





REMatch plus SOS: Machine-learning-accelerated structure prediction for supported metal nanoclusters

Yunyu Zhang (张昀宇) ¹, Keith T. Butler ¹, Michael D. Higham ^{1,2} and C. Richard A. Catlow ^{1,2,3}

¹*Kathleen Lonsdale Materials Chemistry, Department of Chemistry, University College London, 20 Gordon Street, London WC1H 0AJ, United Kingdom*

²*Research Complex at Harwell, Rutherford Appleton Laboratory, Harwell Oxford, Didcot, Oxon OX11 0FA, United Kingdom*

³*Cardiff University, School of Chemistry, Main Building, Park Place, Cardiff CF10 3AT, United Kingdom*



(Received 4 October 2024; accepted 28 January 2025; published 3 March 2025)

Predicting stable structures of nanoclusters is crucial yet computationally demanding. This study presents a machine learning-based methodology designed to accelerate the prediction of stable structures in nanoclusters. By integrating local environment descriptors, with dimensionality reduction, kernel-based similarity measure, and outlier detection, we efficiently screen and select promising configurations, thus accelerating identification of global and local minimum structures. The approach is validated through rigorous optimization, demonstrating its capability to identify low-energy structures while significantly reducing computational costs. This method offers a robust framework for structural screening.

DOI: [10.1103/PhysRevMaterials.9.033801](https://doi.org/10.1103/PhysRevMaterials.9.033801)

I. INTRODUCTION

Structure prediction is critically important in materials science, particularly for supported metal nanocluster systems, which have been identified as promising candidates for materials with tailored properties for applications ranging from catalysis, to electronics, to nanotechnology. Due to their small size and large surface area, isolated metal nanoclusters exhibit chemical and physical properties that differ significantly from those of their corresponding bulk materials [1,2]. Furthermore, in practice, metal nanoclusters are almost invariably prepared as adsorbed clusters on some support material; while the support may be considered merely an inert substrate to prevent sintering or nanocluster agglomeration, and thus preserving the high surface area and low coordination environments of the small nanoclusters, it is well established that strong metal-support interactions can result in synergistic effects at metal-support interfaces, which may enhance or inhibit the properties of the material [3–7]. As such, there is much interest in accurately predicting the structure of low energy structures for supported metal nanoclusters. However, obtaining structural information for small nanoclusters through experiments is challenging, making computational screening a vital tool for predicting these structures. Accurate predictions of nanocluster structures enable the optimization of material performance, providing crucial insights for the design of materials in applications such as catalysis, electronics, and optics.

Traditional structure prediction methods, such as Monte Carlo simulations [8–10], random quenching [11–13], simulated annealing [14–17], genetic algorithms [18–20], particle swarm algorithms [21–23], and other methods, have been the cornerstone of global optimization techniques for exploring energy landscapes. These studies have predominantly focused on exploring energy landscapes by applying specific global optimization techniques to particular systems [2,24–27]. While effective, these methods are often computationally intensive, especially when applied to complex systems with large numbers of atoms. The vast configurational space, even for relatively small nanoclusters (i.e. 2–4 nm in diameter), makes exhaustive searches computationally expensive and time-consuming. The high computational cost and the need for extensive sampling to ensure accurate predictions make these techniques less practical for large-scale or high-throughput studies, highlighting the need for more efficient alternatives.

In recent years, machine learning approaches have emerged as promising tools to enhance computational efficiency and reduce the overall cost of structure prediction for complex systems like nanoclusters. For example, McCandler *et al.* developed a machine-learned interatomic potential to study gold–thiolate nanocluster dynamics, significantly accelerating simulations while maintaining accuracy [28]. Ko *et al.* introduced a fourth-generation high-dimensional neural network potential that incorporates accurate electrostatics, enabling more realistic simulations of charge-transfer phenomena in metal clusters [29]. Behler and Parrinello further demonstrated the effectiveness of neural network potentials in describing high-dimensional potential energy surfaces, transforming how molecular dynamics simulations are performed for bulk materials and nanostructures alike [30]. Such advances demonstrate the potential of machine learning approaches in structural prediction

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

and dynamic studies of nanoclusters, effectively addressing some computational challenges associated with traditional methods.

However, despite these developments, challenges remain in applying machine learning methods to a broader range of complex systems, particularly where training data are scarce, computational costs are still significant, or dealing with complex interactions in heterogeneous systems. Therefore, there is still a need for innovative strategies that can further streamline and accelerate the prediction process while maintaining accuracy.

To address this critical challenge, this study introduces a machine learning-based approach that significantly accelerates the structure prediction process while maintaining high accuracy. We propose a methodology that integrates the Regularized Entropy Match (REMatch) kernel and Stochastic Outlier Selection (SOS) with traditional structural descriptors like Smooth Overlap of Atomic Positions (SOAP). The REMatch kernel offers a robust similarity measure that captures both local and global structural characteristics, allowing for precise comparison of nanocluster configurations. SOS, on the other hand, efficiently filters out irrelevant or less promising configurations, focusing computational resources on the most likely candidates for global minima.

By combining these machine learning techniques, we provide a framework that not only accelerates the prediction of stable structures in nanoclusters but also reduces the overall computational cost. This approach is applied to the study of Cu clusters on various reconstructed polar ZnO surfaces, demonstrating its efficacy in navigating the complex configurational space of these systems.

To validate this framework, we applied it to the study of Cu clusters on various reconstructed polar ZnO surfaces, following the work of Higham *et al.* [31], which provided a comprehensive analysis of the morphology of Cu clusters on ZnO (0001) and (000 $\bar{1}$) surfaces, and thus serves as an ideal case study for verifying our approach. This case study allows us to evaluate the performance of our framework in navigating the complex configurational space and predicting stable structures for Cu/ZnO systems.

This article is structured as follows: in the Methodology section, we detail the machine learning techniques and computational approaches employed in our framework, including the REMatch kernel and SOS. In the Results section, we apply our framework to the Cu/ZnO system, conducting comparative analysis, sensitivity analysis, and divergence analysis. Finally, in the Conclusions section, we summarize our findings, discuss the broader implications of our work, and suggest potential directions for future research in the field of machine learning-assisted structure prediction.

II. METHODOLOGY

A. Data collection

We used the Knowledge-Led Master Code (KLMC) to generate a comprehensive dataset for analyzing the morphology of Cu clusters on various reconstructed polar ZnO (0001) and (000 $\bar{1}$) facets. KLMC is a robust and versatile tool for performing unbiased Monte Carlo (MC) simulations to ex-

plore the energy landscape of complex systems [1,32–34]. The dataset includes configurations of Cu clusters on four distinct ZnO surfaces: Zn-rich O-terminated surface, O-poor O-terminated surface, O-rich Zn-terminated surface, and Zn-poor Zn-terminated surface; these were determined to be the most stable under typical conditions as revealed by the grand canonical ensemble studies performed previously, as reported by Mora-Fonz *et al.* [35].

For each ZnO surface, we considered Cu_n clusters of a range of sizes ($n = 4\text{--}8, 16, 24, 36$). For each cluster size and surface combination, KLMC generated 10 000 initial Cu_n/ZnO random configurations, resulting in a total of 320 000 configurations. This methodology involved the permutation of n Cu atoms across a mesh of 147 sites (distributed across three layers of 49 sites each) atop the two most stable reconstructed polar ZnO surfaces identified from preceding research efforts [35].

Each configuration is stored in a ".gin" file format, which is the input format for the General Utility Lattice Program (GULP). The ".gin" file contains critical structural information for each configuration, including Atomic Types, specifying the element types of the atoms in the structure (e.g., Cu, Zn, O), and Atomic Coordinates, providing the Cartesian coordinates (x, y, z) of each atom within the structure. This format ensures that all necessary atomic and spatial data are accurately preserved for subsequent computational analysis.

B. Assumptions

Our proposed scheme for reduced sampling of the configurational space is based on several assumptions about the shape of the potential energy surface. We now describe and justify these assumptions.

1. Structurally similar initial configurations have a higher probability of converging to similar local minima

We assume that, in most cases, Cu clusters on different polar ZnO surfaces with structurally similar initial configurations have a higher probability of converging to similar local minima after geometry relaxation. This assumption is based on principles from energy landscape theory, acknowledging the probabilistic nature of structural evolution during relaxation processes [36]. Systems that start from similar configurations are more likely to follow similar paths on the potential energy surface, leading to analogous final structures, although this relationship is not strictly deterministic. Our preliminary simulations also indicate that Cu clusters with similar initial configurations on ZnO surfaces frequently relax to similar local minima, supporting this assumption.

This assumption is crucial in the context of our study, where efficient prediction of Cu cluster morphologies on ZnO surfaces relies on capturing the relationship between initial and final configurations. Based on this assumption, we perform our sampling using the initial suggested configurations before any expensive geometry relaxation, enabling us to reduce computational redundancy by filtering out structurally similar initial configurations.

