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.

Crystal-structure design by agentic AI in a language of motifs

Dinh-Khiet Le
Japan Advanced Institute of Science and Technology
Minh-Quyet Ha
Japan Advanced Institute of Science and Technology
Hong-Phuc Vu-Dinh
Japan Advanced Institute of Science and Technology
Viet-Cuong Nguyen
HPC SYSTEMS Inc.
Huan Tran
Georgia Institute of Technology
Truyen Tran
Applied Artificial Intelligence Institute
Takashi Miyake
Institute of Science Tokyo
Hiori Kino
The Institute of Statistical Mathematics
Hieu-Chi Dam
dam@jaist.ac.jp

Japan Advanced Institute of Science and Technology

Article

Keywords: Materials informatics, Artificial intelligence, Inverse design, Rare-earth magnets

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11807841/v1

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

Additional Declarations: No competing interests reported.

Crystal-structure design by agentic AI in a language of motifs

Dinh-Khiet Le[a], Minh-Quyet Ha[a,*], Hong-Phuc Vu-Dinh[a], Viet-Cuong Nguyen[b], Huan Tran[c], Truyen Tran[d], Takashi Miyake[e], Hiori Kino[f] and Hieu-Chi Dam[a,g,*]

[a]Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
[b]HPC SYSTEMS Inc., 3-9-15 Kaigan, Minato-ku, Tokyo 108-0022, Japan
[c]School of Materials Science and Engineering, Georgia Institute of Technology, 771 Ferst Dr. NW, Atlanta, Georgia 30332, United States
[d]Applied Artificial Intelligence Institute, Deakin University, Geelong, Victoria 3216, Australia
[e]Department of Physics, School of Science, Institute of Science Tokyo, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8551, Japan
[f]Research Center for Materials Informatics, Department of Advanced Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan
[g]International Center for Synchrotron Radiation Innovation Smart (SRIS), Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

ARTICLE INFO

Keywords:
Materials informatics
Artificial intelligence
Inverse design
Rare-earth magnets

ABSTRACT

Data-driven materials discovery interpolates more reliably than it extrapolates and seldom reveals new structure types. We present MatEvolve, an agentic-AI framework that designs crystals in an interpretable *language of motifs*, in which each crystal is written as a *motif profile* that describes the recurring geometric patterns—the *motifs*—composing the crystal. The motif profile serves not merely as a description of a material but as the medium for material design: the agent edits the profile with a stated rationale and constructs a crystal from the modified one; the most promising candidates are then validated by first-principles calculations. MatEvolve is built on the state-of-the-art language model Claude Fable 5 without fine-tuning. It is applied to the design of rare-earth-lean permanent-magnet compounds and finds novel (meta)stable crystals that are both new structural prototypes and magnet compounds; under an equal validation budget, it does so an order of magnitude more often than generative models. Beyond design, analysis of the human-readable profiles of the discovered crystals shows which motif combinations are associated with their properties.

1. Introduction

New materials have repeatedly redefined the limits of technology, from semiconductors underpinning modern computing to magnets and electrodes used in clean-energy systems. However, their discovery remains slow and costly: the space of candidate structures is effectively infinite, each candidate is expensive to synthesize and characterize. Moreover, after a promising material is identified, translating it into practical applications can take a decade or more. To accelerate discovery, materials science has progressed through four established paradigms—empirical experiment, theory, simulation, and data-driven learning [1]—and the fifth is now emerging, in which AI agents drive discovery [2, 3, 4]. Setting the empirical baseline aside, we consider the computational paradigms in turn—theory and simulation, data-driven learning, and the emerging agent-based approach—asking what each contributes, where it falls short, and how its limitations motivate the next paradigm.

