Research Square

Preprints are preliminary reports that have not undergone peer review. They should not be considered conclusive, used to inform clinical practice, or referenced by the media as validated information.

TASTET: a Python package for geometry-driven selection of representative atomistic structures from large configurational spaces

Alejandro Cañete-Arché
Trinity College Dublin
Ana McGinley
Trinity College Dublin
Pablo Castro-Latorre
Basque Research and Technology Alliance (BRTA)
Graeme W. Watson
Trinity College Dublin
Max García-Melchor
maxgarcia@cicenergigune.com

Basque Research and Technology Alliance (BRTA)

Article

Keywords: Representative structure selection, Configurational sampling, Atomistic representations, Density functional theory, Conformational landscapes, Computational catalysis

Posted Date: October 7th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11013614/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Additional Declarations: No competing interests reported.

TASTET: a Python package for geometry-driven selection of representative atomistic structures from large configurational spaces

Alejandro Cañete-Arché[a], Ana McGinley[a], Pablo Castro-Latorre[b], Graeme W. Watson[a], and Max García-Melchor [a,b,c,*]

[a]*School of Chemistry, Trinity College Dublin, College Green, D02 PN40, Dublin, Ireland.*

[b]*Center for Cooperative Research on Alternative Energy (CIC energiGUNE), Basque Research and Technology Alliance (BRTA), 01510 Vitoria-Gasteiz, Spain.*

[c]*IKERBASQUE, Basque Foundation for Science, 48009 Bilbao, Spain*

E-mail: maxgarcia@cicenergigune.com

Abstract

Modern atomistic workflows generate large ensembles of molecular and materials structures, yet only a fraction can be evaluated using first-principles methods. Existing down-selection strategies typically rely on estimated energies, implicitly optimizing for energetic ranking rather than representative coverage of configurational space. Here, we present a geometry-based framework for representative structure selection that maximizes configurational diversity prior to electronic-structure calculations. The framework maps local chemical environments into global measures of structural similarity, which space-filling algorithms use to select diverse subsets with uniform configurational coverage. We demonstrate its performance through two case studies spanning conformational and configurational ensembles. Selecting only 3% of a 949-member Rh-complex conformer ensemble recovered the principal energetic regions of the conformational landscape, and selecting just 0.3% of 10,000 oxide-supported metal nanocluster structures likewise revealed the dominant structure–energy relationships. These results establish representative structure selection as a transferable geometry-driven strategy for maximizing information gained from demanding first-principles calculations.

Keywords: Representative structure selection · Configurational sampling · Atomistic representations · Density functional theory · Conformational landscapes · Computational catalysis

Introduction

Advances in atomistic modeling have transformed computational molecular and materials science from the study of individual structures to the exploration of increasingly large chemical and configurational spaces. Electronic-structure methods are now routinely embedded within workflows for high-throughput screening, materials discovery, and molecular design, in which thousands of candidate structures may be generated before any first-principles calculations are performed[1–4]. Instead of relying on a single structural model, these workflows typically produce ensembles corresponding to different conformations, atomic arrangements, or adsorption configurations[5–8]. As methods for generating such ensembles continue to improve, prioritizing candidate structures for subsequent first-principles calculations at the target level of theory is becoming an increasing focus.

Although many candidate structures may be generated, only a subset contains genuinely distinct local atomic environments, whereas many others differ only through minor geometric variations. Evaluating every candidate using density functional theory (DFT) or higher-level electronic-structure methods is therefore rarely feasible. Conversely, reducing an ensemble without explicitly preserving configurational diversity risks overlooking important regions of structural space while expending valuable computational resources on redundant configurations. The central challenge is therefore not simply to reduce the size of a configurational ensemble, but to identify a subset that faithfully captures its structural diversity within a fixed computational budget[9].

This need for principled structure selection extends across a broad range of atomistic applications. The reaction energetics, mechanisms, and regio- or enantioselectivities of flexible molecular systems, for instance, may depend strongly on the conformers selected for electronic-structure calculations. Selecting a structurally diverse conformational ensemble is therefore critical for reliably describing these systems. Modern software packages, including CREST[10], RDKit[11,12], GOAT[13], and OpenBabel[14], efficiently generate large conformer pools, which are commonly reduced using energy-based[15] or root-mean-square deviation (RMSD)-based[16] filters before DFT refinement. Although computationally convenient, these approaches do not necessarily preserve the geometric diversity of the original conformational ensemble. In particular, low-cost energetic estimates derived from force fields or semiempirical methods may deviate substantially from DFT, especially for electronically complex systems such as transition metal complexes[16–19]. Consequently, structurally distinct conformers may be discarded prematurely, whereas multiple structures occupying similar regions of conformational space, and therefore providing little additional structural information, are retained. Such redundancy may be acceptable when the objective is to characterize a low-energy region of conformational space, but it is less desirable when seeking to understand how molecular geometry influences chemical properties, for which broad coverage of conformational space is often more informative than concentrating exclusively on the lowest-energy structures[20,21]. Clustering strategies incorporated into conformer-generation and structure-selection workflows reflect this need[10,16], although their application can depend on the underlying conformer-generation procedure and require prior specification of the structural features used to describe conformational diversity. Extending geometry-based structure selection into a general and accessible framework that does not require predefined system-

specific structural features, is independent of conformer generation, and is applicable across chemically diverse systems therefore remains an attractive strategy.

A closely related problem arises in materials modeling. Structural searches guided by chemical intuition, machine-learning interatomic potentials (MLIPs)[22–26], active-learning[27,28], and global-optimization[8,29–33] workflows have dramatically expanded the configurational spaces that can be explored prior to first-principles calculations. Although these approaches substantially accelerate structure generation, they often require additional model development, specialized expertise, and computational effort while ultimately producing large candidate pools that still require down-selection before DFT calculations[34].

In many of these workflows, candidate structures are selected based on energy predictions from surrogate models[35,36]. Structures whose predicted energies fall within a user-defined energy window relative to the surrogate-predicted global minimum, are typically prioritized for subsequent electronic-structure calculations[37,38]. Such thresholds are usually selected empirically and vary considerably across theoretical studies, often reflecting the available computational budget rather than the structural diversity of the configurational ensemble. More fundamentally, selecting structures according to energy and selecting structures to maximize configurational diversity are distinct objectives. Even if surrogate energetics perfectly reproduced DFT, energy-based filtering would still preferentially retain many geometrically similar low-energy structures while undersampling other regions of configurational space. In practice, the finite accuracy of surrogate models may further amplify this limitation by introducing additional uncertainty into the selection process. Consequently, the quality and representativeness of the final selected set for first-principles evaluation depend not only on the configurational search itself, but also on the accuracy of the surrogate model and the criteria used to select candidate structures.

These observations motivate a strategy in which structure selection is based directly on geometry rather than approximate energetics. Nevertheless, random subsampling and manual inspection remain common in molecular and materials modeling because they are simple, inexpensive, and require no specialized infrastructure. For geometry-based selection to provide a practical alternative to these simple approaches, however, its methodology must be made similarly accessible. An effective implementation should preserve configurational diversity systematically while operating within a predefined DFT budget, remaining independent of surrogate-model accuracy, and interfacing naturally with workflows used to generate candidate structures.

Recent advances in atomistic machine learning provide the conceptual foundations for such an approach. Over the past decade, the field has developed various methods that transform local chemical environments into global representations of atomistic structures through descriptors and similarity kernels[39–42]. Combined with space-filling sampling algorithms, these representations enable systematic exploration of configurational spaces while preserving their structural diversity. Efforts have begun to extend these tools beyond their traditional role in machine learning model development, including applications to automated structural mapping, dataset sparsification, and the identification of representative atomic environments[43–45]. Notably, scikit-matter has sought to lower the barrier to such broader use by providing domain-agnostic implementations of methods originating in computational chemistry and materials

science through an interface designed for integration into diverse workflows[46]. Inspired by these efforts, we seek to further strengthen this bridge by focusing specifically on geometry-driven representative structure selection from atomistic ensembles, formulated independently of how the candidate structures are generated.

Building on these foundations, we present a geometry-based framework for selecting structurally diverse subsets from large atomistic ensembles. Structures are first characterized using local descriptors of their chemical environments, which are subsequently aggregated into global measures of structural similarity and sampled to select subsets that maximize configurational coverage while minimizing redundancy. Because the workflow relies exclusively on geometric information, it is independent of approximate energetic models and is equally applicable to molecular and materials systems. We implement this methodology in the open-source Python package TASTET (Tool for Atomistic Structure Selection Through Efficient Triage) and demonstrate its use through two representative case studies involving conformational sampling of a transition-metal complex and configurational sampling of an oxide-supported metal nanocluster. Together, these examples demonstrate that the geometry-based framework is a transferable approach to representative structure selection across diverse molecular and materials systems, with TASTET offering a user-friendly implementation intended to facilitate its broader adoption by the chemical and materials science communities.

