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.

Exploring the Chemical Space of Metal Clusters via Machine Learning

Sheng-Gui He

shengguihe@iccas.ac.cn

Institute of Chemistry, Chinese Academy of Sciences

Ning-Zheng Li

Key Laboratory for Structural Chemistry of Unstable and Stable Species, Institute of Chemistry, Chinese Academy of Sciences

Zi-Yue Wang

Institute of Software, Chinese Academy of Sciences

Zi-Yu Li

Institute of Chemistry, Chinese Academy of Sciences

Qiang Shi

Institute of Chemistry, Chinese Academy of Sciences

Qing-Yu Liu

Institute of Chemistry, Chinese Academy of Sciences

Article

Keywords:

Posted Date: February 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8784571/v1

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

Additional Declarations: There is **NO** Competing Interest.

Exploring the Chemical Space of Metal Clusters via Machine Learning

Ning-Zheng Li,[1,3] Zi-Yue Wang,[2,3] Zi-Yu Li,[1,3,4*] Qiang Shi,[1,3,4] Qing-Yu Liu,[1,3,4] Sheng-Gui He[1,3,4,5*]

[1]State Key Laboratory for Structural Chemistry of Unstable and Stable Species, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100190, People's Republic of China.

[2]Institute of Software, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

[3]University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China.

[4]Beijing National Laboratory for Molecular Sciences and Center for Carbon Neutral Chemistry, Institute of Chemistry, Chinese Academy of Sciences, Beijing 100190, People's Republic of China.

[5]Nanjing Institute of Atomic Scale Manufacturing, Nanjing 211899, People's Republic of China.

***Corresponding Authors.** E-mail: liziyu2010@iccas.ac.cn; shengguihe@iccas.ac.cn

Abstract

Atomic clusters serve as the embryos of materials, yet their enormous and compositionally complex chemical space has long hindered systematic exploration of thermodynamic stability. Here, a unified first-principles and machine-learning framework is developed to enable large-scale mapping of metal cluster thermodynamics. By introducing a progressive sampling strategy integrated with a composition-based deep learning model, the intrinsic sampling bottleneck associated with exponentially expanding chemical spaces is alleviated. Based on ~45,000 global-minimum structures obtained via automated first-principles calculations, a composition-based deep learning model, the Cluster Transformer Encoder Network, enables reliable predictions of atomization energies (~50 meV/atom accuracy) for 9.13 million metal cluster compositions, covering 30 d-block metals and 4 main-group elements. The resulting thermodynamic trends reveal strong correlations between cluster atomization energies and bulk cohesive energies and highlight heteronuclear stabilization and the stability of noble-metal-doped oxide clusters. This work establishes a general strategy for efficient exploration of chemically complex cluster spaces and advances a holistic understanding of the collective properties of metal clusters.

Introduction

Clusters are widely regarded as a conceptual bridge linking isolated atoms and molecules to condensed matter systems (*1*). As finite aggregates stabilized by covalent or non-covalent interactions, clusters govern a wide range of chemical and physical processes, including catalysis, nucleation, and phase transformations. Among various classes of clusters, metal clusters have attracted sustained interest because of the elemental diversity of metals and the rich spectrum of properties they exhibit. Extensive experimental and theoretical studies have revealed that metal clusters can display exceptional electric (*2, 3*), magnetic (*4, 5*), and chemical reactivities (*6–10*), often distinct from both atomic and bulk limits. To elucidate the origin of these properties, the geometric and electronic structures of numerous metal clusters have been characterized through quantum chemical calculations (*11–13*) and advanced spectroscopic techniques (*12, 14–18*). Despite this progress, the combinatorial explosion arising from variations in metal and ligand composition renders exhaustive experimental or computational characterization infeasible. Consequently, a unified and predictive understanding of thermodynamic and chemical trends across the vast chemical space of metal clusters remains largely absent.

Machine learning (ML) has emerged as a powerful statistical framework for identifying hidden correlations within complex chemical data and has been successfully applied across a wide range of chemical systems. From small organic molecules to extended crystalline solids, ML models trained on carefully curated datasets have enabled the efficient screening of large chemical spaces and the identification of candidates with targeted properties. This paradigm has accelerated discoveries in areas including drug-like molecules (*19, 20*), electrochemical reactions (*21*), organometallic catalysis (*22*), organic optoelectronic materials (*23*), and thermodynamically stable crystalline phases (*24–29*), with several predictions subsequently validated experimentally. Despite this progress, extending ML-based chemical space exploration to metal clusters remains challenging. Existing ML studies on metal clusters have largely focused on constructing high-fidelity potential energy surfaces (PES) (*30–32*) or identifying structure-reactivity relationships within narrowly defined cluster families (*33, 34*), rather than enabling systematic exploration across chemically diverse cluster spaces. Unlike organic molecules, metal clusters feature extensive metal-metal bonding and variable coordination

environments, necessitating descriptors that can faithfully encode elemental identity and compositional complexity. At a more fundamental level, an essential and largely unexplored question persists: how much data is required to achieve reliable and transferable learning as the underlying chemical space grows exponentially? Addressing this question is critical not only for metal clusters but also for the broader goal of scalable chemical space exploration.

To overcome the intrinsic impracticality of exhaustively enumerating metal clusters in an exponentially expanding chemical space, this work established a unified first-principles and machine learning framework centered on a progressive sampling strategy. The chemical space considered spans 34 elements, including 30 d-block metals and 4 common ligands (C, N, O, and S), giving rise to millions of distinct cluster compositions. Instead of attempting comprehensive coverage, physically representative clusters, defined by their global-minimum structures, are selectively sampled through first-principles calculations, providing a reliable foundation for data-driven learning. Building on these data, a composition-based machine learning model is progressively trained as increasingly complex cluster subspaces are incorporated. This strategy enables a direct examination of how predictive accuracy scales with both dataset size and chemical complexity, thereby addressing a central and previously unresolved question in chemical space exploration: the data requirements for reliable and transferable learning in systems governed by combinatorial growth. As demonstrated in this work, progressive sampling reveals a pronounced data efficiency in learning metal-cluster thermodynamics, establishing a general and scalable strategy for navigating chemically extensible spaces and for quantitatively linking clusters and condensed-phase materials.

Results

Framework and Sampling Strategy for Exploring Metal-Cluster Chemical Space. The exploration of metal-cluster thermodynamics is fundamentally constrained by the combinatorial growth of chemical space with cluster size and composition, posing a stringent test of whether reliable thermodynamic trends can be learned from limited data. To define a chemically rich yet tractable search domain, the metal-cluster chemical space considered in this work was constructed according to well-defined compositional constraints. Specifically, 34 elements were included, comprising 30 d-

block metals (Sc–Zn, Y–Cd, La, and Hf–Hg) together with four common nonmetal ligands (C, N, O, and S). For each cluster composition, the number of ligand atoms (N_L) was restricted not to exceed the number of metal atoms (N_M), and clusters were allowed to carry an overall charge of 0, −1, or +1. The resulting chemical space was further partitioned into subspaces according to N_M, with the size of each subspace increasing exponentially with the number of metal atoms. As summarized in **Fig. 1A**, even under these constraints, the number of possible metal cluster compositions increases exponentially with N_M, highlighting the intrinsic vastness of the chemical space. In practice, however, the accessibility of this space is severely limited by the rapidly increasing computational cost of explicit first-principles calculations. As shown in **Fig. 1B**, the average computational time required to locate global minimum (GM) structures increases steeply with cluster size, reflecting the growing complexity of both geometric optimization and spin-multiplicity searches in larger clusters. These constraints render exhaustive enumeration impractical beyond the smallest subspaces. Motivated by this computational bottleneck, controlled sampling schemes were deliberately designed to systematically probe learning behavior under data-limited conditions. As illustrated in **Fig. 1C**, distinct subsets of cluster compositions with varying sampling sizes were selected for successive N_M subspaces, enabling a quantitative assessment of how predictive performance scales with both dataset size and chemical complexity.

To enable systematic exploration under the above sampling design, an automated framework was constructed that integrates first-principles calculations with composition-based machine learning. As summarized in **Fig. 2**, this framework couples the generation of representative cluster candidates, the evaluation of their thermodynamic metrics, and the progressive incorporation of these data into predictive models spanning increasingly large compositional subspaces. The GM structures were obtained using an automated genetic algorithm (GA) combined with density functional theory (DFT) calculations (**Fig. 2A**). The GA workflow performs global geometric optimization together with spin multiplicity searches, ensuring reliable identification of low-energy configurations across chemically diverse clusters. From the resulting structures, atomization energies (AEs), defined as the energy required to dissociate a cluster into its constituent neutral atoms (with the charge assigned to the most favorable atomic fragment for charged species), were computed as a uniform thermodynamic metric characterizing the intrinsic aggregation tendency of each cluster composition (see Methods for more

details).

To learn across the chemically diverse and compositionally complex dataset, we employed a Cluster Transformer Encoder Network (CTEN) (**Fig. 2B)**. In CTEN, each atom is represented by an atomic vector derived from its ground-state electron configuration, providing a transferable encoding of elemental identity. Sequences of atomic vectors corresponding to cluster compositions are processed by a multi-head transformer encoder to capture nonlocal correlations between constituent elements, followed by global pooling and a fully connected neural network to predict the average atomization energy. Notably, this model relies solely on compositional information without requiring explicit structural descriptors, enabling scalable prediction across millions of candidate compositions.

