ABSTRACT: In cluster physics, the determination of the ground-state structure of medium-sized and large-sized clusters is a challenge due to the number of local minimal values on the potential energy surface growing exponentially with cluster size. Although machine learning approaches have had much success in materials science, their application in clusters are often hindered by the geometric complexity clusters. Persistent homology provides a new topological strategy to simplify geometric complexity while retaining important chemical and physical information without having to “downgrade” the original data. We further propose persistent pairwise independence (PPI) to enhance the predictive power of persistent homology. We construct topology-based machine learning models to reveal hidden structure–energy relationships in lithium (Li) clusters. We integrate the topology-based machine learning models, a particle swarm optimization algorithm, and density functional theory calculations to accelerate the search of the globally stable structure of clusters.

The determination of the topography of the potential energy surfaces of various clusters is a subject of intensive research in cluster physics. One of the most challenging tasks is to determine the ground-state structure of medium and large clusters due to the fact that the number of local minima on the potential energy surface grows exponentially with the increase of cluster size. It is commonly believed that the binding energy of a cluster is mainly determined by the geometric structure of the cluster. Therefore, the understanding of the structure-energy relationship of clusters is the holy grail in cluster research. In particular, the prediction of structures with the global minimal energy and near minimal energies is of practical interest.

Several efficient global optimization methods, including random sampling, basin hopping, minimal hopping, simulated annealing, genetic algorithm, and particle swarm optimization (PSO) algorithm, have been developed in the past few decades for searching structures with near-global minimal energies. Among them, the PSO algorithm, first proposed by Kennedy and Eberhart, is a relatively efficient one. Call et al. applied the PSO algorithm to the structure prediction of small clusters or isolated molecules. Recently, Ma et al. have developed a PSO-based structure analysis method, which is widely applied to structure predictions. However, all the aforementioned methods are computationally expensive due to involved first-principle calculations of numerous local minimum values. As shown in a recent review, machine learning (ML) has been proven to be an efficient strategy for estimating the density functional theory (DFT) energy of ground states of materials and for the prediction of material properties. It has the potential for material structure prediction as well.

Recently, machine learning has had tremendous success in science, engineering, finance, medicine, and various other industrial sectors. Initial successes of machine learning approaches were limited to image and video processing, computer vision, etc. For relatively simple data sets, such as images, advanced machine learning algorithms, such as a convolutional neural network (CNN), can automatically extract image patterns without handcrafted input features. Such predictive models involve two essential components, namely, data set and learning algorithm. However, for general data sets, particularly data sets with complex geometric structures in material sciences and molecular biology, the internal structural complexity hinders the performance of the aforementioned two-component machine learning approach. A predictive machine learning model for complex data consists of three essential components: the data set, learning algorithm, and data representation. The representations of data, called “descriptors” or “features”, play an essential role in constructing an efficient machine learning model for data sets with complex internal structures.

For complex three-dimensional (3D) cluster data, the representations used in machine learning vary greatly in their nature. Some representations are generated from Cartesian coordinates, geometric properties, electrostatics, atom types,
and atomic partial charges.\textsuperscript{13–30} The most direct and crude representation is Cartesian coordinates, but as regressive neural networks are numerical fitting methods, the output depends on the absolute values of the input coordinates. Since translation and rotation of a cluster should not change cluster intrinsic properties, Cartesian coordinate representation has a major limitation in energy predictions. The conventional solution is to describe cluster structures by using graph theory or symmetric functions that are invariant to translation and rotation. Graph-based representations have been successfully applied to predict molecular properties.\textsuperscript{31,32} Meanwhile, many symmetric functions based on interatomic distances have been developed,\textsuperscript{33–37} which are equally applicable to very different systems such as bulk metals, clusters or molecules. However, graph-based representations may neglect some geometric details, such as distortion, whereas symmetric function models often contain too complex structural detail and are frequently computationally intractable, due to its rather complicated conversion formula. Therefore, continuous, complete and simple low-dimensional representations that are invariant to translation and rotation are greatly desired.