First-principles theory and simulation reason deductively; they predict how a given structure will behave from the established physical theory. Density-functional theory (DFT) does so through a tractable approximation, computing the stability and properties of a structure by calculation alone [5]. Its distinctive strength is that it can assess a material that does not yet exist both quantitatively and with an accuracy determined by the underlying approximation, thereby helping determine whether it is worth synthesizing and characterizing. However, it faces two limitations as a route to discovery. First, each calculation is expensive; thus, even large-scale parallel computation covers only a negligible fraction of an effectively infinite candidate space. Second, deduction evaluates candidates but does not generate them; that task falls to other means: enumeration by researchers, crystal-chemistry rules such as Pauling's rules [6], or structure-search algorithms such as the evolutionary search [7] and particle-swarm search [8]. However, these approaches are also limited: enumeration and rule-based generation largely recombine known chemistry, while searching remains constrained by the high computational cost, which can make calculations infeasible for complex structures. These limitations have motivated an inductive, data-driven paradigm.

The data-driven paradigm learns statistical structure–property relationships from large, curated datasets, reasoning by induction—generalizing from many examples [9]. These inductive methods take two forms. *Surrogates* predict stability and property orders of magnitude faster than DFT at inference, allowing far larger candidate sets to be screened [10, 11, 12, 13, 14]. *Generators* learn a distribution over known structures and sample new candidates to supply that screen; some are conditioned on a target property [15, 16, 17, 18, 19, 20]. The largest such studies have predicted large numbers of DFT-stable candidates [21]. Active learning and autonomous laboratories add a selection

*Corresponding author
mq-ha@jaist.ac.jp (M. Ha); dam@jaist.ac.jp (H. Dam)
ORCID(s): 0000-0003-4617-0059 (M. Ha); 0000-0002-8008-582X (V. Nguyen); 0000-0002-8093-9426 (H. Tran); 0000-0001-6531-8907 (T. Tran); 0000-0003-2658-3470 (T. Miyake); 0000-0002-8912-686X (H. Kino); 0000-0001-8252-7719 (H. Dam)

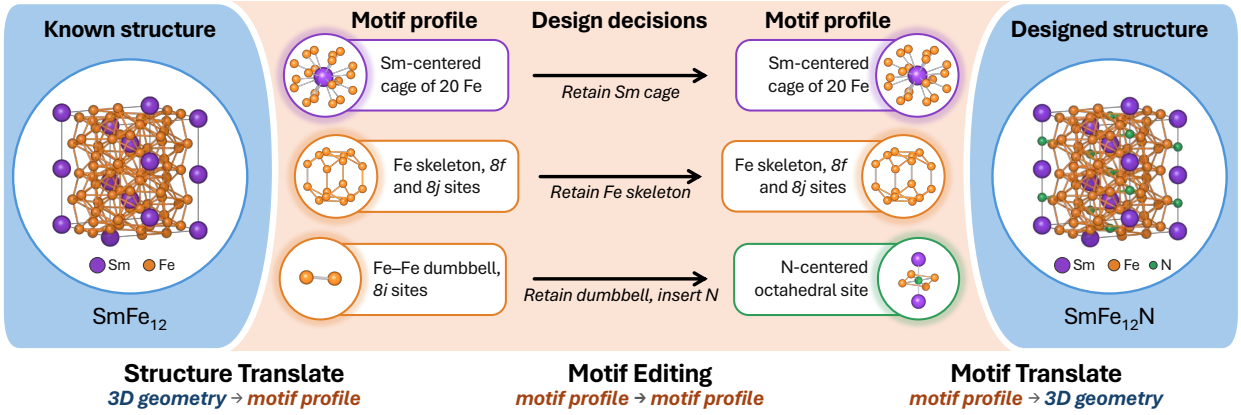


Figure 1: The motif profile as the medium of crystal design. One design step, shown for $ThMn_{12}$-type $SmFe_{12}$ and its interstitial nitride $SmFe_{12}N$. Left: Structure Translate, an agent drawing on crystallographic knowledge, reads the whole known crystal and writes its motif profile, a text that names the motifs from which the crystal is built—here a Sm-centered cage of 20 Fe atoms, Fe–Fe dumbbells, and the Fe skeleton that links them. The images beside the motifs illustrate what the text names; the crystal is not cut into fragments. Center: Motif editing edits the profile; each decision (retain, insert) carries a rationale, in quotation marks, drawn from domain knowledge. Right: Motif Translate builds the whole three-dimensional structure from the edited profile.

layer, ranking candidates using surrogate estimates such as predicted stability or uncertainty to determine which candidates to compute or synthesize next, and can automate the experiments themselves [22, 23].