Building on CTEN, a progressive sampling protocol was implemented to quantify the amount of data required for reliable learning across increasingly complex N_M subspaces (**Fig. 2C**). For clusters containing $N_M = n$ metal atoms, models were trained using all available data from smaller clusters ($N_M \leq n-1$), supplemented by progressively larger subsets drawn from the computed $N_M = n$ species. Model performance was evaluated on a held-out test set of $N_M = n$ clusters as additional training samples were incorporated. When further sampling no longer produced a meaningful reduction in prediction error, the corresponding subspace was considered sufficiently learned, indicating that dominant compositional factors governing thermodynamic stability had been captured.

Learning Curves and Data Efficiency across Increasing N_M Subspaces. Using this protocol, the learning behavior was first examined in the smallest subspace (N_M = 1) and then systematically extended to larger and more complex systems. For N_M = 1, the GM structures and atomization energies were obtained for all 450 species, enabling a complete benchmark. Fifty species were reserved as the test set, and a series of eight models were trained using randomly selected training sets of increasing size (50, 100, …, 400 species) drawn from the remaining data. The resulting learning curve, shown in **Fig. 3(A1)**, plots the mean absolute errors (MAEs) on the testing set as a function of the sampled fractions of the N_M = 1 subspace. With only 50 training samples (11.1% of the subspace), the model yielded a testing MAE of 515.8 meV/atom. As the number of training samples increased, prediction accuracy improved rapidly. When 250 samples (55.6%) were used for training, the testing MAE decreased to 132.8 meV/atom, representing an approximately four-fold

reduction relative to the initial value. Further increases in training data led to only marginal improvements: the minimal testing MAE of 109.9 meV/atom was achieved at a sampling ratio of 66.7% (300 samples), after which the testing MAE fluctuated slightly and reached 138.8 meV/atom when all non-test N_M = 1 samples were included (Pearson r = 0.9879). Comparisons between calculated and predicted AEs for the final model are shown in **Figs. 3(A2) and 3(A3)**. Overall, the learning curve converges to an MAE of approximately 110–130 meV/atom at a sampling ratio of ~50% for the N_M = 1 subspace.

Progressive sampling was then applied to the N_M = 2 subspace, which contains 20,925 possible species. The GM structures were successfully determined for 20,706 clusters using the automated genetic algorithm workflow, while the remaining 219 species were excluded due to imaginary vibrational frequencies. All N_M = 1 samples were included in the training set for every model. A testing set of 200 N_M = 2 species (approximately 1% of the subspace) was reserved. The corresponding learning curve is shown in **Fig. 3(B1)**. When no N_M = 2 samples were included in training, the model yielded a testing MAE of 644.7 meV/atom, indicating limited transferability from N_M = 1 alone. Introducing only 100 N_M = 2 samples (0.48%) reduced the testing MAE sharply to 217.2 meV/atom, underscoring the necessity of including higher-order cluster data. With 5,000 N_M = 2 samples (23.9%), the testing MAE decreased further to 69.6 meV/atom, nearly an order-of-magnitude reduction relative to the initial value. Beyond this point, improvements became progressively smaller, and the best performance of 43.0 meV/atom (Pearson r = 0.9975) was achieved when nearly all available N_M = 2 samples were included. Detailed training and testing results for the final model are shown in **Figs. 3(B2) and 3(B3)**. Notably, reliable prediction accuracy for N_M = 2 clusters was achieved using only ~25% of the subspace, substantially lower than the ~50% required for N_M = 1.

The N_M = 3 subspace is substantially larger, containing 520,800 possible species. Due to its size, the GM structures were computed for a randomly sampled subset of 15,880 clusters (3.05% of the subspace). All N_M = 1 and N_M = 2 samples were again included in the training sets. A testing set of 500 N_M = 3 species (approximately 0.1%) was reserved. The learning curve for this subspace is shown in **Fig. 3(C1)**. Without any N_M = 3 samples in the training data, the testing MAE was 702.2 meV/atom. Adding only 100 N_M = 3 samples (0.02%) reduced the testing MAE to approximately 200 meV/atom.

Remarkably, including 2,604 $N_M = 3$ samples (0.5%) was sufficient to reduce the testing MAE below one-tenth of its initial value (69.8 meV/atom). Further increases in training data yielded diminishing returns: increasing the sampling ratio from 1% to 2.95% reduced the testing MAE only modestly, from 62.8 to 53.6 meV/atom (Pearson $r = 0.9977$). Training and testing comparisons for the final model are shown in **Figs. 3(C2) and 3(C3)**. These results demonstrate that accurate prediction of $N_M = 3$ cluster atomization energies can be achieved using data from only ~1% of the subspace.

The $N_M = 4$ subspace spans 8,593,200 possible species, for which the GM structure calculations remain computationally prohibitive. Atomization energies were available for 8,786 $N_M = 4$ clusters, from which a testing set of 1,500 species was constructed. All training sets included all available samples from $N_M = 1 – 3$. Notably, even in the absence of any $N_M = 4$ training samples, the model achieved a testing MAE of 87.4 meV/atom, substantially lower than the initial errors observed for $N_M = 1 – 3$ subspaces. As $N_M = 4$ samples were progressively added, testing MAE decreased steadily and reached 48.3 meV/atom (Pearson $r = 0.9981$) when all available $N_M = 4$ samples except the testing set were included (sampling ratio of 0.085%), as shown in **Fig. 3D**. The systematic reduction in both the absolute number and fraction of samples required to achieve converged accuracy with increasing N_M highlights the intrinsic data efficiency of the progressive sampling strategy. These results demonstrate that accurate prediction of metal cluster atomization energies at the ~50 meV/atom level is feasible even in extremely large chemical spaces, substantially reducing the computational cost associated with explicit first-principles enumeration.

High-AE Screening as a Test of Model Robustness. Building on the established learning behavior across $N_M = 1 – 4$ subspaces, the model was then applied to prioritize cluster species with the highest average AEs across the full chemical space. To this end, a model trained on a relatively small dataset, denoted as dataset S (**Fig. 4A**), was first employed to screen the entire chemical space, and the top 1,000 high-average-AE candidates were subsequently validated by first-principles calculations. Dataset S consisted of all available $N_M = 1$ and $N_M = 2$ species, together with 14,702 $N_M = 3$ species and 2,730 $N_M = 4$ species. Among the 1,000 top-ranked candidates identified by the model, 999 were not present in dataset S, demonstrating that the screening predominantly targets previously unseen compositions. For the single species overlapping with the training set, the prediction error relative to the DFT value was 130.3 meV/atom (**Fig. 4B**), whereas the remaining 999 unseen species exhibited

a predicted MAE of 66.0 meV/atom (**Fig. 4C**), indicating reliable extrapolation performance in the high-AE regime. To examine whether biasing the training data toward high-average-AE samples could further enhance screening accuracy, additional high-AE species were incorporated to construct an expanded dataset containing 45,834 cluster compositions (**Fig. 4D**). A new model trained on this enriched dataset generated an alternative set of the top 1,000 high-average-AE candidates. Among these, 302 species were absent from the expanded training set, providing a stringent test of generalization. The predicted MAEs for species present in and absent from the expanded dataset were 40.1 meV/atom (**Fig. 4E**) and 63.0 meV/atom (**Fig. 4F**), respectively, representing a modest improvement over the results obtained with dataset S. Overall, these results indicate that while enriching the training set with high-average-AE samples leads to slightly improved accuracy for extreme-AE candidates, the screening performance is only weakly dependent on the detailed distribution of the training data. The modeling framework, therefore, exhibits robust and stable behavior for prioritizing thermodynamically favorable metal clusters across vast chemical spaces, consistent with the intrinsic data efficiency observed in the progressive sampling analysis.

Global Thermodynamic Trends in the N_M = 1–4 Chemical Space. Using the model trained on all available N_M = 1 – 4 samples, the global thermodynamic landscape of metal clusters was subsequently analyzed. The predicted average AEs of all N_M = 1 – 4 species span a broad range from 0 to 6.21 eV/atom, forming a slightly left-skewed, bell-shaped distribution (**Fig. 5A**). The distribution reflects substantial variability in intrinsic aggregation tendencies across the chemical space. To elucidate element-specific trends, mean average AEs were computed for clusters containing at least one atom of a given d-block metal element (**Fig. 5B**). Clusters containing Ta, Zr, Hf, and Nb exhibit the highest mean average AEs, collectively forming a pronounced "stable island" within the d-block region of the periodic table. A similar analysis for ligand elements reveals a clear ordering of mean average AEs, decreasing from carbides to oxides, nitrides, and sulfides. Notably, bare metal clusters without ligand atoms display significantly lower mean average AEs than ligand-containing clusters (**Fig. 5C**), underscoring the critical role of metal-ligand interactions in stabilizing cluster aggregates. Beyond single-element trends, mean average AEs were further analyzed for clusters containing specific pairs of metal elements (**Fig. 5D**). Diagonal entries correspond to clusters containing only one metal species, whereas off-diagonal values represent heteronuclear combinations. Several metal

pairs, including Ta-Zr, Ta-Hf, Ta-Nb, and Hf-Zr, exhibit mean average AEs that exceed those of either constituent element alone, indicating enhanced thermodynamic stability arising from heteronuclear mixing. Extending this analysis to ternary compositions, clusters containing combinations of Ta, Zr, Hf, and Os (**Fig. S2**) frequently display even higher mean average AEs than their unary or binary counterparts, suggesting that multi-metal incorporation can further promote thermodynamic stabilization in metal clusters.

