0.091	0.093	0.037	0.091
$E_{int}(R_i) < 0$	1641	731	0.041	0.052	0.238	0.239	0.253	0.039	0.244
$E_{int}(R_i) < 10$	3045	1365	0.058	0.072	0.226	0.228	0.251	0.054	0.235
					$HO(CH_2)_2OH{\cdot}H_2O$				
$E_{AB}(R_i) < 0$	1457	622	0.110	0.132	0.020	0.111	0.134	0.111	0.113
$E_{AB}(R_i) < 10$	5488	2364	0.209	0.232	0.053	0.220	0.239	0.209	0.220
$E_{int}(R_i) < 0$	3320	1443	0.156	0.162	0.101	0.188	0.190	0.153	0.182
$E_{int}(R_i) < 10$	6190	2664	0.237	0.274	0.110	0.260	0.300	0.237	0.260
					$(HO(CH_2)_2OH)_2$				
$E_{AB}(R_i) < 0$	1340	586	0.241	0.277	0.026	0.243	0.279	0.248	0.251
$E_{AB}(R_i) < 10$	3175	1349	0.342	0.411	0.039	0.342	0.411	0.355	0.355
$E_{int}(R_i) < 0$	2370	1015	0.262	0.289	0.040	0.263	0.290	0.277	0.277
$E_{int}(R_i) < 10$	3219	1365	0.351	0.413	0.040	0.352	0.413	0.363	0.363

[a]Numbers of grid points and RMSEs are given for four groupings of points: those with $E_{AB}(R_i) < 0$, with $E_{AB}(R_i) < 10$ kcal/mol, with $E_{int}(R_i) < 0$, and those with $E_{int}(R_i) < 10$ kcal/mol. The values "V_{int} Final RMSE" and "U_{AB} Final RMSE" are for the final fit to the union of fitting and testing subsets and are given relative to this union. RMSEs for the intramolecular PESs are also given relative to the union of both subsets. All energies are in kcal/mol.

Table IV. Comparison of Geometries (in Å and deg) and Interaction Energies (in kcal/mol) at the Minimum for $(H_2O)_2$ with Literature Values[a]

source	r_{OH} free	r_{OH} donor	r_{OH} acceptor	θ donor	θ acceptor	R_{OO}[b]	α	β	U_{AB}^{min}	V_{int}^{min}
CCpol-8s	0.9716	0.9716	0.9716	104.69	104.69	2.9105	6.0	121.88		−5.104[c]
HBB2									−4.98[d]	
MB-pol									−4.9599[e]	
CCpol-8sf										−5.02[f]
CCpol-8sfIR	0.9575	0.9618	0.9590	104.71	104.67	2.9118	5.5	123.5	−5.0163	−5.0285[g]
Lane FC									−4.9840[h]	
Lane	0.9569	0.9641	0.9584	104.854	104.945	2.9092	5.686	123.458	−5.0208	−5.0516[i]
WW19	0.97565	0.97565	0.97565	104.430	104.430	2.9151	5.9472	121.386		−5.1203[j]
flex-autoPES	0.95680	0.96352	0.95887	104.922	104.975	2.9100	5.8341	122.889	−5.0058	−5.0311[k]
exp/theo[l]						2.91 ± 0.005	2 ± 4	123 ± 10		

[a]The water internal angle is denoted by θ; α is the angle that hydrogen-bonded O−H makes with the O−O axis, and β is the angle between this axis and the bisector of the donor. V_{int}^{min} is the vertical interaction energy at the minimum of U_{AB}, whereas U_{AB}^{min} denotes the interaction energy relative to equilibrium monomer geometries. [b]We use R_{OO} rather than R to conform to literature convention. [c]Ref 89. The ab initio and PES interaction energies are identical to the number of digits listed. [d]Ref 48. Interaction energy from the PES, geometry of minimum not available. [e]Ref 51. Interaction energy from the PES, geometry of minimum not available. [f]Ref 44. Value from the PES listed in ref 44 to 3 significant digits only. [g]Ref 45. Values from the PES. [h]Ref 76. Ab initio value at the CCSD(T)/FC level. [i]Ref 76. Ab initio values at the highest level of theory; V_{int}^{min} estimated in ref 45. [j]Ref 11. At $\langle r \rangle_0$ geometry of $(H_2O)_2$, the interaction energy given by PES, the ab initio value is −5.1040 kcal/mol. [k]Present work. [l]Experimental results.[90] The R_{OO} value was revised in ref 81.

feature an artificially high number of symmetrically distinct local minima if the unsymmetrized monomer geometry were used. The root-mean-square deviation between the symmetrized R_0^A geometry and the PBE0-optimized one is 0.03 Å, and the PBE0 monomer energy of the symmetrized geometry is 0.7 kcal/mol smaller in magnitude.

The OPLS-AA force field used in the initial iteration of the grid generation procedure (see section III.B) includes explicit parametrization for water. Ethylene glycol does not have explicit parameter values, but that system was included in the training set,[54] and so OPLS-AA is fairly accurate for all three of the test dimers used here (see the comparisons later on). To use autoPES for systems that are not included in the OPLS-AA training set, the OPLS-AA atom types must be chosen from those available based on similarity to systems which are parametrized. While OPLS-AA is not expected to have good accuracy in such cases, this does not significantly impact the quality of the grid generation, as only qualitative accuracy is required in that role.

The properties of the fits are summarized in Table II. For all three systems, the orders k, k', and k'' of the polynomials appearing in eq 21 were set to 3, 2, and 2, respectively. The parameter k_q appearing in eq 24 was set to 2 for the $(H_2O)_2$ fit, and for the other two systems, where long-range electrostatic forces are much weaker, we used $k_q = 1$. For the intramonomer fits, the polynomial orders l and l' in eq 11 were set to 4 and 3, respectively, for all three systems. While higher order polynomials could be used in the intramonomer case without incurring a significant increase in the overall computational cost of PES generation, they were not found to significantly improve the fit. The many-body polarizable model has been included for $(H_2O)_2$ but not for the other two systems. The heterogeneous $HO(CH_2)_2OH{\cdot}H_2O$ system requires roughly twice as many free parameters as the other two dimers, and a

