Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?

Joshua Schrier

Department of Chemistry, Fordham University, 441 East Fordham Road, The Bronx, New York, 10458, USA

* jschrier@fordham.edu

Abstract

Coulomb Matrix Eigenvalues (CME) are a global 3-dimensional representation of molecular structure, which have been previously used to predict atomization energies, prioritize geometry searches, and interpret rotational spectra. The properties of the CME representation and its relationship to molecular structure are established using the Gershgorin circle theorem. Numerical bounds are studied using a dataset of 309,000 conformational samples of all constitutional isomers of acyclic alkanes, C_nH_{2n+2}, from methane (n = 1) to undecane (n = 11), to establish the extent to which the CME preserves chemical intuitions about isomer and conformer similarity, and its ability to distinguish constitutional isomers are performed. Neither supervised nor unsupervised machine learning algorithms can perfectly distinguish constitutional isomers as the molecular size increases, but the misclassification rate can be kept below 1%.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

Introduction

Machine learning is now applied to many chemical problems.[1,2] Using an appropriate representation is a first step for applying machine learning to any problem. Although it is possible to learn representations,[3] it is more common to select a representation and then learn a function that uses it as an input. Many molecular representations are now widely available in general packages such as CheML,[4] DScribe,[5] and ChemReps;[6] as recently reviewed by Langer *et al.*[7] Molecular representations can be divided into those that capture 1D (composition), 2D (molecular graph), 3D (shape), and 4D (averaged over conformations or time) descriptions of the molecular structure. Alternatively, they can be divided into local (in the vicinity around each atom) and global (taking into account the entire molecule) types of representations.[8] Local representations lend themselves to additive expansions of properties in terms of atoms, bonds, or groups, whereas global representations capture the *gestalt* of the molecule as a whole, lending them to non-additive properties and long-range structure.

One global 3D descriptor—introduced in the landmark paper by Rupp *et al.* on machine learning for molecular total energy—is the Coulomb matrix (CM) representation.[9] The elements in the CM representation, M_{ij}, are defined using the atomic number, Z_i, of atom i, and the interatomic separations, R_{ij}, between atoms i and j, as,

$$M_{ij} = \begin{cases} 0.5\, Z_i^{2.4} & \text{for } i = j \\ Z_i Z_j / R_{ij} & \text{for i} \neq \text{j} \end{cases}. \qquad (1)$$

The exponents in the diagonal entries correspond to a polynomial fit relating atomic number to the total energies of the free atoms; the off-diagonal entries describe Coulomb's law like repulsions between the bare nuclei. (To avoid confusion, we note that the CM representation is unrelated to the "Coulomb matrix" of electron-electron integrals used in quantum chemistry calculations.) Although the CM is invariant to molecular translations and rotations, it is not invariant to permutations of the atoms. To

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

address this, a plethora of alternative representation schemes deriving from or extending the CM representation have been proposed such as Random and Sorted Coulomb Matrices,[10] Bag of Bonds (BoB),[11] encoded bond schemes for generating a representation independent of the molecule,[12] and Many-Body Tensor Representations.[13] (Local variants of the CM have been devised,[14] but we are only interested in the global description.)

The CM can be made permutationally invariant by taking the sorted vector of Coulomb matrix eigenvalues (CME). This has the additional advantage of reducing the size of the representation, as an N atom system has N CMEs, rather than the N^2 entries in the full CM. The CME representation was introduced by Rupp *et al.* and used as an input for atomization energy prediction models.[9] The CME has also been used for determining molecular similarity, where it has been employed to perform nearest-neighbor matching of molecules[15] and to prioritize and enforce diversity in genetic algorithm searches for low-energy structures.[16] In very recent work, CME has been used as an intermediate representation for decoding rotational spectra.[17] In this approach, a neural network converts observed rotational spectra to an intermediate CME representation, and then other neural networks deduce chemical formulas, functional group information, and SMILES strings from the intermediate CME representation.

Moussa observed a possible disadvantage of the CME representation in a comment on Rupp *et al*.'s paper.[18] Taking the eigenvalues creates a lossy representation, and the N-dimensional CME vector corresponds to a $2N$-dimensional space of N-atom molecules; these homometric molecules are indistinguishable. For example, the CME of stereoisomers are homometric. For the purposes of total energy calculations, losing this distinction is not important, as the effect of parity violation on chemical energies is exceedingly small for organic molecules.[19] Moussa also described an unphysical distortion of acetylene with homometric CMEs to demonstrate how this representation might fail. However, it is

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

unknown whether this is actually a problem if one limits the consideration to physically reasonable molecular geometries.

The effectiveness of the CME as a molecular description depends on its ability to distinguish different molecules in a meaningful way. The title of this paper alludes to the classic mathematics paper by Kac about whether one can "hear the shape of a drum"—i.e., determine its shape from the eigenvalue spectra.[20] The answer is not obvious. Experience suggests that bongos are distinguishable from timpani, and certain geometrical features such as area and perimeter have unique mappings to eigenspectra, but this question stimulated mathematical research in spectral theory and physical applications of isospectrality,[21] ultimately resulting in constructive methods for generating isospectral drums. Likewise, this paper investigates the representational properties of the CME, its interpretability, and its ability to distinguish constitutional isomers. This provides an opportunity to explore how the CME encodes molecular structure, and the extent to which that encoding preserves chemical intuitions about isomer and conformer similarity.

