THE JOURNAL OF PHYSICAL CHEMISTRY A

Article

Cite This: J. Phys. Chem. A XXXX, XXX, XXX−XXX

pubs.acs.org/JPCA

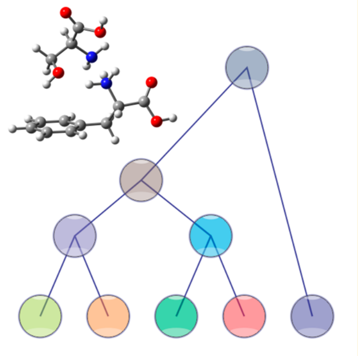
Applying Machine Learning to Vibrational Spectroscopy

Weiqiang Fu and W. Scott Hopkins*

Department of Chemistry, University of Waterloo, Waterloo, ON N2L 3G1, Canada

Supporting Information

ABSTRACT: The low-energy region of the potential energy surface (PES) of the protonated phenylalanine/serine dimer is mapped using the basin-hoping search algorithm, and 37 isomers are identified within 180 kJ·mol^{-1} of the global-minimum structure. Cluster structures are grouped using hierarchical clustering to partition the PES in terms of nuclear configuration. Calculated IR spectra for the various isomers are then compared with the isomer-specific IR spectra by means of the cosine distance metric to facilitate spectral assignment and identify which regions of the PES are populated in the electrospray ionization process.



INTRODUCTION

Infrared spectroscopy of isolated molecular ions and ionic clusters has emerged as an important tool for determining molecular structure and properties.[1−6] However, as molecular ions and clusters become larger, it can be difficult to confidently assign a specific geometry as the carrier of an observed experimental spectrum. The challenge in this regard is two-fold: One must first be able to conduct an exhaustive search of the associated potential energy surface (PES) to identify likely candidate geometries, and then one must accurately calculate the vibrational spectra of the various isomers and/or conformers and determine which calculated spectrum best matches experiment. Many approaches have been suggested to solve the problem of global optimization and identification of local minima on PESs (e.g., see refs 7−10). Monte Carlo-based methods, usually involving low-level model chemistry (e.g., molecular mechanics), are most commonly employed in IR spectroscopic studies.[11−14] These searches are then followed by higher-level electronic structure calculations, often at the density functional level of theory, to predict IR absorption spectra for the various isomers/conformers identified by the search of the PES. The calculated IR spectra are then compared with the experimental spectrum (usually qualitatively by the researcher) to identify the structure(s) most likely to give rise to the experimental observations. To date, this approach has been highly successful. However, there are cases—typically associated with highly complex potential energy landscapes—in which, despite the use of this common methodology, the spectra elude assignment. One such example is the protonated phenylalanine/serine dimer.[12]

In 2012, Lorenz and Rizzo published the results of a beautiful series of experiments wherein UV/IR double-resonance spectroscopy was used to obtain isomer-specific IR spectra of protonated 1:1 clusters of L-Phe and L-Ser.[12] In that work, the authors obtained five distinct IR depletion spectra when exciting various UV transitions in (Phe/Ser + H)$^+$, thus indicating the presence of at least five isomers in the probed ensemble (which they labeled A−E). Careful spectral analysis and ^{15}N isotopic substitution of the Phe moiety led the authors to conclude that species A−C were structures with protonation on the amino group of the Phe moiety, whereas species D and E were protonated on the Ser moiety. Lorenz and Rizzo also attempted to determine detailed structures by conducting a Monte Carlo conformational search using the AMBER force field as implemented in MacroModel, followed by DFT treatment of the candidate structures.[12] However, the PES search of this (very flexible) system proved to be a formidable task, and a definitive assignment for the observed spectra was not possible. Here, we pick up the gauntlet for the protonated phenylalanine/serine dimer. We employ a custom-written basin-hopping (BH) algorithm to search the cluster potential energy landscape and identify candidate structures for treatment at the B3LYP/6-311++G(d,p) level of theory.[11,15−17] The outcomes of these calculations are then treated with agglomerative hierarchical clustering to partition the cluster potential energy surface in terms of nuclear configurations. The calculated IR spectra and the UV/IR double-resonance spectra are then compared using the cosine distance metric to determine which isomers give rise to the spectra recorded by Lorenz and Rizzo.[12]

