THE JOURNAL OF PHYSICAL CHEMISTRY A
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCA

Article

Algebraic Graph-Based Machine Learning Model for Li-Cluster Prediction

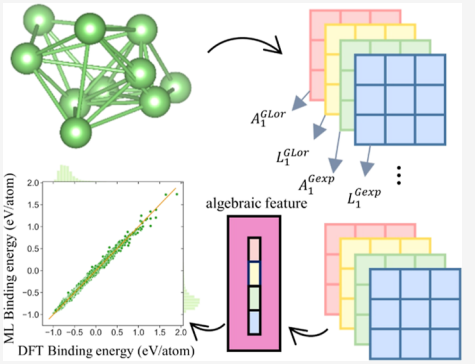
Shengming Ma, Shisheng Zheng, Wentao Zhang, Dong Chen,* and Feng Pan*

Cite This: *J. Phys. Chem. A* 2023, 127, 2051−2059

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In cluster research, determining the ground-state structure of medium-sized clusters is hindered by a large number of local minimum on potential energy surfaces. The global optimization heuristic algorithm is time-consuming due to the use of DFT to determine the relative size of the cluster energy. Although machine learning (ML) is proved to be a promising way to reduce the DFT computational costs, a suitable method to represent clusters as input vectors is one of the bottlenecks in the application of ML to cluster research. In this work, we proposed a multiscale weighted spectral subgraph (MWSS) as an effective low-dimension representation of clusters and build an MWSS-based ML model to discover the structure−energy relationships in lithium clusters. We combine this model with the particle swarm optimization algorithm and DFT calculations to search for globally stable structures of clusters. We have successfully predicted the ground-state structure of Li_{20}



INTRODUCTION

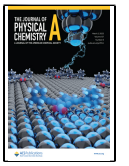
Serving as a bridge between crystals and isolated atoms, clusters are widely studied to unveil the relationship between structure and property in materials science.[1] Determining the ground-state structure of medium and large clusters is a crucial and difficult task in cluster physics. The atomic coordinates are an important factor affecting the binding energy. Thus, the above problem can be considered as the optimization problem of atomic coordinates. Unfortunately, algorithms based entirely on gradient descent generally could not perform very well in the above tasks because there is a large number of local minima on the potential energy surface of such clusters.[2] Therefore, it is of practical significance to search for (or predict) structures with global minimum energy or near minimum energy for clusters.

To efficiently achieve the above objectives, several optimization algorithms including heuristic algorithms are proposed, such as lattice-based search,[3] stochastic surface walking,[4] genetic algorithm,[5−7] simulated annealing,[8] and particle swarm search (PSO).[9] Among all these algorithms, PSO is a relatively efficient one in searching for near-ground-state clusters. Call et al. predicted the global minimum structures of LJ26, $Si2H_5^-$, and $OH^-(H2O)_3$ chemical systems by PSO.[10] Ma et al. successfully developed a PSO-based procedure and finished the structure prediction task on clusters.[9,11] However, the above methods are usually computationally expensive since they involve first-principles calculations of numerous local minimums to determine the formation energy of clusters. Fortunately, machine learning (ML) is expected to provide an efficient approach for structural prediction due to its ability to predict the formation energy of materials.[12]

ML is a statistical method that can extract the pattern of input data. In recent years, ML has achieved great success in various fields, including computer vision, natural language processing, video processing, etc.[13] It seems feasible to train the model to understand the structure−energy relationship pattern. When applying the ML model, feature engineering, which aims to transform the representation of data, is naturally introduced. An appropriate data representation or descriptor is helpful for the learning algorithm to complete the task.

In materials science, descriptors are designed around Cartesian coordinates, atomic types, and physical properties.[14−16] The most straightforward representation of a cluster is the Cartesian coordinates and the element type of each atom in the cluster. This representation loses the information of the translational symmetry and rotational symmetry of the cluster. Feeding a cluster that has been translated or rotated into the model may give inconsistent results. In addition, if the above representation is adopted, the output of the model depends on the order of atomic coordinates inputted.[14] Therefore, we need a feature engineering method that maps multiple inputs with symmetry to the same representation. Moreover, our

Received: January 13, 2023
Revised: February 3, 2023
Published: February 21, 2023



https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

representation needs to have a relatively low dimension to avoid the "curse of dimensionality" in ML. Faced with the above constraints, many effective ML methods have been proposed, such as Gaussian-approximation potentials with the SOAP descriptor,[17] crystal-graph convolutional network (CGCNN),[18] moment tensor potential,[19] atomic cluster expansion,[20] equivariant message-passing NN,[21] and non-parametric many-body force fields.[22] Meanwhile, graph theory, a prime subject of discrete mathematics, provides an effective method.

Graph theory models real things as nodes, or vertices, and pairwise relationships between nodes (i.e., edges), which has been widely used in social network analysis,[23] web page ranking,[24] landscape connectivity,[25] and so on.

In the fields of chemistry and materials science, many problems are naturally suitable for processing with graph theory by treating atoms as nodes and the interaction between atoms as edges.[26−29]

Graph theory has different branches in mathematics, such as geometric graph, algebraic graph, and topological graph.[30,31] Geometric graph theory focuses on graphs drawn on the Euclidean plane and their geometry property. Algebraic graph theory is the combination of graph theory and linear algebra. It studies graphs via eigenvalues, eigenvectors, and characteristic polynomials of the adjacency matrix and the Laplace matrix of a graph. Topological graph theory concerns the embeddings and immersions of graphs. As for application, geometric graph theories are relatively concise and easy to handle. However, treating a graph as a geometric structure is not conducive to numerical calculations. Moreover, the elements in the traditional adjacency matrixes are either 0 or 1, which represents whether the two nodes are connected or not. Such a representation oversimplifies the complex interaction between atoms. For example, it cannot estimate the interaction strength between pairwise atoms. To overcome the above shortcomings, different techniques are proposed. Weighted graphs using radial basis functions to represent graph edges were proposed based on the flexible rigidity index (FRI).[32−34] Mathematically, weighted graphs are complete graphs, and the weight of edges increases with the decrease of the Euclidean distance between two nodes. The physical interpretation of such weighted edges is that the closer the Euclidean distance between the pairwise atoms, the stronger the interaction between them and the greater the weight reflected on the weighted graph. Additionally, multiscale FRI was introduced into weighted graphs. This technique extends the weighted graph to a multiplex graph with multiple edges, giving it the ability to capture multi-scale interatomic interactions.[35] Multi-scale technology has been proven to be feasible in expressing the intramolecular and intermolecular interactions of bio-molecules, including electrostatic interactions, van der Waals interactions, hydrogen bond interactions, etc.[32,36,37]

