Canadian Journal of Chemistry
Revue canadienne de chimie

An Improved Two-Rotor Function for Conformational Potential Energy Surfaces of 20 Amino Acid Diamides

Journal:	<i>Canadian Journal of Chemistry</i>
Manuscript ID	cjc-2017-0571.R2
Manuscript Type:	Article
Date Submitted by the Author:	26-Oct-2017
Complete List of Authors:	Villar, John Justine; University of the Philippines Diliman, Department of Computer Science; Miskolci Egyetem, Institute of Chemistry Valdez, Adrian Roy; University of the Philippines Diliman, Department of Computer Science Setiadi, David H.; University of Toronto, Chemistry Csizmadia, G. Imre; University of Toronto, Department of Chemistry; Miskolci Egyetem, Institute of Chemistry Viskolcz, Béla; Miskolci Egyetem, Institute of Chemistry Rágyanszki, Anita; University of Toronto, Department of Chemistry; Miskolci Egyetem, Institute of Chemistry
Is the invited manuscript for consideration in a Special Issue?:	N/A
Keyword:	potential energy surface, conformational analysis, mathematical modeling

SCHOLARONE™
Manuscripts

An Improved Two-Rotor Function for Conformational Potential Energy Surfaces of 20 Amino Acid Diamides

John Justine S. Villar, Adrian Roy L. Valdez, David H. Setiadi, Imre G. Csizmadia, Béla Viskolcz, and Anita Rágyanszki

Abstract: Predicting the three-dimensional structure of a protein from its amino acid sequence requires a complete understanding of the molecular forces that influences the protein folding process. Each possible conformation has its corresponding potential energy, which characterizes its thermodynamic stability. This is needed to identify the primary intra- and intermolecular interactions, so that we can reduce the dimensionality of the problem, and create a relatively simple representation of the system. Investigating this problem using quantum chemical methods, albeit produces accurate results, this also entails large computational resources needed.

In this study, an improved two-rotor potential energy function is proposed to represent the backbone interactions in amino acids, through a linear combination of a Fourier series and a mixture of Gaussian functions. This function is applied to approximate the 20 amino acid diamide Ramachandran-type PESs, and results yielded an average RMSE of 2.36 kJ·mol⁻¹, which suggest that the mathematical model precisely captures the general topology of the conformational potential energy surface. Furthermore, this paper provides insights on the conformational preferences of amino acid diamides through local minima geometries and energy ranges, using the improved mathematical model.

The proposed mathematical model presents a simpler representation that attempts to provide a framework on building polypeptide models from individual amino acid functions, and consequently, a novel method for rapid but accurate evaluation of potential energies for biomolecular simulations.

Key words: potential energy surface, conformational analysis, mathematical modeling.

John Justine S. Villar. Scientific Computing Laboratory, Department of Computer Science, University of the Philippines, Diliman, 1101 Quezon City, Philippines; Institute of Chemistry, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

Adrian Roy L. Valdez. Scientific Computing Laboratory, Department of Computer Science, University of the Philippines, Diliman, 1101 Quezon City, Philippines

David H. Setiadi. Department of Chemistry, University of Toronto, M5S 3H6 Toronto, Ontario, Canada

Imre G. Csizmadia. Institute of Chemistry, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary; Department of Chemistry, University of Toronto, M5S 3H6 Toronto, Ontario, Canada

Béla Viskolcz. Institute of Chemistry, Faculty of Materials Science and Engineering, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary

Anita Rágyanszki.¹ Institute of Chemistry, University of Miskolc, 3515 Miskolc-Egyetemváros, Hungary; Department of Chemistry, University of Toronto, M5S 3H6 Toronto, Ontario, Canada

¹Corresponding author (e-mail: a.ragyanszki@utoronto.ca).

1. Introduction

Proteins, which are assembled from 20 amino acids, are machines and building blocks of living cells. The macromolecules are vital for structure, function and regulation of various biological processes.¹ After the successful deciphering of the genetic code that defines how the amino acid sequences of proteins are coded in the DNA, one of the major missing steps in understanding the chemical basis of life is the *protein folding problem*—the task of understanding and predicting how the information coded in the (one-dimensional) amino acid sequence of proteins at the time of their formation translates into the three-dimensional structure of the biologically active protein.^{2–4} Because of this, there is a great interest to determine the factors affecting protein folding, and consequently, its connection to its biological function.⁵

Predicting the three-dimensional structure from its amino acid sequence involves a complete understanding of the molecular forces driving the protein to fold. Each possible conformation has its corresponding potential energy, which characterizes its thermodynamic stability. To find the stable conformers of a certain protein, its entire conformational space should be generated and/or sampled, creating an n -dimensional hypersurface, and the set of local minima of these hypersurface may correspond to the energy of the set of most stable conformations. The relationship between the potential energy of a molecule (in this case, a protein conformer) with respect to its geometry is commonly referred to as the potential energy surface.

A potential energy surface (PES) is important in understanding how structure prediction methods locate and characterize its preferred conformations. These are usually generated through electronic structure calculations, following quantum mechanical principles.⁶ However, the computational time and space required to produce PESs increases exponentially with the number of atoms in the system, as well as the level of theory applied.⁷

Conformational analysis of simple organic molecules has been initiated in the 1970s,^{8–12} where it has been found that methyl ($-\text{C}-\text{CH}_3$) and ethyl rotations ($-\text{CH}_2-\text{CH}_2$) were very similar in a group of hydrocarbons, and one-dimensional trigonometric function was used to relate to the topology of the PEC or PES of these compounds.¹³ Furthermore, in 1963, Ramachandran attempted to perform conformational analysis on simple peptides,¹⁴ and the pattern for such a double rotor $\text{CONH}-\text{CH}_2-\text{CONH}$ turned out to be quite different from the propane surface. More recently, it turned out that even a simple compound as n -pentane behaved in non-ideal fashion.¹⁵ Thus, it is reasonable to assume that the surface topology is a function of two dihedral angles ϕ and ψ associated with the rotation about two bonds connected to the α -carbon of the amino acid. This also follows the results from experiments¹⁶ which show that regular patterns in backbone dihedral angles determine its secondary structure, such as α -helices and β -sheets.

The computational investigation of the PESs of the amino acids in terms of backbone dihedral angles ϕ and ψ spans more than twenty years. Several studies include alanine,^{17–20} valine,²¹ glutamic acid,²² serine,^{23–25} phenylalanine,^{26–29} proline,^{30–32} tryptophan,^{33,34} asparagine,^{35–39} glycine,⁴⁰ cysteine,^{41–44} glutamine,^{45–47} isoleucine,⁴⁸ aspartic acid,⁴⁹ histidine,⁵⁰ tyrosine,⁵¹ threonine,^{52,53} methionine,⁵⁴ arginine,⁵⁵ lysine,⁵⁶ leucine,⁵⁷ and selenocysteine,^{58,59} with some papers tackled multiple amino acids.^{60–67}

As performing quantum chemical calculations is inherently computationally expensive, approximations using analytical fitting methods have been widely used in the construction of PESs.^{68,69} For instance, a fitting procedure based on the many-body expansion (MBE) method,⁷⁰ amplified by Varandas⁷¹ is widely used in the so-called double many-body expansion (DMBE) method. The potential is given by a sum of terms corresponding to atoms, diatoms, triatoms, and tetra-atoms and a series of adjustable coefficients. Another fitting procedure has been developed and applied by Bowman et al.,^{72,73} in which the *ab initio* data are globally fitted to a permutational symmetry invariant polynomial.

Although these methods model bonded interactions, these methods are suitable for small polyatomic systems, such as CH_5^+ and CH_2D_3 . In the paper, they modeled the PES with great accuracy but

Villar et al.

using more than 2300 coefficients, entailing vast amount of computational resources for a relatively small molecule. For larger molecules, such as proteins, it is important to understand the various interactions within them and identify primary factors that drive the folding process, so that we can reduce the dimensionality of the problem, and create a relatively simple representation of the system.

The molecular properties and energy of a protein can, in principle, be predicted using smaller fragments, such as amino acids, as local interactions heavily influence the formation of secondary structure. Following this assumption, a previous work⁷⁴ attempted to represent the energy landscape of the amino acids by a two-rotor mathematical model with respect to its backbone dihedral angles, through a linear combination of Fourier series and a mixture of Gaussian functions. In the paper, it assumed that backbone interactions mainly drive protein folding. The function was initially tested only on three amino acid diamides, i.e., glycine, alanine and valine, and results showed that the two-rotor model can be used to explore the relative stability of the conformers, with acceptable precision.

This study presents a further improvement of the mathematical model from a related study,⁷⁴ in which a combination of an alternate form of Fourier series and a mixture of spherical Gaussian functions was used, which performs better in terms of scalability and tractability of parameter estimation.

2. Methods

2.1. The Two-Rotor Potential Energy Function

Aside from providing a visualization of the energy landscape of the protein conformational space, potential energy surfaces can also be used, conversely, to explore the properties of the molecule. The PESs can be represented by a mathematical function that gives the potential energy of a molecule as a function of its geometry, e.g., bond stretches, bends, torsions, among others, and arises from the Schrödinger equation.^{6,75} Due to its inherent complexity, most calculations are performed using the Born-Oppenheimer approximation.^{6,76} However, for large molecular systems, this is still computationally expensive to use, as calculations scale exponentially with respect to the size of the molecule.

