JCTC Journal of Chemical Theory and Computation

ARTICLE

pubs.acs.org/JCTC

Distributed Multipoles and Energies of Flexible Molecules

Hai-Anh Le and Ryan P. A. Bettens*

Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543

S Supporting Information

ABSTRACT: In this work we show that energies and distributed multipoles, up to and including rank two, can be accurately determined via a modified Shepard interpolation of ab initio data for small molecules. The molecules considered here are the amino aldehydes, Gly and Ala, which may be typical smaller fragment molecules in certain molecular energy-based fragmentation schemes. The method is general and should be suitable for applications also involving crystal structure prediction, modeling molecular clusters, and Monte Carlo or molecular/reaction dynamics simulations. The configuration space covered by the interpolation includes that sampled by the Gly and Ala peptides in protein crystal structures, i.e., 12 dimensions for Gly: 3 torsion angles (φ, ψ, ω), 5 bond lengths, and 4 bond angles and 15 dimensions for Ala: 4 torsion angles, 6 bond lengths, and 5 bond angles. In this work we also describe a new method of importance, sampling the relevant configuration spaces, and show that it is possible to interpolate "axis free" multipoles.

1. INTRODUCTION

It is now well established that an inexpensive but accurate approach to determine the electrostatic interaction energy between molecules, or the electrostatic potential about them, is by distributing multipoles at various sites within molecules. Different methods to determine the multipoles are available and include "distributed multipole analysis" (DMA),[1,2] "atoms in molecules" (AIM)[3] and "transferable atom equivalent" (TAE),[4] and "cumulative atomic multipole moments" (CAMM).[5,6]

All of these approaches utilize electronic structure calculations to obtain the distributed multipoles. Necessarily a specific molecular structure must be adopted. The molecular structure is precisely defined in terms of the atomic configuration that can be represented by an appropriate set of internal coordinates. Thus for a specific electronic structure calculation the distributed multipoles are only exact for the specific point in molecular configuration space where the calculation was performed. This is all that is required for applications involving rigid molecules. However, for systems involving flexible molecules different parts of configuration space are visited, so the distributed multipoles that were determined at a single point in this space are no longer exact. One possible option is to assume that the electron density remains unchanged in different molecular conformations. Unfortunately such an assumption is not sufficient for applications involving the modeling of molecular clusters[7] or crystal structure prediction.[8] Aside from performing a new electronic structure calculation at the new molecular configuration, which at the very least is incredibly expensive or at worst impossible depending upon the application, various attempts have been made at predicting the distributed multipoles.

One such attempt includes substantially reducing the dimensionality of configuration space to focus only upon a few degrees of freedom, namely selected torsions, then to either parametrically fit the multipoles to ad hoc functions in the reduced space[7,9] or by interpolation using a grid of precomputed configurations in this reduced space.[10,11] Both of these types of approaches, while very satisfactory for their desired application, are only practical provided the dimensionality of the space remains small. For substantially higher dimensionality neither of these approaches can be adopted. Alternatively, the change in the distributed multipoles with conformation has been dealt with using "intramolecular polarization"[12,13] and then utilized in molecular dynamics.[14]

Here we focus our attention on DMA and present a method that enables accurate predictions of both distributed multipoles and intramolecular electronic energies for generally flexible molecules, i.e., molecules in any arbitrary configuration. Our method utilizes the modified Shepard interpolation.[15−18] It should be noted that integral to this method is the importance sampling of the relevant configuration space to the application of interest, which results in a significant improvement in computational efficiency. Such a method can readily find application in crystal structure prediction, modeling molecular clusters, and Monte Carlo or molecular/reaction dynamics simulations or in an ab initio molecular database.[19] However, depending on the specific application, the resulting sampled surfaces may not be sufficiently accurate if they are directly transferred to a different application. In the latter case the method can still be applied, but the surface should be importance sampled again on the relevant regions of configuration space. The application of most interest to us here is molecular energy-based fragmentation. This particular application is an approximately linear scaling technique that requires the energies, and other properties like the electrostatic potential, of relatively small molecules in a range of configurations. A number of groups have developed various molecular energy-based fragmentation methods by fragmenting a larger molecule, e.g., a protein, into many smaller complete molecules then linearly combining their electronic energies to approximate the total energy, or some other property, of the target molecule.

Received: November 25, 2010
Published: March 18, 2011

dx.doi.org/10.1021/ct100683u | *J. Chem. Theory Comput.* 2011, 7, 921–930

The first type of molecular energy-based fragmentation was attempted by Gadre's group using their molecular tailoring method.[20−24] This method was originally suggested to calculate one-electron properties, such as the electrostatic potential. More recently Zhang and Zhang developed a quite different fragmentation algorithm[25] (molecular fractionation with conjugated caps) and have applied (and extended) their approach successfully to several systems.[26−39] This method was originally designed to accurately compute interaction energies between two molecular systems. Molecular total energies were first attempted using molecular energy-based fragmentation by Li's group[40−46] (from which we have borrowed the term molecular energy-based fragmentation) and the Collins group[47−50] and later by ourselves.[51−54] It should be noted here that other types of linear scaling methods exist, e.g., the density matrix divide and conquer approach of Yang and Lee[55] or Kitaura's fragment molecular orbital approach (ref 56 and recently ref 57 and references therein), which has been implemented in the GAMESS package[58] and Exner and Mezey's field-adapted adjustable density matrix approach.[59−63]

Molecular energy-based fragmentation utilizing the methods of Collins or of our group offers an intriguing possibility when applied to molecules like proteins. Proteins, being composed of around 20 amino acids, when fragmented into their constituent molecules produces a finite number of possible fragment molecules. This particular flavor of fragmentation breaks large molecules up based solely on the primary sequence. Thus if a database of precomputed highly accurate ab initio potential energy surfaces were available for these fragment molecules, then the "bonded" energy of proteins could be determined at comparatively negligible computational expense. The generation of the required highly dimensional potential energy surfaces is feasible because the fragment molecule sizes involved are relatively small. Of course, nonbonded interactions play a crucial role in such systems, but such interactions may be taken into account via electrostatics (through the utilization of distributed multipoles), induction (through distributed or central polarizabilities), and dispersion (through the real dynamic polarizabilities at imaginary frequencies), if fragment molecules are separated far enough from each other. All of these quantities can be determined from first principle calculations without the need of any parametrization and are solely a property of the monomer fragments.

Therefore apart from the electronic energy, distributed electrostatic multipoles, polarizabilities, and the dynamic polarizabilities at imaginary frequencies can also form part of the precomputed database of fragment molecules. Close-contact nonbonded interactions, most importantly H-bonds, could not be dealt with in such a manner. However, for these types of interactions, precomputed potential energy surfaces can also be generated because there is a finite number of possible close-contact interactions. Similarly, the solvent must be taken into account in such systems, and this may be done either implicitly of explicitly. For the latter, precomputed potential energy surfaces for water−water and small water clusters as well as water−protein fragment molecules can be determined for the close-contact interactions. Therefore, in principle at least, it may be possible to produce highly accurate second generation modeling software capable of dealing with protein−substrate interactions in the presence of solvent from first principles, but utilizing the fragmentation approximation, to replace the current empirical force field or QM/MM methods. Such a goal, if even possible, lies far from the present, but a necessary step along the way is to show that highly dimensional potential energy surfaces can be constructed for possible fragment molecules. Furthermore, it is also necessary to show that distributed multipoles can be interpolated well enough to predict accurate electrostatics within the relevant configuration spaces of the fragments.