This work aims to introduce a multiscale weighted spectral graph (MWSS) as an efficiently low-dimensional representation of cluster structures. A series of subgraphs, adjacency matrices, and Laplace matrices are constructed from a single cluster, and the statistics of eigenvalues are used to depict the structural information of that cluster. An ML model is constructed to understand the structure−energy relationship of clusters based on the MWSS descriptor, which is called spectral graph learning (SGL). For the usual force field studies, the model is required to predict the energy of the system and the force on each atom, but for PSO, the model needs to predict and compare the energy of different clusters without force information. Thus, we are more concerned with the energy (or stability) of the cluster, and the model here does not yield forces, only energies. Furthermore, an SGL−PSO−density functional theory (DFT) system is constructed by the combination of SGL, PSO, and DFT calculation to predict structures with global minimum energy or near minimum energy for clusters. Over the past few decades, lithium-ion batteries have been widely used in various portable electronic devices.[38] The formation of lithium dendrites during charging and discharging can cause a short circuit in the battery and bring about safety issues.[39] Thus, Li clusters are used to test in this article. Trained on Li clusters with 3−10 atoms, our model can not only predict the binding energies of clusters of the same number of atoms but also rank the binding energies of clusters with more atoms. By using our model to replace the DFT calculation in the PSO algorithm, we successfully predicted the ground-state structure of Li_{20}. Our work lays the foundation for understanding lithium nucleation and dendrite growth by computational simulation.

METHOD

Multiscale Weighted Geometric Subgraph. In the present work, we focus on the SGL in the representation of pairwise interactions within lithium clusters. For a given Li cluster with N atoms, a graph $G(V,E)$ can be defined by regarding the atoms as the vertices and regarding the interactions between atoms as the edges. The vertex set of G can be defined as follows

$$V = \{r_i | r_i \in R^3; \qquad i = 1,2,3, ..., N\} \tag{1}$$

where V is the set containing N vertices and r_i is the coordinates of the ith atom. All the atoms are uniquely identified by its coordinate so that any atom within the cluster can be considered individually. In order to keep the physical meaning of the interaction, the weighted edge E_k is introduced as follows

$$E_k = \{\Phi(\left\|r_i - r_j\right\|) | \; i = k;\; j \neq k \in [1, N];\; k \in [1, N]\}$$

where $r_i - r_j$ is the Euclidean distance between the ith and jth atoms. $\Phi(r_i - r_j)$ is the radial basis function, which represents the interaction between the ith and jth atoms. Generally, Φ should meet the following requirements

$$\lim_{||r_i - r_j|| \to 0} \Phi(\left\|r_i - r_j\right\| = 1) \tag{3}$$

$$\lim_{||r_i - r_j|| \to \infty} \Phi(\left\|r_i - r_j\right\| = 0) \tag{4}$$

There are many radial basis functions meeting those requirements. In this work, only the generalized exponential functions (eq 5, called Gexp) and generalized Lorentz functions (eq 6, called GLor) are adopted.[32]

$$Gexp(\left\|r_i - r_j\right\|) = e^{\left(-\frac{\left\|r_i - r_j\right\|}{\eta_{i,j}}\right)^k}, \qquad k > 0 \tag{5}$$

$$GLor(\left\|r_i - r_j\right\|) = \frac{1}{1 + \left(\frac{\left\|r_i - r_j\right\|}{\eta_{i,j}}\right)^v}, \qquad v > 0 \tag{6}$$

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

$$\Phi = \text{GexporGLor} \tag{7}$$

The parameter $\eta_{i,j}$ within eqs 5 and 6 is the characteristic distance between atoms and κ and ν are two hyperparameters. These parameters allow the model to construct multi-scale interactions. To obtain the interactions on each atom in the cluster, a set of multiscale weighted geometric subgraphs (MWGSs) denoted as $G_i(V,E_i)$, i = 1,2,3,...,N is introduced for an N-atom Li cluster. The ith vertex in subgraph G_i is the only vertex that has a degree greater than 1. It is named the star in graph theory. Specifically, a star with three edges is called a claw, as shown in Figure 1a. Similar to the *WCG* centrality

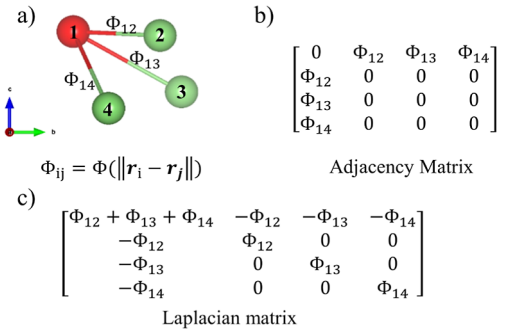


Figure 1. Illustration of weighted subgraph G1 and its matrix representations. (a) G_1 is a subgraph describing the interactions between atom 1 and the other atoms. The vertices with indexes 1, 2, 3, and 4 represent lithium atoms within the cluster; the color of the central atom is red and the others are green. The edges are weighted by the function Φ_{ij}, which represents the interactions between pairs of atoms. (b) Adjacency matrix representation and (c) Laplacian matrix representation of G_1.

proposed by Bramer and Wei,[37] our geometric centrality of the ith vertex in G_i is denoted as

$$C_i^{G_i} = \sum_{j \neq i} \Phi(\|r_i - r_j\|) \tag{8}$$

where $C_i^{G_i}$ characterizes the total interactions between the ith atom and its environment (i.e., all other atoms around it). G_i allows us to consider the interactions in a cluster with fine granularity. In the present work, we use two radial functions, namely, the generalized exponential and generalized Lorentzian functions, simultaneously to extract the complex interaction information between pairs of Li atoms. The distances between the same pair of atoms in the three-dimensional Euclidean space can be projected to different function spaces by this approach, thus enabling the interactions to be carved from different perspectives.