Computational Methods

The 309 acyclic alkane constitutional isomers C_nH_{2n+2}, from methane (n=1) to undecane (n=11), are used as a specific test case. The goal is not to add to the extensive body of ab initio[22] and machine learning[23,24] work on alkanes, but rather to use alkanes as a simple, yet chemically meaningful example when studying fundamental questions about molecular structure.[25,26] The constitutional isomers are recursively enumerated, and the results are in agreement with Henze & Blair.[27] (See Supporting Information.) Alkanes starting with heptane (n=7) have the possibility for stereogenic centers. Stereoisomerism is not explicitly specified, as this is a known limitation of the CM representation.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

Rather, conformers are initialized with random choices for the stereogenic centers, so that the collection of conformers is essentially a racemic mixture. Severe steric packing problems reduce the stability of saturated alkanes starting at $C_{17}H_{36,}$[28] but computational estimates suggest even the most sterically crowded acyclic alkanes up to $C_{15}H_{32}$ should be stable at room temperature.[26] Therefore, limiting our consideration to undecane (n = 11) and smaller keeps us below this threshold, and sufficient to demonstrate the key findings of the paper.

For each of the 309 constitutional isomers, a sample of 1000 random conformations are generated using the ETKDG algorithm (version 1).[29] Benchmarking studies of conformational sampling algorithms indicate that ETKDG is competitive with other conformational sampling programs.[30] The generated conformations are used to create the CM and CME without energy minimization, so as to explore the possible accessible conformational diversity of each isomer. Therefore, this dataset is distinct from previous studies using the CME representation, in which only a single local minimum was considered for each molecule or isomer, neglecting the conformational variety.[9,17] Furthermore, past CME studies on organic molecules considered the GBD-9 dataset (only up to 9 heavy atoms); whereas the current work considers up to undecane (11 carbons). In addition, the sampled conformations avoid unrealistic bond lengths, bond angles, and steric clashes, and thus differ from Moussa's acetylene example.[18]

Supervised machine classification was performed using logistic regression (LR), decision trees (DT), random forests (RF), gradient boosted trees (GBT), support vector machines (SVM), and k-nearest neighbor (k-NN, with k=1, 3, 5). The prediction task is to distinguish an isomer from all other isomers; the misclassification rate is defined as the fraction of incorrectly assigned isomer labels. Each model was trained and evaluated by five-fold cross validation, dividing the total dataset into five 80%-20% training-

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

testing datasets with equal numbers of each isomer. For example, decane has 75 isomers, with 1000 conformers each; the 5-fold cross validation divides the data into 60,000 training examples (800 of each isomer) and 15,000 test examples (200 of each isomer). The training set was used to determine optimal hyperparameters for each model.

A Mathematica 12.1 notebook implementing all calculations is available in the Supporting Information; this notebook includes interactive versions of the figures showing additional statistical characterization of distributions and interactive browsing of the data.

Results and Discussion

Key properties of the CME representation can be understood using the Gershgorin circle theorem.[31] For a real symmetric matrix (such as the CM), each eigenvalue is within an interval whose center is the diagonal element in the matrix and whose radius is the sum of the absolute values of the entries in the row excluding the diagonal element. Carbon atoms have the largest diagonal entries in the CM, not only because of a larger atomic number, but also exacerbated by the exponent in eq 1. Carbon atoms have a diagonal entry of 36.858, whereas hydrogen atoms have a diagonal entry of only 0.5. Thus, the general trend should be to have *n* large eigenvalues corresponding to the *n* carbon atoms in the alkane. Because C-C bonds contribute the largest off-diagonal elements because of the numerator is largest and the short distances in the denominator of eq. (1), the magnitude of each of these eigenvalues will increase as more carbon atoms are added, because of the off-diagonal contributions. Proximal C-C bonds make a larger contribution than distal C-C pairs, because of the inverse distance dependence of the off-diagonal terms. As a result, a more highly substituted carbon center with more C-C bonds to it will have larger off-diagonal entries in its row, and thus *(ceteris*

paribus) will have larger eigenvalues. These properties have the benefit of making the CME vector emphasize changes in C-C bond patterns, while minimizing the difference caused by small changes, such as rotation of a methyl group. In a large alkane, such as decane (*n* = 10), the typical C-C off diagonal elements range from about 3-25, whereas the typical C-H off-diagonal elements range from 0.5-5.6, and the typical H-H off-diagonal elements range from 0.07-0.57. However, the Gershgorin bounds are not particularly tight, and methods for tightening the bounds by adding a constant matrix[32] are of limited use because of the large difference between the typical C-C and H-H off-diagonal matrix elements. As the underlying reason for the predominance of the carbon on-diagonal and C-C bond off-diagonal terms is the large atomic number, Z_i, relative to hydrogen, this trend will also occur for heteroatom-containing molecules.

Turning to numerical calculations on the sampled conformers, to illustrate the general trend, we first examine the largest CME. In Figure 1, each black point is the mean largest eigenvalue (EV1), averaged over conformers, for each constitutional isomer. The general increase with increasing molecular weight is consistent with the Gershgorin trend. These numerical results show the precise magnitude of the variations possible for different constitutional isomers, indicated by the boxed region indicating the minimum and maximum value observed for each molecular formula. (Methane (*n* = 1) through propane (*n* = 3) lack constitutional isomerism.) Although the EV1 for each isomer are distinct for the shorter alkanes, by heptane (*n* = 7) there begin to be isomers whose EV1 is larger than that of some of the corresponding octane (*n* = 8) values. There are clearly *more* isomers as the alkane size increases, but the CME values are also becoming less distinct. The maximum CME eigenvalues grows more slowly than a linear function of molecule size.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

