AutoGraph: Autonomous Graph Based Clustering of Small-Molecule Conformations

Kiyoto A. Tanemura,* Susanta Das, and Kenneth M. Merz, Jr.*

Department of Chemistry, Michigan State University, 578 S. Shaw Lane, East Lansing, Michigan, 48824, United States

E-mail: tanemur1@chemistry.msu.edu; merzjrke@msu.edu

Abstract

While accurately modeling the conformational ensemble is required for predicting properties of flexible molecules, the optimal method of obtaining the conformational ensemble seems as varied as their applications. Ensemble structures have been modeled by generation, refinement, and clustering of conformations with a sufficient number of samples. We present a conformational clustering algorithm intended to automate the conformational clustering step through the Louvain algorithm, which requires minimal hyperparameters and importantly no predefined number of clusters or threshold values. The conformational graphs produced by this method for *O*-succinyl-L-homoserine, oxidized nicotinamide adenine dinucleotide, and 200 representative metabolites each preserved the geometric/energetic correlation expected for points on the potential energy surface. Clustering based on these graphs provide partitions informed by the potential energy surface. Automating conformational clustering in a workflow with AutoGraph may mitigate human biases introduced by guess-and-check over hyperparameter selection while allowing flexibility to the result by not imposing predefined criteria other than optimizing the model's loss function. Associated codes are available at https://github.com/TanemuraKiyoto/AutoGraph .

Introduction

Accurately modeling the distribution of equilibrium conformers is prerequisite in predicting the microscopic and macroscopic properties of flexible molecules. Obtaining a conformational ensemble is foundational to calculating average properties of molecular systems.[1] Methodologies such as ensemble docking of protein-ligand systems,[2] three dimensional quantitative structure-activity relationship,[3] and constructing Markov state models (MSM) from molecular dynamics trajectories[4] rely on sufficiently sampling from the conformational ensemble.

Many methods, algorithms, and their variants exist for conformation generation.[5] The objective of conformation generation protocols is to identify many equilibrium conformers at local minima of the potential energy surface (PES), which would be major contributors among all thermally accessible conformations. This may involve sufficient sampling of nonredundant conformations, followed by refinement of those conformations to local energy minima. Sequential methods such as molecular dynamics and Monte Carlo simulated annealing combine sampling and scoring of conformers to return physically informed, low energy conformers, however are generally computationally intensive compared to knowledge based methods.[6,7] Knowledge based methods such as OMEGA and ETKDG algorithms narrow the search space by using the distributions of observed dihedral angles and ring structures from crystallographic databases.[8,9] The rapid conformation generation by such algorithms should be followed up with physically informed structure refinement.

Geometry optimization by ab initio calculations converge to low energy conformations, however the high computational cost limits its applicability to the numerous conformations which need to be sampled. Recent development and benchmarking of machine learning based potentials such as ANI-2x have prompted its utilization for certain high throughput quantum chemical applications.[10,11] For example, ANI-1ccx potentials were used to accelerate the refinement of generated conformers in a quantum mechanical (QM) NMR spectral prediction workflow.[12,13] The conformers optimized by these models, however, generally do not converge to the same local minima as ab initio methods. Hence, conformational clustering becomes

important in narrowing the number of "full" QM geometry optimization calculations required to obtain a representative set of conformers to estimate the conformational ensemble.

Clustering is a task in unsupervised machine learning, in which individual data are coarse grained into disjoint groups. Clustering is used for purposes such as auto-label generation, dimensionality reduction, image segmentation, and visualization of data. To date, numerous clustering algorithms and their variants have been developed and deployed across disciplinary lines.[14,15] Many conformational clustering algorithms have also been evaluated.[16–20] A majority of these algorithms require the number of clusters or threshold values defined a priori, though these hyperparameters vary by the data under evaluation and its choice may be nontrivial.[21] Unless automated, the iterative guess and check of hyperparameters can render the clustering protocol into one requiring supervision, thus limiting its throughput and integrity from user bias. Highly automated conformational clustering protocols which do not require the number of clusters or threshold value be predefined would be advantageous for applications such as high throughput metabolomics. The performance of several of such autonomous conformational clustering algorithms have been assessed.[17,19]

Here we present an autonomous graph based conformational clustering algorithm named AutoGraph. AutoGraph processes the atomic root-mean-squared deviation (RMSD) matrix between conformers into an affinity matrix using a generic Gaussian kernel. The matrix is processed as a graph object and its nodes are clustered using the Louvain algorithm, which does not require number of clusters or thresholds be predefined.[22] We estimate the conformational ensembles for *O*-succinyl-L-homoserine and nicotinamide adenine dinucleotide as simple examples before exploring the conformational graphs of 200 representative metabolites.

Methods

Definitions

Let there be $n \in \mathbb{N}$ conformers of exactly one molecule.

- A graph $G = (V, E)$ consists of a set of vertices V and a set of edges E.
- A **distance matrix** is an $n \times n$ matrix recording the pairwise dissimilarity between all conformers.
- An **affinity matrix** is an $n \times n$ matrix storing the pairwise similarity between conformers.
- A binary **adjacency matrix** is an $n \times n$ matrix with 1 indicating the presence of an edge and 0 otherwise.
- We define a **filtered matrix** of matrix M be the element-wise product between M and an adjacency matrix.