Multiscale Weighted Spectral Subgraphs. The MWGS gives us an intuitive and effective geometric representation of the cluster. It is very interesting to construct an equally effective description from an algebraic perspective. In mathematics, algebraic graph theory is the study of graphs by introducing the methods of linear algebra. Specifically, the study of graphs is achieved by using matrices to describe the relationships of the vertices of the graph and then by analyzing the characteristic spectrum of the matrices. In this work, we make use of two very important matrices, namely, the adjacency matrix and the Laplace matrix, to describe the MWGS. In a subsequent study, we specifically investigate the structural information stored in the subgraph spectra, which are named MWSSs in this study.

Multiscale Weighted Adjacency Matrix. Based on the discussion about weighted geometric subgraphs above, the adjacency matrix of G_i using weighted edge function Φ is denoted as

$$(A_i^{\phi})_{kj} = \begin{cases} \Phi(\|r_k - r_j\|), & \text{if } k = i;\ j \neq i \Lambda i,\ j \epsilon [1, N] \\ 0, \text{ otherwise} \end{cases} \tag{9}$$

Mathematically, A_i^{Φ} is a symmetric non-negative matrix. The eigenvalues and the corresponding eigenvectors of the adjacency matrix are denoted as λ_j^A, j = 1,2,3,... and μ_j^A, j = 1,2,3,.... Since for each eigenvalue λ_j^A, its opposite $-\lambda_j^A$ is also an

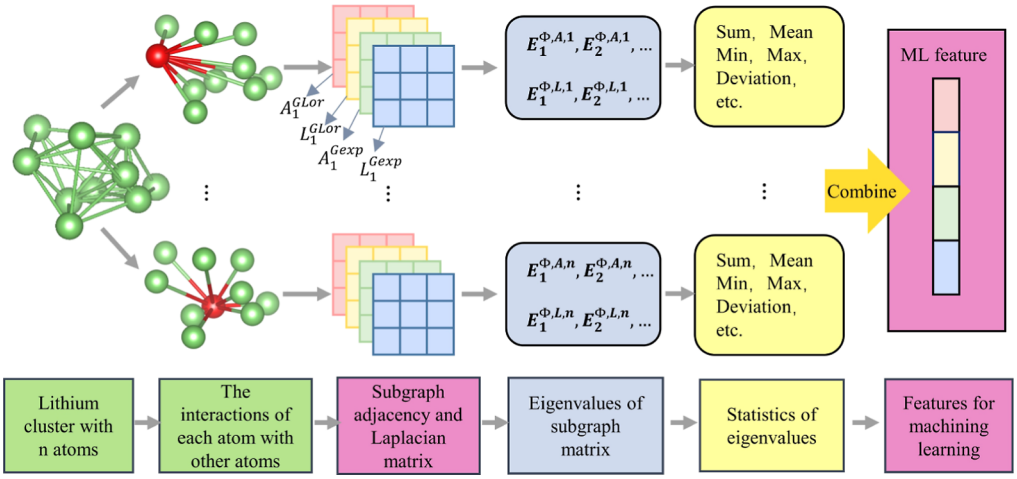


Figure 2. Workflow of the feature generation by using the spectral graph method. The first column describes a Li cluster with n atoms. In the second column, n subgraphs are obtained by considering the interaction of each atom with other atoms separately. In the third column, the adjacency matrix ($A^{\Phi_{i,j}}$) and the Laplacian matrix ($L^{\Phi_{i,j}}$) are constructed using two different subgraph weights, namely, the generalized exponential function (Gexp) and the generalized Lorentz function (GLor). Subsequently, the statistics of the matrix eigenvalues are used to represent the corresponding subgraphs (fourth and fifth columns). In the last column, the statistical information of all subgraphs is combined to become the final features.

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

eigenvalue, only positive eigenvalues are considered in this article.

Multiscale Weighted Laplacian Matrix. The adjacency matrix visually depicts the connection characteristics of the graph. From the perspective of signal processing, this can be thought of as describing the graph from the perspective of the time domain. On the other hand, the Laplacian matrix is related to the frequency domain characteristics of the task and the graph.[40] The Laplacian matrix of G_i is denoted as

$$L_i^{\Phi} = D_i^{\Phi} - A_i^{\Phi} \tag{10}$$

$$(D_i^{\phi})_{kj} = \begin{cases} \sum_j \Phi(\left\|r_k - r_j\right\|), & \text{if } k = i \Lambda k = j \\ 0, \text{ otherwise} \end{cases} \tag{11}$$

$$(L_i^{\phi})_{kj} = \begin{cases} -\sum_{l,l\neq k} L_{lj}, & \text{if } k = j \\ -\Phi(\left\|r_k - r_j\right\|), & \text{if } k = i;\ j \neq i \\ 0, \text{ otherwise} \end{cases} \tag{12}$$

where D_i^{Φ} is the degree matrix, which reflects the degree of each node within the weighted subgraph. Figure 1 shows the adjacency matrix and Laplacian matrix for weighted subgraph G_1 for Li_4. The eigenvalues and the corresponding eigenvectors of the Laplacian matrix are denoted as λ_j^{L}, j = 1,2,3,... and μ_j^{L}, j = 1,2,3,...

Mathematically, λ_j^{L} and μ_j^{L} behave differently compared to λ_j^{A} and μ_j^{A}. Since the Laplacian matrix is positive-semidefinite symmetric and diagonally dominant, all its eigenvalues λ_j^{L} are non-negative. Moreover, the number of zero eigenvalues equals the number of independent components. Also, the second-smallest eigenvalue, i.e., the first non-zero eigenvalue, of the matrix is called the Fiedler value. The Fiedler value is considered to be related to the connectivity of the graph. It has been used to analyze the brain structure,[41] the stability of the power system,[42] and so forth.