To reduce the dimensionality of the problem, it is important to have an understanding of the primary interactions between proteins. These interactions can be grouped into four categories: (i) backbone, (ii) backbone-sidechain, (iii) nearest neighbor, and (iv) long-range interactions.²¹ Several papers attempted to reduce the complexity of the current quantum chemical models to provide an approximation of the landscape of amino acid conformational space, by only looking onto the effects of backbone interactions to the overall energetics.⁷⁴ It has been shown that local energy minima can be deduced from a mathematical model (within an acceptable precision), considering only the backbone interaction effects.

Following these results from a previous work,⁷⁴ the paper is guided by two premises in developing a mathematical representation of the potential energy surface: (1) using global functions that cover a large region with acceptable accuracy, and (2) using functions that perform well in a local region.

Fourier series have been used in various force field functional forms to represent the torsional rotation of four consecutive atoms about a central bond, as the torsional movement is cyclic in nature. However, to account for different bonded and non-bonded interactions in more complex amino acids, it is important to find a suitable function that may account for dissymmetries in the PES.

In this paper, the potential energy function is constructed from a linear combination of

- a Fourier series of the form

$$\Delta E_{f*}(\phi, \psi) = \sum_{\substack{i,j \in \mathbb{Z}_{\geq 0} \\ 0 \leq i+j \leq N_f}} A_{i,j} \cos[i\omega(\phi - u_{i,j})] \cos[j\omega(\psi - v_{i,j})] \quad (2.1)$$

where $A_{i,j}$ be the amplitude, $\omega = \frac{180^\circ}{\pi}$ is the conversion factor from degrees to radians, and $u_{i,j}$

Published by NRC Research Press

Can. J. Chem.

and $v_{i,j}$ are translation parameters for ϕ and ψ , respectively, where i, j are nonnegative integers with $0 \leq i + j \leq n$; and

- a mixture of spherical Gaussian functions, which are fitted at the recognizable peaks of the input grid, of the form

$$\Delta E_g(\phi, \psi) = \sum_{k=1}^{N_g} B_k \exp \left[-\frac{c_{\phi_k}(\phi - \bar{\phi}_k)^2}{2\sigma_{\phi_k}^2} - \frac{c_{\psi_k}(\psi - \bar{\psi}_k)^2}{2\sigma_{\psi_k}^2} \right], \quad (2.2)$$

where N_g is the multiplicity, B_k is the amplitude, $\bar{\phi}_k, \bar{\psi}_k$ and $\sigma_{\phi_k}^2, \sigma_{\psi_k}^2$ is the center and spreads of the ellipsoids on the ϕ and ψ axes, respectively, all with $k = 1, \dots, N_g$. Also, $\exp(\cdot)$ is the exponential function.

Thus,

$$\Delta E_{f^*+g}(\phi, \psi) = \Delta E_{f^*}(\phi, \psi) + \Delta E_g(\phi, \psi). \quad (2.3)$$

Given a grid $G = [-180^\circ, 180^\circ]^2$, and the corresponding relative energy ΔE values over G computed from electronic structure calculations, denoted by $\Delta E(G)$, optimal parameters for $\Delta E_{f^*+g}(G)$ are estimated using the Levenberg-Marquardt algorithm.^{77,78} The parameter estimation were performed using a procedure detailed in Sec. 2.3 and implemented in the Python programming language.

Note that the form of the Fourier series used in this paper (eq. 2.1) is different from what is used in a related work,⁷⁴ as eq. A.1 (shown in Appendix A) assumes independence between the two variables, and these interactions are presumed to be compensated by the mixture of Gaussian functions. In eq. 2.1, on the other hand, there are terms having products of cosines of ϕ and ψ , which reflect the dependence of the two rotors in the function.

Despite of an increase in the number of parameters, the two equations proposed in this section are also scalable and the parameter estimation procedure detailed in Sec. 2.3 is easier to perform, resulting to shorter computation time. It is also apparent that the number of parameters in both equations eq. A.1 and eq. 2.1 can be expressed as a function of N_f , and while the number of parameters for eq. 2.1 is quadratic in N_f , and for eq. A.1 is linear in N_f , the number of parameters do not differ as much for small values of N_f . A discussion on Appendix A details the comparison between the two models in terms of their goodness-of-fit and parameter size.

2.2. Generation of Energy Data

This paper considered 20 amino acid diamides, which include glycine, alanine, valine, cysteine, serine, selenocysteine, aspartic acid, asparagine, threonine, phenylalanine, histidine, tyrosine, isoleucine, leucine, tryptophan, glutamic acid, methionine, glutamine, lysine, and arginine. Proline was not included in the study due to the complication made by the side chain. The molecular structures used are illustrated in Fig. 1, and they are referred in the succeeding sections by their respective three-letter codes.¹

For each of these molecules, the energy $E(\phi, \psi)$ corresponding to the rotations of two peptide backbone dihedral angles, ϕ and ψ , were computed over the interval $[-180^\circ, 180^\circ]^2$, with a step size of 15° . These values were generated from electronic structure calculations under the B3LYP hybrid functional^{79,80} and 6-31G(d) basis set⁸¹ implementation on GAUSSIAN09⁸² software package, and the input Z-matrices were constructed following a previous study.⁸³ Furthermore, the molecular geometries of the amino acid considered were fully relaxed except for the dihedral angles ϕ and ψ .

Published by NRC Research Press

Villar et al.

Fig. 1. Molecular structures of 20 amino acid Diamides, and the dihedral angle locations.

The relative energy, given by the formula

$$\Delta E(\phi_i, \psi_i) = E(\phi_i, \psi_i) - \min_{(\phi, \psi) \in [-180^\circ, 180^\circ]^2} E(\phi, \psi), \quad (2.4)$$

were also subsequently calculated, generating a 625-point grid surface, which was used as an input for the procedure discussed in the previous study.⁷⁴

The goodness-of-fit was determined using the root mean square error (RMSE) and R^2 between the resulting fit and relaxed scan relative energy values. Also, in checking the deviation of the molecular geometries from the local minima of ΔE_{f+g} from those obtained from electronic structure calculations, simple optimizations (using the same hybrid functional and basis set) are performed from the minima of the grid PES obtained to get what is referred to in this paper the DFT-optimized minima.

2.3. Procedure for Analytical Fitting

Given a grid $G \subseteq [-180^\circ, 180^\circ]^2$, and the corresponding relative energy ΔE values over G computed from DFT calculations, denoted by $\Delta E(G)$, the fitting procedure consists of three phases: (1)

Published by NRC Research Press

Can. J. Chem.

nonlinear least squares (NLS) fit on $\Delta E_{f^*}(\phi, \psi)$, (2) enumerating the grid minima of $\Delta E(G)$, and (3) NLS fit on $\Delta E_{f^*+g}(\phi, \psi)$. The Levenberg-Marquardt algorithm^{77,78} was used to obtain the optimal parameter values for ΔE_{f^*} and ΔE_{f^*+g} for the NLS fit procedures in the first and third phases. All calculations performed were implemented in the Python programming language.

The first phase starts with a nonlinear least squares fit on $\Delta E_{f^*}(\phi, \psi)$ given $\Delta E^*(G)$ and an initial parameter vector f . This phase is done to get a partial fit on the PES, i.e., to capture the general topology of the PES. The value for f is determined by assigning values for $A_{i,j}$ to be all one, and values for $u_{i,j}$ and $v_{i,j}$ to be all zero (i.e., initially no translation). The optimal parameter vector is referred to as f^* . In this study, the value for N_f is set to 6.

The next phase enumerates the minima over G with ΔE values of at most $\tau_g \max[\Delta E(G)]$, which are needed for the next phase, where τ_g is the threshold factor supplied by the user. The number of minima, as well as the (ϕ, ψ) and the corresponding ΔE values in the abovementioned list, referred to as N_g and g^* , respectively, will be used as in constructing the initial parameter vector for the next phase. In this paper, τ_g is chosen to be equal to 0.75.

The obtained parameter vectors f^* and g^* in the previous phases are used to perform an NLS fit on ΔE_{f^*+g} as the last phase, in which an initial parameter h is constructed as a vector concatenation of f^* and g^* . The fit will yield an optimal parameter vector h^* , which will be used to describe the topology of the PES being fitted from G .

3. Results and Discussion

In this section, the resulting Ramachandran-type PESs for the twenty amino acid diamides from the potential energy function described in Sec. 2.1 are presented, as well as the discussion on the geometries on the preferred amino acid conformations with respect to local energy minima geometries and energy ranges.

3.1. Topology of the Ramachandran-type PESs

This section analyzes the global topological properties of the Ramachandran-type PESs obtained from Sec. 2.1. Table 1 provides a summary of the goodness-of fit analysis on using the potential energy function ΔE_{f^*+g} for the 20 amino acid diamides considered in the study.

