

Active Phase Discovery in Heterogeneous Catalysis via Topology-Guided Sampling and Machine Learning

Corresponding Author: Professor Feng Pan

This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. The manuscript was considered suitable for publication without further review at Nature Communications.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper proposes a new algorithm for efficiently exploring structural candidates, which is crucial in building machine learning potentials for simulating molecular/atom adsorption and diffusion within catalysis. By applying persistent homology, the authors have developed a scheme that automatically identifies potential reaction sites and generates structural candidates. Using the dataset obtained through this method, the authors have shown that it is possible to generate machine learning potentials suitable for the hydrogen reactions on Pd surfaces and the oxidation reactions of Pt. However, several concerns exist, mainly regarding the significance of using persistent homology in this proposed method, the clarity of its description, the lack of specific comparison with other methods, and the biased citation of previous studies. Therefore, the reviewer cannot recommend publication of this manuscript in Nature Communications. Detailed comments are provided below:

- What is the definition of "active phase"? This should be clearly defined for the broad readers of Nature Communications.
- How can the authors discuss the general applicability of their method to various catalytic reactions by addressing the specific reactions targeted in this study?
- The citation of studies on the application of persistent homology, particularly in the field of materials science, seems biased towards certain research groups. Besides the references mentioned by the authors, there are several other studies that analyze structures and structure-property relationship using persistent homology. In recent years, the idea of applying persistent homology to materials science has gained momentum among multiple researchers. The authors should cite relevant previous works appropriately. Examples include:
<https://www.pnas.org/doi/full/10.1073/pnas.1520877113>
<https://www.science.org/doi/10.1126/sciadv.abc2320>
<https://doi.org/10.1063/5.0093441>
<https://www.nature.com/articles/ncomms15396>
<https://www.nature.com/articles/s41598-021-88027-8>
- The connection between atom aggregation and the persistent homology calculation is unclear. It seems that persistent homology is applied to cluster structures formed by adding up to ten atoms near a specific atom or by introducing a cutoff radius to include surrounding atoms. Furthermore, the description in the Methods section, "Since we start iterating from two atoms, the sites determined by the low-dimensional Betti numbers in the multi-atom aggregate will be included in the Betti number evolution of fewer atoms." suggests that the Betti numbers are calculated for structures containing two, three, four, etc., atoms. The authors should clarify how they merge the Betti number information obtained from each structure, or if they only use the results of the largest cluster.
- The atom aggregation process itself is not clear. How did the authors pick up the octahedron as shown in Figure 2 by picking up atoms near some specific atom one by one?

- In the Methods section, the statement " Through the filtering process, we determine the circumcenters of the geometric shapes formed by the initial positions of the atoms based on the distances obtained from the death point, which potentially represent the active sites." is difficult to understand. At what stage of atom aggregation is the circumcenter defined? It would be helpful to include specific equations or figures.
- In the proposed method, persistent homology is applied to decompose the arrangement of atoms around a specific atom into polyhedral structures for interfacial, bulk, and cluster configurations. Persistent homology is just used to evaluate circumcenter of polyhedral structure. If this is the case, wouldn't it be computationally more efficient to directly compute the circumcenter of the polygons within the cluster structure formed during the atom aggregation process, without using persistent homology, to achieve the same goal?
- Can the authors provide specific comparisons to demonstrate the superiority of the proposed method over other search algorithms?
- Which codes were used to compute persistent homology? GUDHI or ripser or other codes? If the authors used some open-source code, it should be cited.
- What kind of filtration (Vietoris-Rips or alpha) was used to obtain the barcode?

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

Summary:

This work proposes the application of persistence homology to sample active phases of heterogeneous catalysts, and further proposes the use of persistence homology as a data extraction method for machine learning.

Comments:

- The paper justifies the application of persistence homology with a well written and persuasive exposition. The discussion of Figure 2 particularly shows the strength of this approach as a simple feature extractor for machine learning applications.
- The paper fully discusses two applications of their work, showing the versatility of their approach. I generally appreciate the depth of the exposition the authors provide.
- There are several past proposed methodologies for the sampling the paper proposes (e.g. citations 27-32 of the paper). However, no comparisons are made between these methods and the paper's proposed new method. Serious evaluation of this paper requires a comparison with current state of the art techniques to demonstrate improvements in either clock time or evaluation quality. I view this as the most serious weakness of the paper: in its current form, the comparisons to past works are purely qualitative, without data-driven support.
- The authors may be interested in Reinauer et al.'s paper "Persformer: A Transformer Architecture for Topological Machine Learning", which demonstrates a useful method for machine learning with persistence homology inputs. If they are interested in further work with persistency homology, this could provide a useful beginning to their further experiments.
- The authors need to do a more thorough lit review as there are various papers related to PH and chemical applications. For example e.g. (Queen et al. 2024- npj Comp Mat.), (Jones et al., 2023, ACS) and references therein.

Minor comments:

- The example barcode in Figure 1 is not plausible: H_1 and H_2 generators shouldn't be "born" at $\lambda=0$, but at a variety of greater radii. This could present confusion for readers unfamiliar with persistence homology.
- L147: "catch" is presumably supposed to read "capture"
- This is particularly minor, but I think the phrasing of the sentence with begins on L151, while perfectly accurate, could cause confusion. The first half of the sentence puts a parallel between "components, loops, and cavities" with " β_0 , β_1 , and β_2 ", and the second half notes that barcodes record "birth, death, and persistence". It's easy to slip up and read this as " β_0 , β_1 , and β_2 " corresponding to "birth, death, and persistence". Separating this into two sentences would avoid the possibility of confusion.

Overall:

I view the general methodology as sound. However, I do believe additional evidence is needed to validate the claims the authors make. Some comparison to the quality of previous work is necessary to validate these claims. The significance of the paper's proposed method is contingent on improvements to current state-of-the-art sampling methods.

(Remarks on code availability)

Comments on the code:

The available code is helpful, and Jupyter notebooks ensure that the process is easily understandable. However, the code provided is incomplete for the replication of experiments demonstrated in the paper. Only the novel sampling procedure is described, none of the post-processing MLFFs or other systems are shown.

Reviewer #4

(Remarks to the Author)

This is a very interesting application of persistent homology in combination with a configurational exploration that is expedited through machine learning force fields. I agree that PH could be used to uniquely identify configurations that are relevant to catalytic activation and that this could be used in conjunction with DFT and MLFF for the purposes of this paper. My largest concern has to do with the clarity of the methodology such that other may easily understand and reproduce the approach.

My questions arise primarily around the persistent homology. There is quite a bit of information in the method section that I don't believe used within the approach (e.g., homology groups, the cycle and boundary groups) and it makes the methodology more complicated than it needs to be. It is a bit unclear what information is exactly employed and how. Additionally, it would seem to me that the PH data would need added information regarding the atom types if there is a mix of Pd, O, H or other adsorbates being considered. This is not described well.

The 10 Ang cutoff for the PH seems a bit arbitrary for the cutoff of the search of aggregate structures. Did the authors do any systematic searching of how large a cutoff was needed and how this impacted their algorithm?

It is unclear to me how the PH based aggregates are embedded within the slabs that are then employed with the force field enumeration?

In the MPLFF - the AIMD simulations for 2 or 10 ps seems very short. What and how are the AIMD simulations utilized in the MLFF? what is analyzed to understand if there is convergence of the energy and fluctuations in the bond distances, angles, etc? There seems to be a disconnect between the size of the PH aggregate clusters and the size of AIMD simulations. I don't quite understand why different sizes of the information embedding by PH and the AIMD is used. Why wouldn't use use barcodes of the full size of the system?

The DPA and associated models are not well described - I really don't understand this well at all.

At a very fundamental level, the methodology is not written in a very clear and logical framework such that it would be easy to reproduce. It would be nice if it were simplified

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The authors have responded to the reviewer's questions in a thorough and detailed manner. Their explanation using Eqs. 1 and 2 has clarified, to a reproducible level, how reaction sites can be identified from the death positions in the persistence barcode. This explanation also makes it clear that persistent homology can reduce the number of parameters to be optimized and improve the efficiency of the search process. The revised manuscript provides a convincing example of this efficiency improvement using Pt clusters.

Additionally, the comparison with existing methods has been improved by providing a quantitative comparison with the basin hopping method, which strengthens the case for the proposed approach. The issues raised in the initial review, including insufficient explanation of the method, lack of quantitative comparison with existing methods, and inadequate references, have been significantly addressed in this revision, resulting in a notable improvement in the quality of the manuscript.

Considering these improvements, the reviewer recommends this paper for publication in Nature Communications.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)

I would like to thank the authors for their attention to our comments and detailed replies. The paper is much stronger than its initial version, and the updates the authors have made to the organization have improved its readability and flow. The comparison with the basin hopping algorithm is particularly helpful, making clear the improvement their methodology offers, including both higher effectiveness than previous global methods and the ability to more effectively utilize modern hardware for efficient computation.

Given how often the material surface is the primary point of interest, I believe the paper would be made stronger by also offering comparisons to the graph based methodology, so that users could easily determine which algorithm is best suited to their needs when subsurface and bulk interactions are not relevant to the system. However, the current version of the work is sufficiently complete for publication.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

the manuscript is significantly improved and it seems that the authors have address all of the prior reviewer concerns. I recommend publication.

(Remarks on code availability)

the manuscript is significantly improved and it seems that the authors have address all of the prior reviewer concerns. I recommend publication.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer #1 (Remarks to the Author):

This paper proposes a new algorithm for efficiently exploring structural candidates, which is crucial in building machine learning potentials for simulating molecular/atom adsorption and diffusion within catalysis. By applying persistent homology, the authors have developed a scheme that automatically identifies potential reaction sites and generates structural candidates. Using the dataset obtained through this method, the authors have shown that it is possible to generate machine learning potentials suitable for the hydrogen reactions on Pd surfaces and the oxidation reactions of Pt. However, several concerns exist, mainly regarding the significance of using persistent homology in this proposed method, the clarity of its description, the lack of specific comparison with other methods, and the biased citation of previous studies. Therefore, the reviewer cannot recommend publication of this manuscript in Nature Communications. Detailed comments are provided below:

Response:

We are deeply grateful to the reviewer for the professional comments, which provide us the opportunity to enhance the quality of our paper. We have thoroughly revised both the manuscript and the supplementary information according to your constructive feedback. We sincerely hope that our modifications meet with your satisfaction.

- What is the definition of "active phase"? This should be clearly defined for the broad readers of Nature Communications.

Response:

We appreciate the reviewer's insightful comments. A clear definition of the "active phase" is indeed crucial. In our study, we define the active phase as the surface state and overall structure of the catalyst formed under specific operating conditions. This definition is summarized based on previous research literatures (Chem. Commun., 2024, 60, 14346–14359, Nature Communications, (2023) 14:3004, Nature Material 20, 674–682 (2021)).

We have made revision in the manuscript (page 3).

The resulting surface states and overall structures of catalysts under specific conditions are considered the active phases in catalytic reactions. They are typically dynamic and may undergo transitions as conditions change, which are intimately tied to the mechanistic insights and material design within this field.

- How can the authors discuss the general applicability of their method to various catalytic reactions by addressing the specific reactions targeted in this study?

Response:

We thank the reviewers. Regarding the general applicability of the method developed in this work, we demonstrate it from two perspectives: its suitability for different representative catalyst morphologies and its applicability to various typical catalytic reaction processes.

Morphological generality

Fundamentally, any material structure can be considered as a collection of points, making the use of PH a straightforward and intuitive approach for analyzing material structures and exploring interaction sites. In heterogeneous catalysis, two types of atomic models are commonly employed: slab models with fixed x-y lattice parameters and cluster models, which are characterized by their amorphous nature (Chem. Rev. 2021, 121, 1007–1048). Based on this, we selected a Pd slab and a Pt cluster to evaluate the effectiveness of the PH method in identifying potential active sites of catalysts with different morphologies. Here, we would like to kindly highlight that for the Pt-O system, the number of interaction sites was dynamically updated based on the varying oxygen incorporation levels. This process involved enumerating sites in a multi-element system (Pt and O), demonstrating the adaptability of our method to multi-component materials. Given its excellent performance in these two systems, we propose that the method can be intuitively and feasibly applied to other slab or cluster systems.

Reaction type generality

Heterogeneous catalysis typically occurs within either a reduction or oxidation reactive environment. During a reduction atmosphere, catalyst interfaces are often

covered with hydrogen, which originates from H_2 in thermal catalysis or from water activation in electrocatalysis and photocatalysis. Conversely, in an oxidation atmosphere, the catalyst surface tends to adsorb large amounts of oxygen from the surrounding environment.

To represent these two fundamental reaction types, we selected hydrogen adsorption and incorporation on Pd as an example for the reduction atmosphere, and the oxidation process on Pt as an example for the oxidation atmosphere. Notably, although the Pt example is set under oxygen reduction reaction (ORR), the reaction actually occurs at positive potentials, meaning the Pt catalyst effectively operates in an oxidizing environment. By constructing Pourbaix diagrams, we described the variations in hydrogen and oxygen concentrations as functions of external environmental changes, which align closely with experimental observations.

Basically, the challenge of active phase exploration lies in efficiently determining interaction sites in material structure. Once the potential active sites are identified by PH-SA, relevant adsorption configurations can be readily generated by placing adsorbates—not limited to hydrogen and oxygen but also other species—at these sites. For instance, in the CO_2 reduction reaction (CO_2RR) to produce multi-carbon products, the co-adsorption of $*CO$ and $*H$ is considered crucial (Nano Lett. 2022, 22, 3801–3808). In such cases, after determining the active sites using the PH-SA, $*H$ and $*CO$ can be systematically placed at these identified sites to generate possible configurations. This process is highly intuitive and can be applied to any other species of interest in catalytic reactions. Therefore, we believe the PH-SA method possesses broad applicability.

In summary, the two catalyst structures and reaction process we selected are representative, providing strong evidence for the general applicability of the method we developed.

We have revised the manuscript.

(Page 9)

To demonstrate the general applicability of the PH-SA method for diverse material morphologies, we selected two of the most commonly used catalyst structural models,

slab and cluster, for illustration.

(Page 18)

The hydrogen adsorption under reduction environment and oxygen adsorption under oxidation environment were selected as model process. The two catalyst structures and reaction process we demonstrated are both representative, providing strong evidence for the general applicability of the method.