The AutoGraph Conformational Clustering Algorithm

Atomic root mean squared deviation (RMSD) are computed comprehensively between n structures and stored in a symmetric $n \times n$ distance matrix. The Kabsch algorithm is used for finding the minimum pairwise RMSD.[23] An affinity matrix is calculated by applying a generic radial basis function, $\phi(r) = \exp(-r^2)$. A threshold value is applied to remove edges with low valued weights from the affinity matrix. An adjacency matrix is produced such that the threshold is the maximum value for which the filtered affinity matrix contains exactly one component. The resulting filtered affinity matrix encodes a undirected, wighted graph $G = (V, E)$ consisting of vertices V and edges E. Clusters are detected by applying the Louvain algorithm to the filtered affinity matrix.[22] The lowest energy conformer is reported for each cluster as its representative conformer.

Performance Evaluation

Correlation of Actual Energy to Local Weighted Estimation

The goal of conformational clustering is to discretize the PES to identify major contributors to average properties. AutoGraph clusters metabolite conformers based on the conformational graph generated. To justify clustering by the conformational graph, we assessed the geometric/energetic correlation in the graphs. We measured the single point energies of all conformers, which yields the actual measured energy. We also take a local weighed estimate of each node's energy by the average of neighbors' energy values weighted by incident edge weights. The Spearman correlation between actual and local estimated energy therefore provides a metric for the geometric/energetic correlation implicit in the conformational graph.

Formally, let the conformational graph $G = (V, E)$ consist of a set of vertices V connected by edges in set E. Let $v_i \in V$ and let its neighbors be $N_i = \{n_j | \{v_i, n_j\} \in E\} \subseteq V$. Note that $\{v_i, v_i\} \notin E$. Let $U(v)$ be the single point energy of the conformer assigned to vertex $v \in V$. Also, let $w(e)$ be the weight of edge $e \in E$. The actual energy of v_i is $U(v_i)$. The local estimated energy of v_i is given by,

$$\hat{U}(v_i) = \frac{\sum_{j=1}^{p} w(\{v_i, n_j\}) U(n_j)}{\sum_{j=1}^{p} w(\{v_i, n_j\})}$$

The Spearman correlation coefficient was determined between U and $\hat{U}$, calculated in R.[24]

Cluster-wise Variance of Conformer Energy

Geometric similarity within clusters are achieved trivially by the objective function of the conformational clustering algorithm. We must instead consider the variance in the energies of the clustered conformers to evaluate the methodology. Energy as a metric informs us of the conformers' proximity in the PES and is independent of the clustering protocol, thus provides

an independent metric for assessment. Note the AutoGraph algorithm considers conformer energy after partitioning conformers by clusters, thus its independence from assessment method is not compromised. For a distribution of values $X = \{x_i | i = 1, 2, \ldots, N\}$, the sample variance σ^2 is given by,

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$$

Because our conformers are organized by clusters, we can decompose the total variance into variance between clusters and variance within clusters. The size of clusters are heterogeneous, thus we take a weighted mean to compute each terms as follows:

$$\sigma^2 = \sigma^2_{between} + \sigma^2_{within} = \sum_{j=1}^{C} \frac{c_j}{N}(\mu_j - \mu)^2 + \sum_{j=1}^{C} \frac{c_j}{N}\sigma_j^2$$

Suppose the conformers are clustered with no bias. Then $\sigma^2_{between}$ is negligible, thus we would expect $\sigma^2 \approx \sum_{j=1}^{C} \frac{c_j}{N}\sigma_j^2$. The converse of this statement is true, in which $\sigma^2 \neq \sum_{j=1}^{C} \frac{c_j}{N}\sigma_j^2$ implies a bias in clustering the energy. Our null hypothesis H_0 is $\sigma^2 = \sum_{j=1}^{C} \frac{c_j}{N}\sigma_j^2$. Our alternative hypothesis H_1 is $\sum_{j=1}^{C} \frac{c_j}{N}\sigma_j^2 < \sigma^2$. We detect a difference in variances using the F-test of equality of variances.[25]

To evaluate the energy variance among the 200 benchmark metabolites, we compute the variance in energy within clusters and variance across clusters for each of the 200 metabolites. We then apply the paired left-tailed Wilcoxon signed-rank test to determine the statistical significance of the difference between the two distributions for variance values.[26]

Because $0 \leq \sigma^2_{between}$, the F-test is biased toward type-I error. For this reason, we impose a stringent significance level α of 0.005.

Case Studies

O-Succinyl-L-Homoserine

We clustered conformers generated for *O*-succinyl-L-homoserine (OSLH) to illustrate the use of the AutoGraph conformational clustering algorithm. Conformers were generated and refined in a previous study.[13] In summary, the MacroModel/ConfGen protocol (Schrödinger, Inc.) was used to generate 501 conformations unique to 0.1 Å in atomic RMSD.[27,28] This was followed by ANI-1ccx calculations in the gas phase.[12] Only conformers with no imaginary vibrational frequencies were retained, narrowing the conformers to 485. Calculations using ANI potentials were performed using the Atomic Simulation Environment (ASE) interface.[29] Conformations were clustered by AutoGraph and the resultant graph was visualized using Gephi.[30] In addition, a charged structure was prepared using the PrepWizard tool (Schrödinger, Inc.).[31,32] Single point energies were calculated using the Gaussian quantum chemistry software at the HF/6-31G(d) level of theory in the gas phase for both the neutral and charged conformers.[33] ANI-1ccx optimized neutral OSLH conformers were further refined using Gaussian at the B3LYP/6-31G(d,p) level of theory in gas phase. Conformers were confirmed to be at local minima by vibrational analysis. The fully optimized geometries were subjected to clustering by AutoGraph.