Extrapolation Limits toward Larger Metal Clusters. The extrapolation capability of the model was further assessed by testing its predictions for larger clusters beyond the training regime. Owing to the substantially increased computational cost associated with larger systems, the GM structures for clusters containing more than four metal atoms were generated exclusively for testing purposes. Metal clusters with N_M = 5–8 were randomly sampled, and their predicted average AEs were compared against first-principles reference data, as summarized in **Table 1**. A systematic degradation in predictive accuracy is observed with increasing cluster size. The MAE increases from 58.9 meV/atom for N_M = 5 clusters to 116.4 meV/atom for N_M = 8 clusters. While predictions for clusters with N_M = 5–7 remain within a practically acceptable range, the accuracy for N_M = 8 deteriorates markedly, with an MAE more than twice that obtained for N_M = 4 clusters (48.3 meV/atom). This trend indicates a progressive loss of transferability as the cluster size extends beyond the training regime. An analogous analysis with respect to ligand number reveals a similar limitation. The model performs satisfactorily for clusters containing one to four ligands (N_L = 1 – 4) but exhibits a pronounced decline in accuracy for species with five ligands, for which the MAE increases to 93.5 meV/atom. Together, these results suggest that the present model reaches its effective predictive limit for clusters comprising up to approximately eight metal atoms or roughly ten total metal and ligand atoms. To further contextualize these limits, the distributions of predicted average AEs for all species in chemical spaces defined by $N_M \leq$ 1, 2, 3, 4, and 5 were analyzed and are shown in **Fig. S3**. As the upper bound of N_M increases from 1 to 5, the maximum predicted average AE rises from 4.23 to 6.51 eV/atom. This systematic increase reflects the enhanced aggregation tendency of larger metal clusters, consistent with chemical intuition and providing additional validation of the model's physical realism within its application regime.

Electronic Representation as the Origin of CTEN Performance. To elucidate the origin of the

strong predictive performance observed above, the impact of atomic representation and model architecture was systematically examined. In the CTEN framework, each atom is encoded as an atomic vector that reflects its ground-state electronic configuration (see Materials and Methods for details). The influence of descriptor granularity was evaluated by testing CTEN models with progressively refined atomic representations at the final point of the N_M = 4 learning curve [**Fig. 3(D1)**], where all available N_M = 1–3 samples and the maximum number of N_M = 4 samples were included for training. The predictive performance on the 1,500 species N_M = 4 testing set is summarized in **Table 2**. Models employing only coarse atomic information, such as atomic number or periodic table position, exhibit limited predictive capability, yielding large MAEs of 619.3 and 474.7 meV/atom, respectively. Incorporating increasingly detailed electronic-structure information leads to substantial and systematic improvements. Shell-based atomic representations reduce the MAE to 74.6 meV/atom, while subshell-based representations further decrease the MAE to 48.5 meV/atom. The best performance is achieved using orbital-level atomic vectors, which yield a minimal MAE of 48.3 meV/atom. This monotonic improvement demonstrates that fine-grained electronic representations are essential for capturing the intrinsic atomic characteristics governing metal-metal and metal-ligand interactions in cluster systems.

To further benchmark the CTEN framework, several conventional machine-learning algorithms, including gradient boosting regression tree (GBRT), support vector regression (SVR), random forest (RF), and multi-layer perceptron (MLP), were evaluated under the same training and testing conditions corresponding to the final N_M = 4 learning point (**Table 3**). When cluster composition alone was used as the descriptor, all conventional models yielded MAEs exceeding 50 meV/atom. Among them, MLP achieved the best performance (58.4 meV/atom), whereas RF performed the worst (253.6 meV/atom). When orbital-level atomic vectors were adopted for descriptors, the predictive performance of RF and GBRT improved, whereas SVR and MLP exhibited degraded accuracy. For a given descriptor type, the overall performance of these algorithms follows the order MLP > GBRT > SVR > RF. Although the MLP model using composition-based descriptors provides the most accurate predictions among the conventional algorithms, its performance remains inferior to that of CTEN. These comparisons highlight the unique advantage of the CTEN architecture in effectively leveraging atomic-vector representations to model the thermodynamics of chemical complex metal clusters.

Discussion

Utilization Across Diverse Thermodynamic Properties. Beyond the primary application for AE prediction, the robustness of the CTEN framework was further assessed by extending its predictions to additional thermodynamic properties, including reaction energies (REs) with small molecules, adiabatic ionization potentials (AIPs), and adiabatic electron detachment energies (ADEs). Using N_2, O_2, CO, and CO_2 as probe molecules, the predicted REs exhibit MAEs of 268.5, 303.4, 260.9, and 346.2 meV, respectively. Predictions for ADEs and AIPs yielded MAEs of 158.0 meV and 181.4 meV (**Fig. S4**). Although these properties were not explicitly included in model training, the achieved accuracies are comparable to those reported for state-of-the-art machine learning models developed for molecular and metal-containing systems, which typically show MAEs of 0.24–0.40 eV for reaction energies (*35*, *36*) and 0.03–0.30 eV for ionization-related properties (*37*, *38*). Notably, CTEN attains this level of performance despite the substantially greater electronic complexity associated with d-block metal clusters. In this context, the model's accuracy is broadly comparable to that of existing ML frameworks for metal clusters and catalytic surfaces, where reported MAEs commonly fall in the range of 0.10–0.40 eV (*32*, *39*). These results underscore the robustness and transferability of the CTEN architecture in capturing complex interatomic interactions across diverse chemical environments. More broadly, they suggest that composition-based learning with physically grounded atomic representations can provide a viable and scalable route for predicting a wide spectrum of thermodynamic properties in chemically complex systems, extending beyond the specific case of atomization energies.

Comparison with Chemical Space Studies on Molecular and Material Systems. Chemical space exploration using machine learning has been widely pursued across molecular and material domains, yet the present work occupies a distinct niche by focusing on transition-metal clusters. Molecular chemical spaces are typically dominated by light main-group elements (such as C, H, N, O, F, and S), characterized by predominantly covalent bonding and well-defined molecular geometries (*40, 41*). In contrast, material informatics primarily concerns extended periodic lattices with averaged bonding environments and long-range order (*42, 43*). Transition-metal clusters occupy an intermediate regime

between these two limits, exhibiting localized metallic bonding, variable coordination numbers, and pronounced quantum-size effects. Despite their central relevance to catalysis and nanomaterials design, this regime has remained largely inaccessible to systematic chemical-space exploration due to its intrinsic complexity.

A defining challenge of transition-metal clusters lies in their complex electronic structures arising from partially filled d orbitals, which give rise to multiple oxidation and spin states (**Fig. S5**), strong electron correlation, and intricate multi-center bonding motifs. These features render accurate determination of ground-state geometries and energetics computationally demanding, with costs that increase sharply, often near exponentially, with cluster size. As a result, exhaustive first-principles enumeration rapidly becomes infeasible beyond small systems. In this context, machine-learning approaches provide a critical pathway for mapping the thermodynamic landscape of transition-metal clusters, enabling efficient identification of stable compositions and accelerating the discovery of metal-based functional materials.

The atomic vector descriptor introduced in this work represents each atom solely through its elemental and electronic characteristics, including atomic number, periodic table position, valence-shell configuration, and orbital-level features, without relying on explicit geometric, conformational, or isomeric information. This composition-only strategy contrasts sharply with descriptor schemes commonly employed in molecular machine-learning, where structure-resolved representations such as Columb matrix, molecular fingerprints, or graph-based neural networks (e.g., PhysNet) are routinely used to capture detailed geometries but require extensive and computationally intensive conformational sampling (*44*). In materials informatics, composition-based descriptors are often combined with structure-aware graph representations to balance generality and accuracy (*45*). Conceptually, the present approach aligns more closely with materials informatics philosophy by emphasizing compositional trends rather than explicit geometries. This abstraction confers substantial advantages in computational efficiency and transferability, which are particularly valuable for transition-metal clusters where exhaustive structural sampling is impractical, and bonding patterns are often ambiguous. Although this representation cannot distinguish geometric or spin isomers in the manner of fully structure-resolved molecular models, it offers a pragmatic balance between generality and efficiency, effectively bridging molecular and material paradigms for thermodynamic prediction

in metal cluster systems.

Recent years have witnessed remarkable progress in machine learning prediction of energetic properties across molecular and materials domains. In molecular datasets such as QM9, prediction errors for atomization energies have decreased from approximately 340 meV in early deep-learning models to near chemical accuracy (~1 meV) using modern machine-learning interatomic potentials (MLIPs) (*46–48*). In materials informatics, composition-only models such as CrabNet achieve MAEs of ~77 meV/atom for formation energy predictions, while structure-aware frameworks like ALIGNN further reduce errors to 22 meV/atom on the Materials Project dataset (*49, 50*). Within this broader context, the CTEN model developed here achieves an MAE of 48.3 meV/atom for average atomization energies of transition-metal clusters, despite relying solely on compositional information and addressing systems with substantially greater electronic complexity. This performance underscores the robustness and generalizability of CTEN in navigating the composition-driven chemical space of transition-metal clusters governed by intricate d-electron interactions and multi-center bonding. The remaining accuracy gap relative to the most advanced structure-resolved models for molecular systems can be attributed primarily to the absence of explicit geometric and bonding information. Future improvements may therefore benefit from integrating structural descriptors, such as interatomic distances, coordination environments, or graph-derived bonding features, into the atomic vector framework. Such a hybrid architecture would enable joint learning of compositional and geometric effects, potentially enhancing predictive precision while preserving the scalability required for systematic exploration of the vast and diverse chemical landscapes associated with multimetallic clusters and related materials.

Implications for Materials Discovery. The average AE derived from metal clusters serves as a powerful atomic-level source of information for intrinsic bonding strength and local structural robustness. A higher average AE corresponds to a larger energetic penalty for bond cleavage or structural rearrangement, reflecting enhanced intrinsic stability. This interpretation is consistent with condensed-phase behavior, where strong average bonding is generally associated with improved thermal stability. Notably, as shown in **Fig. 6A**, a strong linear correlation (Pearson r = 0.9528) is observed between the average AE of bare homonuclear $M_n^{\pm/0}$ (n = 2–8) clusters and the cohesive energies of the corresponding bulk metals (*51*), defined as the energy required to remove one atom

from the bulk-phase. This correspondence indicates that cluster-level energetics can serve as an effective proxy for assessing thermodynamic stability in condensed-phase materials.