2. METHODOLOGY

2.1. Theory. *2.1.1. Distributed Multipoles.* In this work we utilized Stone's distributed multipole analysis,[1,2] which has been well described elsewhere,[2,64] so only a brief account will be provided here. The electrostatic potential at some location **r** is given by the well-known expression:

$$V(\mathbf{r}) = \sum_{k=1}^{N_{\text{nuclei}}} \frac{Z_k}{|\mathbf{r} - \mathbf{r}_k|} - \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}' \tag{1}$$

where $\rho(\mathbf{r}')$ is the charge density which may incorporate the effects of electron correlation. The charge density takes the form

$$\rho(\mathbf{r}') = \sum_{st} P_{st}\chi_s(\mathbf{r}' - \mathbf{p}_s)\chi_t(\mathbf{r}' - \mathbf{p}_t) \tag{2}$$

where P_{st} is an element of the density matrix, and $\chi_s(\mathbf{r}' - \mathbf{p}_s)$ is a basis function centered at $\mathbf{p}_s$. The contribution made to the charge density from a product of Gaussian basis functions can be represented as a linear combination of multipoles, centered at **p**, from rank zero to rank $l + l'$, where l and l' are the ranks of the original basis functions (l = 0, *s*-function; l = 1, *p*-function, etc.). The location of **p** lies on a line in between the centers *s* and *t* and depends on the exponents of the basis functions. Thus the charge density resulting from the overlap of, say, two *s*-functions can be represented as a single monopole or the overlap of a *d*- and *p*-function as a monopole, dipole, quadrupole, and octapole only. However, the locations of each of these sets of multipoles will be different for each pair of basis functions. Each set of multipoles can have their origin shifted to one of the centers *s* or *t* or some other convenient center. The price paid for shifting the origin is the generation of an infinite number of multipoles at the new origin. The convergence of this infinite series of multipoles at the new center depends strongly on the distance shifted. How this shift is affected is ad hoc, but the smaller the shift, the more rapid the convergence of the multipole series. Thus there is a strong argument to make this shift to the nearest possible center, which may not be either site *s* or *t*.

In this way a set of multipoles can be generated at various locations within the molecule. The total molecular dipole of the molecule is exactly reproduced by the collection of distributed monopoles and dipoles − the molecular quadrupole by the collection of distributed monopoles, dipoles, and quadrupoles, etc. Furthermore the potential, electric field, electric field gradient, etc. of the molecule is very accurately reproduced at locations further away from the molecule than the van der Waals surface using a sufficient number of centers and multipoles.

The deficiencies in the above treatment lay in determining the multipoles piecemeal from the individual products of basis functions rather than from the molecular charge density in physical space and was most pronounced when the basis included diffuse functions. The deficiency was not one of accuracy, but rather the distributed multipoles could vary widely and unpredictably by improving the basis set used to describe the molecule even though the electrostatic potential may have changed little.[2] This deficiency was alleviated in 2005 by modifying the distributed multipole analysis to allow for a numerical integration of the charge density due to diffuse, or extended, basis functions around atomic sites[2] to determine their contribution to the multipole moments at those sites.

2.1.2. Modified Shepard Interpolation. The modified Shepard interpolation[15−18] has been used to accurately represent ab initio potential energy surfaces by the Collins group and others for both classical and quantum reaction dynamics (e.g., refs 16 and 65−67). It has also been utilized in stationary state problems (e.g., refs 68 and 69). In this work we employed the technique in order to interpolate both the energy and the distributed multipoles. Note that the interpolated surface always passes exactly through all data used in the interpolation.

The interpolation of some quantity, $X(\mathbf{Z})$, which is a function of the internal coordinates, $\mathbf{Z}$, proceeds by first expanding $X(\mathbf{Z})$ as a Taylor series about some specific location $\mathbf{Z}_0$:

$$X(\mathbf{Z}) = X_{\mathbf{Z}_0} + \left.\frac{\partial X}{\partial \mathbf{Z}}\right|_{\mathbf{Z}_0} \cdot (\mathbf{Z} - \mathbf{Z}_0) + \frac{1}{2}(\mathbf{Z} - \mathbf{Z}_0)^{\mathrm{T}} \left.\frac{\partial^2 X}{\partial \mathbf{Z}^2}\right|_{\mathbf{Z}_0} (\mathbf{Z} - \mathbf{Z}_0) + ... \quad (3)$$

The Taylor series is then truncated at some order, and in this work it was truncated after the second order. Thus the estimate of X is only expected to be accurate in the vicinity around $\mathbf{Z}_0$. Indeed, for some applications, like the crystal structure prediction of small molecules,[70] a single Taylor series expansion has been utilized, where X represents the electronic energy. However, the estimate of the quantity X can be improved through the addition of further Taylor series expanded about different locations. If we include all of the Taylor series estimates of X at $\mathbf{Z}$, then we may write

$$\chi(\mathbf{Z}) = \sum_{i=1}^{N} w_i(\mathbf{Z}) T_i(\mathbf{Z}) \quad (4)$$

where $T_i(\mathbf{Z})$ is a truncated Taylor series expanded about location $\mathbf{Z}_i$, and $w_i(\mathbf{Z})$ is the normalized weight given to Taylor series i, which depends upon the location $\mathbf{Z}$. N is the total number of Taylor series which constitute the interpolation data set. In our work, the simple "one-part" weight function[15] (see eq 6) was used to add the first 40 data points for both Gly and Ala. After that, the more flexible "two-part" weight function (see eq 7) together with the confidence radius[17] (see eq 8) were employed to improve the accuracy of the data sets:

$$w_i(\mathbf{Z}) = \frac{v_i(\mathbf{Z})}{\sum_{j=1}^{N} v_j(\mathbf{Z})} \quad (5)$$

where

$$v_i(\mathbf{Z}) = |\mathbf{Z} - \mathbf{Z}_i|^{-2p} \quad (6)$$

for the "one-part" weight function and

$$v_i(\mathbf{Z}) = \left\{ \left[\frac{|\mathbf{Z} - \mathbf{Z}_i|}{\mathrm{crad}_i}\right]^{2q} + \left[\frac{|\mathbf{Z} - \mathbf{Z}_i|}{\mathrm{crad}_i}\right]^{2p} \right\}^{-1} \quad (7)$$

for the "two-part" weight function. In this expression crad_i is given by

$$\mathrm{crad}_i^{-6} = \frac{1}{N_{\mathrm{neigh}}} \sum_{k=1}^{N_{\mathrm{neigh}}} \frac{[X(\mathbf{Z}_k) - T(\mathbf{Z}_k)]^2}{E_{\mathrm{tol}}^2 |\mathbf{Z}_k - \mathbf{Z}|^6} \quad (8)$$

N_{neigh} is the number of nearby configurations, $2p = 16$ and $q = 2$ for Gly and Ala, and E_{tol} was set to 0.2 m-Eh. Note that $2p$ must be greater than the number of degrees of freedom in order for the potential to converge using both one- and two-part weight functions. The number of degrees of freedom of Gly and Ala were 12 and 15, respectively (see the Fragment Structures section for an explanation of these numbers). For convenience and accuracy, $q = 2$ was taken.[18] The confidence radius crad_i represents the distance away from Taylor series i in which the average error increases to the value of the error tolerance; crad_i is discussed in ref 17.

2.1.3. Axis Systems and "Axis-Free" Multipoles. The Shepard interpolation, as described above, is utilized to interpolate scalar quantities. In this work we have directly applied it to the interpolation of the total electronic energy and the distributed monopoles. The individual components of the distributed dipoles and quadrupoles might also be Shepard interpolated. That is, treated as though they were scalar quantities, but as we shall see later, imposing an axis system on the molecule and then interpolating the individual components leads to difficulties that can be avoided. It should be noted that multipoles are tensors, and thus they transform according to the tensor transformation law. This transformation affects the values of the components and therefore the interpolation.