Methods

Workflow outline

Fig. 1 summarizes the overall workflow of the representative structure selection strategy developed in this work. Starting from a configurational space generated by an atomistic sampling workflow, the resulting candidate structures are supplied as one or more input files and consolidated into a single Atomic Simulation Environment (ASE) database[47]. The workflow then proceeds through three main stages: (i) encoding selected local chemical environments using the Smooth Overlap of Atomic Positions (SOAP) descriptor[48], as implemented in DScribe[49]; (ii) constructing global similarity matrices (kernels) that quantify structural similarity across the configurational ensemble; and (iii) selecting structurally diverse subsets using space-filling sampling algorithms. Where appropriate, the descriptor and kernel hyperparameters governing stages (i) and (ii) can be optimized through automated grid searches.

The selected subset is subsequently propagated to computationally demanding first-principles calculations, such as DFT. In practice, the number of selected structures is determined by the computational budget available for these calculations. The individual components of the workflow are described in detail below.

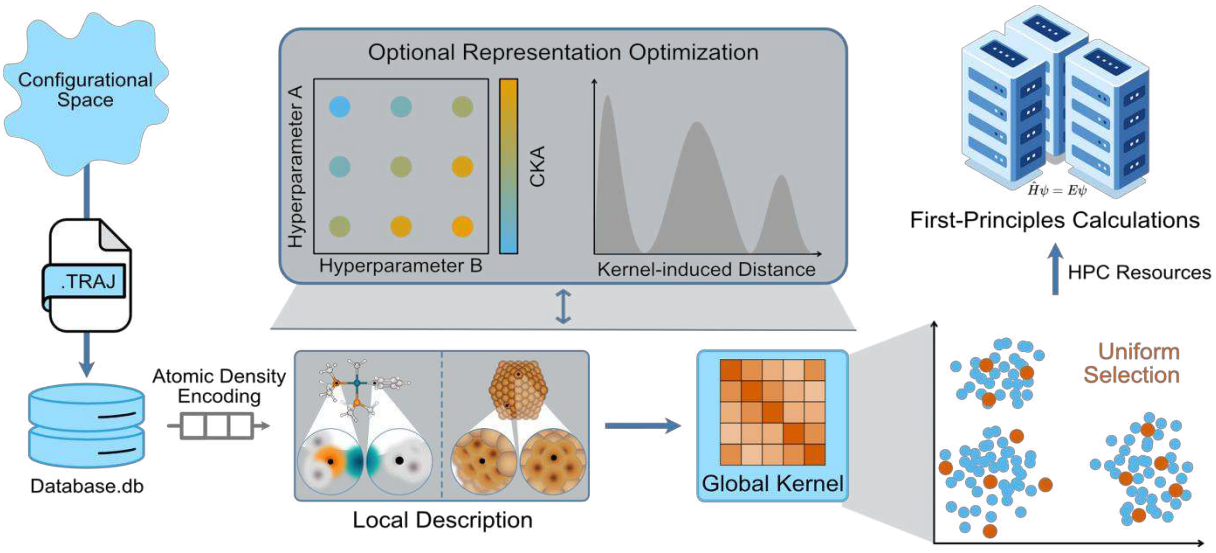

Fig. 1. Schematic overview of the geometry-based representative structure selection workflow. Candidate structures generated by a given atomistic sampling workflow are first consolidated into a centralized ASE database. Selected local chemical environments are then encoded using atomic-density-based descriptors, from which structure-level similarity kernels are constructed. These kernels define the geometry of the configurational space and provide the basis for representative subset selection using space-filling sampling algorithms. When required, descriptor and kernel hyperparameters are optimized through grid searches using criteria such as kernel-induced distance distributions and centered kernel alignment (CKA). The resulting subset is subsequently propagated to first-principles calculations on high-performance computing (HPC) resources.

Representation of local chemical environments using SOAP

The first step of the workflow is to construct a numerical representation of the local chemical environments that can be compared across different atomistic structures. In this work, we employ the SOAP descriptor[48], which provides a representation of atomic environments that is invariant to translations, rotations, and permutations of atoms of the same chemical species. Throughout this section, vector quantities are denoted by bold symbols, whereas scalar quantities are written in italics.

For each structure, the local environment is described through the neighboring atomic density within a finite cutoff radius, r_{cut}, centered on a selected atom. To distinguish between chemical species, an independent atomic density is constructed for each element α. For an environment centered on atom i, the atomic density of species α is represented as

$$\rho_i^{\alpha}(\mathbf{r}) = \sum_{j\in\alpha} \exp\left(-\frac{|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma^2}\right) \tag{1}$$

where $\mathbf{r}_{ij}$ is the position vector of neighboring atom j relative to the central atom i, and σ is the spatial width of the Gaussian functions.

The local density is then expanded in a basis of orthogonal radial functions, $R_n(r)$, which resolve variations in atomic density with distance from the central atom, and spherical harmonics, $Y_{lm}(\hat{\mathbf{r}})$, which resolve its angular dependence, up to a maximum radial and angular indices n_{max} and l_{max}, respectively.

$$\rho_i^{\alpha}(\mathbf{r}) = \sum_{n=1}^{n_{max}} \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} c_{nlm}^{\alpha i} R_n(r) Y_{lm}(\hat{\mathbf{r}}) \tag{2}$$

where $c_{nlm}^{i\alpha}$ are the corresponding expansion coefficients.

To construct a rotationally invariant descriptor, the expansion coefficients are combined into power-spectrum components,

$$p_{nn'l}^{\alpha\beta i} \propto \sum_{m=-l}^{l} \left(c_{nlm}^{\alpha i}\right)^* c_{n'lm}^{\beta i} \tag{3}$$

where * denotes complex conjugation. The complete SOAP descriptor for atom i, denoted by $\mathbf{p}_i$, is obtained by concatenating these components over all unique species pairs (α, β), all combinations of radial basis functions $(n, n' \leq n_{max})$, and all angular channels $(l \leq l_{max})$.

The descriptor construction may be restricted to a chosen subset of chemical elements, such that the representation is sensitive only to the species of interest and insensitive to all others. Similarly, descriptors need not be centered on every atom in the structure. Depending on the application, descriptors may instead be centered on selected elements, functional groups, ligands, adsorption sites, or other local motifs. This makes it possible to tailor the representation to reactive centers, or chemically distinct regions of configurational space that are most relevant to the problem under investigation.

The resulting collection of local descriptors provides the basis for comparing atomistic structures at the global level. In the following section, these local representations are aggregated into global similarity kernels that quantify structural similarity across the configurational ensemble.

Quantifying similarity between whole atomistic structures

Once the local chemical environments in each structure have been characterized, whole structures can be compared by assessing the similarity between the collections of local environments they contain. For a pair of structures, the SOAP vector associated with each local environment in one structure is compared with those in the other using a kernel function, producing a matrix of pairwise environment similarities. These similarities are then aggregated into a single measure of similarity between the two structures.

Given two structures, A and B, let $\mathbf{p}_i$ and $\mathbf{p}_j$ denote the SOAP descriptors of local environments centered on SOAP atoms $i \in A$ and $j \in B$, respectively. Pairwise similarities between local environments are collected in a local similarity matrix, whose elements are given by

$$C_{ij}(A, B) = k\left(\mathbf{p}_i, \mathbf{p}_j\right) \tag{4}$$

where k is a positive-semidefinite kernel that quantifies the similarity between two local descriptors. Thus, each element $C_{ij}(A, B)$ quantifies the similarity between the local environment centered on i in structure A and that centered on j in structure B. In the present

implementation, TASTET provides linear, polynomial, and radial basis function (RBF) kernels, although the framework is readily extensible to alternative kernel choices.

A structure-level similarity is subsequently obtained by aggregating the local similarity matrix $C_{ij}(A,B)$ into a single global kernel value, $K(A,B)$. In this work, we consider two well-established aggregation schemes to construct $K(A,B)$ from $C_{ij}(A,B)$: the average kernel and the regularized entropy match (REMatch) kernel[50]. The average kernel defines the similarity between two structures as the arithmetic mean of all pairwise similarities between their local environments, such that each entry of the local similarity matrix contributes equally to the final structure-level similarity, whereas the REMatch kernel constructs a soft one-to-one matching between the local environments through the regularization parameter $\lambda_{\mathrm{REMatch}}$, enabling more discriminative comparisons between structures with heterogeneous local environments. Because these kernels have been extensively described elsewhere[50], we use them here without further mathematical development and focus instead on their role within the representative structure selection workflow.

Combining multiple structural representations

Some atomistic systems contain chemically distinct motifs for which different structural representations may be appropriate. For example, one may wish to describe the coordination environment of a metal center at a different geometric resolution than its surrounding ligands, or represent different spatial regions of a material at distinct levels of geometric resolution[44]. To accommodate such situations, TASTET allows complementary structural information to be incorporated through multiple representation channels.