This cluster-level perspective established a direct conceptual link to recent experimental advances in chemically complex materials, particularly those developed for thermal barrier coatings (TBCs). Studies of medium- and high-entropy systems have demonstrated that multicomponent cation disorder, especially involving refractory elements such as Ti, Zr, and Hf, enhances thermal robustness by inducing lattice distortion, increasing phonon scattering, and suppressing sintering (*52, 53*). A representative example is provided by Y-Ta-based medium-entropy ceramic, in which Ti, Zr, or Hf doping leads to reduced thermal conductivity and improved high-temperature stability relative to the undoped phase (*54*). The present cluster analysis reveals a fully consistent atomic-scale trend. The mean average AE of Y-Ta-based clusters increases systematically with chemical complexity, from 4.668 eV/atom for Y-Ta to 4.793 eV/atom for Y-Ta-Ti, 4.935 eV/atom for Y-Ta-Zr, and 4.920 eV/atom for Y-Ta-Hf, ultimately reaching 5.113 eV/atom for the quaternary Y-Ta-Zr-Hf-based cluster (**Figs. 5D** and **S2**). This monotonic increase quantitatively reflects strengthened local bonding and elevated energetic barriers to structural rearrangement, providing a mechanistic explanation for the enhanced thermal stability observed in multicomponent Y-Ta-based ceramics. Together, these results demonstrate that stabilizing effects identified in condensed-phase materials are mirrored at the atomic scale by corresponding increases in cluster average AE, reinforcing the value of cluster energetics as predictive indicators for materials design.

Beyond reproducing established experimental trends, the large-scale exploration also uncovers compositional features that would be difficult to infer from intuition or isolated case studies. In addition to elements commonly associated with "stability islands", such as Ta and Hf, clusters containing the noble metal element Os exhibit unexpectedly high average AEs. Indeed, the five highest-average-AE clusters identified across the entire 9.13-million-species space, $OsTaHf_2O_4^-$, $OsTaHfZrO_4^-$, $OsTaZr_2O_4^+$, $OsTa_2ZrO_4^-$, and $OsTaZr_2O_4^-$, all share a common structural motif (**Fig. 6B**), in which a single Os atom is embedded within a four-metal, four-oxygen framework. The repeated emergence of Os-containing oxide clusters at the top of the AE distribution indicates unusually strong intrinsic bonding in these compositions. Such behavior suggests that oxide-supported Os species may exhibit enhanced resistance to thermal degradation at the atomic scale,

highlighting a class of compositions potentially well-suited for high-temperature catalytic environments. Given the growing interest in Os-based species for electrochemical catalysis (*55, 56*), these cluster motifs may also provide a pathway for stabilizing Os single-atom catalytic sites through appropriate oxide supports. Collectively, these findings illustrate how cluster-level energetic descriptors can reveal non-intuitive yet chemically meaningful design principles, extending beyond conventional heuristics and offering a predictive framework for discovering thermodynamically robust materials and catalysts.

Conclusion

In summary, this work establishes a scalable and data-efficient framework for exploring the thermodynamics of transition-metal clusters across an exceptionally large and chemically complex space. Central to this framework is the demonstration that progressive sampling fundamentally reshapes the data-scaling behavior of chemical-space learning, enabling reliable thermodynamic prediction with a vanishing fraction of explicitly computed structures even as compositional complexity grows exponentially. By integrating automated first-principles calculations with a composition-based machine-learning model, the Cluster Transformer Encoder Network (CTEN), thermodynamic stability can be predicted with near chemical accuracy across millions of cluster compositions. Beyond methodological scalability, the learned thermodynamic landscape reveals chemically interpretable stability trends, including enhanced stabilization through heteronuclear mixing, a quantitative correspondence between cluster atomization energies and bulk cohesive energies, and non-intuitive stabilization motifs involving noble-metal incorporation. These results position metal clusters as a predictive intermediate between molecular chemistry and condensed-phase materials and demonstrate that chemically complex cluster spaces can be systematically understood through composition-driven learning. More broadly, this framework provides a generalizable paradigm for data-driven discovery of robust materials and catalysts in chemically extensible systems.

Methods

Dataset Preparation.

Chemical space definition and dataset construction. Metal-cluster compositions were generated within the chemically constrained space defined in the Results. Briefly, 34 elements were considered, including 30 d-block metals (Sc–Zn, Y–Cd, La, and Hf–Hg) and four nonmetal ligands (C, N, O, and S). For each cluster composition, the number of ligand atoms was restricted not to exceed the number of metal atoms ($N_L \leq N_M$), and the total cluster charge was allowed to be 0, −1, or +1. The chemical space was partitioned into subspaces according to the number of metal atoms (N_M), which served as the organizational basis for sampling and model training. Owing to the rapidly increasing computational cost with cluster size, different N_M subspaces were treated with distinct sampling strategies. All compositions with N_M = 1 and 2 were fully enumerated and processed through the global optimization workflow. For larger clusters, representative sampling was performed: complete subspaces were considered for N_M = 3–5 systems, whereas for clusters with N_M = 6–8, only compositions satisfying $N_M + N_L \leq 8$ were included.

Global minimum structure search. Global minimum (GM) structures were determined using an automated workflow combining a genetic algorithm (GA) with density functional theory (DFT) calculations. The general GA framework follows that described in previous work (*11*), with specific modifications introduced to improve the exploration of spin multiplicity. For each newly generated cluster, the initial spin multiplicity (M) was inherited from the lowest-energy parent structure. To diversify the population, each offspring structure underwent two or three local optimizations with spin multiplicities M, $M - 2$ (for $M \geq 3$), and $M + 2$. This procedure enables efficient convergence toward the joint global minimum in both geometry and spin state. During the GA phase, geometry optimizations were performed using the LanL2DZ basis set (*57*) with relatively loose convergence thresholds to reduce computational cost. Low-energy candidates emerging from the GA search were subsequently refined through successive optimization stages with increasingly stringent convergence criteria. In the final optimization stage, the def2-TZVP basis set was employed for Y–Cd, La, and Hf–Hg atoms, while the TZVP basis set was used for Sc–Zn and all ligand atoms (C, N, O, and S) (*58, 59*). All electronic structure calculations were carried out using the TPSS functional (*60*) as

implemented in the Gaussian 16 program package (*61*). Frequency analyses were performed for all fully optimized structures to confirm that they correspond to true minima on the potential energy surface. The global minimum for each cluster composition was identified as the structure with the lowest sum of electronic and zero-point energies.

Target Values. Atomization energies (AEs) were computed using the optimized global minimum structures and the corresponding reference atomic energies, providing a uniform thermodynamic metric for all clusters included in the dataset. Zero-point energy corrections were included in all energy evaluations. For neutral clusters, the AE was defined as the difference between the sum of the energies of the constituent neutral atoms and the total energy of the cluster.

$$\mathrm{AE}_{\mathrm{neutral}} = \sum_{i}^{N} E_{\mathrm{atom},i} - E_{\mathrm{cluster}}$$

For charged clusters, a consistent definition was adopted to ensure that AE reflects the intrinsic aggregation tendency of the constituent atoms without explicitly including the energetic cost of electron removal or attachment. For cationic clusters, the ionization potential (IP) of each constituent atom was first calculated. The AE was then defined by adding the minimum atomic IP among all constituent elements to the neutral-cluster expression, corresponding to dissociation into neutral atoms plus one atom carrying the positive charge that minimizes the total dissociation energy.

$$\mathrm{IP}_i = E_{\mathrm{cation},i} - E_{\mathrm{atom},i}$$

$$\mathrm{AE}_{\mathrm{cation}} = \sum_{i}^{N} E_{\mathrm{atom},i} + \min_i \left(\mathrm{IP}_i\right) - E_{\mathrm{cluster}}$$

Similarly, for anionic clusters, the electron affinity (EA) of each constituent atom was computed. The AE was obtained by subtracting the maximum atomic EA from the neutral-cluster expression, corresponding to dissociation into neutral atoms plus one atom retaining the excess electron that yields the lowest AE.

$$\mathrm{EA}_i = E_{\mathrm{atom},i} - E_{\mathrm{anion},i}$$

$$\mathrm{AE}_{\mathrm{anion}} = \sum_{i}^{N} E_{\mathrm{atom},i} - \max_{i}\left(\mathrm{EA}_{i}\right) - E_{\mathrm{cluster}}$$

Under this definition, charged clusters dissociate into neutral atoms and a single charged atom, with the charge assignment chosen to minimize the atomization energy. This convention ensures that the AE consistently represents the maximum intrinsic aggregating tendency of a given set of atoms, independent of explicit ionization or electron-attachment contributions.

In addition to atomization energies, other thermodynamic quantities were derived from the computed AEs. Reaction energies (REs) with small molecules were calculated from the known AEs of reactant clusters, product clusters, and the corresponding molecular species. Adiabatic electron detachment energies (ADEs) were obtained from the difference between the AEs of neutral and anionic clusters, combined with the maximum atomic electron affinity. Adiabatic ionization potentials (AIPs) were calculated analogously using the AEs of neutral and cationic clusters together with the minimum atomic ionization potential.

Feature Selection.