Topology, free of metrics or coordinates, offers an entirely different approach and could provide the ultimate simplification of structural complexity. In fact, one only needs qualitative topological information to understand many physical properties. However, traditional topology oversimplifies data structures and leads to too much loss in geometric information. Persistent homology is a relatively new method that bridges geometry and topology. It integrates multiscale geometric information and topology, free of metrics or coordinates, or coordinate-free representations, persistent homology is able to embed geometric information into topological invariants so that the “birth”, “death”, and “persistence” of isolated components, circles, holes, and void at all geometric scales can be monitored by topological measurements. The basic idea of size functions was introduced by Frosini and Landi\textsuperscript{38} and by Robins.\textsuperscript{39} Edelsbrunner et al.\textsuperscript{38} formulated persistent homology, and Zomorodian and Carlsson generalized the mathematical theory.\textsuperscript{40} Usually, topological persistence over the filtration is represented either in persistent diagrams or in persistent barcodes,\textsuperscript{35} in which various horizontal line segments or bars indicate homology generators. Persistent homology has been applied to a variety of fields, including image analysis,\textsuperscript{41–47} image retrieval,\textsuperscript{48} chaotic dynamics verification,\textsuperscript{49,50} shape recognition,\textsuperscript{51} and computational biology.\textsuperscript{44,52–54} Wei and workers explored the utility of persistent homology for the quantitative analysis of protein\textsuperscript{54} and fullerene\textsuperscript{55} and proposed some of the first integrations of persistent homology and machine learning.\textsuperscript{56} It has been shown that persistent-homology-based machine learning algorithms outperform other methods in D3R Grand Challenges, a worldwide competition series in computer-aided drug design.\textsuperscript{57}

The objective of the present work is to introduce persistent homology as a unique representation of cluster structures. The persistent homology representation of cluster structures is realized using persistent barcodes. To enhance the predictive power of persistent barcode representation, we propose a new concept, persistent pairwise independence (PPI), as an auxiliary feature. The resulting topological fingerprints, including both persistent barcodes and PPI, are employed in a machine learning model to understand the structure—energy relationship of clusters. In a further combination with the PSO algorithm and DFT calculation, a topology-ML-PSO–DFT protocol is developed for the high throughput screen of cluster structures with the global energy minima. Lithium (Li) is the lightest metallic element in the periodic table and Li cluster is a subject of intense research in Li anode materials in storage battery. So Li cluster were studied by our topology-ML-PSO–DFT protocol as an illustration in this work. In fact, the idea about the combination of ML model with DFT-based structure prediction is not new.\textsuperscript{58–62} Among them, the Gaussian approximation potential (GAP) method is a relatively efficient one and has been successfully tested on many cases. However, these works focused on the acceleration of crystal structure relaxation with ML model learning and interatomic potential. In contrast, the part of acceleration in our work focused on the combination of ML model and stochastic global optimization algorithm to avoid the expensive DFT calculation on screening local minimum values.

In this work, persistent homology is used to characterize topological invariants, such as isolated components, circles, rings, loops, pockets, voids, and cavities, via topological spaces and algebraic group representations. We associate each atom in a cluster with an ever-increasing diameter to systematically generate a multiscale representation. We use persistent barcodes to represent cluster atomic interactions. Topological invariants are combined with PPIs to describe cluster molecules in our machine learning modeling. The global optimization algorithm and the first principle calculation used in this work are also discussed.

**Simplicial Homology.** Let \( \{v_0, v_1, \ldots, v_n\} \) be an affine independent sets in \( \mathbb{R}^n \), and a \( k \)-simplex, \( \sigma^k \), is a \( k \)-dimensional polytope that is the convex hull of above \( k + 1 \) vertices, expressed as

\[
\sigma^k = \left\{ \lambda_0 v_0 + \lambda_1 v_1 + \cdots + \lambda_k v_k \mid \sum_{i=0}^{k} \lambda_i = 1, 0 \leq \lambda_i \leq 1, i = 0, 1, \ldots, k \right\}
\]

To combine these geometric components, such as vertices, edges, triangles, and tetrahedrons together under certain rules, a simplicial complex \( K \) is constructed, which is a set of simplices that satisfies the following two conditions. The first is that any face of a simplex from \( K \) is also in \( K \). The second is that the intersection of any two simplices in \( K \) is either empty or a shared face of both simplices.