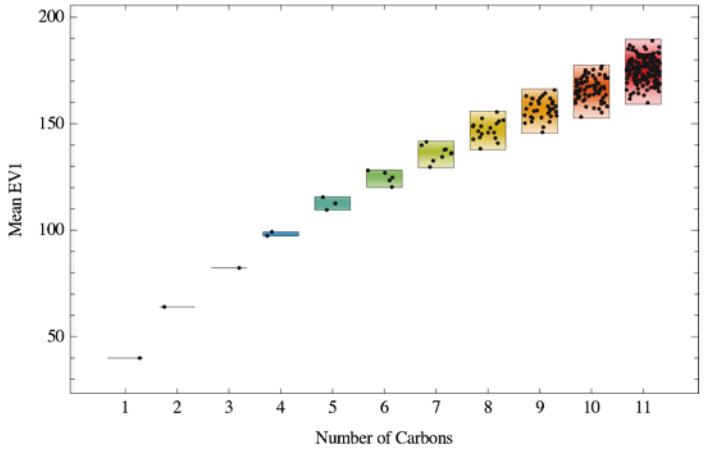


Figure 1: Mean largest eigenvalue (EV1), averaged over 1000 conformers, as a function of molecular size for acyclic alkanes C_nH_{2n+2}. Each black point corresponds to a single constitutional isomer; the boxes indicate the observed minima and maxima.

The results shown in Figure 1 indicate that the maximum CME alone does not distinguish different molecular formulas. However, the molecular formula of acyclic alkanes can be readily determined from the CME vector by counting the number of CME values greater than one. The number of CME entries whose value is greater than 1 is exactly equal to the number of carbon atoms in the saturated alkane for all 309,000 conformers up to n=11 in our study. This is an upper bound on a more general statement for molecules in general. For a random sample of 1000 organic molecules (not restricted to alkanes) from PubChem, 89% have exactly the same number of "CME values greater than 1" as the number of non-hydrogen atoms, and *all* molecules satisfy the looser condition that the number of "CME values greater than 1" is less than or equal to the number of non-hydrogen atoms. Calculation details and structures of the molecules failing the more rigorous criterion are in the electronic Supporting Information file.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

The EV1 is also insufficient to distinguish isomers. Figure 2 shows the distribution of EV1 for the nine isomers of heptane (n = 7); this example is chosen because it has a relatively small number of isomers so that the plot is not too overwhelming. The probability density functions are a Gaussian kernel density estimate using Silverman's rule to determine the bandwidth. The general trend is consistent with the Gershgorin theorem arguments discussed above—n-heptane conformers have the smallest EV1 and 2,2,3-trimethylbutane conformers have the largest—but the overlap of the distributions of EV1 values can be quite significant. For example, the 2,4-dimethylpentane (blue) and 2-ethylpentane (yellow) conformer EV1 values have a large overlap, indicating that EV1 alone is insufficient to distinguish these isomers.

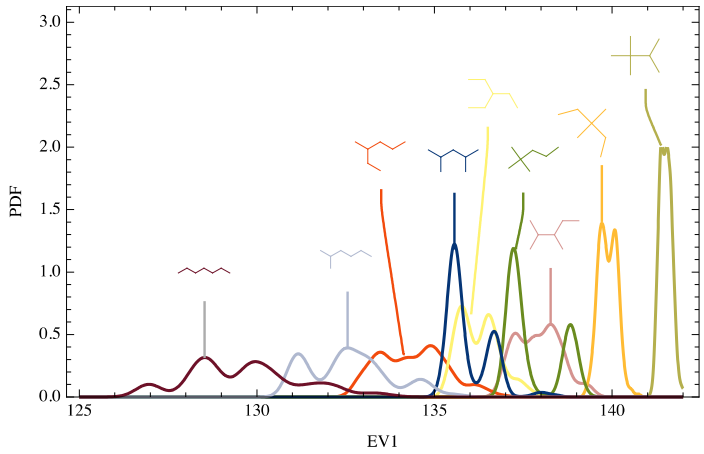


Figure 2: Probability density function (PDF) of the largest coulomb matrix eigenvalue (EV1) of heptane (n=7) isomers.

Expanding the description to include several CMEs makes it possible to distinguish constitutional isomers and conformers. Figure 3 shows the distribution of the first two CMEs (EV1 and EV2) for n-butane (blue) and isobutane (red). The cluster of points associated with each isomer are distinguishable, and the *cis*- and *trans*-like conformers of *n*-butane are also distinct. This is a desirable

behavior, as it suggests the feasibility of hierarchical clustering of nearby constitutional isomer groups followed by larger distinctions between isomers. The eigenvalues pattern in Figure 3 illustrate general considerations from the Gershgorin theorem. Isobutane has a more highly substituted carbon center, resulting in a larger largest eigenvalue. But the second, third, and fourth eigenvalues are smaller, because each methyl-group carbon is farther away from each other). In contrast, only primary and secondary carbons are present in *n*-Butane. In the *trans*-like conformation, the 1- and 4-carbons are farthest from each other, reducing the off-diagonal CM elements (and thus giving rise to the left cluster of points), whereas in the *cis*-like conformation, they are closer (giving rise to the right cluster of points)

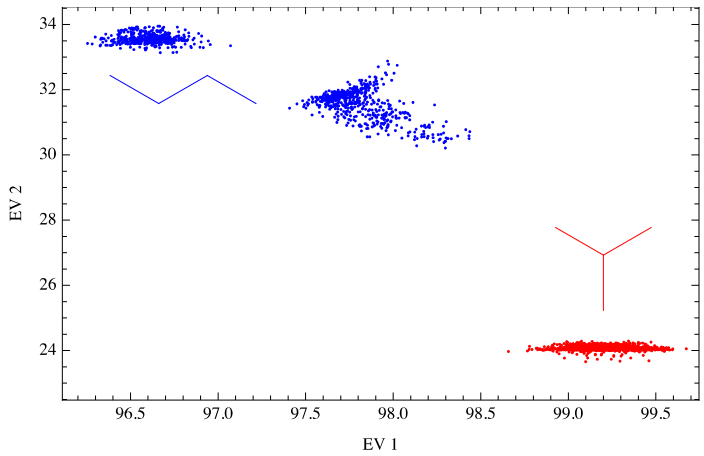


Figure 3: First and second CME for butane. Blue and red points indicate *n*-Butane and isobutane, respectively (depicted by the molecular diagrams).