The overall fit procedure yields an average R^2 of 0.9845 and average RMSE of 2.36 kJ·mol⁻¹ for all 20 amino acids, with the isoleucine PES yielded the lowest RMSE of 1.95 kJ·mol⁻¹.

Note that the results obtained in this study, having an average RMSE of 2.05 kJ·mol⁻¹ for the amino acids glycine, alanine and valine, is superior to results in the previous work,⁷⁴ which yielded an average RMSE of 4.99 kJ·mol⁻¹. The detailed comparison on the results of the two studies is analyzed in Appendix A.

However, it is observed that there are some minima obtained from electronic structure calculations, but are not found by the proposed function, i.e., $\Delta E_{f^*+g}(\phi, \psi)$, are listed in Table 2. Note that the proposed mathematical model is assumed to return relative energy values with an error of around the RMSE value from the relative energies from quantum chemical calculations. Further checking from the fine grid cross-sections of the PESs of the relevant amino acids show that these are shallow minima with barrier heights of less than 1 kJ·mol⁻¹ that is less than the average RMSE in Sec. 3.1, and thus these are not included for comparison in the study.

Furthermore, the deviation of local relative energy minima and its corresponding geometries from the proposed model is also compared from geometry optimizations from GAUSSIAN09 calculations. The resulting correlation plot is shown in Fig. 2.

From the plots in Figure 2, this shows that the R^2 for both ϕ and ψ values is above 0.99, which suggest an almost perfect correlation between the geometries provided by the proposed function and from GAUSSIAN09 calculations. Moreover, the R^2 and maximum deviation for the minima energy

Published by NRC Research Press

Villar et al.

Table 1. Statistics on the 20 amino acid diamides, and its corresponding two-rotor potential energy function ΔE_{f^*+g} . Degrees of freedom (DF) is defined as the sum of the number of bond lengths, bond angles and dihedral angles of a given amino acid. Relative energies are in $\text{kJ}\cdot\text{mol}^{-1}$.

Amino Acid	DF	Parameter Size (N_f, N_g)	RMSE	R^2	Detected Minima
Gly	51	67 (6, 6)	1.7744	0.9890	5
Ala	60	81 (6, 8)	2.2714	0.9829	5
Val	78	95 (6, 10)	2.0918	0.9885	6
Cys	63	88 (6, 9)	2.2508	0.9858	6
Ser	63	74 (6, 7)	2.4351	0.9848	6
Sec	63	95 (6, 10)	2.1510	0.9877	6
Asp	69	74 (6, 7)	2.4692	0.9826	7
Asn	72	88 (6, 9)	2.8164	0.9818	5
Thr	72	74 (6, 7)	2.2935	0.9890	6
Phe	75	88 (6, 9)	2.6900	0.9799	5
His	75	88 (6, 9)	2.1711	0.9881	6
Tyr	81	81 (6, 8)	2.4538	0.9833	5
Ile	87	88 (6, 9)	1.9482	0.9899	6
Leu	87	88 (6, 9)	2.1000	0.9869	5
Trp	87	74 (6, 7)	2.2154	0.9867	7
Glu	78	81 (6, 8)	2.8722	0.9762	6
Met	81	81 (6, 8)	2.8069	0.9782	5
Gln	81	88 (6, 9)	3.3125	0.9741	6
Lys	93	81 (6, 8)	2.1285	0.9866	5
Arg	99	74 (6, 7)	2.0450	0.9878	6

Table 2. DFT-optimized minima structures found with low barrier height (i.e., around shallow well region). Barrier heights are in $\text{kJ}\cdot\text{mol}^{-1}$

Amino Acid	Conformation	ϕ	ψ	Barrier Height
Ala	α_D	67.92	26.68	1.0284
Ile	α_D	50.39	41.94	0.2626
Leu	α_D	64.80	35.03	0.7877
Asn	δ_D	-166.23	-38.55	0.5251
Ser	δ_D	179.85	-20.50	0.2625
Glu	α_D	63.78	38.70	0.7877

values (Figure 2b) are 0.9837 and not larger than $5.1 \text{ kJ}\cdot\text{mol}^{-1}$ from the DFT-optimized structures (shown in blue bands), respectively, which is highly acceptable. This is also supported by the slope and intercept of the respective regression lines, in which the values are close to the ideal values of 1.0 and 0.0, respectively.

In general, the results in Fig. 2 suggest that the two-rotor model ΔE_{f^*+g} precisely captures the overall topology of the amino acid diamide PESs, and may also be used to locate the local energy minima geometries through the model.

3.2. Conformational Preference of Amino Acid Diamides

One of the interests of this study to look onto the conformational preferences of a single amino acid in terms of the backbone dihedral angles ϕ and ψ , since this has an important role in the local

Can. J. Chem.

Fig. 2. Correlation plots of the potential energy function values against the DFT-optimized values for all 20 amino acid diamides: (a) minima backbone geometries, (b) minima relative energies

conformational preferences and determining secondary structures in proteins. While sidechains need to be considered as well, because they also affect the stability of amino acids, this property is not considered in the study and meanwhile, relaxing the sidechain dihedral angles (for more complex amino acids) in the electronic structure calculations compensate this limitation.

Note that in theory, the two backbone angles ϕ and ψ each have three possible configurations, which imply that by topological theory of periodic functions,^{84,85} it is expected to have $3 \times 3 = 9$ minima on the Ramachandran-type PES for each of the amino acid diamides considered in the study. However, due to steric hindrance and other sidechain structure restrictions, the number of energy minima found in a PES may be lower than nine.

For the succeeding discussions, the conformation regions^{74,86} illustrated through Fig. 3 pertains to the simplified conformational notations for the ranges of values for the ordered pair of dihedral angles ϕ and ψ for peptide binding units. This characterizes the molecular geometries of the energy minima and identifying the preferred regions in the Ramachandran-type PES.

Fig. 3. Conformational regions in the Ramachandran-type PES

Fig. 4 shows the Ramachandran-type PES obtained using the potential energy function ΔE_{f*+g} described in Sec. 2.1.

Published by NRC Research Press

Villar et al.

Fig. 4. Ramachandran-type PES and minimum energy backbone geometries of 20 amino acid diamides. Horizontal and vertical axes correspond to the ϕ and ψ angle values, respectively. Dark blue areas correspond to minima regions of the PES. Yellow and red circles correspond to the detected local and global minima of the PESs, respectively.

Published by NRC Research Press

Can. J. Chem.

For the energy minima geometries, Table B.3 (Appendix B) details the dihedral angle values of the local minima for the 20 amino acid diamides considered in the study. As stated in Sec. 2.1, only backbone angles ϕ and ψ are modeled in this study, and the sidechain dihedral angles χ_i , $i = 1, 2, 3, 4, 5$ are approximate values from the nearest conformation from the electronic structure calculations performed in Sec. 2.2.

To examine further, Fig. 5 shows the plot of the local minima geometries with respect to the conformation regions described in Fig. 3.

Fig. 5. Population plot of the local energy minima of the 20 amino acid diamides, with respective relative energy values in $\text{kJ}\cdot\text{mol}^{-1}$

For most amino acids, the backbone ϕ angle predominantly occur both on the far left region of the Ramachandran map around -170° to -120° and midright region between 50° and 80° . The far right region has low density, with a couple clusters in the range between -93° and -75° . With respect to the ψ axis, the map is populated with minima mostly in the central region between -78° and -20° , and between 20° and 78° . There is also a relatively dense range between 143° to 178° .

The amino acids also highly prefer γ_L , β_L , γ_D and δ_D conformations, as most of the amino acids prefer conformations within these regions in the PES, while there are no minima found in the ϵ_L conformation, which was also verified from DFT-optimized structure optimizations. Global minima also lie mostly on the γ_L and β_L regions, which are favored conformations for amino acids to form α -helices and β -sheets, respectively.

There is also low preference for the α_L conformation since only His, Leu, Trp and Gln diamides have minima in this region. From the plot and Table B.3, it can also be deduced that except for some outlier points, most of the minima in every region have similar relative energy values. This confirms that backbone interactions primarily influence the conformational energy landscape of amino acids.

A more extensive study by Yuan et al.,⁶⁷ in which all stable amino acid molecular geometries are searched such that there are only small differences between two conformations with respect to all torsion angles, including sidechain dihedral angles. The authors carried out the calculations in three levels of theory, such as HF/6-31G(d,p), B3LYP/apc-1 and MP2/cc-pVDZ. From their methodology, it is possible to obtain more than one minima within one conformational region since the heteroatom-containing sidechain dihedral angles were also varied. While the present study produced similar results, in terms of energy minima and corresponding geometries, the rotation of the sidechain dihedral angles

Published by NRC Research Press

Villar et al.

was not performed. Moreover, the DFT-optimized conformations either remained similar to the initial anti orientation or moved to the nearest gauche orientation.

It is also interesting to confirm from the results that the L conformers are more stable than the D conformers, since γ_L values have an average relative energy value of 3.85 kJ·mol⁻¹, as compared to γ_D with 16.35 kJ·mol⁻¹. Similarly, α_L and α_D have average relative energy values of 25.09 kJ·mol⁻¹ and 30.18 kJ·mol⁻¹, respectively, and δ_L and δ_D having average relative energy values of 14.42 kJ·mol⁻¹ and 33.99 kJ·mol⁻¹, respectively.