These inductive methods are most reliable where they interpolate: within the range of known materials they predict well, but their accuracy tends to decrease beyond it [24]—and that is where genuinely new materials lie. A secondary trade-off concerns interpretability: the most accurate surrogates are often the least transparent, whereas interpretable descriptors such as symbolic regression [25] can recover physical insight at some cost in accuracy. The catch is what counts as *new*. A generator samples from a distribution fitted to known structures; therefore, a compound merely absent from current databases is readily obtained. A genuinely new *prototype*—an arrangement unlike those in the training set—tends to lie in the low-probability tail of that distribution and is reached only rarely [26]. Moreover, a likely draw provides little rationale for why it should meet the target property. Such methods generate readily within the known space but less often reach a new, better prototype—a gap that the next paradigm may help close.

The emerging agent-based approach addresses that task. At its center is an *agent*—a large language model directed not by a task-specific dataset but by a goal stated in words [2, 27]. It reasons abductively, drawing on the human knowledge absorbed during training to propose a plausible candidate and reason for it, which a subsequent test then confirms or refutes. Such agents already operate at research scale: language-model agents have autonomously identified real software vulnerabilities, agent-proposed scientific hypotheses have been confirmed in the laboratory [28, 29], and agent systems such as ChatMOF answer materials queries and generate metal–organic frameworks by coordinating search, prediction, and generation tools [30]. This approach is widely argued to mark an emerging fifth paradigm [3, 4]; the agent is itself trained on data, but a single agent can serve many goals, whereas the data-driven paradigm fits a new model to each task. Because the agent verbalizes its reasoning, it and the scientist share a medium through which the exchange can run in both directions: a scientist can shape the agent's proposals, and the agent makes its reasoning transparent in return. This reciprocity is something that one-directional couplings of human knowledge to a model, such as physics-informed learning [31], cannot readily offer.

In materials design, the goal is to identify a compound that is both stable and exhibits a target property. At a fixed composition, both are governed largely by how its atoms are arranged—the crystal's geometry. Geometry is something an agent can handle in words: a language model can solve an Olympiad geometry problem from its written statement, proposing the constructions required by its proof [32]. A crystal can be translated into the same medium [33]: as the set of local motifs from which it is built, its *motif profile*. From this text, an agent proposes edits that recombine and mutate parent motifs, as in an evolutionary search, but guided by a stated, target-directed reason rather than the fixed, target-blind operators of structure-search methods such as USPEX [7] and CALYPSO [8]. The agent draws on knowledge far broader than any property-labeled dataset; therefore, its edits are not confined to the structures in that dataset, whereas data-driven generators mostly produce structures close to those they were trained on [34].

Prior work has established each side of this crystal–text translation. Automated description turns a crystal into text: Robocrystallographer writes a structure as prose that reports its space group and prototype [35]. Text-conditioned generation turns a specification back into a crystal: CrysText produces a structure from its composition and symmetry [36]. Textual descriptions further act as machine-readable representations for property prediction [37, 38],

with the choice of representation itself shaping model performance [39]. Agentic loops have been demonstrated as well: FunSearch evolves computer programs [40], LLMatDesign and MatAgent iterate over chemical compositions [41, 42], and MatLLMSearch evolves raw crystal encodings [43]. The closest of these, MatAgent, has a language model reason about which composition to try but leaves the atomic arrangement to a diffusion model trained on known crystals [42]. Together, these studies provide every component: a crystal can be written as text, text can be rebuilt into a crystal, and an agent can steer an iterative search. They point to a natural next step—a framework in which text itself becomes the medium of materials design.