Given the probabilistic relationship between initial and final structures, we utilize Smooth Overlap of Atomic Positions (SOAP) descriptors, which are specifically designed

to capture the spatial distribution and environment of atoms [37]. These descriptors are ideal for representing the initial configurations of Cu clusters on ZnO surfaces, providing a high-dimensional representation that effectively encodes the local atomic environment and allows us to quantitatively compare and analyse structural similarities.

By representing the initial configurations using SOAP descriptors and computing structural similarities with the Regularized Entropy Match (REMatch) kernel, we can identify and exclude structurally similar initial configurations before performing computationally expensive geometry optimizations. This strategy enables us to focus computational resources on a more diverse set of configurations, enhancing the efficiency of our study without significantly compromising structural diversity. The effectiveness of combining SOAP descriptors with the REMatch kernel in capturing structural similarities has been demonstrated in previous research, [38] further supporting our methodology.

2. Global minimum energy structures are typically surrounded by structurally and energetically similar local minima

We assume that global minimum energy structures are typically surrounded by structurally and energetically similar local minima. This assumption is grounded in the well-established concept of energy landscapes in materials science and chemistry, where the global minimum represents the most stable configuration [36]. Typically, this global minimum is embedded within a basin of attraction, consisting of several nearby local minima that share similar structural characteristics and are separated by relatively small energy barriers. These barriers are often low enough to allow transitions between these minima through minor perturbations of atomic positions [39].

This clustering of similar structures around the global minimum is a well-documented phenomenon in the study of atomic clusters and condensed matter systems [40]. Energy landscapes are often funnel-shaped, guiding the system toward the most stable configurations during relaxation processes. As a result, global minima are seldom isolated points but rather central hubs within a network of structurally related local minima.

Empirical evidence supporting this hypothesis includes extensive studies on the structural distribution of metal clusters, where global minima are observed to be at the core of a network of similar configurations [8,41]. This behavior has been demonstrated through computational simulations and experimental studies of various atomic and molecular systems. By assuming the presence of structurally and energetically similar structures around global minima, we can effectively employ clustering analysis to identify and categorize these configurations, thereby improving the efficiency and accuracy of our structure prediction methods.

3. As the cluster size increases, the proportion of low-energy structures decreases

We assume that as the size of Cu clusters increases, the complexity of the energy landscape grows, leading to a relative decrease in the number of distinct low-energy structures. This assumption is based on the principles

of cluster chemistry, where larger systems exhibit more intricate energy landscapes characterized by a higher density of local minima, but a lower proportion of distinct low-energy structures [36].

As the size of a cluster increases, the number of possible atomic configurations increases exponentially, resulting in a more complex energy landscape. This complexity arises from the increased degrees of freedom and the multiple ways atoms can arrange themselves to minimize energy. In smaller clusters, the energy landscape is simpler with fewer configurations, making it easier to identify distinct global minima. However, in larger clusters, the landscape becomes more rugged, with many near-degenerate configurations, making distinct low-energy structures relatively rarer.

Empirical evidence supporting this assumption is found in studies of atomic and molecular clusters that investigate the relationship between cluster size and the number of stable configurations. For instance, in metal clusters, smaller clusters often exhibit well-defined global minima, whereas larger clusters display a greater variety of nearly degenerate configurations [25,42]. This trend is consistent across various types of clusters and materials, underscoring a fundamental aspect of energy landscape theory.

The decrease in the number of distinct low-energy structures as the cluster size increases has significant implications for the stability and properties of these clusters. Smaller clusters, with fewer distinct global minima, tend to have well-defined stable structures, and exhibit less structural diversity. In contrast, larger clusters, with their more complex energy landscapes, are prone to structural fluctuations and a higher degree of polymorphism. This behavior directly influences their chemical reactivity, catalytic properties, and overall stability.

C. Smooth Overlap of Atomic Positions (SOAP) with feature compression

In this study, we employ the Smooth Overlap of Atomic Positions (SOAP) descriptor to represent the local atomic environments of Cu clusters on ZnO surfaces. SOAP characterizes atomic environments by comparing the overlap of Gaussian-smearred atomic densities centered on each atom, effectively capturing both radial and angular structural information. [38,43]

We choose SOAP because it provides a continuous and differentiable representation of atomic structures, which is particularly suited for quantifying similarities and differences between complex configurations [44]. This high-dimensional descriptor encodes detailed information about the spatial distribution of atoms, enabling precise comparison and analysis of various cluster configurations. Its robustness and accuracy make it ideal for identifying subtle variations in atomic arrangements, which is essential for our study's focus on structural prediction.

However, transforming configurations into SOAP descriptors often results in extremely high-dimensional data, potentially involving tens of thousands of features. Such high dimensionality can lead to computational inefficiencies and challenges in discerning meaningful similarities due to the curse of dimensionality. To address this, we apply a

feature compression scheme known as the $\mu = 1, \nu = 1$ feature compression scheme, as introduced by Darby *et al.* [45].

By implementing feature compression within the SOAP framework, we significantly reduce the dimensionality of the feature vectors while preserving essential structural information. Specifically, this compression combines the coefficients associated with each atomic species, resulting in a more manageable number of features. In our case, we set the target element to Cu, focusing on the Cu atoms within the clusters. After compression, each structure is represented as an $n \times 36$ matrix, where n is the number of Cu atoms in the cluster. This compressed representation scales linearly with the number of atoms, enhancing computational efficiency without compromising the descriptor's ability to capture critical structural characteristics.

By employing the SOAP descriptor with feature compression, we achieve a balance between retaining detailed structural information and reducing computational complexity. This approach enables us to effectively compare and cluster configurations of Cu clusters on ZnO surfaces, facilitating the identification of structurally similar configurations. It aligns with our study's objectives to streamline the structure prediction process while maintaining high accuracy.

D. Regularized Entropy Match kernel (REMatch kernel)

After obtaining the compressed SOAP descriptors, where each structure is represented as an $n \times 36$ matrix, we need a method to assess the similarity between different structures effectively. To achieve this, we employ the Regularized Entropy Match (REMatch) kernel.

The REMatch kernel computes the similarity between two structures by optimally matching their local atomic environments, as described by their SOAP descriptors [38]. It compares the sets of atomic environments in each structure, balancing the influence of the best-matching environments with the overall distribution of similarities. This approach makes the REMatch kernel particularly robust against outliers and variations in atomic configurations.

We use the REMatch kernel because it provides a flexible and accurate measure of structural similarity that can handle the complexity of our dataset. By adjusting a parameter α within the kernel, we can control the emphasis between focusing on the most similar local environments and considering the average similarity across all environments. This flexibility allows us to capture meaningful similarities even when structures exhibit minor differences or distortions, which is essential for clustering and analyzing a large set of configurations.

Using the REMatch kernel, we compute the similarity between each pair of structures in our dataset. Given that we have 10 000 structures in each case, this involves pairwise comparisons resulting in a $10\,000 \times 10\,000$ similarity matrix. Each element of this matrix represents the similarity score between a pair of structures, ranging from 0 to 1, where 0 indicates no similarity and 1 indicates identical structures. This provides a comprehensive map of the structural relationships within our dataset.

This similarity matrix serves as the foundation for subsequent clustering and analysis. By quantifying the structural

similarities between configurations, we can efficiently identify and group structurally similar configurations. This enables us to reduce computational redundancy by focusing resources on a diverse set of configurations, aligning with our objective to streamline the structure prediction process while maintaining high accuracy.

E. Approximation algorithm

As previously discussed, computing the REMatch kernel for all pairs of structures in our dataset would result in a $10\,000 \times 10\,000$ similarity matrix. While this exhaustive computation provides comprehensive similarity information, it is computationally intensive and impractical for large datasets. To address this challenge and accelerate the computation, we designed an approximation algorithm that significantly reduces the computational load while maintaining acceptable accuracy.

The primary objective of this approximation algorithm is to reduce the number of pairwise comparisons required, thereby expediting the overall computation without substantially compromising the quality of the similarity measurements. The algorithm operates as follows:

(1) Grouping structures: The initial set of 10 000 structures is divided into 10 groups, each containing 1000 structures. This segmentation simplifies the subsequent computations by enabling pairwise comparisons within and between smaller subsets of structures.

(2) Intragroup REMatch kernel calculation: For each group, the REMatch kernel is calculated within the group, producing a 1000×1000 similarity matrix. This step is relatively efficient due to the smaller size of each group.

(3) Intergroup REMatch kernel calculation: REMatch kernel calculations are performed between pairs of consecutive groups (e.g., between group 1 and group 2, group 2 and group 3, etc.). This results in nine additional 1000×1000 similarity matrices for the adjacent group pairs.

(4) Constructing the approximate similarity matrix: A $10\,000 \times 10\,000$ matrix S is initialized to store the computed similarities. The intragroup and intergroup similarity matrices are placed into the appropriate sections of S , filling in the corresponding entries. At this stage, the similarity matrix S contains computed similarity values within groups and between adjacent groups, while the remaining entries are uncomputed (initially set to zero).