The orientation of a molecule with respect to some lab-based Cartesian frame is defined in terms of the three Euler angles that rotate the lab-based Cartesian frame into the molecule-based Cartesian frame. For rigid molecules, i.e., those with no change in internal coordinates, when the molecule rotates, i.e., changes its orientation, the molecule-based Cartesian framework rotates with it. However, if the molecule can distort via some change in the internal coordinates, then the molecule-based Cartesian framework will generally move both in origin and orientation with respect to the lab-based framework. Thus, the three Euler angles depend upon the internal coordinates. This phenomenon is well-known in spectroscopy and is the origin of rovibrational coupling. It is therefore best to refer all components of the multipole moments to a molecule-based frame. The advantage of referring multipole components to this frame is that these components do not depend upon the Euler angles.

However, the disadvantage of referring the multipole components to a molecule-based frame is that if a poor choice is made for the frame, the components of the multipoles may vary significantly at atomic sites located far from the place in the molecule where a distortion has occurred. This variation in the components would be solely due to the changing of the molecule-based frame rather than any change in electron density at the remote atomic sites. One way to significantly alleviate this effect would be to define a local atomic Cartesian frame at each atomic site, say by using the atoms adjacent to the site in question. Such axis systems have been suggested and utilized before.[12]

Alternatively the molecule-based Cartesian frame can be disposed of entirely, and the distributed dipoles and quadrupoles expressed in terms of the internal coordinates of the molecule. It is this approach that we have adopted here. In order to achieve

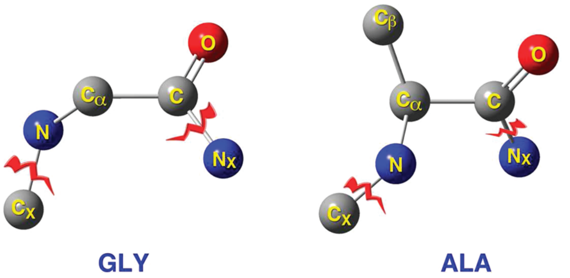

Figure 1. Gly and Ala residues extracted from a pdb file. Also illustrated are the adjacent carbon and nitrogen atoms (labeled C_X and N_X) and the bonds to be cut and capped with hydrogen atoms.

this transformation we envisaged an atomic dipole as a vector which locates the position of a fictitious atom attached to the atomic site where the dipole originated. As such, the magnitude of the dipole represented a "bond length" to the fictitious atom. The direction of the dipole was expressed as a bond angle and a dihedral angle, which was made between the fictitious atom, the atomic site, and one (for angle) or two (for dihedral angle) nearby atoms. The Shepard interpolation then involved interpolating these three strictly scalar quantities in terms of the actual internal coordinates of the real atoms in the molecule.

The distributed quadrupoles were treated similarly but in a slightly more complicated manner. The quadrupole tensor and atomic coordinates in the standard orientation defined in the Gaussian software[71] (utilized in this work) were first transformed into another molecule fixed axis system defined using the three atoms, N, C_α, and C (see Figure 1). This was necessary due to axis switching that arbitrarily occurred during numerical displacements in internal coordinates while calculating the derivatives and between different conformations.

The Cartesian quadrupole tensor in the new axis system was then diagonalized to obtain the principle quadrupole axes. We then considered each of the axes as a vector, and as such, the treatment was similar to that of the dipoles. The first two eigenvalues, with the last eigenvalue being equal to minus their sum, were readily treated as scalar quantities and interpolated as usual. The corresponding first two eigenvectors were described by placing two fictitious atoms at the end points of these vectors from the atomic site where the quadrupole originated. After placing these two fictitious atoms, three independent internal coordinates were defined to locate them: a single "bond angle" and two "dihedral angles". The bond angle was defined to be between the first fictitious atoms, the atom to which it was "bonded", and some adjacent atom in the molecule. Since the two eigenvectors are orthogonal it was unnecessary to define a second "bond angle". The two "dihedral angles" were made to adjacent atoms within the molecule. These three coordinates completely specified the directions of the principle quadrupole axes. The Shepard interpolation then involved interpolating the five quantities: two eigenvalues, a "bond angle", and two "dihedral angles", in terms of the actual internal coordinates of the real atoms in the molecule.

Nevertheless, it should be noted that the eigenvectors are fundamentally bidirectional. As a consequence, there are two possible positions for each of the fictitious atoms. Hence the axes may arbitrarily switch between different conformations. For similar geometries this is readily detected and rectified. Unfortunately, for geometries located far from others in configuration space this becomes problematic. A solution was found by erecting a network of nearest neighbors between all conformations. If a complete weighted graph is associated with the interpolation data set, a vertex being a conformation and the weight of an edge being the Euclidean distance in internal coordinates between them, then the desired network is a minimum spanning tree. In our work, Prim's algorithm[72,73] was used to determine one of the minimum spanning trees. The first vertex was defined as an end point of the first edge added during the construction of this tree. The orientation of the eigenvectors chosen for this starting vertex was then used as a point of reference to assign the orientation of the eigenvectors for subsequent geometries. This pretreatment assures consistency among the eigenvectors prior to interpolation, provided that the data points are not sparsely distributed. Therefore, a sufficient data density was necessary.

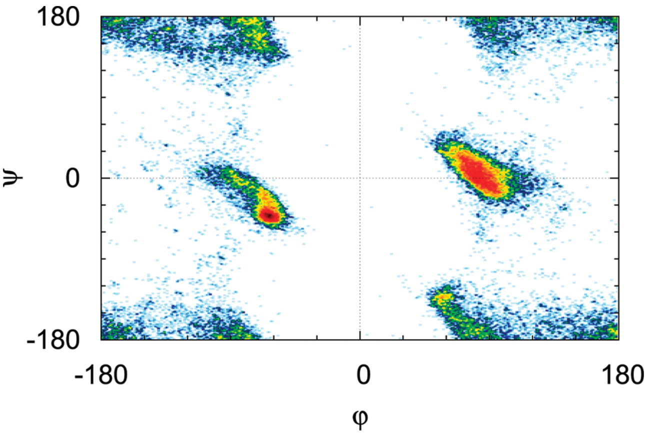

Figure 2. A 2-dimensional projection (φ and ψ) of the 12-dimensional data set extracted from the .pdb files for Gly. Red indicates high data density, while blue is low data density.

For both the dipole and quadrupole, interpolation of the dihedral angles to the fictitious atoms requires an additional comment. Because of the periodic nature of an angle, care must be taken when two or more Taylor series estimated a dihedral to be approximately $\pm\pi$. While the estimated angles may be nearly the same, the numerical values from individual Taylor series may have differed in sign so that application of eq 4 resulted in substantially different interpolated values for the dihedral.