METHODS

The BH algorithm has been described in detail elsewhere, and additional details are available in the Supporting Information (SI).[11,17] Briefly, two separate searches of the (Phe/Ser + H)$^+$

Received: October 17, 2017
Revised: December 5, 2017
Published: December 6, 2017

DOI: 10.1021/acs.jpca.7b10303

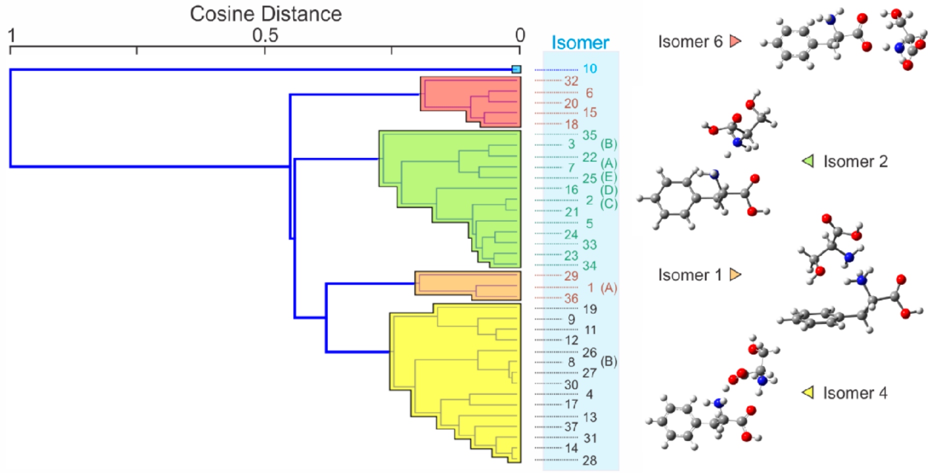


Figure 1. WPGMA dendrogram constructed from the cosine distances between the various cluster structures.

PES were conducted, one in which the site of protonation was on the Phe moiety and one with the protonation site located on the Ser moiety. The neutral and protonated amino acid moieties were first optimized individually at the B3LYP/6-311++G(d,p) level of theory to better approximate their geometries within the cluster, and atomic partial charges were calculated using the ChelpG partition scheme.[18] The individual molecules were then combined to produce the 1:1 dimer for treatment with the BH algorithm. The cluster PES was modeled using the Universal Force Field.[19] To search the PES, the dihedral angles associated with single bonds were randomly distorted by $-5° \geq \theta \geq +5°$ at each iteration of the BH code. Simultaneously, the serine moiety was randomly rotated by $-5° \geq \phi \geq +5°$ about its center of mass, and the serine center of mass was randomly translated by -0.3 Å $\geq \eta \geq +0.3$ Å in the x, y, and z directions. In total, approximately 40000 geometries were sampled. Unique structures were then carried forward for geometry optimization at the B3LYP/6-311++G(d,p) level of theory, where normal-mode analyses were also conducted to predict IR spectra and to ensure that each structure was a local minimum on the PES. Unique cluster structures were identified based on zero-point-corrected energy and geometry. The optimized Cartesian coordinates for each atom in a cluster were converted to mass-weighted distances from the cluster center of mass (i.e., $m_i r_i^{\text{com}}$, where i indicates the atom index). The resulting column vector, which was then sorted in order of increasing values of $m_i r_i^{\text{com}}$, was used as a unique identifier for cluster structure. To compare the various cluster structures, cosine distances were calculated using the Orange Python package according to the equation[20]

$$\text{distance} = \frac{\cos^{-1}(\text{similarity})}{\pi} \tag{1}$$

where

$$\text{similarity} = \cos(\theta) = \frac{\sum_{i=1}^{n} P_i Q_i}{\sqrt{\sum_{i=1}^{n} P_i^2}\sqrt{\sum_{i=1}^{n} Q_i^2}} \tag{2}$$

In eq 2, P_i and Q_i are components of the mass-weighted distance vectors P and Q, which are associated with the two isomers that are being compared. In general, the cosine similarity ranges from +1 (meaning identical) to −1 (meaning exactly opposite), with a value of 0 indicating orthogonality. Thus, two identical structures will exhibit mass-weighted distance vectors with zero angular distance between them. The angular distances between mass-weighted distance vectors increase as the differences between the geometric structures of the associated isomers increase. This procedure resulted in the identification of 37 isomers for $(\text{Phe/Ser + H})^+$ within 180 kJ·mol^{-1} of the global minimum. The Cartesian coordinates and calculated thermochemical data for these species are available in the SI.