Nicotinamide Adenine Dinucleotide

We also clustered conformers generated for nicotinamide adenine dinucleotide in the oxidized form (NAD+). We generated 1000 conformers from the SMILES of NAD+ using RDKit's implementation of the ETKDG algorithm, unique to 0.1 Å in atomic RMSD.[9,34] Structures were optimized using the MMFF94 potential.[35] This was followed by the Austin Model 1 (AM1) semiempirical method in the gas phase using the MOPAC software interfaced through ASE.[29,36,37] Only structures with the original topology were retained, filtering to 785 total conformers. Conformations were clustered by AutoGraph and the graphs were visualized using Gephi.[30] Single point energies were calculated using the Psi4 quantum chemistry package using the B3LYP/6-31G(d) level of theory in the gas phase.[38]

Benchmark Dataset

Metabolite Curation

Molecules were selected from the Human Metabolome Database (HMDB), consisting of 114184 metabolites.[39] Those localized to the cytosol, nucleus, or mitochondria were kept to yield 8249 molecules. We selected metabolites containing only elements which could be subjected to energy calculation by ANI-2x potentials (CHONSFCl) to yield 7547 molecules.[10] The number of rotatable bonds were calculated for each metabolite to remove trivial or unfeasible cases for conformation generation. Metabolites with rotatable bonds on the range of the 50th to 95th quantile were retained, resulting in all metabolites having four to fourteen rotatable bonds. Out of the 3350 remaining metabolites, 200 representative structures were chosen using the following protocol. Morgan fingerprints were calculated for all molecules, using 2048 bits with a connectivity of three. A 3350×3350 matrix of Tanimoto distance between all fingerprints were calculated. A Ward dendrogram was calculated from the dissimilarity matrix and a threshold was applied to produce 200 clusters. Representative metabolites were selected from each cluster by having the greatest in-cluster degree. The metabolites are given in the SI (Table S1).

Conformer Generation

Up to 1000 conformers were generated for each selected metabolite using RDKit's implementation of the ETKDG algorithm, discarding any structures with RMSD below 0.1 Å from any of the previous structures.[9,34] All structures were optimized by the Merck Molecular Force Field 94 (MMFF94) in the gas phase.[35] Further, all structures were optimized using ANI-2x potentials with the BFGS optimizer in the gas phase.[10] The final potential energies calculated with ANI-2x potentials were recorded.

Results and Discussion

Clustering *O*-succinyl-L-homoserine conformers by AutoGraph as an illustrative example

Among the strengths of using a graph to represent the conformers is graphs offer intuitive summaries of the relationships in the data, and an abundance of graph algorithms are available. AutoGraph was designed with the strategy to process the RMSD matrix into a readily interpretable form of a graph, then apply existing graph clustering protocols. We highlight the conformational graph of OSLH, for which each node represents exactly one conformer, and edge weights are proportional to the structural similarity determined by atomic RMSD between conformers (Figure 1).

Since the conformers are geometry optimized, we suspect the densely connected subgraphs represent basins in the PES, thus should be grouped in the same cluster. The AutoGraph protocol identified 28 clusters as shown. Nodes seem to have neighbors with similar energy, particularly in dense regions of the graph. The qualitative geometric/energetic correlation provides preliminary evidence for information regarding the PES is available implicitly in the conformational graph.

The superimposed conformers illustrate, while noisy, the overall molecular shape is similar within each cluster. The intuitive clusters on the graph appear to translate to qualitative conformational similarity for this system.

Unlike other clustering algorithms which take the RMSD matrix as the direct input, AutoGraph first processes the RMSD matrix into a graph representation. We assess whether the resulting conformational graph reasonably preserves the geometric/energetic similarities between conformers as expected for the PES. To do so, we measured the Spearman correlation coefficients ρ between the local estimated energy values on OSLH's conformational graph with their actual energy values, calculated by HF/6-31G(d) in the gas phase (Figure 2). There is a positive correlation for both conformational graphs, suggesting the monotonic relationship between geometry/energy is preserved even after kernelizing and filtering the RMSD matrix input. The local estimated energy appears more responsive to the actual energy when the adaptive threshold is applied. We also observe an improvement in the correlation. The improvement is particularly pronounced for the case of the