(5) Iterative approximation to fill the matrix: To estimate the uncomputed similarities, we apply an iterative process that leverages the known similarities. For each structure i we identify its most similar structure j in the adjacent group using the intergroup similarity matrices. If structure i in group k is most similar to structure j in group $k + 1$, we consider them closely related. We then approximate the similarities between structure j and other structures in Group k by using the similarities between structure i and those structures. This process is repeated iteratively across all groups, effectively propagating similarity information and filling in the uncomputed entries of the similarity matrix S .

(6) Converting similarity to distance: After completing the similarity matrix S , where each element ranges from 0 (no

similarity) to 1 (identical structures), we convert it into a distance matrix D using the relation $D = 1 - S$. This distance matrix is suitable for subsequent analyses, such as outlier detection.

By grouping structures and focusing computations on intragroup and adjacent intergroup similarities, we capture the most relevant similarity information while avoiding unnecessary computations between structures that are likely to be less similar. The iterative approximation effectively propagates similarity information throughout the matrix, enabling us to construct a sufficiently accurate distance matrix for our purposes.

The pseudocode and computational complexity of this algorithm are shown in Secs. S1 and S2 in the Supplemental Material [46]. Through computational complexity analysis, the computational complexity of this algorithm is $O(14 \times 10^6 \times n^2)$, and if the algorithm is not applied, the computational complexity would be $O(5 \times 10^7 \times n^2)$. Therefore, the ratio of complexity reduction achieved by applying this algorithm is $\frac{5 \times 10^7 \times n^2}{14 \times 10^6 \times n^2} \approx 3.57$. This algorithm significantly reduces the computational load compared to a full pairwise comparison while maintaining a sufficiently accurate distance matrix for subsequent outlier detection and structural characterization.

F. Stochastic outlier selection (SOS)

After constructing the approximate distance matrix using the REMatch kernel and our approximation algorithm, the next step is to strategically screen the structures to identify and filter out configurations based on their outlier probabilities. This process is crucial for reducing computational redundancy and focusing on structurally significant variants within our dataset of Cu clusters on ZnO surfaces.

To achieve this, we employ the Stochastic Outlier Selection (SOS) algorithm, an unsupervised method designed to compute an outlier probability for each data point based on its dissimilarity (distance) to others [47]. The core idea of SOS is that a data point is considered an outlier if other data points have insufficient affinity with it. This approach allows us to quantify how atypical each structure is within the context of the entire dataset.

The SOS algorithm begins by converting the distance matrix into an affinity matrix. The affinity between two data points decreases in a Gaussian-like manner relative to their dissimilarity. Each data point has an associated variance that depends on the density of its neighborhood, controlled by a parameter known as perplexity. This ensures that each data point effectively has the same number of neighbors, allowing for consistent comparisons across the dataset. The affinity matrix represents how strongly each structure is connected to others based on structural similarity.

Once the affinity matrix is computed, it is normalized to create a binding probability matrix where each row sums to one. This normalization transforms the affinities into probabilities that reflect the likelihood of each structure being similar to others. The outlier probability for each structure is then calculated as the joint probability that other structures will not bind to it. A high outlier probability indicates that a

structure is not closely related to any other structures and is thus considered an outlier.

Based on the computed outlier probabilities, we categorize the structures to strategically select or discard them. Structures with high outlier probabilities (e.g., greater than 75%) are considered super outliers and are removed, as they represent extreme configurations unlikely to contribute meaningful insights. Structures with low outlier probabilities (e.g., less than 30%) reside at the centers of dense clusters and are also discarded, as they represent redundant information due to high structural similarity with many other configurations. Structures with medium outlier probabilities (e.g., between 55% and 75%) are retained selectively, as they may represent unique structural variants or potential alternative configurations. The specific thresholds for these categories are informed by sensitivity analyses and can be adjusted based on the desired balance between dataset size and structural diversity.

The specific thresholds of 30%, 55%, and 75% were determined based on sensitivity analyses conducted in subsequent sections of our study. Through these analyses, we empirically established thresholds that balance the need to reduce computational redundancy while retaining sufficient structural diversity. The thresholds are not universally fixed and can be adjusted depending on the characteristics of the dataset and the desired balance between dataset size and diversity. By fine-tuning these parameters, we ensure that the selection process aligns with our goals of retaining structurally significant variants and excluding redundant or extreme configurations.

By employing the SOS algorithm in this manner, we strategically filter the dataset to focus on structurally significant variants. This approach ensures that we retain configurations that are likely to contribute valuable insights into the structural landscape of Cu clusters on ZnO surfaces. It aligns with our overarching objective to efficiently explore and predict the morphology of these clusters by emphasizing the most diverse and relevant configurations.

This methodology builds upon the distance matrix derived from the REMatch kernel and our approximation algorithm. By first efficiently computing structural similarities and then strategically filtering the dataset based on outlier probabilities, we create a streamlined and focused set of configurations for further analysis and geometry optimization. This integrated approach enhances computational efficiency without significantly compromising accuracy, allowing us to concentrate computational resources on analyzing structures that are both significant and representative of the potential configurational space.

G. Verification

The flow chart of our methodology is shown in Fig. 1. The goal of our methodology is to efficiently identify a subset of initial configurations that is highly likely to include the global minimum energy structure and other low-energy structures, thereby significantly reducing the computational effort required for full optimization. By focusing on approximately 30% of the initial configurations (around 3000 out of 10 000), we aim to capture low-energy structures while conserving both time and computational resources. To assess the

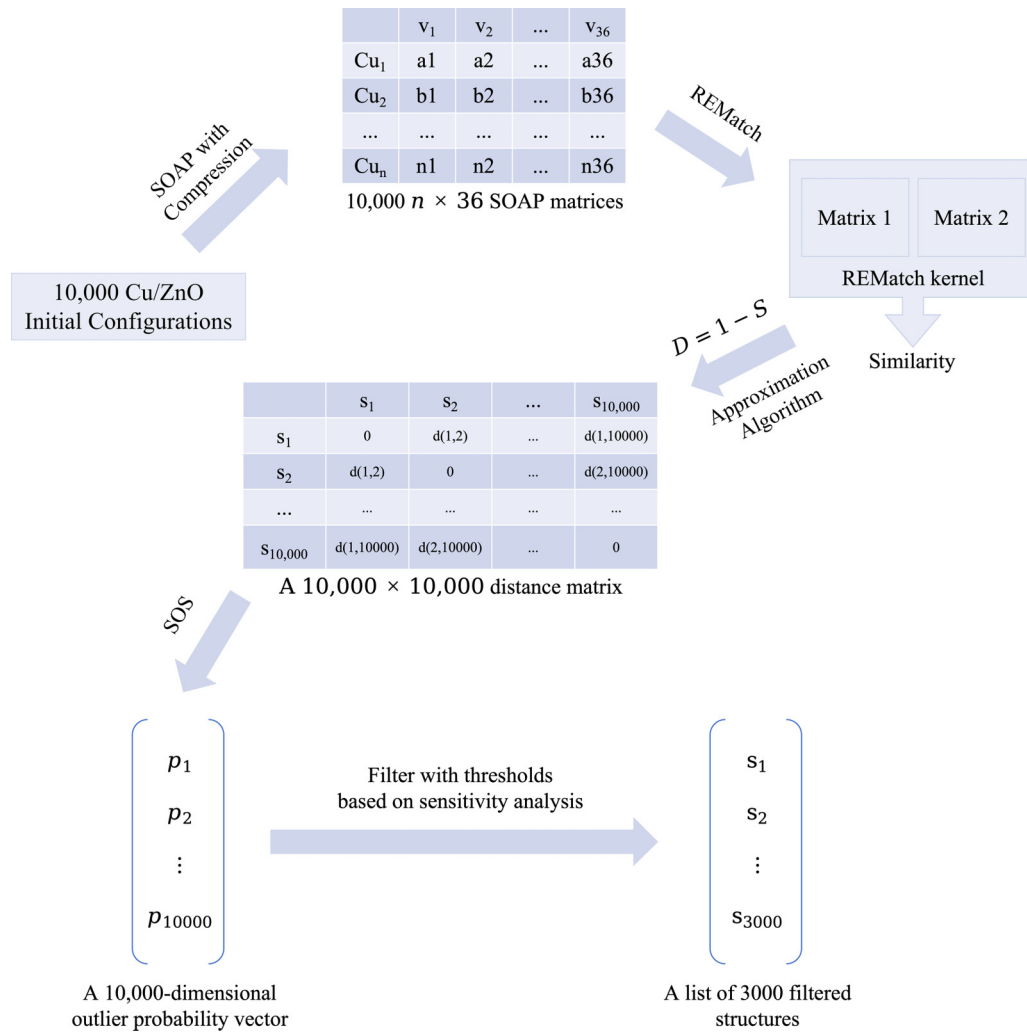


FIG. 1. Flow chart of our methodology workflow.

effectiveness of our approach, we performed a series of verification steps:

(1) Full optimization of initial configurations: To determine the final structures and energies of the 10 000 initial configurations, we performed a comprehensive optimization using the General Utility Lattice Program (GULP). This process involved applying a polarizable shell model potential to the oxygen atoms in ZnO, with interaction potentials for Cu-ZnO derived from Buckingham and Morse potentials for Cu-O and Cu-Zn interactions, respectively. Cu-Cu interactions were modeled using many-body Gupta potentials [48,49]. The potentials applied in the present work were robustly devised in the previous studies by fitting the IP parameters representing the Cu-O and Cu-Zn interactions to a set of data comprising a series of geometries and energies obtained with periodic single-point (SP) DFT [48]. For the ZnO potentials, fitting was performed with reference to a large number of experimentally measured parameters [50], with a similar approach being applied to obtain the Cu-Cu potentials [49]. These potentials have been previously applied to investigate Cu clusters supported on nonpolar and reconstructed polar ZnO surfaces, using unbiased Monte Carlo global optimization techniques to identify low-energy struc-

tures [31,48]; DFT refinement of the IP-obtained structures in both studies showed that the IP structures and their energy rankings are well reproduced upon DFT refinement. Hence, in the present work, the objective is to apply machine learning techniques to expedite the structural search process, rather than to validate the potentials themselves, which has already been demonstrated by the aforementioned previous studies. During optimization, all Cu atoms and the top three layers of ZnO were fully relaxed. The optimization was conducted using a 2D periodic surface model, with charge compensation applied to account for surface reconstruction effects. The BFGS algorithm was used to minimize iteratively the energy of each configuration, yielding the final relaxed structures and their corresponding energies. These optimized results serve as the basic data for evaluating the effectiveness of our structural screening methodology [31,35].