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

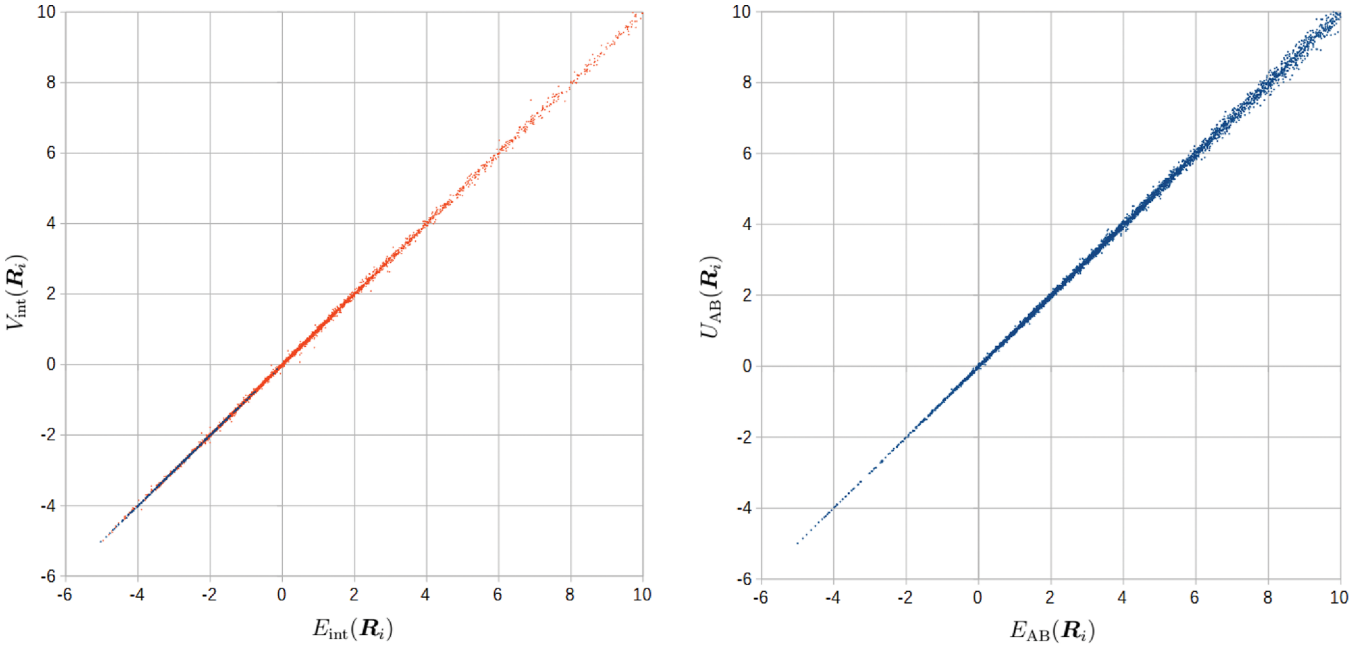


Figure 3. Energy of the final $(H_2O)_2$ PES compared to CCSD(T) energies for all points of the flexible-monomer intermolecular grid. Units are kcal/mol. The left plot compares V_{int} to the supermolecular CCSD(T) energy, while the right plot compares U_{AB} to the sum of supermolecular interaction energy and monomers' CCSD(T) energies. The points shown in blue on the left plot are those with $U_{AB} < 0$. The monomer CCSD(T) energies for points in the intermolecular grid were computed in the monomer basis only for this analysis, and were not used in the normal PES generation procedure (as discussed in section II, the fitting grid for U_C is different than that for V_{int}).

proportionally larger number of grid points. Table II lists also the values of $U_{AB}(\boldsymbol{R}_{min})$ and $V_{int}(\boldsymbol{R}_{min})$. For the water dimer, the two values are close to each other since $\boldsymbol{R}_0^A$ is very close to the equilibrium geometry $\boldsymbol{R}_{eq}^A$ and $\boldsymbol{R}_{min}$ contains monomer geometries close to their equilibrium values. For the ethylene glycol dimer, $\boldsymbol{R}_0^A$ is very different from $\boldsymbol{R}_{eq}^A$, with an energetic difference of 2.3 kcal/mol relative to the minimum, as stated above. The difference between $U_{AB}(\boldsymbol{R}_{min})$ and $V_{int}(\boldsymbol{R}_{min})$, 3.3 kcal/mol, is of the order of $U_A(\boldsymbol{R}_0^A) - U_A(\boldsymbol{R}_{eq}^A) + U_B(\boldsymbol{R}_0^B) - U_B(\boldsymbol{R}_{eq}^B)$, reflecting the fact that $\boldsymbol{R}_{min}$ for the ethylene glycol dimer contains monomers which are actually quite far from $\boldsymbol{R}_{eq}^A$. This illustrates one advantage of using an approach such as flex-autoPES for investigations of systems of this type.

In Table III, RMSEs are given for intermonomer, intramonomer, and total fits for the three systems. As described in section III.B, the fitting weights are dependent on the total energy rather than the interaction energy, and the sets of points with low interaction energy include some points with large total energy and therefore small fitting weights. Note that the fraction of grid points with negative total energy is much lower than the fraction of grid points with negative interaction energy. As we demonstrate below for the water dimer, this is not indicative of poor accuracy in the region of negative total energy. The accuracy of the $(H_2O)_2$ PES is substantially better than that of the other two systems. This is due in part to the relatively large number of off-atomic sites used, and in part to the relatively high rigidity of the water monomer. The large range of conformations which are accessible to the $HO(CH_2)_2OH$ monomer within about 30 kcal/mol of $U_A(\boldsymbol{R}_{eq}^A)$ means that the fits involving this system are required to reproduce a much greater variety of geometries than in the $(H_2O)_2$ case. Note that the quantities actually fitted are $V_{int}(\boldsymbol{R}_i)$ to $E_{int}(\boldsymbol{R}_i)$, so these lines in Table IV characterize the fitting process best. For $(H_2O)_2$, the uncertainty of our final V_{int} fit in terms of RMSE is 0.039 kcal/mol for negative E_{int}. If one considers the points corresponding to $E_{AB} < 0$ (this reduces the number of points from 2372 to 286), the RMSE for negative E_{int} decreases to 0.015 kcal/mol. The reason for this decrease is that the grid points remaining are very close to the bottom of the PES, where the fitting weights are greatest. The sum of monomer fits has RMSE of 0.021 kcal/mol for monomer geometries corresponding to negative E_{AB}, and the total fit U_{AB} has RMSE of 0.026 kcal/mol. In applications of rigid-monomer PESs, the RMSE for negative interaction energy points is the best indicator of the quality of the predictions. It is not obvious that this will be the case for flexible-monomer PESs, so one may have to consider also the RMSEs for $E_{AB} < 10$ kcal/mol, which for the total PES U_{AB} is 0.091 kcal/mol, dominated by the RMSE of $U_A + U_B$. In applications of this PES, we plan to use some literature PES for the monomer, for example, the PTJ2 PES of Polyansky et al.[87] or that of Mizus et al.[88] Substitution of the monomer PESs would not require any modifications to the intermolecular part because of the decomposition of eq 8. With such a choice, the uncertainties of $U_A + U_B$ become negligible and the 0.037 kcal/mol RMSE of V_{int} for points with $E_{AB} < 10$ kcal/mol should characterize the quality of our surface best.

One notable observation from Table III is the large RMSE of $U_A + U_B$ for the water dimer on grid points with $E_{int} < 0$ or $E_{int} < 10$ kcal/mol. This is due to a small number of grid points with low interaction energy, but large intramonomer energy and correspondingly low fitting weights: there are 25 grid points with $E_{int} < 10$ kcal/mol but $U_A + U_B > 30$ kcal/mol, and the RMSE of $U_A + U_B$ on these points is 1.67 kcal/mol. The $(HO(CH_2)_2OH)_2$ PES, in contrast, does not have a particularly high analogous RMSE because the entire intramonomer energy range of that system (within our modeled IDOFs) only spans 23 kcal/mol.

The RMSEs of our fits can also be compared to literature values. The MB-pol[51] PES has an RMSE of V_{int} for all grid

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

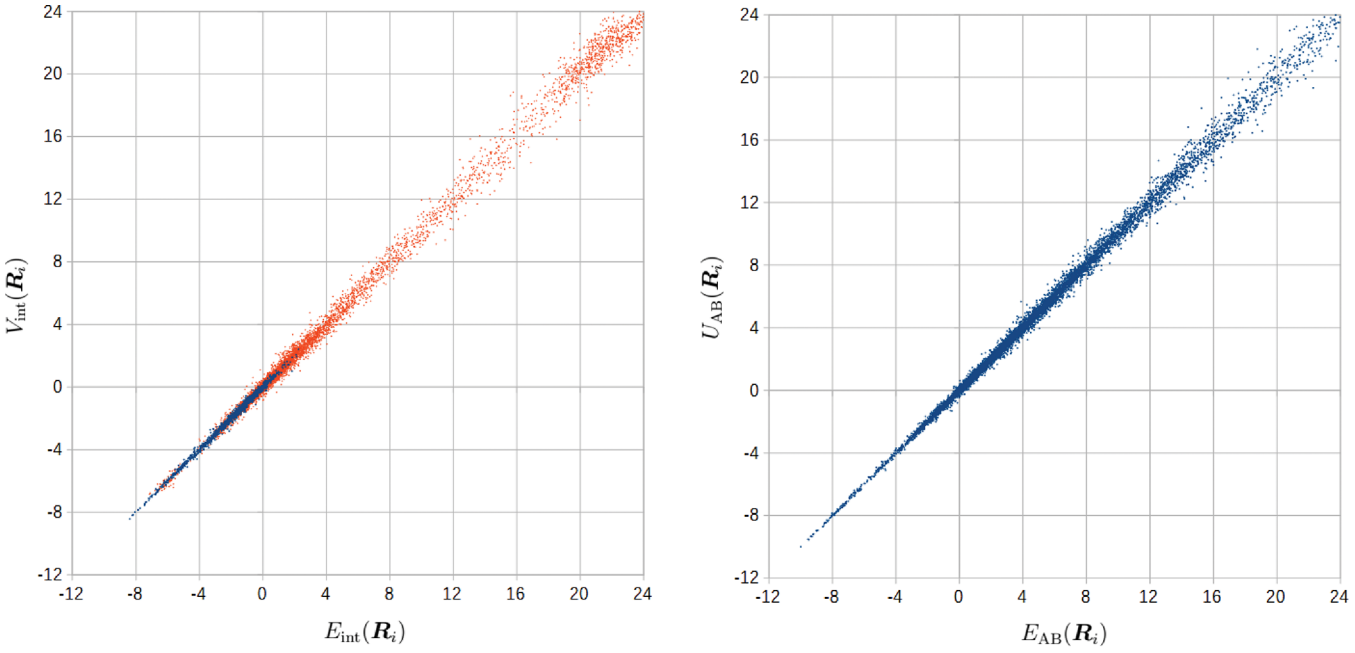


Figure 4. Similar to Figure 3 but comparing energies of the $HO(CH_2)_2OH{\cdot}H_2O$ PES to those computed from SAPT(DFT) and PBE0.