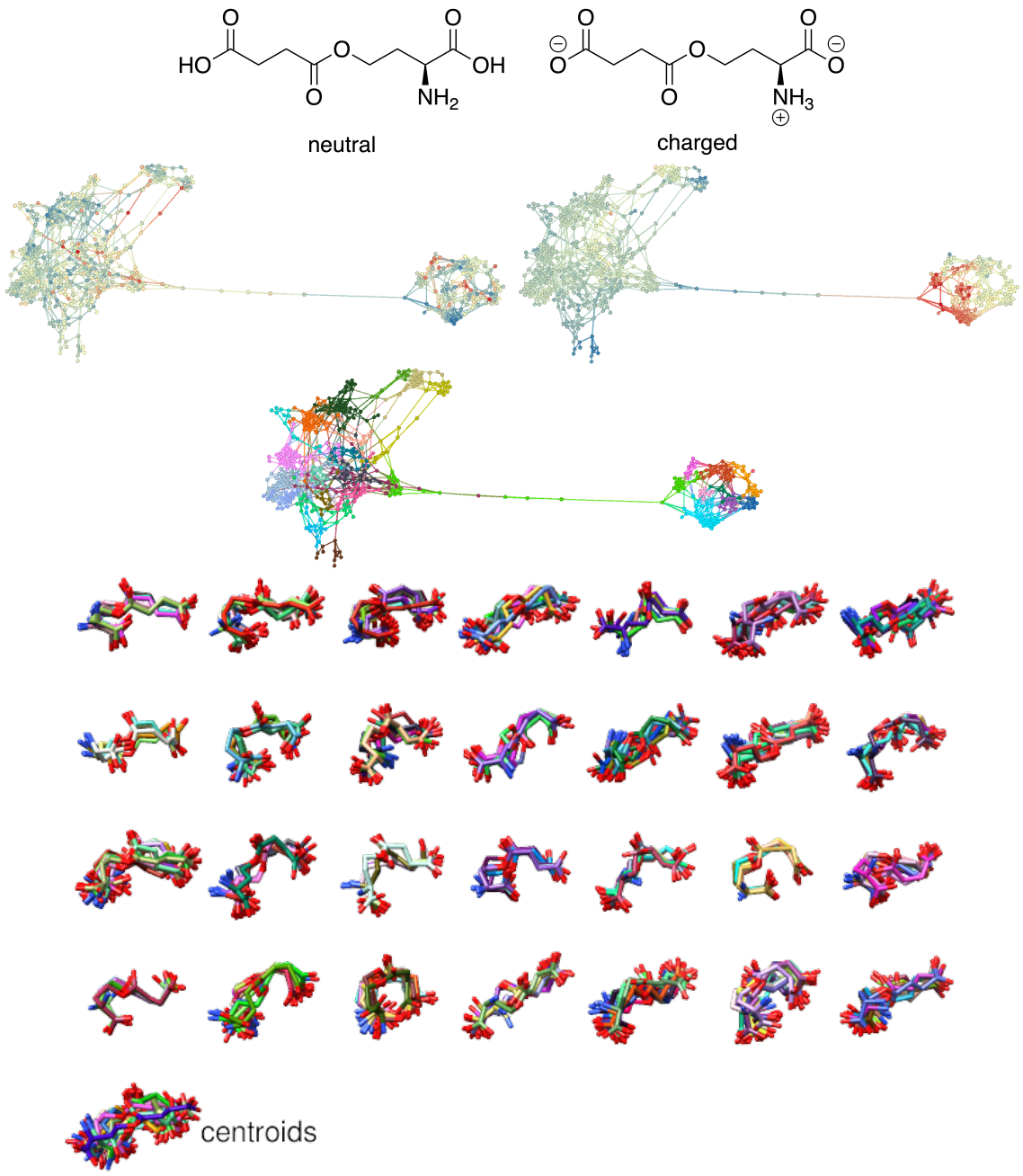

Figure 1: *O*-succinyl-L-homoserine (top) was subjected to conformer generation, geometry optimization, and clustering by AutoGraph. The conformational graphs colored by single point energy values for neutral (middle left) and charged (middle right) are shown using a gradient of blue (low energy) to red (high energy). The conformational graph colored by assigned cluster provide intuitive results (middle center). Conformers within each cluster were superimposed to their centroid (bottom).

charged system. We suspect the magnitude of Coulombic interactions in the gas phase is much greater than that of the noise such that we observe a strong relationship between the geometry and energy of the system. It is promising to observe a correlation in both the charged and neutral case even though their trends in relative energy differ. While the conformational graph represents purely geometric information, we observe we can infer energetic information because the refined conformations are biased to minima in the PES.

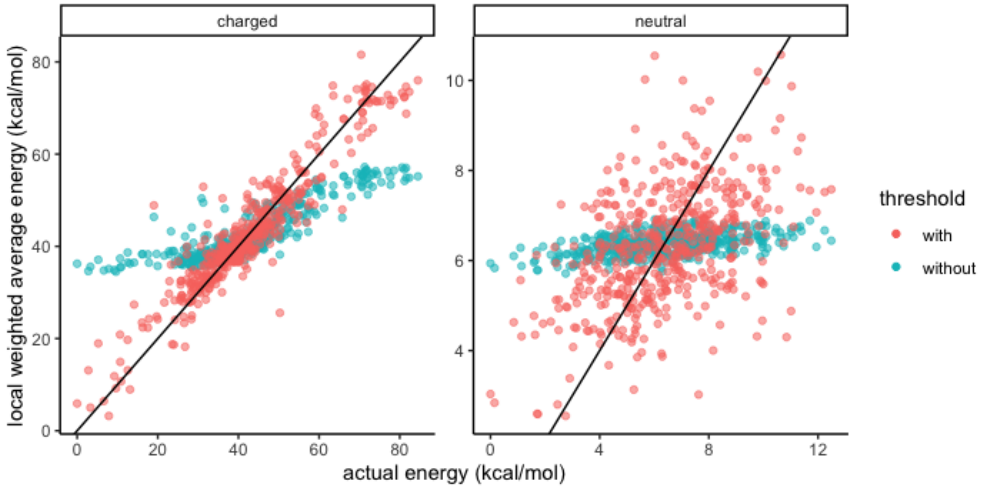

Figure 2: Locally weighted estimated relative energies was plotted against the measured relative single point energy values ($n = 485$). Graphs before and after filtering low weight edges were considered and Spearman correlation coefficients were calculated (charged: $\rho_{with} = 0.916, \rho_{without} = 0.852$; neutral: $\rho_{with} = 0.465, \rho_{without} = 0.457$). The line for $x = y$ is plotted.

All ANI-1ccx optimized neutral OSLH were subjected to geometry optimization at B3LYP/6-31G(d,p) level of theory. The conformational graphs were visualized (Figure 3). In the resulting conformational graph, we indicate the conformers which were chosen as centroids by AutoGraph in the previous step for the ANI-1ccx optimized structures. We observe at least one structure from the centroids chosen from the ANI-1ccx optimized conformers appear in densely connected regions of the B3LYP/6-31G(d,p) optimized conformational graph. If we were to calculate an average property, we may select a representative conformer from each of the clusters of the B3LYP/6-31G(d,p) optimized conformational graph and take a Boltzmann average. We observe we can perform ab initio optimization on only the centroids selected from the ANI-1ccx optimized graph and obtain similar results as if we subjected all conformers to full geometry optimization. However,

we can expedite the workflow in this case by subjecting only 28 starting conformers to the expensive geometry optimization rather than the full collection of 485 conformers.