(2) Identification of global minima and low-energy structures: After optimization, we ranked all structures based on their final energies to identify the global minimum energy structure as well as low-energy structures. These low-energy structures are of particular interest as they provide insights into the stability and diversity of the energy landscape. We compared the global minima and low-energy structures

identified in the full set of 10 000 configurations with those present in our selected subset of approximately 3000 structures. Our methodology is considered successful if the subset includes the global minimum or at least one of the low-energy structures, ensuring that configurations of significant interest are captured.

(3) Comparison with random baseline: To evaluate the performance of our method relative to random sampling, we conducted comparisons across several aspects:

(a) Success Index: We defined a success index that varies according to the size of the Cu clusters, reflecting our third assumption that larger clusters generally have fewer low-energy structures due to increased energy landscape complexity. We compared the success rates of our method and random sampling to assess the effectiveness of our approach in capturing critical low-energy structures.

(b) Local minima capture: We calculated the total number of distinct local minima, the number of distinct local minima captured by our method, and the number of distinct local minima captured by random sampling. By comparing the differences between the total and our method versus the total and random sampling, we assessed our method's ability to capture a broader range of local minima. This comparison demonstrates that our method captures more local minima than random sampling, highlighting its effectiveness in exploring the energy landscape.

(c) Global minima capture: We evaluated whether our method and random sampling could capture the global minimum under different conditions and cluster sizes. This analysis provides insight into the robustness and reliability of our method compared to random sampling.

(4) Distribution consistency evaluation: To assess whether the distribution of the selected subset is consistent with that of the full set of configurations, we calculated Kullback-Leibler (KL) divergence [51] and Jensen-Shannon (JS) divergence [52] between them [53,54]. These measures quantify the differences between the probability distributions of the full dataset and the selected subset. A lower divergence indicates that the subset's distribution closely matches the overall distribution, ensuring that structural diversity is preserved. This evaluation verifies that our method retains the essential characteristics of the original dataset while reducing computational effort.

This verification process allows us to assess rigorously the effectiveness of our methodology. Such validation is essential to ensure that our machine learning-based structural screening method reliably captures the key structural features and stability characteristics of Cu clusters on ZnO surfaces.

III. RESULTS

To validate the effectiveness of our machine learning-based methodology, we applied it to the Cu/ZnO system, which serves as a representative example of complex supported nanocluster systems. By comparing the performance of our method against the results obtained from full optimization of all configurations, we assessed both the accuracy and efficiency of our approach. Specifically, we evaluated whether our selected subset consistently includes the global minimum

energy structures and how well it represents the overall energy landscape of the system.

In this section, we provide a detailed account of the results obtained from applying our methodology. We present comparative evaluations with the full population and with a random baseline, conduct a sensitivity analysis, and perform a divergence evaluation. These analyses demonstrate the robustness and reliability of our approach in efficiently identifying low-energy configurations while preserving structural diversity.

A. Comparative evaluation with the population

In this section, we compare the global minimum structures identified from the original full set of 10 000 configurations with those predicted from the subset of 3000 configurations selected by our method. The goal of this comparative analysis is to assess the accuracy and reliability of our screening approach in predicting the most stable configurations, particularly focusing on adsorption energies and structural characteristics.

Adsorption energy is commonly used to measure the stability and interaction strength of Cu clusters on specific surfaces. It is defined as

$$E_{\text{adsorption}} = E_{\text{cluster_min}} - E_{\text{clean_slab}} - nE_{\text{bulk_Cu}}, \quad (1)$$

where $E_{\text{cluster_min}}$ is the energy of the lowest energy structure of the Cu_n cluster, $E_{\text{clean_slab}}$ is the energy of the clean slab (the surface energy without Cu clusters), n is the number of Cu atoms adsorbed on ZnO, and $E_{\text{bulk_Cu}}$ is the bulk Cu energy per formula unit.

Negative adsorption energy indicates that the adsorption process is exothermic, implying that the presence of Cu clusters adsorbed on the surface is more stable than the isolated components of the same number of Cu atoms in the bulk phase and the clean ZnO surface. Conversely, positive adsorption energy suggests a nonspontaneous adsorption process. In our study, adsorption energy serves as a crucial metric to evaluate whether the predicted global minima from the subset align with those identified from the full set of configurations.

To ensure an accurate comparison between the original global minima obtained from full optimization and the predicted global minima identified by our methodology, we establish size-dependent energy tolerances based on the Rank 2 relative energies reported by Higham *et al.* [31]. Specifically, we use the energy difference between the second lowest energy structure (Rank 2) and the global minimum energy structure as the tolerance for each cluster size and ZnO surface. This approach accounts for the inherent structural complexity and energy variability associated with different cluster sizes and surfaces. We include a table (Table I) summarizing the Rank 2 relative energies for various Cu cluster sizes on different ZnO surfaces, adapted from Higham *et al.* [31].

The following comparative tables present the structural visualizations of both the original and predicted global minima for each ZnO surface type, along with their corresponding adsorption energies: Table II for O-poor O-terminated surface, Table III for Zn-rich O-terminated surface, Table IV for Zn-poor Zn-terminated surface, and Table V for O-rich Zn-terminated surface. To simplify notations, we use O-p-O

TABLE I. Rank 2 relative energies (in eV, tolerance in our case) for adsorbed Cu clusters of various sizes on different ZnO surfaces, where O-p-O represents O-poor O-terminated surfaces, Zn-r-O represents Zn-rich O-terminated surfaces, Zn-p-Zn represents Zn-poor Zn-terminated surfaces, and O-r-Zn represents O-rich Zn-terminated surfaces.

Surface type	No. Cu							
	4	5	6	7	8	16	24	36
O-p-O	0.110	0.072	0.306	0.299	0.206	0.086	0.151	0.194
Zn-r-O	0.446	0.060	0.070	0.068	0.074	0.199	0.142	0.127
Zn-p-Zn	0.357	0.329	0.174	0.078	0.048	0.007	0.049	0.128
O-r-Zn	0.153	0.266	0.353	0.119	0.054	0.463	0.287	0.410

for O-poor O-terminated, Zn-r-O for Zn-rich O-terminated, Zn-p-Zn for Zn-poor Zn-terminated, O-r-Zn for O-rich Zn-terminated.

The comparative evaluation of the original and predicted global minimum structures across different ZnO surfaces reveals that, in most cases, the structures identified by our subset screening approach are either identical to the original ones or differ only by rotations or reflections (symmetry operations) and are thus equivalent and degenerate. This demonstrates the robustness of our methodology in identifying the global minimum energy structures of Cu clusters on ZnO surfaces.

Specifically, we observed that the adsorption energy differences between the original and predicted structures all fall within the predefined tolerance ranges, supporting the validity of the selected tolerances for different cluster sizes. The cases where the predicted structures were merely rotated or reflected versions of the originals include O-p-O-7, Zn-r-O-8, Zn-r-O-16, Zn-p-Zn-4, Zn-p-Zn-5, and Zn-p-Zn-8.

For cases where the predicted structures differ from the originals but still fall within the acceptable adsorption energy tolerance, such as the following:

(i) O-r-Zn-8: The predicted structure differs from the original but has an adsorption energy difference of only 0.012 eV, well within the 0.054 eV tolerance, indicating it is still a valid low-energy configuration.

(ii) O-r-Zn-24: Although the predicted configuration does not visually match the original, the adsorption energy difference is 0.269 eV, within the 0.287 eV tolerance, demonstrating that the method captures relevant structural variants.

Overall, these findings confirm that our method effectively retains essential structural characteristics, even when minor variations in orientation or configuration occur in some cases. The alignment of adsorption energies within the defined tolerances reinforces the conclusion that our subset screening approach is capable of accurately identifying the most stable configurations while significantly reducing computational resources. This evaluation highlights the robustness of our methodology, validating its application in the accelerated prediction of stable structures in nanocluster systems.