Principal components analysis (PCA) can be used to determine the amount of information contained in the CME vector. Figure 4 depicts the number of principal components required to describe 99% of the conformer variance for each isomer. Recall that a *N=3n+2* atom molecule has an *N*-dimensional CME vector description. Therefore, it is unsurprising that methane (*n* = 1, *N* = 5*)* requires 3 components to describe the vector. However, the number of principal components grows more slowly

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

than the number of carbon atoms, *n*. For octane (*n*=8) and above, at most *n-2* PCA components are needed for any isomer, and nonane (*n=9)* and above each have at least one isomer for which only four components suffice. These are all smaller than the *n* components needed to describe the carbon atom chain. This demonstrates that each isomer's CME vectors are highly correlated, and the information contained in the CME vector is much less than the full (*N=3n+2)* dimensionality.

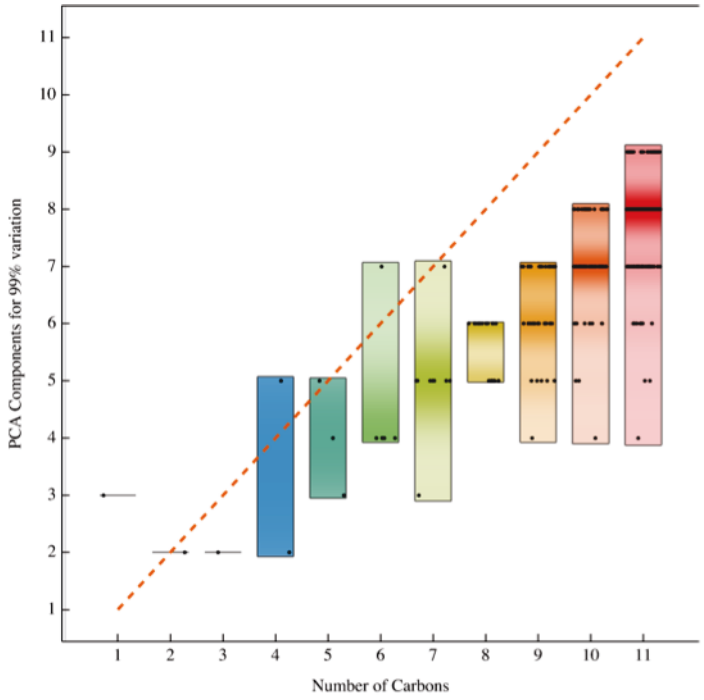


Figure 4: Number of principal components describing at least 99% of the variation in the CME vector. Black points indicate values for each constitutional isomer; boxes indicate observed minima and maxima; hues indicate the density of points at that value. The dotted red bisectrix is a guide to the eye for the linear scaling with number of carbons, *n*.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

If provided with a set of CME vectors (without knowing their true identities), could they be separated into subsets corresponding to each isomer? This is the problem of *clustering*, a form of unsupervised machine learning used to find similarities between unlabeled examples to identify common characteristics and groupings.[33] Clustering is widely used in cheminformatics analysis, e.g., to determine the set of features associated with activity of a drug candidate molecule.[34] An ideal cluster consists of all conformers of a specific isomer. An *intracluster* comparisons is between conformers of the same isomer, and an *intercluster* comparison is between conformers of different isomers. For simplicity, we consider only intercluster comparisons between isomers of the same molecular weight. Because we have the correct labels, we can assess the clusterability of the data, independent of any particular algorithm.

One measure of clusterability is the Dunn index, defined as the ratio of the intercluster *separation* to the intracluster *diameter*, where *separation* is the minimum intercluster distance (considered over all pairs of items in the two clusters), and *diameter* is the maximum intracluster distance (considered over any pair of items in the same cluster).[35] This provides a single value for each collection of data (i.e., all of the isomers and conformers of a given molecular weight). A larger Dunn index indicates compact and well separated clusters that are easier to divide into the proper grouping. The Dunn index approaches zero for unclusterable datasets. Figure 5 shows the Dunn index as a function of carbon atoms, computed using the Manhattan (L_1), Euclidean (L_2), and Chebyshev (L_∞) metrics. (As noted earlier, only butane and above have more than one isomer, so methane, ethane, and propane are not shown in this or subsequent figures.) Results for the three distance metrics are qualitatively the same, but the Euclidean (L_2) metric is slightly larger in all cases, indicating a better separation of clusters, which justifies using the Euclidean metric in subsequent calculations. The Dunn index decreases as the molecular size increases, indicating that isomers of larger alkanes have less

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

distinct CME vectors. Consequently, it becomes more difficult for any unsupervised clustering algorithm separate unlabeled isomers as the molecular size increases.

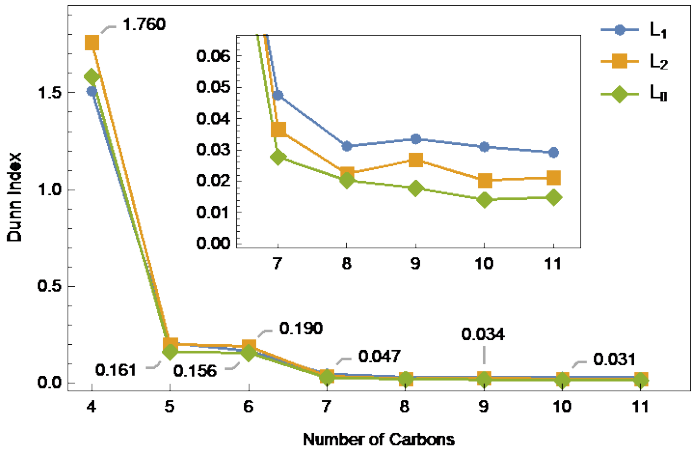


Figure 5: Dunn Index as a function of molecular size for acyclic alkanes. Numerical values for the Euclidean (L_2) metric are labeled. Inset shows the same data focusing on alkanes with 7 and more carbons atoms.