Each representation channel is defined independently through its own structural representation, including the selected chemical species, descriptor centers, SOAP hyperparameters, and the choice of local and global kernels. This modular formulation allows chemically distinct regions of the same structure to be represented independently before being combined into a unified measure of structural similarity. For a given pair of structures, each channel therefore produces an independent structure-level kernel, $K_m(A,B)$, where $m = 1, \dots, M$ indexes the individual channels and M is the total number of channels.

The channel-wise kernels can then be combined into a single global similarity measure in either a multiplicative or an additive manner, depending on how information from different chemically distinct regions is intended to contribute to the overall structural similarity. Specifically, the combined kernel may be constructed through the element-wise product,

$$\prod_m K_m(A,B) \tag{5}$$

or through their arithmetic mean,

$$\frac{1}{M}\sum_m K_m(A,B) \tag{6}$$

The product provides a restrictive measure of similarity: two structures attain a high combined similarity only when they are similar according to all channels, since a low similarity in any

individual channel suppresses the product. In other words, it acts as a geometric intersection, penalizing a mismatch in any individual channel. By contrast, the arithmetic mean allows similarities across channels to compensate for one another, such that the combined kernel reflects their average contribution to the overall structural similarity. This additive formulation is particularly well suited to multiple-kernel learning approaches, where target-property regressions determine the relative importance of each channel[51,52].

Hyperparameter selection

In many applications, it may be useful to tune the hyperparameters of the SOAP representation as well as the choice of local and global kernel types, the corresponding kernel hyperparameters, and the potential use of multi-channel constructions. To facilitate this, TASTET allows for grid searches over combinations of descriptor and kernel settings, either within individual channels or across multiple channel combinations. The appropriate choice of these parameters depends on whether target properties are available to guide the selection process.

When a target property, such as energy, is available for the structures stored in the database, hyperparameter selection can be carried out in a supervised manner by favoring kernel representations whose induced similarity is aligned with variation in that property. In this case, candidate representations are assessed through the centered kernel alignment (CKA)[53–56] between the structure kernel and a target-property kernel constructed from the property vector. CKA values range from 0 to 1, with larger values indicating stronger alignment between the two kernels, and a value of 1 corresponding to perfect alignment. The resulting CKA scores can be visualized as heatmaps over the hyperparameter search space, providing a visually clear supervised route to identify descriptor and kernel combinations for which the kernel representation of the dataset best reflects variation in the target property.

When no target properties are available, candidate representations are instead evaluated through the distribution of kernel-induced pairwise distances.

$$d(A,B) = \sqrt{2(1 - K(A,B))} \tag{7}$$

where A and B are two structures and $K(A,B)$ denotes the corresponding normalized structure-level kernel. This expression follows from the Euclidean distance between the corresponding unit-normalized feature vectors. Accordingly, $d(A,B) = 0$ corresponds to identical representations, $K(A,B) = 1$, whereas $d(A,B) = \sqrt{2}$ corresponds to zero kernel similarity, $K(A,B) = 0$. The distribution of these distances provides an interpretable visual diagnostic of the resolution induced by a representation without requiring target-property information. Distributions concentrated near zero indicate that most structures are represented as highly similar, which may reflect either insufficient discriminatory resolution or an intrinsically redundant configurational ensemble. Conversely, distributions concentrated near $\sqrt{2}$ indicate that most structures are represented as mutually dissimilar, reflecting either genuine structural diversity or an excessively discriminative representation. Broad and multimodal distance distributions can indicate meaningful structural heterogeneity, but are not intrinsically preferable, as they may also arise when small differences between near-degenerate local environments are strongly amplified by the kernel.

Although this provides a practical unsupervised strategy for constructing well-resolved configurational representations, it remains limited by the absence of a ground truth. Consequently, the distance distribution should be interpreted alongside the underlying descriptor-space separations and kernel scale, since apparent multimodality cannot by itself be assumed to reflect physically meaningful structural variation. Whenever target properties are available, introducing supervision is therefore desirable because it provides an independent validation that the selected representation captures physically relevant variation within the configurational ensemble[57].

Diverse selection

Once a kernel representation of the dataset has been established, the similarity between any pair of structures can be expressed as a distance in the corresponding kernel space. These distances provide the basis for selecting a structurally diverse subset. TASTET implements two space-filling sampling algorithms that have recently been shown to outperform alternative approaches for representative subset selection[58]: farthest-point sampling (FPS) and *k*-medoids.

FPS selects structures sequentially by maximizing the minimum Euclidean distance between each newly selected structure and those already selected. By contrast, the *k*-medoids algorithm identifies a representative subset that minimizes the average squared Euclidean distance from every structure in the configurational space to its nearest selected structure. Notably, random subsampling is not considered here because it does not ensure consistently broad coverage of the represented configurational space. FPS and *k*-medoids instead promote consistently diverse selection.

Dimensionality reduction for visualization

To facilitate the interpretation of the configurational representation and the selected subsets, the global kernel matrix is projected into two or three dimensions using kernel principal component analysis (kPCA)[43,59]. This projection is used exclusively for visualization and does not influence the underlying representative subset selection. We employ kPCA rather than neighbor-embedding methods such as t-SNE or UMAP, whose apparent low-dimensional structure can be highly sensitive to hyperparameter choices, such that different visual representations may be obtained from the same underlying data[60]. In contrast, the kPCA coordinates are obtained directly from the eigendecomposition of the global kernel, keeping the visualization directly tied to the similarity measure used for representative selection.

The intrinsic dimensionality of configurational spaces is often high, meaning that the leading principal components may capture only a limited fraction of the total variance. Consequently, the two- or three-dimensional projection cannot always faithfully reproduce the full geometry of the kernel-induced configurational representation and should therefore be interpreted with appropriate caution. Accordingly, kPCA is used solely as an interpretable visual summary, whereas the configurational representation and representative subset selection are performed entirely in the underlying high-dimensional kernel space.

Results and Discussion

Case study 1: Conformational landscape exploration

Flexible molecular and organometallic systems often exhibit highly complex conformational landscapes, in which relatively small geometric changes give rise to numerous conformers spanning a broad range of energies[61,62]. Capturing this diversity is essential because conformational effects can influence the reactivity and stability of reaction intermediates. At the same time, the large number of accessible conformers renders exhaustive exploration using DFT computationally prohibitive. As a first demonstration of the representative structure selection framework, we therefore consider a Rh complex bearing a sterically demanding (R)-BIPHEP-derived diphosphine ligand and a bound acrylate substrate (**Fig. 2a**). The conformational complexity of this system arises from multiple coupled degrees of freedom, including four rotatable methoxy substituents on the ligand biaryl unit, four aryl rings capable of rotating about the P–C bonds, a rotatable substrate C–C bond, and conformational adjustments of the ligand *tert*-butyl substituents[63]. Rather than attempting to enumerate this conformational landscape exhaustively with DFT, our objective is to select a geometrically representative subset that spans conformational space as broadly as possible, thereby enabling systematic exploration of the associated energetic landscape within a realistic DFT budget.

The initial conformational ensemble was generated using OpenBabel[14], yielding 949 candidate structures (see **Supplementary Materials** and **Fig. S1** for details on their generation). Because the conformers were generated from molecular connectivity rather than approximate electronic-structure-based optimization, the ensemble captures variations in molecular conformation while preserving the underlying bonding topology. Structural differences therefore arise mainly from torsional degrees of freedom rather than local bond stretching or compression. This geometric consistency is important for the present SOAP-based analysis because descriptor variations reflect the conformational degrees of freedom of interest rather than distortions introduced by preliminary geometry optimization.

Because the objective is to maximize the structural information gained within a fixed DFT budget, representative selection requires a kernel space that resolves conformational diversity well enough for space-filling sampling to distribute calculations effectively. We therefore screened combinations of descriptor and kernel hyperparameters through a grid search. At this stage, only the geometries of the conformational ensemble were available, with no target properties to guide the optimization, so the search relied on the unsupervised kernel-distance analysis described in the Methods section. This led to the selection of a two-channel representation that provided apparent conformational resolution without excessively amplifying very small differences between otherwise highly similar local environments. Full details of the grid search and comparison between candidate representations are provided in the **Supplementary Materials**.

The two channels describe complementary regions of the complex, reflecting the distinct roles of the Rh coordination environment and the surrounding ligand substituents in shaping conformational diversity. Both channels define structure-level similarity matrices over the same conformer ensemble, such that each corresponding matrix element quantifies the similarity of the same pair of structures according to either the core or peripheral representation. The two similarity measures were combined through their element-wise product, $K_{\mathrm{core}} \circ K_{\mathrm{periphery}}$,

yielding a high overall similarity only when a pair of conformers is similar according to both channels.