2.2. Approach.

2.2.1. Fragment Structures. The Research Collaboratory for Structural Bioinformatics protein data bank[74] was searched on October 12, 2009 for X-ray crystallographic structures of proteins with a resolution in the range of 0−1.3 Å. This resulted in a total of 1745 PDB files (listed in the Supporting Information). All structures were then searched within each file to obtain every occurrence of amino acid bracketed Gly and Ala. Only complete structures were accepted, and if relevant, those with an "alternative location indicator" of type "A". Identical or near identical structures were removed. Our definition of "near identical" structures were those geometries $\mathbf{R}_j$, where $\mathbf{R}_j$ is a vector of interatomic distances between heavy atoms (see Figure 1), which were closer than $|\mathbf{R}_i - \mathbf{R}_j| < 10^{-2}$ Å. Furthermore, "unlikely" geometries were also removed. Here "unlikely" means those geometries so distorted that it would seem to us unlikely that they would appear in any real protein. The criterion used was energy based (at the HF/6-31G level) and were all those structures higher in energy than 35 m-Eh for Gly and 44

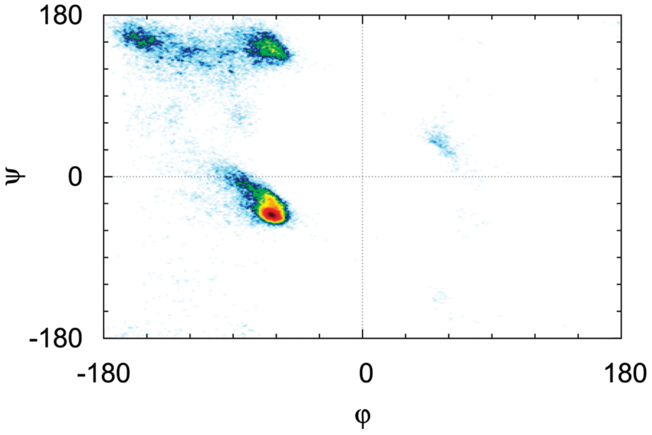

Figure 3. A 2-dimensional projection (φ and ψ) of the 15-dimensional data set extracted from the .pdb files for Ala. Red indicates high data density while blue is low data density.

m-Eh for Ala above the lowest energy structure encountered in the data set. These conditions excluded 37 and 18 geometries for Gly and Ala, respectively. After the above preparation of the data sets, the final number of geometries accepted for Gly and Ala were 41 544 and 42 849, respectively, with standard deviations in their energies of 3.42 and 3.53 m-Eh, respectively. Figures 2 and 3 show the 2-dimensional projections (φ and ψ) of the 12- and 15-dimensional data sets extracted for Gly and Ala, respectively. It is gratifying to note that these figures bear a striking resemblance to "textbook" Ramachandran plots for these residues.

Due to the nature of X-ray crystallography with its propensity to place hydrogen atoms too close to heavy atoms, all H atoms were removed from Gly and Ala residues on those few occasions when they were available. Because the intention of these extracted geometries is to form part of a fragment database, the carbon atom (C_X) adjacent to the nitrogen (N) and the nitrogen atom (N_X) adjacent to the carbonyl carbon (C) were also extracted because it is these two bonds (C_X−N and N_X−C) that will be broken and "capped" with hydrogen atoms to form the fragment residue (see Figure 1), an amino aldehyde.

Addition of H atoms to the heavy atoms was precisely and uniquely determined from the coordinates of the heavy atoms. Addition of H atoms to the Gly residue proceeded as follows. The capping H atom attached to N (see Figure 1) was located along the N−C_X bond at a distance of $r^0_{NH}(r_{NCX}/rNC_x{}^0)$ from N, where $rNC_x{}^0 = 1.32$ Å $= rCN_x{}^0$ and $r^0_{NH} = 0.99$ Å. Likewise for the capping H attached to C, except $r^0_{CH}= 1.07$ Å was used. The remaining H atom attached to N was placed along the negative of the average of the unit vectors $\hat{\mathbf{e}}_{\mathbf{NC_x}}$ and $\hat{\mathbf{e}}_{\mathbf{NC_\alpha}}$ with a bond length of r^0_{NH}. The two H atoms attached to the C_α were placed a distance r^0_{CH} along unit vectors whose end points were equally distant from each other and the end points of the unit vectors of $\hat{\mathbf{e}}_{\mathbf{C_\alpha N}}$ and $\hat{\mathbf{e}}_{\mathbf{C_\alpha C}}$.

Addition of H atoms to the Ala residue differed from Gly at the C_α only. For Ala there is an additional C atom bonded to C_α namely C_β. The single H atom attached to C_α was located r^0_{CH} in the negative direction of the average of the three unit vectors, $\hat{\mathbf{e}}_{\mathbf{C_\alpha N}}$, $\hat{\mathbf{e}}_{\mathbf{C_\alpha C}}$, and $\hat{\mathbf{e}}_{\mathbf{C_\alpha C_\beta}}$. Having placed the H_α atom, the three H atoms of the C_β can be uniquely located. The first H atom, $H_{\beta 1}$, was placed anti to H_α using the tetrahedral angle for $\angle C_\alpha C_\beta H_{\beta 1}$ and at a distance of r^0_{CH} from C_β. The remaining two H atoms, $H_{\beta 2}$ and $H_{\beta 3}$, were located such that the methyl group possessed C_{3v} symmetry with the C_3 axis lying along C_β−C_α.

Note that the numbers of degrees of freedom available to Gly and Ala were reduced because the positions of the H atoms were entirely dependent on the heavy atom coordinates. Thus the numbers of degrees of freedom for Gly and Ala were 12 and 15, respectively. In generating the internal coordinates used in the Taylor series expansion, bond lengths, angles, and dihedrals were all referred to heavy atoms.

2.2.2. Fragment Energies, Distributed Multipoles, and Electrostatic Potentials. To facilitate and assess our method for interpolating energies and distributed multipoles, the energies and multipoles of the entire Gly and Ala data sets were determined at the HF/6-31G level using the Gaussian 03 suite of programs.[71] Note that Shepard interpolation is entirely independent of level of theory, therefore it is only necessary for us to investigate our approach at the above crude level of theory in order to establish its validity and accuracy. Of course in the generation of an actual working database a much higher level of theory, including post-Hartree−Fock effects, would be utilized.

While we may directly compare interpolated energies with the ab initio energies to assess accuracy, comparing interpolated multipoles to those derived directly from the ab initio calculations does not provide much insight into the actual error generated, say, in the potential at meaningful locations in the vicinity of the molecule. In order to assess the success of interpolating the multipoles we have computed the electrostatic potential at the solvent accessible surface using: (a) the interpolated multipoles and (b) those derived from a distributed multipole analysis as well as (c) directly from the electronic wave function. Comparison of (b) with (c) provides us with an indication of the errors associated with using distributed multipoles for predicting the electrostatic potential. Comparison of (a) with (b) provides us with an indication of the errors associated with the interpolation.

For each configuration in the data set, points were placed on the solvent accessible surface at a density of approximately 1 point/Å^2. The solvent accessible surface was located using a probe radius of 1.4 Å and the Bondi van der Waals radii of 1.20, 1.70, 1.55, and 1.52 Å for H, C, N, and O respectively.[75] The algorithm used to locate the solvent accessible surface was essentially that found in Appendix II of ref 76 except applied to the solvent accessible surface rather than the molecular surface. The average surface areas of Gly and Ala were 210 and 236 Å^2, respectively.

A summary of the error in the computed potential at each point on the solvent accessible surface can be expressed as an root mean square (RMS) and an RRMS[77] for an individual fragment or over the entire data set. These quantities are defined below for the entire data set. For an individual fragment, *i*, the sum over *i* is excluded.

$$V_{\mathrm{RMS}} = \left[\frac{1}{M}\sum_{i=1}^{N_{\mathrm{frag}}}\sum_{j=1}^{N_i}(V_{i,j} - v_{i,j})^2\right]^{1/2} \quad (9)$$

$$V_{\mathrm{RRMS}} = \left[\frac{\sum_{i=1}^{N_{\mathrm{frag}}}\sum_{j=1}^{N_i}(V_{i,j} - v_{i,j})^2}{\sum_{i=1}^{N_{\mathrm{frag}}}\sum_{j=1}^{N_i}V_{i,j}^2}\right]^{1/2} \quad (10)$$

Here $V_{i,j}$ is the ab initio potential (or distributed multipoles potential) at point *j* in fragment *i*, and $v_{i,j}$ is the computed potential for the same fragment and point using distributed multipoles (or interpolated distributed multipoles). M is the total number of points in the double summation, i.e., $M = \sum_{i=1}^{N_{\text{frag}}} N_i$, and N_i is the number of points on the solvent accessible surface for fragment i.