A \( k \)-chain \( \{\sigma^k\} \) is a linear combination \( \sum_i a_i \sigma_i^k \) of \( k \)-simplex \( \sigma^k \). The coefficients \( a_i \) can be chosen from rational number field \( \mathbb{Q} \), integers \( \mathbb{Z} \), and prime field \( \mathbb{F}_p \) with prime number \( p \). For simplicity, in this work the coefficients \( a_i \) is chosen from \( \mathbb{Z} \), for which the addition operation between two chains is the modulo 2 addition for the coefficients of their corresponding
The set of all $k$-chains of simplicial complex $K$ together with the addition operation forms an Abelian group $\mathcal{C}_k(K, \mathbb{Z})$. The homology of a topological space is represented by a series of Abelian groups satisfying certain relations.

Let $\partial_k$ be a boundary operator $\partial_k: \mathcal{C}_k \rightarrow \mathcal{C}_{k-1}$. The boundary of a $k$-simplex is given by

$$\sum_{i=0}^{k} \sigma_i$$

where $\{v_0, ..., \hat{v}_i, ..., v_n\}$ means the elimination of vertex $v_i$ from the simplex. The most important topological property is that a boundary has no boundary. In other words, a $k$-chain will be mapped to an empty set by the twice applications of boundary operations $\partial_k \partial_{k-1} = 0$. The $k$th cycle group $Z_k$ and the $k$th boundary group $B_k$ are the subgroups of $\mathcal{C}_k$ defined respectively as

$$Z_k = \text{Ker} \partial_k = \{c \in \mathcal{C}_k | \partial_k c = 0\}$$

and

$$B_k = \text{Im} \partial_{k+1} = \{c \in \mathcal{C}_k | \exists d \in \mathcal{C}_{k+1} : c = \partial_{k+1} d\}$$

The $k$th homology group $H_k$, defined by the $k$th cycle group $Z_k$ and the $k$th boundary group $B_k$, is given by the quotient group

$$H_k = \text{Ker} \partial_k / \text{Im} \partial_{k+1} = Z_k / B_k$$

The homology group extracts low-dimensional topological invariants from original data, which can lie in a very high dimensional space. In practical applications, it is the Betti number ($\beta_k$), one of the most important topological invariants, that is employed in our computation. For a given cluster structure, roughly speaking, the number of independent components, rings and cavities are topological invariants and they are referred to as $\beta_0$, $\beta_1$, and $\beta_2$, respectively. However, as mentioned earlier, the Betti number alone leads to a severe topological abstraction of real-world data.

Persistent Homology, Barcode, and Persistent Pairwise Independence. Persistent homology addresses the aforementioned oversimplification issue by introducing a multiscale geometric analysis to the original homology. Specifically, it systematically creates a family of inclusive topological spaces from a filtration process and then defines a nested family of homology groups over the filtration. In a filtration process, a chain complex, including a family of ordered subcomplexes, is created from a given simplicial complex $K$ by a parameter $\epsilon$,

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K$$

In our case, a simple choice of $\epsilon$ is the atomic diameter, which goes from 0 to a given maximal length $\epsilon_{\text{max}}$. A simplex is a geometric generalization of a triangle to arbitrary dimensions,
namely, a 0-simplex is a vertex, a 1-simplex is an edge, a 2-simplex is a triangle, and a 3-simplex represents a tetrahedron. Simplexes can describe very complicated geometric shapes and are much more computationally tractable than the original shapes that they represent. There are actually many different types of simplicial complex constructions that can be used in persistent homology. The Vietoris Rips complex, a type of simplicial complex derived from the connectivity of ϵ-balls in a given metric space, is used in this work as it is fairly easy to describe and reasonably practical for point cloud data. For example, a 1-simplex is formed if two points with diameter ϵ contact with each other. Vietoris Rips complex models the pairwise interaction of atoms. Specifically, for ϵ sufficiently small, one has only isolated points, i.e., 0-simplexes, whereas, for ϵ sufficiently large, one can have high-dimensional simplex. As such, a nested sequence of subcomplexes is defined when we grow diameter ϵ-ball around each atom in a cluster, as shown in eq 7. For each subcomplex in a chain complex, homology groups and the corresponding topological invariants can be computed. The resulting family of topological invariants in terms of persistent Betti numbers record the “birth” and/or “death” of all topological features in data, giving rise to topological fingerprints. Therefore, the evolution of topological invariants over the filtration process can be recorded as a barcode or a persistence diagram.