- The citation of studies on the application of persistent homology, particularly in the field of materials science, seems biased towards certain research groups. Besides the references mentioned by the authors, there are several other studies that analyze structures and structure-property relationship using persistent homology. In recent years, the idea of applying persistent homology to materials science has gained momentum among multiple researchers. The authors should cite relevant previous works appropriately. Examples include:

<https://www.pnas.org/doi/full/10.1073/pnas.1520877113>

<https://www.science.org/doi/10.1126/sciadv.abc2320>

<https://doi.org/10.1063/5.0093441>

<https://www.nature.com/articles/ncomms15396>

<https://www.nature.com/articles/s41598-021-88027-8>

Response:

We appreciate the reviewer's recommended literature. These studies have significantly expanded the application scenarios of persistent homology in the field of materials science. We have cited them as ref 47-51. Additionally, we have also cited several other articles related to persistent homology (J. Chem. Phys. 159, 084101 (2023), J Cheminform 11, 48 (2019), J. Chem. Inf. Model. 2024, 64, 5242–5252).

- The connection between atom aggregation and the persistent homology calculation is unclear. It seems that persistent homology is applied to cluster structures formed by adding up to ten atoms near a specific atom or by introducing a cutoff radius to include surrounding atoms. Furthermore, the description in the Methods section, " Since we

start iterating from two atoms, the sites determined by the low-dimensional Betti numbers in the multi-atom aggregate will be included in the Betti number evolution of fewer atoms." suggests that the Betti numbers are calculated for structures containing two, three, four, etc., atoms. The authors should clarify how they merge the Betti number information obtained from each structure, or if they only use the results of the largest cluster.

Response:

We sincerely appreciate the reviewer's valuable suggestions.

During the computation of persistent homology for atomic clusters, we start from a single atom and construct atom aggregates consisting of two, three, up to ten atoms within a specified radial distance. This process is implemented using combinatorial enumeration. Additionally, duplicate atomic clusters with the same composition (atom index), which may arise from enumeration starting at different central atoms, are removed.

In the persistent homology analysis of atomic aggregates, we start with combinations of two atoms, which allow us to identify potential active sites between the two atoms. Next, we compute combinations of three atoms. Taking an octahedral structure as an example (Figure 2a), this process also identifies potential sites formed between pairs of atoms (corresponding to the moment when the β_0 vanishes). However, since these sites are already identified during the two-atom combination analysis, we avoid redundant computations by focusing solely on the potential active sites corresponding to the last death point in the barcode, excluding the one that always exists in β_0 . For an octahedral, this corresponds to the moment when the β_2 vanishes. For other atom aggregates, we follow a similar approach, computing only the potential active sites corresponding to the last death point in the barcode in each case.

The underlying mathematical implication of this operation is that the structural information from earlier death points of larger aggregates is generally already encapsulated within smaller aggregates that contain fewer atoms. The last death point of Betti number (excluding the persistent β_0) typically marks the disappearance of the largest void or cavity formed by most of the atoms in the current aggregate. Although

this conclusion has not been rigorously proven, this approach can strike a balance between computational efficiency and the identification of as many potential active sites as possible.

We would like to gently remind that our goal is to develop a method capable of identifying as many potential interaction sites as possible, rather than attempting to exhaustively enumerate all possibilities. The latter is inherently infeasible, as any method can only propose geometrically plausible sites without totally guaranteeing their physical feasibility, which ultimately requires validation through structural optimization based on DFT or MLFF.

In summary, for each atomic aggregate, we focus solely on the last death point in the barcode, excluding the one that always exists in β_0 . The computed active site can be directly mapped back to the original material structure since the atomic indices and coordinates of each aggregate are recorded.

We revisited the relevant expression in manuscript.

(Page 8 in manuscript)

Notably, while each Betti number's death point may correspond to a potential active site, in practice, we focus solely on the last death point (excluding the persistent β_0) for each aggregate. This shift stems from our methodology, which initiates with two atoms and progressively applies persistent homology to larger aggregates. Consequently, the structural information from earlier death points of larger aggregates is generally already captured in smaller aggregates with fewer atoms. Using the octahedral structure as an example, the bridge sites corresponding to the β_0 death points were already identified during the two-atom combination analysis. Using the octahedral structure as an example (Figure 2a), the bridge sites corresponding to the β_0 death points were already identified during the two-atom combination analysis.

(Page 19 in manuscript)

Intuitively, each death point in the barcode represents a potential active site in space, as shown in Figure 2. Here, we focus on the last death point (excluding the persistent β_0) for each aggregate, as it typically marks the disappearance of the largest void or cavity formed by most of the atoms in the aggregate. In material structures, foreign

atoms (e.g., hydrogen or oxygen) typically occupy these interstitial sites. Since our calculation begins with pairs of atoms, the structural information from earlier death points of larger aggregates is generally already captured in smaller aggregates with fewer atoms. For instance, the bridge sites on each edge of an octahedron are determined during the persistent homology analysis of two atoms (Figure2). Although this conclusion has not been rigorously proven, this approach can strike a balance between computational efficiency and the identification of as many potential active sites as possible.

- The atom aggregation process itself is not clear. How did the authors pick up the octahedron as shown in Figure 2 by picking up atoms near some specific atom one by one?

Response:

We appreciate the reviewer's question. As explained in the previous response, we perform combinatorial selections of atoms within a specified radial distance around a given atom, starting with 2 atoms, 3 atoms, and more. This process will naturally include the selection of octahedral structures with 6 atoms as long as all six atoms are within the cutoff range.

- In the Methods section, the statement "Through the filtering process, we determine the circumcenters of the geometric shapes formed by the initial positions of the atoms based on the distances obtained from the death point, which potentially represent the active sites." is difficult to understand. At what stage of atom aggregation is the circumcenter defined? It would be helpful to include specific equations or figures.

Response:

Thank you for the reviewer's thoughtful question. We apologize for the insufficient description of the computational details in this section. Upon careful review, we have decided to refrain from referring to the term "circumcenter" in our explanation. In

simple geometric structures such as squares or octahedra, the last death point of Betti number may coincidentally align with the circumcenter of the circumscribed circle (or sphere). These are special and intuitive cases, as shown in Figure 2 of octahedron and Figure S1 of square. However, as the number of atoms increases and the structures become more complex and disordered, this correspondence may not hold. Notably, when atomic clusters form concave polyhedral, a circumcenter cannot be defined.

Figure 2 provides an intuitive visualization of the intersections of various spheres generated during the persistent homology process in a simple case. However, it does not directly provide the spatial coordinates of these intersection points (what we called "circumcenter" last version) in three-dimensional space. In the following, we detail how the filtration parameter information derived from the death point can be utilized to determine the spatial coordinates of such points. Given that the atomic aggregates are generated through combinatorial arrangements, we can track which persistent barcode corresponds to a specific atomic aggregate. After obtaining the distance information associated with the last death point of the Betti number, we use the following formula to calculate the precise spatial coordinates of the point in three-dimensional space.

$$\min_x \sum_{i=1}^n w_i (\|p_i - x\| - d)^2 \quad (\text{R1})$$

$$w_i = \frac{r_i}{\sum_{j=1}^n r_j} \quad (\text{R2})$$

Here, p represents the three-dimensional coordinates of each atom within the atomic aggregate, recorded during enumeration. The w is the radius weight of different elements composing the atomic aggregate, r is the radius of each atom, i and j are the atom indices, x denotes the coordinates of interaction site to be determined, and d is the filtration parameter derived from the last death point in persistent homology. The underlying rationale of this formulation is as follows: the death point in persistent homology reflects the critical scale at which a ring and/or cavity is filled during the thickening process. This corresponding filtration parameter d can be interpreted as the maximum distance that the "expanding" spheres (centered at the points in the cloud) must travel from their origins to reach and fill this cavity. The d serves as a guiding value in our fitting formula, which aims to determine a geometric point x such that its

distance to all surrounding atoms closely approximates d . By minimizing the cumulative squared error in the objective function, we can determine the spatial coordinates of the interaction site. In this process, we also considered that when the material structure contains different elements, the spatial coordinates of the potential sites should be adjusted based on the radii ratio of the different elements, in order to enhance their physical plausibility. By leveraging this interpretation, the fitting process geometrically approximates the topological information provided by the death point, establishing a connection between the topological scale and spatial geometry.

As a note, our goal is to utilize persistent homology to identify potential initial active sites for atomic adsorption or insertion. It is neither feasible nor necessary to ensure that every identified site is physically optimal. The refinement and optimization of these sites will be achieved through subsequent machine learning force field-based calculations.

We have revised the method section in manuscript

(Page 6)

This is because its approximately equidistant positioning relative to surrounding atoms effectively minimizes repulsive interactions, facilitating the formation of energetically favorable bonding configurations. The geometric coordinates of the contact point can be readily determined through fitting constrained by the topological information of the death points.

(Page 16)

From a mathematical perspective, the "barcode death" point represents a critical spatial location associated with the scale at which a topological feature, such as a cavity or void, disappears during the thickening process. This point reflects a balance in distances to surrounding points, capturing the intrinsic geometry of the structure.

(Page 20)

We can obtain the filtering parameters of the last death point from persistent homology, but the corresponding spatial coordinates of the interaction sites cannot be directly acquired, even though their positions are intuitive, as illustrated in Figure 2. This requires a simple fitting step as follow:

$$\min_x \sum_{i=1}^n w_i (\|p_i - x\| - d)^2 \quad (1)$$

$$w_i = \frac{r_i}{\sum_{j=1}^n r_j} \quad (2)$$

Here, p represents the three-dimensional coordinates of each atom within the atomic aggregate, recorded during enumeration. The w is the radius weight of different elements composing the atomic aggregate, r is the radius of each atom, i and j are the atom indices, x denotes the coordinates of interaction site to be determined, and d is the filtration parameter derived from the last death point in persistent homology. The underlying rationale of this formulation is as follows: the death point in persistent homology reflects the critical scale at which a ring and/or cavity is filled during the thickening process. This corresponding filtration parameter d can be interpreted as the maximum distance that the "expanding" spheres (centered at the points in the cloud) must travel from their origins to reach and fill this cavity. The d serves as a guiding value in our fitting formula, which aims to determine a geometric point x such that its distance to all surrounding atoms closely approximates d . By minimizing the cumulative squared error in the objective function, we can determine the spatial coordinates of the interaction site. In this process, we also considered that when the material structure contains different elements, the spatial coordinates of the potential sites should be adjusted based on the radii ratio of the different elements, in order to enhance their physical plausibility. By leveraging this interpretation, the fitting process geometrically approximates the topological information provided by the death point, establishing a connection between the topological scale and spatial geometry.

- In the proposed method, persistent homology is applied to decompose the arrangement of atoms around a specific atom into polyhedral structures for interfacial, bulk, and cluster configurations. Persistent homology is just used to evaluate circumcenter of polyhedral structure. If this is the case, wouldn't it be computationally more efficient to directly compute the circumcenter of the polygons within the cluster structure formed during the atom aggregation process, without using persistent homology, to achieve the

same goal?

Response:

Thank you for the reviewer's thoughtful question. Persistent homology provides us with the filtration parameter at the death point, which significantly simplifies the computation of the geometric center in equation R1 and R2.

If persistent homology is not employed, equation R1 still remains a fundamental method for determining the spatial point corresponding to the "death point" (previously referred to as the circumcenter). However, in this case, solving the equation would no longer be a single-parameter optimization problem for the point x . Instead, it would become a dual-parameter optimization problem involving both x and the radius, which would substantially increase the computational complexity.

The increased complexity resulting from the absence of topological information poses a challenge in identifying the critical scale (or radius) at which the cavity becomes filled. Persistent homology bridges this gap by efficiently providing a filtration parameter that directly reflects this critical scale.

Using the Pd (100) slab with 64 atoms and Pt cluster with 55 atoms as examples, we compared site calculations with and without persistent homology. As shown in Figure R1, the time required without persistent homology is significantly longer, both approximately ten times greater. We also tested an Pd (711) unit cell with 718 atoms, and the fitting time without persistent homology was found to reach ~30,000 seconds. In contrast, the fitting time was reduced to approximately ~1,300 seconds when persistent homology was utilized. The difference in the time required between them expanded to nearly 25 times. When addressing more complex catalytic problems, models typically require even thousands of atoms (Nat Catal 7, 401–411 (2024), ACS Catal. 2022, 12, 14812–14824), at which the absolute time difference is expected to be much larger. All site enumeration was performed using 16 cores of an Intel Xeon Platinum 8336C central processing unit (CPU).