Herein, we present MatEvolve, a framework whose design medium is a *language of motifs*: a crystal is described as a *motif profile*, an agent edits the profile, and a new crystal is built from the edited profile (Fig. 1). The profile is a human-readable text that records the recurring geometric patterns—the *motifs*—from which the crystal is built. This describe–edit–build cycle repeats as an evolutionary search: parent profiles are recombined or mutated with a stated rationale, an evidence-based recommender decides where the limited first-principles budget is spent, and each validated candidate joins the pool from which the parents of the next generation are drawn. A profile names the building blocks themselves—withholding the coordinates, lattice parameters, space group, and structure-type label that a crystallographic description would report. However, it is accepted only if a crystal containing its motifs can be built and, when the profile is translated from a known crystal, only if that crystal is recovered in a round-trip test. Thus, the agent edits the architecture of a crystal, not its coordinates.

2. Results

2.1. Framework overview

MatEvolve is an agent-based framework for materials *design* through evolutionary search. Rather than screening a fixed or sampled candidate pool, it evolves new crystals over an interpretable text representation—the *motif profile*—across successive generations and validates its designs from first principles. It is designed to be retargetable, with a new objective set by restating it rather than by re-engineering the pipeline. Each generation runs as an ordered pipeline: *Structure Translate* writes each parent crystal as its motif profile; *Motif editing* edits those parent profiles into offspring profiles; *Motif Translate* realizes each offspring profile as a crystal; an evidence-based recommender system (ERS) ranks the offspring; and DFT validates the top-ranked candidates (*Structure Verification*), which then join the population from which the parents of the next generation are drawn. Structure Translate and Motif Translate share two basis agents that translate between a crystal and text in opposite directions—CIF2Text, which reads a structure and writes its profile, and Text2CIF, which generates a structure from a profile. The remainder of this section describes the design of each stage; implementation details are deferred to the Methods.

The motif profile itself is at the core of the framework. A crystal is written as a set of distinct structural motifs from which it is built, together with its composition and no atomic coordinates, lattice parameters, space-group symbol, or structure-type label. A motif is a recurring structural building block, ranging from a local coordination environment—a polyhedron, cage, or bonded cluster—to an extended framework such as a net or layer, each named in a controlled natural-language vocabulary according to its constituent species, geometry, and bond-length regime (Supplementary Information S1). For example, a rare-earth–iron compound may be profiled as a samarium-centered high-coordination (Frank–Kasper-type) cage, an extended iron framework with short Fe–Fe contacts, and an interstitial light-element site. By naming the building blocks themselves rather than encoding coordinates, the profile is human-interpretable and directly editable—a materials scientist can read, edit, and compare it motif by motif—unlike a learned latent vector or a raw coordinate list. A profile must retain sufficient information to specify the crystal to be admissible: we require that the structure can be regenerated from it, satisfying a *bidirectional-consistency* criterion enforced by the process described next.

Structure Translate turns a crystal into an admissible motif profile, and this is where bidirectional consistency—the round-trip test—is enforced (Fig. 2b). Given a structure S, CIF2Text writes a candidate profile T; Text2CIF then reconstructs a structure S' from T alone—the same profile-to-crystal step that Motif Translate performs—and CIF2Text reads S' back as a profile T'. T is accepted only if S' matches S and every motif of T is recovered in T', by the criteria described in Methods 2. Because Text2CIF works from the profile, the agreement reflects the information carried by T; a profile from which no matching structure can be recovered is revised and, if necessary, rejected. Admissibility requires only that a faithful realization be recoverable, not that the profile uniquely determines a structure. CIF2Text and Text2CIF extend the text representation of crystals that we introduced previously [33]; their construction is given in Supplementary Information S1 and Supplementary Information S2, and the accept-or-revise loop is described in Methods 2.

Motif editing turns parent profiles into an offspring profile (Fig. 2a, step 2). A design agent reads the profiles of one or two parents—drawn from the pool according to the design objective—and reasoning over their motifs, proposes an offspring profile T_{off} by recombining or mutating them (an evolutionary operator; Supplementary Information S3) to improve a target property. *Motif Translate* then realizes the offspring profile as a crystal (Fig. 2c): Text2CIF generates a candidate structure S_{off} from T_{off} (Methods 3), CIF2Text reads it back as T'_{off}, and S_{off} is regenerated, for up to five attempts, until every designed motif is recovered, so that the design is faithfully realized.