Atomic vectors. Atomic representations were constructed based on the ground-state electronic configurations of constituent elements at different levels of descriptive granularity. At the finest level, atomic vectors encode the electron populations of individual spatial orbitals. Because orbitals sharing the same principal and angular quantum numbers are degenerate, their electron populations can be summed to yield subshell-level representations. Further aggregation of subshells produces shell-level atomic vectors, representing progressively coarser descriptions of elemental electronic structure. In addition to electronic-structure–resolved representations, two simpler atomic descriptors were also considered. One is the atomic number, which may be viewed as a scalar proxy that implicitly aggregates information from all occupied electron shells. The other is the position of an element in the periodic table, which encodes elemental identity and trends in valence structure in a discrete, low-dimensional form. To systematically assess the influence of descriptor granularity on model performance, five types of atomic vectors spanning a range from coarse to fine descriptions were evaluated: atomic number, periodic-table position, shell-level, subshell-level, and orbital-level representations. The dimensionality of these vectors ranges from 1 to 56, reflecting the increasing

level of electronic detail. **Table 4** summarizes the definitions of these atomic vectors, using iridium (Ir) as a representative example.

Model Construction.

Architecture of cluster transformer encoder network (CTEN). In the Cluster Transformer Encoder Network (CTEN), each atom is represented by a vector encoding its ground-state electronic configuration, which serves as a compact and physically motivated descriptor of elemental identity. This representation maps discrete atomic species into a continuous embedding space, analogous to embedding techniques commonly used in natural language processing, where discrete tokens are embedded as vectors. As illustrated in **Fig. S1**, a cluster composition is first converted into a sequence of atomic vectors, with one vector assigned to each constituent atom. This sequence is processed by a transformer encoder consisting of two stacked layers with four-head self-attention, enabling the model to capture nonlocal correlations among different atomic species. The output of the transformer encoder is subsequently aggregated using a global average pooling layer to generate a fixed-length feature vector that represents the overall cluster composition. The pooled feature vector is then passed to a fully-connected neural network (FNN) to predict the target property. To ensure discrimination between clusters with identical compositions but different sizes or charge states, the total number of atoms and the overall cluster charge are included as additional scalar inputs to the FNN. The FNN comprises four hidden layers with dimensions of 512, 256, 128, and 64 neurons, respectively, followed by a single linear output node. The model is trained to predict the average AE, defined as the total AE divided by the total number of atoms [AE/(N_M+N_L)].

Baseline models. To benchmark the performance of CTEN, several conventional machine-learning models were also evaluated, including random forest (RF), gradient boosting regression trees (GBRT), support vector regression (SVR), and multilayer perceptron (MLP). For the MLP baseline, the neural network architecture was chosen to be identical to the FNN module used in CTEN to enable a fair comparison. Two types of input descriptors were tested for these baseline models. The first was a composition-based descriptor, represented by the counts of each element present in the cluster together with the overall cluster charge. The second descriptor employed orbital-level atomic vectors. For this representation, atomic vectors of all atoms in a cluster were concatenated into a single input

vector and padded with zeros to ensure consistent dimensionality across clusters of different sizes. The total cluster charge was included as an additional input feature in both descriptor types.

Model training and validation. For each training run, one-ninth of the available training data was randomly selected and reserved as a validation set. CTEN models were trained for 800 epochs, and the model parameters corresponding to the epoch with the lowest mean absolute error (MAE) on the validation set were retained for evaluation. For the baseline models, hyperparameters were optimized by minimizing the validation MAE using standard tuning procedures appropriate to each algorithm.

References

1. P. Jena, A. W. Castleman, Clusters: A bridge across the disciplines of physics and chemistry. *Proc. Natl. Acad. Sci.* **103**, 10560–10569 (2006).

2. L. E. VanGelder, A. M. Kosswattaarachchi, P. L. Forrestel, T. R. Cook, E. M. Matson, Polyoxovanadate-alkoxide clusters as multi-electron charge carriers for symmetric non-aqueous redox flow batteries. *Chem. Sci.* **9**, 1692–1699 (2018).

3. S. G. Duggan, S. M. G. Rabbani, P. Miró, Nucleation roadmap of reduced polyoxovanadate-alkoxide clusters. *Inorg. Chem.* **64**, 3790–3795 (2025).

4. T.-H. Shin, Y. Choi, S. Kim, J. Cheon, Recent advances in magnetic nanoparticle-based multi-modal imaging. *Chem. Soc. Rev.* **44**, 4501–4516 (2015).

5. N. B. Singh, U. Sarkar, A density functional study of chemical, magnetic and thermodynamic properties of small palladium clusters. *Mol. Simul.* **40**, 1255–1264 (2014).

6. O. V. Lushchikova, S. Reijmer, P. B. Armentrout, J. M. Bakker, IR spectroscopic characterization of methane adsorption on copper clusters Cu_n^+(n = 2 – 4). *J. Am. Soc. Mass Spectrom.* **33**, 1393–1400 (2022).

7. H. Schwarz, Doping effects in cluster-mediated bond activation. *Angew. Chem. Int. Ed.* **54**, 10090–10100 (2015).

8. L. Geng, T. Sengupta, X. Li, C. Cui, S. Lin, X.-L. Xu, A. C. Reber, S. N. Khanna, W.-J. Zheng, Z. Luo, Unusually high-spin $Fe_{12}C_{12}^-$ metallo-carbohedrene clusters. *J. Am. Chem. Soc.* **145**, 26908–26914 (2023).

9. P. Ferrari, J. Vanbuel, E. Janssens, P. Lievens, Tuning the reactivity of small metal clusters by heteroatom doping. *Acc. Chem. Res.* **51**, 3174–3182 (2018).

10. G. Zou, L. Huang, W. Liu, J. Hu, X. Xing, Silicon dopants create unique active sites on gold clusters toward reactions with O_2. *J. Phys. Chem. C* **129**, 1206–1215 (2025).

11. X.-L. Ding, Z.-Y. Li, J.-H. Meng, Y.-X. Zhao, S.-G. He, Density-functional global optimization of $(La_2O_3)_n$ clusters. *J. Chem. Phys.* **137**, 214311 (2012).

12. N. Kavka, M. Förstel, K. Pollow, T. Studemund, R. Mitric, O. Dopfer, The electronic structure of the superatom Au_3^+. *J. Am. Chem. Soc.* **147**, 31111–31120 (2025).

13. D. R, T. Sengupta, S. N. Khanna, D. Kumar, Ab initio investigation on the role of ligation of metal chalcogenide clusters for CO oxidation. *ACS Appl. Nano Mater.* **8**, 6223–6233 (2025).

14. A. Fielicke, Probing the binding and activation of small molecules by gas-phase transition metal clusters *via* IR spectroscopy. *Chem. Soc. Rev.* **52**, 3778–3841 (2023).

15. M. Wang, L.-Y. Chu, Z.-Y. Li, A. M. Messinis, Y.-Q. Ding, L. Hu, J.-B. Ma, Dinitrogen and carbon dioxide activation to form C–N bonds at room temperature: A new mechanism revealed by experimental and theoretical studies. *J. Phys. Chem. Lett.* **12**, 3490–3496 (2021).

16. Y. Li, F. Müller, W. Schöllkopf, K. R. Asmis, J. Sauer, Gas-phase mechanism of $O^{\cdot-}/Ni^{2+}$-mediated methane conversion to formaldehyde. *Angew. Chem. Int. Ed.* **61**, e202202297 (2022).

17. S.-T. Yan, Z.-C. Long, X.-L. Xu, H.-G. Xu, W.-J. Zheng, Anion photoelectron spectroscopy and quantum chemical calculations of bimetallic niobium–aluminum clusters $NbAl_n^{-/0}$(n = 3 – 12): identification of a half-encapsulated symmetric structure for $NbAl_{12}^-$. *Phys. Chem. Chem. Phys.* **25**, 6498–6509 (2023).

18. D. M. Neumark, Spectroscopy of radicals, clusters, and transition states using slow electron velocity-map imaging of cryogenically cooled anions. *J. Phys. Chem. A* **127**, 4207–4223 (2023).

19. S. Takeda, H. Kaneko, K. Funatsu, Chemical-space-based de Novo design method to generate drug-like molecules. *J. Chem. Inf. Model.* **56**, 1885–1893 (2016).

20. J. Wang, J. Mao, M. Wang, X. Le, Y. Wang, Explore drug-like space with deep generative models. *Methods* **210**, 52–59 (2023).

21. Z. Zheng, F. Florit, B. Jin, H. Wu, S. Li, K. Y. Nandiwale, C. A. Salazar, J. G. Mustakis, W. H. Green, K. F. Jensen, Integrating machine learning and large language models to advance exploration of electrochemical reactions. *Angew. Chem. Int. Ed.* **64**, e202418074 (2025).

22. I. Kevlishvili, R. G. St. Michel, A. G. Garrison, J. W. Toney, H. Adamji, H. Jia, Y. Román-Leshkov, H. J. Kulik, Leveraging natural language processing to curate the tmCAT, tmPHOTO, tmBIO, and tmSCO datasets of functional transition metal complexes. *Faraday Discuss.* **256**, 275–303 (2025).

23. B. Basha, M. H. Tahir, T. Kadyrov, N. S. Alsaiari, M. S. Al-Buriahi, Machine learning assisted prediction of melting points of non-fullerene electron acceptors, chemical space generation and visualization. *Chem. Eng. Sci.* **319**, 122258 (2026).

24. F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, R. Armiento, Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).

25. S. Tang, H. Yuan, R. Wang, D. Xu, Accelerating the discovery of stable metal–organic frameworks for ethane/ethylene separation using data mining and multitarget screening. *Sep. Purif. Technol.* **376**, 133864 (2025).

26. B. Ma, C. Dang, J. Song, Z. Chen, Y. Zhou, Machine learning–accelerated discovery of novel high-entropy spinel oxide cathodes for solid oxide fuel cells. *Fuel* **404**, 136190 (2026).