Calculation of the cosine distances between the various cluster structures facilitated analysis through agglomerative hierarchical clustering using the weighted pair group method with arithmetic mean (WPGMA), developed by Sokal and Michener.[21,22] The WPGMA algorithm can be used to prepare a dendrogram that reflects the structure present in the pairwise distance matrix. For our purposes, this groups the $(\text{Phe/Ser + H})^+$ clusters based on structural similarity as defined by the cosine similarity of the mass-weighted distance vector, thereby providing a visual representation of the PES partitioned in terms of nuclear coordinates. At each step in the WPGMA algorithm, the nearest two clusters (P and Q) are combined into a higher-level group $P \cup Q$, thereby reducing the $m \times m$ distance matrix by one column and one row. The distance between this group and another cluster R is the arithmetic mean of the distances between R and the members of $P \cup Q$

$$d_{(P\cup Q),R} = \frac{d_{P,R} + d_{Q,R}}{2} \tag{3}$$

■ RESULTS

The dendrogram for $(\text{Phe/Ser + H})^+$ is shown in Figure 1. The top five groups based on structural similarity are highlighted. Note that isomers are numbered in order of increasing relative zero-point-corrected energy. The two major groupings—highlighted in green and yellow—are associated with compact cluster structures. In the green group, the Ser moiety is oriented in a bridging fashion with respect to the N and carbonyl O atoms of the Phe moiety. In general, this group is characterized by a N···H···N binding motif, however, there are examples wherein an internal rotation of the Ser moiety results in a N···H···O or O···H···O binding motif (e.g., isomers 16 and 25; *vide infra*). The structures of the isomers in the yellow group are similar to those in the green group, but in this case the Ser

DOI: 10.1021/acs.jpca.7b10303

moiety is bound slightly out-of-plane with respect to the plane formed by the Phe N−C−C═O atoms. In contrast to the green group, the yellow group is predominantly characterized by a N···H···O binding motif, however, there are examples wherein an internal rotation of the Ser moiety results in an O···H···O binding motif. The orange group also exhibits compact structures, but not with the Ser moiety oriented in a bridging fashion to the Phe N and carbonyl O atoms. Instead, the Ser moiety is oriented above the Phe ring (for isomers 1 and 36) or extending away from the Phe ammonium group, perpendicular to the plane of the ring (isomer 29). It is likely that isomers 1 and 36 are stabilized by partial charge/ring quadrupole interactions. The same is true for isomer 10 (the only member of the blue group), but the structure of isomer 10 is somewhat unique in comparison with the structures of the rest of the (Phe/Ser + H)$^+$ isomers. In this case, the Ser moiety is oriented above the Phe ring with the plane of the Ser COOH group nearly parallel to the ring plane. Isomer 10 is bound by a Phe—NH_3^+···O═C interaction. Finally, the isomers of the red group are all extended/elongated structures in which the COOH group of the Phe moiety binds with Ser in a bidentate fashion and the Ser moiety is oriented away from the Phe ring. All isomer structures are available in the SI; those of important isomers as determined by spectroscopy are shown in Figure 1.

Having mapped the PES of (Phe/Ser + H)$^+$, we turned our attention to the vibrational spectra recorded by Lorenz and Rizzo.[12] First, the experimental spectra that were plotted in their original *J. Am. Chem. Soc.* communication[12] were digitized and interpolated across the 2800−3750 cm^{-1} region to produce five *XY* data sets (A−E) that had intensity measurements at 0.5 cm^{-1} intervals. The calculated IR spectra for the 37 isomers identified by our BH search were convoluted with a Gaussian distribution of fwhm = 5 cm^{-1} and similarly interpolated. The intensities of the experimental and theoretical spectra were then individually normalized to a maximum of 1, and cosine distances were calculated for the various *Y* vectors (i.e., columns of intensity values). Owing to the fact that the National Institute of Standards and Technology (NIST) recommends an anharmonic scaling factor of 0.967 ± 0.021 for harmonic frequencies calculated at the B3LYP/6-311++G(d,p) level of theory, the process of calculating the cosine distance matrix was repeated several times as the scaling factor was stepped from 0.9200 to 1.0000. In this way, we were able to identify a scaling factor of ca. 0.952 for (Phe/Ser + H)$^+$ (see SI for details). Using this scaling factor, the cosine distances between the experimental and calculated spectra were calculated, then renormalized across the [0,1] interval, and plotted as the heat map shown in Figure 2. The red-colored panels in Figure 2 indicate the closest distance (i.e., best match) between experiment and theory, whereas the black panels indicate the greatest distance (i.e., worst match).