B. Comparative evaluation with random baseline

1. Success Index and global minima capture

The Success Index is defined here as a measure of how effectively our method recovers known low-energy structures,

including the global minimum. It is conceptually similar to success probabilities or hit rates commonly employed in studies evaluating the performance of global optimization algorithms and structure prediction methodologies [25,36]. While the exact term ‘‘Success Index’’ may not be widely used in the literature, analogous metrics are often reported in terms of how frequently a given approach identifies known global minima or near-global minima in repeated runs or in subsets of selected structures. These benchmarks help assess the method’s reliability in capturing the most physically significant configurations, which is central to understanding the potential energy landscape of a system.

In our implementation, the Success Index is presented as a fraction, for example, ‘‘6/20’’ indicating that out of the top 20 low-energy structures (identified via exhaustive optimization), six were included in our retained subset after screening. This fraction offers a direct, transparent way to convey performance. Even if only one of the top 20 structures is captured (i.e., ‘‘1/20’’), we consider the prediction a success. By setting the reference set of low-energy structures and quantifying how many are successfully recovered, we impose a stringent test of our method’s capability. The fraction format underscores that we do not merely calculate a ratio as a numerical value but instead highlight how many target structures are found out of how many were sought.

The chosen thresholds for the Success Index (e.g., 20 for Cu₄) represent a stringent criterion. Typically, a cluster of size 4 might have around 100 low-energy structures that could be considered global minima due to their close energy and structural similarity. However, by setting a more stringent Success Index, we ensure that our method is robust and reliable in identifying the most stable configurations, even under challenging conditions.

We further validate our method by comparing these fractions against a random baseline of equal sample size. As shown in Table VI and Table VII, for example, random selection typically results in only ‘‘3/20’’ while our method obtains ‘‘6/20’’ for the O-p-O-4 case, which demonstrates a significant improvement and justifies the computational effort and complexity of our approach. Through multiple cases involving different Cu cluster sizes and ZnO surface types, we consistently find that our methodology outperforms random baselines in capturing global minima or near-global minima. This consistent advantage reinforces the notion that integrating structural descriptors, similarity measures, and machine learning-assisted selection leads to a more informed exploration of the configuration space.

By comparison, we find that our approach outperforms random baselines in general, both in terms of validity and robustness, and especially in the performance of obtaining global minima. The Success Index results indicate that our method is generally effective across various cluster sizes and surface types. Although the Success Index varies depending on the specific case, the retained sample sizes are consistently around 3000, demonstrating that our method efficiently reduces the configuration space while still capturing a significant portion of the low-energy structures. This balance between efficiency and accuracy underscores the potential of our approach for large-scale structural screening in materials science.

TABLE II. Comparison of the original and predicted graphics for the lowest energy structures obtained for adsorbed Cu clusters of various sizes on O-poor O-terminated surface. Blue spheres represent Cu, and red and gray spheres represent O and Zn, respectively. The top ZnO layer (i.e., involved in the reconstruction) atoms are highlighted with darker spheres, and the subsurface ZnO atoms are represented by faded spheres. Thin green lines indicate cell boundaries.

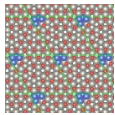
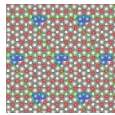
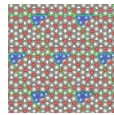
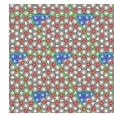
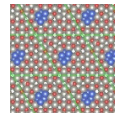
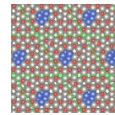
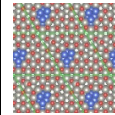
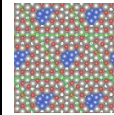
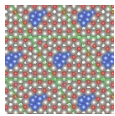
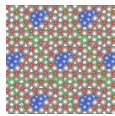
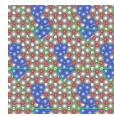
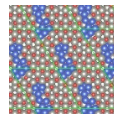
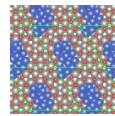
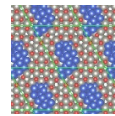
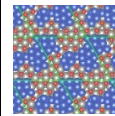
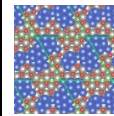
No. Cu							
4		5		6		7	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
-2.930	-2.930	-3.902	-3.902	-4.808	-4.808	-5.450	-5.333
No. Cu							
8		16		24		36	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
-6.133	-6.133	-9.193	-9.193	-9.852	-9.852	-10.655	-10.655

TABLE III. Comparison of the original and predicted graphics for the lowest energy structures obtained for adsorbed Cu clusters of various sizes on Zn-rich O-terminated surface. Blue spheres represent Cu, and red and gray spheres represent O and Zn, respectively. The top ZnO layer (i.e., involved in the reconstruction) atoms are highlighted with darker spheres, and the subsurface ZnO atoms are represented by faded spheres. Thin green lines indicate cell boundaries.

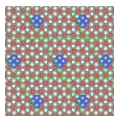
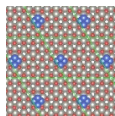
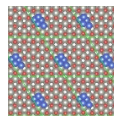
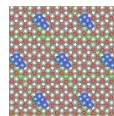
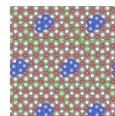
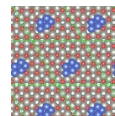
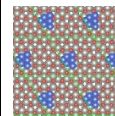
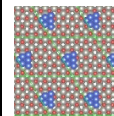
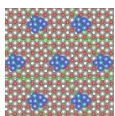
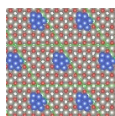
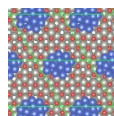
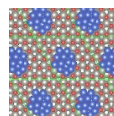
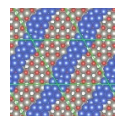
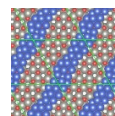
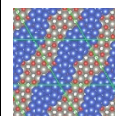
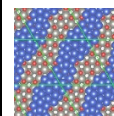
No. Cu							
4		5		6		7	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
1.352	1.352	1.515	1.515	1.709	1.709	1.630	1.630
No. Cu							
8		16		24		36	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
1.414	1.414	1.113	1.397	-0.981	-0.981	-0.547	-0.547

TABLE IV. Comparison of the original and predicted graphics for the lowest energy structures obtained for adsorbed Cu clusters of various sizes on Zn-poor Zn-terminated surface. Blue spheres represent Cu, and red and gray spheres represent O and Zn, respectively. The top ZnO layer (i.e., involved in the reconstruction) atoms are highlighted with darker spheres, and the subsurface ZnO atoms are represented by faded spheres. Thin green lines indicate cell boundaries.

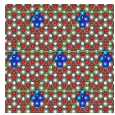
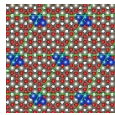
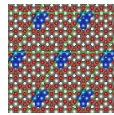
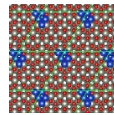
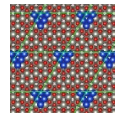
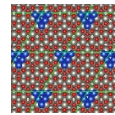
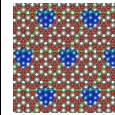
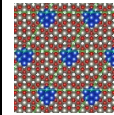
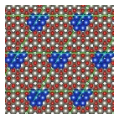
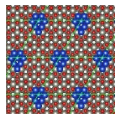
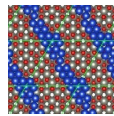
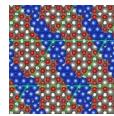
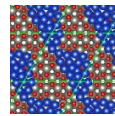
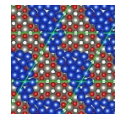
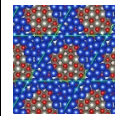
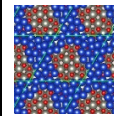
No. Cu							
4		5		6		7	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
0.209	0.209	-0.111	-0.111	-0.366	-0.366	-0.465	-0.465
No. Cu							
8		16		24		36	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
-0.354	-0.352	-0.366	-0.366	-1.070	-1.070	-2.943	-2.943

TABLE V. Comparison of the original and predicted graphics for the lowest energy structures obtained for adsorbed Cu clusters of various sizes on O-rich Zn-terminated surface. Blue spheres represent Cu, and red and gray spheres represent O and Zn, respectively. The top ZnO layer (i.e., involved in the reconstruction) atoms are highlighted with darker spheres, and the subsurface ZnO atoms are represented by faded spheres. Thin green lines indicate cell boundaries.