The Silhouette index is another way to quantify the intrinsic clusterability and separability of a dataset.[36] Unlike the Dunn index, which assigns a value to the entire collection of data (i.e., all isomers of a given molecular weight), a Silhouette index is assigned to *each* individual data item (i.e., each conformer). The Silhouette index is defined using each item's average *intracluster* distance, a_i (taken over all other items in the same cluster), and its average *intercluster* distance, b_i, (taken over every point in the other cluster). When there are multiple clusters (i.e., isomers), the cluster with the smallest average intercluster distance is used (i.e., only consider the nearest isomer for the intercluster distances. Finally, the Silhouette index is given by $S_i = (b_i - a_i)/\max(a_i, b_i)$. Items with $S_i \approx 1$ are in a well separated cluster, where the average intercluster distance is much greater than the average intracluster distance. Items with $S_i \approx 0$ are located between two clusters, and those with negative S_i are closer (on

average) to points in a different cluster than to points in its own cluster. Points with negative S_i will be misclassified by unsupervised algorithms, and a large fraction of examples with negative silhouette scores indicates the difficulty of unsupervised classification for the dataset.

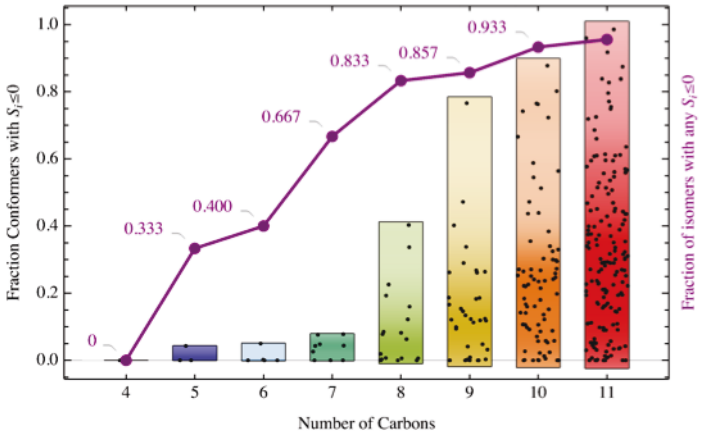


Figure 6: Negative Silhouette index occurrence as a function of molecular size. Black points indicate the fraction of conformers with a negative Silhouette index, S_i, for each isomer. Bars indicate the observed minima and maxima, and color intensity shows the density of occurrences. The purple line shows the fraction of isomers having any negative Silhouette index conformers.

Figure 6 summarizes the silhouette index results for alkane isomers, using the Euclidean distance. Only butane (n = 4) has all perfectly separable isomers. Although all alkanes have *some* isomers without any negative S_i, both the median and maximum fraction of conformers with negative S_i grows monotonically as a function of size. For example, 4.3% of n-pentane conformers have negative S_i, and two isomers of hexane have negative Silhouette scores: 5% of n-hexane conformers, and 0.2% of 3-methylpentane conformers with negative S_i. Considering the n-alkane series, the fraction of conformers with negative S_i increases monotonically with chain length, growing to 20% of n-undecane conformers. This is less than most other undecane isomers: 152/159 isomers have at least one point with negative S_i, and the median undecane isomer has 29% of its conformers with a negative S_i. The worst case is 4,4,5-

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

trimethyloctane for which 98.6% conformers have a negative S_i. These results indicate that unsupervised classification algorithm will fail to distinguish unlabeled isomers using the CME vectors. Preliminary calculations using common unsupervised classification algorithms (DBSCAN, Spectral clustering, and k-Means clustering) failed to give any meaningful clusters, but instead separated conformers and merged different isomers, even when provided with the correct total number of isomer clusters to find. This suggests the need for caution when using CMEs for diversity-oriented exploration, like that described in ref [16]. If only the CMEs are used to identify whether a compound is "novel" or not, there is a high likelihood that some isomers will be excluded because of a similarity to conformers of different isomers.

Although unsupervised classification may fail, *supervised* classification algorithms—specifically, logistic regression (LR), decision trees (DT), random forests (RF), gradient boosted trees (GBT), support vector machines (SVM), and k-nearest neighbor (*k*-NN, with *k*=1, 3, 5)—are more successful at labeling a CME to the correct isomer. The results of five-fold cross-validation for the isomer distinguishing task for decane (*n* = 10) are shown in Figure 7; corresponding figures for butane (*n = 4)* through nonane (*n = 9)* are shown in the Supporting Information. First, we will discuss models using the entire *N*-dimensional CME vector as input. All model types perfectly classify butane (*n = 4)* isomers. LR, SVM and the k-NN models also have perfect classification across all folds for pentane (*n* = 5) and hexane (*n = 6)*, whereas DT, RF, and GBT models only achieve this for some of the folds. For the *n*=7-9 alkanes, there is at least one fold for which the SVM model achieves a zero classification error, but the median misclassification rate is non-zero over the folds. Consistent with the Dunn Index and Silhouette Index analyses (*vide supra*), nearest-neighbor approaches have limited ability to separate the conformers and misclassify about 1.5% of examples, which is worse than all other methods (except decision trees) for all alkanes. Surprisingly, the simple LR classifier is competitive with RF and GBT methods for all alkanes, and all of these (with the exception of DT) achieve better performance than the k-NN models, although still the