3.3. Magnitude of Dispersion Forces in Selected Amino Acid Diamides

Non-covalent interactions give major contributions to the stability of various macromolecules, including proteins. Among these interactions, dispersion (also called London dispersion) is the most difficult to describe. Dispersion interactions originate from correlated motion of electrons on two different atoms. These type of interactions, therefore, require explicitly correlated methods.

The B3LYP functional, which is a DFT method, is a widely used hybrid functional in quantum chemistry because of its high accuracy with less computational resources required.^{87,88} However, in past twenty years, the DFT methods were found to fail in describing the molecular dispersion interactions, since these methods does not correctly include dispersion, and often underestimate the stabilization from dispersion interactions.^{89–92}

In this section, the effects of dispersion forces in selected amino acids, such as valine, serine, phenylalanine and aspartic acid, is studied. The amino acids are chosen due to the presence of intramolecular interactions, such as hydrogen bonding. The chosen amino acid optimized β_L and γ_L conformers obtained the B3LYP/6-31G(d) level of theory is compared to optimized structures using the correlated second-order Møller-Plesset (MP2) method⁹³ with the Dunning correlation consistent basis set aug-cc-pVDZ.⁹⁴ The energies and the differences between respective β_L and γ_L conformers are detailed in Table 3.

Table 3. Relative energy differences between selected amino acid energy minima geometries, with respect to B3LYP/6-31G(d) and MP2/aug-cc-pVDZ level of theories. ΔE_B and ΔE_M are relative energies between the γ_L and β_L conformations, under the B3LYP/6-31G(d) and MP2/aug-cc-pVDZ level of theories, respectively.

Amino Acid	B3LYP/6-31G(d) Energies			MP2/aug-cc-pVDZ Energies			$\Delta(\Delta E_{M-B})$ (kJ·mol ⁻¹)
	E(γ_L) (hartrees)	E(β_L) (hartrees)	ΔE_B (kJ·mol ⁻¹)	E(γ_L) (hartrees)	E(β_L) (hartrees)	ΔE_M (kJ·mol ⁻¹)	
Val	-574.48010	-574.47936	1.94	-572.91121	-572.91035	2.25	0.31
Ser	-571.06494	-571.06420	1.94	-569.59917	-569.59813	2.24	0.30
Phe	-726.90485	-726.90497	-0.33	-724.90834	-724.90725	2.86	3.19
Asp	-684.41787	-684.41850	-1.65	-682.69321	-682.69284	0.97	2.62

In summary, the results show that there is no significant difference in the relative energies, with just at most 3.19 kJ·mol⁻¹ difference between the results from the two functionals. This value is considered small, since it is less than 4.1 kJ·mol⁻¹, the acceptable threshold for chemical accuracy.

Note also that in phenylalanine, from the number of interactions found for all stable conformers at the B3LYP/6-31G(d) level of theory involving the aromatic ring and the peptide bond, a previous study⁹⁵ observed that the p-OH substituent plays no role in the geometric constraints for these types of dispersion forces.

Published by NRC Research Press

3.4. Effects of Sidechain Orientation on the Energy Minima Geometry

The influence of the sidechain orientation on the geometry of the global minimum for each of the 20 amino acid diamides is explored in this section, because the size and geometry of sidechains mainly affect the stabilization of amino acids. It is interesting to compare the energy barrier between the global minimum and the other local minima for each of the amino acids. The energy range, which is defined as the energy difference between the global minimum and the (local) minimum with the highest energy,⁶⁷ is shown in Fig. 6.

Fig. 6. Energy range of the 20 amino acid diamides between the global minimum and the local minimum with the highest energy. Amino acids were ranked according to increasing relative energy values.

Aspartic acid has the largest energy range, while glycine has the smallest energy range. The former is suggested to be flexible due to an -OH in the sidechain, which can form hydrogen bonds between the sidechain and the backbone, while the latter is somehow restricted as there are only backbone interactions. The results above were also compared to a previous study⁶⁷ which analyzed the energy ranges under HF/6-31G(d,p), B3LYP/apc-1 and MP2/cc-pVDZ levels, and observed that the obtained energy ranges vary, which thus locally alter the ordering of the amino acids. It can also be seen from Fig. 6 that most of the amino acids have very small differences in the energy ranges that sidechain interactions are not taken into account in the generation of the PES. However, the study assumed that the sidechain dihedral angles to be fully relaxed, i.e., sidechains are freely moving, in the quantum chemical calculations, which might hint that the sidechain geometry also has a significant influence into the overall potential energy landscape.

4. Conclusion and Future Directions

Analyzing the conformational preferences of the individual amino acids in terms of its backbone dihedral angles ϕ and ψ is a primary step in understanding how secondary structure is determined based on the amino acid sequence. A full conformational energy landscape is important to observe this preferences, but constructing the associated PES requires significant computational resources that a simple but precise approximation is needed.

This study presents an two-rotor mathematical model that represents the amino acid conformational PESs with respect to backbone dihedral angles ϕ and ψ , through a linear combination of a Fourier series and mixture of Gaussian functions. Results show that it precisely captures the global topology of the Ramachandran-type PESs, as it yielded only an average RMSE of 2.36 kJ·mol⁻¹ and R^2 of 0.9845.

Villar et al.

Compared to the previous model,⁷⁴ the improved function returned an average RMSE of 2.05 kJ·mol⁻¹ for the same group of amino acids, which is less than half the average RMSE of the former model, or 4.99 kJ·mol⁻¹. Moreover, in terms of detecting local energy minima geometries, it returned an almost perfect correlation of 0.9946 and 0.9928 R² for the backbone dihedral angles between the minima obtained from the proposed function and from the electronic structure calculations. This suggests that the proposed model can also be used to reliably locate the local minima of the PES from a coarse grid of input energies.

Furthermore, it has been observed that the γ_L , β_L , γ_D and δ_D conformations are highly favored because amino acid local minima are predominantly located in these regions in the PES, while there are no minima found in the ϵ_L conformation, which is also verified from DFT-optimized structures. The model also confirms the preference of L-type over the D-type conformations, as shown from relative energies of obtained local minima conformations.

The influence of dispersion forces and sidechain orientation to the topology of the PES were also explored. While the energy differences with respect to the level of theory is significantly small, at least for the two lowest minima, the chosen level of theory affects the PES landscape, which alters the ordering of the amino acids with respect to the energy range.

For future work, as the study used relative energies with associated conformations with relaxed sidechains, it is suggested to also include sidechain effects on the extension of the mathematical model. The proposed model is also constructed invariant of the level of theory used, but it is suggested to test the model to PES generated with higher level of theory and with solvent models, to reflect more realistic folding environment and dynamics.

5. Acknowledgements

JJV is supported by Philippine Council for Industry, Energy and Emerging Technologies Research and Development through the Engineering Research and Development for Technology HRDP and “Bridging the Human Resource Competency Gaps in Support of the National R&D Agenda” Program. Part of this research was also carried out using the HPC facilities under the CoARE Project, DOST Advanced Science and Technology Institute, Diliman, Quezon City, Philippines, and of the Department of Chemical Informatics, University of Szeged, Hungary.

AR acknowledges the National Talent Programme (Project ID: NTP-NFTÖ-17-B-0352) for the financial support.

This research is also supported by the European Union and the Hungarian State, co-financed by the European Regional Development Fund through GINOP-2.2.1-15-216-00005 project.

The authors also thank Dr. Milan Szőri for the helpful discussions.

References

- (1) Reece, J. B.; Urry, L. A.; Cain, M. L.; Wasserman, S. A.; Minorsky, P. V.; Jackson, R. B., *Campbell Biology*, 10th; Pearson: Boston, MA, 2014.
- (2) Anfisen, C. B., *The Molecular Basis of Evolution*; Wiley: New York, NY, 1959.
- (3) Gierasch, L. M.; King, J., *Protein Folding, Deciphering the Second Half of the Genetic Code*; American Association for the Advancement of Science: Washington, DC, 1990.
- (4) Garrett, R. H.; Grisham, C. M., *Biochemistry*, 5th; Cengage Learning: Boston, MA, 2014.
- (5) Hardin, C.; Pogorelov, T. V.; Luthey-Schulten, Z. *Current Opinion in Structural Biology* **2002**, 12, 176–181.
- (6) Lewars, E. G., *Computational Chemistry—Introduction to the Theory and Applications of Molecular and Quantum Mechanics*, 3rd; Springer International Publishing: 2016.

Can. J. Chem.