Each generation proposes more candidates than can be afforded by first-principles validation; therefore, the ERS [44, 34] ranks them beforehand. It scores a candidate

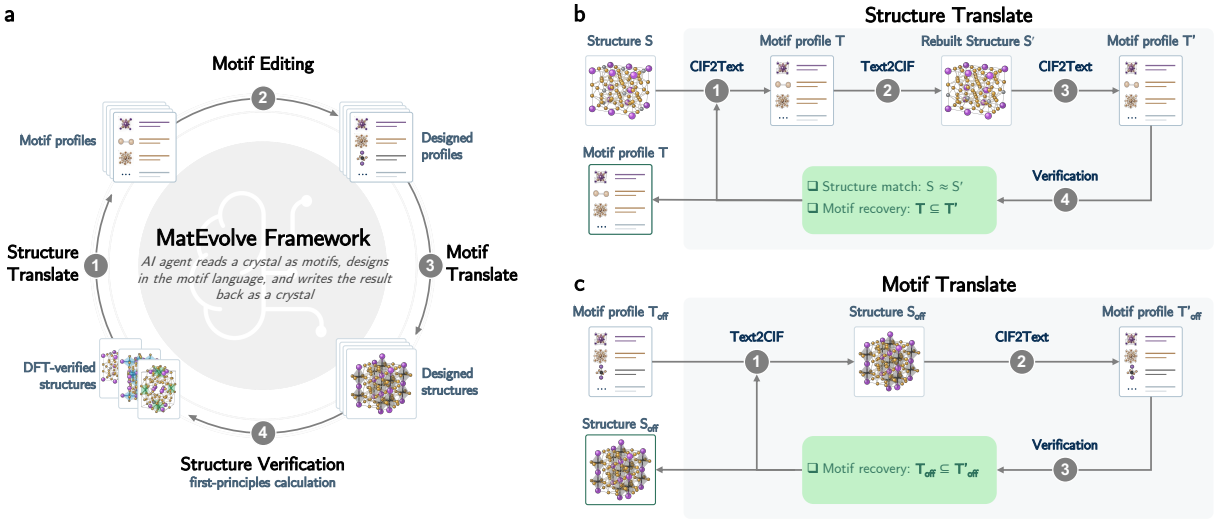


Figure 2: The MatEvolve loop and its two translation steps. (a) One generation. Parent structures from the pool are written as motif profiles (1, Structure Translate), edited by the design agent with a stated rationale (2, Motif editing), realized as three-dimensional structures (3, Motif Translate), and validated by first-principles calculation (4, Structure Verification); validated structures join the pool from which the parents of the next generation are drawn. The evidence-based recommender that ranks candidates before validation is not drawn. (b) Structure Translate of a reference crystal S: CIF2Text writes its profile T (1), Text2CIF rebuilds S' from T alone (2), and CIF2Text reads S' back as T' (3). T is accepted if S' matches S (StructureMatcher) and every motif of T is recovered in T' (4); if either check fails, T is redrafted, for up to five attempts. (c) Motif Translate of an offspring profile T_{off}: Text2CIF builds a candidate structure S_{off} (1), and CIF2Text reads it back as T'_{off} (2). S_{off} is accepted if every motif of T_{off} is recovered in T'_{off} (3); otherwise it is rebuilt, for up to five attempts.

not from its individual motifs but from its *motif groups*—families of related motifs obtained by clustering them across the dataset (Methods 4). Through Dempster–Shafer evidence theory [45, 46], it accumulates evidence of how reliably those groups and their combinations have been associated with the design target across the labeled structures. Candidates are ranked by this evidence so that costly validation is focused on the most promising ones, and DFT then relaxes and evaluates each selected candidate. The evidence model and the first-principles validation protocol are described in Methods 5 and Methods 6.

