Data-driven candidate structure ensemble generation beyond the convex hull with crystALL approach

Adam M. Krajewski [a ⋆ †], **Ricardo Amaral** [b ⋆], Jonathan W. Siegel [c], Luke A. Myers [a], Rushi Gong [a ‡], Shun-Li Shang [a], Zi-Kui Liu [a]

a. Department of Materials Science and Engineering, The Pennsylvania State University, USA

b. Department of Energy and Mineral Engineering, The Pennsylvania State University, USA

c. Department of Mathematics, Texas A&M University, USA

⋆ equal contributions | † now at Google | ‡ now at Microsoft

Corresponding Author: A. M. Krajewski | PO Box 60022, Sunnyvale, CA 94088 | adam@phaseslab.org

November 2, 2026

Abstract

Real materials consist of multiple stable or metastable states under real world conditions, existing as statistical ergodic ensembles of configurations rather than ideal crystals. Yet, data-driven computational materials discovery increasingly biases itself toward finding 'new stable structures' based *only* on DFT ground-state calculations. This can limit our long-term ability to discover materials whose properties emerge from metastable phases, explain experimental observations, or model thermodynamic behavior comprehensively. Predicting ergodic sets of atomic configurations of existing and hypothetical solids is a fundamental challenge in theoretical crystallography, approached through many complementary methods. We discuss this conceptually and take a step forward by introducing crystALL, a pragmatic and powerful data-driven approach to the discovery of atomic configuration ensembles beyond the convex hull, embracing structural diversity over stability alone, thus enabling robust experiment matching in novel cases evading traditional methods. crystALL leverages millions of relaxed structures from Material-Property-Descriptor Database (MPDD) and OPTIMADE, to rapidly generate a comprehensive set of candidates. It achieves this by: (1) species substitutions; (2) machine learning (ML) feature space analysis with pySIPFENN; (3) representative structure selection through clustering and energy-minimization or property matching; and (4) construction of final candidates ensemble for direct validation against experiments or ab initio calculations. Furthermore, we uniquely address the common experimental challenge where experimental composition measurements carry uncertainty spanning many integer stoichiometries, introducing novel and efficient searches across anonymized compositional ranges rather than fixed formulas. Overall, crystALL's ensemble-focused philosophy provides critical advantages over the prevailing convex-hull-structures-searching paradigm, which often cannot account for or miss important experimental observations and increases AI/ML discovery bias over time - akin to model collapse seen in Large Language Models (LLMs) when retrained on iteratively generated content.

Contents

1 New Stable Structure Prediction? 1

2 crystALL Approach 3

2.1 Searching for Experimentally Observed Target Stoichiometries 3

2.2 Searching for Composition Ranges 6

3 Case Studies 7

3.1 Identifying $NdBi_2$ and Other Structures 7

3.2 Guided Search of Novel Hydrides 8

3.3 Catalysis Materials and Molten Salts 9

4 Summary 9

5 Software and Data Availability 11

1. New Stable Structure Prediction?

The ability to determine the low-energy atomic configurations of existing and hypothetical solids formed in chemical systems is one of the fundamental challenges in theoretical crystallography and in materials science as a whole. For the last several decades, the two remaining challenges are predictions of (a) compositions forming compounds of interest and (b) candidate topologies or approximate atomic arrangements they can form [1]. Given these two, the exact structure can be solved experimentally, by matching X-ray diffraction profile to candidates using the method by Le Bail *et al.* [2] and interpreting the profile, or more recently using deep learning [3]–

[5] including initial crystallographic indexing via machine learning (ML) methods [6]. Similar analysis can be done with different experiments, such as the Raman spectrum analysis [7]. For hypothetical compounds in exactly defined states, the corresponding atomic configuration can also be solved using ab initio quantum mechanical methods, such as Kohn-Sham density functional theory (DFT) [8], which are computationally expensive but can provide accurate results and continue to improve [9], [10].

The latter problem of proposing the topologies has been traditionally approached by finding geometries satisfying a set of constraints, like the "principles" systematized by Pauling [11], which included then-novel concepts like coördination [*sic*]. Over decades, this approach has evolved into automated computational tools like the `GRINSP` by Le Bail [12] which can propose such structures. More recently, increasing computational power enabled the shift in focus towards testing many options, with some of the methods starting from random arrangements and collapsing them into local minima of the high-dimensional energy landscape, like `AIRSS` by Pickard and Needs [13], or by evolutionary techniques, like `USPEX` by Glass *et al.* [14], `CALYPSO` by Wang *et al.* [15], or `XtalOpt` by Hajinazar and Zurek [16].

Several efforts have also been made to tackle both topological and compositional challenges simultaneously, such as high-throughput construction of the `AFLOW`'s [17] DFT Database [18], [19] (and several others [20]–[25]) built by systematically populating the earlier, yet continuously growing, libraries of prototype structures [26], [27] with new chemistries based on the expert knowledge. However, even for a limited set of elements, attempting to completely explore the materials space becomes combinatorially challenging, if not intractable, due to the vast number of possible atomic configurations these elements can form [28], which is severely amplified by the numerous ways complex structures can be occupied [29]. This issue has prompted recent efforts into (1) systematic similarity-driven prediction of substitutions which would result in finding new low-energy structures [30]; (2) large-scale brute-force substitutions into prototypes, then filtered by ML models [31]; and (3) increasingly more advanced approaches that combine these strategies, such as ones by Ye *et al.* [32], Merchant *et al.* [33], or Zeni *et al.* [34] leading to the rapid discovery of new structures *predicted* to be *stable against a known set* of competing atomic arrangements *at absolute zero and in the absence of any external fields or perturbations*.

Despite these advances in finding new *candidates* for ground-state configurations across many chemical systems, which certainly open the world for new discoveries, key questions remain regarding the inherent limitations of such methods that must be considered when interpreting their results.

Firstly, *"To what degree have we biased the discovery of new structures?"* through many iterations of biasing the search to our expectations, which is philosophically analogous to how retraining Large Language Models (LLMs) and other ML models on generated data makes them underestimate improbable events causing so-called *Model Collapse* by Shumailov *et al.* [35], and drift away from the distribution of the real data [36]. In the context of materials, the latter is depicted well by Fig. 5 in Ye *et al.* [32] showing relatively small overlap between experimentally known structures from ICSD and new ones predicted to be stable - something that can be attributed to either (a) novelty or (b) sampling in regions where there is less energetic competition from known low-energy structures so many iterations of "new stable compounds" can be found with increasing focus. Similar concerns were raised by Cheetham and Seshadri [37], who examined a sample of generatively-predicted "new stable" compounds and found many to be known structure types, trivial substitutions to complete the task, or samples in chemically implausible regions, questioning the novelty and practical value of such predictions driven by the stability alone.

Secondly, even in the case of the results domain becoming strongly biased towards the true solutions of stable configurations, one needs to also ask *"Do we actually need the diverse energy landscapes to discover new materials?"* to which the answer is a multifaceted yes. The metastable materials are critical to evaluating thermodynamic model parameters, as so-called protodata, which was highlighted in the study by Olson and Liu [38]. That study has also stressed that computational thermodynamics techniques generally model *phases* which, at finite temperatures, represent statistical ensembles of underlying atomic configurations within a material.

Furthermore, one has to consider that going beyond individual highly-stable configurations is needed to create a realistic model matching experiments, as concisely explained by Zunger [39], because of the several complexities involved in the synthesis of new materials, such as defect modeling or disorder contributions. The latter necessitates accounting for entropic contributions, especially in compositionally complex materials, as shown by Divilov *et al.* [40] using ensembles of metastable partially representative configurations to describe high entropy ceramics and match experiments accurately; or by Grau-Crespo *et al.* [41] showcasing the need to account for complex compositional fluctuations in otherwise relatively simple materials.

Once such realistic models are obtained, the coexis-

tence of many configurations can be used to accurately (a) tune material properties, including to extreme behavior, such as infinitely negative thermal expansion predicted by Liu *et al.* [42] that, critically, matched the experiments without any fitting parameters, and (b) model complex phenomena, such as superconductivity [43]. Thus, to achieve, or at least approximate ergodicity, efforts to generate ensembles of candidate structures must sample diverse populations of polymorphs, such as the one shown in Fig. 6 in Pickard and Needs [13] generated through AIRSS for the simple case of pure silicon.

Combined together, the existing materials discovery approaches mentioned throughout this section cover prediction of (1) focused, narrow sets of likely (to be observed) candidates over singular targeted chemistries based on expert intuition and DFT calculations, such as USPEX [14], AIRSS [13], or XtalOpt [16]; (2) narrow sets of *individually* unlikely candidates based on DFT for many chemistries in large-scale mostly unbiased enumerations, like AFLOW [18]; (3) narrow sets of individually more likely candidates for many chemistries, although often less likely than expected due to pronounced stability bias, based on ML and high-throughput enumerations [37], such as GNoME [33]; and (4) wide sets of candidates typically for a single targeted material, considered together and approximately ergodic, created based on the expert intuition and systematic enumeration of configurations, like Zentropy approach by Liu *et al.* [42]. These leave out the essential niche of wide sets of candidates for many chemistries, which remains, as mentioned earlier, infeasible to be solved by DFT or experiments alone, yet it can be approached through fully-data-driven DFT/experiment-validated ML.

2. crystALL Approach

Generating a wide set of diverse candidates requires the opposite philosophy from one that biases the candidate population toward stability based on expert intuition - human or machine alike. Rather, all possibilities coming from historically realized or relaxed structures should be considered during the initial steps. In this way, the search is not constrained by prior assumptions about stability, but is instead allowed to reveal unexpected regions of configurational space.

This work presents a concrete approach and associated open-source software tool based in this philosophy, called crystALL, which does so by unbiased consideration of all past 3D periodic atomic structures (i.e., crystals) data, including different deviations from ideal positions from DFT-based relaxations in different chemical systems, in an attempt to accommodate targeted composition or its range into it. This is akin to populating prototypes from a library like the AFLOW *Library of Crystallographic Prototypes* [26], [27]; however, instead of a highly curated set of around 2,000 (as of March 2026) prototypes with ideal sites and commonly occurring crystallographic parameters, three orders of magnitude more prototype atomic configurations are used, including denser sampling of crystallographic free parameters determining ideal site positions, as well as symmetry-broken configurations, which have been metastable for other chemistries.

At its core, crystALL first connects to an expansive collection of configurations in the MPDD [44], [45], developed by the authors with this application in mind, which spans over 4.9 million *relaxed* structures (as of March 2026) coming from (1) experimental observations processed through Crystallography Open Database (COD) developed by Gražulis *et al.* [46]–[48], and (2) various DFT calculations through OQMD [20]–[22], AFLOW [17]–[19], Materials Project [23], [49], NIST-JARVIS [25], Alexandria [24], CAMD [32], GNoME [33], and in-house DFT calculations done by the authors. All data in MPDD can be accessed both through a limited-access high-performance API and through an open-access OPTIMADE API [45] endpoint at optimade.mpdd.org. Furthermore, MPDD also includes an additional "community" dataset, accessible by crystALL, which also accepts purely hypothetical atomic configurations, providing a pipeline for systematically proposing and sharing unconventional materials, such as superionic phases of different dimensionalities, as in Liu *et al.* [50], usually not covered by other databases. These configurations were of interest to the authors and different external contributors, but were not calculated through ab initio methods. These include different disordered occupancies (e.g. SQS by Zunger *et al.* [51] and Walle *et al.* [52]) or hypothetical structurally complex polymorphs. Finally, one can also add prototypes to crystALL locally, after constructing them in any fashion of choice, such as systematic generation of atomic configurations with aflow++ based on an AFLOW prototype selected using --proto option adjusted by scanning through ranges of prototype-specific free parameters set by --params option [53].

2.1 Searching for Experimentally Observed Target Stoichiometries

The combined pool of diverse atomic configurations provides the starting-point *prototypes* for the rest of the crystALL implementation, depicted schematically in Figure 1 for compounds with a defined stoichiometry and

expanded upon later in subsection 2.2 to cover compositional ranges.