It is clear from Figure 2 that experimental spectrum E can be unambiguously assigned to isomer 25. A comparison of experimental spectrum E and the calculated spectrum for isomer 25 is shown in Figure 3. The calculated geometry of isomer 25 exhibits protonation on the amino group of the Ser moiety, which is consistent with expectations based on the ^{15}N isotopic substitution study conducted by Lorenz and Rizzo.[12] Experimental spectrum D matches well with the calculated spectra of isomers 16 and 18. The calculated spectra for both isomers are plotted along with experimental spectrum D in Figure 3. Although the calculated spectrum for isomer 18 yields a slightly better cosine distance match (because of a slightly better alignment of major spectral features), we favor isomer 16 as the carrier for spectrum D because it exhibits protonation on the amino group of the Ser moiety, which is consistent with ^{15}N isotopic substitution results.[12] Note that isomer 16 and isomer 25 both belong to the green subgroup shown in Figure 1. Their geometries differ slightly in the orientation of the Ser moiety: Isomer 16 exhibits an intermolecular hydrogen bond between the Ser carboxylic acid OH and the Phe carbonyl O, whereas isomer 25 exhibits a hydrogen bond between the Ser carboxylic acid OH and the Phe amino group.

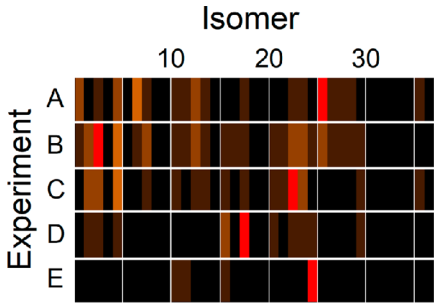


Figure 2. Heat map showing the relative distances between the experimental IR spectra recorded by Lorenz and Rizzo[12] and the calculated IR spectra for the 37 isomers of (Phe/Ser + H)$^+$ identified in our search. Red indicates the closest cosine distance between experiment and theory; black indicates the farthest cosine distance.

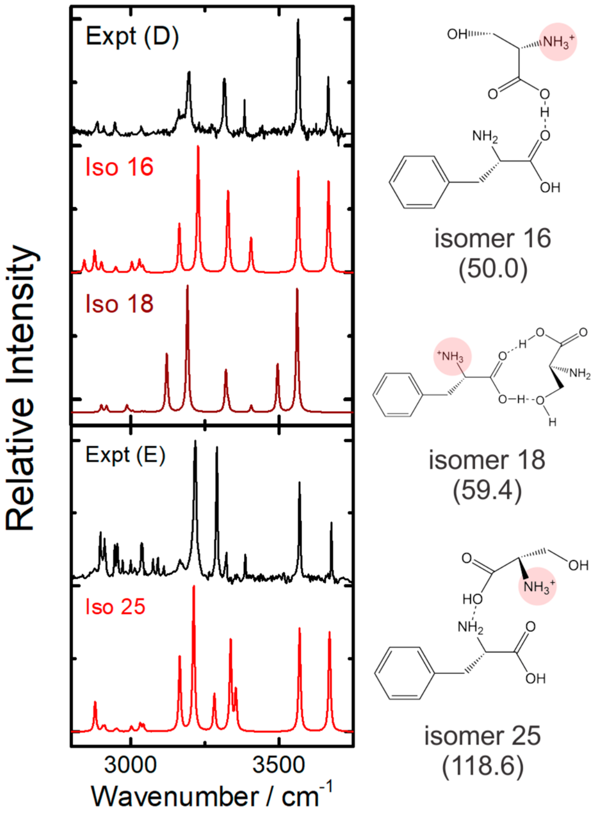


Figure 3. (Black traces) Experimental (Phe/Ser + H)$^+$ UV/IR double-resonance spectra D and E, adapted from ref 12. (Red traces) Calculated IR spectra for isomers 16, 18, and 25. Calculations were conducted at the B3LYP/6-311++G(d,p) level of theory. A scaling factor of 0.952 was applied to the calculated spectra. Zero-point-corrected energies (in parentheses; relative to the global-minimum structure) are reported in kJ mol^{-1}.