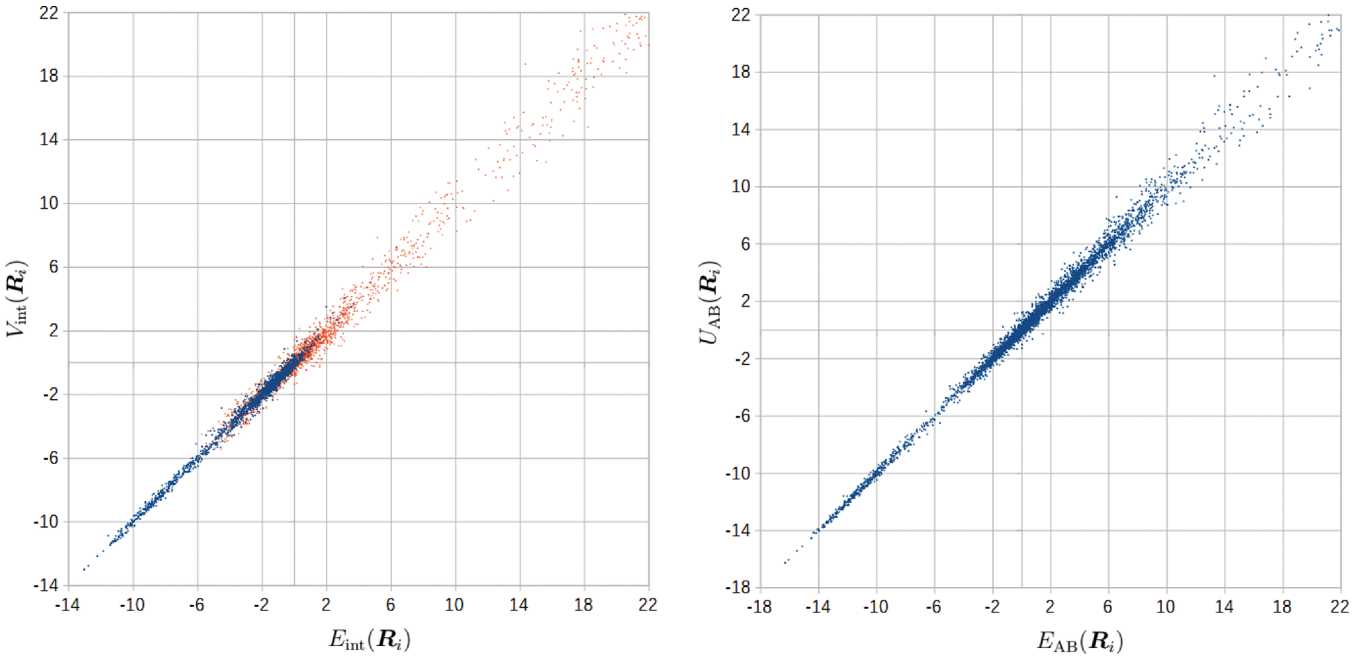


Figure 5. Similar to Figure 4 but for the $(HO(CH_2)_2OH)_2$ PES.

points of 0.054 kcal/mol, whereas the corresponding value for our PES is 0.112 kcal/mol. On the other hand, MB-pol's RMSE of V_{int} for E_{int} < 20 kcal/mol is equal to 0.0008 kcal/mol (0.3 cm^{-1}) on fitting data, whereas our RMSE for E_{int} < 10 kcal/mol amounts to 0.054 kcal/mol. However, the authors of ref 51 do not report error values for points outside of their fitting set. The only test of this type is the RMSE of MB-pol on the stationary points (see the discussion later on) relative to the benchmarks of ref 82, which amounts to 10.0 cm^{-1}, 33 times RMSE of MB-pol in low energy region. The CCpol-8sfIR[45] PES has RMSE of 0.54 kcal/mol for points in the training set with E_{int} < 10 kcal/mol, and HBB2[48] has RMSE of 0.017 kcal/mol for points in the training set with E_{int} < 12 kcal/mol. Thus, the errors of the flex-autoPES fit are similar to those of the best literature 12D PESs for the water dimer, except for MB-pol in the low energy region. [One should also point out that the training set of HBB2 included a fair number of grid points with the total energies of the order of 100 kcal/mol above the global minimum level, while in our case the highest such energy was 108 kcal/mol, but only a few grid points had such high energies. Fitting high-energy points may reduce accuracy of a fit at low energies.] We could have lowered the RMSE of our fit by introducing additional off-atomic sites, but this was, in our opinion, not worth the greater number of required grid points, considering the uncertainties of the ab initio values.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

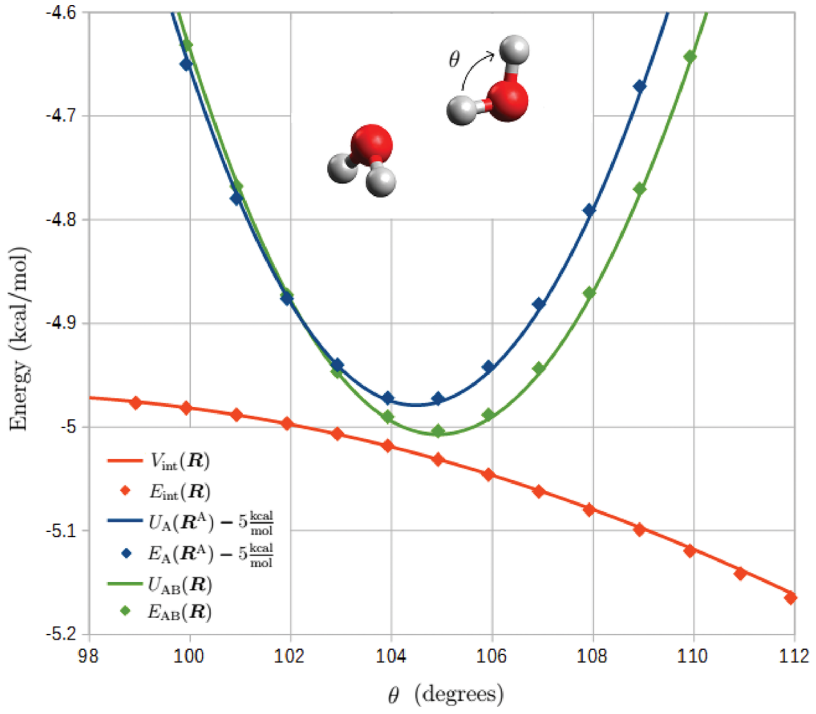


Figure 6. Energy of the water dimer computed from the fit (lines) and using CCSD(T) (symbols) near the minimum dimer configuration as functions of the H−O−H bond angle θ of the hydrogen-bond donor monomer (shown in the inset). All atoms are kept fixed except for the free hydrogen of the donor molecule. For ease of illustration, the isolated monomer energies are shifted by −5 kcal/mol in this figure.