27. Y.-F. Hu, J.-L. Li, Z. Guo, Z.-P. Liu, Y.-F. Li, Global potential energy surface exploration of In–Ga–Zn–

O (IGZO) reveals Ga-rich layered channel material. *J. Phys. Chem. Lett.* **16**, 11851–11860 (2025).

28. H. Park, A. Onwuli, A. Walsh, Exploration of crystal chemical space using text-guided generative artificial intelligence. *Nat. Commun.* **16**, 4379 (2025).

29. C. Wen, H. Shen, Y. Tian, G. Lou, N. Wang, Y. Su, Accelerated discovery of refractory high-entropy alloys for strength-ductility co-optimization: An exploration in Nb/Ta/Zr/Hf/Mo system by machine learning. *Scr. Mater.* **252**, 116240 (2024).

30. H. Zhai, A. N. Alexandrova, Ensemble-average representation of Pt clusters in conditions of catalysis accessed through GPU-accelerated deep neural network fitting global optimization. *J. Chem. Theory Comput.* **12**, 6213–6226 (2016).

31. Q. Yang, Z.-Y. Li, P. L. Rodríguez-Kessler, S.-G. He, Accelerated global optimization of metal cluster structures via the deep neural network complemented with transfer learning and genetic algorithm. *Chin. J. Chem. Phys.* **37**, 321–329 (2024).

32. L. Fu, Q. Du, L. Sai, J. Zhao, Accelerating global search of large-sized silver clusters using cluster graph attention network. *J. Phys. Chem. Lett.* **15**, 9160–9166 (2024).

33. X.-G. Zhao, Q. Yang, Y. Xu, Q.-Y. Liu, Z.-Y. Li, X.-X. Liu, Y.-X. Zhao, S.-G. He, Machine learning for experimental reactivity of a set of metal clusters toward C–H activation. *J. Am. Chem. Soc.* **146**, 12485–12495 (2024).

34. L.-H. Mou, G.-D. Jiang, C. Wang, X. Cheng, J. Yu, Z.-Y. Li, J. Jiang, Unveiling universal reactivity descriptors of metal clusters toward dinitrogen activation: A machine learning protocol with three-level feature extraction. *ACS Catal.* **15**, 6618–6627 (2025).

35. X. Kang, W. Zhou, Z. Li, Z. Chu, H. Yin, S. Gao, A. Du, X. Duan, Prediction of CO_2 reduction reaction intermediates and products on transition metal-doped r-GeSe monolayers: A combined DFT and machine learning approach. arXiv:15710 [physics.chem-ph] (2025).

36. P. Van Gerwen, K. R. Briling, C. Bunne, V. R. Somnath, R. Laplaza, A. Krause, C. Corminboeuf, 3DReact: Geometric deep learning for chemical reactions. *J. Chem. Inf. Model.* **64**, 5771–5785 (2024).

37. R. Zubatyuk, J. S. Smith, B. T. Nebgen, S. Tretiak, O. Isayev, Teaching a neural network to attach and detach electrons from molecules. *Nat. Commun.* **12**, 4870 (2021).

38. F. Pereira, Machine learning for the prediction of ionization potential and electron affinity energies obtained by density functional theory. *ChemistrySelect* **8**, e202300036 (2023).

39. S. Pablo-García, S. Morandi, R. A. Vargas-Hernández, K. Jorner, Ž. Ivković, N. López, A. Aspuru-Guzik, Fast evaluation of the adsorption energy of organic molecules on metals via graph neural networks. *Nat. Comput. Sci.* **3**, 433–442 (2023).

40. B. Sanchez-Lengeling, A. Aspuru-Guzik, Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).

41. J.-L. Reymond, The Chemical Space Project. *Acc. Chem. Res.* **48**, 722–730 (2015).

42. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

43. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).

44. S. Wang, R. Zhang, X. Li, F. Cai, X. Ma, Y. Tang, C. Xu, L. Wang, P. Ren, L. Liu, S. Wu, Q. Qian, F. Bai, Recent advances in molecular representation methods and their applications in scaffold hopping. *npj Drug Discov.* **2**, 14 (2025).

45. J. Damewood, J. Karaguesian, J. R. Lunger, A. R. Tan, M. Xie, J. Peng, R. Gómez-Bombarelli, Representations of materials for machine learning. *Annu. Rev. Mater. Res.* **53**, 399–426 (2023).

46. O. A. Von Lilienfeld, K.-R. Müller, A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).

47. D. P. Kovács, I. Batatia, E. S. Arany, G. Csányi, Evaluation of the MACE force field architecture: From medicinal chemistry to materials science. *J. Chem. Phys.* **159**, 044118 (2023).

48. N. J. Williams, L. Kabalan, L. Stojanovic, V. Zólyomi, E. O. Pyzer-Knapp, Hessian QM9: A quantum chemistry database of molecular Hessians in implicit solvents. *Sci. Data* **12**, 9 (2025).

49. L. Jin, Z. Du, L. Shu, Y. Cen, Y. Xu, Y. Mei, H. Zhang, Transformer-generated atomic embeddings to enhance prediction accuracy of crystal properties with machine learning. *Nat. Commun.* **16**, 1210 (2025).

50. K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 185 (2021).

51. C. Kittel, *Introduction to Solid State Physics* (J. Wiley & sons, New York, 8th ed., 2005).

52. L. Chen, B. Li, J. Feng, Rare-earth tantalates for next-generation thermal barrier coatings. *Prog. Mater. Sci.* **144**, 101265 (2024).

53. J. H. Kim, S. Yoon, D. S. Baek, J. Kim, J. Kim, K. An, S. H. Joo, Boosting thermal stability of volatile Os catalysts by downsizing to atomically dispersed species. *JACS Au* **2**, 1811–1817 (2022).

54. X. Chen, L. Chen, J. Tian, C. Xu, J. Liao, T. Li, J. Wang, J. Feng, Thermal and mechanical properties of MO_2 (M = Ti, Zr, Hf) co-doped $YTaO_4$ medium-entropy ceramics. *Int. J. Miner. Metall. Mater.* **32**, 1441–1450 (2025).

55. W. Xue, Q. Zhou, P. Wang, S. Zuo, F. Li, J. Jiang, F. Mo, C. Yin, G. Liu, Z. Ban, Y. Wei, W. An, X. Huang, H. Wang, S. Zhan, Sulfur-mediated transformation from osmium nanocrystals to single atoms for efficient alkaline hydrogen evolution reaction. *Proc. Natl. Acad. Sci.* **122**, e2426207122 (2025).

56. Y. Sun, Z. Qian, M. Sun, Y. Li, Q. Zhang, Y. Nie, L. Gu, M. Luo, J. Liu, B. Huang, S. Guo, Osmium atomic sites on CuS nanoplates for efficient two-electron oxygen reduction into H_2O_2. *Chem* **11**, 102393 (2025).

57. P. J. Hay, W. R. Wadt, *Ab initio* effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.* **82**, 299–310 (1985).

58. F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297 (2005).

59. A. Schäfer, C. Huber, R. Ahlrichs, Fully optimized contracted Gaussian basis sets of triple zeta valence

quality for atoms Li to Kr. *J. Chem. Phys.* **100**, 5829–5835 (1994).

60. J. Tao, J. P. Perdew, V. N. Staroverov, G. E. Scuseria, Climbing the density functional ladder: Nonempirical meta–generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **91**, 146401 (2003).

61. Frisch, M. J., Trucks, G. W., Schlegel, H. B., *et al.*, Gaussian 16, Revision C.01; Gaussian, Inc.: Wallingford, CT, 2016.

Acknowledgements

Funding: This work was supported by the National Natural Science Foundation of China (Nos. 92461313, 92570205, 22527804, and 22173111) and the Center for Carbon Neutral Chemistry, ICCAS (CCNC-202501).

Author contributions: S.G.H. conceptualized the project. Z.Y.W., N.Z.L., and S.G.H. developed computational and modelling programs. N.Z.L. performed the computations and modelling. N.Z.L., Z.Y.L., Q.Y.L., Q.S., and S.G.H. analyzed the results. N.Z.L. and Z.Y.L. wrote the manuscript. Z.Y.L. and S.G.H. revised the manuscript.

Competing interests: All other authors declare they have no competing interests.

Data and materials availability: The Supporting Information (SI) is available free of charge. Details of CTEN model architecture and additional results are given in SI. The codes and computational results used in this work can be found on GitHub at https://github.com/IridiumLINCH-SK/iccas-cluster-cten.

Figures and Tables

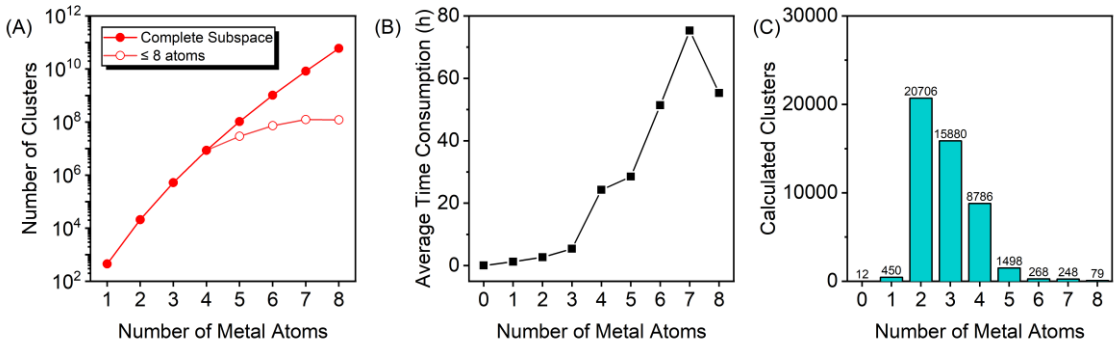


Fig. 1. Subspaces of metal clusters and the computational costs. (A) Exponentially expanding numbers of metal clusters in our study. (B) Average core time consumption for determining the GM structure of a cluster. (C) Numbers of metal clusters computationally determined in the database.