Within this representation, both channels employed a SOAP basis with $n_{max} = l_{max} = 8$, but differed in the chemical regions they described and the corresponding kernel constructions. In the core channel, the local environments centered on the Rh atom, the coordinating P atoms, and the substrate C and O atoms were first compared using an RBF local kernel ($r_{cut} = 3.0$ Å, $\sigma = 0.1$ Å). These pairwise local-environment similarities were then averaged to obtain the structure-level core kernel. In the periphery channel, local environments centered on the ligand carbon atoms bearing the methoxy substituents were likewise compared using an RBF local kernel ($r_{cut} = 2.0$ Å, $\sigma = 0.5$ Å) but were aggregated using the REMatch kernel to obtain the structure-level periphery kernel (**Fig. S2**). The resulting two-channel representation is visualized by the two-dimensional kPCA projection of the combined kernel, $K_{\text{core}} \circ K_{\text{periphery}}$, shown in **Fig. 2b**.

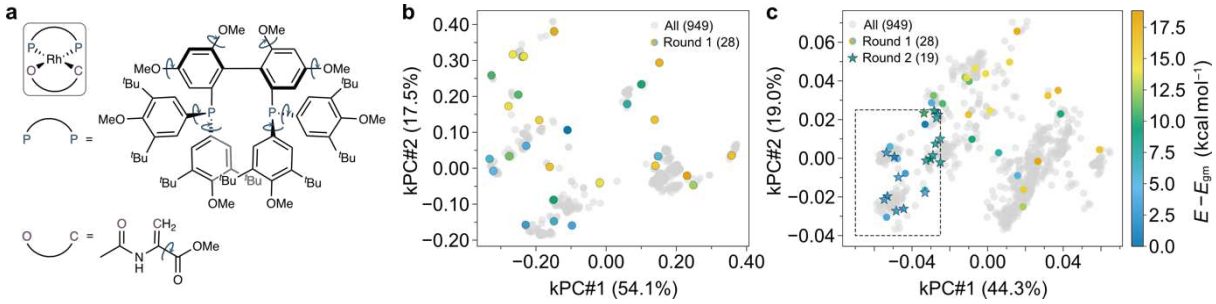

Fig. 2 | Representative structure selection for the Rh-complex conformational landscape. a Structure of the square-planar Rh complex comprising the (R)-BIPHEP-derived diphosphine ligand and the coordinated acrylate substrate. Blue curved arrows indicate the principal rotatable bonds contributing to conformational flexibility. **b** Two-dimensional kPCA projection of the initial unsupervised representation. Gray points correspond to the full ensemble of 949 conformers, whereas colored circles denote the 28 conformers selected by FPS for the first round of DFT calculations. Colors indicate energies relative to the DFT global minimum, E_{gm}, with **b** and **c** sharing a common color scale. **c** kPCA projection of the CKA-refined representation obtained upon supervision from the first round of DFT calculations. Circles correspond to the initial 28 conformers, whereas stars indicate the additional 19 conformers selected by FPS during the second selection round within the low-energy region highlighted by the dashed rectangle.

We next investigated whether a geometrically representative subset spanning the conformational landscape could be selected with a realistic DFT budget. As an illustrative example, we imposed an upper limit of 47 DFT calculations, corresponding to 5% of the 949-member conformer ensemble. Although the specific budget is user-defined, it reflects a realistic computational constraint for exploring conformational space. We then divided the budget in two rounds, allowing the initial unsupervised representation to be subsequently refined once DFT information became available. In the first round, FPS was applied to the unsupervised two-channel representation to select 28 conformers (3% of the ensemble) for DFT evaluation. The resulting DFT energies then provided limited supervision to re-optimize the representation, which was used to allocate the remaining 19 calculations.

The 28 conformers selected in the first round were relaxed with DFT, and their relative energies with respect to the identified DFT global minimum, E_{gm}, are shown in **Fig. 2b**. This structurally

diverse subset spans an energy window of 18.8 kcal mol^{-1}. Inspection of the DFT-relaxed structures further reveals that much of the energetic variation can be attributed to changes in only four dihedral angles, corresponding to the two methoxy substituents adjacent to the biaryl linkage, the ligand aryl ring closest to the substrate ester group, and the substrate C–C bond. Despite the apparent complexity of the conformational landscape, these few degrees of freedom account for much of the observed energetic diversity, reducing an otherwise high-dimensional conformational problem to a small number of geometrical features.

Having obtained this first set of DFT energies, we could introduce supervision into the representation. Using the DFT energies of the 28 relaxed conformers, we repeated the hyperparameter grid search to maximize the CKA between the global kernel and the DFT energy vector (**Fig. S4**). The resulting two-channel representation achieved a CKA of 0.51, compared with 0.20 for the initial unsupervised representation, indicating substantially stronger alignment between structural similarity and DFT-energy variation. Both channels now employed linear local kernels, with the core channel becoming more local ($r_{cut} = 2.0$ Å) and the periphery channel more extended ($r_{cut} = 5.0$ Å), whereas the remaining hyperparameters were unchanged. These changes suggest that energetic variation across the conformational landscape is better aligned with highly local geometric information around the Rh coordination sphere and more extended geometric information within the ligand environment. Increasing the spatial range of the former introduces additional conformational variation that does not improve its alignment with energy and may overlap with information already captured by the ligand representation. Details of the optimized representation are provided in the **Supplementary Materials**. The improved structure–energy alignment is also visually reflected in the corresponding two-dimensional kPCA projection shown in **Fig. 2c**, where low- and high-energy regions become better organized across the configurational landscape.

Once the representation had been refined using the first round of DFT calculations, the remaining budget of 19 calculations could be allocated according to different scientific objectives. To illustrate this flexibility, we considered three alternative selection strategies implemented in TASTET: (i) global exploration, in which a space-filling sampling algorithm (FPS or *k*-medoids) is applied again to the full conformer ensemble; (ii) regional refinement, in which a region of interest is defined in the energy-colored kPCA projection and conformers belonging to that region are selected; and (iii) local exploration, in which selection is restricted to the neighborhood of a conformer of particular interest. Together, these strategies progressively shift the emphasis from broad coverage of the conformational landscape to increasingly targeted exploration guided by the information obtained during the first round of DFT calculations.

The broadest strategy applies FPS a second time using the CKA-refined representation to select an additional 19 conformers from the full ensemble for DFT relaxation. These structures are shown as stars in the kPCA projection in **Fig. S5a** and colored by their relative DFT energies. Their energies remained within the range already sampled by the initial 28 conformers, with the lowest additional conformer lying +4.2 kcal mol^{-1} above E_{gm}. Moreover, structures sampled from regions predicted to be energetically favorable remained among the lowest-energy conformers after DFT relaxation, whereas structures selected from higher-energy regions consistently relaxed to less favorable conformations. Since this second selection was performed

using a representation re-optimized with DFT-energy supervision, the absence of substantial new energetic variation provides retrospective support that the initial representative subset, comprising only 3% of the conformer ensemble and selected using a representation optimized without any DFT energies, had already captured the principal energetic regions of the conformational landscape.

A second way of allocating the remaining 19 DFT calculations is to focus them on a specific region of interest from the energy-colored kPCA projection. This provides an intuitive, user-accessible way of defining a region for further exploration, while the subsequent representative selection is still performed using distances in the full kernel space. Because a two-dimensional projection may not faithfully preserve the geometry of the underlying high-dimensional representation, this strategy should be applied with consideration of the fraction of kernel-space variance captured by the displayed kPCA components.

In the present example, the selected region corresponds to an apparent low-energy region identified after the first round of DFT calculations (dashed rectangle, **Fig. 2c**). FPS was therefore applied only to the structures contained within this region, preserving structural diversity while concentrating the remaining computational budget on a chemically relevant portion of the conformational landscape. The resulting conformers span energies between +1.5 and +10.6 kcal mol^{-1} relative to the global minimum, with nine structures lying within 6 kcal mol^{-1}. This targeted sampling therefore reveals that the apparent low-energy basin is not energetically homogeneous and illustrates how the remaining DFT budget can instead be used to resolve energetic variation within a selected region of conformational space.

A third way of allocating the remaining DFT budget is to investigate the candidates most geometrically similar to a specific conformer of interest, identified as its nearest neighbors in the kernel-induced configurational space. Rather than defining a region through the kPCA projection, the user selects a reference conformer and its nearest neighbors are identified directly from distances in the full kernel space. As a representative example, we selected the starting geometry that ultimately relaxed to the DFT global minimum (black circle, **Fig. S5b**) and selected its 19 nearest neighbors. Because DFT relaxation can substantially alter the initial conformer geometries, this neighborhood is defined solely by similarity between the unrelaxed structures. This provides a direct test of whether conformers that are close in the initial kernel space remain related after DFT relaxation or instead evolve toward different regions of the final conformational space. Following DFT relaxation, the 19 selected structures converged to only seven distinct minima, indicating that many neighboring starting geometries ultimately converge to the same local minimum. Their relative energies range from +6.2 to +11.4 kcal mol^{-1} (**Fig. S5b**), and one structure relaxed to the previously identified DFT global minimum. Therefore, geometrically similar starting structures may either converge to a common minimum or diverge toward distinct energetic states during relaxation. Neighborhood sampling is therefore particularly valuable for systematically characterizing individual regions of configurational space once they have been identified through broader representative exploration.