The proposed persistent pairwise independence (PPI) counts the independence of each pair of atoms (or points) over the filtration. Initially, all atoms are not connected. The number of PPI bars is equal to the number of pairs of independent atoms. As the filtration parameter increases, some pairs of atoms become connected and their persistent bars terminate. The proposed PPI barcode is more informative than the ϵ Persistent homology barcode. As shown in Figure 1, it can be used together with topological invariants to describe material structures.

**Table 1. List of Mean Absolute Error (MAE), Pearson Correlation Coefficient (PCC), and ϵ_max Used in the Present Work**

<table>
<thead>
<tr>
<th>structure of TFs</th>
<th>ϵ_max</th>
<th>MAE (eV/atom)</th>
<th>PCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>p(β_i)+p(β_j)+p(β_k)</td>
<td>10</td>
<td>0.039</td>
<td>0.979</td>
</tr>
<tr>
<td>p(β_i)+p(β_j)+p(β_j)+p(PPI)</td>
<td>10</td>
<td>0.028</td>
<td>0.986</td>
</tr>
<tr>
<td>p(β_i)+p(β_j)+p(β_j)+d(PPI)</td>
<td>10</td>
<td>0.026</td>
<td>0.990</td>
</tr>
<tr>
<td>p(β_i)+p(β_j)+d(PPI)</td>
<td>10</td>
<td>0.031</td>
<td>0.981</td>
</tr>
<tr>
<td>bd(β_i)+bd(β_j)+bd(β_j)+d(PPI)</td>
<td>10</td>
<td>0.030</td>
<td>0.983</td>
</tr>
<tr>
<td>bpd(β_i)+bpd(β_j)+bpd(β_j)+d(PPI)</td>
<td>10</td>
<td>0.026</td>
<td>0.988</td>
</tr>
<tr>
<td>p(β_i)+p(β_j)+p(β_j)+d(PPI)</td>
<td>5</td>
<td>0.028</td>
<td>0.986</td>
</tr>
<tr>
<td>p(β_i)+p(β_j)+p(β_j)+d(PPI)</td>
<td>15</td>
<td>0.026</td>
<td>0.989</td>
</tr>
</tbody>
</table>

*b(·)*, *p(·)*, and *d(·)* represent the birth, the persistence, and the death of topological invariant *β*, respectively. Here, bd(·) represents *b(·)+d(·)* and bpd(·) represents *b(·)+p(·)+d(·)*.

Figure 3. Flowchart of a topology-ML-PSO-DFT protocol
represents the filtration parameter $\epsilon$. From top to bottom, the behaviors of $\beta_0$, $\beta_1$, $\beta_2$, and PPI are depicted in four individual subfigures. It is seen that as $\epsilon$ increases, initially isolated atoms will gradually grow into $\epsilon$-balls with increasing diameter. Once two $\epsilon$-ball overlap with each other, one $\beta_0$ bar is terminated. As shown in Figure 2, at the very beginning, all $\epsilon$-balls do not overlap with one another. There are four purple bars in the first subfigure, meaning that there are four isolated $\beta_0$ bars. When the filtration parameter reaches 3 Å, each $\epsilon$-ball overlaps with its closest two balls, forming a connected square. Therefore, four bars of $\beta_0$ (purple bars) turn into one bar and a $\beta_1$ (blue bars) bar appears, which represents the emergence of an independent noncontractible quadrilateral structure (loops). In the end when the filtration parameter reaches around 4.2 Å, the $\beta_1$ bar disappears, leaving only a $\beta_0$ bar persistent in all subfigures. As can be seen from the last subfigure, Li$_4$ cluster has two different kinds of bars of PPI with lengths around 3 and 4.2 Å, respectively, indicating its two types of interatomic distances. Similarly, an octahedral structure is captured by bars $\beta_0$ and $\beta_2$ (orange bars) in Figure 2b. First, there are $\beta_0$ bars at the beginning. As the filtration progress, there are only one $\beta_0$ bar and one $\beta_1$ bar, which corresponds to void structure in the center of the cage. The initial six $\beta_0$ bars imply this void structure having six vertices. It is obvious that polygon and polyhedron structures, even with perturbations, could be described by $\beta_0$, $\beta_1$, and $\beta_2$ bars. More specific examples of barcode representation are shown in the Supporting Information.