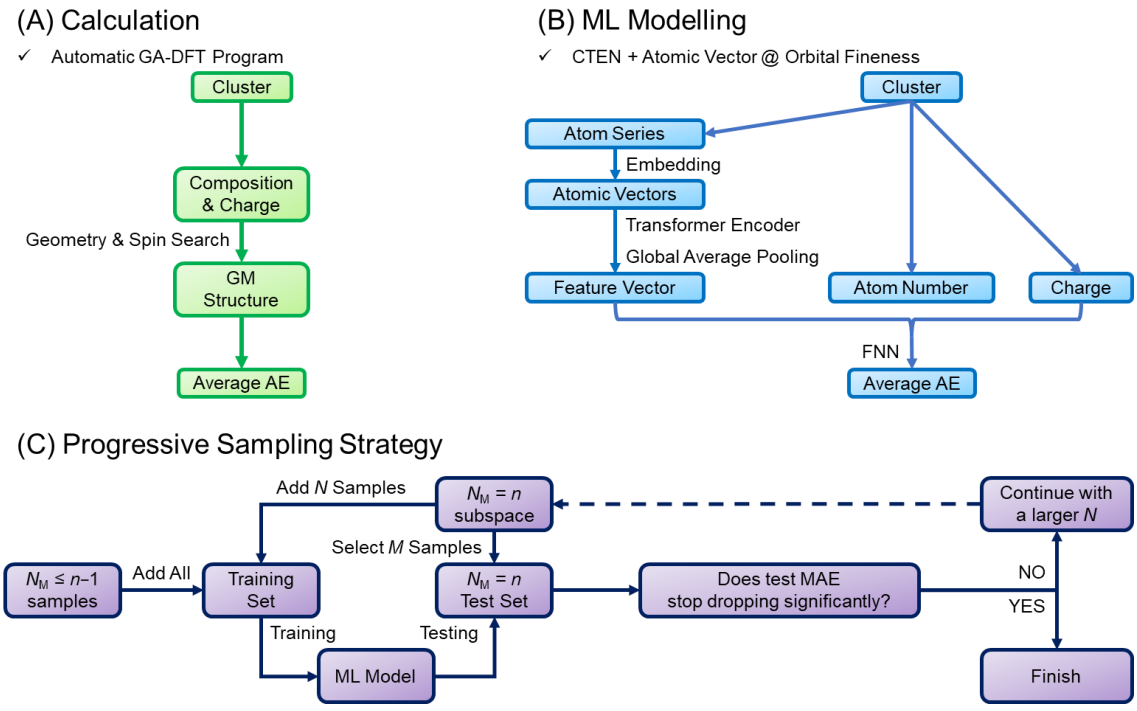


Fig. 2. Workflows for exploring the chemical space of metal clusters. (A) Calculation and (B) ML modelling for the average AE of metal clusters. (C) The progressive sampling strategy.

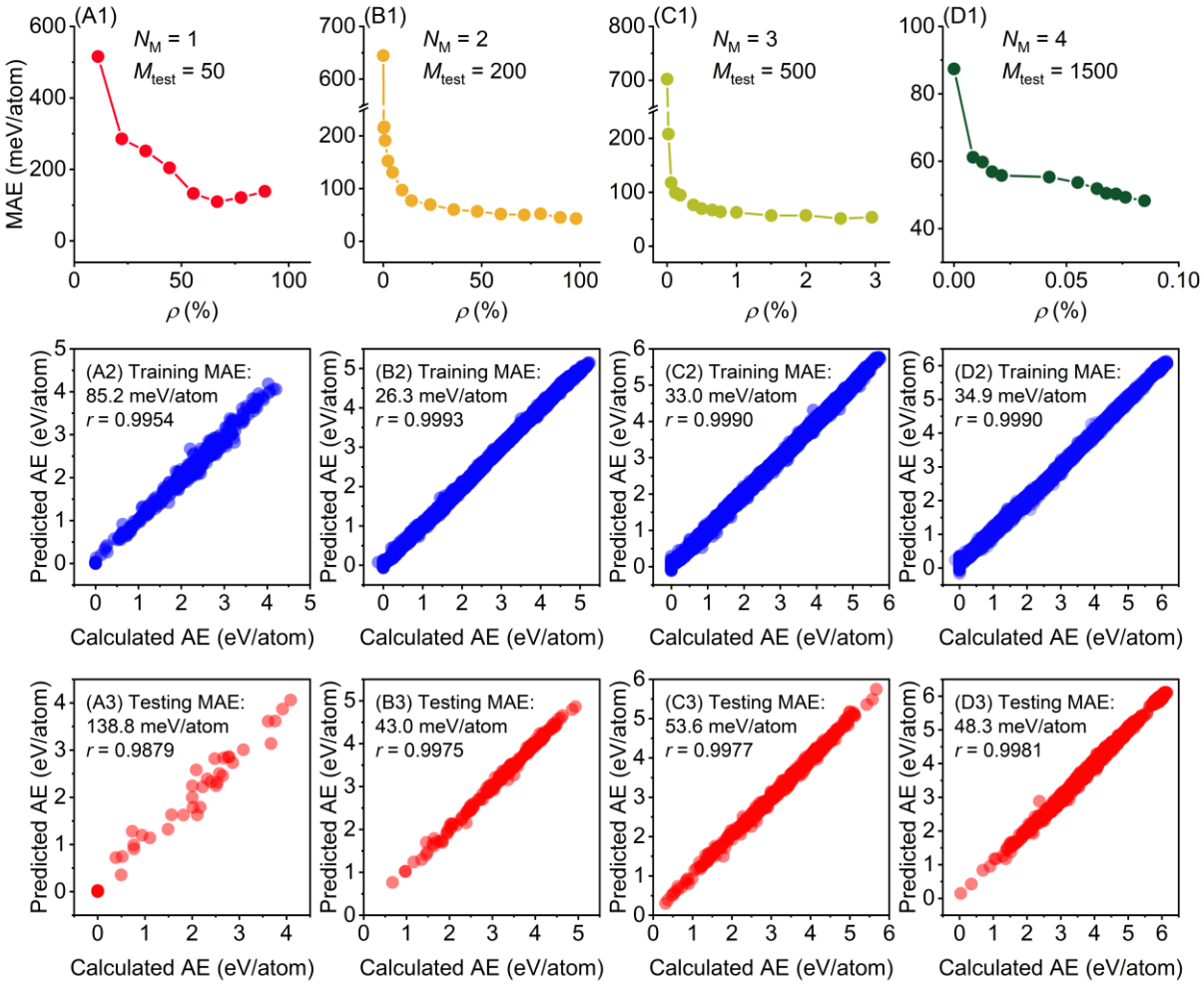


Fig. 3. Learning curves in $N_M = 1 – 4$ subspaces. (A1 – D1) ρ is defined as the fraction of training samples to the size of the subspace. The scatter plots of training sets (A2 – D2) and testing sets (A3 – D3) of the last point in learning curves are also shown.

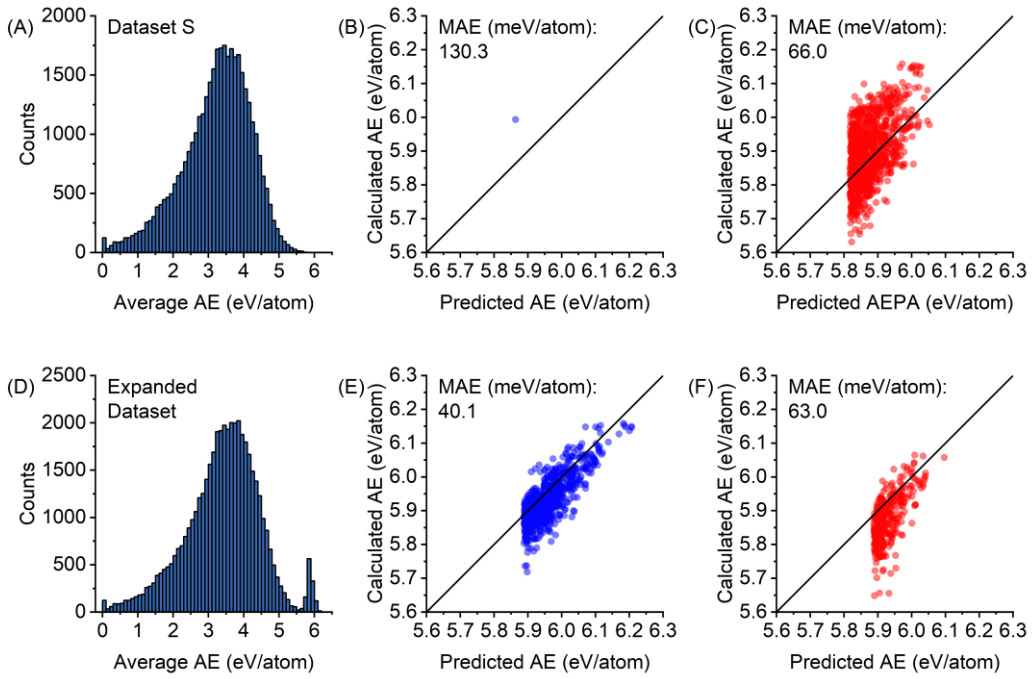


Fig. 4. Effect of high-average-AE-directed sampling. Calculated average AE distribution of (A) a relatively small dataset (dataset S) and (D) an expanded dataset. Calculated and predicted samples of the top 1000 average AE in the $N_M = 1 – 4$ space from the model trained on dataset S, in which data in (B) is present (1 sample) and data in (C) is not present (999 samples). Calculated and predicted samples of the top 1000 average AE in the $N_M = 1 – 4$ space from the model trained on the expanded dataset, in which data in (E) is present (698 samples) and data in (F) is not present (302 samples).