2.2.3. Importance Sampling. Having selected the first geometry about which to expand a Taylor series, addition of further Taylor series proceeded as follows. The most efficient approach would seem to be one which adds a Taylor series that maximally reduces the interpolation errors. We define the energy interpolation error as an RMS of the residuals associated with the energy. Thus

$$E_{\text{RMS}} = \left[\frac{1}{M}\sum_{i=1}^{M}\gamma_i^2\right]^{1/2} \tag{11}$$

where $\gamma_i = E_i - \varepsilon_i$, and M is the number sampled points, which for Gly is 41 544. Addition of a Taylor series at geometry $\mathbf{Z}_i = \mathbf{Z}_a$ will at least eliminate all error associated with point i. If this Taylor series is in the near vicinity of many other similar geometries with significant error associated with them, then by adding this Taylor series to the interpolation data set we can also expect to reduce the interpolation error associate with the nearby geometries as well. Thus our goal is to select a geometry, about which we will expand a new Taylor series, that has significant error associated with it as well as many other neighboring geometries.

To proceed we note that the energy error in our truncated second-order Taylor series expanded about point a is

$$E_{\mathbf{Z}} - T_{\mathbf{Z}_a}(Z) = \mathcal{O}(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \approx f(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \tag{12}$$

where $d_{\mathbf{Z}_a}(\mathbf{Z}) = |\mathbf{Z} - \mathbf{Z}_a|$ in the vicinity of $\mathbf{Z}_a$. We now imagine we have added a new Taylor series to our interpolation data set at configuration $\mathbf{Z} - \mathbf{Z}_a$ and compute our new RMS energy error, E'_{RMS} which can readily be shown to be

$$M \times {E'_{\text{RMS}}}^2 = \sum_{i=1}^{M}\left\{\frac{s(i)}{s(i) + v_a(i)}\gamma_i + w_a(i) f(d_{\mathbf{Z}_a}(\mathbf{Z})^3)\right\}^2 \tag{13}$$

where $s(i) = \sum_{k=1}^{N} v_k(i)$, and N is the cardinality of the interpolation data set. If we imagine our new Taylor series to be particularly accurate, then we can set $f(d_{\mathbf{Z}_a}(\mathbf{Z})^3) \approx 0$, and we arrive at the expression used to select a new geometry about which we will expand a Taylor series:

$$t_a = \sum_{i=1}^{M}\left(\frac{s(i)}{s(i) + v_a(i)}\right)^2 \gamma_i^{\,2} \tag{14}$$

Using the above expression we compute a t_a for each value of i, i.e., there will be a total of M different values of t_a. The best point to choose to perform a new Taylor series expansion is about that point corresponding to the smallest value of t_a. Such a location would be one where there is a large number of similar structures each possessing a relatively large value of γ_i^2. In addition, once the point is determined that leads to the smallest value of t_a, the geometry could be further refined by minimizing t_a with respect to $\mathbf{Z}_a$. The geometry that minimized t_a would be chosen as the next point to add to the interpolation data set. By choosing such a point to add to the interpolation data set, we expect the greatest possible reduction in the RMS error.

Nevertheless, as the new Taylor series was assumed to be accurate, the sampling can become relatively inefficient later on in the growing process. This is because already sampled points may lie in regions of high data density, but the sampled point has nonzero error associated with it. In this case the smallest value of t_a is obtained by replacing the already sampled point with the assumed zero error Taylor series. That is, the point selected to be added to the interpolation data set already exists in the data set. This problem was resolved by introducing an expression for $f(d_{\mathbf{Z}_a}(\mathbf{Z})^3)$ into the formula for t_a once the cardinality of the interpolation data set was large enough.

$$f(d_{\mathbf{Z}a}(\mathbf{Z})^3) \approx E_{\text{tol}}\left(\frac{|\mathbf{Z}_a - \mathbf{Z}_i|}{\text{crad}_a}\right)^3 \tag{15}$$

where crad_a is the confidence radius at geometry $\mathbf{Z}_a$, which was assumed to be equal to $\min\{\text{crad}_i\}$, where i runs over the interpolation data set. The nonzero term $f(d_{\mathbf{Z}_a}(\mathbf{Z})^3)$ offers a means to incorporate into the t_a formula an approximate level of reliability of our existing Taylor series.

$$t_a = \sum_{i=1}^{N_{\text{neigh}}}\left[\frac{s(i)}{s(i) + v_a(i)}\gamma_i + w_a(i) E_{\text{tol}}\left(\frac{|\mathbf{Z}_a - \mathbf{Z}_i|}{\text{crad}_a}\right)^3\right]^2 \tag{16}$$

and N_{neigh} is the number of neighboring configurations, here set to 1000. In our work, the first 80 data points for both Gly and Ala were added using eq 14; the rest were added using eq 16.

The above-described "t_a method" of selecting a data point to add to the interpolation data set warrants further comment. It would seem that such a method of importance sampling a potential energy surface has little utility outside the present application. This is because the above expression requires the actual energy errors at all sampled geometries, while in general, such information is not available. However, related information is available in the form of an energy variance associated with each sampled point and has been described elsewhere as the "RMS method". The expression that provides the variance in the predicted energy at a given location is[78]

$$\sigma_E^2(\mathbf{Z}) = \sum_{j=1}^{N} w_j(\mathbf{Z})[\varepsilon_{\mathbf{Z}} - T_j(\mathbf{Z})]^2 \tag{17}$$

Thus locations where the energy is predicted to be very different values by Taylor expansions possessing high weights are locations of high energy uncertainty. By substitution of eq 17 for γ_i^2 in t_a, we are able to importance sample in regions of configuration space that contributes most to the uncertainty of the interpolated potential energy surface. That is

$$\tau_a = \sum_{i=1}^{M}\left(\frac{s(i)}{s(i) + v_a(i)}\right)^2 \sum_{j=1}^{N} w_j(i)[\varepsilon_{\mathbf{Z}} - T_j(\mathbf{Z})]^2 \tag{18}$$

By selecting the point, i, that produces the smallest value of τ_a then minimizing with respect to $\mathbf{Z}_a$, we expect to obtain the best reduction in overall uncertainty in the interpolated potential energy surface.

It is of note that this approach may be superior to the previous RMS method of importance sampling a potential energy surface in reaction dynamics[79] and stationary state calculations.[68] This is because it will select a configuration associated with significant uncertainty in the interpolated surface at locations of high configuration density encountered while sampling. As such the above τ_a sampling method incorporates both the RMS- and h-weight[80] sampling methods within a single method and does not require any constraints in the minimization carried out in ref 79, so this should improve the efficiency of sampling and thus reduce computational expense. We are planning to further investiage this method in a later publication.

3. RESULTS AND DISCUSSION

3.1. Interpolation Data Sets. The first point added to each of the interpolation data sets was the lowest energy configuration from the corresponding sample sets. The next 39 points were added using the one-part weight function and the t_a sampling method on the electronic energies. Continuing with t_a sampling on the electronic energy, the next 40 points utilized the two-part weight function. Following this, 30 points utilized t_a sampling in the electronic potential at the solvent accessible surface, then 20 points were added by sampling on the energy, and finally 20 further points were added by sampling on the potential. Additionally for Gly a further five points were sampled using the t_a method on the electronic energy. Thus the final cardinality of the interpolation data sets was 150 for Ala and 155 for Gly. Figures 4 and 5 show the RMS error in the energy, E_{RMS}, and the maximum absolute error in the energy, $|\gamma_E|_{\max}$, respectively, as a function of the cardinality of the interpolation data sets.