Together, these three sampling strategies demonstrate that representative structure selection need not be viewed as a single down-selection step. An initial structurally diverse subset can be used to introduce first-principles information and refine the structural representation, after

which the remaining computational budget can be allocated according to the scientific objective: continued global exploration, targeted refinement of a selected region, or neighborhood exploration around a structure of interest. This flexibility allows TASTET to support different modes of representative selection across configurational space as new information becomes available.

For comparison, we also evaluated the conventional energy-based conformer-screening strategy, in which a low-cost force-field calculation is used to prioritize candidate structures before DFT refinement. The full ensemble of 949 conformers was first relaxed using the Universal Force Field (UFF)[64], and the 25 lowest-energy conformers were then selected for DFT relaxation, corresponding to a very permissive energy window of 54 kcal mol^{-1} relative to the UFF global minimum (**Table S2**). Despite this broad energy window, the selected conformers occupy relatively restricted regions of conformational space (**Fig. S6**), resulting in substantial structural redundancy. In addition, several structures identified as favorable after UFF relaxation subsequently converged to high-energy minima upon DFT optimization, indicating that their energetic ranking at the force-field level does not necessarily transfer to the target level of theory. Thus, even with a broad energy window, low-level energy filtering can restrict configurational coverage while retaining structures that become energetically unfavorable after higher-level refinement.

Case study 2: Supported nanoclusters

Having demonstrated the framework for molecular conformational sampling, we next consider a different problem involving oxide-supported metal nanoclusters, where the large number of possible cluster-support arrangements gives rise to the structural diversity characteristic of fluxional cluster systems[65–68]. Such fluxionality motivates descriptions that account for ensembles of accessible structures rather than relying on a single optimized configuration, making efficient exploration of the underlying configurational space particularly important. As a representative example, we consider Cu_{12} nanoclusters supported on a ZnO(10–10) surface[69,70], providing a stringent test of whether the same geometry-based representative-selection strategy can be transferred across molecular and materials systems.

The large configurational spaces required to construct such ensembles are commonly explored using surrogate-assisted global optimization methods[8,33]. Among the current state-of-the-art approaches is Global Optimization with First-principles Energy Expressions (GOFEE)[71,72], which combines Bayesian optimization with an iteratively trained machine-learning surrogate to efficiently explore the configurational landscape. In the present case study, we use GOFEE, as implemented in AGOX[73], exclusively to generate a configurational ensemble of Cu_{12}/ZnO(10–10) structures. Within such workflows, geometry-based representative selection can operate at two distinct stages: by selecting structurally diverse configurations for DFT evaluation during surrogate-model training, and by prioritizing structures from the resulting configurational ensemble for subsequent high-accuracy DFT refinement. Here, we focus on the latter application and use the GOFEE-generated ensemble as the dataset for representative structure selection.

Consistent with common practice in surrogate-assisted global optimization, lower-accuracy DFT settings were deliberately employed during the GOFEE searches to accelerate surrogate-

model training and configurational exploration. Higher-accuracy DFT calculations were subsequently reserved exclusively for refinement of the representative structures selected by the present workflow. Complete computational details are provided in the **Supplementary Materials**.

Ten independent GOFEE runs, differing only in their random initialization seeds, were performed for Cu_{12}/ZnO(10–10), each producing 1,000 candidate structures (**Fig. S7**). The resulting configurational ensemble therefore comprises 10,000 structures, four representative examples of which are illustrated in **Fig. 3a**. In conventional surrogate-assisted workflows, structures from such an ensemble are typically prioritized for high-accuracy DFT refinement according to their predicted energies, for example by selecting those lying within a predefined window around the surrogate-predicted global minimum[35,36,38]. This approach parallels the low-level energy-based conformer screening examined in Case Study 1 but replaces force-field energies with predictions from a machine-learning surrogate. In both cases, the selection depends on the extent to which approximate energetics provide a reliable basis for deciding which structures should be carried forward to the target level of theory. As observed for UFF in Case Study 1, errors in the surrogate predictions can alter the energetic ordering of structurally distinct configurations and consequently determine which regions of configurational space are retained for subsequent refinement.

We address this limitation by shifting the selection criterion from surrogate energetic ranking to geometrical diversity. The objective is to identify a structurally diverse subset that maximizes coverage of the configurational landscape within a fixed DFT budget. Once the configurational ensemble has been mapped into a geometry-based representation, this subset can be selected systematically using space-filling sampling without relying on surrogate energetic ranking.

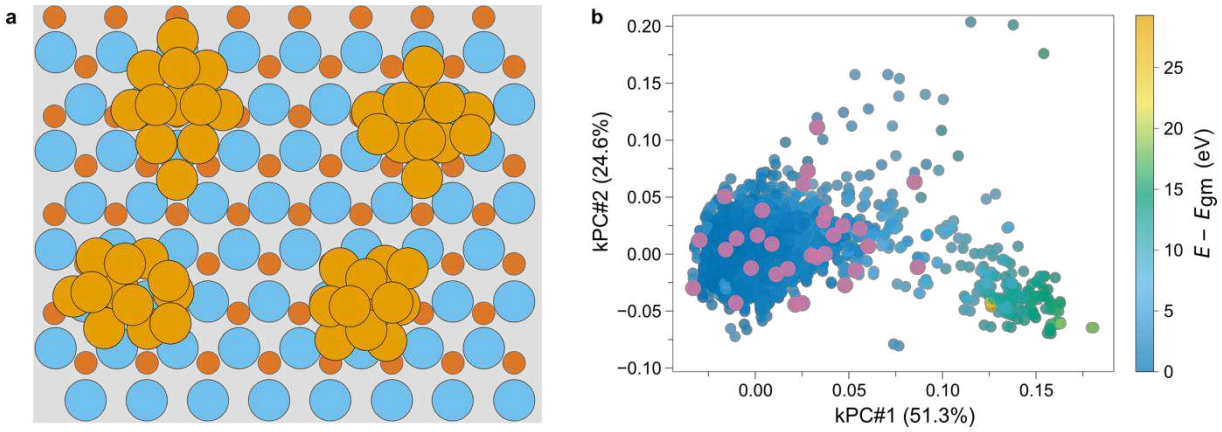

Fig. 3. Representative structure selection for the ZnO(10–10)-supported Cu_{12} nanocluster configurational landscape. a Four representative configurations (top view) from the configurational ensemble of 10,000 structures generated by ten independent GOFEE runs, illustrating the structural diversity of Cu_{12} clusters supported on ZnO(10–10). **b** Two-dimensional kPCA projection of the CKA-optimized representation. Structures (circles) are colored according to their GOFEE-predicted energies relative to the surrogate-predicted global minimum, E_{gm}. Magenta circles denote the 30 representative structures selected by FPS for high-accuracy DFT refinement.

Unlike in the first case study, we adopted a more pragmatic, single-stage approach to hyperparameter optimization. Here, the GOFEE-predicted energies were deliberately used as the supervisory target to identify the structural representation that best aligns with the energetic

variation learned by the surrogate model. This serves two purposes: it reveals how the surrogate energetic landscape is organized in terms of structural similarity and provides a basis for testing whether its predictions are reliable for structure selection. Crucially, the surrogate energies were not used to rank, filter, or impose an energy window on the candidate structures. Representative structures were instead selected by FPS across the resulting kernel space without energetic bias. High-accuracy DFT calculations on this broadly distributed subset can therefore establish whether the energetic relationships predicted by the surrogate persist at the target level of theory and, consequently, whether an energy-window-based selection would have been reliable.

SOAP descriptors were computed for local environments centered on all Cu atoms, so that structural similarity reflected the geometry of the nanocluster itself. The descriptor and kernel hyperparameters were then optimized using the supervised workflow described in the Methods Section. To make this optimization computationally tractable while retaining coverage across the surrogate energy range, a subsample of the configurational ensemble was selected by inverse-density weighting with respect to the GOFEE-predicted energies (**Fig. S8**).

The grid search explored combinations of n_{max}, l_{max}, σ, and r_{cut}, together with average global kernels (**Fig. S9**). The selected representation employed an average SOAP kernel with $r_{cut} =$ 4 Å, $\sigma = 0.1$ Å, $n_{max} = 8$, and $l_{max} = 4$, achieving a CKA value of 0.73 while favoring a comparatively simple representation that limits computational cost and descriptor sparsity.