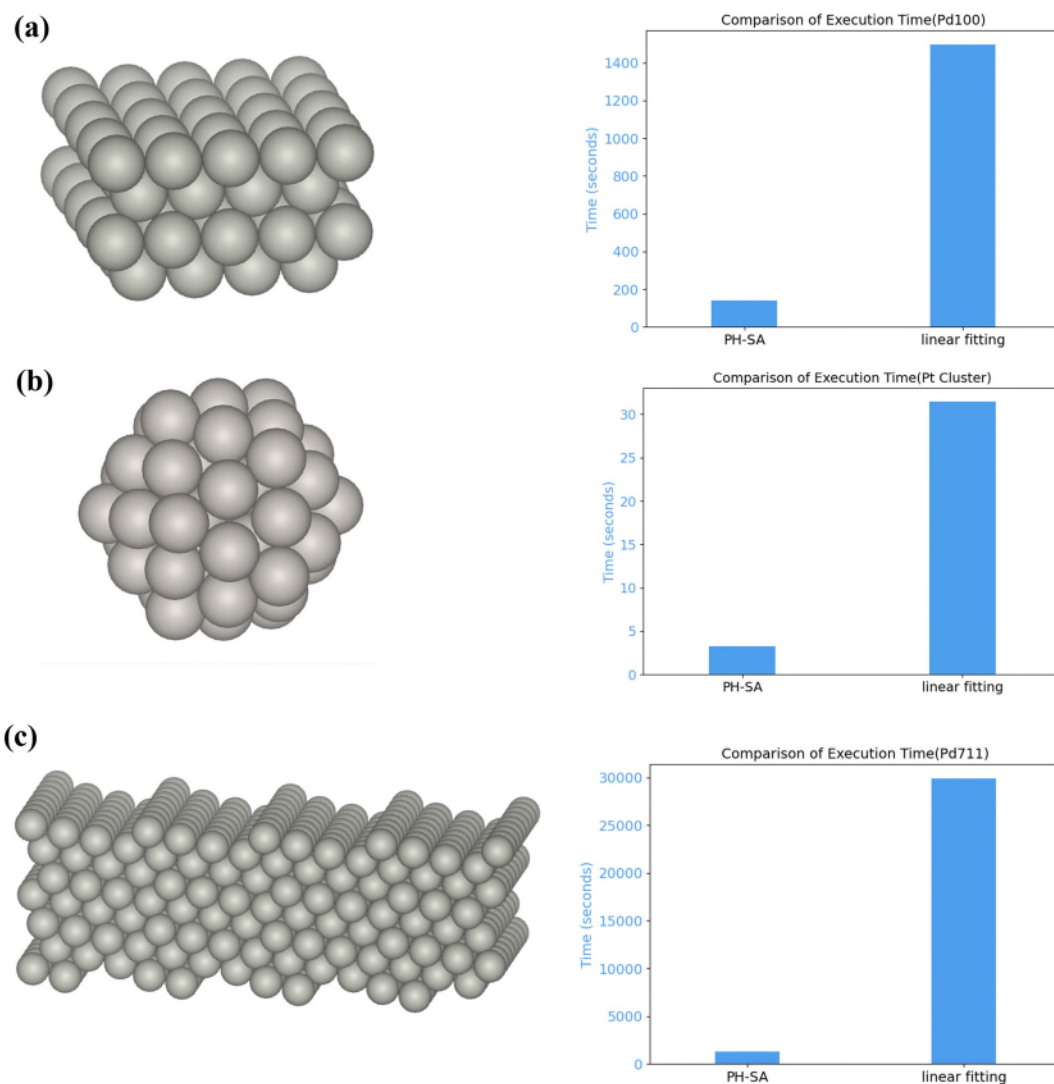


Figure R1. Comparison of the time required to determine the coordinates of binding sites using and without using persistent homology on Pd (100) slab with 64 atoms, Pt cluster with 55 atoms and Pd (711) slab with 718 atoms.

We have revised the method section and added the Figure R1 as Figure S14 with annotation.

(Page 20 in manuscript)

While it is possible to directly optimize the function for each atomic aggregate to identify coordinate without using persistent homology, this approach would transform the optimization into a two-parameter problem. As a result, the computational time would increase significantly, as demonstrated in Figure S14.

(Page 21-22 in SI)

Persistent homology provides us with the filtration parameter at the death point, which significantly simplifies the computation of the geometric center in equation 1 and 2.

If persistent homology is not employed, equation 1 still remains a fundamental method for determining the spatial point corresponding to the "death point". However, in this case, solving the equation would no longer be a single-parameter optimization problem for the point x . Instead, it would become a dual-parameter optimization problem involving both x and the radius, which would substantially increase the computational complexity.

The increased complexity resulting from the absence of topological information poses a challenge in identifying the critical scale (or radius) at which the cavity becomes filled. Persistent homology bridges this gap by efficiently providing a filtration parameter that directly reflects this critical scale.

Using the Pd (100) slab with 64 atoms and Pt cluster with 55 atoms as examples, we compared site calculations with and without persistent homology. As shown in Figure S14, the time required without persistent homology is significantly longer, both approximately ten times greater. We also tested a Pd (711) unit cell with 718 atoms, and the fitting time without persistent homology was found to reach ~30,000 seconds. In contrast, the fitting time was reduced to approximately ~1,300 seconds when persistent homology was utilized. The difference in the time required between them expanded to nearly 25 times. When addressing more complex catalytic problems, models typically require even thousands of atoms, at which the absolute time difference is expected to be much larger. All site enumeration was performed using 16 cores of an Intel Xeon Platinum 8336C central processing unit (CPU).

- Can the authors provide specific comparisons to demonstrate the superiority of the proposed method over other search algorithms?

Response:

We appreciate the reviewer's insightful question. Indeed, a direct quantitative comparison with previous methods is crucial. As we discuss in both the introduction and the discussion sections, global optimization methods are a major approach in material configuration search. Among these, basin hopping is one of the most popular and classical methods (The Journal of Physical Chemistry A, 1997, 101 (28), 5111-5116, Nano Lett. 2015, 15 (12), 8044-8, Chem. Sci., 2021, 12, 15543, J. Am. Chem. Soc. 2023, 145, 392–401), just a few cases. Therefore, we compare our approach with this well-established method.

First, we provide a brief introduction to the basic logic of the basin hopping method. It begins with generating a set of initial configurations randomly. These configurations are then optimized using an optimizer to find the most optimal configuration. In our case, the BFGS optimizer is used. Once the optimal configuration is found, a "hop" is performed from this configuration. This hopping step is achieved through random perturbations with step size at 0.5 Å, allowing the algorithm to jump out of local minima and explore new regions. The new configuration is then optimized, and according to the Metropolis criterion (equation R3), a decision is made on whether to accept the new configuration as the current optimal one. These steps are repeated until certain stopping criteria are met, such as maximum iterations or target energy thresholds.

$$P = \min(1, e^{-\Delta U/k_B T}) \quad (\text{R3})$$

Here, P refers to the probability of accepting an attempted "move" step, where ΔU is the potential energy change after the action and k_B is the Boltzmann constant. T is the temperature, which is a parameter that controls the acceptance probability and usually gradually reduced as iterations proceed. By following this criterion, the configurations are ensured to evolve toward lower energy states. More theoretical details of basin hopping can be found in the reference mentioned above. We implement the basin hopping algorithm by the atomic simulation environment (ASE) package (Ask Hjorth

Larsen et al 2017 J. Phys.: Condens. Matter 29 273002). All parameters, except for the optimizer and step size mentioned above, remain at their default settings.

Next, we proceed to compare the outcomes of the basin hopping approach with those derived from PH-SA, as applied to the Pd slab and Pt cluster systems. The comparison will be conducted in two aspects:

1. When the same number of samples are computed, we will contrast which method yields the most stable configurations with lower average adsorption energies. A configuration with a lower energy indicates that it is thermodynamically more likely to occur.
2. Under the condition of computing the same number of samples, we will also assess which method consumes more time.

The comparison of energy

The Pd-H system. For each hydrogen concentration ($\text{H}:\text{Pd} = 0.1\text{--}1$), we selected 10 initial seed configurations and then applied the basin-hopping algorithm, iterating 500 times to obtain 5000 configurations. During this process, we calculated the energies of the 5000 configurations of each concentration gradient. To minimize the impact of randomness, we repeated the process three times with different initial configurations. In the PH-SA, we directly enumerated 5000 configurations for each of the 10 concentration gradients and calculated their energies. The energy trends obtained by basin hopping are shown in Figure R2. As observed, for all three repetitions, the lowest-energy configurations identified by basin hopping were almost all more positive in energy than those found by the PH-SA method. Furthermore, the energy variation curve obtained from basin hopping does not reflect the compensatory effect of the hex phase transition on energy that occurs when $\text{H}:\text{Pd}=0.5$ (Figure 4c). This demonstrates the superiority of the PH-SA.

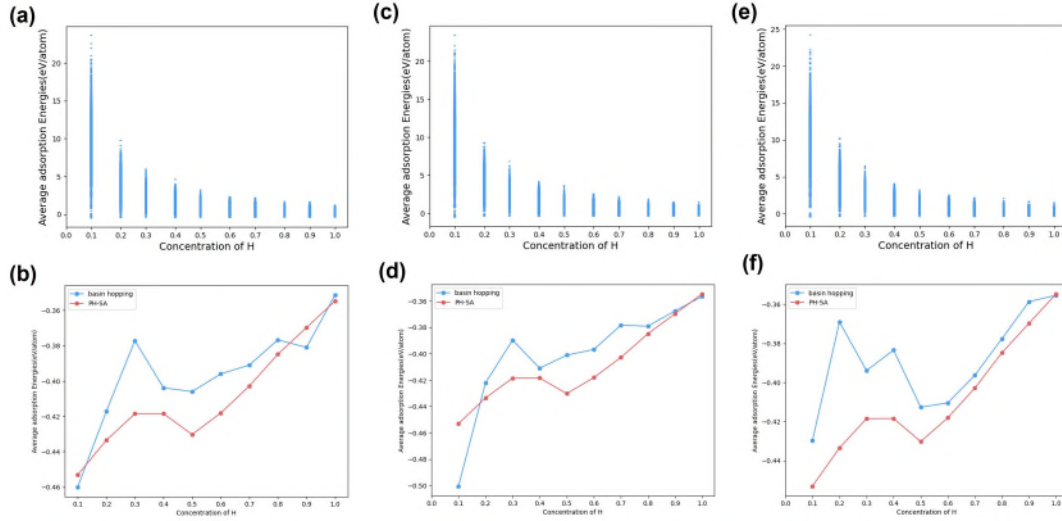


Figure R2. A comparison of sampling results based on the PH-SA versus those obtained using the basin-hopping method on Pd-H system is presented. Figures (a), (c), and (e) depict the average hydrogen (H) adsorption energy distributions derived from different random initial structures. Figures (b), (d), and (f) contrast the lowest adsorption energies at each H concentration achieved by the PH-SA method against those obtained using the basin-hopping method. The results obtained by the PH-SA method are represented by the red lines, which are sourced from Figure S4.

The Pt-O system. For the Pt-O system, when using the PH-SA, we first computed the energy distribution for the 10% oxygen concentration. We then identified potential insertion sites for new oxygen atoms based on the lowest-energy 5% of configurations obtained from the 10% oxygen concentration, and used these configurations to generate structures at a 20% oxygen concentration, repeating this process stepwise. This approach was adopted because Pt₅₅ clusters undergo continuous deformation and expansion as oxygen is inserted, which enhances the accuracy of sampling. To facilitate a fair comparison with the PH method, we directly used the lowest-energy 5% of configurations at 10% oxygen concentration as seed structures, randomly added 10% oxygen to form initial configurations for 20% oxygen, and applied the basin-hopping algorithm to obtain configurations and energies at this concentration, continuing in the same manner for other concentrations. The results (Figure R3) indicate that the average adsorption energy obtained through the basin hopping method is generally positive than

that obtained via the PH-SA method.

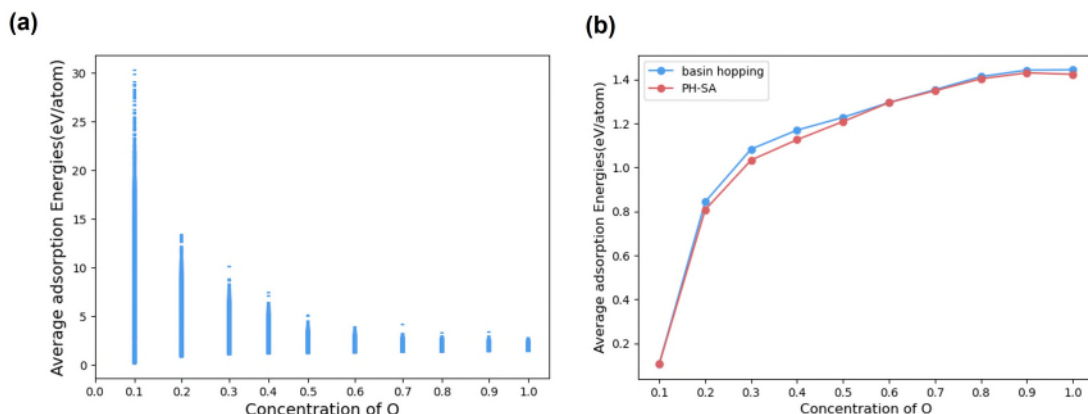


Figure R3. A comparison of sampling results based on the PH-SA versus those obtained using the basin-hopping method on Pt-O system is presented. Figures (a) depict the average oxygen (O) adsorption energy distributions with different oxygen concentration based on basin hopping. Figures (b) contrast the lowest adsorption energies at each oxygen concentration achieved by the PH-SA method against those obtained using the basin-hopping method. The results obtained by the PH-SA method are represented by the red lines, which are sourced from Figure 6.

For these two examples, although the absolute values of the average adsorption energies for the lowest-energy configurations obtained by the two methods appear close, this is due to normalization by the number of atoms. If we instead consider the total energy differences, the gap expands by 64 to 128 times for the Pd-H system and by 55 to 110 times for the Pt-O system, depending on the concentration of active species. This difference can reach several eVs in magnitude. Based on these results, while we cannot guarantee that our method will always outperform basin hopping in all systems, it has shown advantages in these specific cases.

The comparison of time

Before initiating the quantitative comparison, we would like to gently highlight that the PH-SA and the basin hopping method represent two fundamentally distinct technical approaches. The PH-SA excels in rapidly identifying and enumerating potential interaction sites across materials with diverse morphologies, ensuring that the

identified sites are geometrically plausible. This approach allows for an initial generation of configuration dataset, followed by the selective refinement of configurations of interest through structural optimization. In contrast, the basin hopping method integrates configuration generation with structural optimization in a sequential manner. New configurations are generated by applying perturbations to the current optimized configurations, requiring each step to undergo optimization before proceeding. Without the optimization step, the reliance on random perturbations would easily generate invalid, high-energy configurations, making the process less efficient and less directed.

The fundamental difference in logic between the two methods leads to a significant disparity in the time required for active phase exploration. Taking the Pd-H system as an example, in the basin hopping approach, we select 10 seed structures at each gradient step, resulting in a total of 100 seed structures across 10 gradient steps. Due to its sequential nature, only 100 tasks can be computed at a time, regardless of the computational resources available. In contrast, the PH-SA method directly generates 50000 configurations for ten hydrogen concentration gradients, which can be optimized simultaneously in a fully parallel manner, provided sufficient computational resources. When employed 600 computational cores (Intel Xeon Platinum 8336C), the results of time comparisons demonstrate that the PH-SA method offers an advantage in terms of computational efficiency on both systems on Pd-H system (Figure R4 a). Specifically, when employing the PH-SA method, each structure is optimized using a single core, allowing all 600 cores to work simultaneously. In contrast, for the basin hopping method, each seed structure uses 6 cores for exploration. Since the optimization time using 6 cores is not six times faster than using a single core, the overall time required for the PH-SA method is significantly reduced. A similar situation occurs in the Pt-O system, although the effect is less pronounced due to the larger number of seed structures involved (Figure R4 b). It is foreseeable that, with access to more computational resources, the time required for the PH-SA could be further reduced.