Figures 3−5 compare energies computed from the fits to those computed from the respective ab initio methods for all points in the close-range intermolecular fitting grid. Accuracy of the fit decreases at higher total energies, due to the weighting function. Note the increased density of points in the high energy region in the plots of U_{AB} relative to the plots of V_{int} for all three systems, consistent with the values in Table III. This is due to the large numbers of grid points with U_A + U_A in the energy range between roughly 0 and 10 kcal/mol (or between −2.3 and 10 kcal/mol in the case of $HO(CH_2)_2OH$). As mentioned in section II.B, this is expected due to the large region of physically important monomer configurations within this energy range. Since the interaction energy at the $(H_2O)_2$ minimum is −5.03, the addition of U_A + U_B in this range makes the sum positive in a large number of cases. Note that the plot of interaction energy contains more points than the plot of total energy (4410 versus 3500), although in the former case the bulk of the points are concentrated at low energy very close to the $V_{int}(R_i) = E_{int}(R_i)$ line (this may give the illusion that there are more points in the right panel). For $(HO(CH_2)_2OH)_2$, a low density of points is observed near the minimum due to the small region of configuration space in this energy range: while the global minimum of this system is at $U_{AB}(R_{min}) = -16.3$ kcal/mol, the next lowest minimum is at −14.2 kcal/mol. The number of grid points with energy greater than 10 kcal/mol is much larger for $HO(CH_2)_2OH{\cdot}H_2O$ than for $(HO(CH_2)_2OH)_2$, primarily due to the large number of points added during hole-fixing iterations in the former case: 2635 hole-fixing points were added to the repulsive wall over 8 iterations for $HO(CH_2)_2OH{\cdot}H_2O$, while 1395 points over 5 iterations were used for $(HO(CH_2)_2OH)_2$.

The inter- and intramolecular components of the water dimer PES are shown in Figure 6 as functions of the H−O−H bond angle of the hydrogen-bond donor near the minimum configuration. The angle $\theta = 104.922°$ corresponds to the exact minimum on U_{AB} and $\theta = 104.478°$ to the monomer equilibrium geometry. The geometry R^B of the acceptor molecule is fixed and $U_B(R^B) = 0.007$ kcal/mol at this geometry. As the bond angle increases away from the equilibrium value, the interaction energy grows in magnitude, but the monomer energy increases more rapidly. This gives an angle of the minimum of the total energy that is shifted by 0.44° from the equilibrium monomer. Figure 6 is also a good illustration of the phenomenon discussed earlier that most of the grid points with negative E_{int} correspond to positive E_{AB}. We have to use a reasonably wide range of monomer deformations to describe at least the ground state vibrations of monomers. For the water molecule, classical turning points for θ in ground-state vibration (zero-point energy of 13.2 kcal/mol) are 76.7° and 136.7°,[45] whereas we sampled the range 82° to 143° (only a part of the sampled range is shown in Figure 6). One can see in Figure 6 that while E_{int} remains relatively flat in the range plotted (varying from −4.97 to −5.16 kcal/mol), E_{AB} quickly decreases in magnitude as θ departs from the equilibrium value (E_{AB} reaches −4.26 and −4.31 kcal/mol in the same range, whereas its value at θ = 104.922° is −5.004 kcal/mol).

Table IV compares the minimum of our water dimer PES to literature data. The vertical interaction energy at the minimum geometry $R_{min} = (R^{rigid}_{min}, R^A_{min}, R^B_{min})$ is defined as $V^{min}_{int} = U_{AB}(R_{min}) - U_A(R^A_{min}) - U_B(R^B_{min}) = V_{int}(R_{min})$. Note that $R^C_{min} \neq R^C_0$ and, in general, in our approach $R^C_0 \neq R^C_{eq}$, where R^C_{eq} is the equilibrium geometry of the isolated monomer C on the surface $U_C(R^C)$. Note also that $V^{int}_{min} \neq U_{AB}(R_{min})$, as the former is independent of the convention of the zero of energy for U_{AB}. With our assumptions, the zero of energy follows from $\lim_{R\to\infty} U_{AB}(R^{rigid}, R^A_0, R^B_0) = 0$, assuming $U_C(R^C_0) = 0$, for any

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

Table V. Energies (in cm^{-1}) of SPs of the $(H_2O)_2$ PES Relative to the Global Minimum Energy, Compared to Literature Results and Ab Initio Computed Values[a]

no.	flex-autoPES	CCSD(T)	CCSD(T)/aug-cc-pCV(TQ)Z	CCpol-8sfIR[45]	MB-pol[51]	HBB2[48]	Tschumper et al.[82]
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	176.4	179.6	179.3	175.6	177.7	185.3	181 ± 5
3	185.6	196.1	195.6	198.6	196.6	210.2	198 ± 6
4	244.2	243.9	242.9	245.4	235.2	260.7	245 ± 15
5	329.0	331.2	329.8	322.0	315.4	354.4	333 ± 15
6	343.0	347.9	346.0	343.4	332.5	375.5	348 ± 18
7	632.5	624.3	623.6	642.8	623.8	629.2	625 ± 18
8	643.0	636.2	635.7	646.2	630.7	640.9	634 ± 16
9	960.0	946.5	946.2	974.0	935.3	946.1	948 ± 23
10	1243.7	1245.6	1245.8	1238.5	1239.2	1230.4	1249 ± 21
11	1935.8	1937.6	1935.9				
12	1943.0	1944.8	1943.5				
RMSE	7.6	1.8	2.3	12.6	10.1	15.1	

[a]SPs 11 and 12 were not reported in the cited works. Energies computed at the geometries of the flex-autoPES SPs at the same level of theory that was used used in flex-autoPES are labeled "CCSD(T)". Energies labeled "CCSD(T)/aug-cc-pCV(TQ)Z" are the same as "CCSD(T)", but computed using core−valence optimized basis sets. RMSEs are relative to results of Tschumper et al.[82]

values of the Euler angles. If, for a given PES, $\boldsymbol{R}_0^C \neq \boldsymbol{R}_{eq}^C$ for C = A or B and an application of this PES requires $\lim_{R\to\infty} U_{AB}(\boldsymbol{R}^{rigid}, \boldsymbol{R}_{eq}^A, \boldsymbol{R}_{eq}^B) = 0$, this can be achieved by an overall shift of the whole PES. Some authors do not list the value of V_{int}^{min}, but instead give U_{AB}^{min} defined as $U_{AB}^{min} = U_{AB}(\boldsymbol{R}_{min}) - U_A(\boldsymbol{R}_{eq}^A) - U_B(\boldsymbol{R}_{eq}^B)$, and therefore, we also compute this quantity for our PES. In the case of the water dimer, $U_A(\boldsymbol{R}_{eq}^A) = -0.0013$ kcal/mol, so that the difference between $\boldsymbol{R}_0^A$ and $\boldsymbol{R}_{eq}^A$ is almost negligible. It is worth noting that $U_{AB}(\boldsymbol{R}^{rigid},\boldsymbol{R}_0^A,\boldsymbol{R}_0^B) = V_{int}(\boldsymbol{R}^{rigid},\boldsymbol{R}_0^A,\boldsymbol{R}_0^B)$ has the minimum with respect to $\boldsymbol{R}^{rigid}$ of quite different value than V_{int}^{min}. For the case of the water dimer, the two values are −4.9712 and −5.0311 kcal/mol. To get closer to V_{int}^{min} in calculations with rigid monomers, one should use the $\langle \boldsymbol{R}^C \rangle_0$ monomer geometry, that is, the geometry averaged in the ground vibration.[91] Compared to the benchmark quantities computed by Lane,[76] our U_{AB}^{min} and V_{int}^{min} differ by 0.02 kcal/mol. This is less than the 0.05 kcal/mol difference resulting from differences in theory level and basis set, cf. Table III. Apparently, the uncertainty of the fit partly cancels the discrepancies at the ab initio level. Even more impressive is the agreement between the geometries: all distances differ by less than 0.0008 Å and angles by 0.1 degrees. The values of U_{AB}^{min} of the best published full-dimensional PESs differ from Lane's value by 0.04, 0.06, and 0.005 kcal/mol for HBB2,[48] MB-pol,[51] and CC-pol-8sfIR[45] PESs, respectively.