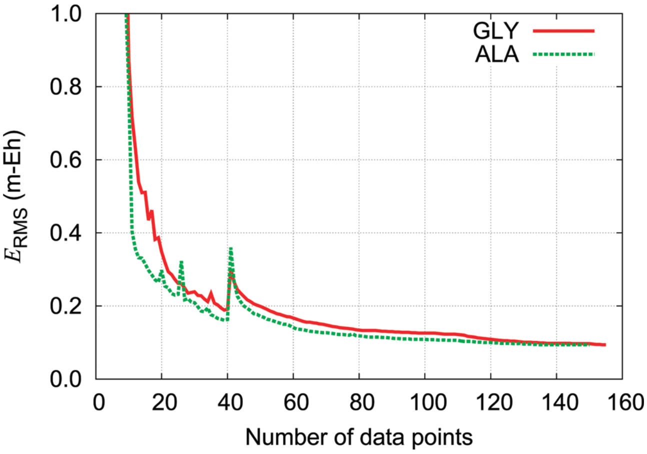

Figure 4. E_{RMS} for Gly and Ala as a function of the number of data points.

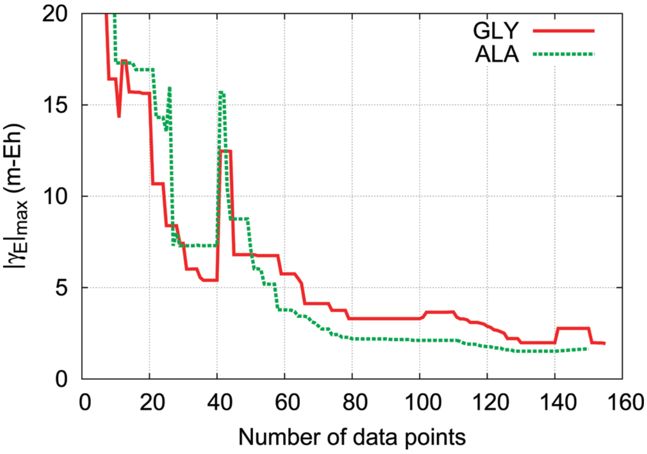

Figure 5. $|\gamma_E|_{\max}$ for Gly and Ala as a function of the number of data points.

As expected, the t_a method smoothly reduces the E_{RMS} and $|\gamma_E|_{\max}$. A peak was observed in both plots at 40 data points as a result of switching from one- to two-part weight function. The RMS error as well as the maximum absolute error in the energy using our final interpolation data sets were evaluated to be respectively 0.094 and 1.942 m-Eh for Gly and 0.094 and 1.654 m-Eh for Ala. Thus using less that 0.4% of the sample set, the modified Shepard interpolation is capable of reproducing the vast majority of electronic energies to better than 0.1 m-Eh with no sampled configuration possesses an electronic energy error greater than 2 m-Eh. This level of accuracy should be sufficient for even the most demanding applications.

3.2. The Electrostatic Potential. The accuracy in reproducing the ab initio electrostatic potential using the modified Shepard interpolation depends upon several factors. First, since we are utilizing distributed multipoles to compute the electrostatic potential, we need to evaluate their accuracy at doing so for a given multipole rank and site selection. Second, the errors associated with interpolating the multipoles also impacts on how well the ab initio potential can be reproduced. We require that this second contribution to the errors in the potential at the solvent accessible surface to be minimized or even negligible in comparison to the errors associated with the first.

Previously we concluded from the results of several test molecules that rank two multipoles were sufficient to obtain errors less than or about equal to 1 m-Eh in the predicted electrostatic potential around the solvent accessible surface.[54] We wish to verify that this is the case for the two selected sample sets here. As such, distributed multipoles up to rank five were computed from the HF/6-31G wave function using the GDMA2 program[2] for all geometries in each sample set. As described in the Approach Section, the potential was computed at points on the solvent accessible surface using a density of about 1/Å, and the V_{RMS} and V_{RRMS} were evaluated. For Gly and Ala this amounted to about 9×10^6 and 10×10^6 points, respectively. A summary of the results is provided in Table 1.

Not surprisingly, Table 1 shows that the distributed multipoles are capable of producing near exact agreement with the potential

Table 1. V_{RMS} **and** V_{RRMS} **Error at the Solvent Accessible Surface in the Electrostatic Potential between That Computed with Distributed Multipoles to the Rank Indicated and the ab Initio Potential**

	Gly all atoms		Gly heavy and cap hydrogens		Ala all atoms		Ala heavy and cap hydrogens	
rank	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)
0	4.12	27.45	21.88	145.49	4.11	29.86	22.29	162.13
1	3.91	26.02	4.41	29.39	3.61	26.26	3.98	28.95
2	0.64	4.27	0.89	5.94	0.64	4.64	0.85	6.21
3	0.09	0.59	0.23	1.51	0.11	0.77	0.27	1.94
4	0.04	0.28	0.11	0.74	0.05	0.34	0.15	1.06
5	0.02	0.11	0.05	0.34	0.02	0.16	0.08	0.58

Table 2. V_{RMS} and V_{RRMS} Errors between the Electrostatic Potential at the Solvent Accessible Surface Computed with the Interpolated Distributed Multipoles and That Computed by the Exact Distributed Multipoles as well as ab Initio Potential to the Rank Indicated

	compared to Stone's exact DMs				compared to ab initio potential			
	Gly		Ala		Gly		Ala	
rank	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)	V_{RMS}	V_{RRMS} (%)
0	0.044	0.23	0.050	0.24	21.881	145.79	22.289	162.12
1	0.085	0.57	0.085	0.61	4.413	29.40	3.982	28.96
2	0.157	1.04	0.241	1.75	0.904	6.02	0.895	6.51

Table 3. V_{RMS} Error at the Solvent Accessible Surface in the Electrostatic Potential for Each Atom between That Computed with Distributed Multipoles at Rank Two

Gly		Ala	
N	0.204	N	0.313
Cα	0.223	Cα	0.452
C	0.149	C	0.293
O	0.108	O	0.174
H_{CX}	0.160	C_{β}	0.225
$H_{\alpha2}$	0.163	H_{CX}	0.223
$H_{\alpha3}$	0.155	Hα	0.334
H	0.203	$H_{\beta1}$	0.213
H_{NX}	0.132	$H_{\beta2}$	0.207
		$H_{\beta3}$	0.214
		H	0.284
		H_{NX}	0.243
V_{RMS} (m-au)	0.157		0.241

obtained from the electronic wave function at the solvent accessible surface. V_{RMS} errors as low as 20 μ-au can be obtained with rank five multipoles located on all nuclei. It is noted that only a small reduction in accuracy is obtained if multipoles are centered on heavy and capping hydrogens, except in the case of distributed charges. However, in a situation where a database is to be used, say, to perform molecular dynamics or Monte Carlo simulations, it would seem that reducing the rank of the multipoles and the number of sites, while still provide adequate accuracy, would be best to select. It is evident from Table 1 that rank two with multipole sites placed on heavies and capping hydrogens represents a good trade-off, as the accuracy is still better than 1 m-au.