- (7) Jákli, I.; Perczel, A.; Viskolcz, B.; Csizmadia, I. G. In *Protein Modelling*, Náray-Szabó, G., Ed.; Springer International Publishing: Cham, 2014, pp 5–50.
- (8) Radom, L.; Hehre, W. J.; Pople, J. A. *J Am Chem Soc* **1972**, *94*, 2371–2381.
- (9) Radom, L.; Lathan, W.; Hehre, W.; Pople, J. A. *J Am Chem Soc* **1973**, *95*, 693–698.
- (10) Demare, G. R.; Peterson, M. R.; Csizmadia, I. G.; Strausz, O. P. *J Comp Chem* **1980**, *1*, 141–148.
- (11) Peterson, R.; Mare, R. D. E.; Roosevelt, A. F. D. *J Mol Struct (THEOCHEM)* **1981**, *86*, 131–147.
- (12) Kehoe, T. A. K.; Peterson, M. R.; Chass, G. A.; Viskolcz, B.; Stachod, L.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2003**, *666-667*, 79–87.
- (13) Ragyanszki, A.; Suranyi, A.; Csizmadia, I. G.; Kelemen, A.; Jensen, S. J. K.; Uysal, S. Y.; Viskolcz, B. *Chem Phys Lett* **2014**, *599*, 169–174.
- (14) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. *J Mol Biol* **1963**, *7*, 95–99.
- (15) Tasi, G.; Nagy, B.; Matisz, G.; Tasi, T. S. *Comp. Theor. Chem.* **2011**, *963*, 378–383.
- (16) Creighton, T. E., *Proteins, Structures and Molecular Properties*, 2nd; W.H. Freeman and Co.: New York, 1993.
- (17) McAllister, M. A.; Perczel, A.; Csaszar, P.; Viviani, W.; Rivail, J. L.; Csizmadia, I. G. *J. Mol Struct* **1993**, *288*, 161–180.
- (18) Endredi, G.; Liegner, C.-M.; McAllister, M.; Perczel, A.; Ladik, J.; Csizmadia, I. *J. Mol Struct* **1994**, *306*, 1–7.
- (19) Rodriguez, A.; Baldoni, H. A.; Suvire, F.; Nieto-Vasquez, R.; Zamarbide, G.; D.Enriz, R.; Farkas, O.; Perczel, A.; Csizmadia, I. *J Mol Struct (THEOCHEM)* **1998**, *455*, 275–302.
- (20) Stepanian, S. G.; Reva, I. D.; Radchenko, E. D.; Adamowicz, L. *J. Phys. Chem. A* **1998**, *102*, 4623.
- (21) Viviani, W.; Rivail, J.-L.; Perczel, A.; Csizmadia, I. G. *J. Am. Chem. Soc.* **1993**, *115*, 8321.
- (22) Sapse, A.-M.; Mezei, M.; Jain, D. C.; Unson, C. *J Mol Struct (THEOCHEM)* **1994**, *306*, 225.
- (23) Farkas, O.; Perczel, A.; Marcoccia, J. F.; Hollosi, M.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **1995**, *331*, 27.
- (24) Perczel, A.; Farkas, O.; Csizmadia, I. G. *J. Am. Chem. Soc.* **1996**, *118*, 7809.
- (25) Jakli, I.; Perczel, A.; Farkas, O.; Csaszar, A. G.; Sosa, C.; Csizmadia, I. G. *J. Comput. Chem.* **2000**, *21*, 626.
- (26) Farkas, O.; McAllister, M.; Ma, J. H.; Perczel, A.; Hollosi, M.; Csizmadia, I. *J Mol Struct (THEOCHEM)* **1996**, *369*, 105–114.
- (27) Jakli, I.; Perczel, A.; Farkas, O.; Hollosi, M.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **1998**, *455*, 303.
- (28) Chass, G. A.; Mirasol, R. S.; Setiadi, D. H.; Tang, T. H.; Chin, W.; Mons, M.; Dimicoli, I.; Dognon, J. P.; Viskolcz, B.; Lovas, S.; Penke, B.; Csizmadia, I. G. *J. Phys. Chem. A* **2005**, *109*, 5289–5302.
- (29) Von Helden, G.; Compagnon, I.; Blom, M. N.; Frankowski, M.; Erlekam, U.; Oomens, J.; Brauer, B.; Gerber, R. B.; Meijer, G. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1248.
- (30) Baldoni, H. A.; Rodriguez, A. M.; Zamora, M. A.; Zamarbide, G. N.; Enriz, R. D.; Farkas, O.; Csaszar, P.; Torday, L. L.; Sosa, C. P.; Jakli, I.; Perczel, A.; Papp, J. G.; Hollosi, M.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **1999**, *465*, 79.

Published by NRC Research Press

Villar et al.

- (31) Lam, J. S. W.; Koo, J. C. P.; Hudaky, I.; Varro, A.; Papp, J.; Penke, B.; Csizmadia, I. *J Mol Struct (THEOCHEM)* **2003**, 666-667, 285-289.
- (32) Sahai, M. A.; Kehoe, T. A. K.; Koo, J. C. P.; Setiadi, D. H.; Chass, G. A.; Viskolcz, B.; Penke, B.; Pai, E. F.; Csizmadia, I. G. *J. Phys. Chem. A* **2005**, 109, 2660-2679.
- (33) Huang, Z.; Lin, Z. *J. Phys. Chem. A* **2005**, 109, 2656.
- (34) Kaczor, A.; Reva, I. D.; Proniewicz, L. M.; Fausto, R. *J. Phys. Chem. A* **2007**, 111, 2957.
- (35) Berg, M. A.; Salpietro, S. J.; Perczel, A.; Farkas, O.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2000**, 504, 127-139.
- (36) Koo, J. C. P.; Chass, G. A.; Perczel, A.; Farkas, O.; Varro, A.; Torday, L. L.; Papp, J. G.; Csizmadia, I. G. *The Eur. Phys. J. D* **2002**, 20, 499-511.
- (37) Rassolian, M.; Chass, G. A.; Setiadi, D. H.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2003**, 666-667, 273-278.
- (38) Koo, J. C. P.; Lam, J. S. W.; Chass, G. A.; Setiadi, D. H.; Law, J. M. S.; Papp, J. G.; Penke, B.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2003**, 666-667, 279-284.
- (39) Gerlei, K. Z.; Jakli, I.; Szori, M.; Jensen, S. J. K.; Viskolcz, B.; Csizmadia, I. G.; Perczel, A. *J. Phys. Chem. B* **2013**, 117, 12402-12409.
- (40) Baldoni, H. A.; Zamarbide, G. N.; Enriz, R. D.; Jauregui, E. A.; Farkas, O.; Perczel, A.; Salpietro, S. J.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2000**, 500, 97.
- (41) Zamora, M. A.; Baldoni, H. A.; Bombasaro, J. A.; Mak, M. L.; Perczel, A.; Farkas, O.; Enriz, R. D. *J Mol Struct (THEOCHEM)* **2001**, 540, 271.
- (42) Sheraly, A. R.; Chass, G. A.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2003**, 666-667, 243-249.
- (43) Dobrowolski, J. C.; Jamroz, M. H.; Koos, R.; Rode, J. E.; Sadlej, J. *Chem Phys Chem* **2007**, 8, 1085.
- (44) Boeckx, B.; Ramaekers, R.; Maes, G. *J. Mol. Spectrosc.* **2010**, 261, 73.
- (45) Tarditi, A. M.; Klipfel, M. W.; Rodriguez, A. M.; Suvire, F. D.; Chasse, G. A.; Farkas, O.; Perczel, A.; Enriz, R. D. *J Mol Struct (THEOCHEM)* **2001**, 545, 29.
- (46) Masman, M. F.; Zamora, M. A.; Rodriguez, A. M.; Fidanza, N. G.; Peruchena, N.; Enriz, R. D.; Csizmadia, I. G. *The Eur. Phys. J. D* **2002**, 20, 531-542.
- (47) Klipfel, M. W.; Zamora, M. A.; Rodriguez, A. M.; Fidanza, N. G.; Enriz, R. D.; Csizmadia, I. G. *J. Phys. Chem A* **2003**, 107, 5079-5091.
- (48) Barroso, M. N.; Cerutti, E. S.; Rodriguez, A. M.; Jauregui, E. A.; Farkas, O.; Perczel, A.; Enriz, R. D. *J Mol Struct (THEOCHEM)* **2001**, 548, 21.
- (49) Koo, J. C. P.; Chass, G. A.; Perczel, A.; Farkas, O.; Torday, L. L.; Varro, A.; Papp, J. G.; Csizmadia, I. G. *J. Phys. Chem. A* **2002**, 106, 6999.
- (50) Hudaky, P.; Beke, T.; Perczel, A. *J Mol Struct (THEOCHEM)* **2002**, 583, 117.
- (51) Ramaekers, R.; Pajak, J.; Rospenk, M.; Maes, G. *Spectrochim. Acta, Part A: Mol. Biomol. Spectrosc.* **2005**, 61, 1347.
- (52) Sahai, M. A.; Motiwala, S. S.; Chass, G. A.; Pai, E. F.; Penke, B.; Csizmadia, I. G. *J Mol Struct (THEOCHEM)* **2003**, 666-667, 251-267.
- (53) Sahai, M. A.; Fejer, S. N.; Viskolcz, B.; Pai, E. F.; Csizmadia, I. G. *J. Phys. Chem. A* **2006**, 110, 11527.

Can. J. Chem.