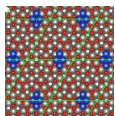
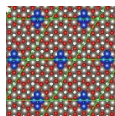
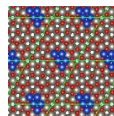
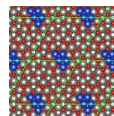
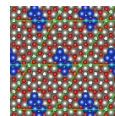
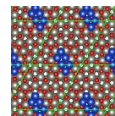
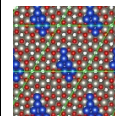
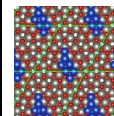
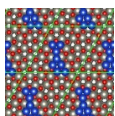
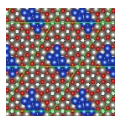
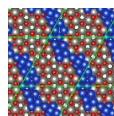
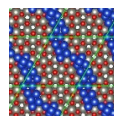
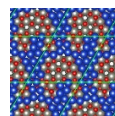
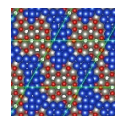
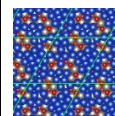
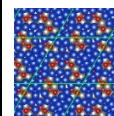
No. Cu							
4		5		6		7	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
-1.459	-1.459	-1.926	-1.926	-2.307	-2.307	-2.460	-2.460
No. Cu							
8		16		24		36	
Original	Predicted	Original	Predicted	Original	Predicted	Original	Predicted
							
Adsorption energy w.r.t. bulk Cu/eV							
-2.728	-2.716	-5.828	-5.828	-8.402	-8.133	-12.307	-12.307

TABLE VI. Evaluation results where O-p-O represents O-poor O-terminated surfaces, Zn-r-O represents Zn-rich O-terminated surfaces, Zn-p-Zn represents Zn-poor Zn-terminated surfaces, and O-r-Zn represents O-rich Zn-terminated surfaces.

Case	O-p-O-4	O-p-O-5	O-p-O-6	O-p-O-7	O-p-O-8	O-p-O-16	O-p-O-24	O-p-O-36
Success Index	6/20	6/18	5/15	6/12	4/10	3/8	3/7	3/5
Sample size	2866	3021	3053	3276	3234	3554	3481	3533
Global minima	√	√	√	√	√	√	√	√
Case	Zn-r-O-4	Zn-r-O-5	Zn-r-O-6	Zn-r-O-7	Zn-r-O-8	Zn-r-O-16	Zn-r-O-24	Zn-r-O-36
Success Index	6/20	7/18	4/15	3/12	4/10	3/8	4/7	1/5
Sample size	3108	3086	3213	3614	3464	3597	3708	3753
Global minima	√	√	√	√	√	√	√	√
Case	Zn-p-Zn-4	Zn-p-Zn-5	Zn-p-Zn-6	Zn-p-Zn-7	Zn-p-Zn-8	Zn-p-Zn-16	Zn-p-Zn-24	Zn-p-Zn-36
Success Index	4/20	7/18	2/15	3/12	4/10	4/8	5/7	3/5
Sample size	2759	2864	3179	3388	3481	3657	3592	3665
Global minima	√	√	√	√	√	√	√	√
Case	O-r-Zn-4	O-r-Zn-5	O-r-Zn-6	O-r-Zn-7	O-r-Zn-8	O-r-Zn-16	O-r-Zn-24	O-r-Zn-36
Success Index	3/20	5/18	5/15	1/12	2/10	1/8	2/7	3/5
Sample size	2678	2988	3114	3205	3340	3556	3662	3676
Global minima	√	√	√	√	×	√	×	√

2. Local minima capture

In addition to evaluating our method’s ability to identify global minima and related low-energy structures, it is also critical to assess how effectively it captures the broader landscape of local minima. Local minima represent configurations that are stable at the local level, often differing subtly in structure and slightly in energy from the global minimum. While these minima may not be the most energetically favorable overall, their presence provides valuable insights into the complexity and richness of the potential energy surface. Moreover, different local minimum structures can exhibit different local structural environments that may be of practical importance (for example, as active catalyst sites). In a real sample, it would be expected that structures resemble not just the predicted global minimum, but also potentially those corresponding to close local minima, especially if there are large kinetic barriers for transitions from one local minimum

to another. Hence, by understanding how well our method recovers these local minima compared to a purely random selection of configurations, we can gauge its ability to map out a more complete and nuanced picture of the structural landscape.

The rationale behind this evaluation is twofold. First, the presence of numerous local minima can make it challenging to rely solely on global minima or a few low-energy structures for a comprehensive understanding of a system’s stability and morphological variety. By ensuring that our method effectively captures a significant portion of the local minima, we increase the likelihood that we are not missing important structural variants that could influence properties like catalytic activity, surface reactivity, or thermodynamic stability. Second, comparing our results against a random baseline helps illustrate the added value and guidance that our methodology provides. If our approach consistently outperforms random sampling in capturing local minima, it affirms that the struc-

TABLE VII. Random baseline where O-p-O represents O-poor O-terminated surfaces, Zn-r-O represents Zn-rich O-terminated surfaces, Zn-p-Zn represents Zn-poor Zn-terminated surfaces, and O-r-Zn represents O-rich Zn-terminated surfaces.

Case	O-p-O-4	O-p-O-5	O-p-O-6	O-p-O-7	O-p-O-8	O-p-O-16	O-p-O-24	O-p-O-36
Success Index	3/20	5/18	1/15	1/12	3/10	1/8	1/7	3/5
Sample size	2866	3021	3053	3276	3234	3554	3481	3533
Global minima	×	×	×	×	×	×	√	√
Case	Zn-r-O-4	Zn-r-O-5	Zn-r-O-6	Zn-r-O-7	Zn-r-O-8	Zn-r-O-16	Zn-r-O-24	Zn-r-O-36
Success Index	8/20	6/18	4/15	3/12	3/10	3/8	1/7	0/5
Sample size	3108	3086	3213	3614	3464	3597	3708	3753
Global minima	×	√	×	×	√	×	√	×
Case	Zn-p-Zn-4	Zn-p-Zn-5	Zn-p-Zn-6	Zn-p-Zn-7	Zn-p-Zn-8	Zn-p-Zn-16	Zn-p-Zn-24	Zn-p-Zn-36
Success Index	6/20	5/18	2/15	4/12	2/10	0/8	2/7	1/5
Sample size	2759	2864	3179	3388	3481	3657	3592	3665
Global minima	×	×	√	×	×	×	√	√
Case	O-r-Zn-4	O-r-Zn-5	O-r-Zn-6	O-r-Zn-7	O-r-Zn-8	O-r-Zn-16	O-r-Zn-24	O-r-Zn-36
Success Index	5/20	6/18	7/15	3/12	1/10	0/8	1/7	2/5
Sample size	2678	2988	3114	3205	3340	3556	3662	3676
Global minima	√	×	×	×	√	×	×	√

TABLE VIII. Proportions of total local minima captured by our method and random sampling for different Cu cluster sizes on various ZnO surfaces, using 30% of the data points from 10 000 initial structures, where O-p-O represents O-poor O-terminated surfaces, Zn-r-O represents Zn-rich O-terminated surfaces, Zn-p-Zn represents Zn-poor Zn-terminated surfaces, O-r-Zn represents O-rich Zn-terminated surfaces, Method represents our methodology, and Random represents random sampling.

Surface type	Comparison	No. Cu							
		4	5	6	7	8	16	24	36
O-p-O	Method	40.98%	41.88%	40.17%	41.63%	40.23%	40.79%	39.08%	38.98%
	Random	22.77%	43.16%	40.10%	38.96%	37.86%	33.20%	33.87%	33.29%
Zn-r-O	Method	40.86%	43.07%	43.51%	46.40%	44.31%	41.22%	42.01%	41.76%
	Random	42.19%	43.04%	42.13%	40.98%	39.35%	35.25%	34.57%	33.75%
Zn-p-Zn	Method	39.83%	41.84%	42.29%	42.89%	43.07%	41.08%	39.97%	40.29%
	Random	44.74%	44.38%	41.18%	39.38%	38.43%	34.24%	33.72%	33.30%
O-r-Zn	Method	38.52%	41.67%	41.00%	41.25%	41.49%	41.26%	40.90%	40.74%
	Random	45.15%	42.82%	40.90%	39.67%	38.47%	35.60%	34.11%	33.46%

tural descriptors, similarity measures, and selection strategies we employ genuinely enhance exploration of the energy landscape, rather than simply adding computational overhead.

The data shown in Table VIII, aggregated over various surfaces and cluster sizes, show clear trends. For small clusters, differences between our method and random sampling may appear modest, with occasional instances where random selection performs comparably or even slightly better. However, as cluster size increases, our method's advantage becomes more pronounced. Across multiple surfaces, our method consistently outperforms random sampling in capturing a larger proportion of the local minima. This pattern suggests that the structural features and informed selection strategies at the core of our methodology scale effectively with complexity, enabling it to navigate the energy landscape more intelligently than an unguided approach.

In essence, while random sampling may occasionally stumble onto local minima in simpler scenarios, our method's more systematic, data-driven selection process ensures that as the system grows in size and complexity, it continues to identify a broader and richer array of local minima. This superiority in capturing local minima reinforces the notion that our approach is not just about finding the very best structures but also about preserving the intrinsic diversity of the energy landscape. Ultimately, this leads to a more comprehensive and meaningful understanding of the supported Cu/ZnO nanocluster systems under study.