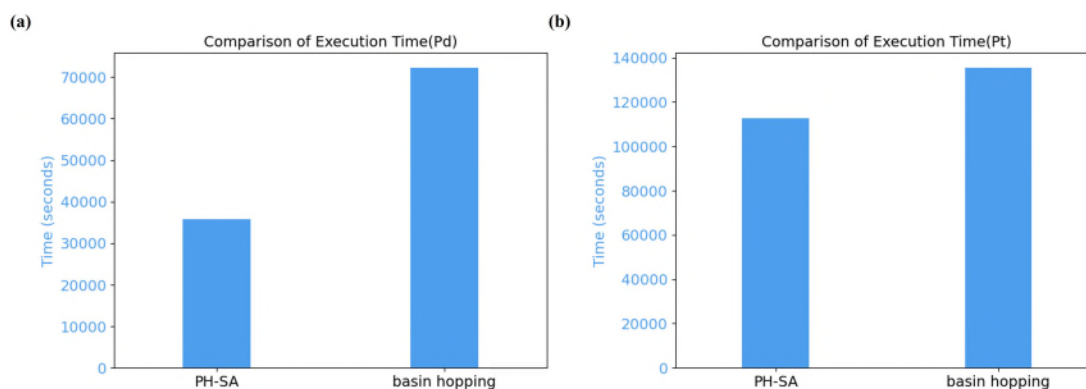


Figure R4. Comparison of the computational time required to explore (a) 50,000 Pd-H configurations and (b) 100,000 Pt-O configurations using the persistent homology-based approach and the basin hopping method, respectively.

In the paper, we also mentioned other mathematics-based structural enumeration methods, among which a representative approach is using graph theory to enumerate the coverage of surface adsorbates (npj Computational Materials 2020, 6 (1). Nat. Commun. 2022, 13 (1), 5788.). This method employs graph theory to identify surface sites, thereby enabling the enumeration of surface adsorbate configurations. In terms of its overall logic, it resembles the PH-SA method. However, due to the insensitivity of the graph representation to three-dimensional information, it is unable to account for subsurface and bulk structure sampling, which can alternatively be achieved by the PH method.

We have revised the manuscript and SI. We have added Figure R2- R4 as Figure S8-S10. We have also updated the code of basin hopping to GitHub.

(Page 16-17 in manuscript)

We compared the PH-SA sampling process with the classical basin hopping algorithm for the Pd-H and Pt-O systems. The generally lower average adsorption energy achieved by PH-SA highlights its superior ability to identify more stable structures (Figure S8 and Figure S9). Fundamentally, the PH-SA and the basin hopping method represent two distinct technical approaches. The PH-SA excels in rapidly identifying and enumerating potential interaction sites across materials with diverse morphologies, ensuring that the identified sites are geometrically plausible. This approach allows for an initial exploration of configuration distributions, followed by

the selective refinement of configurations of interest through structural optimization. In contrast, the basin hopping method integrates configuration generation with structural optimization in a sequential manner. New configurations are generated by applying perturbations to the current optimized configurations, requiring each step to undergo optimization before proceeding. Without the optimization step, the reliance on random perturbations would easily generate invalid, high-energy configurations, making the process less efficient and less directed. The difference in logic also endows the PH-SA method with greater potential for computational efficiency, as verified in the Pd-H and Pt-O systems (Figure S10).

(Page 12-17 in SI)

The comparison with basin hopping

Indeed, a direct quantitative comparison with previous methods is crucial. As we discuss in both the introduction and the discussion sections, global optimization methods are a major approach in material configuration search. Among these, basin hopping is one of the most popular and classical methods. Therefore, we compare our approach with this well-established method.

First, we provide a brief introduction to the basic logic of the basin hopping method. It begins with generating a set of initial configurations randomly. These configurations are then optimized using an optimizer to find the most optimal configuration. In our case, the BFGS optimizer is used. Once the optimal configuration is found, a "hop" is performed from this configuration. This hopping step is achieved through random perturbations with step size at 0.5 Å, allowing the algorithm to jump out of local minima and explore new regions. The new configuration is then optimized, and according to the Metropolis criterion (equation R3), a decision is made on whether to accept the new configuration as the current optimal one. These steps are repeated until certain stopping criteria are met, such as maximum iterations or target energy thresholds.

$$P = \min(1, e^{-\Delta U/k_B T})$$

Here, P refers to the probability of accepting an attempted “move” step. where ΔU is the potential energy change after the action and k_B is the Boltzmann constant. T is the temperature, which is a parameter that controls the acceptance probability and usually

gradually reduced as iterations proceed. By following this criterion, the configurations are ensured to evolve toward lower energy states. More theoretical details of basin hopping can be found in the reference mentioned above. We implement the basin hopping algorithm by the atomic simulation environment (ASE) package. All parameters, except for the optimizer and step size mentioned above, remain at their default settings.

Next, we proceed to compare the outcomes of the basin hopping approach with those derived from PH-SA, as applied to the Pd slab and Pt cluster systems. The comparison will be conducted in two aspects:

1. When the same number of samples are computed, we will contrast which method yields the most stable configurations with lower average adsorption energies. A configuration with a lower energy indicates that it is thermodynamically more likely to occur.
2. Under the condition of computing the same number of samples, we will also assess which method consumes more time.

The comparison of energy

The Pd-H system. For each hydrogen concentration ($\text{H}:\text{Pd} = 0.1\text{--}1$), we selected 10 initial seed configurations and then applied the basin-hopping algorithm, iterating 500 times to obtain 5000 configurations. During this process, we calculated the energies of the 5000 configurations of each concentration gradient. To minimize the impact of randomness, we repeated the process three times with different initial configurations. In the PH-SA, we directly enumerated 5000 configurations for each of the 10 concentration gradients and calculated their energies. The energy trends obtained by basin hopping are shown in Figure S8. As observed, for all three repetitions, the lowest-energy configurations identified by basin hopping were almost all more positive in energy than those found by the PH-SA method. Furthermore, the energy variation curve obtained from basin hopping does not reflect the compensatory effect of the hex phase transition on energy that occurs when $\text{H}:\text{Pd}=0.5$ (Figure 4c). These demonstrate the superiority of the PH-SA.

The Pt-O system. For the Pt-O system, when using the PH method, we first computed

the energy distribution for the 10% oxygen concentration. We then identified potential insertion sites for new oxygen atoms based on the lowest-energy 5% of configurations obtained from the 10% oxygen concentration, and used these configurations to generate structures at a 20% oxygen concentration, repeating this process stepwise. This approach was adopted because Pt_{55} clusters undergo continuous deformation and expansion as oxygen is inserted, which enhances the accuracy of sampling. To facilitate a fair comparison with the PH method, we directly used the lowest-energy 5% of configurations at 10% oxygen concentration as seed structures, randomly added 10% oxygen to form initial configurations for 20% oxygen, and applied the basin-hopping algorithm to obtain configurations and energies at this concentration, continuing in the same manner for other concentrations. The results (Figure S9) indicate that the average adsorption energy obtained through the basin hopping method is generally positive than that obtained via the PH-SA method.

For these two examples, although the absolute values of the average adsorption energies for the lowest-energy configurations obtained by the two methods appear close, this is due to normalization by the number of atoms. If we instead consider the total energy differences, the gap expands by 64 to 128 times for the Pd-H system and by 55 to 110 times for the Pt-O system, depending on the concentration of active species. This difference can reach several eVs in magnitude. Based on these results, while we cannot guarantee that our method will always outperform basin hopping in all systems, it has shown advantages in these specific cases.

The comparison of time

Before initiating the quantitative comparison, we would like to gently highlight that the PH-SA and the basin hopping method represent two fundamentally distinct technical approaches. The PH-SA excels in rapidly identifying and enumerating potential interaction sites across materials with diverse morphologies, ensuring that the identified sites are geometrically plausible. This approach allows for an initial generation of configuration dataset, followed by the selective refinement of configurations of interest through structural optimization. In contrast, the basin hopping

method integrates configuration generation with structural optimization in a sequential manner. New configurations are generated by applying perturbations to the current optimized configurations, requiring each step to undergo optimization before proceeding. Without the optimization step, the reliance on random perturbations would easily generate invalid, high-energy configurations, making the process less efficient and less directed.

The fundamental difference in logic between the two methods leads to a significant disparity in the time required for active phase exploration. Taking the Pd-H system as an example, in the basin hopping approach, we select 10 seed structures at each gradient step, resulting in a total of 100 seed structures across 10 gradient steps. Due to its sequential nature, only 100 tasks can be computed at a time, regardless of the computational resources available. In contrast, the PH-SA method directly generates 50000 configurations for ten hydrogen concentration gradients, which can be optimized simultaneously in a fully parallel manner, provided sufficient computational resources. When employed 600 computational cores (Intel Xeon Platinum 8336C), the results of time comparisons demonstrate that the PH-SA method offers an advantage in terms of computational efficiency on both systems on Pd-H system (Figure S10 a). Specifically, when employing the PH-SA method, each structure is optimized using a single core, allowing all 600 cores to work simultaneously. In contrast, for the basin hopping method, each seed structure uses 6 cores for exploration. Since the optimization time using 6 cores is not six times faster than using a single core, the overall time required for the PH-SA method is significantly reduced. A similar situation occurs in the Pt-O system, although the effect is less pronounced due to the larger number of seed structures involved (Figure S10 b). It is foreseeable that, with access to more computational resources, the time required for the PH-SA could be further reduced.

- Which codes were used to compute persistent homology? GUDHI or ripser or other codes? If the authors used some open-source code, it should be cited.

Response:

We thank the reviewers. We use the GUDHI for the computation of persistent homology (https://doi.org/10.1007/978-3-662-44199-2_28).

We have cited it as ref 86 and we have revised the method section in page 19

The implementation of the persistent homology relies on the open-source package GUDHI.

- What kind of filtration (Vietoris-Rips or alpha) was used to obtain the barcode?

Response:

In this study, we utilized both Alpha complexes and Vietoris–Rips (VR) complexes to construct filtrations and compute barcodes for atomic aggregates of varying sizes and complexities.

Specifically:

Small-scale systems (up to four atoms): Alpha complexes were employed for smaller and simpler structures. These types of atomic aggregate data are the most abundant in material structures, and the high computational efficiency of Alpha complexes can significantly reduce the time required to process these structures.

Larger-scale systems (more than four atoms): For larger aggregates, VR complexes were used. VR complexes account for all pairwise distances within the point cloud, making them particularly effective at capturing subtle topological features in complex or amorphous structures, which often involve more atoms. Although VR complexes incur higher computational costs, their ability to preserve detailed connectivity information facilitates detecting intricate topological features, especially in non-crystalline or highly disordered systems.

The choice of four atoms as the dividing threshold is primarily due to the fact that VR complexes do not generate β_1 classes when processing tetrahedral structures (Figure R5). As a result, they cannot directly identify the central positions of tetrahedra, which is an intuitive and common interaction site in material structure. This limitation is also one of the key reasons why we prioritize the use of Alpha complexes for small-scale systems.

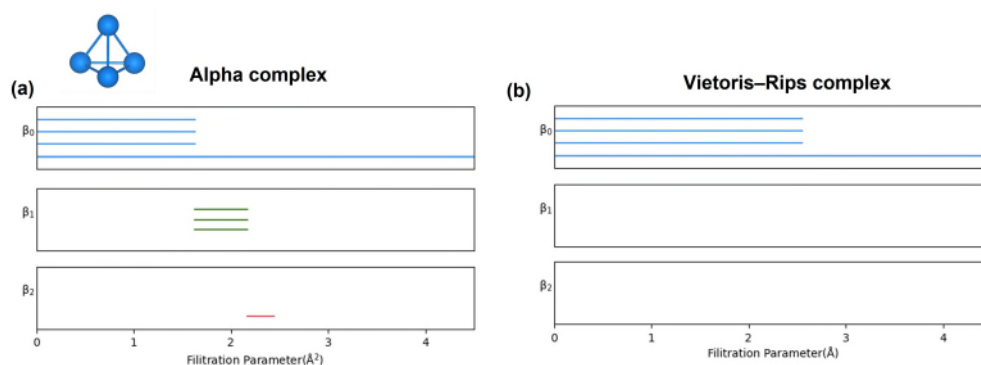


Figure R5. The filtration process of tetrahedral structure based on Alpha complex and Vietoris–Rips complex.

Moreover, we have offered an option in our code to perform analyses entirely using Alpha complexes. For scenarios requiring rapid exploration of potential active sites, using Alpha complexes exclusively can yield results in a shorter time. However, the obtained sites may be fewer compared to those identified using the VR complex method, as demonstrated by the examples of pure Pd and PtO_{0.5} with different morphology (Figure R6).

In summary, we can flexibly select Alpha or VR complexes based on the size and morphology of the atomic aggregates to balance the number of active sites and the computational efficiency.

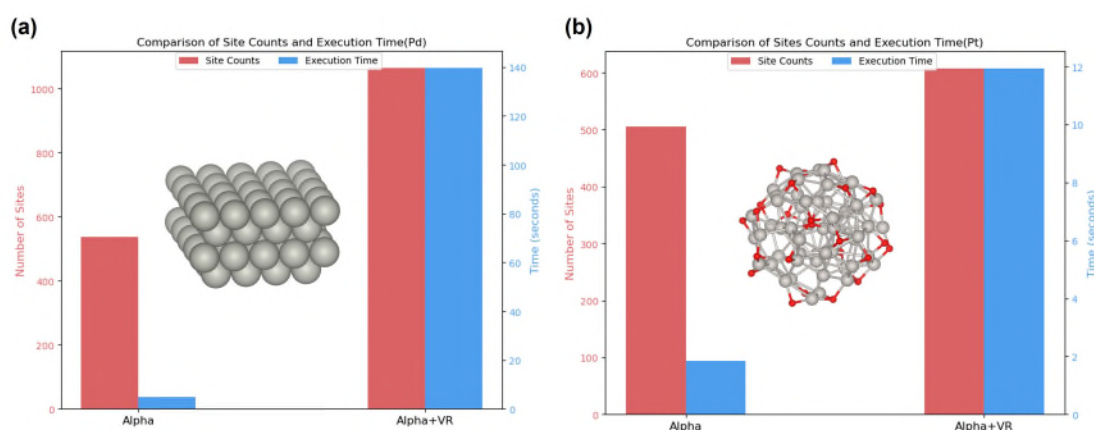


Figure R6. A comparison of the number of identified interaction sites and the time taken using either Alpha complex or both Alpha and VR complex on (a) pure Pd and (b) PtO_{0.5}.