Figure 2 shows that the spectra of several isomers provide reasonable matches with experimental spectra A–C. Upon examining the five closest matches for experimental spectrum C (isomers 2, 3, 5, 23, 24), we find that all five of these isomers are associated with the green subgroup of the dendrogram (Figure 1) and exhibit N···H···N binding. Consequently, these species all exhibit similar spectra. Even though isomers 23 and 5 have slightly smaller cosine distances to spectrum C, we favor the assignment of spectrum C to isomer 2 because of its lower calculated relative energy (assuming that structures within 3 kJ·mol^{-1} of the global minimum are populated). Note that the breadth of the spectral features and the peak at ca. 3605 cm^{-1}, which was identified by Lorenz and Rizzo as a contribution from spectrum A, suggests that multiple isomers might be contributing to the observed spectrum. Our analysis shows that multiple isomers are responsible for experimental spectra A and B. The five best matches for spectrum B are isomers 2, 3, 5, 8, and 24. With the exception of isomer 8, which belongs to the yellow subgroup, these isomers are all associated with the green subgroup of the dendrogram and exhibit N···H···N binding. The spectrum of isomer 3, which exhibited the best match to spectrum B, is plotted in Figure 4. Although the spectrum of isomer 3 is a good match in the NH and OH regions of spectrum B, there are additional features in the experimental spectrum that remain unassigned. These features are not well-predicted by isomers 2, 5, or 24, which all have structures and spectra similar to those of isomer 3. However, the spectrum for isomer 8 does capture these features and is overlaid with the spectrum of isomer 3 in Figure 4. In the case of experimental spectrum A, the top five matches were isomer 1 (orange group), isomers 5 and 7 (green group), and isomers 13 and 26 (yellow group). The spectrum of isomer 1, the global-minimum structure, is plotted in Figure 4 along with the spectrum of isomer 7. Although isomer 5 is calculated to be lower in energy than isomer 7, we discounted isomer 5 because its spectrum exhibits an additional feature to higher wavenumber of the highest observed band. A convolution of the calculated spectra of isomers 1 and 7 captures all of the features in spectrum A, except for the bands observed at 3115 and 3203 cm^{-1} (marked with an × in Figure 4). The spectrum of isomer 13 does not exhibit bands at either of these positions; however, the spectrum of isomer 26 (not plotted) does show a peak at 3200 cm^{-1}. Thus, similarly to spectrum B, spectrum A appears to be a convolution of spectra for multiple isomers. Note that all five of the calculated spectra shown in Figure 4 are associated with isomers that exhibit protonation at the Phe nitrogen atom or shared between the amino groups of the two amino acids. This accords with expectations for spectra A–C based on the ^{15}N isotopic substitution study by Lorenz and Rizzo.[12]