Table V lists the stationary points (SPs) of the $(H_2O)_2$ surface. Energies of each SP are compared to values from the CCpol-8sfIR,[45] MB-pol,[51] and HBB2[48] potentials and to those optimized directly in ab initio calculations of ref 82. Overall accuracy of the flex-autoPES surface at these points as compared to the values from ref 82 is somewhat better than that of the CCpol-8sfIR, HBB2, and MB-pol PESs. This is despite the flex-autoPES surface using only about 5000 grid points, as compared to the 240 000 points, 40 000 points, and 30 000 points needed to fit the CCpol-8sfIR, MB-pol, and HBB2 surfaces, respectively. While the flex-autoPES grid generation algorithm places grid points near any minima of the surface, other SPs are not specifically sampled, so the accuracy of the PES at these points is representative of its overall accuracy. Our PES gives SPs barriers within the Tschumper et al. uncertainties except for SP3 where the deviation is two times the uncertainty. However, the uncertainties given by Tschumper et al.[82] are likely overly conservative, which can be seen by comparing the RMSE of these uncertainties amounting to 16 cm^{-1} to the RMSE of the ab initio results in Table V, which amount to only 2 cm^{-1} and to 8 cm^{-1} uncertainty of our fit. In fact, these ab initio results could be more accurate than Tschumper's et al.[82] benchmarks. Recall the RMSE of our fit from Table III (U_{AB} final for $E_{AB} < 0$) is equal to 0.026 kcal/mol = 9.1 cm^{-1}. Thus, the uncertainty due to fitting is about the same as the largest deviations from the results of Tschumper et al.[82]

Also shown in Table V are CCSD(T) barriers, computed at the locations of the SPs on the flex-autoPES surface, at the same level of theory that was used to generate the fit. We compare these to CCSD(T) barriers, which are extrapolated from the aug-cc-pCVTZ and aug-cc-pCVQZ basis sets of ref 78, but are otherwise identical. In all cases, the differences are within 2 cm^{-1}, which is much less than the fitting error of the PES. This justifies the use of the smaller aug-cc-pVTZ and aug-cc-pVQZ basis sets with all-electron calculations in this application. Comparison of either set of the ab initio barriers to fit values and to the benchmarks show that our PES predicts geometries of SPs very accurately even when the fit values are relatively far from the benchmarks.

In Figure 7, the intramolecular fit $U_A(\boldsymbol{R}^A)$ and the PBE0 values $E_A(\boldsymbol{R}^A)$ for the ethylene glycol monomer are shown as functions of the angle of the rotational IDOF through the carbon−carbon axis. The shape of the monomer PES is reproduced almost perfectly by our fit of the form of eqs 11−13 despite not including any terms dependent on the sine or cosine of the rotation angle, as is commonly done in empirical force-fields, such as OPLS-AA, and also in some PESs of small monomers (e.g., in ref 87). It can be seen from this figure that this entire range of deformations gives monomer geometries with energy between −1.7 and 8.4 kcal/mol (the three-dimensional monomer surface varies from −2.3 to 20.8 kcal/mol). This means that, in contrast to H_2O, $HO(CH_2)_2OH$ has a very broad range of geometries that are energetically accessible and need to be reproduced accurately by the fit. As discussed above, this contributes to the lower RMSE of the water dimer PES in comparison to the other two.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

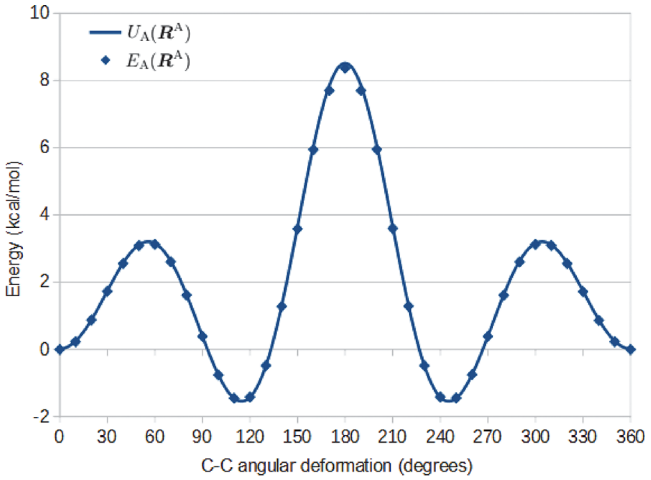


Figure 7. Energy of the $HO(CH_2)_2OH$ monomer computed from the fit $U_A(\boldsymbol{R}^A)$ and the PBE0 values $E_A(\boldsymbol{R}^A)$ as functions of the angular IDOF about the central C−C bond, with the other two IDOFs kept fixed. Angle zero corresponds to the $\boldsymbol{R}_0^A$ geometry.

In Figure 8, we compare the interaction energy component of the $HO(CH_2)_2OH{\cdot}H_2O$ and $(HO(CH_2)_2OH)_2$ PESs to OPLS-AA and to the recently published IPML method.[55] As a generic potential with a simple functional form, OPLS-AA is not expected to be competitive with the system-specific PESs developed here but still provides a useful benchmark. As noted above, water is explicitly parametrized in OPLS-AA, whereas ethylene glycol is included in the training set, so OPLS-AA should have reasonable accuracy for all three dimers. For $HO(CH_2)_2OH{\cdot}H_2O$, the RMSE of OPLS-AA and the flex-autoPES potentials relative to SAPT(DFT) for the five points in the figure with $E_{int} < 0$ are 1.0 and 0.1 kcal/mol, respectively. For $(HO(CH_2)_2OH)_2$, OPLS-AA performs well in the asymptotic and minimum regions, with an RMSE relative to SAPT(DFT) of 0.6 kcal/mol for points with $E_{int} < 0$, compared to 0.16 kcal/mol for flex-autoPES. However, OPLS-AA has an error of 41 kcal/mol for the point at 2.5 Å.

The IPML method, which gives interaction energies between molecules consisting of H, C, N, and O, uses kernel-ridge regression to predict atom-partitioned properties of each molecule by comparing the chemical environment of each atom to those in a training set. These properties include polarizabilities, electrostatic multipole coefficients up to quadrupoles, and parameters of a repulsive model for each atom. The authors report a mean absolute error of 0.7 kcal/mol on the S22 × 5 data set,[92] with somewhat larger error on strongly hydrogen-bonded systems. It can be seen from Figure 8 that IPML works quite well in the asymptotic region but fails badly for these systems near the minima and repulsive walls. This could likely be improved by including ethylene glycol in the training set used by IPML (water is already included). These results demonstrate the importance of system-specific PESs when high accuracy and reliability are required.

IX. CONCLUSIONS

We have demonstrated a method for the generation of accurate nonreactive intermolecular PESs for a wide range of systems, given minimal user input. This work naturally generalizes the earlier autoPES method to enable treatment of systems with an arbitrary number of IDOFs. This allows a more accurate description of systems, such as the water dimer, for which the rigid-monomer approximation already works well, but accuracy standards are so high that monomer-flexibility effects have to be included. It also enables computational studies of effects such as the dependencies of intramonomer rovibrational transition frequencies on interactions with other molecules. Additionally, it extends the class of systems that can be automatically modeled to include those with soft degrees of freedom, such as ethylene glycol, for which rigid-monomer potentials cannot give an accurate description of the entire