We have revised the method section and added Figure R5-R6 as Figure S12-

S13 with annotation.

(Page 18 in manuscript)

When performing persistent homology, it is also necessary to specify the type of complex used. For atomic aggregates consisting of four or fewer atoms, we employ the Alpha complex, while for aggregates with more than four atoms, we utilize the Vietoris–Rips complex. The advantages of this combined strategy are discussed in Figure SX.

(Page 19-20 in SI)

In this study, we utilized both Alpha complexes and Vietoris–Rips (VR) complexes to construct filtrations and compute barcodes for atomic aggregates of varying sizes and complexities.

Specifically:

Small-scale systems (up to four atoms): Alpha complexes were employed for smaller and simpler structures. These types of atomic aggregate data are the most abundant in material structures, and the high computational efficiency of Alpha complexes can significantly reduce the time required to process these structures.

Larger-scale systems (more than four atoms): For larger aggregates, VR complexes were used. VR complexes account for all pairwise distances within the point cloud, making them particularly effective at capturing subtle topological features in complex or amorphous structures, which often involve more atoms. Although VR complexes incur higher computational costs, their ability to preserve detailed connectivity information facilitates detecting intricate topological features, especially in non-crystalline or highly disordered systems.

The choice of four atoms as the dividing threshold is primarily due to the fact that VR complexes do not generate β_1 classes when processing tetrahedral structures (Figure S12). As a result, they cannot directly identify the central positions of tetrahedra, which is an intuitive and common interaction site in material structure. This limitation is also one of the key reasons why we prioritize the use of Alpha complexes for small-scale systems.

Moreover, we have offered an option in our code to perform analyses entirely using

Alpha complexes. For scenarios requiring rapid exploration of potential active sites, using Alpha complexes exclusively can yield results in a shorter time. However, the obtained sites may be fewer compared to those identified using the VR complex method, as demonstrated by the examples of pure Pd and $\text{PtO}_{0.5}$ with different morphology (Figure S13).

In summary, we can flexibly select Alpha or VR complexes based on the size and morphology of the atomic aggregates to balance the number of active sites and the computational efficiency.

Reviewer #2 (Remarks to the Author):

Summary:

This work proposes the application of persistence homology to sample active phases of heterogeneous catalysts, and further proposes the use of persistence homology as a data extraction method for machine learning.

Comments:

- The paper justifies the application of persistence homology with a well written and persuasive exposition. The discussion of Figure 2 particularly shows the strength of this approach as a simple feature extractor for machine learning applications.
- The paper fully discusses two applications of their work, showing the versatility of their approach. I generally appreciate the depth of the exposition the authors provide.

Response:

We sincerely thank the reviewer for the valuable time dedicated to our work and for the acknowledgment of our efforts. We have carefully revised the article according to the reviewer's requirements.

- There are several past proposed methodologies for the sampling the paper proposes (e.g., citations 27-32 of the paper). However, no comparisons are made between these methods and the paper's proposed new method. Serious evaluation of this paper requires a comparison with current state of the art techniques to demonstrate improvements in either clock time or evaluation quality. I view this as the most serious weakness of the paper: in its current form, the comparisons to past works are purely qualitative, without data-driven support.

Response:

We appreciate the reviewer's insightful question. Indeed, a direct quantitative comparison with previous methods is crucial. As we discuss in both the introduction and the discussion sections, global optimization methods are a major approach in material configuration search. Among these, basin hopping is one of the most popular and classical methods (The Journal of Physical Chemistry A, 1997, 101 (28), 5111-5116, Nano Lett. 2015, 15 (12), 8044-8, Chem. Sci., 2021, 12, 15543, J. Am. Chem. Soc. 2023, 145, 392–401), just a few cases. Therefore, we compare our approach with this well-established method.

First, we provide a brief introduction to the basic logic of the basin hopping method. It begins with generating a set of initial configurations randomly. These configurations are then optimized using an optimizer to find the most optimal configuration. In our case, the BFGS optimizer is used. Once the optimal configuration is found, a "hop" is performed from this configuration. This hopping step is achieved through random perturbations with step size at 0.5 Å, allowing the algorithm to jump out of local minima and explore new regions. The new configuration is then optimized, and according to the Metropolis criterion (equation R3), a decision is made on whether to accept the new configuration as the current optimal one. These steps are repeated until certain stopping criteria are met, such as maximum iterations or target energy thresholds.

$$P = \min(1, e^{-\Delta U/k_B T}) \quad (\text{R3})$$

Here, P refers to the probability of accepting an attempted “move” step. where ΔU is the potential energy change after the action and k_B is the Boltzmann constant. T is the temperature, which is a parameter that controls the acceptance probability and usually gradually reduced as iterations proceed. By following this criterion, the configurations are ensured to evolve toward lower energy states. More theoretical details of basin hopping can be found in the reference mentioned above. We implement the basin hopping algorithm by the atomic simulation environment (ASE) package (Ask Hjorth Larsen et al 2017 J. Phys.: Condens. Matter 29 273002). All parameters, except for the optimizer and step size mentioned above, remain at their default settings.

Next, we proceed to compare the outcomes of the basin hopping approach with those derived from PH-SA, as applied to the Pd slab and Pt cluster systems. The comparison will be conducted in two aspects:

1. When the same number of samples are computed, we will contrast which method yields the most stable configurations with lower average adsorption energies. A configuration with a lower energy indicates that it is thermodynamically more likely to occur.
2. Under the condition of computing the same number of samples, we will also assess which method consumes more time.

The comparison of energy

The Pd-H system. For each hydrogen concentration ($H: Pd = 0.1-1$), we selected 10 initial seed configurations and then applied the basin-hopping algorithm, iterating 500 times to obtain 5000 configurations. During this process, we calculated the energies of the 5000 configurations of each concentration gradient. To minimize the impact of randomness, we repeated the process three times with different initial configurations. In the persistent homology-based sampling approach (PH-SA), we directly enumerated 5000 configurations for each of the 10 concentration gradients and calculated their energies. The energy trends obtained by basin hopping are shown in Figure R2. As observed, for all three repetitions, the lowest-energy configurations identified by basin hopping were almost all more positive in energy than those found by the PH-SA method. Furthermore, the energy variation curve obtained from basin hopping does not reflect the compensatory effect of the hex phase transition on energy that occurs when $H: Pd=0.5$ (Figure 4c). This demonstrates the superiority of the PH-SA.

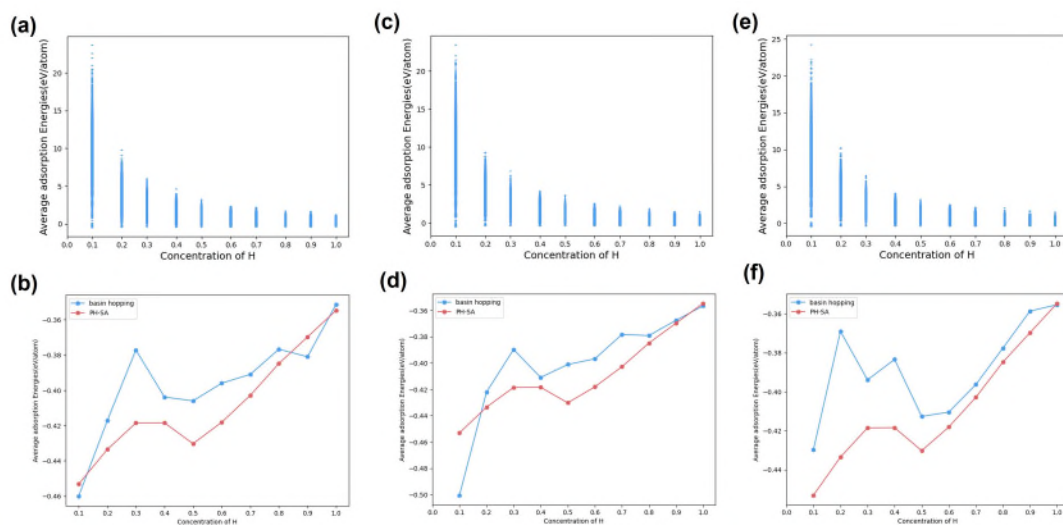


Figure R2. A comparison of sampling results based on the PH-SA versus those obtained using the basin-hopping method on Pd-H system is presented. Figures (a), (c), and (e) depict the average hydrogen (H) adsorption energy distributions derived from different random initial structures. Figures (b), (d), and (f) contrast the lowest adsorption energies at each H concentration achieved by the PH-SA method against those obtained using the basin-hopping method. The results obtained by the PH-SA method are represented by the red lines, which are sourced from Figure S4.

The Pt-O system. For the Pt-O system, when using the PH-SA, we first computed the energy distribution for the 10% oxygen concentration. We then identified potential insertion sites for new oxygen atoms based on the lowest-energy 5% of configurations obtained from the 10% oxygen concentration, and used these configurations to generate structures at a 20% oxygen concentration, repeating this process stepwise. This approach was adopted because Pt₅₅ clusters undergo continuous deformation and expansion as oxygen is inserted, which enhances the accuracy of sampling. To facilitate a fair comparison with the PH method, we directly used the lowest-energy 5% of configurations at 10% oxygen concentration as seed structures, randomly added 10% oxygen to form initial configurations for 20% oxygen, and applied the basin-hopping algorithm to obtain configurations and energies at this concentration, continuing in the same manner for other concentrations. The results (Figure R3) indicate that the average adsorption energy obtained through the basin hopping method is generally positive than that obtained via the PH-SA method.

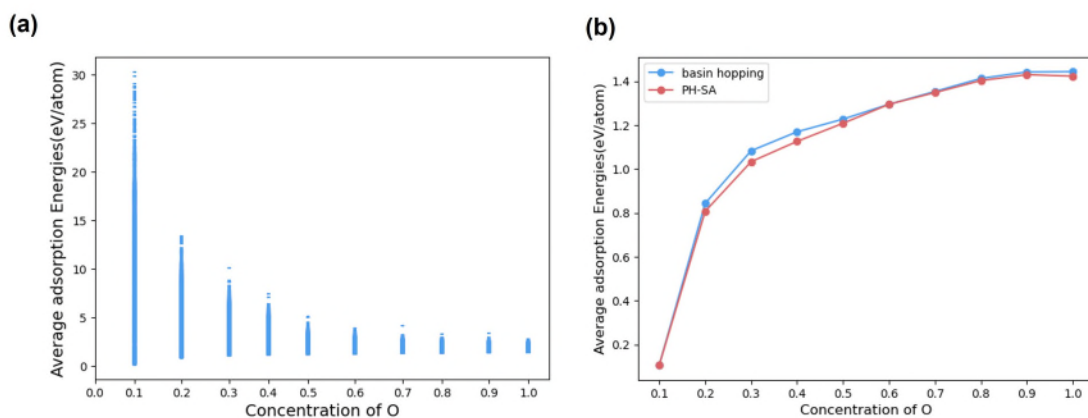


Figure R3. A comparison of sampling results based on the PH-SA versus those obtained using the basin-hopping method on Pt-O system is presented. Figures (a) depict the average oxygen (O) adsorption energy distributions with different oxygen concentration based on basin hopping. Figures (b) contrast the lowest adsorption energies at each oxygen concentration achieved by the PH-SA method against those obtained using the basin-hopping method. The results obtained by the PH-SA method are represented by the red lines, which are sourced from Figure 6.

For these two examples, although the absolute values of the average adsorption energies for the lowest-energy configurations obtained by the two methods appear close, this is due to normalization by the number of atoms. If we instead consider the total energy differences, the gap expands by 64 to 128 times for the Pd-H system and by 55 to 110 times for the Pt-O system, depending on the concentration of active species. This difference can reach several eVs in magnitude. Based on these results, while we cannot guarantee that our method will always outperform basin hopping in all systems, it has shown advantages in these specific cases.

The comparison of time

Before initiating the quantitative comparison, we would like to gently highlight that the PH-SA and the basin hopping method represent two fundamentally distinct technical approaches. The PH-SA excels in rapidly identifying and enumerating potential interaction sites across materials with diverse morphologies, ensuring that the identified sites are geometrically plausible. This approach allows for an initial generation of configuration dataset, followed by the selective refinement of configurations of interest through structural optimization. In contrast, the basin hopping method integrates configuration generation with structural optimization in a sequential manner. New configurations are generated by applying perturbations to the current optimized configurations, requiring each step to undergo optimization before proceeding. Without the optimization step, the reliance on random perturbations would easily generate invalid, high-energy configurations, making the process less efficient and less directed.

The fundamental difference in logic between the two methods leads to a significant disparity in the time required for active phase exploration. Taking the Pd-H system as an example, in the basin hopping approach, we select 10 seed structures at each gradient step, resulting in a total of 100 seed structures across 10 gradient steps. Due to its sequential nature, only 100 tasks can be computed at a time, regardless of the computational resources available. In contrast, the PH-SA method directly generates 50000 configurations for ten hydrogen concentration gradients, which can be optimized

simultaneously in a fully parallel manner, provided sufficient computational resources. When employed 600 computational cores (Intel Xeon Platinum 8336C), the results of time comparisons demonstrate that the PH-SA method offers an advantage in terms of computational efficiency on both systems on Pd-H system (Figure R4 a). Specifically, when employing the PH-SA method, each structure is optimized using a single core, allowing all 600 cores to work simultaneously. In contrast, for the basin hopping method, each seed structure uses 6 cores for exploration. Since the optimization time using 6 cores is not six times faster than using a single core, the overall time required for the PH-SA method is significantly reduced. A similar situation occurs in the Pt-O system, although the effect is less pronounced due to the larger number of seed structures involved (Figure R4 b). It is foreseeable that, with access to more computational resources, the time required for the PH-SA could be further reduced.