Machine Learning Models. The success of a machine learning model for material sciences depends not only on the quality of original data, appropriate descriptors but also on learning algorithms. Gradient boosting regression (GBR) and neural network (NN) have been the methods of choice for modeling atomic interactions. In particular, the GBR is known to be robust to small data sets and less prone to overfitting. Compared with NN, GBR is very easy to train and much more efficient. Essentially, GBR constructs a strong regressor from an ensemble of weak regressors. Here, we choose GBR to optimize feature parameters, including the structure of the TFs, the number of subintervals in each angstrom, and $\epsilon_{\text{max}}$.

Once the optimized feature parameters are obtained, we further choose the neural network method for deep learning. Neural networks are modeled after the function of neurons in the brain. A neural network applies activation functions, called perceptrons, to inputs. During the training, weights of the neurons are updated through the backpropagation to minimize a loss function over many epochs, or passes of an entire training data set. The detail of the machine learning models used in this work was listed in the Supporting Information. In this work, to evaluate the performance of our models, mean absolute error (MAE) and Pearson correlation coefficient (PCC) were used. The MAE is given by

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} | f_i - y_i |$$

where $N$ is the number of samples and $f_i$ and $y_i$ are the prediction and true value of sample $i$, respectively. Given a pair of random variables $(X, Y)$, the formula of PCC is

$$\text{PCC} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where the cov$(X, Y)$ is the covariance of $X$ and $Y$ and $\sigma_X$ and $\sigma_Y$ are the standard deviations of $X$ and $Y$, respectively.

Particle Swarm Optimization (PSO). The next vital ingredient required in the prediction of the global stable structure of a cluster is a stochastic global optimization algorithm. Particle
swarm optimization (PSO) strategy is based on behaviors of swarms of flocks of birds, where a displacement velocity updates the position of each candidate particle. In this work, a candidate cluster structure is regarded as a particle. The \( j \)-th dimension of \( i \)-th particle is updated over a unit time according to

\[
x_{ij}^{t+1} = x_{ij}^t + v_{ij}^{t+1}
\]

where \( t \) denotes the generation index, \( x_{ij}^t \) is the coordinate of the \( j \)-th dimension of the \( i \)-th particle, and \( v_{ij}^t \) is its correspond displacement velocity per unit time. Each velocity \( v_{ij}^t \) is determined according to the formula

\[
v_{ij}^{t+1} = w v_{ij}^t + c_1 r_1 (x_{ij}^{sbest} - x_{ij}^t) + c_2 r_2 (x_{ij}^{ibest} - x_{ij}^t)
\]

where \( w \) is the inertia coefficient, \( x_{ij}^{sbest} \) is the value of the given coordinate from the best solution seen by the swarm, \( x_{ij}^{ibest} \) is the value of the given coordinate from the best solution seen by the \( i \)-th particle, and \( c_1 \) and \( c_2 \) are constants. Here, \( r_1 \) and \( r_2 \) are random numbers. The methodology has been demonstrated to be efficient and powerful for global structural convergence.\textsuperscript{12,66,67} The first generation of clusters are stochastically generated from crystal structures.

**Topology-ML-PSO-DFT Protocol.** Our topology-ML-PSO-DFT protocol consists of mainly four steps as depicted in the flowchart of Figure 3. First, the initial structures are randomly generated from crystal structures. Once a new structure generated, the TFs are calculated and used to examine the consistency of this structure with all the previous ones. When the TFs of the two structures are the same, we conclude that they are identical and discard the new structure. In this work, the number of subintervals in each angstrom is the essential parameter of our model and ultimately determines the length of the TFs and the precision for deduplication. The larger the number of subintervals per unit length, the more compatible the deduplication process is. Then, the PSO procedure is applied to generate new structures for the next generation. After all the structures of the final generation have been generated, the three lowest energy structures of every generation determined by the machine learning model were collected in a low-energy set. Eventually, the structures in the low-energy set are further optimized using DFT calculation to obtain more stable structures, among them the structure with the lowest binding energy was picked as the final globally stable structure.