The resulting configurational representation is visualized through the two-dimensional kPCA projection shown in **Fig. 3b**. Coloring the structures according to their GOFEE-predicted energies relative to the surrogate-predicted global minimum, E_{gm}, reveals an energy range of approximately 30 eV and a clear separation between two regions of configurational space. The smaller region on the right-hand side contains highly unfavorable, nonphysical structures lying approximately 5–30 eV above the surrogate global minimum. Inspection of the GOFEE trajectories shows that these structures originate predominantly from the random initial configurations generated at the beginning of each optimization run. By contrast, the larger region on the left-hand side comprises approximately 9,700 chemically meaningful structures generated during surrogate-guided exploration of the configurational landscape. Importantly, excluding the former structures does not require an arbitrarily chosen energetic cutoff. The two populations are clearly separated in configurational space and by an energetic gap substantially larger than the expected uncertainty of the surrogate model. FPS was therefore applied to the chemically meaningful ensemble of approximately 9,700 structures, selecting 30 representative configurations for high-accuracy DFT refinement (magenta circles in **Fig. 3b**), corresponding to approximately 0.3% of this ensemble. Their DFT energies relative to the identified global minimum, $E_{\mathrm{gm}}^{\mathrm{DFT}}$, are analyzed in **Fig. 4**.

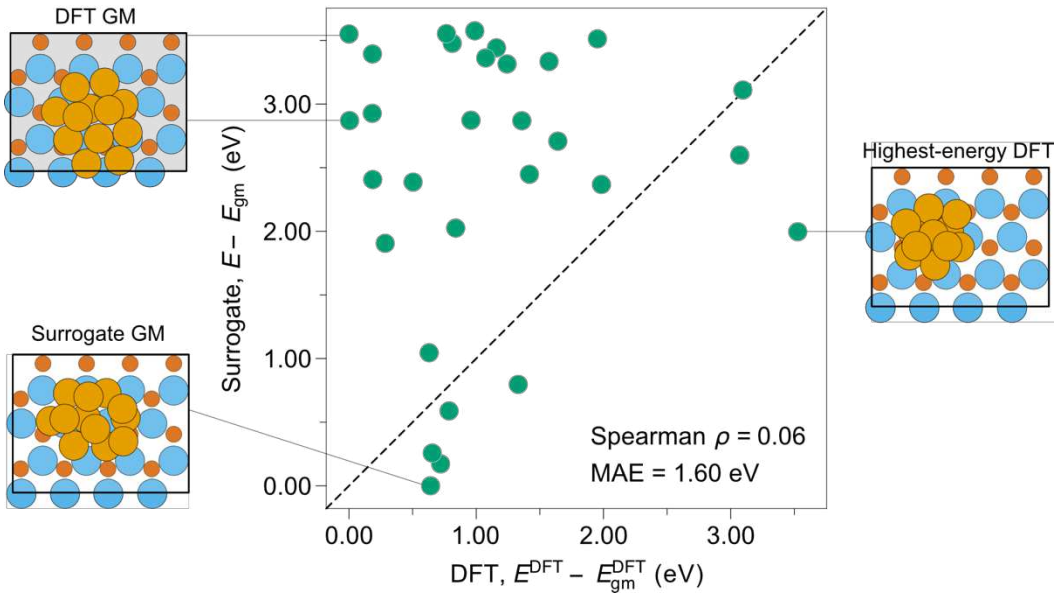

Fig. 4. Comparison of surrogate-predicted and DFT-computed relative energies for the 30 FPS-selected Cu_{12}/ZnO(10–10) structures. Each point represents one selected structure, with its DFT-computed relative energy, $E^{\mathrm{DFT}} - E_{\mathrm{gm}}^{\mathrm{DFT}}$, on the x-axis and its surrogate-predicted relative energy, $E - E_{\mathrm{gm}}$, on the y-axis. The DFT and surrogate energies are referenced to their respective global minima. The dashed line indicates perfect agreement between surrogate and DFT relative energies, with vertical deviations reflecting errors in the surrogate predictions. The Spearman rank correlation coefficient, ρ, and mean absolute error (MAE) quantify the agreement in energetic ranking and relative energies, respectively. The DFT-relaxed structures corresponding to the DFT global minimum and the surrogate-predicted global minimum are shown at the top left and bottom left, respectively, while the highest-energy DFT structure is shown on the right.

Following high-accuracy DFT relaxation, the 30 representative structures span an energy range of approximately 3.5 eV. Notably, two starting configurations that were structurally distinct in the kernel-induced space converged to the same DFT global-minimum structure, as indicated by the two points connected to the top-left inset in **Fig. 4**. This convergence is particularly informative because the structures were selected by FPS to maximize configurational coverage, yet distinct starting geometries can still reach the same minimum upon DFT relaxation. The full ensemble, which contains many more geometrically similar configurations, is therefore likely to exhibit even greater redundancy after DFT relaxation. Nevertheless, the remaining FPS-selected structures converge to a variety of metastable configurations distributed across the sampled energy range, demonstrating that geometry-based selection retains substantial structural and energetic diversity while requiring DFT refinement of only 0.3% of the configurational ensemble.

Comparison of the GOFEE-surrogate-predicted relative energies with the corresponding high-accuracy DFT values reveals poor agreement in the energetic ordering of the selected structures (**Fig. 4**). This is reflected in a Spearman rank correlation coefficient of $\rho = 0.06$ (p-value $=$ 0.74), indicating essentially no correlation between the surrogate and DFT energetic rankings. Notably, the surrogate-predicted global minimum ranks ninth out of 30 at the DFT level and lies 0.64 eV above the DFT global minimum. Conversely, the DFT global-minimum structure ranks only 28th according to the surrogate and lies 3.55 eV above the surrogate-predicted minimum, illustrating the substantial mismatch between the two energetic orderings. To quantify discrepancies in the energies relevant to energy-window-based screening, the

surrogate and DFT energies were referenced to their respective minima, yielding a mean absolute error (MAE) of 1.60 eV. A reasonably permissive surrogate-energy window of 1 eV is therefore narrower than the average absolute discrepancy between the surrogate-predicted and DFT-computed relative energies. Selection based on such a window could consequently exclude configurations that are subsequently found to be energetically favorable at the target level of theory.

These results highlight the complementary roles of surrogate-assisted global optimization and geometry-based representative sampling within the overall workflow. Surrogate-assisted optimization enables efficient exploration of large configurational spaces and generation of chemically meaningful candidate structures, while geometry-based sampling provides a means of selecting from the resulting ensemble without relying on surrogate energetic ranking as the final criterion. High-accuracy DFT calculations can then be concentrated on a structurally diverse subset, allowing the energetics of the selected configurations to be established directly at the target level of theory.

Further inspection of the DFT-refined structures reveals a clear relationship between cluster morphology and stability. Lower-energy configurations generally exhibit greater wetting of the ZnO(10–10) surface by the Cu_{12} cluster (top-left inset in **Fig. 4**), consistent with stabilization through increased metal–support interactions. By contrast, higher-energy configurations tend to adopt more vertically extended morphologies (right inset in **Fig. 4**), in which the cluster protrudes away from the surface rather than spreading across it. Representative sampling therefore not only avoids dependence on surrogate energetic ranking for final structure selection but also provides a diverse structural basis for identifying geometric features associated with energetic stability across the configurational landscape.

Conclusions

This work establishes representative structure selection as a coverage problem in configurational space rather than an energetic filtering problem. We implement this framework in TASTET, an open-source Python package that encodes candidate structures through local atomic-environment descriptors and structure-level similarity kernels and distributes a fixed first-principles computational budget across configurational space using space-filling sampling. The separation between structural coverage and energetic ranking is central to the framework. Approximate energies can guide optimization of the kernel representation when available, but they are not used as hard exclusion criteria during subset selection. TASTET can therefore complement conformer-generation, global-optimization, and surrogate-model workflows at different stages, ranging from selecting configurations for model training to prioritizing structures for final high-accuracy refinement.

The two case studies demonstrate this principle across substantially different configurational problems. For the Rh-complex ensemble, an initial DFT evaluation of only 28 of 949 conformers (3%) recovered the principal energetic regions of the conformational landscape and provided sufficient supervision to refine the configurational representation. Expanding the sampling to a predefined 5% DFT budget enabled either continued global exploration or increasingly targeted sampling of selected regions and local neighborhoods. Much of the energetic variation across this complex conformational landscape could also be associated with

only four key torsional degrees of freedom, providing an interpretable connection between molecular structure and energetics. Comparison with conventional UFF energy-based screening further showed that even a permissive low-level energy window can produce structurally redundant selections and relaxation outcomes that differ substantially from those obtained directly at the DFT level.