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

error rates are greater than the SVM. Overall, the SVM classifier delivered the best performance (i.e., lowest misclassification rate) for all alkanes. In contrast, DT has the highest misclassification rate for all alkanes, and its performance decreases with increasing molecule size. Focusing on the decane (n = 10) case shown in Figure 7a, the SVM only misclassifies 16-28 of the 15,000 test items in the 75-way classification task; the median misclassification rate is merely 0.15%, distributed across different isomers. In general, the errors rates of all other classifiers perform about an order of magnitude worse, about 1% error.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

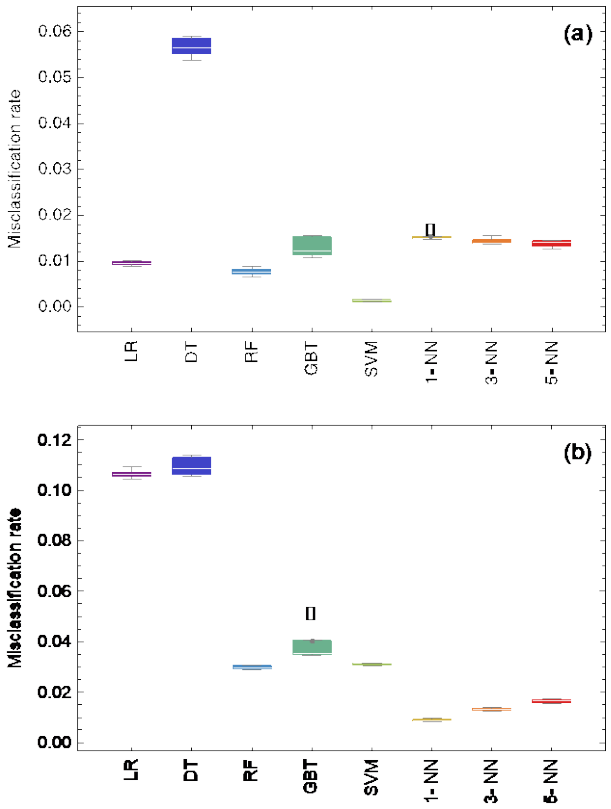


Figure 7: Misclassification error rate for supervised classification of decane (*n* = 10) isomers; errors bars indicate the results of 5-fold cross validation; solid boxes indicate the 25% and 75% quantiles, and whiskers show the minima and maxima, and white lines indicate the median. (a) Using the full 32-dimensional CMEs for each example; (b) using the 10 largest CMEs.

The largest *n* CME values are dominated by information about the carbon atoms (*vide supra*), and that the information is contained in a subset of the full CME vector. This suggests that lower-dimensional inputs suffice to distinguish isomers. To test this hypothesis, a parallel set of machine learning calculations was performed using only the largest *n* CMEs; results are shown in Figure 7b. The different vertical axis scales of Figures 7a and 7b indicate that truncating the CME vector dramatically increases the misclassification error rate, except for 1-NN (whose median error rate *decreases* to 0.86%

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

from 1.5% with the full CME vector), and 3-NN (decreasing to 1.3% from 1.4%). The improved performance of the 1-NN and 3-NN methods is unsurprising, as exemplar-based models such as *k*-NN perform best for low dimensional problems. (The intuition is that the ratio between the closest (1-NN) distance and the average distance between random examples decreases rapidly as the number of dimensions is increased, so the predictive power of the nearest example decreases relative to any average example in the dataset.) Reducing the dimensionality to the 10 largest CMEs preserves information primarily about the C-C bonding structure, allowing for a successful classification. The error rate for 5-NN is about the same (1.6%) as with the full CME vector (1.4%), and for all other models the error rate is increased dramatically, sometimes by an order of magnitude. The RF and SVM models achieve median error rates of 2.9% and 3.1%, respectively, about thrice the error rate of the simple 1-NN model. This is surprising, as it indicates that the successes of the SVM and RF models using the full CME vector (depicted in Figure 7a) depend upon the information in the *2n+2* deleted entries dominated by H-H interactions. This use of unexpected information is reminiscent of recent machine-learning based infrared spectroscopic interpretation algorithms that use the fingerprint region in addition to the typical functional group peaks taught to human chemists.[37]

Conclusion

This paper examined the properties of the CME representation and its ability to distinguish realistic, chemically plausible molecular structures. The central result is that the CME representation does not perfectly distinguish constitutional isomers with 10-heavy atoms or more. This has several important consequences for cheminformatics. First, it helps rationalize limitations of CME-based machine learning for spectral interpretation. For example, neural networks can successfully deduce molecular formulas from an intermediate CME representation, but cannot reliably extract molecular structures (i.e.,