- (54) Lang, A.; Csizmadia, I. G.; Perczel, A. *Proteins: Struct., Funct., Bioinf.* **2005**, 58, 571.
- (55) Ling, S.; Yu, W.; Huang, Z.; Lin, Z.; Haranczyk, M.; Gutowski, M. *J. Phys. Chem. A* **2006**, 110, 12282.
- (56) Boeckx, B.; Maes, G. *J. Phys. Chem. B* **2012**, 116, 12441.
- (57) Stepanian, S. G.; Y., I. A.; Adamowicz, L. *Chem. Phys.* **2013**, 423, 20.
- (58) Vank, J. C.; Sosa, C. P.; Perczel, A.; Csizmadia, I. *Can. J. Chem.* **2000**, 78, 395–408.
- (59) Meng, L.; Wu, W.; Zhang, J. *J. Phys. Chem. A* **2014**, 118, 1684.
- (60) Perczel, A.; Angyan, J. G.; Kajtar, M.; Viviani, W.; Rivail, J.; Marcoccia, J. F.; Csizmadia, I. *J. Am. Chem. Soc.* **1991**, 113, 6256–6265.
- (61) Viviani, W.; Rivail, J.-L.; Csizmadia, I. G. *Theoretica Chimica Acta* **1992**, 85, 189–197.
- (62) Gronert, S.; O’Hair, R. A. J. *J. Am. Chem. Soc.* **1995**, 117, 20712081.
- (63) Smolyar, A.; Wong, C. F. *J Mol Struct (THEOCHEM)* **1999**, 488, 51.
- (64) Lang, A.; Gyorgy, K.; Csizmadia, I. G.; Perczel, A. *J Mol Struct (THEOCHEM)* **2003**, 666–667, 251–267.
- (65) Lakard, B. *J Mol Struct (THEOCHEM)* **2004**, 681, 183.
- (66) Hernandez, B.; Pflueger, F.; Nsangou, M.; Ghomi, M. *J. Phys. Chem. B* **2009**, 113, 3169.
- (67) Yuan, Y.; Mills, M. J. L.; Popelier, P. L. A.; Jensen, F. *The Journal of Physical Chemistry A* **2014**, 118, 7876–7891.
- (68) Murrell, J. N.; Carter, S.; Farantos, S. C.; Huxley, P.; Varanda, A. J. C., *Molecular Potential Energy Functions*; Wiley: New York, 1984.
- (69) Schatz, G. C. *Rev. Mod. Phys.* **1989**, 61, 669.
- (70) Sorbie, K.; Murrell, J. N. *Mol Phys* **1976**, 31.
- (71) Varandas, A. J. C. *Int Rev Phys Chem* **2000**, 19.
- (72) Braams, B. J.; Bowman, J. M. *Int Rev Phys Chem* **2009**, 28.
- (73) Bowman, J. M.; Czako, G.; Fu, B. *Phys Chem Chem Phys* **2011**, 13.
- (74) Villar, J. J. S.; Valdez, A. R. L.; Setiadi, D. H.; Fiser, B.; Viskolcz, B.; Csizmadia, I. G.; Ragyanszki, A. *Can J Chem* **2017**, 95, 830–836.
- (75) Schrödinger, E. *Phys. Rev.* **1926**, 28, 1049–1070.
- (76) Born, M.; Oppenheimer, R. *Annalen der Physik* **1927**, 389, 457–484.
- (77) Levenberg, K. *Quart. Appl. Math.* **1944**, 2, 164–168.
- (78) Marquardt, D. *SIAM J. Appl. Math.* **1963**, 11, 431–441.
- (79) Becke, A. D. *J. Chem. Phys.* **1993**, 98, 5648–52.
- (80) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Phys. Chem.* **1994**, 98, 11623–11627.
- (81) Hehre, W. J.; Ditchfield, R.; Pople, J. A. *J. Chem. Phys.* **1972**, 56, 2257–2261.
- (82) Frisch, M. J.; et al. Gaussian09 Revision A.02., 2009.
- (83) Chass, G. A.; Sahai, M. A.; Law, J. M. S.; Lovas, S.; Farkas, O.; Perczel, A.; Rivail, J.-L.; Csizmadia, I. G. *Int. J. Quant. Chem.* **2002**, 90, 933–968.
- (84) Mezey, P. G. *Chemical Physics Letters* **1981**, 82, 100–104.

Published by NRC Research Press

Villar et al.

- (85) Csizmadia, I. *Journal of Molecular Structure: THEOCHEM* **1986**, *138*, 1–11.
- (86) Perczel, A.; Angyan, J. G.; Kajtar, M.; Viviani, W.; Rivail, J.-L.; Marcoccia, J.-F.; Csizmadia, I. G. *J. Am. Chem. Soc.* **1991**, *113*, 6256.
- (87) Gresh, N.; Kafafi, S. A.; Truchon, J. F.; R., S. D. *J. Comput. Chem.* **2004**, *25*, 823–834.
- (88) Jurecka, P.; Cerny, J.; Hobza, P.; R., S. D. *J. Comput. Chem.* **2007**, *28*, 555–569.
- (89) Van Mourik, T.; Gdanitz, R. J. *J. Chem. Phys.* **2002**, *116*, 9620–9623.
- (90) Morgado, C.; A., V. M.; Hillier, I. H.; Shan, X. *Phys. Chem. Chem. Phys.* **2007**, *9*, 448–451.
- (91) Von Lilienfeld, O. A.; Tavernelli, I.; Rothlisberger, U.; Sebastiani, D. *Phys. Rev. Lett.* **2004**, *93*, 153004–153007.
- (92) Du, Q.-S.; Liu, P.-J.; Deng, J. *J. Chem. Theory Comput.* **2007**, *3*, 1665–1672.
- (93) Møller, C.; Plesset, M. S. *Phys. Rev.* **1934**, *46*, 618–622.
- (94) Dunning, T. H. J. *The Journal of Chemical Physics* **1989**, *90*, 1007–1023.
- (95) Chass, G.; Lovas, S.; Murphy, R.; Csizmadia, I. *Eur. Phys. J. D* **2002**, *20*, 481–497.
- (96) Case, D. A. et al. AMBER 12., 2012.
- (97) Salomon-Ferrer, R.; Case, D. A.; Walker, R. C. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 198–210.
- (98) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725.

Appendix A: Comparison Between Two Potential Energy Functions

In the previous work,⁷⁴ a two-rotor mathematical function to represent amino acid diamide backbone PES, referred in this paper as $\Delta E_{f+g}(\phi, \psi)$, has been presented, and is defined as a linear combination of two functions given by the following, where $(\phi, \psi) \in [-180^\circ, 180^\circ]^2$:

- a Fourier series of the form

$$\Delta E_f(\phi, \psi) = k + \sum_{i=1}^{N_f} (w_i \cos i\omega\phi + x_i \cos i\omega\psi + y_i \sin i\omega\phi + z_i \sin i\omega\psi), \quad (\text{A.1})$$

where N_f be the multiplicity (i.e., number of the terms), k is the constant term in the series, ω is the conversion factor from degrees to radians, i.e., $\omega = \frac{\pi}{180}$, and $w_i, x_i, y_i, z_i \in \mathbb{R}$, with $i = 1, \dots, N_f$, and

- a mixture of spherical Gaussian functions, which are fitted at the recognizable peaks of the input grid, of the form

$$\Delta E_g(\phi, \psi) = \sum_{j=1}^{N_g} A_j \exp \left[-\frac{c_{\phi_j}(\phi - \bar{\phi}_j)^2}{2\sigma_{\phi_j}^2} - \frac{c_{\psi_j}(\psi - \bar{\psi}_j)^2}{2\sigma_{\psi_j}^2} \right], \quad (\text{A.2})$$

where N_g be the multiplicity, A_j be the amplitude, $(\bar{\phi}_j, \bar{\psi}_j)$ and $\sigma_{\phi_j}^2, \sigma_{\psi_j}^2$ be the center and axes of the ellipsoids, respectively, all with $j = 1, \dots, N_g$. Also, $\exp(\cdot)$ is the exponential function.

The full model is defined as

$$\Delta E_{f+g}(\phi, \psi) = \Delta E_f(\phi, \psi) + \Delta E_g(\phi, \psi). \quad (\text{A.3})$$

Note that the Fourier series used in⁷⁴ and the function detailed in Sec. 2.1 is different, as the former assumed independence of two variables, and assumed that the interactions between these variables are captured by the Gaussian mixtures. The function presented in this paper also includes the interaction between ϕ and ψ in the Fourier series submodel.

To analyze the goodness-of-fit of the two models, the RMSE and R^2 are evaluated for both functions, with respect to the energy data from Sec. 2.2, over $G = [-180^\circ, 180^\circ]^2$. Table A.1 shows the summary of the global topology error analysis for the two models.

Table A.1. Goodness-of-fit comparison between ΔE_{f+g} and ΔE_{f^*+g}

Amino Acid	ΔE_{f+g} Fit Results			ΔE_{f^*+g} Fit Results		
	No. Parameters	RMSE	R^2	No. Parameters	RMSE	R^2
Glycine	76	4.8363	0.9190	67	1.7744	0.9890
Alanine	87	4.9544	0.9186	81	2.2714	0.9829
Valine	102	5.1930	0.9280	95	2.0918	0.9885

As reflected in the table, the average RMSE of the proposed model is $2.05 \text{ kJ}\cdot\text{mol}^{-1}$, which is less than half the error obtained from the previous study, or $4.99 \text{ kJ}\cdot\text{mol}^{-1}$. Similarly, the R^2 obtained using the proposed function is higher than the values obtained on the previous paper. This result imply that

Villar et al.

the proposed model yields more precise relative energy values than the model presented in the previous work.