For the Cu_{12}/ZnO(10–10) system, high-accuracy DFT refinement of only 30 structures, roughly 0.3% of the configurational ensemble, exposed substantial discrepancies between surrogate and DFT-level energetics while retaining configurations spanning approximately 3.5 eV. The surrogate yielded errors larger than energy windows that might reasonably be used for final structure prioritization, illustrating the potential limitations of applying hard energetic cutoffs when their resolution exceeds the predictive accuracy of the underlying model. The refined structures further revealed a clear relationship between cluster morphology and stability, with lower-energy configurations exhibiting greater surface wetting and more dispersed morphologies than their higher-energy, vertically extended counterparts. Together with the UFF benchmark, these results illustrate how geometry-based representative selection can preserve structural diversity when approximate energetic rankings are insufficiently reliable for final structure prioritization.

The framework nevertheless has well-defined limitations. Representative selection is constrained by both the diversity of the input ensemble and the ability of the chosen representation to resolve that diversity. No selection algorithm can recover configurations absent from the ensemble or distinguish structures that the representation maps as identical. In TASTET, this latter limitation is inherited from the use of SOAP descriptors. Distinct local environments can in some cases yield identical or nearly identical representations[74,75], while the standard SOAP power spectrum is invariant to spatial inversion, such that enantiomeric environments cannot be distinguished[76]. Applications in which chirality or other structural distinctions not resolved by SOAP are essential therefore require a representation designed to retain the relevant information[77]. Conversely, increasing the resolution of a representation cannot create meaningful diversity in an intrinsically redundant ensemble and may instead amplify negligible geometric differences. Visual inspection of the candidate structures and resulting representation is therefore advisable before committing computational resources to first-principles calculations.

A separate limitation arises because representative selection is performed on the initial, unrelaxed structures. Structural relaxation can substantially reorganize configurational space, such that similarity between initial geometries does not necessarily translate into similarity between the corresponding relaxed structures or their properties. Indeed, both case studies show that initially distinct structures can converge to the same minimum, while geometrically similar starting structures may undergo different relaxation outcomes.

In addition, as ensemble size increases, hyperparameter optimization can become computationally demanding because it requires repeated construction of descriptor and kernel matrices. As demonstrated here, this cost can be reduced by optimizing the representation on a suitably chosen subsample and subsequently applying the selected hyperparameters in a single full-ensemble calculation for representative selection.

Representative sampling should therefore be regarded as complementary to global optimization and other structure-generation methods. These approaches can efficiently explore configurational space and generate candidate structures, while TASTET addresses the subsequent problem of allocating a limited first-principles computational budget across the resulting ensemble according to structural coverage. By integrating atomistic representations, kernel-based similarity measures, hyperparameter optimization, and space-filling sampling within an accessible Python workflow, TASTET provides a general toolset for incorporating representative structure selection into molecular and materials modeling. Its independence from any particular structure-generation method or energetic model allows it to interface with existing computational workflows while reducing reliance on energetic assumptions during final structure selection. By making these capabilities available through an open-source implementation, we hope TASTET will facilitate systematic representative structure selection and support the construction of physically meaningful models in future computational studies.

Acknowledgements

The authors gratefully acknowledge financial support from the European Union's Horizon program under the Marie Skłodowska-Curie Doctoral Networks (MSCA-DN) grant agreement No. 101072830 (A.C.-A. and M.G.-M.); the European Union's Horizon Europe research and innovation program under grant agreement No. 101115204 (P.C.-L. and M.G.-M.) and No. 101126600 (M.G.-M.); Taighde Eireann – Research Ireland under grant agreement GOIPG/2025/8993 (A.M.); and the Basque Government through the ELKARTEK program (CIC2025, KK2025-00054). Computational resources and technical support were provided by the Spanish Supercomputing Network (RES) and the MareNostrum cluster hosted at the Barcelona Supercomputing Center (BSC) under Project ID QHS-2026-1-0008.

Author contributions

A.C.-A.: Conceptualization, methodology, software, formal analysis, investigation, data curation, writing – original draft, and visualization. A.M.: Formal analysis, investigation, data curation, writing – original draft, and visualization. P.C.-L.: Data curation and writing – review and editing. G.W.W.: Supervision, project administration, and writing – review and editing. M.G.-M.: Conceptualization, supervision, project administration, funding acquisition, and writing – review and editing.

Conflicts of interest

The authors declare no competing interests.

Data availability

The DFT data underlying the two case studies, including the atomic structures and corresponding energies, are openly available through the ioChem-BD repository at https://iochem-bd.bsc.es/browse/review-collection/100/503816/cfd0f1b32a3219b12ea3b14a.

Code availability

The TASTET source code, together with the scripts and input structural ensembles required to reproduce the analyses reported in this study, are openly available through the TASTET GitHub

repository at https://github.com/CCEMGroupTCD/TASTET. TASTET is also distributed through the Python Package Index (PyPI) under the package name `tastet` and can be installed directly using `pip`.

Notes and references

1. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* **111**, 218–230 (2016).
2. Ganose, A. M. *et al.* Atomate2: modular workflows for materials science. *Digit. Discov.* **4**, 1944–1973 (2025).
3. Grygorenko, O. O. *et al.* Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **23**, 101681 (2020).
4. Barros-Luque, L. *et al.* The Open Materials 2024 (OMat24) inorganic materials dataset and models. *Nat. Comput. Sci.* **6**, 642–652 (2026).
5. Okhotnikov, K., Charpentier, T. & Cadars, S. Supercell program: a combinatorial structure-generation approach for the local-level modeling of atomic substitutions and partial occupancies in crystals. *J. Cheminformatics* **8**, 17 (2016).
6. Mamun, O., Winther, K. T., Boes, J. R. & Bligaard, T. A Bayesian framework for adsorption energy prediction on bimetallic alloy catalysts. *Npj Comput. Mater.* **6**, 177 (2020).
7. Yeo, B. C. *et al.* High-throughput computational-experimental screening protocol for the discovery of bimetallic catalysts. *Npj Comput. Mater.* **7**, 137 (2021).
8. Han, S., Lysgaard, S., Vegge, T. & Hansen, H. A. Rapid mapping of alloy surface phase diagrams via Bayesian evolutionary multitasking. *Npj Comput. Mater.* **9**, 139 (2023).
9. Yu, B., Lordi, V. & Schwalbe-Koda, D. Maximizing efficiency of dataset compression for machine learning potentials with information theory. *J. Chem. Phys.* **164**, 164111 (2026).
10. Pracht, P. *et al.* CREST—A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.* **160**, 114110 (2024).
11. Riniker, S. & Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
12. RDKit: Open-source cheminformatics. https://doi.org/10.5281/zenodo.18797641 (2026).
13. de Souza, B. GOAT: A Global Optimization Algorithm for Molecules and Atomic Clusters. *Angew. Chem. Int. Ed.* **64**, e202500393 (2025).
14. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminformatics* **3**, 33 (2011).
15. Zuccarello, G. *et al.* Enantioselective Catalysis with Pyrrolidinyl Gold(I) Complexes: DFT and NEST Analysis of the Chiral Binding Pocket. *JACS Au* **3**, 1742–1754 (2023).
16. Finta, S., Kalikadien, A. V. & Pidko, E. A. Data-Driven Virtual Screening of Conformational Ensembles of Transition-Metal Complexes. *J. Chem. Theory Comput.* **21**, 5334–5345 (2025).
17. Minenkov, Y., Sharapa, D. I. & Cavallo, L. Application of Semiempirical Methods to Transition Metal Complexes: Fast Results but Hard-to-Predict Accuracy. *J. Chem. Theory Comput.* **14**, 3428–3439 (2018).