distinguish different constitutional isomers).[17] The present study indicates that this limit is inevitable as the molecular size increases, and may not be solved by a better model, but instead is because of a limitation in the underlying CME representation itself. This is also consistent with Langer *et al*.'s observations that local representations perform better than global representations for molecular energy prediction as system size increases.[7] Second, the CME representation has been used to prioritize diversity-oriented searches for global minima (e.g., of Lennard-Jones clusters[16]). This is of broad relevance as conformational sampling is important for generating augmented datasets for cheminformatic-based drug design[38] and as a benchmark for computational chemistry methods.[39] The present study confirms that the CME representation preserves the general closeness of related isomers (c.f., Figure 2) and related conformers (c.f*.,* Figure 3), even if it fails to perfectly distinguish isomers for larger species. Therefore, CME may suffice for roughly guide sampling a diverse space, without perfectly distinguishing the species. For this purpose, the PCA results (c.f., Figure 4) and the results of 1-NN classification using the largest *n* CMEs only (c.f., Figure 7b) suggest that the entire CME vector is unnecessary, and restricting the representation to the largest eigenvalues for the heavy (non-hydrogen) atoms encodes sufficient conformation and hydrogen position in a more compact form. This can enable a reduction in storage space and computing time. Third, the large difference in magnitudes of the carbon and hydrogen entries in the CM diminish the influence of hydrogen atom positions. The original motivation was to make the diagonal terms correspond to atomization energies, as this was the original prediction task in Rupp *et al.*[9] Although this captures the intuition that the heavy-atom framework defines the molecule, it results in the large-value CMEs being dominated by C-C bond information. That information is sufficient for a 1-NN classifier, but in general, machine learning models can take advantage of information present in the small components. This suggests that alternate functional forms for the CM that emphasize hydrogen atom positions might improve machine learning model performance. Variations proposed to date, such as the London matrix,[40] *deemphasize* the long-distance

components and hydrogens, and thus may be exactly the opposite of what is desired for this purpose. Finally, although this paper focused on alkanes, the analysis using the Gershgorin theorem and the numerical sample from PubChem described here both suggest that these key results are equally relevant to more general classes of molecules.

Associated Content

Supporting Information

A Mathematica 12.1 computational notebook implementing the calculations described in this paper is available at [ACS Information], and the same notebook has been deposited at https://notebookarchive.org/2020-06-1h9pgc7

Acknowledgements

I thank Rodolfo Keesey for a careful reading of the manuscript and acknowledge support from the National Science Foundation (DMR-1709351) and the Henry Dreyfus Teacher-Scholar Award (TH-14-010).

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

References

(1) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555. https://doi.org/10.1038/s41586-018-0337-2.

(2) Dral, P. O. Quantum Chemistry in the Age of Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 2336–2347. https://doi.org/10.1021/acs.jpclett.9b03664.

(3) Chuang, K. V.; Gunsalus, L. M.; Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2020**, acs.jmedchem.0c00385. https://doi.org/10.1021/acs.jmedchem.0c00385.

(4) Haghighatlari, M.; Vishwakarma, G.; Altarawy, D.; Subramanian, R.; Kota, B. U.; Sonpal, A.; Setlur, S.; Hachmann, J. ChemML: A Machine Learning and Informatics Program Package for the Analysis, Mining, and Modeling of Chemical and Materials Data. *WIREs Comput. Mol. Sci.* **2020**. https://doi.org/10.1002/wcms.1458.

(5) Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of Descriptors for Machine Learning in Materials Science. *Comput. Phys. Commun.* **2020**, *247*, 106949. https://doi.org/10.1016/j.cpc.2019.106949.

(6) Folmsbee, D.; Upadhyay, S.; Dumi, A.; Hiener, D.; Mulvey, D. *Chemreps/Chemreps: Molecular Machine Learning Representations*; Zenodo, 2019. https://doi.org/10.5281/ZENODO.3333855.

(7) Langer, M. F.; Goeßmann, A.; Rupp, M. Representations of Molecules and Materials for Interpolation of Quantum-Mechanical Simulations via Machine Learning. **2020**. *arXiv:2003.12081*

(8) Leach, A. R.; Gillet, V. J. *An Introduction to Chemoinformatics*, Rev. ed.; Springer: Dordrecht, 2007.

(9) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301. https://doi.org/10.1103/PhysRevLett.108.058301.

(10) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419. https://doi.org/10.1021/ct400195d.

(11) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331. https://doi.org/10.1021/acs.jpclett.5b00831.

(12) Collins, C. R.; Gordon, G. J.; von Lilienfeld, O. A.; Yaron, D. J. Constant Size Descriptors for Accurate Machine Learning Models of Molecular Properties. *J. Chem. Phys.* **2018**, *148*, 241718. https://doi.org/10.1063/1.5020441.

(13) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. **2018**. *arXiv:1704.06439*

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

(14) Barker, J.; Bulin, J.; Hamaekers, J.; Mathias, S. LC-GAP: Localized Coulomb Descriptors for the Gaussian Approximation Potential. In *Scientific Computing and Algorithms in Industrial Simulations*; Griebel, M., Schüller, A., Schweitzer, M. A., Eds.; Springer International Publishing: Cham, 2017; pp 25–42. https://doi.org/10.1007/978-3-319-62458-7_2.

(15) Nguyen, V.-D.; Khiet, L. D.; Lam, P. T.; Chi, D. H. Chemical Bond-Based Representation of Materials. **2017**. *arXiv:1712.01663*

(16) Dittner, M.; Hartke, B. Conquering the Hard Cases of Lennard-Jones Clusters with Simple Recipes. *Comput. Theor. Chem.* **2017**, *1107*, 7–13. https://doi.org/10.1016/j.comptc.2016.09.032.

(17) McCarthy, M.; Lee, K. L. K. Molecule Identification with Rotational Spectroscopy and Probabilistic Deep Learning. *J. Phys. Chem. A* **2020**, *124*, 3002–3017. https://doi.org/10.1021/acs.jpca.0c01376.

(18) Moussa, J. E. Comment on "Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning." *Phys. Rev. Lett.* **2012**, *109*, 059801. https://doi.org/10.1103/PhysRevLett.109.059801.

(19) Bast, R.; Koers, A.; Gomes, A. S. P.; Iliaš, M.; Visscher, L.; Schwerdtfeger, P.; Saue, T. Analysis of Parity Violation in Chiral Molecules. *Phys Chem Chem Phys* **2011**, *13*, 864–876. https://doi.org/10.1039/C0CP01483D.