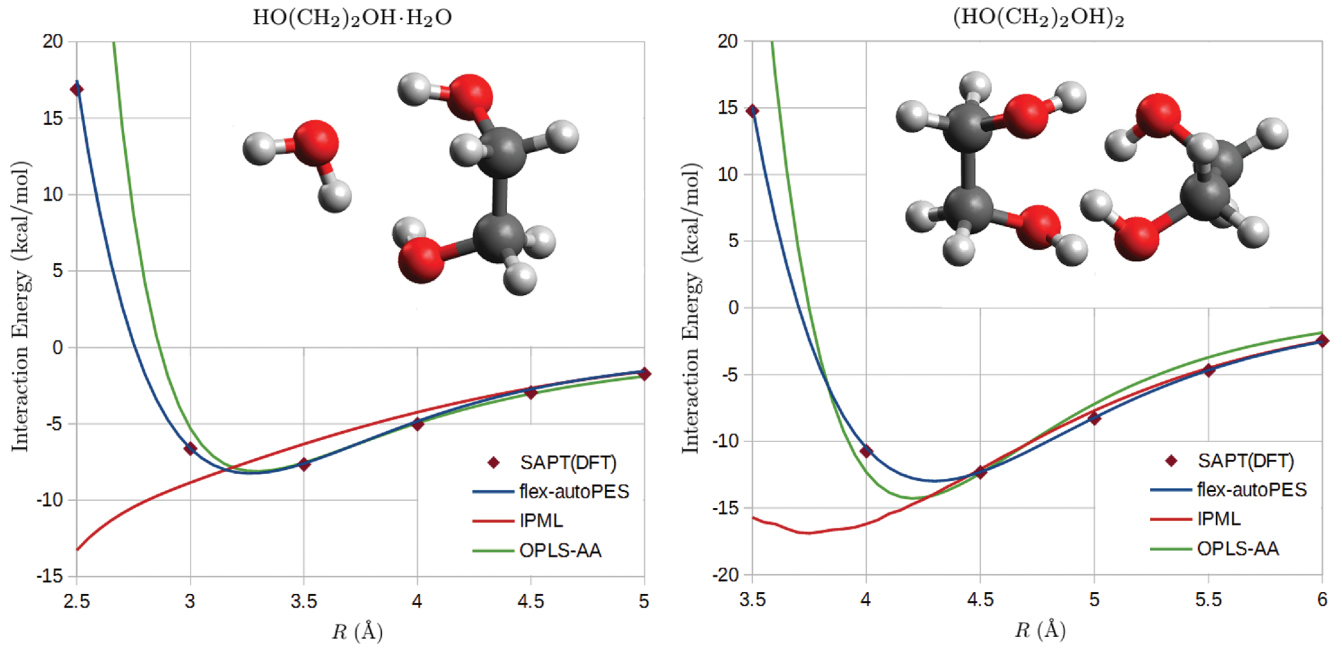


Figure 8. Interaction energies of the $HO(CH_2)_2OH{\cdot}H_2O$ and $(HO(CH_2)_2OH)_2$ PESs compared to SAPT(DFT) reference values, and to the OPLS-AA and IPML methods. In both cases, the radial coordinate R (see section II.B) is varied through the global minimum dimer configuration of the PES (shown in the insets), with the other coordinates kept fixed.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

range of physically relevant configurations. The flex-autoPES method retains the fully automatic character of the original version and requires only the specification of monomer geometries, types of sites, and of IDOFs that are to be modeled. After submitting a flex-autoPES job, no further user action is needed. On output, the user gets a complete PES, which can be used, for example, in MD simulations. Basis sets and electronic structure methods can be selected by the user by simple keyword input, but defaults are used if a keyword is not set. The basis sets are read by the program from a built-in repository. Inputs for the electronic structure programs are generated by autoPES, the jobs performing computations at grid points are submitted by autoPES, and results are extracted in a fully automatic way. The program will continue this way until the convergence criteria described in section VII are satisfied. Use of the program is described by an accompanying software manual which is available online.[13] A copy of the manual is included in the SI.

The account for monomer flexibility was performed differently for each of the three elements affected by it: coupling of the asymptotic interaction model to the IDOFs, coupling of the short-range interaction model to the IDOFs, and the inclusion of an intramonomer model, which gives the energy due to deformations of the isolated monomers. Each of these elements requires its own separate design for the functional form of the model, for generation of the fitting grid, and for parametrization. Unlike in approaches such as PIP as applied in refs 33−35, IMLS as applied in refs 30−32, or many applications of ML methods,[37−39] we treat intermolecular degrees of freedom as distinct from internal ones, which allows greater leeway when choosing physically appropriate functional forms for different components of the system. This, together with other design factors, results in models which are highly accurate, despite using small numbers of fitting data as compared to those needed in existing applications of IMLS, PIP, and ML approaches.

The method was tested by developing a full-dimensional PES for the water dimer and reduced-dimensionality surfaces for ethylene glycol interacting with water and with another ethylene glycol, with 3 out of 24 IDOFs of the ethylene glycol monomer considered. For the water dimer, the quality of the PES is validated by its small RMSE compared to literature fits, by an excellent agreement of the parameters of the vdW minimum with Lane's[76] benchmarks, and by a comparison to accurate ab initio benchmarks for SPs by Tschumper et al.[82] In the latter case, the present PES gives lower RMSE with respect to the benchmarks than any literature PES. In fact, the ab initio computed values of the barriers at the flex-autoPES geometries may be more accurate estimates of the exact barriers than the results of ref 82. For dimers containing ethylene glycol, no literature data are available for comparisons. In fact, there are no literature data for any system of this type as our work is quite pioneering in this respect.

Future work includes extending the reduced-dimensionality PESs to make qualitatively reasonable predictions for configurations whose monomer geometries are significantly outside the range specified by the IDOFs used in the model, that is, when IDOFs not used in the model are significantly deformed. This would permit the use of models which are fitted with reduced dimensionality, such as the ethylene glycol PES presented here, in software which does not support constrained motion (albeit with reduced accuracy if any soft IDOFs are not explicitly parametrized). This would likely be accomplished by combining the current models with generic biomolecular force fields. Alternative functional forms or more refined grid generation procedures may also result in further reduction of the number of grid points required to achieve a given accuracy. The flex-autoPES functionality is available in the recently released version of the autoPES software package.[13] While the developments presented here are restricted to nonreactive surfaces, our approach can be extended to reactive ground-state ones by merging some elements of the present method with the method developed recently by Gyori and Czako[36] for reactive surfaces.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.9b01241.

Programs to evaluate the PESs, plus fitting data, parameter values, and lists of local minima for each of the three surfaces and a copy of the autoPES user manual (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Krzysztof Szalewicz − *Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716, United States*; Email: szalewic@udel.edu

Author

Michael P. Metz − *Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716, United States*; orcid.org/0000-0001-7083-7285

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.9b01241

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the U.S. Army Research Laboratory and Army Research Office under Grants W911NF-13-1-0387 and W911NF-19-1-0117 and by the National Science Foundation under Grants CHE-1566036 and CHE-1900551. Computer resources were provided by the University of Delaware Research Computing Center and by a DoD DURIP grant.

■ REFERENCES

(1) Tadmor, E. B.; Miller, R. E. *Modeling Materials: Continuum, Atomistic and Multiscale Techniques*; Cambridge University Press, 2011.

(2) Taylor, D. E.; Rice, B. M. Quantum-Informed Multiscale M&S for Energetic Materials. *Adv. Quantum Chem.* **2014**, *69*, 171−219.

(3) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. Predictions of the Properties of Water from First Principles. *Science* **2007**, *315*, 1249−1252.

(4) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. Polarizable interaction potential for water from coupled cluster calculations. I. Analysis of dimer potential energy surface. *J. Chem. Phys.* **2008**, *128*, 094313.

(5) Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; van der Avoird, A. Polarizable interaction potential for water from coupled cluster calculations. II. Applications to dimer spectra, virial