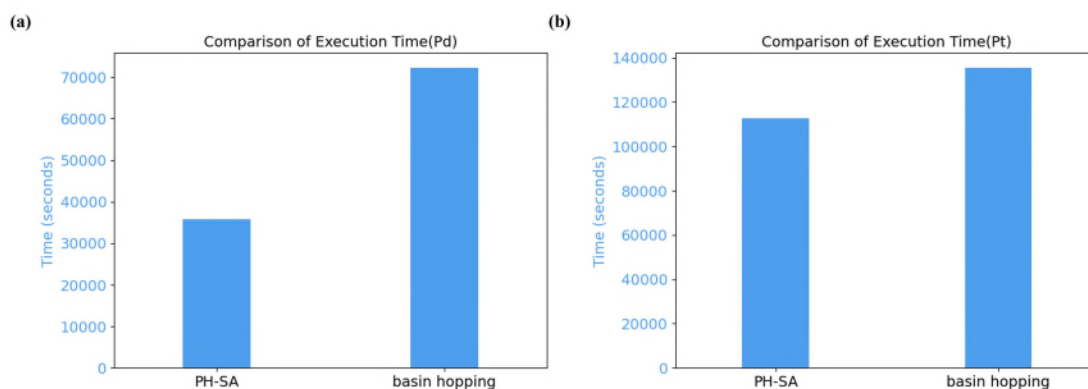


Figure R4. Comparison of the computational time required to explore (a) 50,000 Pd-H configurations and (b) 100,000 Pt-O configurations using the persistent homology-based approach and the basin hopping method, respectively.

In the paper, we also mentioned other mathematics-based structural enumeration methods, among which a representative approach is using graph theory to enumerate the coverage of surface adsorbates (npj Computational Materials 2020, 6 (1). Nat. Commun. 2022, 13 (1), 5788.). This method employs graph theory to identify surface sites, thereby enabling the enumeration of surface adsorbate configurations. In terms of its overall logic, it resembles the PH-SA method. However, due to the insensitivity of the graph representation to three-dimensional information, it is unable to account for

subsurface and bulk structure sampling, which can alternatively be achieved by the PH method.

We have revised the manuscript and SI. We have added Figure R2- R4 as Figure S8-S10. We have also updated the code of basin hopping to GitHub.

(Page 16-17 in manuscript)

We compared the PH-SA sampling process with the classical basin hopping algorithm for the Pd-H and Pt-O systems. The generally lower average adsorption energy achieved by PH-SA highlights its superior ability to identify more stable structures (Figure S8 and Figure S9). Fundamentally, the PH-SA and the basin hopping method represent two distinct technical approaches. The PH-SA excels in rapidly identifying and enumerating potential interaction sites across materials with diverse morphologies, ensuring that the identified sites are geometrically plausible. This approach allows for an initial exploration of configuration distributions, followed by the selective refinement of configurations of interest through structural optimization. In contrast, the basin hopping method integrates configuration generation with structural optimization in a sequential manner. New configurations are generated by applying perturbations to the current optimized configurations, requiring each step to undergo optimization before proceeding. Without the optimization step, the reliance on random perturbations would easily generate invalid, high-energy configurations, making the process less efficient and less directed. The difference in logic also endows the PH-SA method with greater potential for computational efficiency, as verified in the Pd-H and Pt-O systems (Figure S10).

(Page 12-17 in SI)

The comparison with basin hopping

Indeed, a direct quantitative comparison with previous methods is crucial. As we discuss in both the introduction and the discussion sections, global optimization methods are a major approach in material configuration search. Among these, basin hopping is one of the most popular and classical methods. Therefore, we compare our approach with this well-established method.

First, we provide a brief introduction to the basic logic of the basin hopping method.

It begins with generating a set of initial configurations randomly. These configurations are then optimized using an optimizer to find the most optimal configuration. In our case, the BFGS optimizer is used. Once the optimal configuration is found, a "hop" is performed from this configuration. This hopping step is achieved through random perturbations with step size at 0.5 Å, allowing the algorithm to jump out of local minima and explore new regions. The new configuration is then optimized, and according to the Metropolis criterion (equation R3), a decision is made on whether to accept the new configuration as the current optimal one. These steps are repeated until certain stopping criteria are met, such as maximum iterations or target energy thresholds.

$$P = \min(1, e^{-\Delta U/k_B T})$$

Here, P refers to the probability of accepting an attempted “move” step. where ΔU is the potential energy change after the action and k_B is the Boltzmann constant. T is the temperature, which is a parameter that controls the acceptance probability and usually gradually reduced as iterations proceed. By following this criterion, the configurations are ensured to evolve toward lower energy states. More theoretical details of basin hopping can be found in the reference mentioned above. We implement the basin hopping algorithm by the atomic simulation environment (ASE) package. All parameters, except for the optimizer and step size mentioned above, remain at their default settings.

Next, we proceed to compare the outcomes of the basin hopping approach with those derived from PH-SA, as applied to the Pd slab and Pt cluster systems. The comparison will be conducted in two aspects:

1. When the same number of samples are computed, we will contrast which method yields the most stable configurations with lower average adsorption energies. A configuration with a lower energy indicates that it is thermodynamically more likely to occur.
2. Under the condition of computing the same number of samples, we will also assess which method consumes more time.

The comparison of energy

The Pd-H system. For each hydrogen concentration (H: Pd = 0.1–1), we selected 10

initial seed configurations and then applied the basin-hopping algorithm, iterating 500 times to obtain 5000 configurations. During this process, we calculated the energies of the 5000 configurations of each concentration gradient. To minimize the impact of randomness, we repeated the process three times with different initial configurations. In the PH-SA, we directly enumerated 5000 configurations for each of the 10 concentration gradients and calculated their energies. The energy trends obtained by basin hopping are shown in Figure S8. As observed, for all three repetitions, the lowest-energy configurations identified by basin hopping were almost all more positive in energy than those found by the PH-SA method. Furthermore, the energy variation curve obtained from basin hopping does not reflect the compensatory effect of the hex phase transition on energy that occurs when H: Pd=0.5 (Figure 4c). These demonstrate the superiority of the PH-SA.

The Pt-O system. For the Pt-O system, when using the PH method, we first computed the energy distribution for the 10% oxygen concentration. We then identified potential insertion sites for new oxygen atoms based on the lowest-energy 5% of configurations obtained from the 10% oxygen concentration, and used these configurations to generate structures at a 20% oxygen concentration, repeating this process stepwise. This approach was adopted because Pt₅₅ clusters undergo continuous deformation and expansion as oxygen is inserted, which enhances the accuracy of sampling. To facilitate a fair comparison with the PH method, we directly used the lowest-energy 5% of configurations at 10% oxygen concentration as seed structures, randomly added 10% oxygen to form initial configurations for 20% oxygen, and applied the basin-hopping algorithm to obtain configurations and energies at this concentration, continuing in the same manner for other concentrations. The results (Figure S9) indicate that the average adsorption energy obtained through the basin hopping method is generally positive than that obtained via the PH-SA method.

For these two examples, although the absolute values of the average adsorption energies for the lowest-energy configurations obtained by the two methods appear close, this is due to normalization by the number of atoms. If we instead consider the total energy differences, the gap expands by 64 to 128 times for the Pd-H system and by 55

to 110 times for the Pt-O system, depending on the concentration of active species. This difference can reach several eVs in magnitude. Based on these results, while we cannot guarantee that our method will always outperform basin hopping in all systems, it has shown advantages in these specific cases.

The comparison of time

Before initiating the quantitative comparison, we would like to gently highlight that the PH-SA and the basin hopping method represent two fundamentally distinct technical approaches. The PH-SA excels in rapidly identifying and enumerating potential interaction sites across materials with diverse morphologies, ensuring that the identified sites are geometrically plausible. This approach allows for an initial generation of configuration dataset, followed by the selective refinement of configurations of interest through structural optimization. In contrast, the basin hopping method integrates configuration generation with structural optimization in a sequential manner. New configurations are generated by applying perturbations to the current optimized configurations, requiring each step to undergo optimization before proceeding. Without the optimization step, the reliance on random perturbations would easily generate invalid, high-energy configurations, making the process less efficient and less directed.

The fundamental difference in logic between the two methods leads to a significant disparity in the time required for active phase exploration. Taking the Pd-H system as an example, in the basin hopping approach, we select 10 seed structures at each gradient step, resulting in a total of 100 seed structures across 10 gradient steps. Due to its sequential nature, only 100 tasks can be computed at a time, regardless of the computational resources available. In contrast, the PH-SA method directly generates 50000 configurations for ten hydrogen concentration gradients, which can be optimized simultaneously in a fully parallel manner, provided sufficient computational resources. When employed 600 computational cores (Intel Xeon Platinum 8336C), the results of time comparisons demonstrate that the PH-SA method offers an advantage in terms of computational efficiency on both systems on Pd-H system (Figure S10 a). Specifically,

when employing the PH-SA method, each structure is optimized using a single core, allowing all 600 cores to work simultaneously. In contrast, for the basin hopping method, each seed structure uses 6 cores for exploration. Since the optimization time using 6 cores is not six times faster than using a single core, the overall time required for the PH-SA method is significantly reduced. A similar situation occurs in the Pt-O system, although the effect is less pronounced due to the larger number of seed structures involved (Figure S10 b). It is foreseeable that, with access to more computational resources, the time required for the PH-SA could be further reduced.

- The authors may be interested in Reinauer et al.'s paper "Persformer: A Transformer Architecture for Topological Machine Learning", which demonstrates a useful method for machine learning with persistence homology inputs. If they are interested in further work with persistency homology, this could provide a useful beginning to their further experiments.

- The authors need to do a more thorough lit review as there are various papers related to PH and chemical applications. For example e.g. (Queen et al. 2024- npj Comp Mat.), (Jones et al., 2023, ACS) and references therein.

Response:

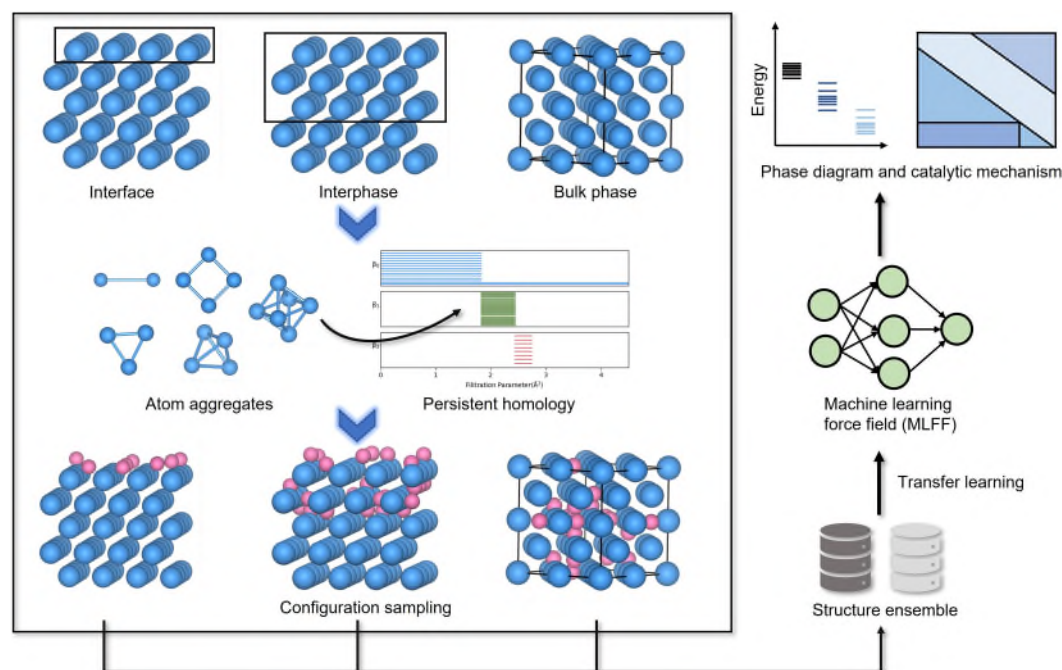
Thank you to the reviewer for recommending the articles. We have carefully read Reinauer et al.'s paper and believe that their work is highly innovative. Our subsequent research will attempt to apply the Persformer to the field of materials science. We have cited this literature as ref 39. Regarding the two articles by Queen et al. (2024 - npj Computational Materials) and Jones et al. (2023, ACS), we were unable to locate the original texts based on the information provided. We kindly request the reviewer to provide further information, such as the DOI or complete journal page details, to help us improve our literature review.

Minor comments:

- The example barcode in Figure 1 is not plausible: H_1 and H_2 generators shouldn't be "born" at $\lambda=0$, but at a variety of greater radii. This could present confusion for readers unfamiliar with persistence homology.

Response:

We thank the reviewer's suggestions. We have revised the schematic diagram in Figure 1 as follows.



- L147: "catch" is presumably supposed to read "capture"

Response:

We have revised the "catch" to "capture".

- This is particularly minor, but I think the phrasing of the sentence with begins on L151, while perfectly accurate, could cause confusion. The first half of the sentence puts a parallel between "components, loops, and cavities" with " β_0 , β_1 , and β_2 ", and the second half notes that barcodes record "birth, death, and persistence". It's easy to slip up and read this as " β_0 , β_1 , and β_2 " corresponding to "birth, death, and persistence". Separating this into two sentences would avoid the possibility of confusion.

Response:

We thank the reviewer's suggestions. We have revised the phrasing of the sentence as follows:

(page 6 in manuscript)

For a given structure aggregate, the number of independent components, loops and cavities are topological invariants and they are referred to as β_0 , β_1 , and β_2 , respectively. The “birth,” “death,” and “persistence” of β_0 , β_1 , and β_2 are recorded in the persistent barcodes.

Overall:

I view the general methodology as sound. However, I do believe additional evidence is needed to validate the claims the authors make. Some comparison to the quality of previous work is necessary to validate these claims. The significance of the paper's proposed method is contingent on improvements to current state-of-the-art sampling methods.

Response:

We appreciate the reviewer's acknowledgment of our work. We have conducted a detailed comparison of the results obtained using the sampling method developed in this paper with those from previous classic sampling methods in the Pd and Pt systems. This comparison demonstrates the superiority of our method. We hope that our new results will satisfy the reviewer.

Reviewer #2 (Remarks on code availability):

Comments on the code:

The available code is helpful, and Jupyter notebooks ensure that the process is easily understandable. However, the code provided is incomplete for the replication of experiments demonstrated in the paper. Only the novel sampling procedure is described, none of the post-processing MLFFs or other systems are shown.

Response:

We have provided more details related to the code. Besides the sampling process, we further offer the codes for training machine learning force fields and the code for optimizing the sampled structures using these force fields. Subsequent analyses, such as the Pourbaix diagram analysis and free energy path analysis, are relatively routine

processes that can be completed without specialized code. See <https://github.com/JFLigroup/PH-SA>

Reviewer #4 (Remarks to the Author):

This is a very interesting application of persistent homology in combination with a configurational exploration that is expedited through machine learning force fields. I agree that PH could be used to uniquely identify configurations that are relevant to catalytic activation and that this could be used in conjunction with DFT and MLFF for the purposes of this paper. My largest concern has to do with the clarity of the methodology such that other may easily understand and reproduce the approach.