Given a target chemical composition/stoichiometry, which can be anonymized to, for instance, $ABCD_3$, `crystALL` first filters incoming `MPDD` data to match it by selecting the appropriate number of components (i.e., 4) and component ratios (i.e., 1:1:1:3). Then, using atomic structure representations and manipulation routines from `pymatgen` [54], it generates *all* possible permutations of how atoms in the prototype can be substituted with target chemical elements. Depending on the ratios, this can be just 1 for compounds with non-repeating ratios (e.g., in Sec. 3.1), 2 for a single matching ratio (e.g., in Sec. 3.2), 6 for the $ABCD_3$ example, highlighted in the orange-colored part of Figure 1, or many more, such as 48 permutations possible in complex compounds like $Ba_2CaCeGd_2Ti_2(CuO_7)_2$ (mp-1229095). At this stage, prototypes become an ensemble of *candidates* that can be passed into AI/ML and other materials informatics frameworks.

In addition to restricting searches by target anonymized composition, users may refine the prototype pool using any queryable field within `MPDD`, including dimensionality, crystal system, or any combination of MPDD fields. For example, `crystALL` may be used with `MPDD` to identify two-dimensional prototypes with a given composition, providing a route to construct wide and diverse datasets of two-dimensional materials. `MPDD` supports this functionality using two popular algorithms by Cheon *et al.* [55] and by Larsen *et al.* [56], which assign a label (`0D` / `1D` / `2D` / `3D`) corresponding to the identified dimensionality of material (Cheon) or an array of labels identifying dimensionality of all features (Larsen). `MPDD` provides a field combining

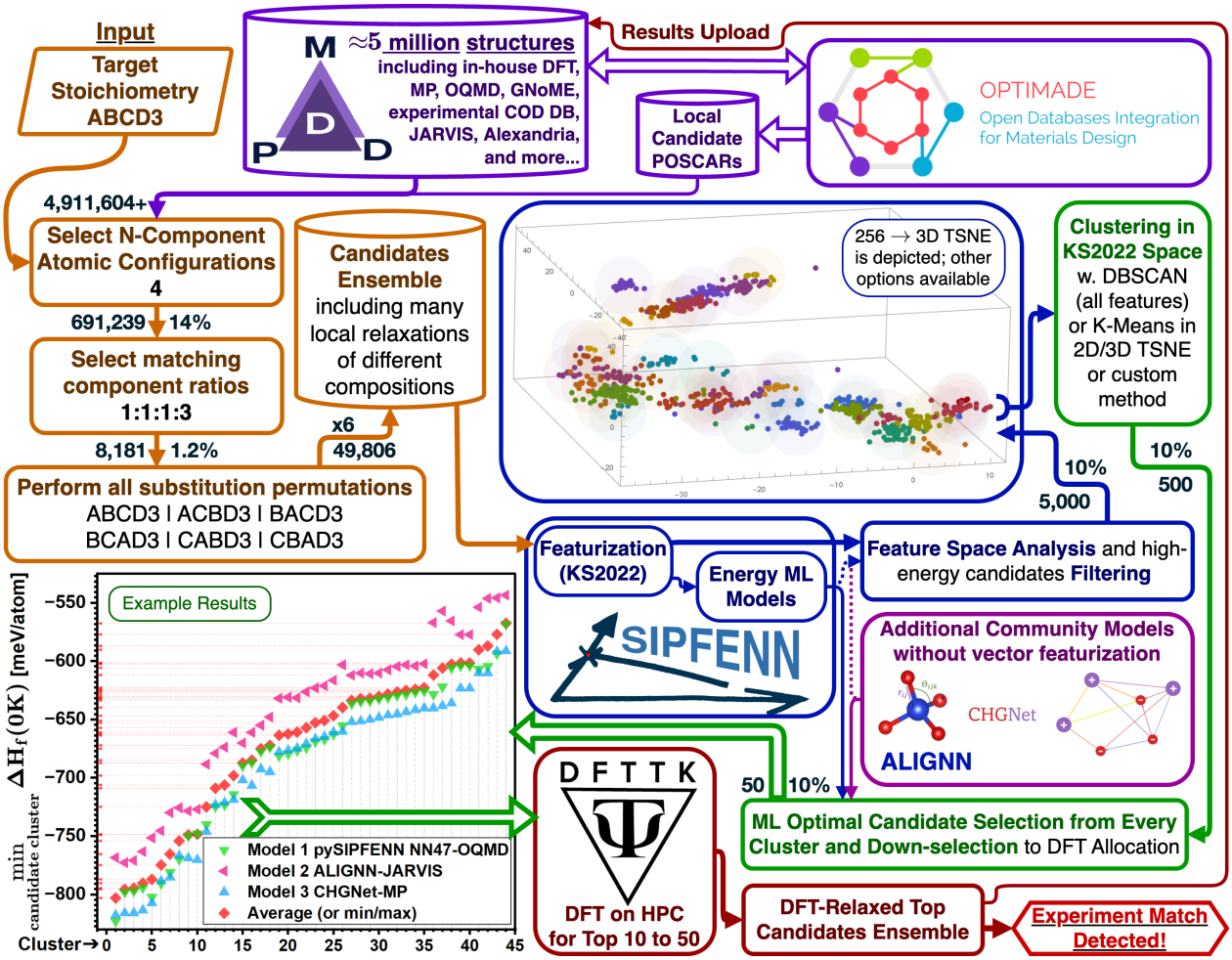


Figure 1: Core schematic of the `crystALL` approach over a specific stoichiometry, illustrating the complete workflow from candidate generation (orange), featurization and ML with `pySIPFENN` [29] (blue), addition of community ML models (purple), clustering to select top candidates from unique structures/energy basins (green), and final DFT relaxations leading to match with experimental observations (red). The example depicts a dramatic filtering cascade - 5M structures from `MPDD` are harnessed to generate around 50k $ABCD_3$ candidates, then ML-filtered to 5k in 500 clusters, culminating in DFT calculations for top 50 cluster representatives to construct an approximately ergodic ensemble and match experimental observations; while feedback loop to `MPDD` creates a self-improving ecosystem for materials discovery that grows with `crystALL` use.

these properties and an agreement score that quantifies confidence across the methods at various settings, but users interested in mixed-dimensional structures (e.g., layered materials with intercalated ions) can query the raw per-algorithm dimensionality fields directly. As of writing of this work, MPDD contains approximately 4.2 million 3D, 430,000 2D, 50,000 1D, and 20,000 0D entries. Although dimensionality is emphasized here, the same filtering functionality applies to any combination of queryable fields in MPDD.

These prototypical configurations, handled as pymatgen [54] Structure objects, are then passed to the open-source pySIPFENN (python toolset for Structure-Informed Property and Feature Engineering with Neural Networks) [57] where they are featurized into high-dimensional vectors (256D) in pySIPFENN's KS2022 space [29], as of the current crystALL version, which is a computationally and scientifically improved version of Ward2017 pure-Python space featurizer from earlier pySIPFENN versions. Both were implemented based on statistics over global structure features and local chemical environments building upon the approach first introduced in Magpie by Ward *et al.* [58]. Such a description of each candidate configuration captures both structural and compositional information, which can be used to analyze the large pool of candidates in terms of uniqueness and to represent different basins in the energy landscape.

The recognition of similar configurations of the same structure energy basins is accomplished through different strategies of clustering candidates in the KS2022 feature space, depicted in blue-colored part of Figure 1. In the current crystALL version, this can be done either (a) across all dimensions encoding structural information, scaled based on min/max values, using DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [59], or (b) via a two-step process of nonlinear dimensionality reduction into 2D or 3D (e.g. Fig. 3) through t-distributed Stochastic Neighbor Embedding (tSNE) [60], [61] followed by k-means partitioning the data into k clusters, with the number selected by the user based on expert intuition over the candidate pool or through optimization of silhouette scores, allowing the clustering to reflect the natural groupings in the data. The former approach generally allows capturing more complex relationships and pathway information, while the latter approach provides more intuitive visualization and interpretation in the context of where prototypes came from, allowing researchers to adapt the clustering approach to the specific characteristics of their problem. The accompanying software documentation details the tunable parameters for both clustering methods, the rationale for default selections, and practical guidance for users.

Parallel to clustering, each candidate is passed through ML models to predict its energy and other properties of interest. For rapid static property evaluations, crystALL provides built-in access to multi-property models of pySIPFENN alongside MPDD-ALIGNN, a lightweight, author-maintained fork of ALIGNN [62] optimized for deployment over MPDD, which includes a default set of individual property models based on the Materials Project [23], [49], NIST-JARVIS [25], and Alexandria [24].

Furthermore, in the upcoming v0.6.0 release (currently available as an active development branch on GitHub; see section 5), users can utilize the ASE (Atomic Simulation Environment) [63] interface, allowing users to perform ML-based relaxation of candidate structures utilizing machine-learned interatomic potentials (MLIP), such as CHGNet [64], MACE [65], [66], Orb-v3 [67], MatterSim [68], or UPET [69], [70]. Appendix B provides further discussion on the 24 available MLIPs. crystALL's extensibility stems from not only the seamless incorporation of any state-of-the-art model compatible with ASE, ALIGNN, or pySIPFENN via a simple change of config files, but also architectural choices allowing for easy code changes to incorporate any custom model the user may want to include. At the end of the ML pipeline of crystALL, all predictions are combined and can be used individually or based on min/max/mean statistics. This multi-objective capability provides flexibility and convenience to test different model implementations, while allowing for a more nuanced analysis considering other specified physical properties beyond just stability.

Then, using the ML modeling and clustering results, crystALL selects the *top candidate* configuration from each cluster based on user-selected optimality metric, as depicted in the green-colored part of Figure 1. Optimality may correspond to lowest predicted energy, closest match to an experimentally observed property (for example band gap), or a composite objective defined symbolically based on multiple ML models. For example, one can effortlessly pass an equation like $\sqrt[3]{3/\left((M_A/V_0)^{3/2}\left[(B+\frac{4}{3}G)^{-3/2}+2G^{-3/2}\right]\right)}$ to compute polycrystalline speed of sound v_m from known mass per atom M_A and ML-predicted shear modulus G, bulk modulus B, and volume per atom V_0. In a typical use case, such as those discussed in section 3, optimality corresponds to the lowest energy in each cluster. Thus, top candidates can be interpreted as structurally diverse atomic configurations at the local minima of respective ML energy basins within a high-dimensional energy landscape spanning beyond the initial convex hull, often both

above and below.

Finally, for materials in well-explored regions of the chemical space, where the accuracy of ML models is reliably high, the *ensemble of top candidates* can be used directly. For novel or complex regions, the candidates are passed to a final validation step using DFT [8]–[10], which can be executed outside of `crystALL` via integrated tools like `DFTTK` [71] on high-performance computing clusters. The latter is typically performed in batches on top 10 to 200 configurations until either (1) an approximately ergodic ensemble is constructed, in the case of materials discovery use, or (2) a match or matches (see Sec. 3.1) to the experimental observations are found, as depicted in red-colored part of Figure 1. Importantly, the new DFT-calculated structures are contributed back into `MPDD` allowing their re-use in next iterations significantly improving future results over similar compounds.

2.2 Searching for Composition Ranges

To recognize limitations of single-composition-targeting `crystALL` searches, discussed in terms of implementation in subsection 2.1 and examples later in section 3, one has to consider consequences of their operation on exact chemical formulas, also called stoichiometries. These can be set precisely, based on (a) highly accurate and precise experimental composition measurements; (b) sufficient composition measurements, combined with expert intuition; or (c) simple assumption that something may be of interest, e.g., enough to be screened [17], [31], [32]. Additionally, the exact/nominal nature of compounds, allows researchers to access atomistic data on them readily from nearly any atomistic database thanks to the `OPTIMADE` API [45].

However, this apparent precision masks a fundamental challenge to purely data-driven structure identification and discovery under point *b*. Namely, the vast majority of the experimental observations do not provide unambiguous stoichiometry due to both: (1) mean/nominal compositions never exactly matching any "low-index" (small integer ratios) compound atomic ratios like AB_3, A_2B_7, or $A_2BC_3D_3$; and (2) the measurement uncertainty bounds spanning many of such ratios.