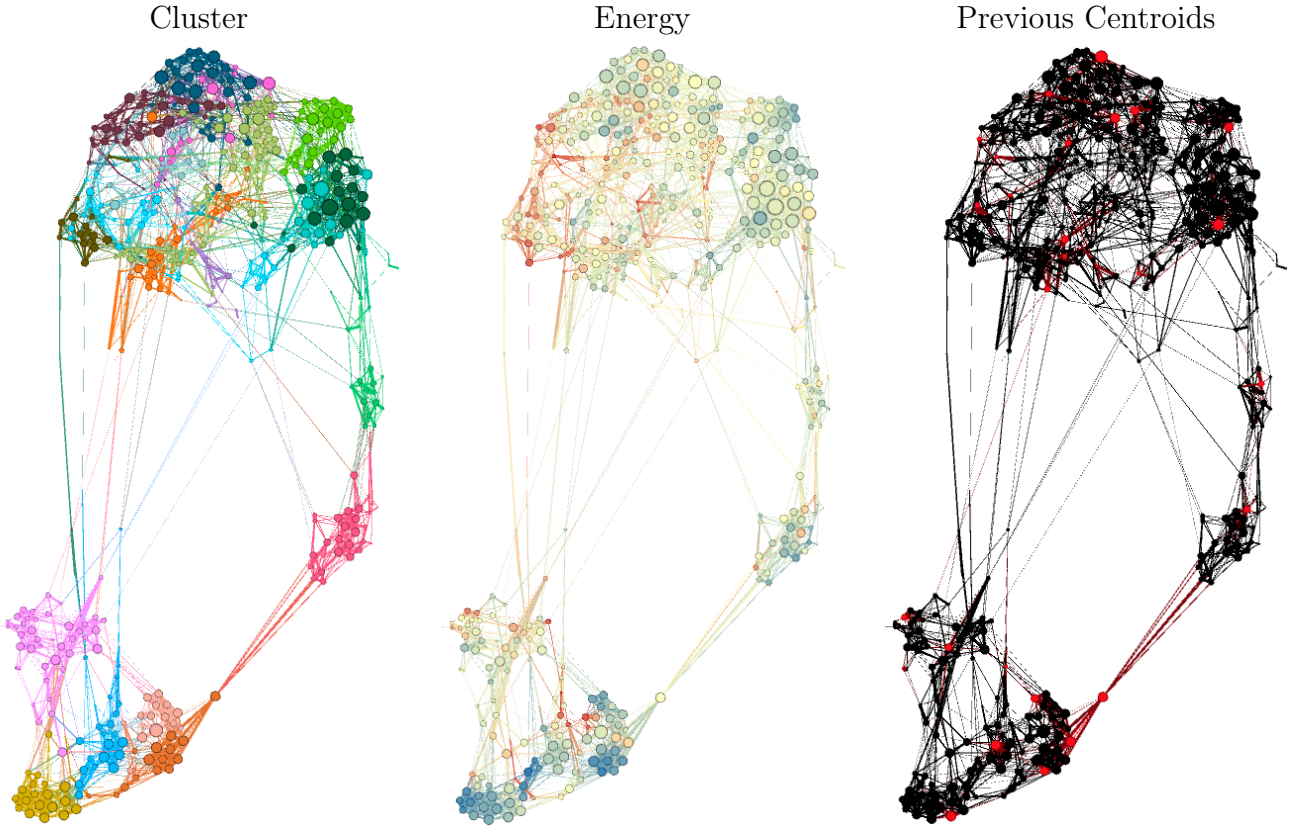

Figure 3: The conformational graph for B3LYP/6-31G(d,p) optimized neutral OSLH is colored by cluster assigned by AutoGraph, relative energy values from low (blue) to high (red), and centroids selected by AutoGraph from ANI-1ccx optimized OSLH conformers (red). Size of nodes are proportional to the weighted degree of each node.

Clustering nicotinamide adenine dinucleotide conformers by AutoGraph

We next examine the clustering result of NAD+ conformations. NAD+ is a well-known co-factor that mediates the redox currency in the cell.[40] Importantly, NAD+ contains two phosphorus atoms, which is an atom type not represented in the ANI-2x potential. We require an alternative protocol for refinement, and we can probe the robustness of the AutoGraph protocol to the method for conformer generation and refinement. After conformer generation and refinement with MMFF94,

we optimized the NAD+ conformations using AM1 in gas phase. The 785 conformers which retained the original topology were subjected to conformational clustering by AutoGraph.

The AutoGraph protocol identified 31 clusters as shown (Figure 4). Overall, the graph appeared more globally connected than the OSLH example. The semiempirical QM method parameterized model 7 (PM7) has a lesser agreement to coupled cluster energy values when compared to ANI potentials, so the AM1 energies as a semiempirical method may have exhibited less convergence to local optima for NAD+ than ANI-2x did to OSLH.[11] The graph colored by single point energy values show an overall gradient, in which the densely connected region also appear to be low energy conformers. Meanwhile the higher energy side of the graph seems more sparsely connected. Superimposed conformers are also visually sound. While a positive correlation was observed between actual and local weighted energy estimates for conformational graphs before and after applying an edge weight threshold ($\rho_{with} = 0.845, \rho_{without} = 0.842$), no notable change in the correlation coefficient was observed (Figure S1). Because energetic information is inferred from the conformational graph, which only encodes geometric information explicitly, the success of the clustering results may depend on the convergence of optimized structures, which in turn depend on the accuracy of the energy calculation method. The throughput achieved by deep learning potentials like ANI potentials provide a unique opportunity for AutoGraph to interface between deep learning and ab initio methods in high throughput applications.