Next we consider the contribution of the errors directly associated with the modified Shepard interpolation. The potential on the solvent accessible surface was computed from the interpolated multipoles and then compared to both the ab initio potential and the potential computed from the exact multipoles at the same rank. V_{RMS} and V_{RRMS} were evaluated. Indeed, it is evident from Table 2 that the potential computed from the interpolated multipoles agrees well with that using the exact multipoles. Carefully comparing Tables 1 and 2 shows that our interpolated potential leads to negligible additional error other than that produced by using exact rank two multipoles placed on heavy atoms and capping hydrogens.

Finally we examine how well the electrostatic potential can be reproduced compared to Stone's exact distributed multipoles in regions associated with each atom on the solvent accessible surface. The results are provided in Table 3. It is clear from this table that there are no regions around either molecule that are described particularly poorly with our method.

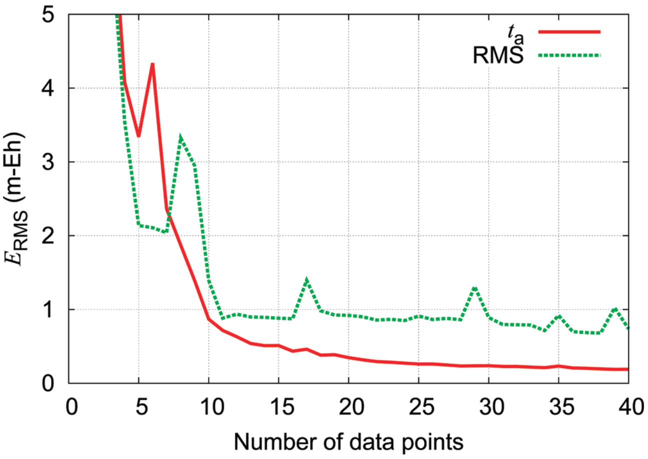

Figure 6. E_{RMS} using t_a and RMS methods as a function of the number of data points.

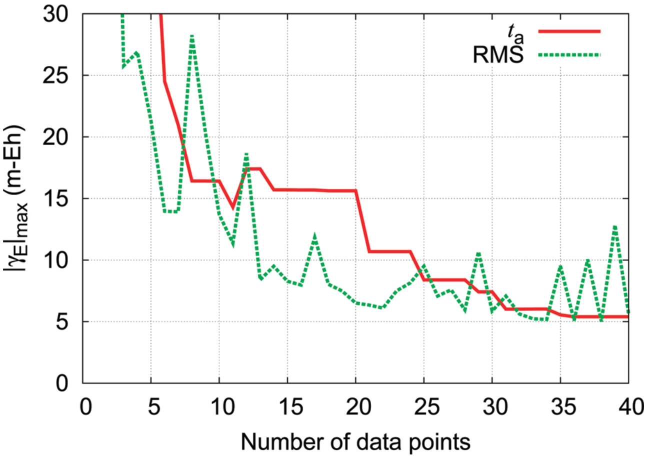

Figure 7. $|\gamma_E|_{max}$ using t_a- and RMS methods as a function of the number of data points.

3.3. The Importance Sampling Method. To verify that our t_a method is superior to the previous RMS method, E_{RMS} and $|\gamma_E|_{max}$ were computed for the first 40 data points of the Gly potential energy surface using these two methods. The results are illustrated in Figures 6 and 7. It is apparent that the t_a method achieves much lower E_{RMS} and comparable $|\gamma_E|_{max}$. Moreover, the fluctuations appearing in both figures for the RMS method are more frequent and pronounced, which implies more instability in the early stages of "growing" the potential energy surface. Worse is that for the same number of data points, beyond a very small number, the E_{RMS} using the RMS method is approximately four times greater than that obtained using the t_a method, implying greater computational expense in generating the potential energy surface.

4. CONCLUSION

We showed that for the 12- and 15-dimensional systems of the amino aldehydes, Gly and Ala, that the RMS energy error in electronic energies can be interpolated to better than 0.1 m-Eh for more than 41 000 different configurations encountered in protein X-ray structures. We also showed that distributed multipoles up to and including rank two can be interpolated very accurately so that negligible additional error is introduced into the calculation of electrostatic potential generated at the solvent accessible surface by the exact distributed multipoles. Rank two distributed multipoles lead to less that 1 m-au error in the potential at the solvent accessible surface. Considerable improvement in this accuracy was obtained by including rank three multipoles. The modified Shepard interpolation used in determining the interpolated energies and distributed multipoles required a small number of configurations selected using a newly described efficient sampling method, the "t_a method". This small number of points was selected from a set of over 41 000 different configurations encountered in protein X-ray data. Multipoles were also interpolated in an "axis-free" manner, which alleviated difficulties encountered in interpolating Cartesian components.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. A complete listing of all the . pdb files used in this work can be found in Table S1. Also included in the Supporting Information are the Cartesian coordinates of the heavy atoms of all of the molecules included in the interpolation data sets for Gly (Table S2) and Ala (Table S3).This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: chmbrpa@nus.edu.sg.

■ ACKNOWLEDGMENT

The authors thanks the National University of Singapore's support from the Academic Research Fund, grant number R-143-000-402-112. The authors also thank the Centre for Computational Science and Engineering for the use of their computers.

■ REFERENCES

(1) Stone, A. J.; Alderton, M. *Mol. Phys.* **1985**, *56*, 1047–1064.

(2) Stone, A. J. *J. Chem. Theo. Comput.* **2005**, *1*, 1128–1132.

(3) Kosov, D. S.; Popelier, P. L. A. *J. Phys. Chem. A* **2000**, *104*, 7339–7345.

(4) Whitehead, C. E.; Breneman, C. M.; Sukumar, N.; Ryan, M. D. *J. Comput. Chem.* **2003**, *24*, 512–529.

(5) Sokalski, W. A.; Poirier, R. A. *Chem. Phys. Lett.* **1983**, *98*, 86–92.

(6) Sokalski, W. A.; Sawaryn, A. *J. Chem. Phys.* **1987**, *87*, 526–534.

(7) Koch, U.; Stone, A. J. *J. Chem. Soc. Faraday Trans.* **1996**, *92*, 1701–1708.

(8) Brodersen, S.; Wilke, S.; Leusen, F. J. J.; Engel, G. *Phys. Chem. Chem. Phys.* **2003**, *5*, 4923–4931.

(9) Koch, U.; Popelier, P. L. A.; Stone, A. J. *Chem. Phys. Lett.* **1995**, *238*, 253–260.

(10) Karamertzanis, P. G.; Price, S. L. *J. Chem. Theo. Comput.* **2006**, *2*, 1184–1199.

(11) Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.

(12) Ren, P. Y.; Ponder, J. W. *J. Comput. Chem.* **2002**, *23*, 1497–1506.

(13) Ren, P. Y.; Ponder, J. W. *J. Phys. Chem. B* **2003**, *107*, 5933–5947.

(14) Liang, T.; Walsh, T. R. *Phys. Chem. Chem. Phys.* **2006**, *8*, 4410–4419.

(15) Ischtwan, J.; Collins, M. A. *J. Chem. Phys.* **1994**, *100*, 8080–8088.

(16) Collins, M. A. *Theo. Chem. Acc.* **2002**, *108*, 313–324.

(17) Bettens, R. P. A.; Collins, M. A. *J. Chem. Phys.* **1999**, *111*, 816–826.

(18) Thompson, K. C.; Jordan, M. J. T.; Collins, M. A. *J. Chem. Phys.* **1998**, *108*, 8302–8316.

(19) Devereux, M.; Popelier, P. L. A.; McLay, I. M. *J. Comput. Chem.* **2009**, *30*, 1300–1318.

(20) Rahalkar, A. P.; Ganesh, V.; Gadre, S. R. *J. Chem. Phys.* **2008**, *129*, 234101.