This compositional ambiguity can often be overcome with educated guesses based on similar elements and sets of rules, or extensive ab initio efforts until a match is found; however, there are at least two cases where these often fail. First case covers compounds with uncommon or complex structures, such as one called $Mg_{32}Al_{47}Cu_7$ based on its approximate composition which has been reported for years as the minority phase in several aerospace Al alloys [72] and more recently in lightweight high entropy alloys (L-HEAs) [73]. These typically have a known XRD pattern but no corresponding crystal structure with a well-defined stoichiometry. The second case represents an epistemic challenge underexamined by the community - when exploratory studies investigate uncommon chemical systems without crystallographic measurements and report observing compounds of nominal compositions which are later picked by computational researchers who propose some candidates, identify lowest-energy ones, but conclude that the compound is metastable. Such assessment may be correct, but cannot be proven. To illustrate this epistemic trap, consider a hypothetical first-time study of the Al−Cu system where Energy Dispersive Spectroscopy (EDS) yields an $Al_{32.1}Cu_{67.9}$ composition. Based on this, one could calculate many candidates for $AlCu_2$, which are all metastable, without ever considering Al_4Cu_9 even though it is equally distant in the compositional space from the EDS result and actually stable in a γ-brass-like structure.

The preceding discussion clarifies two critical needs: (a) robust pools of prototypes that span a wide variety of compositions, extending the concept in Figure 1; and (b) the removal of bias toward only "low-index" compositions, which allows harvesting of all available community data (i.e., approximately 11,393 distinct stoichiometric ratios in `MPDD` as of March 2026). To address this need, `crystALL` enables searches for compositional ranges, where the user specifies both the nominal/mean composition and uncertainty in all of the components, individually, as shown in the top left of Figure 2.

To enable this kind of approach, one small yet elegant innovation had to be implemented within `crystALL` and `MPDD`. That's because, while most atomistic datasets and `OPTIMADE` API [45] include both anonymized chemical formulas and dictionaries of element fractions, these only allow queries for limited predefined sets or for fraction bounds over specific elements. To assure completeness, the former would require the user to fetch all anonymous formulas every time and filter them for a given bound. The latter can, in principle, work when searching for 2/3-component prototypes, but fails for higher order systems, like 5-component ones, where one has to set $N \cdot N! \cdot 2$, i.e. $5 \cdot 5! \cdot 2 = 1,200$ bounds for each filtered chemical system to look for range of anonymized compositions. In `crystALL`, this issue has been addressed by including a sorted list of atomic fractions in `MPDD`, under the `material.sortedFractionsDict` field. This construction enables both (a) single queries to cover all N-component systems and (b) the collapse of symmetrically-equivalent regions of the composition space $N!$-fold, as depicted in the pink-colored part of Fig-

ure 2; thus, making queries feasible for any system complexity. To put it in perspective, querying for a modest 5-component composition range over all 84,000+ quinary systems in MPDD, requires over 10,000,000 times fewer bounds set compared to using the typical dictionary of defined elements and their fractions.

Once all atomic ratios within set bounds and corresponding prototype configurations are obtained, they are processed through crystALL's atomic substitution steps of permutations, featurization, clustering, ML modeling, candidate selection, and, optionally, ab initio steps (e.g., DFT/DFTTK); in the same fashion as described in Section 2.1, with two additional considerations. Firstly, since there are variations in candidate compositions, formation energies alone can no longer be used for direct stability comparisons between them. Thus, all candidates are referenced to the self-consistent ML-computed energy convex hull from MPDD either yielding stability values that can be analyzed directly or allowing for the introduction of a composition-dependent correction factor based on the hull depth at each stoichiometry. Secondly, such an approach generates output with additional dimension of the compositional distance to the nominal composition value, as shown in the example chart in Figure 2, enabling researchers or AI agents to make a better-informed decision on what to pass into the validation steps, balancing the predicted stability with closeness to the experimental match.

3. Case Studies

3.1 Identifying $NdBi_2$ and Other Structures

The crystALL approach was first deployed by the authors to address an unsolved materials challenge within the study by Im *et al.* [74]. That study reassessed thermodynamic models for the Nd−Bi chemical system, representing an essential step towards the rational design of rare-earth alloys for clean energy applications, such as the electrochemical recovery of rare-earth elements and liquid metal batteries. As highlighted by Im *et al.* [74], while the $NdBi_2$ compound has long been known to be thermodynamically stable [75], its crystal structure has remained unknown for five decades, posing a significant obstacle to ab initio investigations. By deploying crystALL's data-driven approach on MPDD, an ensemble of candidate configurations for $NdBi_2$ was generated and later validated by DFT-based methods, following an earlier development version of the workflow depicted in Figure 1.

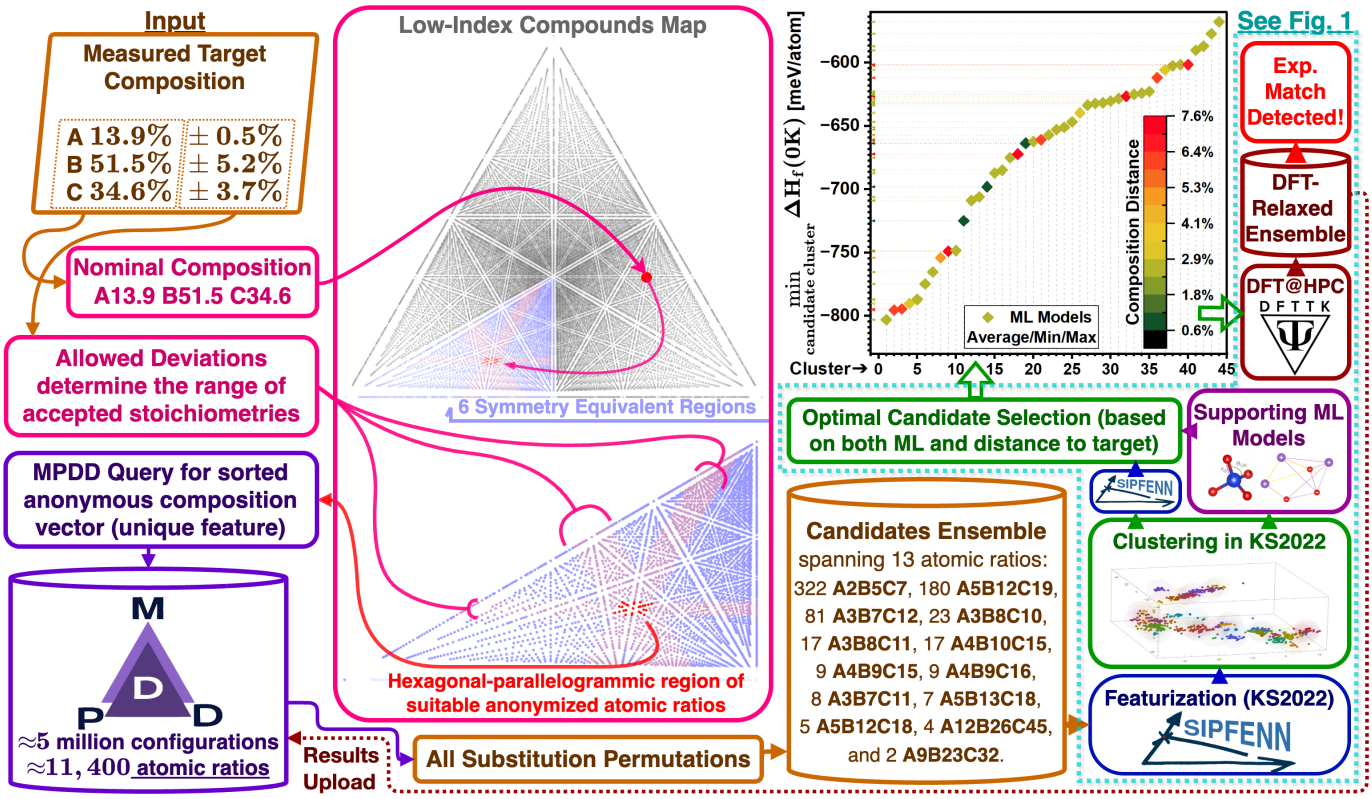


Figure 2: Core schematic of the crystALL approach over compositions with defined ranges, illustrating the complete workflow from definition of the search query (pink) and multi-composition candidate ensemble generation (orange), and further steps described in Figure 1. The example depicts a composition is matched to 13 compounds within the defined bounds and then analyzed both in terms of stability and distance.

First, 26,055 AB_2 prototype configurations were extracted from MPDD, which at the time had approximately 1.3 million DFT-relaxed or experimentally observed configurations covering all structural records contained in OQMD [21], Materials Project [23], [49], NIST's JARVIS [25], and the Crystallography Open Database (COD) [48]. Following the extraction and substitution, all generated candidates were featurized using the Ward2017 featurizer in pySIPFENN, and their energies were predicted via the SIPFENN-NN20 "Novel Materials Model" optimized for extrapolations [57]. The top 1,000 lowest-energy candidates ($\simeq 4\%$) were then selected and clustered using k-means (following the 3D tSNE approach described in subsection 2.1) into an intentionally high number of clusters to capture smaller variations, as depicted in Figure 3.

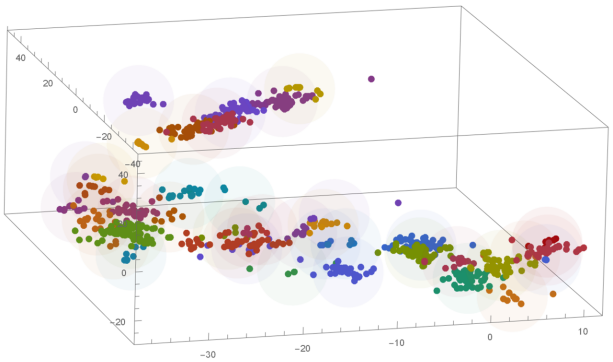


Figure 3: Visualization of crystALL's tSNE low-dimensional (3D) embedding of Ward2017 feature vectors [57], [58] of 1,000 lowest energy $NdBi_2$ candidates, as predicted by SIPFENN, clustered with k-means approach to determine the ensemble of unique basins.

Next, the SIPFENN-NN20-predicted formation energies were used to select the most stable candidate from each cluster, which were each relaxed using three different DFT exchange-correlation functionals (PBE [76], PS/PBESol [77], and SCAN [10]) to assess their order of predicted stability. For comparison, an alternative evolutionary structure search was performed using USPEX by Lyakhov *et al.* [78], discussed earlier in section 1.

The DFT results, shown in Figure 4 (reproduced from Im *et al.* [74]), indicated that USPEX indeed found the stable structure under SCAN and PBE functionals, relative to all of the 20 structures generated by crystALL. However, as explained in section 1, the zero-temperature lowest energy structure is not necessarily the one experimentally observed. This has been precisely the case in Im *et al.* [74], as the XRD pattern did not match the USPEX-generated structure. Instead, it was successfully matched to a mixture of two related configurations (anti-La_2Sb and La_2Te) generated by crystALL, each predicted to be metastable by all three DFT methods, which are marked by arrows in Figure 4.

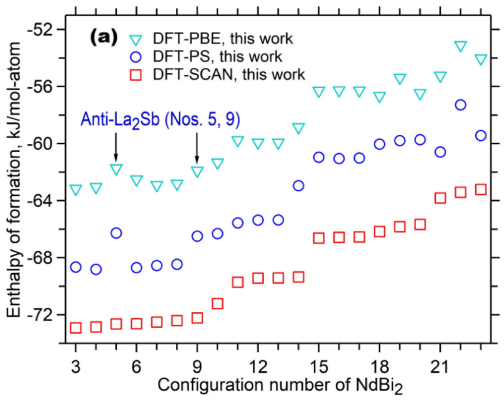


Figure 4: 0K formation energies of structure candidates predicted by USPEX (#3) and crystALL (#4-23). Pure element reference configurations #1 and #2 are omitted. The mixture of 2nd and 6th lowest-energy candidates has been matched to the experimental observations of $NdBi_2$ demonstrating both crystALL's predictive ability and importance of studying of ensembles of metastable structures. Reproduced under license from Im *et al.* [74].

A similar approach and results were also obtained in a computational study of high-pressure intermetallics formed during welding process in the Al−Fe system by Shang *et al.* [79], wherein crystALL has been used to propose a set of candidates for Al_2Fe, including its ground state configuration, which was further validated via DFT and phonon calculations.

Furthermore, following the successful deployment of crystALL over Nd−Bi in Im *et al.* [74], the compositional range search described in subsection 2.2, has been deployed over Nd−Bi−Sn chemical system, in a follow-up study that remains unpublished. There, a stable ternary compound has been observed, contrary to the current literature on that system [80], and its structure was rapidly proposed through crystALL.