The cluster-wise variance in energy values were considered as an evaluation metric. The variance in energy within clusters was compared against the variance of energy among all conformers for OSLH and NAD+ individually (Table 1). We detect the within-cluster variance in energy is lesser than the overall variance for charged OSLH and for NAD+. The trend for neutral OSLH was not validated, likely due to its negligible effect size of $-0.65 \left(\frac{kcal}{mol}\right)^2$ for the difference in variances.

While conformational clustering of OSLH and NAD+ both produced reasonable results, the results of this case studies are anecdotal. Validation of the AutoGraph protocol should be performed over a diverse data set chosen in a manner that minimizes bias. For this reason we constructed a benchmark conformation set for 200 metabolites chosen from the HMDB using a fairly automated protocol.[39]

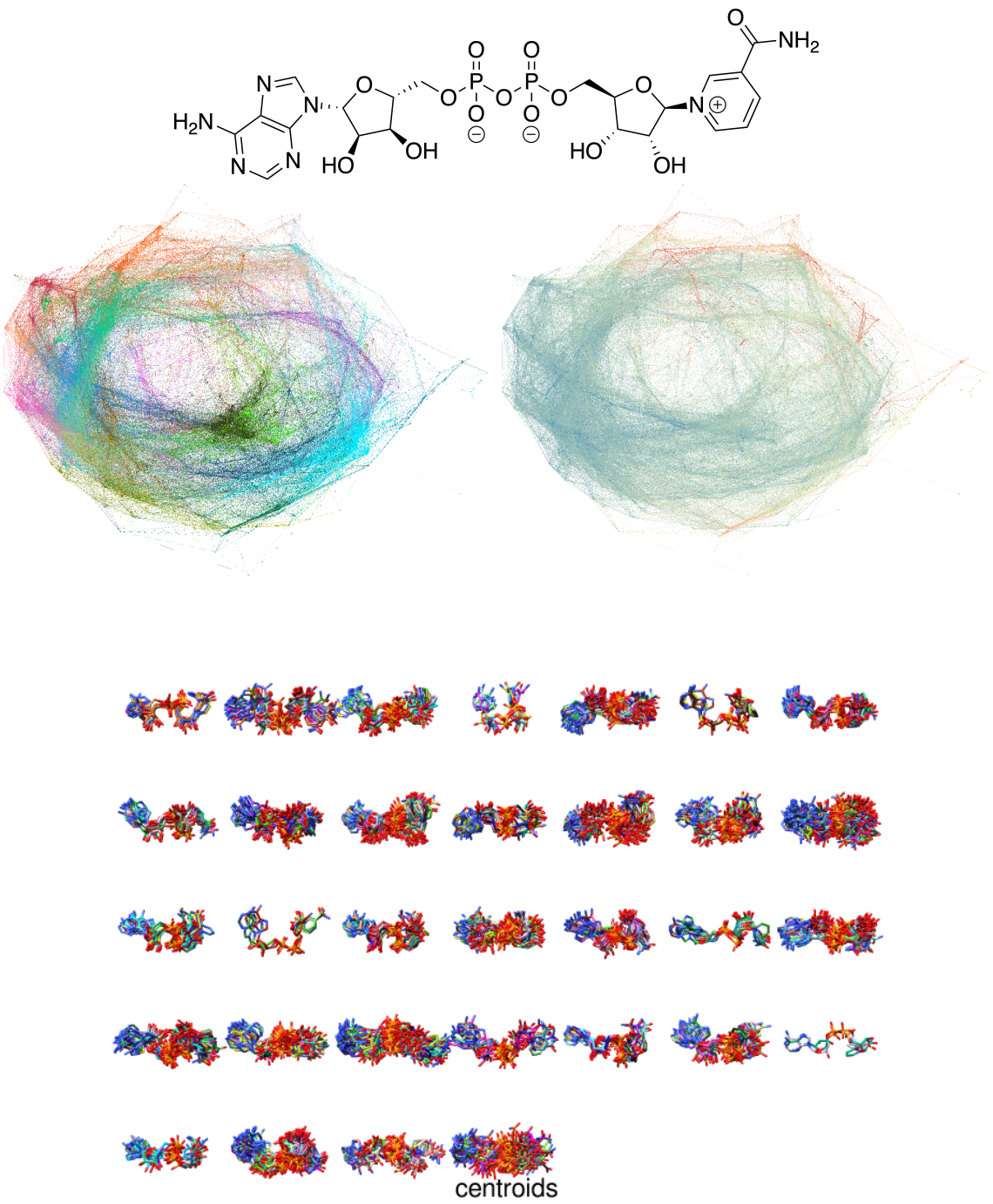

Figure 4: Nicotinamide adenine dinucelotide (top) was subjected to conformer generation, geometry optimization, and clustering by AutoGraph. The conformational graphs colored by cluster (middle left) and single point energy values (middle right) are shown using a gradient of blue (low energy) to red (high energy). Conformers within clusters were superimposed to their centroids (bottom).