MWSS-Assisted Feature Generation. Figure 2 shows the process of transforming a lithium cluster with n atoms into ML input features, which is called feature engineering. When we obtain a cluster with n atoms, the interaction of each atom with the surrounding $n-1$ atoms is considered and n subgraphs that contain the same vertices but with different connections are constructed, as shown in Figure 2 (second column). Then, multiscale weighted adjacency matrices (A_i^{Φ}) and multiscale weighted Laplacian matrices (L_i^{Φ}) can be established according to eqs 9 and 12. Note that two different weight functions, e.g., the exponential function and the generalized Lorentz function, are used in this work, so that four matrices can be constructed, including two adjacency matrices and two Laplacian matrices with different weight functions, as shown in the third column of Figure 2. Suppose that we get all the eigenvalues $\{\lambda\}$ of a certain matrix, A_i^{Φ} or L_i^{Φ}. A statistical representation of a matrix is adopted by calculating nine statistical values, including summation, minimum, maximum, mean, median, standard deviation, variance, and the numbers and the sum of squares of the eigenvalues of that matrix. Note that for both the adjacency matrix and Laplacian matrix, only positive eigenvalues are considered during the statistic. All these statistical values can be divided into four groups, e.g., A_i^{Gexp}, A_i^{GLor}, L_i^{Gexp}, and L_i^{GLor}, where A and L mean adjacency matrix and Laplacian matrix, subscript i is the subgraph of the ith atom, and superscript indicates radial basis function. The nine statistical values of each group's statistical representation of matrices are calculated. Such statistical values are combined to form the 36-dimension ML feature of a cluster as shown in the last column of Figure 2.

ML Algorithm. Assume that the feature engineering above encodes structural information of the Li cluster into a low-dimension feature, $x \in R_n$. The ML algorithm maps the feature to the binding energy of the cluster, which is regarded as a regression problem in ML. This task can be formally expressed as

$$\underset{\theta}{\operatorname{argmin}} \sum_{i\in\text{train}} \mathcal{L}(y_i, f(x_i;\ \theta)) \tag{13}$$

where i is the ith sample in the training set, $f{:}R_n \rightarrow f$ is a reflection given by the learning algorithm, θ is the learnable parameter that the learning algorithm learns from the training data, and $\mathcal{L}$ is the loss function that determines the difference between the predicted value and the true value. Different learning algorithms are proposed for regression problems. The support vector machine (SVM) is mostly used in classification problems. Support vector regression (SVR) extends SVM from a classification problem to regression problem.[43] The regression tree is another technique for this problem. It uses a process called binary recursive partitioning to estimate the output. Usually, a single regression tree is considered a weak learner and ensemble learning methods are then used to improve its performance. Ensemble learning uses multiple weak regression models to jointly determine the final output. Ideally, in order to benefit as much as possible from ensemble learning, different weak regression models need sufficient performance and large diversity, but the two are usually in conflict. Ensemble learning is divided into two main directions in regression problems: the gradient boosting regression (GBR),[44] which aims to train a series of interrelated strong learners, and the random forest regression (RFR),[45] which aims to increase the diversity of learners as much as possible. As for loss functions, several functions are used in regression problems, such as mean absolute error (MAE), mean squared error, and so on. Loss functions can also be used to evaluate the performance of the model, and there are other metrics of the performance that could not be loss functions because they cannot perform gradient optimization, such as classification accuracy. Here, two metrics, MAE and Pearson correlation coefficient (PCC), are used. The MAE is given as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - f(x_i;\ \theta)| \tag{14}$$

where N is the number of samples and y_i and $f(x_i;\theta)$ are the label and the prediction value of the ith sample, respectively. Also, PCC measures the linear correlation between two sets of data X and Y, and it is given as

$$\mathrm{PCC} = \frac{\mathrm{COV}(X, Y)}{\sigma_x \sigma_y} \tag{15}$$

where COV (X, Y) is the covariance between random variables X and Y and σ_x and σ_y are the standard deviation of X and Y, respectively.

Here, the goal of our model is to predict the binding energy of clusters. The binding energy represents the energy at which

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

isolated atoms in a vacuum combine to form a cluster, which is denoted as

$$E_{\text{bind}} = (E_{\text{cluster}} - nE_{\text{atom}})/n \tag{16}$$

where E_{cluster} and E_{atom} are the energy of a cluster and an isolated atom (here, Li atom), respectively, and n is the number of atoms in the cluster. Since different clusters have different atomic numbers, we use the average energy, making the clusters with different atomic numbers comparable.

PSO Algorithm. The PSO algorithm is a heuristic method that simulates a biological population searching for food. Individuals in the population decide the direction of the next search based on the information they have obtained and the information obtained by the entire population. Here, every cluster is regarded as a particle. The above process can be described as

$$x_{i,j}^{t+1} = x_{i,j}^{t} + v_{i,j}^{t+1} \tag{17}$$

$$v_{i,j}^{t+1} = wv_{i,j}^{t} + c_1 r_1(s\text{best}_{i,j}^{t} - x_{i,j}^{t}) + c_2 r_2(i\text{best}_{i,j}^{t} - x_{i,j}^{t}) \tag{18}$$

where $x_{i,j}^{t}$ represents the jth dimension of the ith individual in the time step t. w, c_1, and c_2 are constants and r_1 and r_2 are random numbers, $s$best is the coordinate of the best structure found by the swarm, and $i$best is the best structure found during the first t-generation of the ith cluster.

DFT Calculation. The plane-wave projector augmented wave (PAW) method, which is implemented in the Vienna Ab initio Simulation Package, is used to perform DFT calculations. The energy cut-off is 520 eV. The exchange−correlation potential is the generalized gradient approximation with the Perdew−Burke−Ernzerhof parameterization. The PAW pseudo-potential is used.