3.2 Guided Search of Novel Hydrides

Another successful case study of crystALL is demonstrated in the search for novel complex hydrides for H_2 storage, where compositions with higher hydrogen content are extensively researched to enhance volumetric and gravimetric capacities and to match desired structural properties. Within this rationale, specific structures such as perovskites have traditionally been targeted for showcasing improved H_2 adsorption and desorption kinetics associated with their high tolerance to defects, large interstices, and high diffusivity.

Several perovskite hydrides — such as $CsCaH_3$ and $NaMgH_3$ [81]–[84] — have already been synthesized, and their stability and structures are predicted using the empirical tolerance factor between A and B cations, first introduced by Goldschmidt in 1926 [85]. In these ex-

amples, structural predictions are based on the idealized spatial arrangement of atoms in the ABH_3 structure, which is often oversimplified for more complex perovskites, such as $KMn(BH_4)_3$ [86], [87], with large ionic groups.

Consequently, crystallographic details of such complex perovskites are still lacking, with only a few experimental cases reported [88]–[90], given the challenges associated with their synthesis and the refinement of their structural characterization. In addition, perovskite structures can undergo a series of temperature-induced polymorphic transformations that are of great importance for their emerging physical properties. This sequence of structural transitions is dominated by lattice instabilities, corresponding to ensembles of both octahedral rotations and polar displacements, as well as some degree of orbital-induced effects in most compounds.

Following the target stoichiometry search methodology in subsection 2.1, the authors investigated the $NaMg(BH_4)_3$ composition system, considering all its stable and metastable candidate structures using `crystALL`. Appendix Figure A1 summarizes the 100 most stable candidates, generated based on prototypes from Materials Project [23], [49], GNoME [33], AFLOW [19], and the Crystallography Open Database (COD) [48], while Figure A2 displays their 2D tSNE clustering. The seven clusters identified group together candidates with similar atomic environments, representing primarily hexagonal structures (clusters 1 and 2), triclinic structures (cluster 3), trigonal structures (clusters 4, 5, and 7), and perovskite-type structures (cluster 6). The most stable structure predicted in cluster 6 is based on the $NbV(PO_4)_3$ entry from the GNoME dataset (a9b189292f), and comprises a highly distorted perovskite structure, wherein the A sites are coordinated by 12 BH_4^- anions, with half pointing inwards and half pointing outwards. The corner-sharing coordination exhibited by this prototype represents a previously unobserved borohydride structure, distinct from any previously reported ones, which are characterized by edge-sharing [89] and face-sharing [90] configurations (see Figure A3).

The $NaMg(BH_4)_3$ case highlights the significant potential of `crystALL` not only for the rapid and robust prediction of candidate structures to aid experimental validation, but also for extracting valuable insights into relative stabilities and ensemble configurations. A detailed description and dataset from the $NaMg(BH_4)_3$ study are provided as an interactive tutorial included with the `crystALL` software (see section 5).

3.3 Catalysis Materials and Molten Salts

Ru-based catalysts are critical to enhancing manufacturing efficiency of nylon, thanks to being highly effective in synthesizing the key chemical intermediate through the benzene partial hydrogenation reaction [91]. Achieving precise control over hydrogenation selectivity involves designing diverse active site ensembles on the surfaces of intermetallic catalysts [92]. The Ru−Zn system is of particular interest to the community, but it is typically modeled strictly as FCC/HCP solid solutions [93] without the consideration of intermetallic compounds. To address this, Gong [94] used `crystALL` alongside `pySIPFENN` [57] ML models and DFT validation to explore over 165,000 candidate configurations of Ru−Zn intermetallic compounds, including Ru_3Zn, RuZn, Ru_4Zn_9, $RuZn_3$, Ru_3Zn_{10}, $RuZn_6$, $RuZn_{10}$, and $RuZn_{13}$; simultaneously modeling active site ensembles of top candidates to design novel catalysts with high activity and selectivity.

Another application where accurate modeling necessitates the use of ensembles to fully capture the complexity of the underlying materials is in the study of chloride molten salts in advanced Molten Salt nuclear Reactors (MSRs) and pyroprocessing techniques to separate lanthanide fission products [95], both of which are critical in the journey towards clean energy and carbon footprint reduction. In this example, `crystALL` facilitated the prediction of structures for the KLa_3Cl_{10} and $K_3La_5Cl_{18}$ compounds in the $KCl-LaCl_3$ system described by Gong *et al.* [95], thereby explaining their stability reported in various literature sources [96], [97] and allowing for further investigation using ab-initio-based methods.

4. Summary

Section 1 presents a high-level analysis of the field of computational structure prediction, where for the last several decades, two fundamental challenges have persisted: predicting (a) which compositions form compounds of interest experimentally and (b) their candidate topologies or atomic arrangements. Databases like `AFLOW` [17], `OQMD` [20], and Materials Project [23], [49] have made invaluable contributions by systematically populating prototype libraries with diverse chemistries, creating rich repositories of both stable and metastable structures. However, some of the recent ML-driven approaches have shifted focus exclusively toward finding "new stable compounds", potentially creating a form of model collapse where iterative bias toward lowest-energy structures progressively eliminates the diversity essential for both (a) understanding real materials formed by sta-

tistical ensembles of configurations and (b) robustness on the training datasets allowing discovery of novel structures. Methods like USPEX [14], AIRSS [13], and XtalOpt [16] can in principle generate diverse polymorphs for individual compositions, but they are much too expensive to efficiently generate wide, approximately ergodic ensembles across many chemistries. This gap is critical because metastable configurations in phases often govern interesting properties (e.g., negative thermal expansion or occurrence of catalytic sites), thermodynamic models require diverse configurational data, and experimental structures frequently correspond to higher-energy configurations. crystALL addresses this niche by leveraging the structural diversity already present across millions of entries in MPDD to generate comprehensive configurational ensembles in the least biased fashion.

The beginning of Section 2 establishes the foundation of the crystALL approach, emphasizing how it leverages MPDD's 4.9 million relaxed structures as diverse prototypes spanning configurations, including many symmetry-broken ones. This bias-minimizing approach of using all available data generates orders of magnitude more candidates than traditional prototype libraries, which can be handled thanks to the use of ML for their characterization. Section 2 also notes how users can expand this pool further through community datasets and systematic generation tools like ATAT [52] and aflow++ [53]. Subsequently, subsection 2.1 details the implementation workflow of crystALL where, typically, thousands of prototypes that can accommodate a given stoichiometry are used to generate candidates through all permutations of atomic substitutions, then followed by featurization in pySIPFENN's KS2022 space [29]. Such representations are then clustered through a method of choice to identify unique representative candidates picked by ML models based on the lowest energy in the cluster, i.e., the lowest sampled configuration in given basin of energy landscape, or based on match to observed properties. Finally, the top candidates can be passed to DFT calculations or matched with experiments directly. The resulting ensemble captures diverse configurations across the energy landscape, with validated structures feeding back into MPDD to enrich future searches.

Section 2.2 addresses a critical gap in computational materials discovery by extending crystALL to handle compositional uncertainty inherent to experimental measurements. The section addresses how techniques like EDS yield uncertain, non-stoichiometric compositions, like $Al_{32.1}Cu_{67.9}$ ± 1.5 $at.$ $\%$, that could correspond to many compounds. Traditional search techniques typically either (a) define a set of low-index anonymous formulas, risking missing prototypes with complex high-index ratios or (b) search for compositional ranges based by querying for bounds over individual chemical elements, which works well for compounds with low number of components but fails in more complex cases. MPDD implements sorted fractional composition queries allowing crystALL to query on sortedFractionsDict field, enabling efficient searches across anonymized and symmetry-collapsed compositional bounds; reducing query complexity 10^7 times for quinary compounds. All prototypes within bounds undergo the standard pipeline (substitutions, featurization, clustering), with formation energies referenced to ML-predicted convex hulls for composition-corrected stability comparisons. The output provides both predicted stability and compositional distance from the nominal value, enabling informed selection for validation.

Lastly, Section 3 demonstrates crystALL's practical use across diverse materials systems, including published validations of its effectiveness in structure discovery for experimentally observed compounds in subsection 3.1. It showcases how crystALL was used in Im *et al.* [74] to find $NdBi_2$ structure, unresolved for five decades; it identified it as a mixture of two metastable configurations through crystALL, while ground-state-searching methods like USPEX converged only to the lowest-energy structure that did not match experiments. Another validation was also achieved in Shang *et al.* [79] for Al_2Fe intermetallics in high-pressure welding applications. Additionally, several yet-unpublished and example use cases have been included, such as identification of observed Nd−Bi−Sn ternary compound through compositional search from subsection 2.2. subsection 3.2 illustrates crystALL's potential for discovering complex structures, predicting diverse $NaMg(BH_4)_3$ configurations. Lastly, subsection 3.3 also includes examples of use over functional materials through systematic analysis of Ru−Zn intermetallics (165,000+ candidates across several observed compounds) for benzene hydrogenation catalyst design and successful structure prediction of KLa_3Cl_{10} and $K_3La_5Cl_{18}$ compounds essential for molten salt reactor technologies and lanthanide separation [95].

Overall, the crystALL approach represents a shift in computational materials discovery to fill an interesting niche - from focusing on singular ground states to embracing the full configurational complexity, metastability, and uncertainty that defines real materials at finite temperatures. It conceptually bridges the gap between idealized computational predictions and experimental observations. By leveraging millions of existing configurations relaxed for hundreds of thousands of different chemical

compositions, it forms a minimally-biased and ensemble-focused framework. The crystALL approach generates intermediate results, which do not find "the answer" but generate comprehensive landscapes of possibilities to explain experimental results and empower novel materials discovery.

5. Software and Data Availability

The crystALL software described in this work has been published as free open-source software (FOSS) under the MIT license. It is available through the GitHub repository at github.com/PhasesResearchLab/crystALL and selected versions are archived in Zenodo under all-versions DOI 10.5281/zenodo.23872006, including `v0.4.0` used in Im *et al.* [74] and Shang *et al.* [79], and the latest stable release `v0.5.4` which is also attached to this publication.

End-users can also obtain it from the PyPI package repository under the name `crystall-search`.

Additionally, the upcoming `v0.6.0` version featuring the `MLIPAdapter` relaxation architecture covering 24 MLIPs, described in Appendix B, is currently available for cloning and testing via the `v0.6.0` development branch on the GitHub repository.

Data used within this study and the past studies employing crystALL has been obtained from MPDD, which can be accessed via an open-access OPTIMADE endpoint at optimade.mpdd.org indexed by the consortium, or via a closed-access API authors can share upon reasonable request.

Contributions CRediT

Adam M. Krajewski: Conceptualization, Methodology, Software, Writing - Original Draft, Data curation, Validation, Visualization; **Ricardo Amaral:** Methodology, Software, Investigation, Writing - Original Draft, Visualization; **Jonathan W. Siegel:** Software, Validation, Writing - Review & Editing; **Luke A. Myers:** Methodology, Software, Writing - Original Draft; **Rushi Gong**: Investigation, Writing - Original Draft; **Shun-Li Shang:** Investigation, Writing - Review & Editing; **Zi-Kui Liu:** Funding acquisition, Supervision, Writing - Review & Editing, Resources.

Acknowledgments

This work has been funded through grants: U.S. National Science Foundation - Pathways to Enable Open-Source Ecosystems (POSE) with Award No. **FAIN-2229690**, NSF Grant **DMS-2424305**, DOE Advanced Research Projects Agency-Energy (ARPA-E) with Award No. **DE-AR0001435**, and DOE Nuclear Energy University Programs (NEUP) with Award No. **DE-NE0009288**.

We would like to thank **Dr. Hongyeun Kim** at PSU for valuable feedback on the development of the compositional search method, and **Dr. Hojong Kim**'s research group at PSU for conducting experiments in the Bi-Nd and Bi-Nd-Sn systems, which inspired the initial method development in 2020.

Additionally, Ricardo Amaral gratefully acknowledges financial support from the **Cuomo Foundation** and the **Intergovernmental Panel on Climate Change (IPCC)**. The contents of this document are solely the responsibility of the authors and under no circumstances may be considered as reflecting the position of the Cuomo Foundation and/or the IPCC.

References

[1] F. C. Hawthorne, "Structural aspects of oxide and oxysalt crystals," *Acta Crystallographica Section B Structural Science*, vol. 50, no. 5, pp. 481–510, Oct. 1994. DOI: 10.1107/S0108768193014466. [Online]. Available: https://scripts.iucr.org/cgi-bin/paper?S0108768193014466.