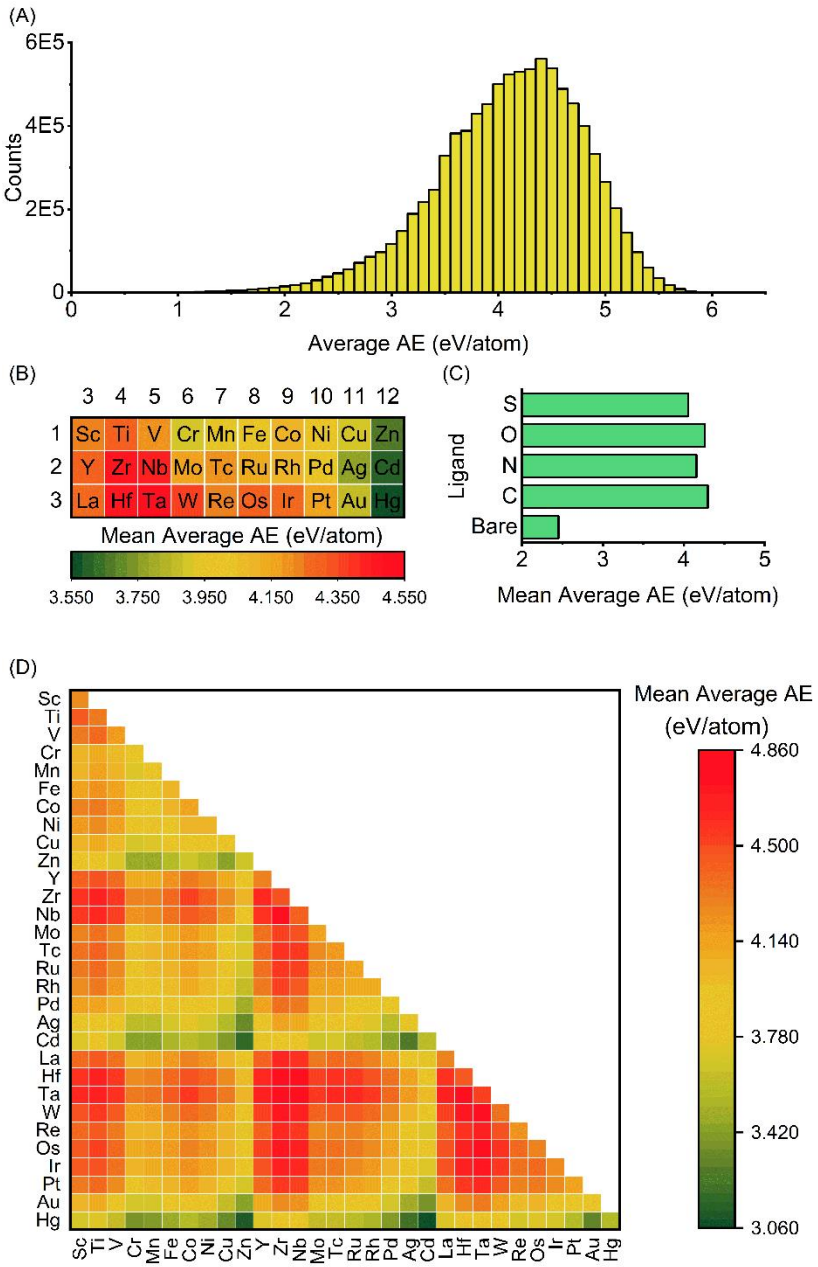


Fig. 5. Thermodynamic landscapes in the metal-cluster space. (A) Distribution of predicted average AE of all clusters of $N_M \leq 4$. (B) Mean values of predicted average AE values of clusters containing the specific d-block metal element with at least one atom. (C) Mean values of predicted average AE values of clusters of bare metal clusters and clusters containing the specific ligand element with at least one atom. (D) Mean average AE values of clusters containing two specific d-block metal elements.

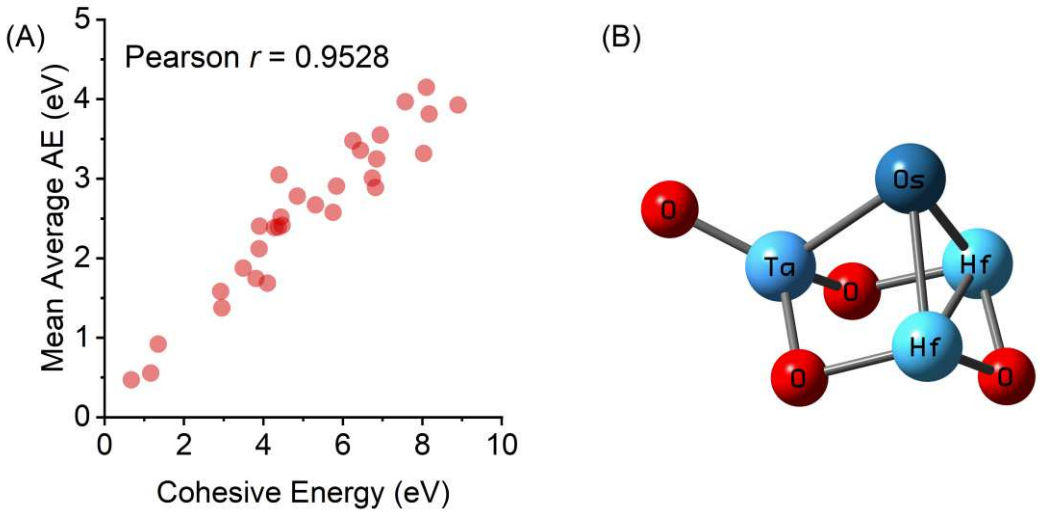


Fig. 6. Connection between metal clusters and bulk phase materials. (A) Correlation between mean average AE of bare metal clusters $M_{2-8}^{0/+/-}$ and the cohesive energy of bulk-phase metals. (B) Geometric structure of $OsTaHf_2O_4^-$, the species with the highest average AE among the DFT calculations of ML predicted top-1000 metal clusters for $N_M \leq 4$.

Table 1. Testing MAEs (meV/atom) of larger metal clusters.

N_M	Overall	$N_L = 0$	$N_L = 1$	$N_L = 2$	$N_L = 3$	$N_L = 4$	$N_L = 5$
5	58.9	72.2 (170)*	51.6 (702)	49.4 (161)	51.2 (149)	57.0 (146)	93.5 (170)
6	67.6	75.5 (85)	68.5 (115)	56.1 (68)	-	-	-
7	81.0	76.9 (103)	83.9 (145)	-	-	-	-
8	116.4	116.4 (79)	-	-	-	-	-

* Numbers of samples are labeled in parentheses.

Table 2. MAE (meV/atom) and Pearson coefficient (r) on the 1500 $N_M = 4$ testing set of the CTEN model using different atomic descriptors.

Fineness	MAE	r
Atomic number	619.3	0.6738
Periodic table position	474.7	0.8065
Shell	74.6	0.9954
Subshell	48.5	0.9980
Orbital	48.3	0.9981

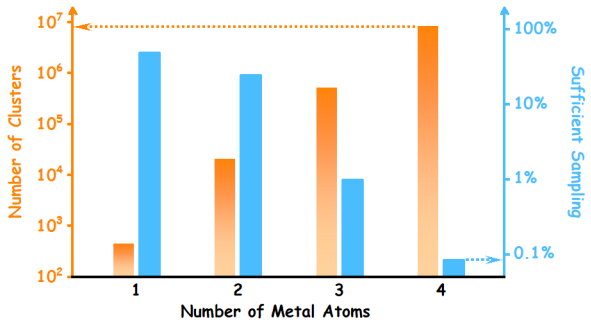
Table 3. MAE (meV/atom) and Pearson coefficient (r) on the 1500 $N_M = 4$ testing set of conventional ML algorithms with different descriptors.

ML algorithm	Composition		Orbital-Level Atomic Vectors	
	MAE	r	MAE	r
MLP	58.4	0.9978	73.7	0.9960
GBRT	134.4	0.9835	79.8	0.9936
SVR	144.5	0.9808	153.5/0.9759	0.9759
RF	253.6	0.9584	193.0/0.9665	0.9665

Table 4. Atomic vectors of an element. e.g. Ir, in different fineness.

Fineness	Description	Vector
Atomic number	Total electron number of the neutral atom	77
Periodic table position	Row and column position of the element in the periodic table	6,9
Shell	Population of electrons in K, L, M, N, O, P, Q shells	2,8,18,32,15,2,0
Subshell	Population of electrons in subshells (1s, 2s, 2p, 3s, 3p, 3d, 4s, 4p, 4d, 4f, 5s, 5p, 5d, 5f, 6s, 6p, 6d, 7s)	2,2,6,2,6,10,2,6,10,14,2,6,7,0,2,0,0,0
Orbital	Population of electrons in individual orbitals (1s, 2s, $2p_{1-3}$, 3s, $3p_{1-3}$, $3d_{1-5}$, 4s, $4p_{1-3}$, $4d_{1-5}$, $4f_{1-7}$, 5s, $5p_{1-3}$, $5d_{1-5}$, $5f_{1-7}$, 6s, $6p_{1-3}$, $6d_{1-5}$, 7s)	2,2,2,

Graphical abstract



Supplementary Files

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

- SupportingInformationNatChem02024.pdf