Response:

We would like to express our sincere gratitude to the reviewer for the affirmation of our work and the constructive suggestions provided. We have carefully revised the manuscript in accordance with the requirements of the reviewer.

My questions arise primarily around the persistent homology. There is quite a bit of information in the method section that I don't believe used within the approach(e.g., homology groups, the cycle and boundary groups) and it makes the methodology more complicated than it needs to be. It is a bit unclear what information is exactly employed and how.

Response:

We sincerely appreciate the reviewer's feedback and understand your concerns regarding whether the homological concepts introduced in the Methods section (such as homology groups, cycle groups, and boundary groups) are genuinely utilized in our approach. Here, we provide further clarification:

The inclusion of these homological concepts in the Methods section was intended to offer a theoretical framework for readers without a background in topology. It serves to mathematically explain why the barcodes generated by persistent homology can reveal the topological features of atomic structures. We acknowledge that these concepts are not explicitly manifested in the computational implementation, as we primarily relied on standard codes (e.g., the GUDHI

library(https://doi.org/10.1007/978-3-662-44199-2_28) to compute persistent homology.

However, this does not imply that these concepts are absent from the underlying calculations. The GUDHI library's implementation of persistent homology fundamentally relies on these homological principles. It processes chain groups derived from simplicial complexes, and by constructing and reducing boundary matrices, it implicitly distinguishes between cycle groups and boundary groups to identify homology classes and their persistence. In other words, concepts such as cycle groups and boundary groups form the indispensable mathematical foundation of GUDHI's calculations, even though users are not required to directly manipulate or perceive these structures. What is ultimately presented to users are higher-level outputs, such as barcode diagrams and persistent Betti numbers, rather than the algebraic groups themselves.

The inclusion of this theoretical background may have made the Methods section unnecessarily lengthy. To improve readability, we have relocated this content to the Supplementary Information. In Methods section of the main text, we now focus on key practical details that directly contribute to the interpretation and results of our study. This ensures that readers can concentrate on the core steps and concepts most relevant to the outcomes of our work.

We have revised the page 18 of manuscript

The mathematical foundation of persistent homology is provided in the supplementary information. The implementation of the persistent homology relies on the open-source package GUDHI. Here, we focus on describing the practical operational process.

Additionally, it would seem to me that the PH data would need added information regarding the atom types if there is a mix of Pd, O, H or other adsorbates being considered. This is not described well.

Response:

We thank the reviewer's insightful question. Our entire configuration enumeration

process is divided into two steps. The first step involves obtaining the filtration parameter corresponding to the death point of a specific atomic aggregates using persistent homology. The second step determines the three-dimensional spatial coordinates of potential sites based on this filtration parameter. When the atomic cluster contains different elements, such as Pt and O, it is indeed inappropriate to treat Pt and O as indistinguishable points. If treated as such, the resulting potential sites might be too close to Pt and too far from O. In the second step, the formula we employ takes into account the handling of different elements within the atomic cluster. As shown in Equation R1 and R2

$$\min_x \sum_{i=1}^n w_i (\|p_i - x\| - d)^2 \quad (R1)$$

$$w_i = \frac{r_i}{\sum_{j=1}^n r_j} \quad (R2)$$

Here, p represents the three-dimensional coordinates of each atom within the atomic aggregate, recorded during enumeration. The w is the radius weight of different elements composing the atomic aggregate, r is the radius of each atom, i and j are the atom indices, x denotes the coordinates of interaction site to be determined, and d is the filtration parameter derived from the last death point in persistent homology. By assigning weights to different elements based on their atomic radii, we can achieve a reasonable determination of potential sites when different types of atoms are present in the material.

As a note, the determination of these sites mainly reflects the topological characteristics of the material structure itself. After identifying the sites, regardless of the type of atom/molecule we embed or adsorb, we only need to place the corresponding atom or molecule at the corresponding sites.

We have revised the page 19 of manuscript

In this process, we also considered that when the material structure contains different elements, the spatial coordinates of the potential sites should be adjusted based on the radii ratio of the different elements, in order to enhance their physical plausibility.

The 10 Ang cutoff for the PH seems a bit arbitrary for the cutoff of the search of

aggregate structures. Did the authors do any systematic searching of how large a cutoff was needed and how this impacted their algorithm?

Response:

Thank you for the reviewer's question. The selection of 10 Å is mainly based on general experience, namely that an active species usually only interacts directly with its nearest neighboring atoms. Choosing a radius of 5 Å, which corresponds to a diameter of 10 Å, typically ensures that most systems include the nearest neighboring atoms, since chemical bonds generally do not exceed 3.5 Å

Essentially, there is always a conflict between time and accuracy in any search algorithm. Choosing a small radius can speed up the process but may omit some sites, while selecting an excessively large radius can include as many sites as possible but will increase the search time. Here, we used the Pd slab and Pt cluster systems as examples to test the site search results and corresponding time costs for radii ranging from 3 Å to 6 Å (corresponding to the diameter of 6 Å to 12 Å). The results (Figure R7) indicate that 5 Å is indeed a preferable choice, as it allows for the identification of as many potential sites as possible while ensuring computational efficiency. Extending to 6 Å may yield a few additional sites, but the computation time increases by nearly 10 times. The comparisons were performed using 16 cores of an Intel Xeon Platinum 8336C central processing unit (CPU).

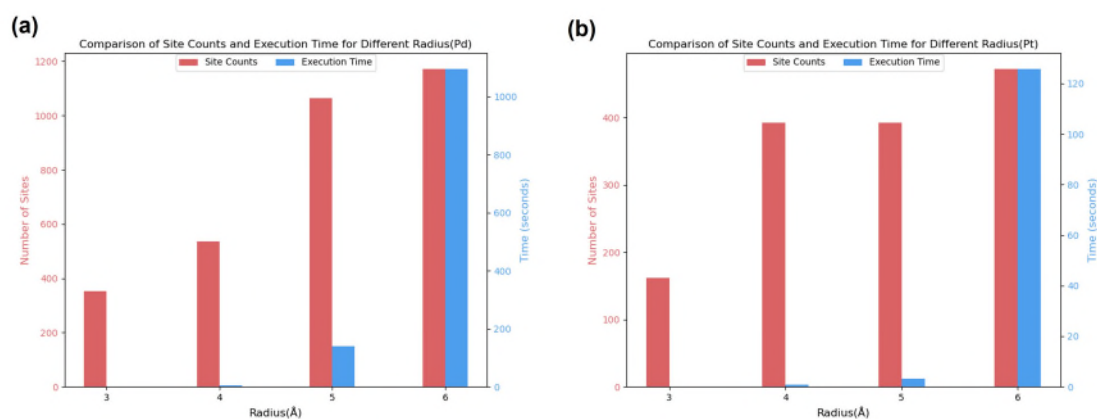


Figure R7. Determination of the number of potential interaction sites and corresponding enumeration time for (a) Pd-H and (b) Pt-O systems using different cutoff radii.

We have revised the manuscript and added Figure R7 as Figure S11

(Page 20 in manuscript)

The effect of threshold variation on site determination is tested in S11.

(Page 18 in SI)

Choosing a radius of 5 Å, which corresponds to a diameter of 10 Å, typically ensures that most systems include the nearest neighboring atoms, since chemical bonds generally do not exceed 3.5 Å

Essentially, there is always a conflict between time and accuracy in any search algorithm. Choosing a small radius can speed up the process but may omit some sites, while selecting an excessively large radius can include as many sites as possible but will increase the search time. Here, we used the Pd slab and Pt cluster systems as examples to test the site search results and corresponding time costs for radii ranging from 3 Å to 6 Å (corresponding to the diameter of 6 Å to 12 Å). The results (Figure S11) indicate that 5 Å is indeed a preferable choice, as it allows for the identification of as many potential sites as possible while ensuring computational efficiency. Extending to 6 Å may yield a few additional sites, but the computation time increases by nearly 10 times. The comparisons were performed using 16 cores of an Intel Xeon Platinum 8336C central processing unit (CPU).

It is unclear to me how the PH based aggregates are embedded within the slabs that are then employed with the force field enumeration?

Response:

We thank the reviewer for the insightful question. The procedure for embedding the PH-based atomic aggregates within slabs, and their subsequent use in force field enumeration, follows a three-step process.

Step 1: Atomic Aggregate Construction. We begin by enumerating atomic clusters within a slab. We enumerate the aggregates formed by each atom and the surrounding

atoms within a specified radial distance. This is accomplished using combinatorial enumeration. Moreover, the atomic indices of each atomic aggregate and their three-dimensional coordinates within the slab are recorded. To avoid redundancy, we remove duplicate atomic clusters with the same composition (atomic indices) that may arise from different central atoms.

Step 2: Persistent Homology is utilized to determine the critical filtration parameters for these atomic aggregates, and further identify potential active sites. Since we have recorded the atomic indices and three-dimensional coordinates of these atomic aggregates within the slab during enumeration, the active sites we identify can be directly correlated to those in the slab.

Step 3: Once the three-dimensional coordinates of these potential sites are determined, we proceed to enumerate adsorption and intercalation configurations. These configurations are then evaluated using machine learning-based force field calculations to assess the stability and reactivity of these sites.

In the MPLFF - the AIMD simulations for 2 or 10 ps seems very short. What and how are the AIMD simulations utilized in the MLFF? what is analyzed to understand if there is convergence of the energy and fluctuations in the bond distances, angles, etc? There seems to be a disconnect between the size of the PH aggregate clusters and the size of AIMD simulations. I don't quite understand why different sizes of the information embedding by PH and the AIMD is used. Why wouldn't use use barcodes of the full size of the system?

Response:

We thank the reviewer's insightful opinions. We will address the different aspects of the inquiry:

1. The utilization and coverage of AIMD simulation

The AIMD simulations were conducted to generate high-quality DFT reference data for training the machine learning force field (MLFF). In this context,

it is sufficient to ensure that the energy and forces at each ionic step meet appropriate convergence criteria to provide structural-energy/force pairs for MLFF training, rather than to achieve full energy convergence over extended timescales. Specifically, the structure generated at each AIMD ionic step is characterized and fed into the neural network to fit the electronic energy and forces of that configuration, thereby modeling the interatomic interaction potential. More details about the training of MLFF can be found in the response to DPA.

While traditional AIMD studies emphasize long-term energy convergence, MLFF training prioritizes the ability to accurately predict energy and forces relative to DFT data. Short AIMD simulations are widely accepted for MLFF training because they efficiently sample a sufficient variety of structural-energy/force pairs to improve the MLFF performance even if the individual AIMD simulations are not fully relaxed. Recent studies have demonstrated the effectiveness of such approaches (10 steps AIMD for initial DFT data, *Adv. Funct. Mater.* 2023, 33, 2303936; 10ps AIMD for initial DFT data, *Angew. Chem. Int. Ed.* 2023, 62, e202304205; 10ps AIMD for initial DFT data, *Angew. Chem. Int. Ed.* 2024, e202411849, ~1000 structure generated from AIMD, *J. Chem. Phys.* 155, 224701 (2021).) Based on this, we conclude that AIMD simulations with time lengths of 2 ps or 10 ps are acceptable as initial data for training our MLFF model regardless of whether they have long-term converged or not, which is of lesser importance in the training of MLFF.

After using the AIMD data to fine-tuning of DPA, the RMSE of energy is 0.008 eV/atom, and the RMSE of force is 0.21 eV/Å for our Pt-O system. We believe this precision is sufficient for subsequent structural optimizations, and it is consistent with the accuracy levels reported in previous work (*ACS Catal.* 2022, 12, 14812–14824).

Notably, structural optimization can also be used to generate data for training the MLFF, as demonstrated in the Pd-H system. Since the convergence criteria for structural optimization are well-defined and straightforward, we ensured that each structure was fully converged. The different approaches we employed to generate

datasets for various systems were primarily intended to demonstrate the flexibility of the machine learning force field training framework, emphasizing its capability to address diverse requirements.

2. Discrepancy between PH aggregate clusters and AIMD simulations

This is primarily because PH and AIMD serve different roles within the overall framework. PH is used to identify potential active sites where species might interact. At this stage, we decompose the slab or Pt system into a series of smaller clusters to pinpoint these sites, starting with two-atom clusters, which ensures a systematic enumeration of potential sites. We will elaborate on this point in the next question.

On the other hand, AIMD, as discussed earlier, is primarily used to generate the initial data required for training the MLFF. There is no strict definition regarding the size of the clusters that should be selected for this process. Generally, larger clusters may be more advantageous for accurately fitting the Pt-O interactions, as they encompass richer chemical environments. However, this also increases the computational burden of the AIMD simulations. Consequently, we chose Pt₁₃ and Pt₅₅ as the representative clusters for our AIMD simulations. This choice was based on the fact that 13-atom and 55-atom clusters are typical cluster sizes, and the 55-atom cluster also serves as the model for downstream tasks, such as phase diagram analysis.

3. Why not use full-size barcodes for the entire system

We decompose the atomic structure into smaller clusters before performing persistent homology to ensure that we capture as many potential active sites as possible. If we were to apply persistent homology directly to the entire structure, many voids or loops might be obscured by interactions with other atoms, making it difficult to detect them in the barcode of the whole system. By focusing on smaller clusters, we can more precisely identify these features, as the barcode of each smaller aggregate reflects the topological characteristics of that subset of atoms. This approach allows for a more systematic and thorough exploration of the potential active sites in the system. Here, we take a simple octahedron structure as an example (Figure R8). If we perform persistent homology on the entire

octahedron, we can only identify 12 bridge sites and 1 central site. However, by decomposing it into combinations of 2 atoms, 3 atoms, 4 atoms, 5 atoms and 6 atoms aggregates, we can additionally identify 8 more sites on the center of each face of octahedron.