(20) Kac, M. Can One Hear the Shape of a Drum? *Am. Math. Mon.* **1966**, *73*, 1. https://doi.org/10.2307/2313748.

(21) Giraud, O.; Thas, K. Hearing Shapes of Drums: Mathematical and Physical Aspects of Isospectrality. *Rev. Mod. Phys.* **2010**, *82*, 2213–2255. https://doi.org/10.1103/RevModPhys.82.2213.

(22) Karton, A.; Gruzman, D.; Martin, J. M. L. Benchmark Thermochemistry of the C_nH_{2n+2} Alkane Isomers (n = 2−8) and Performance of DFT and Composite Ab Initio Methods for Dispersion-Driven Isomeric Equilibria. *J. Phys. Chem. A* **2009**, *113*, 8434–8447. https://doi.org/10.1021/jp904369h.

(23) Yalamanchi, K. K.; van Oudenhoven, V. C. O.; Tutino, F.; Monge-Palacios, M.; Alshehri, A.; Gao, X.; Sarathy, S. M. Machine Learning to Predict Standard Enthalpy of Formation of Hydrocarbons. *J. Phys. Chem. A* **2019**, *123*, 8305–8313. https://doi.org/10.1021/acs.jpca.9b04771.

(24) Li, G.; Hu, Z.; Hou, F.; Li, X.; Wang, L.; Zhang, X. Machine Learning Enabled High-Throughput Screening of Hydrocarbon Molecules for the Design of next Generation Fuels. *Fuel* **2020**, *265*, 116968. https://doi.org/10.1016/j.fuel.2019.116968.

(25) Goodman, J. M. What Is the Longest Unbranched Alkane with a Linear Global Minimum Conformation? *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 876–878. https://doi.org/10.1021/ci9704219.

(26) Nalin de Silva, K. M.; Goodman, J. M. What Is the Smallest Saturated Acyclic Alkane That Cannot Be Made? *J. Chem. Inf. Model.* **2005**, *45*, 81–87. https://doi.org/10.1021/ci0497657.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

(27) Henze, H. R.; Blair, C. M. The Number of Isomeric Hydrocarbons of the Methane Series. *J. Am. Chem. Soc.* **1931**, *53*, 3077–3085. https://doi.org/10.1021/ja01359a034.

(28) Davies, R. E.; Freyd, P. J. $C_{167}H_{336}$ Is the Smallest Alkane with More Realizable Isomers than the Observed Universe Has "Particles." *J. Chem. Educ.* **1989**, *66*, 278. https://doi.org/10.1021/ed066p278.

(29) 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. https://doi.org/10.1021/acs.jcim.5b00654.

(30) Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators. *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728. https://doi.org/10.1021/acs.jcim.7b00505.

(31) Weisstein, E. W. Gershgorin Circle Theorem https://mathworld.wolfram.com/GershgorinCircleTheorem.html (accessed May 29, 2020).

(32) DeVille, L. Optimizing Gershgorin for Symmetric Matrices. *Linear Algebra Its Appl.* **2019**, *577*, 360–383. https://doi.org/10.1016/j.laa.2019.04.034.

(33) Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*, 4. ed.; Elsevier Acad. Press: Amsterdam, 2009.

(34) Voicu, A.; Duteanu, N.; Voicu, M.; Vlad, D.; Dumitrascu, V. The Rcdk and Cluster R Packages Applied to Drug Candidate Selection. *J. Cheminformatics* **2020**, *12*, 3. https://doi.org/10.1186/s13321-019-0405-0.

(35) Dunn, J. C. Well-Separated Clusters and Optimal Fuzzy Partitions. *J. Cybern.* **1974**, *4*, 95–104. https://doi.org/10.1080/01969727408546059.

(36) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

(37) Fine, J. A.; Rajasekar, A. A.; Jethava, K. P.; Chopra, G. Spectral Deep Learning for Prediction and Prospective Validation of Functional Groups. *Chem. Sci.* **2020**, *11*, 4618–4630. https://doi.org/10.1039/C9SC06240H.

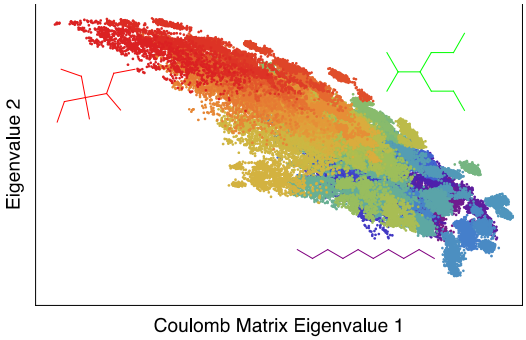
(38) Hemmerich, J.; Asilar, E.; Ecker, G. F. COVER: Conformational Oversampling as Data Augmentation for Molecules. *J. Cheminformatics* **2020**, *12*, 18. https://doi.org/10.1186/s13321-020-00420-z.

(39) Folmsbee, D.; Hutchison, G. *Assessing Conformer Energies Using Electronic Structure and Machine Learning Methods*; preprint; 2020. https://doi.org/10.26434/chemrxiv.11920914.v2.

(40) Huang, B.; von Lilienfeld, O. A. Understanding Molecular Representations in Machine Learning: The Role of Uniqueness and Target Similarity. *J. Chem. Phys.* **2016**, *145*, 161102. https://doi.org/10.1063/1.4964627.

Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631

Graphical Table of Contents



Accepted manuscript: J. Schrier, "Can one Hear the Shape of a Molecule (from its Coulomb Matrix Eigenvalues)?" *J. Chem. Inf. Model.* (2020) doi:10.1021/acs.jcim.0c00631