18. Otlyotov, A. A., Rozov, T. P., Moshchenkov, A. D. & Minenkov, Y. 16TMCONF543: An Automatically Generated Data Set of Conformational Energies of Transition Metal Complexes Relevant to Catalysis. *Organometallics* **43**, 2232–2242 (2024).
19. Nguyen, L. T. T., Larsson, E. D., Niklasson, K. M. F., Wieduwilt, E. K. & Hedegård, E. D. Efficient Calculation of Absorption Spectra of Platinum Complexes Used as Luminescent Probes for Cancer Detection. Preprint at https://doi.org/10.48550/arXiv.2602.18284 (2026).
20. Laplaza, R., Wodrich, M. D. & Corminboeuf, C. Overcoming the Pitfalls of Computing Reaction Selectivity from Ensembles of Transition States. *J. Phys. Chem. Lett.* **15**, 7363–7370 (2024).
21. Zhou, Y., Limbu, I., Garson, M. J. & Krenske, E. H. Conformational Sampling in Computational Studies of Natural Products: Why Is It Important? *J. Nat. Prod.* **87**, 2543–2549 (2024).
22. Ghanekar, P. G., Deshpande, S. & Greeley, J. Adsorbate chemical environment-based machine learning framework for heterogeneous catalysis. *Nat. Commun.* **13**, 5788 (2022).
23. Lan, J. *et al.* AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *Npj Comput. Mater.* **9**, 172 (2023).
24. Radova, M., Stark, W. G., Allen, C. S., Maurer, R. J. & Bartók, A. P. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. *Npj Comput. Mater.* **11**, 237 (2025).
25. Schwalbe-Koda, D., Govindarajan, N. & Varley, J. B. Comprehensive sampling of coverage effects in catalysis by leveraging generalization in neural network models. *Digit. Discov.* **4**, 234–251 (2025).
26. Loveday, O., Kaźmierczak, K. & López, N. Challenges and Opportunities of Pretrained Machine Learning Interatomic Potentials in Heterogeneous Catalysis. *ACS Catal.* **16**, 4113–4124 (2026).
27. Tran, K. & Ulissi, Z. W. Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. *Nat. Catal.* **1**, 696–703 (2018).
28. Gu, G. H., Lee, M., Jung, Y. & Vlachos, D. G. Automated exploitation of the big configuration space of large adsorbates on transition metals reveals chemistry feasibility. *Nat. Commun.* **13**, 2087 (2022).
29. Deaven, D. M. & Ho, K. M. Molecular Geometry Optimization with a Genetic Algorithm. *Phys. Rev. Lett.* **75**, 288–291 (1995).
30. Wales, D. J. & Doye, J. P. K. Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
31. Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **120**, 9911–9917 (2004).
32. Vilhelmsen, L. B. & Hammer, B. A genetic algorithm for first principles global structure optimization of supported nano structures. *J. Chem. Phys.* **141**, 044711 (2014).
33. Jung, H., Sauerland, L., Stocker, S., Reuter, K. & Margraf, J. T. Machine-learning driven global optimization of surface adsorbate geometries. *Npj Comput. Mater.* **9**, 114 (2023).
34. Xie, W., Han, Y., Wu, C. & Hu, P. Smarter Data: Rethinking Data Generation for Machine Learning Potentials in Heterogeneous Catalysis. *JACS Au* **6**, 3081–3091 (2026).

35. Kempen, L. H. E. & Andersen, M. Inverse catalysts: tuning the composition and structure of oxide clusters through the metal support. *Npj Comput. Mater.* **11**, 8 (2025).
36. Pham, T. N. *et al.* Stability of PdxOy Particles Supported on Strontium Titanate Perovskite under Three-Way Catalyst Operating Conditions: Implications for Sintering Resistance. *ACS Catal.* **14**, 1443–1458 (2024).
37. Hamamoto, Y., Pham, T. N., Bisbo, M. K., Hammer, B. & Morikawa, Y. Machine-learned search for the stable structures of silicene on Ag(111). *Phys. Rev. Mater.* **7**, 124002 (2023).
38. Quinlivan Domínguez, J. E., Neyman, K. M. & Bruix, A. Stability of oxidized states of freestanding and ceria-supported PtOx particles. *J. Chem. Phys.* **157**, 094709 (2022).
39. Shapeev, A. V. Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2016).
40. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
41. Deringer, V. L. *et al.* Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
42. Musil, F. *et al.* Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **121**, 9759–9815 (2021).
43. Cheng, B. *et al.* Mapping Materials and Molecules. *Acc. Chem. Res.* **53**, 1981–1991 (2020).
44. Lai, K. C., Matera, S., Scheurer, C. & Reuter, K. A fuzzy classification framework to identify equivalent atoms in complex materials and molecules. *J. Chem. Phys.* **159**, 024129 (2023).
45. Li, R. *et al.* Local-environment-guided selection of atomic structures for the development of machine-learning potentials. *J. Chem. Phys.* **160**, 074109 (2024).
46. Goscinski, A. *et al.* scikit-matter : A Suite of Generalisable Machine Learning Methods Born out of Chemistry and Materials Science. *Open Res. Eur.* **3**, 81 (2023).
47. Hjorth Larsen, A. *et al.* The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
48. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
49. Himanen, L. *et al.* DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
50. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).
51. Bartók, A. P. *et al.* Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **3**, e1701816 (2017).
52. Willatt, M. J., Musil, F. & Ceriotti, M. Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements. *Phys. Chem. Chem. Phys.* **20**, 29661–29668 (2018).
53. Cortes, C., Mohri, M. & Rostamizadeh, A. Algorithms for Learning Kernels Based on Centered Alignment. *J. Mach. Learn. Res.* **13**, 795–828 (2012).
54. Kornblith, S., Norouzi, M., Lee, H. & Hinton, G. Similarity of Neural Network Representations Revisited. Preprint at https://doi.org/10.48550/arXiv.1905.00414 (2019).
55. Welsch, M., Hirte, S. & Kirchmair, J. Deciphering Molecular Embeddings with Centered Kernel Alignment. *J. Chem. Inf. Model.* **64**, 7303–7312 (2024).

56. Welsch, M., Jiang, E., Papantonis, I. & Kirchmair, J. Relating Model Performance to Embedding Distributions in Molecular Machine Learning. *J. Chem. Inf. Model.* **66**, 5125–5134 (2026).
57. Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *J. Chem. Phys.* **150**, 150901 (2019).
58. Gallagher, Q. M. & Webb, M. A. Data efficiency of classification strategies for chemical and materials design. *Digit. Discov.* **4**, 135–148 (2025).
59. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **10**, 1299–1319 (1998).
60. Chari, T. & Pachter, L. The specious art of single-cell genomics. *PLOS Comput. Biol.* **19**, e1011288 (2023).
61. Crawford, J. M. & Sigman, M. S. Conformational Dynamics in Asymmetric Catalysis: Is Catalyst Flexibility a Design Element? *Synthesis* **51**, 1021–1036 (2019).
62. McNutt, A. T. *et al.* Conformer Generation for Structure-Based Drug Design: How Many and How Good? *J. Chem. Inf. Model.* **63**, 6598–6607 (2023).
63. Baidun, M. S., Kalikadien, A. V., Lefort, L. & Pidko, E. A. Impact of Model Selection and Conformational Effects on the Descriptors for In Silico Screening Campaigns: A Case Study of Rh-Catalyzed Acrylate Hydrogenation. *J. Phys. Chem. C* **128**, 7987–7998 (2024).
64. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A., III & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
65. Zhai, H. & Alexandrova, A. N. Fluxionality of Catalytic Clusters: When It Matters and How to Address It. *ACS Catal.* **7**, 1905–1911 (2017).
66. Zhang, Z., Zandkarimi, B. & Alexandrova, A. N. Ensembles of Metastable States Govern Heterogeneous Catalysis on Dynamic Interfaces. *Acc. Chem. Res.* **53**, 447–458 (2020).
67. Poths, P., Vargas, S., Sautet, P. & Alexandrova, A. N. Thermodynamic Equilibrium versus Kinetic Trapping: Thermalization of Cluster Catalyst Ensembles Can Extend Beyond Reaction Time Scales. *ACS Catal.* **14**, 5403–5415 (2024).
68. Poths, P., Lai, K. C., Scheurer, C., Matera, S. & Reuter, K. Kinetic restructuring of catalyst active sites: a MACE-APE study of fluxional Pdn/MgO (n = 3–11) clusters. *Faraday Discuss.* https://doi.org/10.1039/D5FD00138B (2026).
69. Liu, L. & Corma, A. Metal Catalysts for Heterogeneous Catalysis: From Single Atoms to Nanoclusters and Nanoparticles. *Chem. Rev.* **118**, 4981–5079 (2018).
70. Guo, Y., Wang, M., Zhu, Q., Xiao, D. & Ma, D. Ensemble effect for single-atom, small cluster and nanoparticle catalysts. *Nat. Catal.* **5**, 766–776 (2022).
71. Bisbo, M. K. & Hammer, B. Global optimization of atomic structure enhanced by machine learning. *Phys. Rev. B* **105**, 245404 (2022).
72. Bisbo, M. K. & Hammer, B. Efficient Global Structure Optimization with a Machine-Learned Surrogate Model. *Phys. Rev. Lett.* **124**, 086102 (2020).
73. Christiansen, M.-P. V., Rønne, N. & Hammer, B. Atomistic Global Optimization X: A Python package for optimization of atomistic structures. *J. Chem. Phys.* **157**, 054701 (2022).
74. Pozdnyakov, S. N. *et al.* Incompleteness of Atomic Structure Representations. *Phys. Rev. Lett.* **125**, 166001 (2020).

75. Nigam, J., Pozdnyakov, S. N., Huguenin-Dumittan, K. K. & Ceriotti, M. Completeness of atomic structure representations. *APL Mach. Learn.* **2**, 016110 (2024).
76. Grisafi, A., Wilkins, D. M., Csányi, G. & Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **120**, 036002 (2018).
77. Xu, L.-C. *et al.* A Molecular Stereostructure Descriptor Based On Spherical Projection. *Synlett* **32**, 1837–1842 (2021).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SMnpjCompMat20260912.docx