coefficients, and simulations of liquid water. *J. Chem. Phys.* **2008**, *128*, 094314.
(6) Szalewicz, K.; Leforestier, C.; van der Avoird, A. Towards complete understanding of water by first-principle computational approach. *Chem. Phys. Lett.* **2009**, *482*, 1−14.
(7) Yang, B.; Zhang, P.; Wang, X.; Stancil, P. C.; Bowman, J. M.; Balakrishnan, N.; Forrey, R. C. Quantum dynamics of CO−H_2 in full dimensionality. *Nat. Commun.* **2015**, *6*, 6629.
(8) Faure, A.; Jankowski, P.; Stoecklin, T.; Szalewicz, K. On the importance of full-dimensionality in low-energy molecular scattering calculations. *Sci. Rep.* **2016**, *6*, 28449.
(9) Góra, U.; Cencek, W.; Podeszwa, R.; van der Avoird, A.; Szalewicz, K. Predictions for water clusters from a first-principles two- and three-body force field. *J. Chem. Phys.* **2014**, *140*, 194101.
(10) Wang, X.-G.; Carrington, T. Using monomer vibrational wavefunctions to compute numerically exact (12D) rovibrational levels of water dimer. *J. Chem. Phys.* **2018**, *148*, 074108.
(11) Metz, M. P.; Szalewicz, K.; Sarka, J.; Tóbiás, R.; Császár, A. G.; Mátyus, E. Molecular dimers of methane clathrates: ab initio potential energy surfaces and variational vibrational states. *Phys. Chem. Chem. Phys.* **2019**, *21*, 13504−13525.
(12) Metz, M. P.; Piszczatowski, K.; Szalewicz, K. Automatic generation of intermolecular potential energy surfaces. *J. Chem. Theory Comput.* **2016**, *12*, 5895−5919.
(13) Metz, M. P.; Piszczatowski, K.; Szalewicz, K. *autoPES Automatic Intermolecular Potential Energy Surface Generation Software*. http://www.physics.udel.edu/~szalewic/SAPT/index.html.
(14) Williams, H. L.; Chabalowski, C. F. Using Kohn-Sham Orbitals in Symmetry-Adapted Perturbation Theory to Investigate Intermolecular Interactions. *J. Phys. Chem. A* **2001**, *105*, 646−659.
(15) Misquitta, A. J.; Szalewicz, K. Intermolecular forces from asymptotically corrected density functional description of monomers. *Chem. Phys. Lett.* **2002**, *357*, 301−306.
(16) Heßelmann, A.; Jansen, G. First-order intermolecular interaction energies from Kohn−Sham orbitals. *Chem. Phys. Lett.* **2002**, *357*, 464−470.
(17) Heßelmann, A.; Jansen, G. Intermolecular induction and exchange-induction energies from coupled-perturbed Kohn-Sham density functional theory. *Chem. Phys. Lett.* **2002**, *362*, 319−325.
(18) Misquitta, A. J.; Jeziorski, B.; Szalewicz, K. Dispersion energy from density-functional theory description of monomers. *Phys. Rev. Lett.* **2003**, *91*, 033201.
(19) Heßelmann, A.; Jansen, G. Intermolecular dispersion energies from time-dependent density functional theory. *Chem. Phys. Lett.* **2003**, *367*, 778−784.
(20) Misquitta, A. J.; Szalewicz, K. Symmetry-Adapted Perturbation Theory Calculations of Intermolecular Forces Employing Density Functional Description of Monomers. *J. Chem. Phys.* **2005**, *122*, 214109.
(21) Misquitta, A. J.; Podeszwa, R.; Jeziorski, B.; Szalewicz, K. Intermolecular potentials based on symmetry-adapted perturbation theory with dispersion energies from time-dependent density-functional calculations. *J. Chem. Phys.* **2005**, *123*, 214103.
(22) Hesselmann, A.; Jansen, G.; Schütz, M. Density-functional theory-symmetry-adapted intermolecular perturbation theory with density fitting: A new efficient method to study intermolecular interaction energies. *J. Chem. Phys.* **2005**, *122*, 014103.
(23) Bukowski, R.; Podeszwa, R.; Szalewicz, K. Efficient calculation of coupled Kohn−Sham dynamic susceptibility functions and dispersion energies with density fitting. *Chem. Phys. Lett.* **2005**, *414*, 111−116.
(24) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Density-fitting method in symmetry-adapted perturbation theory based on Kohn-Sham description of monomers. *J. Chem. Theory Comput.* **2006**, *2*, 400−412.
(25) Podeszwa, R.; Cencek, W.; Szalewicz, K. Efficient Calculations of Dispersion Energies for Nanoscale Systems from Coupled Density Response Functions. *J. Chem. Theory Comput.* **2012**, *8*, 1963−1969.
(26) Taylor, D. C.; et al. Blind test of density-functional-based methods on intermolecular interaction energies. *J. Chem. Phys.* **2016**, *145*, 124105.
(27) Jankowski, P.; McKellar, A. R. W.; Szalewicz, K. Theory untangles high-resolution infrared spectrum of the *ortho*H_2-CO van der Waals complex. *Science* **2012**, *336*, 1147−1150.
(28) Jankowski, P.; Surin, L. A.; Potapov, A.; Schlemmer, S.; McKellar, A. R. W.; Szalewicz, K. A comprehensive experimental and theoretical study of H_2-CO spectra. *J. Chem. Phys.* **2013**, *138*, 084307.
(29) Akin-Ojo, O.; Szalewicz, K. How well can polarization models of pairwise nonadditive forces describe liquid water? *J. Chem. Phys.* **2013**, *138*, 024316.
(30) Dawes, R.; Thompson, D. L.; Guo, Y.; Wagner, A. F.; Minkoff, M. Interpolating moving least-squares methods for fitting potential energy surfaces: Computing high-density potential energy surface data from low-density ab initio data points. *J. Chem. Phys.* **2007**, *126*, 184108.
(31) Majumder, M.; Ndengue, S. A.; Dawes, R. Automated construction of potential energy surfaces. *Mol. Phys.* **2016**, *114*, 1−18.
(32) Quintas-Sanchez, E.; Dawes, R. AUTOSURF: a freely available program to construct potential energy surfaces. *J. Chem. Inf. Model.* **2019**, *59*, 262−271.
(33) Huang, X.; Braams, B. J.; Bowman, J. M. *Ab initio* potential energy and dipole moment surfaces of $(H_2O)_2$. *J. Phys. Chem. A* **2006**, *110*, 445−451.
(34) Braams, B. J.; Bowman, J. M. Permutationally invariant potential energy surfaces in high dimensionality. *Int. Rev. Phys. Chem.* **2009**, *28*, 577−606.
(35) Qu, C.; Yu, Q.; Bowman, J. M. Permutationally Invariant Potential Energy Surfaces. *Annu. Rev. Phys. Chem.* **2018**, *69*, 151−175.
(36) Gyori, T.; Czako, G. Automating the Development of High-Dimensional Reactive Potential Energy Surfaces with the ROBOSURFER Program System. *J. Chem. Theory Comput.* **2020**, *16*, 51−66.
(37) Manzhos, S.; Dawes, R.; Carrington, T. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem.* **2015**, *115*, 1012−1020.
(38) Qu, C.; Yu, Q.; Van Hoozen, B. L.; Bowman, J. M.; Vargas-Hernández, R. A. Assessing Gaussian Process Regression and Permutationally Invariant Polynomial Approaches To Represent High-Dimensional Potential Energy Surfaces. *J. Chem. Theory Comput.* **2018**, *14*, 3381−3396.
(39) Nguyen, T. T.; Szekely, E.; Imbalzano, G.; Behler, J.; Csanyi, G.; Ceriotti, M.; Goetz, A. W.; Paesani, F. Comparison of permutationally invariant polynomials, neural networks, and Gaussian approximation potentials in representing water interactions through many-body expansions. *J. Chem. Phys.* **2018**, *148*, 241725.
(40) Abbott, A. S.; Turney, J. M.; Zhang, B.; Smith, D. G. A.; Altarawy, D.; Schaefer, H. F. PES-Learn: An Open-Source Software Package for the Automated Generation of Machine Learning Models of Molecular Potential Energy Surfaces. *J. Chem. Theory Comput.* **2019**, *15*, 4386−4398.
(41) Williams, H. L.; Szalewicz, K.; Jeziorski, B.; Moszynski, R.; Rybak, S. Symmetry-Adapted Perturbation Theory Calculation of the Ar-H_2 Intermolecular Potential Energy Surface. *J. Chem. Phys.* **1993**, *98*, 1279−1292.
(42) Jeziorska, M.; Jankowski, P.; Szalewicz, K.; Jeziorski, B. On the Optimal Choice of Monomer Geometry in Calculations of Intermolecular Potentials. Rovibrational Spectrum of Ar−HF Generated from Two- and Three-Dimensional SAPT Potentials. *J. Chem. Phys.* **2000**, *113*, 2957−2968.
(43) Szalewicz, K.; Murdachaew, G.; Bukowski, R.; Akin-Ojo, O.; Leforestier, C. In *Lecture Series on Computer and Computational Science: International Conference of Computational Methods in Science and Engineering (ICCMSE 2006)*; Maroulis, G., Simos, T., Eds.; Brill Academic Publishers, Leiden, 2006; Vol. *6*; pp 482−491.
(44) Leforestier, C.; Szalewicz, K.; van der Avoird, A. Spectra of water dimer from a new *ab initio* potential with flexible monomers. *J. Chem. Phys.* **2012**, *137*, 014305.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