C. Sensitivity analysis

To further validate the robustness of our methodology, we conducted a sensitivity analysis by adjusting the boundary points' SOS threshold ranges. The Success Index, global minima, and sample size were evaluated as we varied the lower and upper limits of the SOS range to examine how different selections of boundary points affect the performance of our method. This analysis helps demonstrate the impact of the SOS thresholds on our results, guiding us to select the most efficient balance between accuracy and computational cost.

a. Success Index sensitivity analysis. The Success Index is a broader standard that measures the proportion of low-energy structures captured by our method. By adjusting the SOS threshold (both lower and upper bounds), we evaluate how

the Success Index changes as we alter the range of retained boundary points.

As shown in Table IX, the results show that by adjusting the SOS threshold ranges, the Success Index remains high across most selections. This demonstrates the robustness of our method and suggests that our SOS threshold selection provides flexibility without sacrificing accuracy.

b. Global minima sensitivity analysis. The global minima metric is more stringent than the Success Index, as it focuses on identifying the actual lowest-energy structures from the full set. We analyse how changes in the SOS threshold affect the ability of our method to capture these critical configurations.

As shown in Table X, for the global minima, a lower threshold around 55% ensures that we capture 100% of the true global minima, and increasing the upper bound to 75% further strengthens this capture. Selecting 55% as the lower bound and 75% as the upper bound provides the optimal balance between accurately capturing the global minima and limiting the computational cost. This is why we chose these particular values for our methodology, as they strike the best trade-off between efficiency and performance.

c. Sample size sensitivity analysis. We also examine the impact of varying the SOS thresholds on the average sample size selected from the initial set. This is crucial for understanding the computational savings our method provides.

As shown in Table XI, the sample size grows significantly as the lower bound of the SOS range decreases, with a lower

TABLE IX. Success index sensitivity analysis: The vertical values of 45%–65% represent the lower limit (lower boundary point) of the threshold. The horizontal values of 65%–85% represent the upper limit (upper boundary point) of the threshold. The success index inside represents the success index when the thresholds take specific values.

Success Index	65%	70%	75%	80%	85%
65%	0%	75%	100%	100%	100%
60%	87.5%	100%	100%	100%	100%
55%	100%	100%	100%	100%	100%
50%	100%	100%	100%	100%	100%
45%	100%	100%	100%	100%	100%

TABLE X. Global minima sensitivity analysis: The vertical values of 45%–65% represent the lower limit (lower boundary point) of the threshold. The horizontal values of 65%–85% represent the upper limit (upper boundary point) of the threshold. The percentage inside represents the probability that the subset contains the global minima when the thresholds take specific values.

Global minima	65%	70%	75%	80%	85%
65%	0%	25%	37.5%	37.5%	37.5%
60%	37.5%	62.5%	75%	75%	75%
55%	62.5%	87.5%	100%	100%	100%
50%	62.5%	87.5%	100%	100%	100%
45%	62.5%	87.5%	100%	100%	100%

threshold of 55% yielding approximately 3252 structures on average when combined with an upper bound of 75%. This sample size aligns well with the desired goal of reducing the computational load while maintaining accuracy in capturing global minima and structural diversity.

By performing this sensitivity analysis, we demonstrate that the method is robust and flexible across different parameter ranges. The selection of 55% as the lower bound and 75% as the upper bound strikes the optimal balance, ensuring the identification of critical structures while keeping the sample size manageable and the computational load reasonable.

D. Divergence evaluation

In addition to evaluating how well our method captures low-energy structures, it is equally important to determine whether the selected subset of configurations adequately represents the overall energy landscape. To address this, we employ Kullback-Leibler (KL) divergence and Jensen-Shannon (JS) divergence as distributional similarity metrics. Both KL and JS divergences measure how closely the probability distribution of the selected subset’s energy levels matches that of the full configuration set. Lower values of these metrics indicate a closer match, implying that our method preserves not just the stability characteristics (i.e., identifying global and near-global minima) but also the broader structural diversity of the system.

While there are no universally fixed “acceptable” divergence thresholds specific to nanocluster energy landscapes,

TABLE XI. Sample size sensitivity analysis: The vertical values of 45%–65% represent the lower limit (lower boundary point) of the threshold. The horizontal values of 65%–85% represent the upper limit (upper boundary point) of the threshold. The number inside represents the average sample size when the thresholds take specific values.

Average sample size	65%	70%	75%	80%	85%
65%	0.00	758.13	1317.38	1692.13	1916.75
60%	924.75	1682.88	2242.13	2616.88	2841.50
55%	1934.88	2693.00	3252.25	3627.00	3851.63
50%	2956.50	3714.63	4273.88	4648.63	4873.25
45%	3840.38	4598.50	5157.75	5532.50	5757.13

comparative benchmarks can be gleaned from related fields where distributional similarity measures have been employed. In information theory, the JS divergence ranges between 0 (identical distributions) and 1 (completely disjoint distributions), and values below 0.4 often indicate reasonably similar distributions [55]. Within computational chemistry and materials science, studies have seldom defined strict cutoffs, but a JS divergence near or below this range is typically viewed as reflecting a representative sampling of the underlying distribution [36,56]. Similarly, KL divergence, which can range from 0 to infinity, does not have a universally accepted threshold; however, smaller values (e.g., below five) suggest that the selected distribution does not drastically deviate from the original, especially given the complexity and high dimensionality of energy landscapes in nanocluster systems [57].

As shown in Table XII, KL divergence values range from approximately 1.0 to 5.6, and JS divergence values generally fall between 0.2 and 0.5. Although the upper end of the JS divergence extends slightly beyond the 0.4, most cases remain at or near levels consistent with relatively close resemblance to the full energy landscape distribution. Given the inherent complexity and variability of supported nanocluster systems, these values are indicative of a reasonably faithful representation. In other words, while we aggressively reduce the original 10 000 configurations to about 3000, we still maintain a distribution of energy states that does not heavily skew away from the full set’s profile. The relatively low JS divergence across many scenarios implies that the subset includes a broad spectrum of low- and moderate-energy configurations, rather than focusing solely on the absolute lowest energy states.

These observations suggest that our method strikes a practical balance: it efficiently narrows down the configuration space while retaining a structurally and energetically representative subset. By not diverging significantly from the original distribution, the methodology ensures that important structural variations and potential intermediate states are not entirely lost. This, in turn, enhances the value of the screened subset for further analyses, such as electronic structure calculations, dynamical simulations, or catalytic reactivity assessments.

Overall, the low divergence values support the conclusion that our methodology is robust, capturing not only the key stable configurations but also preserving a representative cross section of the energy landscape. Thus, the approach maintains its utility for a wide array of applications requiring a comprehensive understanding of complex nanocluster systems.

IV. CONCLUSIONS

We have developed and validated a machine learning-based methodology for accelerating the prediction of stable structures in nanoclusters, with a specific application to Cu clusters on various polar ZnO surfaces. By integrating SOAP descriptors, feature compression, REMatch kernel similarity measurement, and Stochastic Outlier Selection (SOS), we effectively reduced the computational burden associated with global optimization techniques while maintaining a high level of accuracy in identifying low-energy structures.

Our approach demonstrated significant efficiency gains by focusing on approximately 30% of the initial configurations,

TABLE XII. KL&JS divergence results where O-p-O represents O-poor O-terminated surfaces, Zn-r-O represents Zn-rich O-terminated surfaces, Zn-p-Zn represents Zn-poor Zn-terminated surfaces, and O-r-Zn represents O-rich Zn-terminated surfaces.

Case	O-p-O-4	O-p-O-5	O-p-O-6	O-p-O-7	O-p-O-8	O-p-O-16	O-p-O-24	O-p-O-36
KL divergence	4.3711	3.5466	2.4759	1.3512	2.0751	1.6046	1.7458	1.5762
JS divergence	0.4399	0.4163	0.3365	0.2264	0.3202	0.3954	0.2955	0.2826
Case	Zn-r-O-4	Zn-r-O-5	Zn-r-O-6	Zn-r-O-7	Zn-r-O-8	Zn-r-O-16	Zn-r-O-24	Zn-r-O-36
KL divergence	1.1165	5.6184	2.3643	2.2178	1.9296	1.1920	1.7201	3.2653
JS divergence	0.2144	0.5127	0.3731	0.2958	0.2893	0.3161	0.2993	0.4755
Case	Zn-p-Zn-4	Zn-p-Zn-5	Zn-p-Zn-6	Zn-p-Zn-7	Zn-p-Zn-8	Zn-p-Zn-16	Zn-p-Zn-24	Zn-p-Zn-36
KL divergence	1.4935	2.4711	2.6437	3.0683	2.0300	1.9037	1.6362	2.2937
JS divergence	0.2251	0.3506	0.3662	0.4204	0.3292	0.3042	0.3343	0.4017
Case	O-r-Zn-4	O-r-Zn-5	O-r-Zn-6	O-r-Zn-7	O-r-Zn-8	O-r-Zn-16	O-r-Zn-24	O-r-Zn-36
KL divergence	4.2210	2.2109	4.4510	1.8765	2.8953	1.7315	2.1357	1.0093
JS divergence	0.4626	0.3026	0.4451	0.2784	0.3798	0.3553	0.3358	0.2147

successfully capturing the essential low-energy structures that are of primary interest. The success of our methodology was confirmed through a rigorous evaluation process, where the selected subsets consistently included the global minimum energy structures across different cluster sizes. This validation underscores the potential of our method to streamline computationally intensive structural screening processes in materials science. Furthermore, the use of KL and JS divergences provided additional evidence that the distribution of the retained structures closely matches that of the original dataset, reinforcing the robustness of our approach.