**DFT Calculation.** For \( Li_n \) clusters, to match the results of previous studies,\textsuperscript{12,67} the DFT calculations were performed using the same method, namely, the plane-wave projector-augmented wave method\textsuperscript{68,69} implemented in the Vienna ab initio simulation package (VASP)\textsuperscript{70} with an energy cutoff of 520 eV. The generalized gradient approximation (GGA) with the Perdew–Burke–Enzerhof parametrization (PBE)\textsuperscript{71} was chosen as the exchange–correlation potential. The binding energy, the energy to form a cluster with joint atoms from isolated atoms in a vacuum, is defined as

![Figure 5. Comparison between persistent homology prediction and DFT calculation of the binding energy of (a) TFs on Li\(_{20}\), (b) GAP on Li\(_{40}\), and (c) TFs on Li\(_{40}\).](image_url)

![Figure 6. (a) Putative global stable structure of Li\(_{20}\). Two metastable structures of Li\(_{20}\) are depicted in (b), (c). (d) Putative global stable structure of Li\(_{40}\). The internal and external atoms are shown in red and green, respectively.](image_url)
where $E_{\text{cluster}}$, $E_{\text{total}}$, and $n$ are the energy of cluster, the energy of the sole atom, and the numbers of atoms, respectively. The energy was divided by $n$ is to make clusters with a different number of atoms comparable.

Data and Model Availability. To ensure the reproducibility of the results, the code and data used in this work were uploaded to Github. More details are given in the Supporting Information.

First, we generate some data of small-size clusters for machine learning and optimize essential parameters. With generated data and optimized parameters, the comparisons of the algorithm of GBR and NN were also carried out. Then, we apply the model to predict the most stable isomer for medium-size clusters.

Data Generation. Li is the lightest metallic element in the periodic table. Li clusters are thus considered to be prototype systems for understanding the various physical properties of simple metal clusters. Numerous theoretical studies have been performed to understand the structures of Li clusters. Here, we use the PSO method to generate the structures of Li clusters, where $n = 3, 4, ..., 10$, for optimizing the essential parameters and training our machine learning models. For each cluster, a search with a population size of 50 was performed and stopped at the 21st generations. All the above-mentioned DFT calculations were performed with ion step below 20. A total of 136 617 structures was produced. Meanwhile, it has also carried on the processing of duplication removal for similar structures with identical TFs. Finally, 70% of structures are used as the training set, with the other 30% being used as the testing set.

Parameter Optimization. First, the structure of TFs, i.e., the number of subintervals in each angstrom and $\epsilon_{\text{max}}$, should be optimized. The number of subintervals in each angstrom, $n_s$, is an essential parameter of our model and ultimately determines the length of the TFs and the precision for deduplication, which means that the $n_s$ affects the number of distinct structures we finally collect. Thus, we first optimize the structure of TFs and $\epsilon_{\text{max}}$ with $n_s = 10$. Subsequently, $n_s$ could be optimized. As for the structure of TFs, the topological invariants and PPI could be described by birth, death, and persistence.

We are interested in examining the utility of the newly proposed PPI. In the comparison of the statistics in the first three lines in Table 1, the MAE of the TFs without PPI is 0.039 eV/atom, and with PPI used the MAE is below 0.030 eV/atom. Therefore, it can be concluded that the accuracy of the present model has been significantly improved by the application of PPI as an auxiliary feature. More specifically, the death of PPI is more effective than the persistence of PPI.

As shown in Table 1, the combination of the persistence of $\beta_1$, $\beta_5$, and $\beta_7$ and the death of PPI would provide the lowest MAE and the highest PCC when the value of $\epsilon_{\text{max}}$ was the same. However, it is obvious when the structures of TFs were the same, the PCC would be increased in accordance with the increase of the value of $\epsilon_{\text{max}}$, but the increase was not obvious when $\epsilon_{\text{max}}$ was greater than 10 Å. Thus, we choose the combination of the persistence of $\beta_1$, $\beta_5$, and $\beta_7$ and death of PPI with $\epsilon_{\text{max}}$ of 10 Å as our TFs. These results are further compared with previous ML model, GAP, with GBR as the base ML algorithm. More details are shown in Supporting Information. The PCC of GAP the GAP model is 0.990.