[2] A. Le Bail, H. Duroy, and J. Fourquet, "Ab-initio structure determination of LiSbWO6 by X-ray powder diffraction," *Materials Research Bulletin*, vol. 23, no. 3, pp. 447–452, Mar. 1988. DOI: 10.1016/0025-5408(88)90019-0. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0025540888900190.

[3] W. Yue, P. K. Tripathi, G. Ponon, *et al.*, "Phase Identification in Synchrotron X-ray Diffraction Patterns of Ti–6Al–4V Using Computer Vision and Deep Learning," *Integrating Materials and Manufacturing Innovation*, vol. 13, no. 1, pp. 36–52, Mar. 2024. DOI: 10.1007/s40192-023-00328-0. [Online]. Available: https://link.springer.com/10.1007/s40192-023-00328-0.

[4] R. Mehdi, R. Chawla, E. I. Barcelos, M. A. Willard, R. H. French, and F. Ernst, "2D-diffractogram analysis: Kinematic-diffraction simulator for neural-network training-data generation," *Computational Materials Science*, vol. 252, p. 113 777, Apr. 2025. DOI: 10.1016/j.commatsci.2025.113777.

[5] Z. Qin, Q. Luo, W. Qin, X. Chen, H. Zhang, and C. U. I. Wong, "GCN-Based Framework for Materials Screening and Phase Identification," *Materials 2025, Vol. 18, Page 959*, vol. 18, no. 5, p. 959, Feb. 2025. DOI: 10.3390/MA18050959. [Online]. Available: https://www.mdpi.com/1996-1944/18/5/959/htmhttps://www.mdpi.com/1996-1944/18/5/959.

[6] K. Shu, D.-Y. Gui, W.-X. Yan, and C.-H. Wang, "Machine Learning Tackles the Challenge of Powder X-ray Diffraction Indexing for All Crystal Systems," *Journal of Chemical Information and Modeling*, vol. 65, no. 19, pp. 10 025–10 036, Oct. 2025. DOI: 10.1021/acs.jcim.5c01506. [Online]. Avail-

able: https://pubs.acs.org/doi/10.1021/acs.jcim.5c01506.

[7] A. C. Ferrari and D. M. Basko, "Raman spectroscopy as a versatile tool for studying the properties of graphene," *Nature Nanotechnology*, vol. 8, pp. 235–246, 4 Apr. 2013. DOI: 10.1038/nnano.2013.46. [Online]. Available: https://www.nature.com/articles/nnano.2013.46.

[8] W. Kohn and L. J. Sham, "Self-Consistent Equations Including Exchange and Correlation Effects," *Physical Review*, vol. 140, no. 4A, A1133–A1138, Nov. 1965. DOI: 10.1103/PhysRev.140.A1133.

[9] M. Kothakonda, A. D. Kaplan, E. B. Isaacs, *et al.*, "Testing the r ² SCAN Density Functional for the Thermodynamic Stability of Solids with and without a van der Waals Correction," *ACS Materials Au*, vol. 3, no. 2, pp. 102–111, Mar. 2023. DOI: 10.1021/acsmaterialsau.2c00059. [Online]. Available: https://pubs.acs.org/doi/10.1021/acsmaterialsau.2c00059.

[10] J. Sun, A. Ruzsinszky, and J. Perdew, "Strongly Constrained and Appropriately Normed Semilocal Density Functional," *Physical Review Letters*, vol. 115, no. 3, pp. 1–6, 2015. DOI: 10.1103/PhysRevLett.115.036402.

[11] L. Pauling, "The principles determining the structure of complex ionic crystals," *Journal of the American Chemical Society*, vol. 51, no. 4, pp. 1010–1026, Apr. 1929. DOI: 10.1021/ja01379a006. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ja01379a006.

[12] A. Le Bail, "Inorganic structure prediction with GRINSP," *Journal of Applied Crystallography*, vol. 38, no. 2, pp. 389–393, Apr. 2005. DOI: 10.1107/S0021889805002384. [Online]. Available: www.iza-structure.org/databases/.

[13] C. J. Pickard and R. J. Needs, "Ab initio random structure searching," *Journal of Physics: Condensed Matter*, vol. 23, no. 5, p. 053 201, Feb. 2011. DOI: 10.1088/0953-8984/23/5/053201. [Online]. Available: https://iopscience.iop.org/article/10.1088/0953-8984/23/5/053201.

[14] C. W. Glass, A. R. Oganov, and N. Hansen, "USPEX - Evolutionary crystal structure prediction," *Computer Physics Communications*, vol. 175, no. 11, pp. 713–720, 2006. DOI: 10.1016/j.cpc.2006.07.020.

[15] Y. Wang, J. Lv, L. Zhu, and Y. Ma, "CALYPSO: A method for crystal structure prediction," *Computer Physics Communications*, vol. 183, no. 10, pp. 2063–2070, Oct. 2012. DOI: 10.1016/J.CPC.2012.05.008.

[16] S. Hajinazar and E. Zurek, "XtalOpt version 14: Variable-composition crystal structure search for functional materials through Pareto optimization," *Computer Physics Communications*, vol. 320, p. 109 910, Mar. 2026. DOI: 10.1016/j.cpc.2025.109910. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0010465525004114.

[17] S. Curtarolo, W. Setyawan, G. L. Hart, *et al.*, "AFLOW: An automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, Jun. 2012. DOI: 10.1016/j.commatsci.2012.02.005. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0927025612000717.

[18] C. Toher, C. Oses, D. Hicks, *et al.*, "The AFLOW Fleet for Materials Discovery," in *Handbook of Materials Modeling*, Cham: Springer International Publishing, 2018, pp. 1–28. DOI: 10.1007/978-3-319-42913-7{_}63-1. [Online]. Available: http://link.springer.com/10.1007/978-3-319-42913-7_63-1.

[19] M. Esters, C. Oses, S. Divilov, *et al.*, "aflow.org: A web ecosystem of databases, software and tools," 2022. DOI: 10.1016/j.commatsci.2022.111808. [Online]. Available: www.elsevier.com/locate/commatsci.

[20] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD)," *JOM*, vol. 65, no. 11, pp. 1501–1509, Nov. 2013. DOI: 10.1007/s11837-013-0755-4. [Online]. Available: http://link.springer.com/10.1007/s11837-013-0755-4.

[21] S. Kirklin, J. E. Saal, B. Meredig, *et al.*, "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies," *npj Computational Materials*, vol. 1, no. 1, p. 15 010, Dec. 2015. DOI: 10.1038/npjcompumats.2015.10. [Online]. Available: www.oqmd.org/download.http://www.nature.com/articles/npjcompumats201510.

[22] J. Shen, S. D. Griesemer, A. Gopakumar, *et al.*, "Reflections on one million compounds in the open quantum materials database (OQMD)," *Journal of Physics: Materials*, vol. 5, no. 3, p. 031 001, Jul. 2022. DOI: 10.1088/2515-7639/ac7ba9. [Online]. Available: https://iopscience.iop.org/article/10.1088/2515-7639/ac7ba9.

[23] A. Jain, S. P. Ong, G. Hautier, *et al.*, "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation," *APL Materials*, vol. 1, no. 1, p. 011 002, Jul. 2013. DOI: 10.1063/1.4812323. [Online]. Available: http://aip.scitation.org/doi/10.1063/1.4812323.

[24] J. Schmidt, H. C. Wang, T. F. Cerqueira, S. Botti, and M. A. Marques, "A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals," *Scientific Data 2022 9:1*, vol. 9, no. 1, pp. 1–8, Mar. 2022. DOI: 10.1038/s41597-022-01177-w. [Online]. Available: https://www.nature.com/articles/s41597-022-01177-w.

[25] K. Choudhary, K. F. Garrity, A. C. Reid, *et al.*, "The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design," *npj Computational Materials*, vol. 6, no. 1, pp. 1–13, Dec. 2020. DOI: 10.1038/s41524-020-00440-1. [Online]. Available: https://doi.org/10.1038/s41524-020-00440-1.

[26] H. Eckert, S. Divilov, M. J. Mehl, *et al.*, "The AFLOW library of crystallographic prototypes: Part 4," *Computational Materials Science*, vol. 240, p. 112 988, May 2024. DOI: 10.1016/J.COMMATSCI.2024.112988.

[27] N. H. Anderson, M. J. Mehl, H. Eckert, S. Divilov, X. Campilongo, and S. Curtarolo, "The AFLOW library of crystallographic prototypes: Part 5," *Computational Materials Science*, vol. 275, p. 115 022, Oct. 2026. DOI: 10.1016/j.commatsci.2026.115022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0927025626005410.

[28] A. M. Krajewski, A. M. Beese, W. F. Reinhart, and Z.-K. Liu, "Efficient generation of grids and traversal graphs in compositional spaces towards exploration and path planning," *npj Unconventional Computing*, vol. 1, no. 1, p. 12, Nov. 2024. DOI: 10.1038/s44335-024-00012-2. [Online]. Available: https://www.nature.com/articles/s44335-024-00012-2.

[29] A. M. Krajewski, J. W. Siegel, and Z.-K. Liu, "Efficient structure-informed featurization and property prediction of ordered, dilute, and random atomic structures," *Computational Materials Science*, vol. 247, p. 113 495, Jan. 2025. DOI: 10.1016/j.commatsci.2024.113495. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S092702562400716X.

[30] H.-C. Wang, S. Botti, and M. A. L. Marques, "Predicting stable crystalline compounds using chemical similarity," *npj Computational Materials*, vol. 7, no. 1, p. 12, Jan. 2021. DOI: 10.1038/s41524-020-00481-6. [Online]. Available: https://www.nature.com/articles/s41524-020-00481-6.

[31] J. Schmidt, N. Hoffmann, H.-C. Wang, *et al.*, "Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials," *Advanced Materials*, vol. 35, no. 22, p. 2 210 788, Jun. 2023. DOI: 10.1002/ADMA.202210788. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/adma.202210788https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.202210788https://onlinelibrary.wiley.com/doi/10.1002/adma.202210788.

[32] W. Ye, X. Lei, M. Aykol, and J. H. Montoya, "Novel inorganic crystal structures predicted using autonomous simulation agents," *Scientific Data*, vol. 9, no. 1, p. 302, Jun. 2022. DOI: 10.1038/s41597-022-01438-8. [Online]. Available: https://www.nature.com/articles/s41597-022-01438-8.

[33] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, and E. Dogus Cubuk, "Scaling deep learning for materials discovery," *80 | Nature |*, vol. 624, 2023. DOI: 10.1038/s41586-023-06735-9. [Online]. Available: https://doi.org/10.1038/s41586-023-06735-9.

[34] C. Zeni, R. Pinsler, D. Zügner, *et al.*, "A generative model for inorganic materials design," *Nature*, vol. 639, no. 8055, pp. 624–632, Mar. 2025. DOI: 10.1038/s41586-025-08628-5. [Online]. Available: https://www.nature.com/articles/s41586-025-08628-5.

[35] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, no. 8022, pp. 755–759, Jul. 2024. DOI: 10.1038/s41586-024-07566-y. [Online]. Available: https://www.nature.com/articles/s41586-024-07566-y.

[36] S. Alemohammad, J. Casco-Rodriguez, L. Luzi, *et al.*, "Self-Consuming Generative Models Go MAD," Jul. 2023. DOI: 10.48550/arXiv.2307.01850. [Online]. Available: http://arxiv.org/abs/2307.01850.

[37] A. K. Cheetham and R. Seshadri, "Artificial Intelligence Driving Materials Discovery? Perspective on the Article: Scaling Deep Learning for Materials Discovery," *Chemistry of Materials*, vol. 36, no. 8, pp. 3490–3495, Apr. 2024. DOI: 10.1021/acs.chemmater.4c00643. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.chemmater.4c00643.

[38] G. Olson and Z. Liu, "Genomic materials design: CALculation of PHAse Dynamics," *Calphad*, vol. 82, p. 102 590, Sep. 2023. DOI: 10.1016/j.calphad.2023.102590. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0364591623000627.

[39] A. Zunger, "Beware of plausible predictions of fantasy materials," *Nature 2021 566:7745*, vol. 566, no. 7745, pp. 447–449, Feb. 2019. DOI: 10.1038/d41586-019-00676-y. [Online]. Available: https://www.nature.com/articles/d41586-019-00676-y.