2.2. Experimental design and settings

We evaluated MatEvolve in two experiments. Experiment 1 investigates whether a crystal can be written as a motif profile and rebuilt from that profile alone. Experiment 2 verifies whether editing profiles yields stable, novel crystals that are magnet compounds and new prototypes at rates that are comparable to generative models under an equal validation budget, and how much of that performance comes from the recommender that selects candidates for validation. A descriptive analysis of the motif content of the resulting structures, prompted by the outcome of Experiment 2, follows in Section 2.5. All agents ran on a single frozen commercial language model (Claude Fable 5, API identifier `claude-fable-5`; decoding temperature 0.5) that was identical across the experiments.

Datasets Two datasets underpinned the experiments: one seeded the search, the other defined novelty.

- **Initial population,** $\mathcal{D}_{\mathrm{seed}}$ — the 948 magnetic structures from which the search started were curated from the Materials Project [5] to cover permanent-magnet and magnetic-intermetallic families: compounds pairing a rare-earth element (Ce, Pr, Nd, Sm, Gd, Tb, Dy, Ho, Er, Y) with a transition metal (Fe, Co, Mn, Ni), with total magnetization above 10 μ_B per formula unit, energy above the convex hull of at most 0.2 eV/atom, and converged DFT energies (tolerance below 10^{-5} eV/atom). Of these, 208 are thermodynamically stable ($E_{\mathrm{hull}} = 0$); by magnetic ordering, it included 282 ferromagnetic and 616 ferrimagnetic structures, with the remaining 50 unresolved in the source data. Compositions ranged from binary to quinary with 5–100 atoms per cell, including $ThMn_{12}$-type intermetallics, RCo_5 compounds, $R_2T_{14}B$ compounds, Heusler alloys, and perovskite oxides.

- **Novelty reference,** $\mathcal{D}_{\mathrm{ref}}$ — the set against which novelty was judged: the entire Materials Project [5] together with the ∼5 million structures in the Alexandria database [47]. A candidate was counted as novel only if it matched nothing in $\mathcal{D}_{\mathrm{ref}}$, which included the seed; thus, rediscovering a seed structure earned no credit.

Experiment 1: translation fidelity Experiment 1 tests the translation between crystal and text in both directions: from a crystal to its motif profile, and from that profile back to a crystal. Each of the 948 seed structures was translated

in both directions (Methods 2): Structure Translate wrote its motif profile, Motif Translate rebuilt a structure from that profile alone, and the profile was revised, up to five times, until the rebuilt structure passed the consistency tests of Methods 2. Each final reconstruction, as generated and without relaxation by DFT, was then compared with its original, the DFT-relaxed Materials Project structure, in three respects: the match rate, by pymatgen's `StructureMatcher` at the standard tolerance under which generative models report match rates (`ltol` 0.3, `stol` 0.5, `angle_tol` 10°); geometric and shape fidelity, by the root-mean-square displacement (RMSD) and the continuous shape measure (CShM) [48, 49]; and the local atomic environments, by the orbital-field matrix (OFM) [50, 51]; the measures are defined in Supplementary Information S4.2.

Experiment 2: design benchmark Experiment 2 ran the full loop for four generations. Each generation produced 200 proposals, of which the ERS, ranking on stability (Methods 5), selected 100 for DFT validation; the run length and both budgets were fixed in advance by the available DFT resources. Therefore, four generations resulted in a proposal pool of 800 and benchmark set of 400 validated candidates. The remaining 400 proposals, which the ERS had not selected, were relaxed after the run had completed and excluded from the hull reference of the benchmark; comparing the selected and unselected halves of the same pool indicates how much of the benchmark performance comes from the recommender and how much from motif editing.

MatEvolve was compared with two recent generative models: MatterGen [17], in both its magnetization-conditioned and unconditioned (Materials Project + Alexandria) forms, and the unconditioned Chemeleon2 [52]. Every method received the same budget of 400 DFT-validated candidates, drawn from a disclosed proposal pool by its own ranking procedure: each baseline generated 1,000 candidates by its native sampling procedure, ranked them with CHGNet based on formation energy and predicted magnetic moment, and had its top 400 candidates validated.