(45) Jankowski, P.; Murdachaew, G.; Bukowski, R.; Akin-Ojo, O.; Leforestier, C.; Szalewicz, K. Ab initio water pair potential with flexible monomersz. *J. Phys. Chem. A* **2015**, *119*, 2940−2964.
(46) Murdachaew, G.; Szalewicz, K.; Bukowski, R. Efficient generation of flexible-monomer intermolecular potential energy surfaces. *Phys. Rev. Lett.* **2002**, *88*, 123202.
(47) Murdachaew, G.; Szalewicz, K. Intermolecular potentials with flexible monomers. *Faraday Discuss.* **2001**, *118*, 121−142.
(48) Shank, A.; Wang, Y.; Kaledin, A.; Braams, B. J.; Bowman, J. M. Accurate *ab initio* and "hybrid" potential energy surfaces, intramolecular vibrational energies, and classical IR spectrum of the water dimer. *J. Chem. Phys.* **2009**, *130*, 144314.
(49) Babin, V.; Medders, G. R.; Paesani, F. Toward a universal water model: first principles simulations from the dimer to the liquid phase. *J. Phys. Chem. Lett.* **2012**, *3*, 3765−3769.
(50) Medders, G. R.; Babin, V.; Paesani, F. A critical assessment of two-body and three-body interactions in water. *J. Chem. Theory Comput.* **2013**, *9*, 1103−1114.
(51) Babin, V.; Leforestier, C.; Paesani, F. Development of a 'first principles' water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient. *J. Chem. Theory Comput.* **2013**, *9*, 5395−5403.
(52) Garberoglio, G.; Jankowski, P.; Szalewicz, K.; Harvey, A. H. Fully quantum calculation of the second and third virial coefficients of water and its isotopologues from *ab initio* potentials. *Faraday Discuss.* **2018**, *212*, 467−497.
(53) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **1996**, *118*, 11225−11236.
(54) Damm, W.; Frontera, A.; Tirado-Rives, J.; Jorgensen, W. L. OPLS all-atom force field for carbohydrates. *J. Comput. Chem.* **1997**, *18*, 1955−1970.
(55) Bereau, T.; DiStasio, R. A., Jr; Tkatchenko, A.; Von Lilienfeld, O. A. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *J. Chem. Phys.* **2018**, *148*, 241706.
(56) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent radii revisited. *Dalt. Trans.* **2008**, 2832−2838.
(57) Sobol', I. M. On the distribution of points in a cube and the approximate evaluation of integrals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* **1967**, *7*, 86.
(58) Geman, S.; Bienenstock, E.; Doursat, R. Neural networks and the bias/variance dilemma. *Neural computation* **1992**, *4*, 1−58.
(59) Sacks, J.; Welch, W. J.; Mitchell, T. J.; Wynn, H. P. Design and analysis of computer experiments. *Stat. Sci.* **1989**, *4*, 409−423.
(60) Metz, M. P.; Szalewicz, K. A statistically guided grid generation method and its application to intermolecular potential energy surfaces. *J. Chem. Phys.*, **2020**, DOI: 10.1063/1.5141777, in press.
(61) Rowland, R. S.; Taylor, R. Intermolecular nonbonded contact distances in organic crystal structures: Comparison with distances expected from van der Waals radii. *J. Phys. Chem.* **1996**, *100*, 7384−7391.
(62) Szalewicz, K.; Bukowski, R.; Jeziorski, B. In *Theory and Applications of Computational Chemistry: The First Fourty Years*; Dykstra, C. E., Frenking, G., Kim, K. S., Scuseria, G. E., Eds.; Elsevier, Amsterdam, 2005; Chapter 33, pp 919−962.
(63) Gora, U.; Podeszwa, R.; Cencek, W.; Szalewicz, K. Interaction energies of large clusters from many-body expansion: water clusters. *J. Chem. Phys.* **2011**, *135*, 224102.
(64) Akin-Ojo, O.; Song, Y.; Wang, F. Developing ab initio quality force fields from condensed phase quantum-mechanics/molecular-mechanics calculations through the adaptive force matching method. *J. Chem. Phys.* **2008**, *129*, 064108.
(65) Nandi, A.; Qu, C.; Bowman, J. M. Using Gradients in Permutationally Invariant Polynomial Potential Fitting: A Demonstration for CH_4 Using as Few as 100 Configurations. *J. Chem. Theory Comput.* **2019**, *15*, 2826−2835.
(66) Tang, K. T.; Toennies, J. P. An improved simple model for the van der Waals potential based on universal damping functions for the dispersion coefficients. *J. Chem. Phys.* **1984**, *80*, 3726−3741.
(67) Powell, M. J. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comp. J.* **1964**, *7*, 155−162.
(68) Brent, R.; Burkardt, J. *PRAXIS*, 2006. http://people.sc.fsu.edu/~jburkardt/f_src/praxis/praxis.html.
(69) Rob, F.; Szalewicz, K. Distributed molecular polarisabilities and asymptotic intermolecular interaction energies. *Mol. Phys.* **2013**, *111*, 1430−1455.
(70) Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **1990**, *11*, 361−373.
(71) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.
(72) Hartke, B. Global optimization. *Wiley Interdis. Rev.: Comp. Mol. Sci.* **2011**, *1*, 879−887.
(73) Pereira, F. B.; Marques, J. M. A study on diversity for cluster geometry optimization. *Evolutionary Intelligence* **2009**, *2*, 121.
(74) Charbonneau, P.; Knapp, B. *A User's Guide to PIKAIA 1.0*; NCAR Tech, 1995.
(75) Charbonneau, P. Genetic algorithms in astronomy and astrophysics. *Astrophys. J., Suppl. Ser.* **1995**, *101*, 309.
(76) Lane, J. R. CCSDTQ optimized geometry of water dimer. *J. Chem. Theory Comput.* **2013**, *9*, 316−323.
(77) Werner, H.-J. et al. *MOLPRO, version 2010.1, A Package of Ab Initio Programs*, 2010. http://www.molpro.net.
(78) Kendall, R. A.; Dunning, T. H., Jr.; Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **1992**, *96*, 6796−6806.
(79) Podeszwa, R.; Bukowski, R.; Szalewicz, K. Potential energy surface for the benzene dimer and perturbational analysis of π-π interactions. *J. Phys. Chem. A* **2006**, *110*, 10345−10354.
(80) Boys, S. F.; Bernardi, F. The calculation of small molecular interactions by the differences of separate total energies. Some procedures with reduced errors. *Mol. Phys.* **1970**, *19*, 553−566.
(81) Mas, E. M.; Bukowski, R.; Szalewicz, K.; Groenenboom, G. C.; Wormer, P. E. S.; van der Avoird, A. Water pair potential of near spectroscopic accuracy. I. Analysis of potential surface and virial coefficients. *J. Chem. Phys.* **2000**, *113*, 6687−6701.
(82) Tschumper, G. S.; Leininger, M. L.; Hoffman, B. C.; Valeev, E. F.; Schaefer, H. F., III; Quack, M. Anchoring the water dimer potential energy surface with explicitly correlated computations and focal point analyses. *J. Chem. Phys.* **2002**, *116*, 690−701.
(83) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.
(84) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158−6170.
(85) Jeziorska, M.; Jeziorski, B.; Cizek, J. Direct calculation of the Hartree-Fock interaction energy via exchange perturbation expansion - the He−He interaction. *Int. J. Quantum Chem.* **1987**, *32*, 149−164.
(86) Patkowski, K.; Szalewicz, K.; Jeziorski, B. Third-order interactions in symmetry-adapted perturbation theory. *J. Chem. Phys.* **2006**, *125*, 154107.
(87) Polyansky, O. L.; Jensen, P.; Tennyson, J. The potential energy surface of $H_2{}^{16}O$. *J. Chem. Phys.* **1996**, *105*, 6490−6497.
(88) Mizus, I. I.; Kyuberis, A. A.; Zobov, N. F.; Makhnev, V. Y.; Polyansky, O. L.; Tennyson, J. High-accuracy water potential energy surface for the calculation of infrared spectra. *Philos. Trans. R. Soc., A* **2018**, *376*, 20170149.
(89) Cencek, W.; Szalewicz, K.; Leforestier, C.; van Harrevelt, R.; van der Avoird, A. An accurate analytic representation of the water pair potential. *Phys. Chem. Chem. Phys.* **2008**, *10*, 4716−4731.
(90) Odutola, J. A.; Dyke, T. R. *J. Chem. Phys.* **1980**, *72*, 5062.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339

(91) Mas, E. M.; Szalewicz, K. Effects of monomer geometry and basis set saturation on Depth of Water Dimer Potential. *J. Chem. Phys.* **1996**, *104*, 7606−7614.
(92) Gráfová, L.; Pitonak, M.; Rezac, J.; Hobza, P. Comparative study of selected wave function and density functional methods for noncovalent interaction energy calculations using the extended S22 data set. *J. Chem. Theory Comput.* **2010**, *6*, 2365−2376.

https://dx.doi.org/10.1021/acs.jctc.9b01241
J. Chem. Theory Comput. 2020, 16, 2317−2339