SGL−ML−PSO−DFT System. Using the methods described above, an SGL−ML−PSO−DFT algorithm is proposed. The workflow of the system is shown in Figure 3. The red blocks represent the start point or the end point of the system. First, the training data are generated and an ML model is trained as described in the blue blocks. Once the model is trained, the ML−PSO procedure could be applied. In the PSO procedure, we generated different cluster structures from crystal structures and iteratively generated new structures according to eqs 17 and 18. After N iterations, ML models are used to select clusters with lower energy from the resulting structures. Finally, such clusters are optimized using DFT calculations to obtain more stable structures.

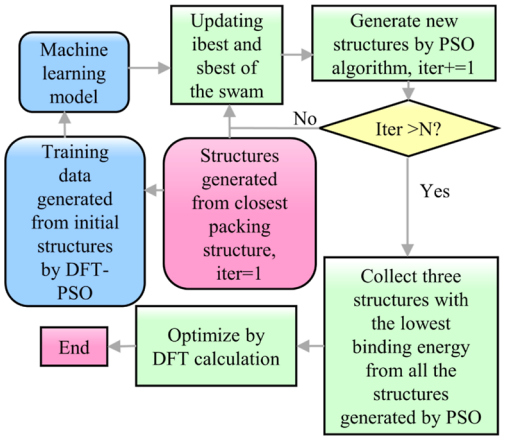


Figure 3. Flowchart of the ML−PSO−DFT algorithm. Here, the function of the ML model is to predict and compare the binding energy of the structures generated by PSO. After iterations, the final selected candidate structures are further optimized by DFT calculations.

Data Sets. In this article, the data set published by Chen[46] is adopted. There are 138,617 Li clusters, and their binding energy is generated by PSO−DFT. Among them, there are 16,617 lithium clusters containing 3−10 Li atoms (denoted as Li_{3-10}), 1000 lithium clusters containing 20 atoms (denoted as Li_{20}) and 1000 lithium clusters containing 40 atoms (denoted as Li_{40}).

Here, Li_{3-10} clusters are first used to train the model and test the effectiveness of the ML methods. A train-validation-test data splitting scheme is applied to optimize and test the model. All Li_{3-10} clusters are randomly divided into an 80% training set, a 10% validation set, and a 10% test set. The training set is used to optimize the parameter of the model. The validation set is used to optimize the hyperparameters including hyperparameters of the ML model and parameters of SGL. The testing set is used to test the performance of our method.

RESULTS AND DISCUSSION

SGL Parameterization. According to eqs 5 and 6, generally our MWSSs could be characterized by three parameters τ, κ and ν. κ and ν refer to the kernel order of generalized exponential and generalized Lorentz kernels, respectively. Additionally, τ is used to characterize $\eta_{kk'}$, which is defined as $\eta_{kk'} = \tau(r_k + r_{k'})$, where r_k and $r_{k'}$ are the van der Waals radii of element type k and element type k', respectively. In our situation, $\eta_{\text{LiLi}} = 2\tau(r_{\text{Li}})$. Since r_{Li} is a constant, η is regarded as a parameter to be optimized. Based on the previous work, the recommended values for the three parameters are shown in Table 1:

Table 1. Ranges of Hyperparameters Recommended for SGL

parameter	domain
H	{0.5,1.0,...,6}
κ	{0.5,1.0,...,6}∪{10,15,20}
N	{0.5,1.0,...,6}∪{10,15,20}

Here, the train-validation scheme was used to find the best-fit parameters. Parameters $\eta = 2$, $\kappa = 2.5$, and $\nu = 2$ were used in the following experiment. The details for determining the combination of parameters are described in the Supporting Information.

ML Model. After determining the parameters of SGL, the atomic position information of each Li cluster is encoded as a vector, e.g., features. The ML model was trained to learn the relation between the features and their corresponding binding energy. The regression tree is a classic ML model and ensemble learning coupling a series of decision trees for better performance. Here, two different ensemble-learning models GBR and FRF were used. We trained the two different models on the same Li_{train}, hyperparameter optimization on Li_{val} and finally measured the performance of the model on Li_{test}. In order to reduce the stochastic error within the ML model, each model was trained 10 times repeatedly with all the hyperparameters unchanged and the output was averaged. Feature normalization was also implemented when training the model. The hyperparameters of GBR and FRF are described in the

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

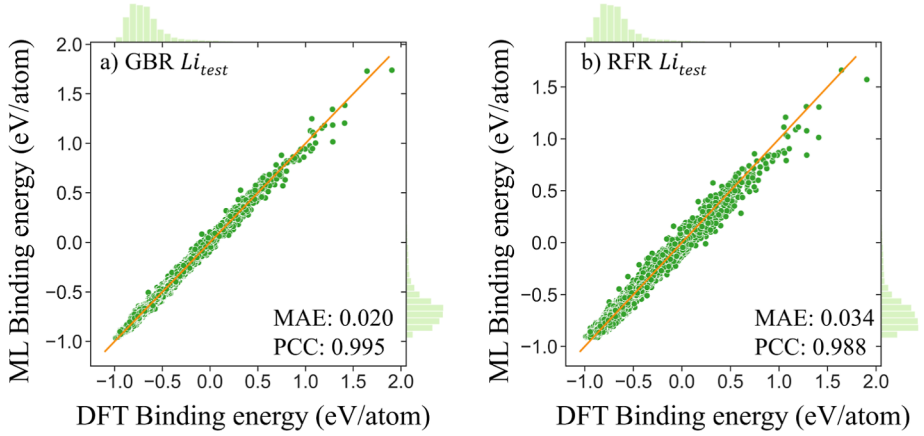


Figure 4. Comparison between ML predictions and DFT calculations of binding energy. (a) GBR model on the test set of Li_{3-10}. (b) RFR model on the test set of Li_{3-10}.

Supporting Information, and the performances of the two models on Li_{test} are shown in Figure 4.

According to Figure 4, the MAE loss and Pearson's correlation coefficient (PCC) of the GBR model on the test set were 0.020 and 0.995, while the RFR model had a higher MAE loss (0.034) and lower PCC (0.988). This indicated that the GBR model was more effective for the organization of regression trees on this issue. Since the GBR model can achieve a better performance, we used the GBR model in the following section.