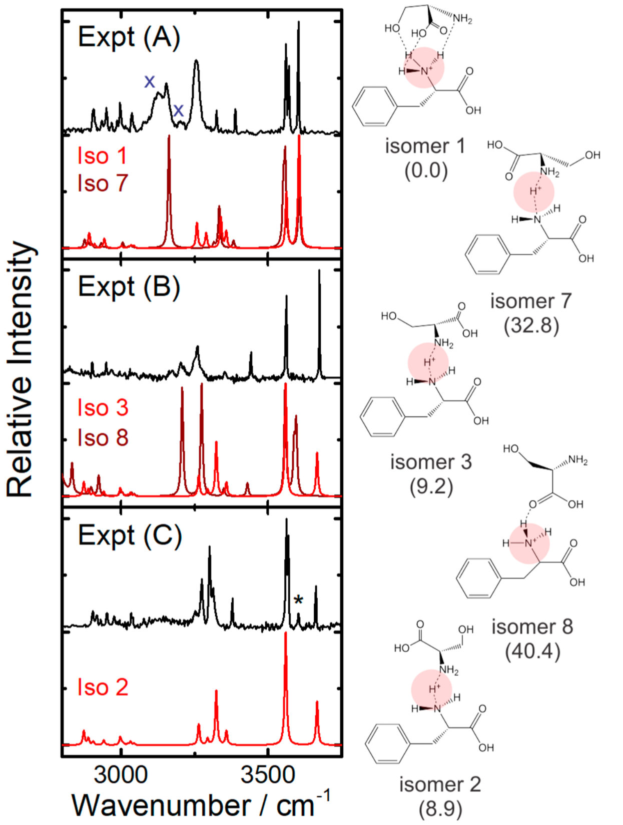


Figure 4. (Black traces) Experimental (Phe/Ser + H)$^+$ UV/IR double-resonance spectra A–C, adapted from ref 12. (Red traces) Calculated IR spectra for isomers 1–3, 7, and 8. Calculations were conducted at the B3LYP/6-311++G(d,p) level of theory. A scaling factor of 0.952 was applied to the calculated spectra. Zero-point-corrected energies (in parentheses; relative to the global-minimum structure) are reported in kJ mol^{-1}. The band marked with an asterisk is attributed to spectrum A.[12] Bands marked with an × are attributed to higher-energy structures.

■ CONCLUSIONS

By applying hierarchical clustering to the results of our BH search of the (Phe/Ser + H)$^+$ PES, we are able to partition the PES in terms of nuclear configuration. Although the resulting dendrogram does not provide information on barriers to isomerization (as would a disconnectivity graph),[23,24] this exercise does provide some insight into likely regions of kinetic trapping. Indeed, when we compare our calculated spectra to the experimental IR spectra recorded by Lorenz and Rizzo,[12] we find that the spectral carriers are associated with three relatively different regions of the PES. It should be noted that the assignment of the experimental spectra was enabled by calculating cosine distances between the experimental spectra and calculated harmonic spectra. It is possible that the matching algorithm could be improved by using another distance metric or by introducing anharmonic corrections to the calculated frequencies. Such considerations might be necessary for larger, more complex molecular clusters. Nevertheless, this work does demonstrate the utility of introducing aspects of unsupervised machine learning to the analysis of vibrational spectra.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpca.7b10303.

Cluster Cartesian atomic coordinates and thermochemical data (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: scott.hopkins@uwaterloo.ca.

ORCID

W. Scott Hopkins: 0000-0003-1617-9220

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors gratefully acknowledge helpful discussions with Prof. Roger Melko (University of Waterloo). Furthermore, we acknowledge high-performance computing support from the SHARCNET consortium of Compute Canada. The generous financial support of this work from the Natural Sciences and Engineering Research Council of Canada is also acknowledged.

■ REFERENCES

(1) Duncan, M. A. Infrared spectroscopy to probe structure and dynamics in metal ion−molecule complexes. *Int. Rev. Phys. Chem.* **2003**, *22*, 407−435.

(2) Hammer, N. I.; Diken, E. G.; Roscioli, J. R.; Johnson, M. A.; Myshakin, E. M.; Jordan, K. D.; McCoy, A. B.; Huang, X.; Bowman, J. M.; Carter, S. The vibrational predissociation spectra of the $H_5O_2^+\cdot RG_n$(RG = Ar,Ne) clusters: Correlation of the solvent perturbations in the free OH and shared proton transitions of the Zundel ion. *J. Chem. Phys.* **2005**, *122*, 244301.

(3) Lemaire, J.; Boissel, P.; Heninger, M.; Mauclaire, G.; Bellec, G.; Mestdagh, H.; Simon, A.; Caer, S. L.; Ortega, J. M.; Glotin, F.; Maitre, P. Gas phase infrared spectroscopy of selectively prepared ions. *Phys. Rev. Lett.* **2002**, *89*, 273002.

(4) Oomens, J.; Sartakov, B. G.; Meijer, G.; Von Helden, G. Gas-phase infrared multiple photon dissociation spectroscopy of mass-selected molecular ions. *Int. J. Mass Spectrom.* **2006**, *254*, 1−19.