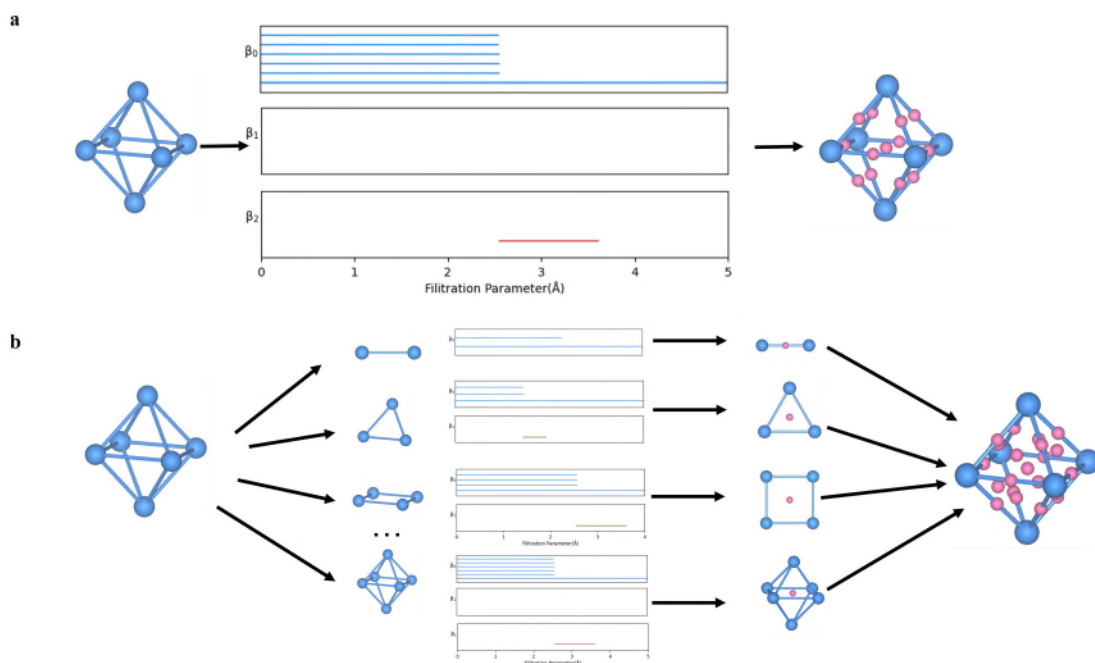


Figure R8. (a) Performing persistent homology on the entire octahedron allows the determination of potential spatial sites based on the death points of each Betti number. (b) Alternatively, decomposing the octahedron into atomic aggregates of varying sizes and identifying potential spatial sites using only the last death point of each aggregate's highest-dimensional Betti number represents a bottom-up approach. This latter method enables the identification of a greater number of potential sites.

We have revised the manuscript and SI. We have added the Figure R8 as Figure S2.

(Page 8 in manuscript)

Instead of performing persistent homology directly on the entire structure, we adopt this bottom-up approach to avoid interactions within the whole system that might obscure topological features in the barcode (Figure S2).

(Page 22 in manuscript)

In general, even relatively short AIMD simulation durations can provide an initial training dataset for MLFF, as we primarily focus on the structural-energy/force pairs at DFT accuracy obtained from each ionic step, which serve as input data for training the MLFF model.

(Page 7 in SI)

We decompose the atomic structure into smaller clusters before performing persistent homology to ensure that we capture as many potential active sites as possible. If we were to apply persistent homology directly to the entire structure, many voids or loops might be obscured by interactions with other atoms, making it difficult to detect them in the barcode of the whole system. By focusing on smaller clusters, we can more precisely identify these features, as the barcode of each smaller aggregate reflects the topological characteristics of that subset of atoms. This approach allows for a more systematic and thorough exploration of the potential active sites in the system. Here, we take a simple octahedron structure as an example (Figure S2). If we perform persistent homology on the entire octahedron, we can only identify 12 bridge sites and 1 central site. However, by decomposing it into combinations of 2 atoms, 3 atoms, 4 atoms, 5 atoms and 6 atoms aggregates, we can additionally identify 8 more sites on the center of each face of octahedron.

The DPA and associated models are not well described - I really don't understand this well at all.

Response:

We thank the reviewer for their insightful question. Machine learning-assisted modeling of inter-atomic potential energy surfaces, namely machine learning force fields (MLFF), is transforming the field of catalyst simulation (Chemical Engineering Journal 494 (2024) 152757). The purpose of machine learning force fields (MLFFs) is to model the interatomic interaction potential using machine learning methods, which typically require a substantial amount of DFT data for training. With the emergence of high-quality DFT datasets in material science, researchers have developed pre-trained

MLFF models that learn interatomic interactions for a wide range of elements based on extensive DFT data. Those universal pre-trained models provide a start baseline for specific material systems (e.g., Pd-H and Pt-O in our work), which can be further optimized by “fine-tuning” technique to improve prediction accuracy. This fine-tuning process to construct MLFF based on pre-train model can effectively reduce the amount of DFT calculations required compared to training an MLFF from scratch. DPA, a Deep Potential model with a gated attention mechanism, is a representative example of pre-trained models. Its training dataset is based on the OC20 database, which includes 56 elements and millions of atomic configurations based on DFT. (npj Comput Mater 10, 94 (2024)). A variety of slab and adsorbed molecule configurations in this database cover the catalytic field to a large extent.

Technically, the fine-tuning process for MLFF based on pre-trained models starts with the weights of the pre-trained model, rather than the random initialization weights of the from scratch training. Then, partial of the configurations sampling by our PH methods are used as the training set based on DFT calculations (AIMD or structure optimization), adjusting the model weights by means of supervised learning. These DFT data is first featurized into a format compatible with the pre-trained model, and then used to calculate a loss function, typically based on discrepancies in predicted energies and forces. Through backpropagation, the model parameters are updated to minimize this loss, similar to traditional neural network training. A full weight fine-tuning approach without freezing any components is implemented to maintain the convergence and accuracy of the model during training.

To construct the MLFF for Pd-H and Pt-O systems, we generated DFT-based structure-energy datasets through structural optimization and molecular dynamics simulations. These datasets were then used to fine-tune the DPA pre-trained model, resulting in a high-precision force field tailored to the unique properties of Pd-H and Pt-O. The fine-tuned MLFF was subsequently employed for configuration optimization.

The hyperparameter settings used in the fine-tuning process are detailed below. For both Pd and Pt systems, the same hyperparameters were used, except for the number of training steps. This consistent approach ensured that the model was properly fine-tuned

for both systems, balancing computational efficiency with high precision.

Table S1. The hyperparameters of fine-tuning of DPA towards Pd-H and Pt-O system.

Hyperparameters	Value
Type_embedding	"trainable": true
Descriptor	"trainable": true
Fitting_net	"trainable": true
Start_lr	0.001
Decay_steps	50
Stop_lr	3.51e-08

We have revised the page 21-22 of manuscript and added the Table R1 as Table S1. The related code was uploaded on GitHub.

DPA, a deep potential model with a gated attention mechanism, is a pre-trained machine learning force field based on the OC20 database, covering 56 elements and millions of atomic configurations. Fine-tuning DPA for specific tasks enables high-accuracy force fields more efficiently than training MLFF from scratch. The fine-tuning process for MLFFs leverages pre-trained weights and adapts them to target systems using high-quality DFT data. After featurizing the DFT data into a compatible format, a loss function based on discrepancies in predicted energies and forces is minimized through backpropagation. We employed global fine-tuning, updating all model parameters without freezing any components. The hyperparameter settings used in the fine-tuning process are detailed below. For both Pd-H and Pt-O systems, the same hyperparameters were used (Table S1), except for the number of training steps. Next, we present the DFT datasets and the results of force field fine-tuning for the Pd-H and Pt-O systems, respectively.

At a very fundamental level, the methodology is not written in a very clear and logical framework such that it would be easy to reproduce. It would be nice if it were simplified

Response:

We greatly thank the reviewer's suggestions. To enhance the readability of the article, we have reorganized the Methods section. For the persistent homology part, we

have moved the theoretical knowledge to the Supplementary Information and focused on describing the practical operational aspects. We have rewritten the description of how to determine the interaction sites. We have also polished the detailed description of the machine learning training process. To further enhance reproducibility, we have uploaded additional code related to the aforementioned sections.

Reviewer #3

I would like to thank the authors for their attention to our comments and detailed replies. The paper is much stronger than its initial version, and the updates the authors have made to the organization have improved its readability and flow. The comparison with the basin hopping algorithm is particularly helpful, making clear the improvement their methodology offers, including both higher effectiveness than previous global methods and the ability to more effectively utilize modern hardware for efficient computation.

Given how often the material surface is the primary point of interest, I believe the paper would be made stronger by also offering comparisons to the graph based methodology, so that users could easily determine which algorithm is best suited to their needs when subsurface and bulk interactions are not relevant to the system. However, the current version of the work is sufficiently complete for publication.

Response:

We appreciate the reviewer's recognition and have made comparisons with graph-based methods. Through our research, we found several graph-theory-based methods can determine surface adsorption sites, including the SurfGraph algorithm (npj Computational Materials, 2020, 6(1): 79., Nature Communications (2022) 13:5788), Catkit (J. Phys. Chem. A 2019, 123, 2281–2285), and ACAT (npj Computational Materials, 2023, 9(1): 139.). These methods essentially convert surface atoms into graphs and then identify the possible adsorption sites. After experimentation, we found that ACAT can handle both slab and cluster models, and we primarily compared this method. In the two systems designed for this study, namely the Pd (100) slab and Pt₅₅ clusters, the number of surface sites identified by both the PH-SA and ACAT methods is exactly the same: 64 sites for the Pd system and 242 sites for the Pt system (Figure R1). The time spent is around 1 second on a regular personal computer for both methods. This demonstrates that our PH-SA method is highly robust and comparable to previous methods when only enumerating surface structures. We would like to emphasize that graph-theory-based methods, due to their focus solely on atomic connectivity, lose three-dimensional information and therefore cannot consider subsurface and bulk

structure sampling, which is the strengths that the PH-SA method can achieve.

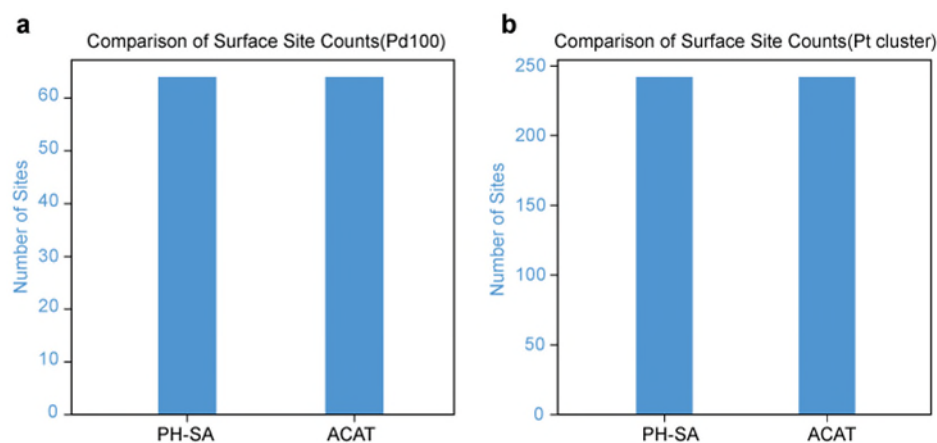


Figure R1. The surface adsorption sites determination by different methods. The comparison between persistent homology-based sampling algorithm (PH-SA) and alloy catalysis automated toolkit (ACAT) for surface adsorption sites determination in (a) the Pd (100) and (b) Pt₅₅ cluster.

We have made revision in manuscript and added Figure R1 as Figure S11.

(Page4 in manuscript)

The graph theory-based algorithms, which can enumerate configurations for a specific adsorbate coverage on the surface with graph isomorphism algorithms, even on an asymmetric one. Nevertheless, these methods can only study the adsorbate coverage effect on the surface because the graph representation is insensitive to three-dimensional information, making it unable to consider subsurface and bulk structure sampling.

(Page 19 in manuscript)

In addition, when considering only the surface adsorption sites, the number of sites identified by the PH-SA method is consistent with that determined by previous graph-theory-based methods in both systems (Figure S11), thereby demonstrating the robustness of the PH-SA method.

(Page 18 in SI)

The comparison with graph-based method for surface adsorption sites determination

In the two systems designed for this study, namely the Pd (100) slab and Pt₅₅ clusters, the number of surface sites identified by both the PH-SA and graph-theory-based alloy catalysis automated toolkit (ACAT) methods is exactly the same: 64 sites for the Pd system and 242 sites for the Pt system (Figure S11). The time spent is around 1 second on a regular personal computer for both methods. This demonstrates that our PH-SA method is highly robust and comparable to previous graph-theory-based methods when only enumerating surface structures. We would like to emphasize that graph-theory-based methods, due to their focus solely on atomic connectivity, lose three-dimensional information and therefore cannot consider subsurface and bulk structure sampling, which is the strengths that the PH-SA method can achieve.

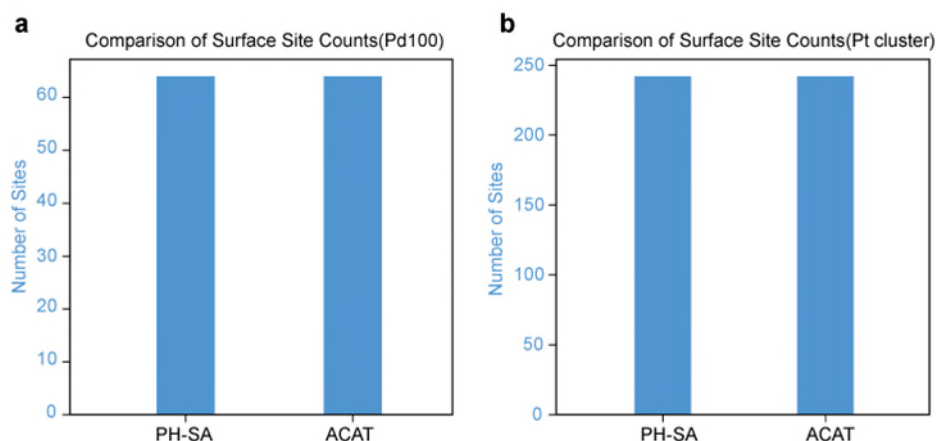


Figure S11. The surface adsorption sites determination by different methods. The comparison between persistent homology-based sampling algorithm (PH-SA) and alloy catalysis automated toolkit (ACAT) for surface adsorption sites determination in (a) the Pd (100) and (b) Pt₅₅ cluster.