In this work, since the structure of TFs and $\epsilon_{\text{max}}$ were settled, 119 219, 130 099, and 132 874 distinct structures linked to the binding energy were acquired after deduplication process from initial 136 617 structures with $n_s$ of 5, 10, and 20, respectively. We compared the MAE and PCC of the algorithm of the GBR and NN and found the latter algorithm had a better predictive ability. As shown in Figure 4, the MAEs of $n_s = 10$ and $n_s = 20$ by NN were minimal, and their PCCs were higher than others. A higher value of $n_s$ indicates a longer time training model needs. Thus, the followed work was all based on the algorithm of NN with $n_s$ of 10.

Performance on Li$_n$. Additionally, we use the same method to generate only two generation of Li$_{20}$ and Li$_{40}$ clusters for further testing our model. The learning task is to predict the cluster binding energy of Li$_{20}$ and Li$_{40}$ clusters with model learning on Li$_n$ ($n = 3, 4, ..., 10$) clusters. As shown in Figure 5, the model trained on Li$_n$ ($n = 3, 4, ..., 10$) cluster data performed well on Li$_{20}$ ($n = 3, 4, ..., 10$) (PCC = 0.993) and Li$_{40}$ clusters (PCC = 0.990) but did not do well on Li$_{40}$ clusters (PCC = 0.950). The MAEs of both Li$_{20}$ and Li$_{40}$ are larger than MAE of Li$_n$ ($n = 3, 4, ..., 10$). The larger MAEs were caused by the systematical error (overall migration of prediction compared to true value). Since we only concerned about relative energy values rather than absolute energy values (it has no effect on global stable structure search when all prediction energy values shift to lower or higher values overall), the quality of the model was evaluated by the PCC. The results showed that our model could be applied in a certain range (Li$_{20}$, $n$ was the maximum value of number of atoms learned) with wide adaptability and exactitude. The reason for its large deviation on Li$_{40}$ is that there are some contributions from long-range interactions in large clusters that cannot be learned from small-sized clusters. These results are further compared with GAP, as shown in Figure 5, on Li$_{20}$ TF model performs better.

Putative Globally Stable Structures of Li$_{20}$ and Li$_{40}$. As a case study, we test our structure prediction protocol on Li$_{20}$ and Li$_{40}$ clusters. Our topology-ML-PSO-DFT protocol mainly consists of two steps. First, a search with the population size of 2000 was performed and stopped at the 15th generation through the machine learning model, which avoided the expensive DFT calculation on screening local minimum values. Second, three predicted lowest energy structures (determined by our machine learning model) of every generation were collected for deduplication and optimized through DFT calculations. For each cluster, two independently topology-ML-PSO-DFT structural searches were performed and both searches found the same lowest energy structure.

Our predicted lowest energy structures of Li$_{20}$ is composed of three centered trigonal prisms with five additional capped atoms, as shown in Figure 6a. This is in agreement with the result of Ma et al. obtained by CALYPSO and Fournier et al. obtained by using the Tabu search in the descriptor space. The second and the third lowest energy structures are very similar. The only difference is that the middle layer of the first one is an isoceles trapezoid (Figure 6b) whereas the other is hexagonal (see Figure 6c). Ma et al. found the latter structure but not the former. As shown in Figure 6d, our predicted lowest energy structure of Li$_{40}$ is a 45-atom polyicosahedron with five missing vertex atoms. This is also in agreement with the previous work. More detail of this structure could be found in the Supporting Information. In order to demonstrate the role of the ML model in the structure prediction, three
independent topology-ML-PSO-DFT searches of Li_{30} and three independently PSO-DFT searches of Li_{30} with a population size of 20 were performed and stopped when the lowest energy structures were found. Since the DFT calculation is far more computationally expensive than the ML model prediction, only the DFT calculation is used to measure the efficiency. The average number of DFT calculations taken of topology-ML-PSO-DFT is 21 and of PSO-DFT is 480. The ratio of average CPU time used of topology-ML-PSO-DFT/PSO-DFT is 1:25. Of course, a large DFT-based data set is required for training the ML model, which is even more computationally expensive than the structure search. However, if we did not carry out just one, but two or more structure searches, the ML-PSO-DFT would reach orders of magnitude speedup compared to PSO-DFT. For example, ML model training on Li_{n} (n = 3, 4, ..., 10) could be applied to the structure search of Li_{n} (n = 11, 12, ..., 20).