(5) Pribble, R. N.; Zwier, T. S. Size specific infrared spectra of benzene-$(H_2O)_N$ clusters (N = 1 - 7) - Evidence for noncyclic $(H_2O)_N$ structures. *Science* **1994**, *265*, 75−79.

(6) Rizzo, T. R.; Stearns, J. A.; Boyarkin, O. V. Spectroscopic studies of cold, gas-phase biomolecular ions. *Int. Rev. Phys. Chem.* **2009**, *28*, 481−515.

(7) Piela, L.; Olszewski, K. A.; Pillardy, J. On the stability of conformers. *J. Mol. Struct.: THEOCHEM* **1994**, *308*, 229−239.

(8) Scheraga, H. A. Some approached to the multiple minima problem in the calculation of polypeptide and protein structures. *Int. J. Quantum Chem.* **1992**, *42*, 1529−1536.

(9) Wales, D. J.; Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **1997**, *101*, 5111−5116.

(10) Wales, D. J.; Scheraga, H. A. Global optimization of clusters, crystals, and biomolecules. *Science* **1999**, *285*, 1368−1372.

(11) Hopkins, W. S.; Marta, R. A.; McMahon, T. B. Proton-Bound 3-Cyanophenylalanine Trimethylamine Clusters: Isomer-Specific Fragmentation Pathways and Evidence of Gas-Phase Zwitterions. *J. Phys. Chem. A* **2013**, *117*, 10714−10718.

(12) Lorenz, U. J.; Rizzo, T. R. Multiple Isomers and Protonation Sites of the Phenylalanine/Serine Dimer. *J. Am. Chem. Soc.* **2012**, *134*, 11053−11055.

(13) Martens, J.; Grzetic, J.; Berden, G.; Oomens, J. Structural identification of electron transfer dissociation products in mass spectrometry using infrared ion spectroscopy. *Nat. Commun.* **2016**, *7*, 11754.

(14) Poutsma, J. C.; Martens, J.; Oomens, J.; Maitre, P.; Steinmetz, V.; Bernier, M.; Jia, M. X.; Wysocki, V. Infrared Multiple-Photon Dissociation Action Spectroscopy of the b_2^+ Ion from PPG: Evidence of Third Residue Affecting b_2^+ Fragment Structure. *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 1482−1488.

(15) Becke, A. D. Density functional exchange energy approximation with correct asymptotic behavior. *Phys. Rev. A: At., Mol., Opt. Phys.* **1988**, *38*, 3098−3100.

(16) Becke, A. D. Density functional thermochemistry 0.3. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.

(17) Lecours, M. J.; Chow, W. C. T.; Hopkins, W. S. Density Functional Theory Study of $Rh_nS^{0,\pm}$ and $Rh_{n+1}{}^{0,\pm}$ (n = 1−9). *J. Phys. Chem. A* **2014**, *118*, 4278−4287.

(18) Wiberg, K. B.; Rablen, P. R. Comparison of atomic charged derived via different procedures. *J. Comput. Chem.* **1993**, *14*, 1504−1518.

(19) Rappe, A. K.; Casewit, C. J.; Colwell, K. S.; Goddard, W. A.; Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **1992**, *114*, 10024−10035.

(20) Demsar, J.; Curk, T.; Erjavec, A.; Gorup, C.; Hocevar, T.; Milutinovic, M.; Mozina, M.; Polajnar, M.; Toplak, M.; Staric, A.; Stajdohar, M.; Umek, L.; Zagar, L.; Zbontar, J.; Zitnik, M.; Zupan, B. Orange: Data Mining Toolbox in Python. *J. Mach. Learn. Res.* **2013**, *14*, 2349−2353.

(21) Michener, C. D.; Sokal, R. R. A quantitive approach to a problem in classification. *Evolution* **1957**, *11*, 130−162.

(22) Sokal, R. R.; Michener, C. D. A statistical method for evaluating systematic relationships. *Univ. Kansas Sci. Bull.* **1958**, *38*, 1409−1438.

(23) Becker, O. M.; Karplus, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.* **1997**, *106*, 1495−1517.

(24) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758−760.

DOI: 10.1021/acs.jpca.7b10303