As mentioned above, we applied the multi-scale technique by using two different radial basis functions eqs 5 and 6 to describe the interaction between atoms simultaneously. For each radial basis function, we constructed a group of adjacency matrices and a group of Laplacian matrices to extract the feature. In all, our feature generates from four groups of matrices: A_i^{Gexp}, A_i^{GLor}, L_i^{Gexp}, and L_i^{Lor}. (The meaning of the symbol is given above.) We consider different combinations of matrices to generate structural features to test the impact of these features on the model performance and thus validate the effectiveness of this multi-scale approach. Table 2 presents the result of our ablation study. When only considering one type of matrix group, the model could not obtain a very good performance. Then, the features generated from matrix groups that used the same radial basis were combined ($A^{Gexp} + L^{Gexp}$ and $A^{GLor} + L^{GLor}$). Such a composite combination reduces the MAE of the Gexp and GLor radial basis functions to 0.050 and 0.041, respectively, compared to using only one type of the matrix group. Furthermore, the PCC was increased to 0.979 and 0.986 for these two functions. For a graph, its adjacency matrix and its Laplacian matrix contain the same amount of information, but their eigenvalues are very different. Therefore, considering the eigenvalues of the adjacency matrix and the Laplace matrix at the same time can better reflect the structural information of clusters even without using the multiscale technique. The last three rows in Table 2 show the effects of the multiscale technique. Compared with the absence of the multi-scale technique, the MAE of the model decreased from 0.050 (using G^{exp} function alone) and 0.041 (using G^{Lor} function alone) to 0.020, respectively, and the PCC increased from 0.979 and 0.986–0.995. The above ablation experiments showed that it was reasonable to consider the eigenvalues of the adjacency matrix and the Laplacian matrix at the same time and to use the multi-scale technique to describe the interatomic interaction.

Table 2. Result of the Ablation Study[a]

composition of features	MAE	PCC
A^{exp}	0.049	0.970
L^{exp}	0.155	0.876
A^{Lor}	0.060	0.928
L^{Lor}	0.072	0.905
$A^{exp} + L^{exp}$	0.050	0.979
$A^{Lor} + L^{Lor}$	0.041	0.986
$A^{exp} + L^{exp} + A^{Lor} + L^{Lor}$	0.020	0.995

[a]The first column denotes the features generated by A^{exp}, L^{exp}, A^{Lor}, and L^{Lor}, respectively, where $A^{exp} + L^{exp}$ denotes the combination of A^{exp} features and L^{exp} features and $A^{exp} + L^{exp} + A^{Lor} + L^{Lor}$ denotes the combination of features generated by all four sets of matrices. MAE and PCC are used as metrics.

Performance on Li_n. The previous section indicated that our model is capable of predicting the binding energy of small Li clusters (atom number ranges from 3 to 10). To improve the accuracy of the model, the model was trained on a large amount of labeled (binding energy) data. However, for a cluster with a particular structure, DFT calculations are usually required to obtain its binding energy. In addition, the time of DFT calculation increases rapidly with the increase in the number of atoms contained in the cluster. It is difficult to build data sets that contain binding energies of a large number of clusters of medium and large size. So, we used Li_{20} and Li_{40} to further test our model trained on Li_{3-10}. Additionally, Chen[46] developed a topological fingerprint (TF) to estimate the binding energy of clusters using the topological graph. Because Chen's way of dividing data sets was different from ours, we reproduced his method as a comparison. The performances of SGL and TF are shown in Figure 5. Both methods had larger MAEs compared to the MAE of Li_{3-10}. The larger MAEs for both two methods were due to systematical errors. However, in our work, it is more important to judge the relative value of the binding energy of different clusters than to predict the absolute value of the binding energy; thus, PCC is a more reasonable metric. From this point of view, the SGL method outperformed the TF method. Meanwhile, the feature dimension of SGL was 36 dimensions, lower than that of TF with 400 dimensions. An SVR model was also implemented for the same tasks, and its performance is described in the Supporting

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

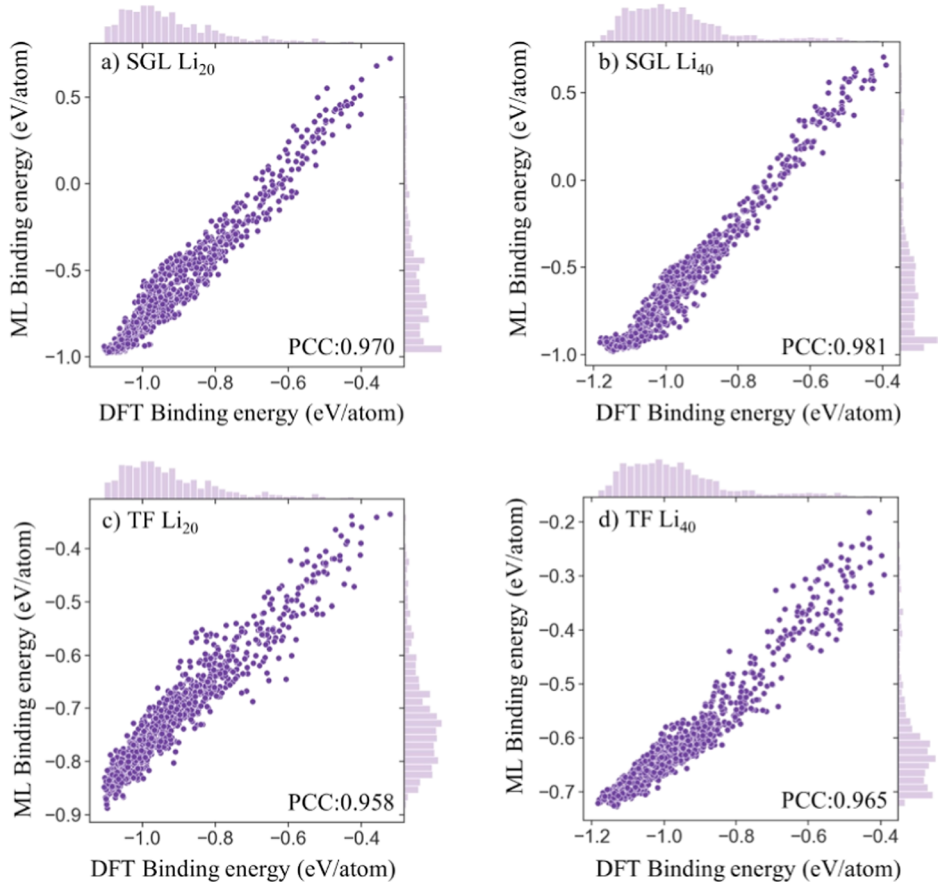


Figure 5. Performance comparison of models using different feature engineering methods. (a) SGL method on Li_{20}. (b) SGL method on Li_{40}. (c) TF method on Li_{20}. (d) TF method on Li_{40}.