Looking ahead, there is substantial potential to further enhance this methodology. One promising direction is the incorporation of supervised learning techniques to refine the selection process. By training models on known datasets, it may be possible to predict the likelihood of specific configurations being low-energy structures with even greater accuracy. This could further reduce the computational demands of structural optimization, making the process more efficient and scalable.

In summary, our work represents a significant step forward in the application of machine learning to structural prediction in nanoclusters. The framework we have developed not only improves the efficiency of identifying stable configurations but also opens avenues for the integration of advanced machine learning techniques in the future.

ACKNOWLEDGMENTS

The authors acknowledge the use of ARCHER2 UK National Supercomputing Service [58] via membership of UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (Grant No. EP/L000202/1); this work used the UK Materials and Molecular Modelling Hub for computational resources, MMM Hub, which is partially funded by EPSRC (Grants No. EP/T022213/1, No. EP/W032260/1, and No. EP/P020194/1). We also acknowledge the use of the UCL Kathleen and Myriad High Performance Computing Facility (Kathleen@UCL, Myriad@UCL), and associated support services, in the completion of this work. Y.Z. acknowledges a UCL Research Excellence Scholarship for a Ph.D. studentship. K.T.B. acknowledges financial support from the United Kingdom Research and Innovation (UKRI) Engineering and Physical Sciences Research Council (EPSRC), under projects Grant No. EP/Y000552/1 and No. EP/Y014405/1. M.D.H. acknowledges the EPSRC/UKRI (EP/T028629/1) for financial support and the UK Catalysis Hub Consortium (funded by EPSRC, Grant No. EP/R026815/1) for the provision of additional resources.

DATA AVAILABILITY

A repository containing the codes used in this work has been made available on GitHub [58].

- [1] M. Farrow, Y. Chow, and S. Woodley, *Phys. Chem. Chem. Phys.* **16**, 21119 (2014).
- [2] S. M. Woodley and R. Catlow, *Nat. Mater.* **7**, 937 (2008).
- [3] D. Astruc, *Chemical Reviews* **120**, 461 (2020).
- [4] L. Liu and A. Corma, *Chem. Rev.* **118**, 4981 (2018).
- [5] M. J. Ndolomingo, N. Bingwa, and R. Meijboom, *J. Mater. Sci.* **55**, 6195 (2020).
- [6] Z. Luo, G. Zhao, H. Pan, and W. Sun, *Adv. Energy Mater.* **12**, 2201395 (2022).
- [7] M. Xu, M. Peng, H. Tang, W. Zhou, B. Qiao, and D. Ma, *J. Am. Chem. Soc.* **146**, 2290 (2024).
- [8] D. J. Wales and J. P. Doye, *J. Phys. Chem. A* **101**, 5111 (1997).
- [9] D. J. Wales and H. A. Scheraga, *Science* **285**, 1368 (1999).
- [10] M. A. Zwijnenburg and S. T. Bromley, *Phys. Rev. B* **83**, 024104 (2011).
- [11] C. J. Pickard and R. J. Needs, *J. Phys.: Condens. Matter* **23**, 053201 (2011).
- [12] C. J. Pickard and R. J. Needs, *J. Chem. Phys.* **127**, 244503 (2007).
- [13] J. M. McMahon, *Phys. Rev. B* **84**, 220104(R) (2011).
- [14] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, *Science* **220**, 671 (1983).
- [15] J. Schön and M. Jansen, *Z. Kristall.-Crystal. Mater.* **216**, 307 (2001).

- [16] J. Schön and M. Jansen, *Z. Kristall.-Crystal. Mater.* **216**, 361 (2001).
- [17] D. Zagorac, K. Doll, J. C. Schön, and M. Jansen, *Phys. Rev. B* **84**, 045206 (2011).
- [18] A. R. Oganov and C. W. Glass, *J. Chem. Phys.* **124**, 244704 (2006).
- [19] D. M. Deaven and K.-M. Ho, *Phys. Rev. Lett.* **75**, 288 (1995).
- [20] Y. Zeiri, *Phys. Rev. E* **51**, R2769(R) (1995).
- [21] R. C. Eberhart and Y. Shi, *IEEE Trans. Evol. Comput.* **8**, 201 (2004).
- [22] R. Poli, D. Bratton, T. Blackwell, and J. Kennedy, in *2007 IEEE Congress on Evolutionary Computation* (IEEE, Singapore, 2007), pp. 1955–1962.
- [23] Y. Wang, J. Lv, L. Zhu, and Y. Ma, *Phys. Rev. B* **82**, 094116 (2010).
- [24] S. Heiles and R. L. Johnston, *Int. J. Quantum Chem.* **113**, 2091 (2013).
- [25] R. L. Johnston, *Dalton Trans.* 4193 (2003).
- [26] C. R. A. Catlow, S. T. Bromley, S. Hamad, M. Mora-Fonz, A. A. Sokol, and S. M. Woodley, *Phys. Chem. Chem. Phys.* **12**, 786 (2010).
- [27] T. Lazauskas, A. A. Sokol, J. Buckeridge, C. R. A. Catlow, S. G. Escher, M. R. Farrow, D. Mora-Fonz, V. W. Blum, T. M. Phaahla, and H. R. Chauke, *Phys. Chem. Chem. Phys.* **20**, 13962 (2018).
- [28] C. A. McCandler, A. Pihlajamäki, S. Malola, H. Häkkinen, and K. A. Persson, *ACS Nano* **18**, 19014 (2024).
- [29] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, *Nat. Commun.* **12**, 398 (2021).
- [30] J. Behler and M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007).
- [31] M. D. Higham, D. Mora-Fonz, A. A. Sokol, S. M. Woodley, and C. R. A. Catlow, *J. Mater. Chem. A* **8**, 22840 (2020).
- [32] S. M. Woodley, *J. Phys. Chem. C* **117**, 24003 (2013).
- [33] A. Walsh and S. M. Woodley, *Phys. Chem. Chem. Phys.* **12**, 8446 (2010).
- [34] W. Jee, A. A. Sokol, C. Xu, B. Camino, X. Zhang, and S. M. Woodley, *Chem. Mater.* **36**, 8737 (2024).
- [35] D. Mora-Fonz, T. Lazauskas, M. R. Farrow, C. R. A. Catlow, S. M. Woodley, and A. A. Sokol, *Chem. Mater.* **29**, 5306 (2017).
- [36] R. Podgornik, *J. Stat. Phys.* **126**, 423 (2007).
- [37] A. P. Bartók, R. Kondor, and G. Csányi, *Phys. Rev. B* **87**, 184115 (2013).
- [38] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, *Phys. Chem. Chem. Phys.* **18**, 13754 (2016).
- [39] F. H. Stillinger and T. A. Weber, *Science* **225**, 983 (1984).
- [40] B. Hartke, *J. Comput. Chem.* **20**, 1752 (1999).
- [41] B. Hartke, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **1**, 879 (2011).
- [42] F. Baletto and R. Ferrando, *Rev. Mod. Phys.* **77**, 371 (2005).
- [43] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010).
- [44] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, *Phys. Rev. Lett.* **108**, 058301 (2012).
- [45] J. P. Darby, J. R. Kermode, and G. Csányi, *npj Comput. Mater.* **8**, 166 (2022).
- [46] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevMaterials.9.033801> for the pseudocode and computational complexity analysis.
- [47] J. Janssens, E. O. Postma, and J. van den Herik, in *MAD 2011 Workshop Proceedings* (Tilburg, 2011), p. 21.
- [48] D. Mora-Fonz, T. Lazauskas, S. M. Woodley, S. T. Bromley, C. R. A. Catlow, and A. A. Sokol, *J. Phys. Chem. C* **121**, 16831 (2017).
- [49] F. Cleri and V. Rosato, *Phys. Rev. B* **48**, 22 (1993).
- [50] L. Whitmore, A. A. Sokol, and C. R. A. Catlow, *Surf. Sci.* **498**, 135 (2002).
- [51] S. Kullback and R. A. Leibler, *Ann. Math. Stat.* **22**, 79 (1951).
- [52] M. Thomas and A. T. Joy, *Elements of information theory* (Wiley-Interscience, New York, 2006).
- [53] M. Thomas and A. T. Joy, *Elements of Information Theory* (Wiley-Interscience, New York, 2006).
- [54] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to Information Retrieval* (Cambridge University Press, Cambridge, 2008), Vol. 39.
- [55] D. M. Endres and J. E. Schindelin, *IEEE Trans. Inf. Theory* **49**, 1858 (2003).
- [56] J. Doye and D. Wales, *Z. Phys. D* **40**, 194 (1997).
- [57] F. H. Stillinger, *Science* **267**, 1935 (1995).
- [58] <https://github.com/yunyuzh/REMatch-SOS>.