This is supported by the correlation of the energy values from the two models with respect to energy data obtained from Sec. 2.2. These are also compared with the results from the force field calculations on the previous study.⁷⁴ For the force field data, the molecular mechanics minimizations were performed with the sander module of the AMBER12 software suite.^{96,97} Constraints were used to keep the dihedral angles of the peptides at the desired values $[-180^\circ, 180^\circ]^2$, while structures were defined using the ff99SB force field parameter set.⁹⁸ The plots comparing the results of the three methods are shown in Fig. A.1.

Fig. A.1. Correlation plots of the potential energy function relative energy values (vertical axis) against the relaxed scan relative energy values (horizontal axis): (a) Glycine, (b) Alanine, (c) Valine. The three columns pertain to results from ΔE_{f+g} (previous result⁷⁴), ΔE_{f^*+g} (proposed) and ff99SB force field,⁷⁴ respectively. R^2 values are indicated in the corresponding scatter plots.

It can be inferred from the results above that the additional terms involving products of the cosines of ϕ and ψ in eq. 2.1 have a positive influence on the precision of the predicted (relative) energy values, as shown in Figure A.1.

As pointed out in the main text, the number of parameters in eq. A.1 and eq. 2.1 can be expressed as a function of the number of terms, i.e. N_f . Now, define f to be this function with respect to the number of terms N_f . Then, eq. A.1 has $f(N_f) = 4N_f$, while eq. 2.1 has $f(N_f) = 0.5(N_f^2 + 7N_f + 2)$. From these functions, it is apparent that eq. 2.1 requires more parameters to fit the PES for all $N_f \geq 1$, but the new Fourier function, with the mixture of Gaussian functions, improves the fit to decrease the fit

Can. J. Chem.

RMSE to around half the RMSE of the fit from using eq. A.1, as seen in Table A.1.

The backbone dihedral angles of the minima obtained from the two functions are also compared against the DFT-optimized structures to check the deviations between the relative energies obtained from the three methods. The summary is presented in Table A.2.

Table A.2. Comparison of potential energy function values from analytical fitting results against DFT (B3LYP/6-31G(d)) optimized structures. Dihedral angle values are in degrees and ΔE values are in $\text{kJ}\cdot\text{mol}^{-1}$.

Amino Acid	Unique Minima	DFT Results			ΔE_{f+g} Results		ΔE_{f^*+g} Results	
		ϕ	ψ	ΔE_{dft}	ΔE_{old}	$\Delta\Delta E_{old-dft}$	ΔE_{new}	$\Delta\Delta E_{new-dft}$
Gly	γ_L	-82.15	68.54	0.00	0.00	0.00	0.53	0.53
	γ_D	82.15	-68.54	0.00	0.00	0.00	3.36	3.36
	β_L	179.99	180.00	2.58	2.01	-0.57	5.68	3.10
	δ_L	-122.78	21.93	10.24	14.36	4.12	10.28	0.04
	δ_D	122.78	-21.93	10.24	14.36	4.12	11.92	1.68
Ala	γ_L	-82.87	72.60	0.00	3.37	3.37	0.45	0.00
	β_L	-158.33	163.41	5.97	3.30	-2.67	3.58	-2.39
	γ_D	73.52	-57.39	10.88	11.06	0.18	12.15	1.27
	δ_L	-126.55	21.42	13.04	11.35	-1.69	13.73	0.69
	α_D	67.92	26.68	24.21	27.63	3.42	27.77	3.56
	δ_D	-165.06	-30.94	29.28	27.25	-2.03	31.36	2.08
Val	γ_L	-83.49	78.96	0.00	3.33	3.33	0.88	0.88
	β_L	-132.08	159.92	5.11	8.49	3.38	4.18	-0.93
	γ_D	73.51	-60.81	9.10	12.37	3.27	10.18	1.08
	δ_L	-122.94	18.74	10.99	13.87	2.88	13.93	2.94
	ϵ_L	-122.84	18.70	10.99	13.85	2.86	13.94	2.95
	α_L	-86.81	-22.80	22.46	20.05	-2.41	16.95	-5.51
	α_D	62.93	39.08	29.08	30.28	1.20	32.35	3.27
	δ_D	-160.47	-36.22	29.76	27.09	-2.67	29.24	-0.52
	ϵ_D	77.03	142.69	40.48	43.91	3.43	45.82	5.34

From Table A.2, the mean absolute deviation of relative energy values from the proposed model with respect to the DFT-optimized energies is $2.11 \text{ kJ}\cdot\text{mol}^{-1}$, which is lower than $2.38 \text{ kJ}\cdot\text{mol}^{-1}$ from the values in the previous study.

Overall, results show that the proposed model also captures both the global and local topologies of the PESs of the three amino acid diamides, in terms of RMSE, R^2 and deviations of the local energy minima geometries from the results from DFT calculations.

Villar et al.

Appendix B: Local Energy Minimum Molecular Geometries for the 20 Amino Acid Diamides

Table B.3 provides the dihedral angle values for the 20 amino acid diamides, with backbone dihedral angle ϕ and ψ values are obtained from the potential energy function ΔE_{f^*+g} in Sec. 2.1 while sidechain dihedral angle $\chi_i, i = 1, 2, 3, 4, 5$ values are approximate values from the nearest grid geometry from GAUSSIAN09 calculations.

Table B.3. Local energy minimum geometries for the 20 amino acid diamides. Dihedral angle values are in degrees, and ΔE values are in $\text{kJ}\cdot\text{mol}^{-1}$.

Conformation	ϕ	ψ	χ_1	χ_2	χ_3	χ_4	χ_5	E	ΔE
Glycine									
γ_L	-79.76	69.35	—	—	—	—	—	-456.53830	0.0000
γ_D	80.46	-72.42	—	—	—	—	—	-456.53774	1.4607
β_L	-166.92	175.52	—	—	—	—	—	-456.53670	4.1981
δ_L	-132.20	22.64	—	—	—	—	—	-456.53387	11.6456
δ_D	138.67	-28.60	—	—	—	—	—	-456.53351	12.5797
Alanine									
γ_L	-77.82	69.58	—	—	—	—	—	-495.85437	0.0000
β_L	-159.24	163.58	—	—	—	—	—	-495.85301	3.5663
γ_D	72.84	-59.02	—	—	—	—	—	-495.84976	12.1067
δ_L	-115.89	13.38	—	—	—	—	—	-495.84943	12.9689
δ_D	-173.60	-43.35	—	—	—	—	—	-495.84259	30.9207
Valine									
γ_L	-83.34	69.23	-146.62	—	—	—	—	-574.48038	0.0000
β_L	-156.45	149.18	-169.21	—	—	—	—	-574.47956	2.1337
δ_L	-141.08	20.68	-159.47	—	—	—	—	-574.47570	12.2721
γ_D	63.69	-46.15	-155.92	—	—	—	—	-574.47407	16.5573
α_D	57.85	28.48	-153.32	—	—	—	—	-574.46886	30.2410
ϵ_D	76.60	159.99	-130.09	—	—	—	—	-574.46371	43.7527
Cysteine									
γ_L	-82.33	76.89	-170.01	-70.82	—	—	—	-894.03857	0.0000
β_L	-157.41	-175.55	-162.55	69.70	—	—	—	-894.03792	1.7050
γ_D	69.39	-64.94	-178.51	57.33	—	—	—	-894.03149	18.5833
α_D	65.52	32.89	-154.47	-64.79	—	—	—	-894.02979	23.0484
ϵ_D	62.94	-170.29	-158.90	58.95	—	—	—	-894.02841	26.6600
δ_D	-159.60	-78.47	178.05	65.13	—	—	—	-894.02502	35.5730
Serine									
γ_L	-78.32	65.71	-178.14	-63.40	—	—	—	-571.06556	0.0000
β_L	-170.93	-175.46	-169.81	176.97	—	—	—	-571.06239	8.3253
δ_L	-143.34	29.81	-172.50	-56.98	—	—	—	-571.06107	11.7847
α_D	60.76	30.91	-165.97	-60.12	—	—	—	-571.05773	20.5563
γ_D	59.04	-12.10	-160.93	-39.46	—	—	—	-571.05686	22.8248
ϵ_D	65.76	-166.54	-161.27	-172.78	—	—	—	-571.05558	26.2060
Selenocysteine									
β_L	-162.94	171.19	165.36	67.24	—	—	—	-2895.23046	0.0000
γ_L	-80.40	74.86	-173.97	96.31	—	—	—	-2895.22542	13.2475

(continued on next page)

Published by NRC Research Press

Can. J. Chem.