Information. The CGCNN model was also implemented for the same task. Compared to models using deep learning, our ML approach was comparable in PCC. In addition, deep learning models are often sensitive to the choice of hyperparameters and may require more efforts to tune the parameters for a particular task. The details and results of the model are presented in the Supporting Information.

Searching for the Ground-State Structure of Li_{20}. We implemented the ML−PSO−DFT system to search for the Li_{20} structure with the lowest energy as a case study. First, we generated different structures using ML−PSO with a population of 2000 and a generation of 15 and the ML model trained above. The ML model determined the energy of each structure and guided the generation. After generation, the ML model determined the three clusters with the lowest energy among all the structures. Finally, all three clusters were further optimized using DFT. Figure 6 shows the output of the ML−PSO−DFT system. Our predicted ground-state structure of Li_{20}, as shown in Figure 6a, is composed of three centered trigonal prisms with five additional capped atoms. This result is consistent with the structure of the ground state obtained by DFT−PSO[9] and by Tabu's search.[47]

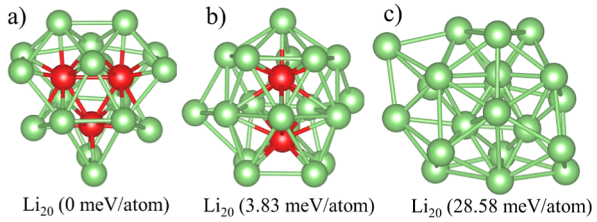


Figure 6. Three clusters obtained by ML−PSO−DFT. (a) Ground-state structure of Li_{20}. (b,c) Metastable structures of Li_{20}.

CONCLUSIONS

In conclusion, ML for cluster physics critically relies on appropriate data representations, which should not only reflect the collective characteristics of the structure but also maintain the physical invariance. In this work, MWSSs were proposed as a low-dimension representation of clusters. Unlike conventional adjacency matrices, we introduced the multiscale technique to extract information on the interaction between pairwise atoms. We combined SGL with different ML algorithms and tested their performance on binding energy prediction. By incorporating the GBR ML algorithm, the binding energy of lithium (Li) clusters can be accurately determined, and expensive DFT calculations can be avoided. Then, in a practical application, we embed SGL−GBR as an ML model into the PSO algorithm and find the global ground state of Li_{20}. Since the morphology of Li clusters affects the nucleation of Li dendrites, our work may have positive implications for the study of Li dendrite formation.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpca.3c00272.

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059

SGL parameter optimization; heat map of MAE in Li_{val} with different SGL parameters; GBR and RFR models; performance of SGL with the SVR model; implementation and results of CGCNN; dihedral distortions and features; and effect of dihedral angle on features (PDF)

AUTHOR INFORMATION

Corresponding Authors

Dong Chen − *School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*; Email: chend@pku.edu.cn

Feng Pan − *School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*; orcid.org/0000-0002-8216-1339; Email: panfeng@pkusz.edu.cn

Authors

Shengming Ma − *School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Shisheng Zheng − *School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Wentao Zhang − *School of Advanced Materials, Peking University Shenzhen Graduate School, Shenzhen 518055, People's Republic of China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpca.3c00272

Author Contributions

D.C. designed the project and revised the manuscript. S.M. finished the coding, data analysis, and first draft writing. W.Z. and S.Z. revised the manuscript. F.P. supervised the project and acquired funding.

Notes

The authors declare no competing financial interest.

For reproducibility, all the implementation of SGL and the data used to train the model are publicly available at https://github.com/arronsma/SGL-COD.

ACKNOWLEDGMENTS

This work was financially supported by the Soft Science Research Project of Guangdong Province (no. 2017B030301013) and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform.

REFERENCES

(1) Brack, M. The physics of simple metal clusters: self-consistent jellium model and semiclassical approaches. *Rev. Mod. Phys.* **1993**, *65*, 677−732.

(2) Stillinger, F. H. Exponential multiplicity of inherent structures. *Phys. Rev. E* **1999**, *59*, 48−51.

(3) Ali, M.; Smith, R. The structure of small clusters ejected by ion bombardment of solids. *Vacuum* **1993**, *44*, 377−379.

(4) Shang, C.; Liu, Z.-P. Stochastic Surface Walking Method for Structure Prediction and Pathway Searching. *J. Chem. Theory Comput.* **2013**, *9*, 1838−1845.

(5) Deaven, D. M.; Ho, K. M. Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.* **1995**, *75*, 288−291.

(6) Hartke, B. Global geometry optimization of clusters using genetic algorithms. *J. Phys. Chem.* **1993**, *97*, 9973−9976.

(7) Hobday, S.; Smith, R. Optimisation of carbon cluster geometry using a genetic algorithm. *J. Chem. Soc., Faraday Trans.* **1997**, *93*, 3919−3926.

(8) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(9) Lv, J.; Wang, Y.; Zhu, L.; Ma, Y. Particle-swarm structure prediction on clusters. *J. Chem. Phys.* **2012**, *137*, 084104.

(10) Call, S. T.; Zubarev, D. Y.; Boldyrev, A. I. Global minimum structure searches via particle swarm optimization. *J. Comput. Chem.* **2007**, *28*, 1177−1186.

(11) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063−2070.

(12) Ward, L.; Wolverton, C. Atomistic calculations and materials informatics: A review. *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 167−176.

(13) Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition 2014, p. arXiv:1409.1556. https://ui.adsabs.harvard.edu/abs/2014arXiv1409.1556S (accessed Sept 01, 2014).