In conclusion, data representation is one of three major ingredients of machine learning (ML) studies. It becomes more important as the data complexity increases. A wide variety of data representation has been proposed for material sciences. However, topology has been hardly introduced to the field. In this work, we introduce persistent homology, a new branch of algebraic topology, as a new tool for representing metal cluster structures. Unlike conventional topology, persistent homology bridges geometry and topology, achieving an interplay between geometry and topology without downgrading important geometric information of structures. Persistent homology extracts low-dimensional topological invariants from high-dimensional data. It has been successfully applied to various data science problems. To enhance the predictive power of persistent homology, we propose persistent pairwise independence (PPI) as an auxiliary feature. Persistent barcodes of both topological invariants and PPI are utilized as cluster descriptors. Our topology-based ML models are found to offer highly accurate predictions of lithium (Li) cluster binding energies. Then, a systematic predictive protocol for generating the globally stable cluster structure is constructed by integrating a topology-based ML, particle swarm optimization (PSO), and density functional theory (DFT) calculation. The proposed topology-ML-PSO-DFT protocol is validated on medium-sized Li_{n} clusters. The predicted globally stable structure is in agreement with that in the literature. Therefore, the current methodology has been proved to be a reliable approach for cluster structure prediction. It is worthy to mention that this is the first time that topology and ML are applied to avoid the expensive DFT calculation on screening local minimum values. However, there are many aspects to be improved, such as the precision and generalization. Moreover, the method accelerating crystal structure relaxation by ML could be integrated with our method to further accelerate structure prediction in future work. This work is a precursor for the prediction of multielement clusters and the structure prediction of complex crystal surface adsorption.

ASSOCIATED CONTENT

The Supporting Information is available free of charge on the ACS Publications Web site. The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.0c00974.

Parameters of machine learning models, more specific examples of barcode representation, putative globally stable structures of Li_{40} (PDF)

AUTHOR INFORMATION

Corresponding Authors
Guo-Wei Wei — Department of Mathematics, Michigan State University, East Lansing, Michigan 48824, United States; Email: weig@msu.edu
Feng Pan — School of Advanced Materials, Shenzhen Graduate School, Peking University, Shenzhen 518055, People’s Republic of China; Email: panfeng@pkusz.edu.cn

Authors
Xin Chen — School of Advanced Materials, Shenzhen Graduate School, Peking University, Shenzhen 518055, People’s Republic of China
Dong Chen — School of Advanced Materials, Shenzhen Graduate School, Peking University, Shenzhen 518055, People’s Republic of China
Mouyi Weng — School of Advanced Materials, Shenzhen Graduate School, Peking University, Shenzhen 518055, People’s Republic of China
Yi Jiang — School of Advanced Materials, Shenzhen Graduate School, Peking University, Shenzhen 518055, People’s Republic of China

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c00974

Author Contributions
Xin Chen designed the project and carried out the calculations. Xin Chen, Dong Chen, Yi Jiang, Mouyi Weng, Guo-Wei Wei, and Feng Pan discussed the results, analyzed the data, and drafted the manuscript. Guo-Wei Wei and Feng Pan conceptualized the project and obtained funding.

Notes
The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The research was financially supported by Soft Science Research Project of Guangdong Province (2017B030301013), National Key R&D Program of China (2016YFB0700600) and Shenzhen Science and Technology Research Grant (ZDSYS201707282106184). The work of Guo-Wei Wei was supported in partial by NSF Grants DMS1721024, DMS1761320, IIS1900473, NIH grants GM126189 and GM129004, Bristol-Myers Squibb, and Pfizer.

REFERENCES