[40] S. Divilov, H. Eckert, D. Hicks, *et al.*, "Disordered enthalpy–entropy descriptor for high-entropy ceramics discovery," *Nature*, vol. 625, no. 7993, pp. 66–73, Jan. 2024. DOI: 10.1038/s41586-023-06786-y. [Online]. Available: https://www.nature.com/articles/s41586-023-06786-y.

[41] R. Grau-Crespo, S. Hamad, S. R. Balestra, *et al.*, "Capturing local compositional fluctuations in NMR modelling of solid solutions," *Chemical Science*, vol. 16, no. 41, pp. 19 357–19 369, Oct. 2025. DOI: 10.1039/D5SC04550A. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2025/sc/d5sc04550ahttps://pubs.rsc.org/en/content/articlelanding/2025/sc/d5sc04550a.

[42] Z.-K. Liu, Y. Wang, and S.-L. Shang, "Zentropy Theory for Positive and Negative Thermal Expansion," *Journal of Phase Equilibria and Diffusion*, Feb. 2022. DOI: 10.1007/s11669-022-00942-z. [Online]. Available: https://link.springer.com/10.1007/s11669-022-00942-z.

[43] P. P. Ferreira, L. J. Conway, A. Cucciari, *et al.*, "Search for ambient superconductivity in the Lu-N-H system," *Nature Communications*, vol. 14, no. 1, p. 5367, Sep. 2023. DOI: 10.1038/s41467-023-41005-2. [Online]. Available: https://www.nature.com/articles/s41467-023-41005-2.

[44] A. M. Krajewski, J. W. Siegel, S.-L. Shang, Y. Wang, J. Xu, and Z.-K. Liu, *MPDD: The Material-Property-Descriptor Database*, 2021. [Online]. Available: https://phaseslab.com/mpdd.

[45] M. Evans, J. Bergsma, A. Merkys, *et al.*, "Developments and applications of the OPTIMADE API for materials discovery, design, and data exchange," *Digital Discovery*, Feb. 2024. DOI: 10.1039/D4DD00039K. [Online]. Available: http://pubs.rsc.org/en/Content/ArticleLanding/2024/DD/D4DD00039K.

[46] S. Gražulis, D. Chateigner, R. T. Downs, *et al.*, "Crystallography Open Database – an open-access collection of crystal structures," *Journal of Applied Crystallography*, vol. 42, no. 4, pp. 726–729, Aug. 2009. DOI: 10.1107/S0021889809016690.

[47] S. Gražulis, A. Daškevič, A. Merkys, *et al.*, "Crystallography Open Database (COD): An open-access collection of crystal structures and platform for world-wide collaboration," *Nucleic Acids Research*, vol. 40, no. D1, pp. D420–D427, Jan. 2012. DOI: 10.1093/nar/gkr900. [Online]. Available: https://academic.oup.com/nar/article/40/D1/D420/2903497.

[48] S. Gražulis, A. Merkys, A. Vaitkus, *et al.*, "Crystallography Open Database: History, Development, and Perspectives," in *Materials Informatics*, Wiley, Oct. 2019, pp. 1–39. DOI: 10.1002/9783527802265.ch1. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/9783527802265.ch1.

[49] M. K. Horton, P. Huck, R. X. Yang, *et al.*, "Accelerated data-driven materials science with the Materials Project," *Nature Materials*, Jul. 2025. DOI: 10.1038/s41563-025-02272-0. [Online]. Available: https://www.nature.com/articles/s41563-025-02272-0.

[50] C. Liu, R. E. Cohen, and J. Sun, "Prediction of thermally driven quasi-1D superionic states in carbon hydride under giant planetary conditions," *Nature Communications*, Mar. 2026. DOI: 10.1038/s41467-026-70603-z. [Online]. Available: https://www.nature.com/articles/s41467-026-70603-z.

[51] A. Zunger, S.-H. Wei, L. G. Ferreira, and J. E. Bernard, "Special quasirandom structures," *Physical Review Letters*, vol. 65, no. 3, pp. 353–356, Jul. 1990. DOI: 10.1103/PhysRevLett.65.353. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevLett.65.353.

[52] A. van de Walle, P. Tiwary, M. de Jong, *et al.*, "Efficient stochastic generation of special quasirandom structures," *Calphad*, vol. 42, pp. 13–18, Sep. 2013. DOI: 10.1016/j.calphad.2013.06.006. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0364591613000540.

[53] C. Oses, M. Esters, D. Hicks, *et al.*, "aflow++: A C++ framework for autonomous materials design," *Computational Materials Science*, vol. 217, p. 111 889, Jan. 2023. DOI: 10.1016/J.COMMATSCI.2022.111889.

[54] J. George, M. K. Horton, A. D. Kaplan, *et al.*, "pymatgen: 15 years of community growth, new functionality, and future prospects," Oct. 2026. [Online]. Available: http://arxiv.org/abs/2610.23538.

[55] G. Cheon, K.-A. N. Duerloo, A. D. Sendek, C. Porter, Y. Chen, and E. J. Reed, "Data Mining for New Two- and One-Dimensional Weakly Bonded Solids and Lattice-Commensurate Heterostructures," *Nano Letters*, vol. 17, no. 3, pp. 1915–1923, Mar. 2017. DOI: 10.1021/acs.nanolett.6b05229. [Online]. Available: https://pubs.acs.org/doi/10.1021/acs.nanolett.6b05229.

[56] P. M. Larsen, M. Pandey, M. Strange, and K. W. Jacobsen, "Definition of a scoring parameter to identify low-dimensional materials components," *Physical Review Materials*, vol. 3, no. 3, p. 034 003, Mar. 2019. DOI: 10.1103/PhysRevMaterials.3.034003. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevMaterials.3.034003.

[57] A. M. Krajewski, J. W. Siegel, J. Xu, and Z.-K. Liu, "Extensible Structure-Informed Prediction of Formation Energy with improved accuracy and usability employing neural networks," *Computational Materials Science*, vol. 208, p. 111 254, Jun. 2022. DOI: 10.1016/j.commatsci.2022.111254. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0927025622000593.

[58] L. Ward, R. Liu, A. Krishna, *et al.*, "Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations," *Physical Review B*, vol. 96, no. 2, p. 024 104, Jul. 2017. DOI: 10.1103/PhysRevB.96.024104. [Online]. Available: http://link.aps.org/doi/10.1103/PhysRevB.96.024104.

[59] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," 1996. [Online]. Available: www.aaai.org.

[60] G. Hinton and S. Roweis, "Stochastic Neighbor Embedding," in *Neural Information Processing Systems*, 2002.

[61] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[62] K. Choudhary and B. DeCost, "Atomistic Line Graph Neural Network for improved materials property predictions," *npj Computational Materials*, vol. 7, no. 1, p. 185, Nov. 2021. DOI: 10.1038/s41524-021-00650-1. [Online]. Available: https://www.nature.com/articles/s41524-021-00650-1.

[63] A. Hjorth Larsen, J. JØrgen Mortensen, J. Blomqvist, *et al.*, "The atomic simulation environment—a Python library for working with atoms," *Journal of Physics: Condensed Matter*, vol. 29, no. 27, p. 273 002, Jun. 2017. DOI: 10.1088/1361-648X/AA680E. [Online]. Available: https://iopscience.iop.org/article/10.1088/1361-648X/aa680ehttps://iopscience.iop.org/article/10.1088/1361-648X/aa680e/meta.

[64] B. Deng, P. Zhong, K. Jun, *et al.*, "CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling," *Nature Machine Intelligence* |, vol. 5, pp. 1031–1041, 2023. DOI: 10.1038/s42256-023-00716-3. [Online]. Available: https://doi.org/10.1038/s42256-023-00716-3.

[65] I. Batatia, S. Batzner, D. P. Kovács, *et al.*, "The design space of E(3)-equivariant atom-centred interatomic potentials," *Nature Machine Intelligence*, vol. 7, no. 1, pp. 56–67, Jan. 2025. DOI: 10.1038/s42256-024-00956-x. [Online]. Available: https://www.nature.com/articles/s42256-024-00956-x.

[66] I. Batatia, D. P. Kovács, G. N. C. Simm, C. Ortner, and G. Csányi, "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 11 423–11 436. arXiv: 2206.07697.

[67] B. Rhodes, S. Vandenhaute, V. Šimkus, *et al.*, "Orb-v3: atomistic simulation at scale," Apr. 2025. [Online]. Available: http://arxiv.org/abs/2504.06231.

[68] H. Yang, C. Hu, Y. Zhou, *et al.*, "MatterSim: A Deep Learning Atomistic Model Across Elements, Temperatures and Pressures," May 2024. [Online]. Available: http://arxiv.org/abs/2405.04967.

[69] C. Malosso, F. Bigi, P. Pegolo, *et al.*, "High-quality, high-information datasets for universal atomistic machine learning," Sep. 2026. [Online]. Available: http://arxiv.org/abs/2603.02089.

[70] F. Bigi, P. Pegolo, A. Mazitov, J. Schmidt, and M. Ceriotti, "Pushing the limits of unconstrained machine-learned interatomic potentials," *Machine Learning: Science and Technology*, vol. 7, no. 3, p. 035 051, 2026. DOI: 10.1088/2632-2153/ae6417. arXiv: 2601.16195.

[71] N. L. E. Hew, L. A. Myers, A. van de Walle, S.-L. Shang, and Z.-K. Liu, "Density Functional Theory ToolKit (DFTTK) to automate first-principles thermodynamics via the quasi-harmonic approximation," *Computational Materials Science*, vol. 258, p. 114 072, Aug. 2025. DOI: 10.1016/j.commatsci.2025.114072. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S092702562500415X.

[72] J. F. Li, Z. Ziqiao, J. Na, and T. Chengyu, "Localized corrosion mechanism of 2xxx-series Al alloy containing S(Al2CuMg) and θ'(Al2Cu) precipitates in 4.0% NaCl solution at pH 6.1," *Materials Chemistry and Physics*, vol. 91, no. 2-3, pp. 325–329, Jun. 2005. DOI: 10.1016/J.MATCHEMPHYS.2004.11.034.

[73] M. Sadeghi and B. Niroumand, "Design and characterization of a novel MgAlZnCuMn low melting point light weight high entropy alloy (LMLW-HEA)," *Intermetallics*, vol. 151, p. 107 658, Dec. 2022. DOI: 10.1016/J.INTERMET.2022.107658.

[74] S. Im, S. L. Shang, N. D. Smith, *et al.*, "Thermodynamic properties of the Nd-Bi system via emf measurements, DFT calculations, machine learning, and CALPHAD modeling," *Acta Materialia*, vol. 223, p. 117 448, Jan. 2022. DOI: 10.1016/J.ACTAMAT.2021.117448.

[75] K. Yoshihara, J. Taylor, L. Calvert, and J. Despault, "Rare-earth bismuthides," *Journal of the Less Common Metals*, vol. 41, no. 2, pp. 329–337, Jul. 1975. DOI: 10.1016/0022-5088(75)90038-7. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0022508875900387.

[76] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized Gradient Approximation Made Simple," *Physical Review Letters*, vol. 77, no. 18, pp. 3865–3868, Oct. 1996. DOI: 10.1103/PhysRevLett.77.3865.

[77] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, *et al.*, "Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces," *Physical Review Letters*, vol. 100, no. 13, p. 136 406, Apr. 2008. DOI: 10.1103/PhysRevLett.100.136406.

[78] A. O. Lyakhov, A. R. Oganov, H. T. Stokes, and Q. Zhu, "New developments in evolutionary structure prediction algorithm USPEX," *Computer Physics Communications*, vol. 184, no. 4, pp. 1172–1182, 2013. DOI: 10.1016/j.cpc.2012.12.009.

[79] S.-L. Shang, H. Sun, B. Pan, *et al.*, "Forming mechanism of equilibrium and non-equilibrium metallurgical phases in dissimilar aluminum/steel (Al–Fe) joints," *Scientific Reports*, vol. 11, no. 1, p. 24 251, Dec. 2021. DOI: 10.1038/s41598-021-03578-0. [Online]. Available: https://www.nature.com/articles/s41598-021-03578-0.

[80] F Xu, Y. T. Chen, R Ye, *et al.*, "Thermodynamic Study of Sn-Bi-Nd, Tb Ternary Systems," *IOP Conference Series: Materials Science and Engineering*, vol. 230, p. 012 033, Sep. 2017. DOI: 10.1088/1757-899X/230/1/012033.