(14) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(15) Court, C. J.; Yildirim, B.; Jain, A.; Cole, J. M. 3-D Inorganic Crystal Structure Generation and Property Prediction via Representation Learning. *J. Chem. Inf. Model.* **2020**, *60*, 4518−4535.

(16) Nouira, A.; Sokolovska, N.; Crivello, J.-C. CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks, 2018, p. arXiv:1810.11203. https://ui.adsabs.harvard.edu/abs/2018arXiv181011203N (accessed Oct 01, 2018).

(17) Fujikake, S.; Deringer, V. L.; Lee, T. H.; Krynski, M.; Elliott, S. R.; Csányi, G. Gaussian approximation potential modeling of lithium intercalation in carbon nanostructures. *J. Chem. Phys.* **2018**, *148*, 241714.

(18) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(19) Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **2016**, *14*, 1153−1173.

(20) Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **2019**, *99*, 014104.

(21) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.

(22) Glielmo, A.; Zeni, C.; De Vita, A. Efficient nonparametric $n $-body force fields from machine learning. *Phys. Rev. B* **2018**, *97*, 184307.

(23) Scott, J. Social Network Analysis. *Sociology* **1988**, *22*, 109−127.

(24) Abedin, B.; Sohrabi, B. Graph theory application and web page ranking for website link structure improvement. *Behav. Inf. Technol.* **2009**, *28*, 63−72.

(25) Bunn, A. G.; Urban, D. L.; Keitt, T. H. Landscape connectivity: A conservation application of graph theory. *J. Environ. Manage.* **2000**, *59*, 265−278.

(26) Weng, M.; Wang, Z.; Qian, G.; Ye, Y.; Chen, Z.; Chen, X.; Zheng, S.; Pan, F. Identify crystal structures by a new paradigm based on graph theory for building materials big data. *Sci. China Chem.* **2019**, *62*, 982−986.

(27) Li, S.; Chen, Z.; Wang, Z.; Weng, M.; Li, J.; Zhang, M.; Lu, J.; Xu, K.; Pan, F. Graph-based discovery and analysis of atomic-scale one-dimensional materials. *Natl. Sci. Rev.* **2022**, *9*, nwac028.

(28) Jie, J.; Weng, M.; Li, S.; Chen, D.; Li, S.; Xiao, W.; Zheng, J.; Pan, F.; Wang, L. A new MaterialGo database and its comparison with other high-throughput electronic structure databases for their predicted energy band gaps. *Sci. China Technol. Sci.* **2019**, *62*, 1423−1430.

(29) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.
(30) Carlsson, G. Topological methods for data modelling. *Nat. Rev. Phys.* **2020**, *2*, 697−708.
(31) Ewing, J.; Solomon, L. Applications of algebraic topology. *Bull. Am. Math. Soc.* **1976**, *82*, 676−682.
(32) Opron, K.; Xia, K.; Wei, G.-W. Fast and anisotropic flexibility-rigidity index for protein flexibility and fluctuation analysis. *J. Chem. Phys.* **2014**, *140*, 234105.
(33) Opron, K.; Xia, K.; Wei, G.-W. Communication: Capturing protein multiscale thermal fluctuations. *J. Chem. Phys.* **2015**, *142*, 211101.
(34) Nguyen, D. D.; Xia, K.; Wei, G.-W. Generalized flexibility-rigidity index. *J. Chem. Phys.* **2016**, *144*, 234106.
(35) Xia, K.; Opron, K.; Wei, G.-W. Multiscale Gaussian network model (mGNM) and multiscale anisotropic network model (mANM). *J. Chem. Phys.* **2015**, *143*, 204106.
(36) Nguyen, D. D.; Xiao, T.; Wang, M.; Wei, G.-W. Rigidity Strengthening: A Mechanism for Protein−Ligand Binding. *J. Chem. Inf. Model.* **2017**, *57*, 1715−1721.
(37) Bramer, D.; Wei, G.-W. Multiscale weighted colored graphs for protein flexibility and rigidity analysis. *J. Chem. Phys.* **2018**, *148*, 054103.
(38) Dunn, B.; Kamath, H.; Tarascon, J.-M. Electrical Energy Storage for the Grid: A Battery of Choices. *Science* **2011**, *334*, 928.
(39) Wen, J.; Yu, Y.; Chen, C. A Review on Lithium-Ion Batteries Safety Issues: Existing Problems and Possible Solutions. *Mater. Express* **2012**, *2*, 197−212.
(40) Shuman, D. I.; Narang, S. K.; Frossard, P.; Ortega, A.; Vandergheynst, P. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **2013**, *30*, 83−98.
(41) Cai, J.; Liu, A.; Mi, T.; Garg, S.; Trappe, W.; McKeown, M. J.; Wang, Z. J. Dynamic Graph Theoretical Analysis of Functional Connectivity in Parkinson's Disease: The Importance of Fiedler Value. *IEEE J. Biomed. Health Inf.* **2019**, *23*, 1720−1729.
(42) Wei, G. W.; Zhan, M.; Lai, C. H. Tailoring Wavelets for Chaos Control. *Phys. Rev. Lett.* **2002**, *89*, 284103.
(43) Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A.; Vapnik, V. Support vector regression machines. *Adv. Neural Inf. Process. Syst.* **1997**, *9*, 155−161.
(44) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(45) Liaw, A.; Wiener, M. Classification and regression by randomForest. *R. News* **2002**, *2*, 18−22.
(46) Chen, X.; Chen, D.; Weng, M.; Jiang, Y.; Wei, G.-W.; Pan, F. Topology-Based Machine Learning Strategy for Cluster Structure Prediction. *J. Phys. Chem. Lett.* **2020**, *11*, 4392−4401.
(47) Fournier, R.; Bo Yi Cheng, J.; Wong, A. Theoretical study of the structure of lithium clusters. *J. Chem. Phys.* **2003**, *119*, 9444−9454.

https://doi.org/10.1021/acs.jpca.3c00272
J. Phys. Chem. A 2023, 127, 2051−2059