Conformation	ϕ	ψ	χ_1	χ_2	χ_3	χ_4	χ_5	E	ΔE
γ_D	70.13	-68.09	179.64	68.12	—	—	—	-2895.22172	22.9402
ϵ_D	59.18	-163.37	-162.25	50.39	—	—	—	-2895.21996	27.5770
α_D	55.96	48.57	-166.36	113.49	—	—	—	-2895.21622	37.3804
δ_D	178.73	-55.93	167.52	61.63	—	—	—	-2895.21296	45.9548
Aspartic Acid									
β_L	-162.42	-176.53	-158.91	165.53	—	—	—	-684.42389	0.0000
γ_L	-77.08	72.08	-176.13	26.61	—	—	—	-684.41855	14.0186
ϵ_D	70.64	-177.71	-149.74	161.42	—	—	—	-684.41454	24.5412
δ_L	-155.56	44.89	-175.00	34.50	—	—	—	-684.41423	25.3402
γ_D	72.59	-60.37	-176.35	63.03	—	—	—	-684.41370	26.7518
α_D	70.01	25.15	-165.66	37.71	—	—	—	-684.40953	37.7005
δ_D	173.89	-39.68	178.15	65.08	—	—	—	-684.40506	49.4225
Asparagine									
β_L	-167.51	-174.32	-136.86	91.23	—	—	—	-664.56265	0.0000
γ_L	-78.38	64.12	-157.03	-87.23	—	—	—	-664.55598	17.5024
γ_D	71.44	-50.12	-153.76	-15.79	—	—	—	-664.55198	27.9952
ϵ_D	70.88	-178.87	-146.70	153.15	—	—	—	-664.54860	36.8781
α_D	57.12	46.27	-151.11	-92.79	—	—	—	-664.54683	41.5174
Threonine									
γ_L	-78.66	57.30	-177.69	-62.99	—	—	—	-610.38195	0.0000
β_L	-168.84	178.56	-168.99	178.05	—	—	—	-610.38108	2.3000
δ_L	-136.11	21.33	-168.07	-52.36	—	—	—	-610.37885	8.1374
γ_D	50.77	-20.54	-154.41	-45.52	—	—	—	-610.37546	17.0393
δ_D	131.30	-16.08	-160.00	-51.82	—	—	—	-610.36916	33.5903
ϵ_D	55.08	-170.95	-161.57	63.23	—	—	—	-610.36748	37.9921
Phenylalanine									
γ_L	-83.73	76.46	-161.65	-88.14	—	—	—	-726.90531	0.0000
β_L	-160.87	167.11	-158.22	-109.93	—	—	—	-726.90514	0.4367
γ_D	68.90	-67.50	-171.23	-95.50	—	—	—	-726.89969	14.7511
ϵ_D	63.83	-163.20	-157.13	-122.37	—	—	—	-726.89482	27.5370
δ_D	-167.79	-58.88	-174.35	-101.68	—	—	—	-726.89308	32.0916
Histidine									
β_L	-166.28	-173.57	-146.33	-61.24	—	—	—	-720.88143	0.0000
γ_L	-79.25	68.65	-165.05	98.03	—	—	—	-720.87612	13.9395
ϵ_D	61.64	-161.92	-144.84	-65.00	—	—	—	-720.87054	28.5807
γ_D	70.51	-56.89	-164.22	167.35	—	—	—	-720.87163	25.7048
α_L	-86.38	-22.46	-150.77	141.17	—	—	—	-720.86717	37.4351
α_D	54.65	49.99	-152.34	95.13	—	—	—	-720.86529	42.3660
Tyrosine									
β_L	-162.09	167.74	-156.08	-108.55	—	—	—	-802.12160	0.0000
γ_L	-77.78	77.12	-160.20	-86.27	—	—	—	-802.12118	1.1140
γ_D	71.09	-69.14	-169.73	-96.92	—	—	—	-802.11552	15.9583
ϵ_D	62.08	-167.11	-157.91	-122.81	—	—	—	-802.11035	29.5252
δ_D	-166.97	-55.29	-175.06	-102.21	—	—	—	-802.10856	34.2311
Isoleucine									
γ_L	-83.08	74.11	75.00	-174.60	—	—	—	-613.79194	0.0000

(continued on next page)

Published by NRC Research Press

Villar et al.

Conformation	ϕ	ψ	χ_1	χ_2	χ_3	χ_4	χ_5	E	ΔE
β_L	-156.42	148.68	-167.97	165.63	—	—	—	-613.79133	1.5957
δ_L	-136.80	25.56	-161.66	165.82	—	—	—	-613.78740	11.9047
γ_D	60.76	-48.32	-155.41	171.63	—	—	—	-613.78529	17.4597
δ_D	-164.70	-51.36	-174.34	167.93	—	—	—	-613.78085	29.1061
ϵ_D	74.44	169.54	-128.19	171.62	—	—	—	-613.77510	44.2142
Leucine									
γ_L	-81.16	74.57	-168.91	-171.05	—	—	—	-613.79609	0.0000
β_L	-143.41	143.76	-173.95	-174.21	—	—	—	-613.79235	9.8359
γ_D	70.23	-57.94	-161.24	-168.17	—	—	—	-613.79306	7.9632
α_L	-92.99	-17.65	-174.69	-179.07	—	—	—	-613.78950	17.2964
δ_D	-170.94	-46.54	178.32	-178.12	—	—	—	-613.78496	29.2380
Tryptophan									
β_L	-155.44	166.67	-160.18	68.15	—	—	—	-858.47665	0.0000
γ_L	-82.64	73.61	-166.05	93.80	—	—	—	-858.47446	5.7623
γ_D	69.40	-66.94	-169.12	83.53	—	—	—	-858.46881	20.5916
α_L	-75.04	-24.68	-160.55	160.42	—	—	—	-858.46505	30.4746
α_D	65.13	41.55	-150.85	97.88	—	—	—	-858.46369	34.0455
ϵ_D	64.50	-170.19	-157.04	63.37	—	—	—	-858.46481	31.1046
δ_D	-160.63	-59.59	-165.50	75.54	—	—	—	-858.46128	40.3757
Glutamic Acid									
β_L	-155.94	146.14	-171.44	176.74	-178.53	—	—	-723.73761	0.0000
γ_L	-80.16	70.84	-171.70	175.35	-179.83	—	—	-723.73574	4.9044
γ_D	68.71	-45.07	-172.07	174.36	173.14	—	—	-723.72997	20.0558
δ_L	-148.23	41.62	-165.38	176.67	-176.89	—	—	-723.73034	19.0963
ϵ_D	58.47	-141.39	-161.101	-124.786	-145.31	—	—	-723.72427	35.0215
δ_D	177.44	-40.44	-171.64	-168.67	-162.95	—	—	-723.72190	41.2386
Methionine									
γ_L	-83.10	76.67	-172.25	174.33	174.14	—	—	-972.66678	0.0000
β_L	-151.36	143.42	-172.52	175.97	178.63	—	—	-972.66163	13.5400
δ_L	-140.13	28.80	-161.04	175.44	175.81	—	—	-972.66064	16.1336
γ_D	74.25	-57.58	-72.32	65.902	98.108	—	—	-972.66013	17.4790
δ_D	-152.08	-53.23	-169.27	-75.43	112.01	—	—	-972.65349	34.8903
Glutamine									
β_L	-153.58	161.10	-175.74	52.30	96.37	—	—	-703.87368	0.0000
γ_L	-79.44	66.65	-179.09	45.46	-109.49	—	—	-703.86686	17.8979
α_L	-121.90	6.25	-176.73	60.75	-97.12	—	—	-703.86527	22.0872
γ_D	68.30	-41.23	-72.64	65.78	98.28	—	—	-703.86391	25.6547
δ_D	-138.42	-72.02	-8.58	70.63	-171.83	—	—	-703.86093	33.4709
ϵ_D	76.62	148.51	-171.80	57.29	102.13	—	—	-703.86017	35.4637
Lysine									
γ_L	-80.94	74.78	-167.64	179.44	-179.76	-179.53	—	-669.13712	0.0000
β_L	-150.11	152.15	-172.88	174.24	179.49	179.74	—	-669.13333	9.9484
γ_D	71.06	-62.40	-172.12	178.70	-179.96	-179.84	—	-669.13340	9.7676
α_D	75.09	20.39	-151.27	179.51	-178.73	-179.28	—	-669.12584	29.6252
δ_D	-167.27	-59.20	-176.68	176.76	-179.77	179.63	—	-669.12447	33.2255
Arginine									

(continued on next page)

Published by NRC Research Press

Can. J. Chem.

Conformation	ϕ	ψ	χ_1	χ_2	χ_3	χ_4	χ_5	E	ΔE
γ_L	-78.78	78.52	-168.85	177.90	178.52	169.72	171.22	-778.63981	0.0000
β_L	-135.13	144.62	-173.57	174.91	179.64	170.10	171.85	-778.63589	10.2883
γ_D	71.60	-56.53	-171.85	178.32	-179.66	171.89	171.47	-778.63583	10.4696
α_D	60.27	32.80	-156.40	175.91	178.44	169.76	170.72	-778.62812	30.7052
δ_D	-168.57	-53.63	-175.53	-178.25	-177.84	175.77	172.42	-778.62622	35.6872
ϵ_D	72.09	167.24	-161.39	166.27	176.89	165.81	170.99	-778.62251	45.4397

Draft