Table 1: Variance in energy within clusters (σ^2_{within}) was compared against the variance among all conformers (σ^2) for OSLH and NAD+ individually. The F-test of equality of variances was used to assess the statistical significance of the difference in variances. The numbers of total conformers N and of clusters C are shown.

metabolite	N	C	σ^2	σ^2_{within}	p
$\text{OSLH}_{\text{neutral}}$	485	28	4.59	3.94	0.048
$\text{OSLH}_{\text{charged}}$	485	28	194.32	64.21	$< 0.001^*$
NAD+	785	31	105.67	49.15	$< 0.001^*$

Variances have units of $\left(\frac{kcal}{mol}\right)^2$. p-values were rounded to three decimal places. All other values were rounded to the second decimal place. Tests returning a p-value below 0.005 were considered significant.

Conformational graphs retain geometry/energy correlation for metabolites

The distribution of Spearman correlation coefficient obtained between actual and local weighted energy estimates on conformational graphs from all 200 metabolites were plotted (Figure 5). While the distributions exhibited a large range, the majority of ρ were positive, with median values of 0.36 before and 0.38 after applying the edge weight threshold. This indicates most graphs exhibited a positive monotonic relationship between actual and local estimated energy values to varying degrees. Proximity in the conformational graph therefore translates to similarity in energy. No significant enhancement was observed in applying a threshold to the conformational graphs. We observe energetic information was inferred from the conformational graphs over a large, representative set of metabolites, therefore clustering by the conformational graphs may yield energetically informed partitions of the PES by thermally accessible local minima.

The variance within clusters were determined to be smaller than the total variance for charged OSLH and NAD+. We assessed the difference in cluster-wise and total variance between the 200 benchmark structures. A paired left-tailed Wilcoxon signed rank test determined the median values for $\sigma^2_{within} - \sigma^2$ of $-1.11\left(\frac{kcal}{mol}\right)^2$ was significant ($df = 199,\ p < 0.001$). We should note the effect size is notably smaller than the results observed for charged OSLH and NAD+ (Table 1). While we consistently observe clustering by AutoGraph reduces the variance in energy within clusters relative to the overall variance, the small effect size highlights AutoGraph as a tool to draw preliminary trends from a large number of metabolite conformations to reduce the search space for downstream

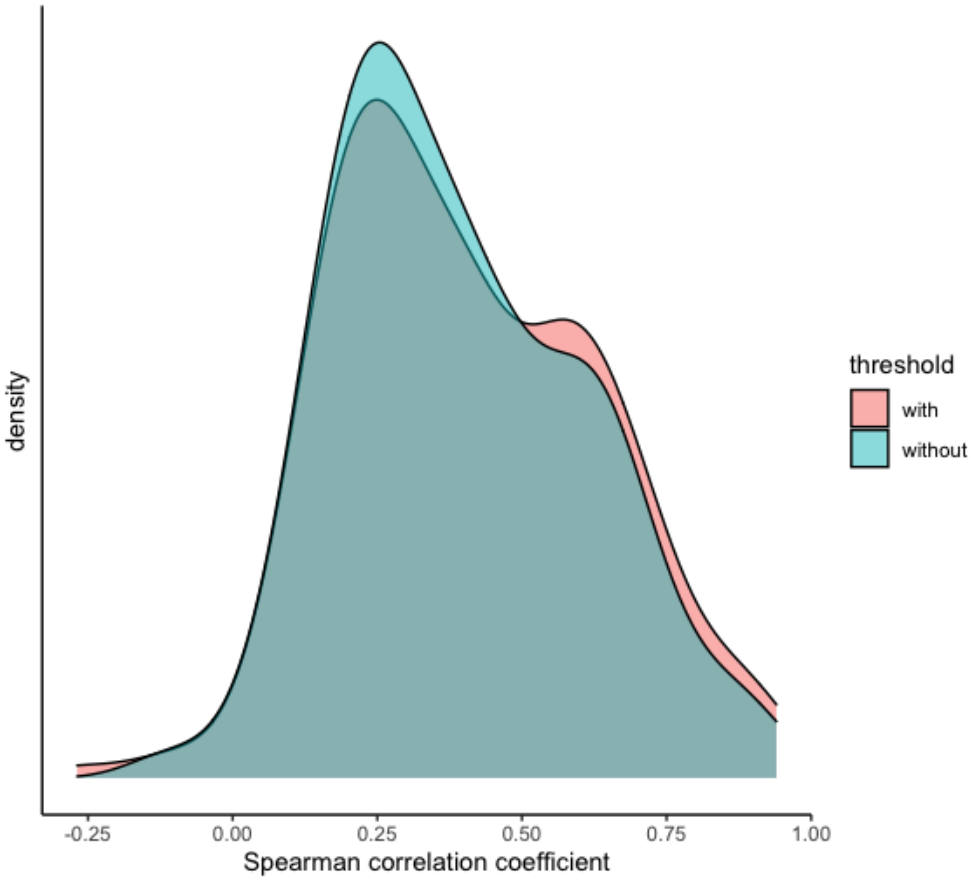

Figure 5: Distribution of Spearman correlation coefficients between actual and local estimated relative energy values by applying a with and without threshold value as part of the AutoGraph protocol to produce the conformational graphs ($n = 200$ for each distribution).

processing and prediction, and not intended to make predictions itself.

Conclusion

We presented use cases for an automated conformational clustering algorithm on OSLH and NAD+. Due to the throughput of ANI-2x potential and high degree of automation of AutoGraph, we could generate, refine, and cluster the conformers for 200 representative metabolites. Further validation of the algorithm as a strategy to obtain an approximation of the underlying conformational ensemble is underway. The AutoGraph protocol can be integrated into computational workflows handling metabolite or other small-molecule conformations using a short Python script, or run as an interactive program with no coding required. We anticipate the application of AutoGraph will narrow down the search space in order to generate a representative conformational ensemble of collections of small-molecules.