(21) Ganesh, V.; Dongare, R. K.; Balanarayan, P.; Gadre, S. R. *J. Chem. Phys.* **2006**, *125*, 104109.

(22) Gadre, S. R.; Shirsat, R. N.; Limaye, A. C. *J. Phys. Chem.* **1994**, *98*, 9165–9169.

(23) Babu, K.; Gadre, S. R. *J. Comput. Chem.* **2003**, *24*, 484–495.

(24) Babu, K.; Ganesh, V.; Gadre, S. R.; Ghermani, N. E. *Theo. Chem. Acc.* **2004**, *111*, 255–263.

(25) Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2003**, *119*, 3599–3605.

(26) Mei, Y.; Ji, C.; Zhang, J. Z. H. *J. Chem. Phys.* 2006, *125*, 094906.

(27) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2006**, *125*, 044903.

(28) Chen, X. H.; Zhang, Y. K.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*, 184105.

(29) Zhang, D. W.; Zhang, J. Z. H. *Int. J. Quantum Chem.* **2005**, *103*, 246–257.

(30) He, X.; Zhang, J. Z. H. *J. Chem. Phys.* **2005**, *122*.

(31) Mei, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Phys. Chem. A* **2005**, *109*, 2–5.

(32) Chen, X. H.; Zhang, J. Z. H. *J. Theo. Comput. Chem.* **2004**, *3*, 277–289.

(33) Gao, A.; Zhang, D. W.; Zhang, J. Z. H.; Zhang, Y. K. *Chem. Phys. Lett.* **2004**, *394*, 293–297.

(34) Xiang, Y.; Zhang, D. W.; Zhang, J. Z. H. *J. Comput. Chem.* **2004**, *25*, 1431–1437.

(35) Chen, X. H.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 11386–11391.

(36) Zhang, D. W.; Xiang, Y.; Gao, A. M.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 1145–1148.

(37) Chen, X. H.; Zhang, D. W.; Zhang, J. Z. H. *J. Chem. Phys.* **2004**, *120*, 839–844.

(38) Zhang, D. W.; Xiang, Y.; Zhang, J. Z. H. *J. Phys. Chem. B* **2003**, *107*, 12039–12041.

(39) Zhang, D. W.; Chen, X. H.; Zhang, J. Z. H. *J. Comput. Chem.* **2003**, *24*, 1846–1852.

(40) Li, S. H.; Li, W.; Fang, T. *J. Am. Chem. Soc.* **2005**, *127*, 7215–7226.

(41) Dong, H.; Hua, S.; Li, S. *J. Phys. Chem. A* **2009**, *113*, 1335–1342.

(42) Li, W.; Dong, H.; Li, S. Relative Energies of Proteins and Water Clusters Predicted with the Generalized Energy-Based Fragmentation

Approach. 12th European Workshop on Quantum Systems in Chemistry and Physics, London, England, August 30—September 5, 2007; Springer: 2008.

(43) Hua, W.; Fang, T.; Li, W.; Yu, J.-G.; Li, S. *J. Phys. Chem. A* **2008**, *112*, 10864–10872.

(44) Li, H.; Li, W.; Li, S.; Ma, J. *J. Phys. Chem. B* **2008**, *112*, 7061–7070.

(45) Li, W.; Li, S.; Jiang, Y. *J. Phys. Chem. A* **2007**, *111*, 2193–2199.

(46) Li, W.; Fang, T.; Li, S. H. *J. Chem. Phys.* **2006**, *124*, 154102.

(47) Deev, V.; Collins, M. A. *J. Chem. Phys.* **2005**, *122*, 154102.

(48) Collins, M. A.; Deev, V. A. *J. Chem. Phys.* **2006**, *125*, 104104.

(49) Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 024104.

(50) Netzloff, H. M.; Collins, M. A. *J. Chem. Phys.* **2007**, *127*, 134113.

(51) Bettens, R. P. A.; Lee, A. M. *J. Phys. Chem. A* **2006**, *110*, 8777–8785.

(52) Lee, A. M.; Bettens, R. P. A. *J. Phys. Chem. A* **2007**, *111*, 5111–5115.

(53) Bettens, R. P. A.; Lee, A. M. *Chem. Phys. Lett.* **2007**, *449*, 341–346.

(54) Le, H.-A.; Lee, A. M.; Bettens, R. P. A. *J. Phys. Chem. A* **2009**, *113*, 10527–10533.

(55) Yang, W. T.; Lee, T. S. *J. Chem. Phys.* **1995**, *103*, 5674–5678.

(56) Nakano, T.; Kaminuma, T.; Sato, T.; Akiyama, Y.; Uebayasi, M.; Kitaura, K. *Chem. Phys. Lett.* **2000**, *318*, 614–618.

(57) Fedorov, D. G.; Kitaura, K. *J. Phys. Chem. A* **2007**, *111*, 6904–6914.

(58) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. J.; Koseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. *J. Comput. Chem.* **1993**, *14*, 1347–1363.

(59) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2002**, *106*, 11791–11800.

(60) Exner, T. E.; Mezey, P. G. *J. Comput. Chem.* **2003**, *24*, 1980–1986.

(61) Exner, T. E.; Mezey, P. G. *J. Phys. Chem. A* **2004**, *108*, 4301–4309.

(62) Exner, T. E.; Mezey, P. G. *Phys. Chem. Chem. Phys.* **2005**, *7*, 4061–4069.

(63) Eckard, S.; Exner, T. E. *Int. J. Res. Phys. Chem. Chem. Phys.* **2006**, *220*, 927–944.

(64) Stone, A. J. *The Theory of Intermolecular Forces*; Clarendon: Oxford, U.K., 2002.

(65) Frankcombe, T. J.; Collins, M. A.; Worth, G. A. *Chem. Phys. Lett.* **2010**, *489*, 242–247.

(66) Cao, J. W.; Zhang, Z. J.; Zhang, C. F.; Liu, K.; Wang, M. H.; Bian, W. S. *Proc. Nat. Acad. Sci. U.S.A.* **2009**, *106*, 13180–13185.

(67) Wu, T.; Werner, H. J.; Manthe, U. *J. Chem. Phys.* **2006**, *124*, 164307.

(68) Bettens, R. P. A. *J. Am. Chem. Soc.* **2003**, *125*, 584–587.

(69) Yagi, K.; Oyanagi, C.; Taketsugu, T.; Hirao, K. *J. Chem. Phys.* **2003**, *118*, 1653–1660.

(70) Kazantsev, A. V.; Karamertzanis, P. G.; Pantelides, C. C.; Adjiman, C. S. In *Molecular Systems Engineering*; Adjiman, C. S., Galindo, A., Eds.; Wiley: Weinheim, Germany, 2010; Vol. 6.

(71) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb,M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; ; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen,W.;Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*; Gaussian, Inc.: Wallingford, CT, 2004.

(72) Prim, R. C. *Bell Syst. Tech. J.* **1957**, 1389–1401.

(73) Jarník, V. *Acta Soc. Sci. Nat. Moravicae* **1930**, *6*, 57–63.

(74) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(75) Bondi, A. *J. Phys. Chem.* **1964**, *68*, 441–451.

(76) Connolly, M. L. *J. Appl. Crystallogr.* **1983**, *16*, 548–558.

(77) Bayly, C.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269–10280.

(78) Thompson, K. C.; Collins, M. A. *J. Chem. Soc. Faraday Trans.* **1997**, *93*, 871–878.

(79) Moyano, G. E.; Collins, M. A. *J. Chem. Phys.* **2004**, *121*, 9769–9775.

(80) Jordan, M. J. T.; Thompson, K. C.; Collins, M. A. *J. Chem. Phys.* **1995**, *102*, 5647–5657.