[81] H. Park, M Pezat, and B Darriet, "A new ternary hydride: Cscah3," *Revue de chimie minérale*, vol. 23, no. 3, pp. 323–328, 1986.

[82] A. Bouamrane, J. Laval, J.-P. Soulie, and J. Bastide, "Structural characterization of namgh2f and namgh3," *Materials Research Bulletin*, vol. 35, no. 4, pp. 545–549, 2000. DOI: 10.1016/S0025-5408(00)00249-X.

[83] K. Ikeda, Y. Kogure, Y. Nakamori, and S. Orimo, "Reversible hydriding and dehydriding reactions of perovskite-type hydride namgh3," *Scripta Materialia*, vol. 53, no. 3, pp. 319–322, 2005. DOI: 10.1016/j.scriptamat.2005.04.010.

[84] D. Pottmaier, E. R. Pinatel, J. G. Vitillo, *et al.*, "Structure and thermodynamic properties of the namgh3 perovskite: A comprehensive study," *Chemistry of Materials*, vol. 23, no. 9, pp. 2317–2326, 2011. DOI: 10.1021/cm103204p.

[85] V. M. Goldschmidt, "Die gesetze der krystallochemie," *Die Naturwissenschaften*, vol. 14, no. 21, pp. 477–485, 1926. DOI: 10.1007/BF01507527.

[86] P. Schouwink, V. D'Anna, M. B. Ley, *et al.*, "Bimetallic borohydrides in the system m(bh4)2–kbh4 (m = mg, mn): On the structural diversity," *The Journal of Physical Chemistry C*, vol. 116, no. 20, pp. 10 829–10 840, 2012. DOI: 10.1021/jp212318s.

[87] P. Schouwink, M. B. Ley, A. Tissot, *et al.*, "Structure and properties of complex hydride perovskite materials," *Nat Commun*, vol. 5, p. 5706, 2014. DOI: 10.1038/ncomms6706. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/25490884.

[88] D. B. Ravnsbaek, L. H. Sorensen, Y. Filinchuk, F. Besenbacher, and T. R. Jensen, "Screening of metal borohydrides by mechanochemistry and diffraction," *Angew Chem Int Ed Engl*, vol. 51, no. 15, pp. 3582–6, 2012. DOI: 10.1002/anie.201106661. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22383379.

[89] P. Schouwink, F. Morelle, Y. Sadikin, Y. Filinchuk, and R. Černý, "Increasing hydrogen density with the cation-anion pair bh4–nh4+ in perovskite-type nh4ca(bh4)3," *Energies*, vol. 8, no. 8, pp. 8286–8299, 2015. DOI: 10.3390/en8088286.

[90] K. T. Moller, M. B. Ley, P. Schouwink, R. Cerny, and T. R. Jensen, "Synthesis and thermal stability of perovskite alkali metal strontium borohydrides," *Dalton Trans*, vol. 45, no. 2, pp. 831–40, 2016. DOI: 10.1039/c5dt03590b. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/26646319.

[91] P. Zhang, T. Wu, T. Jiang, *et al.*, "Ru–Zn supported on hydroxyapatite as an effective catalyst for partial hydrogenation of benzene," *Green Chemistry*, vol. 15, no. 1, pp. 152–159, Jan. 2013. DOI: 10.1039/C2GC36596K. [Online]. Available: https://pubs.rsc.org/en/content/articlehtml/2013/gc/c2gc36596khttps://pubs.rsc.org/en/content/articlelanding/2013/gc/c2gc36596k.

[92] A. Dasgupta, H. He, R. Gong, *et al.*, "Atomic control of active-site ensembles in ordered alloys to enhance hydrogenation selectivity," *Nature Chemistry*, vol. 14, no. 5, pp. 523–529, May 2022. DOI: 10.1038/s41557-021-00855-3. [Online]. Available: https://www.nature.com/articles/s41557-021-00855-3.

[93] L. Mu, J. Han, and Y. Yang, "A Theoretical Study on the Structural Evolution of Ru–Zn Bimetallic Nanoparticles," *Nanomaterials*, vol. 15, no. 8, p. 568, Apr. 2025. DOI: 10.3390/nano15080568. [Online]. Available: https://www.mdpi.com/2079-4991/15/8/568.

[94] R. Gong, "Investigation of Atomic Environments by Computational Thermodynamics: Applications in Intermetallic Catalysts and Molten Salts," Ph.D. dissertation, The Pennsylvania State University, Aug. 2025. [Online]. Available: https://etda.libraries.psu.edu/catalog/20570rfg5281.

[95] R. Gong, S.-L. Shang, V. G. Goncharov, *et al.*, "Thermodynamic modeling of the LiCl-KCl-LaCl3 system with Bayesian model selection and uncertainty quantification," *Journal of Molecular Liquids*, vol. 432, p. 127 811, Aug. 2025. DOI: 10.1016/j.molliq.2025.127811. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0167732225009882.

[96] H. J. Seifert, H. Fink, and G. Thiel, "Thermodynamic properties of double chlorides in the systems ACl/LaCl3 (A = Na, K, Rb, Cs)," *Journal of The Less-Common Metals*, vol. 110, no. 1-2, pp. 139–147, 1985. DOI: 10.1016/0022-5088(85)90315-7.

[97] X. Song and C. Zheng, "An Investigation on the Phase Diagram of Ternary System LaCl3-KCl-LiCl (in Chinese)," *Acta Chimica Sinica*, vol. 53, pp. 978–984, 1995.

[98] H. Wang, L. Zhang, J. Han, and W. E, "DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics," *Computer Physics Communications*, vol. 228, pp. 178–184, 2018. DOI: 10.1016/j.cpc.2018.03.016.

[99] G. Simeon and G. de Fabritiis, "TensorNet: Cartesian tensor representations for efficient learning of molecular potentials," in *Advances in Neural Information Processing Systems*, vol. 36, 2023, pp. 37 334–37 353. DOI: 10.48550/arXiv.2306.06482. arXiv: 2306.06482 [cs.LG].

[100] A. Musaelian, S. Batzner, A. Johansson, *et al.*, "Learning local equivariant representations for large-scale atomistic dynamics," *Nature Communications*, vol. 14, p. 579, 2023. DOI: 10.1038/s41467-023-36329-y.

[101] B. Yin, J. Wang, W. Du, *et al.*, "AlphaNet: Scaling up local-frame-based neural network interatomic potentials," *npj Computational Materials*, vol. 11, 2025. DOI: 10.1038/s41524-025-01817-w. arXiv: 2501.07155.

[102] Y. Chen, L. Fu, S. Yang, and L. Chen, *Eqnorm: Generalized machine learning potential*, https://github.com/yzchen08/eqnorm, Software, 2025.

[103] X. Fu, B. M. Wood, L. Barroso-Luque, *et al.*, *Learning smooth and expressive interatomic potentials for physical property prediction*, 2025. DOI: 10.48550/arXiv.2502.12147. arXiv: 2502.12147 [physics.comp-ph].

[104] S. Batzner, A. Musaelian, L. Sun, *et al.*, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," *Nature Communications*, vol. 13, p. 2453, 2022. DOI: 10.1038/s41467-022-29939-5.

[105] T. Koker, M. Kotak, and T. Smidt, *Training a foundation model for materials on a budget*, 2025. DOI: 10.48550/arXiv.2508.16067. arXiv: 2508.16067 [physics.comp-ph].

[106] I. Batatia, W. J. Baldwin, D. Kuryla, *et al.*, *MACE-POLAR-1: A polarisable electrostatic foundation model for molecular chemistry*, 2026. DOI: 10.48550/arXiv.2602.19411. arXiv: 2602.19411 [physics.chem-ph].

[107] T. W. Ko, R. Liu, A. R. Mishra, Z. Yu, J. Qi, and S. P. Ong, *A linear-scaling, charge-aware foundation potential for atomistic simulations*, 2025. DOI: 10.48550/arXiv.2511.07249. arXiv: 2511.07249 [cond-mat.mtrl-sci].

[108] Y. Park, J. Kim, S. Hwang, and S. Han, "Scalable parallel algorithm for graph neural network interatomic potentials in molecular dynamics simulations," *Journal of Chemical Theory and Computation*, vol. 20, no. 11, pp. 4857–4868, 2024. DOI: 10.1021/acs.jctc.4c00190.

[109] Y.-L. Liao, A. J. Hoffman, S. C. Shen, A. Duval, S. W. Norwood, and T. Smidt, *EquiformerV3: Scaling efficient, expressive, and general SE(3)-equivariant graph attention transformers*, 2026. DOI: 10.48550/arXiv.2604.09130. arXiv: 2604.09130 [cs.LG].

[110] Y. Zhou, S. Hu, X. Zhang, H. Wang, G. Tan, and W. Jia, "MatRIS: Toward reliable and efficient pretrained machine learning interatomic potentials," in *The Fourteenth International Conference on Learning Representations*, 2026. arXiv: 2603.02002 [cs.LG].

[111] B. M. Wood, M. Dzamba, X. Fu, *et al.*, *UMA: A family of universal models for atoms*, 2025. DOI: 10.48550/arXiv.2506.23971. arXiv: 2506.23971 [cs.LG].

[112] A. Bochkarev, Y. Lysogorskiy, and R. Drautz, "Graph atomic cluster expansion for semilocal interactions beyond equivariant message passing," *Physical Review X*, vol. 14, no. 2, p. 021 036, 2024. DOI: 10.1103/PhysRevX.14.021036.

[113] Z. Xu, W. Xie, and P. Hu, *Spectral/spatial tensor atomic cluster expansion with universal embeddings in Cartesian space*, 2025. DOI: 10.48550/arXiv.2509.14961. arXiv: 2509.14961 [stat.ML].

[114] K. Yan, M. Bohde, A. Kryvenko, *et al.*, *A materials foundation model via hybrid invariant-equivariant architectures*, 2025. DOI: 10.48550/arXiv.2503.05771. arXiv: 2503.05771 [cs.LG].

[115] N. Kuroda, K. Ishihara, T. Shiota, and W. Mizukami, *Expanding universal machine learning interatomic potentials to 97 elements towards nuclear applications*, 2026. DOI: 10.48550/arXiv.2603.03223. arXiv: 2603.03223 [physics.chem-ph].

Appendix A. crystALL clustering of structures

To illustrate the clustering workflow implemented in crystALL, we present results for the 100 most stable $NaMg(BH_4)_3$ structures, as shown in Figure A1. The clustering through tSNE and k-means identified seven distinct clusters (Fig. A2), effectively separating structures with fundamentally different atomic arrangements while grouping those with similar local environments. Specifically, clusters 1 and 2 primarily contained hexagonal structures, cluster 3 grouped triclinic structures, clusters 4, 5 and 7, trigonal structures, and cluster 6 contained perovskite-type structures.

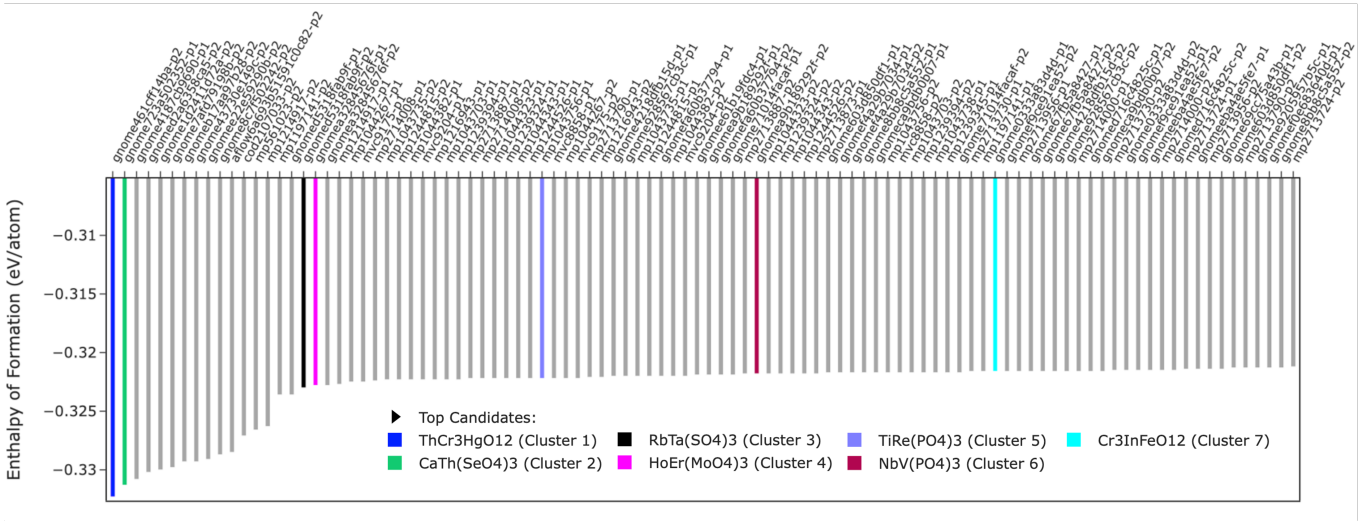


Figure A1: Bar chart with the 100 most stable prototypes and 7 top candidates per cluster identified by crystALL for the $NaMg(BH_4)_3$ composition. Top candidate structures identified on the graph are based on the prototypes: $ThCr_3HgO_{12}$ (gnome461cff14ba), $CaTh(SeO_4)_3$ (gnome723a502352), $RbTa(SO_4)_3$ (gnomed5318fab9f), $HoEr(MoO_4)_3$ (gnomea32845676f), $TiRe(PO_4)_3$ (mp1043756), $NbV(PO_4)_3$ (gnomea9b189292f), and Cr_3InFeO_{12} (gnome0333a83d4d).