Acknowledgement

The authors thank the high-performance computing center (HPCC) at Michigan State University for providing their computational resources.

Supporting Information Available

The following files are available free of charge.

- Supporting Information: Figures and tables mentioned in manuscript.

References

(1) Kubo, R.; McQuarrie, D. A. Statistical Mechanics. *Phys. Today* **1965**, *18*, 74–75.

(2) Amaro, R. E.; Baudry, J.; Chodera, J.; Demir, ; McCammon, J. A.; Miao, Y.; Smith, J. C. Ensemble Docking in Drug Discovery. *Biophys. J.* **2018**, *114*, 2271–2278.

(3) Verma, J.; Khedkar, V.; Coutinho, E. 3D-QSAR in Drug Design - A Review. *CTMC* **2010**, *10*, 95–115.

(4) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386–2396.

(5) Hawkins, P. C. Conformation Generation: The State of the Art. *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

(6) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

(7) Wilson, S. R.; Cui, W.; Moskowitz, J. W.; Schmidt, K. E. Applications of simulated annealing to the conformational analysis of flexible molecules. *J. Comput. Chem.* **1991**, *12*, 342–349.

(8) Hawkins, P. C.; Nicholls, A. Conformer Generation with OMEGA: Learning from the Data Set and the Analysis of Failures. *J. Chem. Inf. Model.* **2012**, *52*, 2919–2936.

(9) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(10) Devereux, C.; Smith, J. S.; Davis, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(11) Folmsbee, D.; Hutchison, G. Assessing conformer energies using electronic structure and machine learning methods. *Int J Quantum Chem* **2020**, *121*, e26381.

(12) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat Commun* **2019**, *10*, 1–8.

(13) Das, S.; Edison, A. S.; Merz, K. M. Metabolite Structure Assignment Using In Silico NMR Techniques. *Anal. Chem.* **2020**, *92*, 10412–10419.

(14) Xu, D.; Tian, Y. A Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci.* **2015**, *2*, 165–193.

(15) Solorio-Fernández, S.; Carrasco-Ochoa, J. A.; Martínez-Trinidad, J. F. A review of unsupervised feature selection methods. *Artif Intell Rev* **2019**, *53*, 907–948.

(16) Shao, J.; Tanner, S. W.; Thompson, N.; Cheatham, T. E. Clustering Molecular Dynamics Trajectories: 1. Characterizing the Performance of Different Clustering Algorithms. *J. Chem. Theory Comput.* **2007**, *3*, 2312–2334.

(17) Yongye, A. B.; Bender, A.; Martínez-Mayorga, K. Dynamic clustering threshold reduces conformer ensemble size while maintaining a biologically relevant ensemble. *J Comput Aided Mol Des* **2010**, *24*, 675–686.

(18) Li, Y.; Dong, Z. Effect of Clustering Algorithm on Establishing Markov State Model for Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2016**, *56*, 1205–1215.

(19) Kim, H.; Jang, C.; Yadav, D. K.; Kim, M.-h. The comparison of automated clustering algorithms for resampling representative conformer ensembles with RMSD matrix. *J Cheminform* **2017**, *9*, 21.

(20) Husic, B. E.; Pande, V. S. Ward Clustering Improves Cross-Validated Markov State Models of Protein Folding. *J. Chem. Theory Comput.* **2017**, *13*, 963–967.

(21) Dubes, R. C. How many clusters are best? - An experiment. *Pattern Recognit.* **1987**, *20*, 645–663.

(22) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, *2008*, P10008.

(23) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Cryst Sect A* **1976**, *32*, 922–923.

(24) R Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria, 2017.

(25) Snedecor, G. W.; Cochran, W. G. Statistical Methods, eight edition. *Iowa state University press, Ames, Iowa* **1989**,

(26) Wilcoxon, F. *Springer Series in Statistics*; Springer New York, 1992; pp 196–202.

(27) Watts, K. S.; Dalal, P.; Murphy, R. B.; Sherman, W.; Friesner, R. A.; Shelley, J. C. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *J. Chem. Inf. Model.* **2010**, *50*, 534–546.

(28) ConfGen. 2020; `https://www.schrodinger.com/Protein-Preparation-Wizard/`, Schrödinger LLC.

(29) Hjorth Larsen, A. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(30) Jacomy, M.; Venturini, T.; Heymann, S.; Bastian, M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS One* **2014**, *9*, e98679.

(31) Wizard. 2017; `https://doi.org/10.5040/9781472500403.article-0000185`, Schrödinger LLC.

(32) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **2013**, *27*, 221–234.

(33) Frisch, M. et al. Gaussian˜16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(34) Ebejer, J.-P.; Morris, G. M.; Deane, C. M. Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.* **2012**, *52*, 1146–1158.

(35) Tosco, P.; Stiefl, N.; Landrum, G. Bringing the MMFF force field to the RDKit: Implementation and validation. *J Cheminform* **2014**, *6*, 37.

(36) Dewar, M. J.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(37) Stewart, J. J. MOPAC: A semiempirical molecular orbital program. *J Computer-Aided Mol Des* **1990**, *4*, 1–103.

(38) Smith, D. G. et al. Psi4 1.4: Open-source software for high-throughput quantum chemistry. *J. Chem. Phys.* **2020**, *152*, 184108.

(39) Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2017**, *46*, D608–D617.

(40) Fessel, J. P.; Oldham, W. M. Nicotine adenine dinucleotides: The redox currency of the cell. 2018.