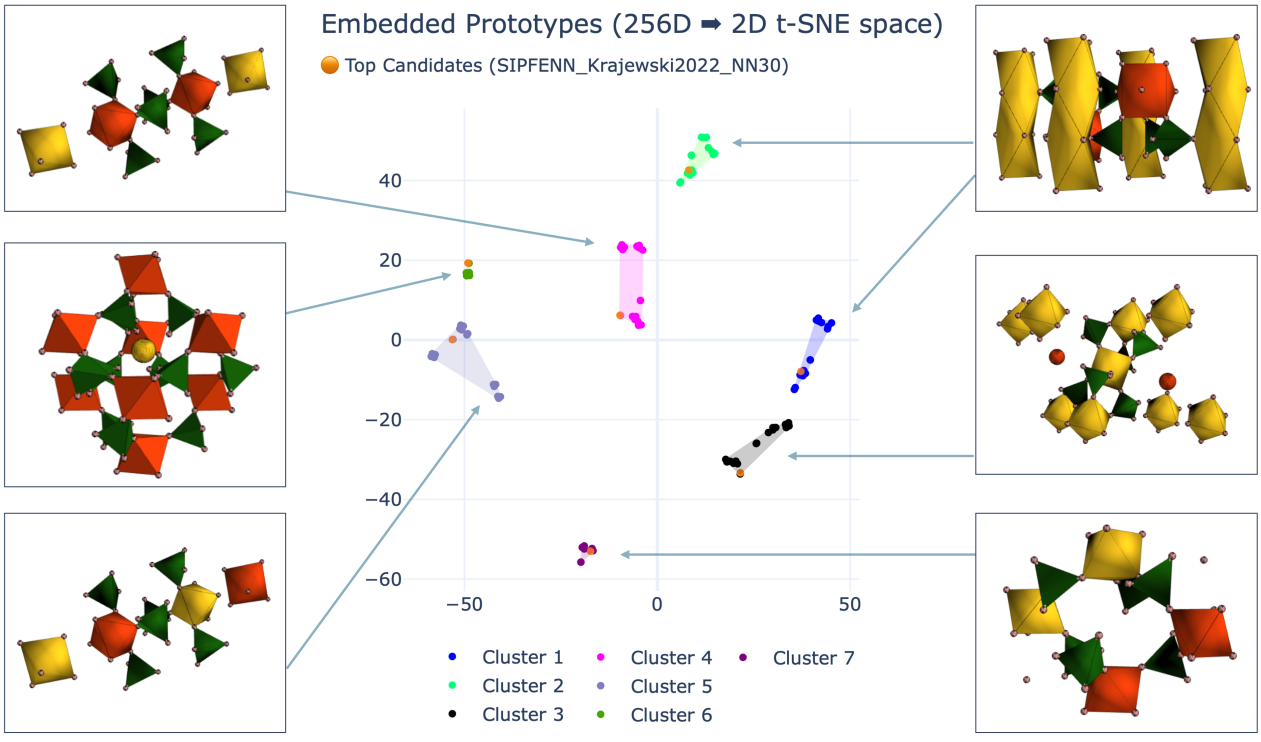


Figure A2: 2D tSNE and k-means clustering of the 100 most stable $NaMg(BH_4)_3$ structures predicted by crystALL, revealing seven distinct clusters: hexagonal (clusters 1 and 2), triclinic (cluster 3), trigonal (clusters 4, 5, and 7), and perovskite-type (cluster 6) symmetries.

We compared the most stable $NaMg(BH_4)_3$ prototypes predicted by crystALL with previously reported borohydride perovskite structures from the literature, as shown side-by-side in Figure A3. The literature examples include both edge-sharing and face-sharing configurations described by Schouwink *et al.* [89] and Moller *et al.* [90]. In contrast, the most stable candidate from cluster 6, derived from the $NbV(PO_4)_3$ prototype (GNoME dataset entry a9b189292f), adopts a highly distorted perovskite structure with unique corner-sharing coordination of the BH_4^- anions, distinguishing it from previously reported borohydride frameworks.

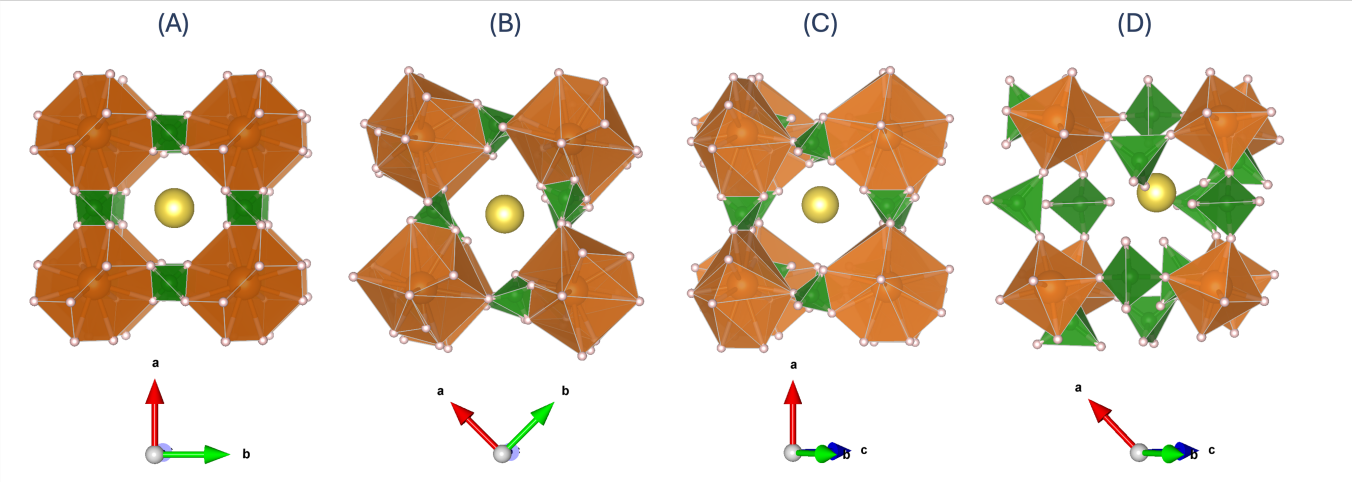


Figure A3: Side-by-side comparison of perovskite-type $NaMg(BH_4)_3$ structures based on: (A) $NH_4Ca(BH_4)_3$ cubic prototype with edge-sharing octahedra (Schouwink *et al.* [89]); (B) $CsSr(BH_4)_3$ prototype with edge- and face-sharing octahedra (Moller *et al.* [90]); (C) $RbSr(BH_4)_3$ prototype with edge- and corner-sharing octahedra (Moller *et al.* [90]); and (D) the most stable candidate from cluster 6 predicted by crystALL, based on the $NbV(PO_4)_3$ prototype (GNoME dataset entry a9b189292f), featuring unique corner-sharing coordination not observed in previous reports.

Appendix B. Machine Learning Interatomic Potentials (MLIPs) Supported in v0.6.0

The upcoming crystALL v0.6.0 release (currently accessible via the v0.6.0 branch on GitHub) introduces a fully modular architecture driven by the MLIPAdapter registry, enabling automated structure relaxations and multi-property predictions across diverse machine-learned interatomic potentials. Each potential manages its own execution micro-environment with support for CPU and CUDA acceleration, full cell/volume optimization, and optional piping into downstream static property models.

A summary of the machine-learned interatomic potentials integrated into crystALL v0.6.0 is given in Table B1. In the table, the "Checkpoint Pinned" column refers to the default model checkpoint variant adopted. However, users can easily access other checkpoint variants within each model family (such as models with different parameter sizes, training datasets, or dispersion corrections) by updating the MLIPAdapter configuration file or supplying custom parameters at initialization. Newly released potentials can be added in a similar fashion by updating the same configuration file, and newly vetted models will continuously be featured in upcoming minor releases. The release column distinguishes between models distributed via official PyPI packages and those installed from source repositories. Official PyPI releases are preferred for their stable dependencies and guaranteed forward compatibility, whereas "From Source" configurations are provisional and should be replaced as official packages become available.

Unlike static predictors such as pySIPFENN and ALIGNN, which evaluate fixed geometries, MLIPs iteratively optimize atomic positions and cell parameters. In crystALL v0.6.0, multiple MLIPs can be evaluated as separate objectives for the same candidate structures, with each potential relaxing independently to prevent coordinate conflicts. During relaxation, adapters extract converged observables such as relaxed total and formation energies, residual forces, stresses, cell volume, and the optimized geometry. Furthermore, static adapters can be piped into any MLIP to evaluate the relaxed structures directly, pairing the chosen potential with multiple downstream property models.

Table B1: Summary of machine-learned interatomic potentials (MLIPs) supported in `crystALL v0.6.0`. For each model family, "Checkpoint Pinned" denotes the default pre-trained variant loaded by the adapter (user-configurable). Release indicates distribution through official PyPI packages or installation from source repositories. Hardware lists supported execution devices (CPU and CUDA).

Model Family	Checkpoint Pinned	Hardware	Distribution	Citation
DeePMD	DPA-4.0.1-Pro-MPtrj	CPU, CUDA	From source	[98]
CHGNet	0.3.0	CPU, CUDA	PyPI	[64]
TensorNet	PES-MatPES-PBE-2025.2	CPU, CUDA	PyPI	[99]
ORB	v3-omat:float32-high	CPU, CUDA	PyPI	[67]
ORB-D3	v3-omat:float32-high:PBE:BJ	CPU, CUDA	PyPI	[67]
Allegro	OAM-L:0.1	CUDA	From source	[100]
AlphaNet	alex-0410:oma	CPU, CUDA	From source	[101]
Eqnorm	eqnorm-mptrj	CPU, CUDA	From source	[102]
eSEN	eSEN-30M-OAM	CPU, CUDA	From source	[103]
MACE	mh-1:omat-pbe	CPU, CUDA	PyPI	[66]
MatterSim	v1.0.0-5M	CPU, CUDA	PyPI	[68]
NequIP	OAM-XL:0.1	CUDA	PyPI	[104]
Nequix	oam-1	CPU, CUDA	From source	[105]
PolarMACE	polar-1-l	CPU, CUDA	From source	[106]
QET	PES-MatPES-PBE-2025.2	CPU, CUDA	From source	[107]
SevenNet	omni-i12:omat24	CPU, CUDA	PyPI	[108]
UPET	oam-xl	CPU, CUDA	PyPI	[69], [70]
EquiformerV3	omat24-mptrj-salex-gradient	CPU, CUDA	From source	[109]
MatRIS	10m-oam:efsm	CPU, CUDA	From source	[110]
UMA	s-1p2:omat	CPU, CUDA	PyPI	[111]
GRACE	2L-SMAX-OMAT-large	CPU, CUDA	PyPI	[112]
TACE	OAM-L	CPU, CUDA	From source	[113]
HIENet	0	CPU, CUDA	From source	[114]
OSAKA	osaka26:small	CPU, CUDA	PyPI	[115]
Total	**24 Models**			