Evaluation metrics Experiment 2 scored every method, and both MatEvolve halves, on the same four rates, each reported as a number of candidates and as a percentage of the method's 400 relaxations; where two methods come close, a footnote gives Fisher's exact test for their difference. The S.U.N. rate, the standard criterion of the structure-generation literature [15, 17], counts candidates that are *stable* (energy above the convex hull $E_{\mathrm{hull}} \leq 0.1$ eV/atom, that is, on the hull or metastable within 0.1 eV/atom of it), *unique* (within the method's own set), and *novel* (against $\mathcal{D}_{\mathrm{ref}}$). The design target, a rare-earth-lean magnet compound beyond the known structure types, adds two requirements, assessed separately. The *magnetic rate* counts the S.U.N. candidates with a saturation magnetization $\mu_0 M_s$ above 1 T, a value of permanent-magnet relevance. The *new-prototype rate* counts the S.U.N. candidates whose AFLOW prototype label [53, 54]—anonymized composition, Pearson symbol, space group and Wyckoff positions—does not appear in $\mathcal{D}_{\mathrm{ref}}$; therefore, substituting elements on a known lattice is invalid. Because the design target asks for both at once, a fourth rate—the *joint rate*—counts the S.U.N. candidates that are new prototypes and magnetic at the same time.

2.3. From crystal to motif profile and back

We first tested the translation on which the framework rests in both directions: every crystal of $\mathcal{D}_{\mathrm{seed}}$ was written as a motif profile by Structure Translate, and a structure was rebuilt from that profile alone by Motif Translate. Because the profile withholds the coordinates, lattice parameters, space group, and structure-type label, the question is whether what remains suffices to recover the crystal. Of the 948 reconstructions of $\mathcal{D}_{\mathrm{seed}}$ (Section 2.2), 59% matched their originals at the tolerances conventionally used to evaluate deep-learning structure generators. Those generators report match rates of approximately 34% to 64% under the same settings on MP-20, a standard benchmark of Materials Project structures with at most 20 atoms per cell [15, 16, 18]; the lower end is set by autoencoder reconstruction, the task closest to ours [15]. The rate lies within that range; however, the comparison is only indicative: our reconstructions benefit from up to five profile revisions but involve cells of up to 100 atoms, five times the MP-20 limit.

Figure 3a shows eight reconstructions beside their originals, spanning the range of outcomes. Four were rebuilt almost exactly: the $ThMn_{12}$-type $Gd(Al_2Fe)_4$, the filled skutterudite $Pr(FeSb_3)_4$, Tb_6FeBi_2, and the layered $Ce_5(FeB_3)_2$, with RMSDs of 0.013–0.056 Å. The boride $DyFeB_4$ was displaced further, by 0.174 Å, yet reproduced the boron network of its original. The three largest departures ($Gd_2Fe_{12}P_7$, the 2:17-type hydride $Tb_2Fe_{17}H_3$, and $SmEr_3Fe_{34}$) have RMSDs of 0.75, 0.98, and 1.13 Å, respectively: their atoms sit visibly off the original positions, but each reconstruction still contains the motifs of its original, as the motif-coverage test of Structure Translate requires.

These displacements measure mainly what the profile leaves free. A motif profile names the motifs a crystal contains, how they are arranged, and the bond-length regime of each, but gives no numerical bond length, coordinate, or lattice parameter. Therefore, Motif Translate must place the atoms, and its placement need not match the original. Scoring was also conservative: reconstructions were taken as Motif Translate emitted them, whereas the originals are DFT-relaxed; thus, part of each displacement may reflect relaxation alone. Over the matched reconstructions, the median RMSD was 0.025 Å and the mean 0.16 Å (Fig. 3b, left). Because the RMSD is an absolute measure, we also report a size-normalized measure of shape.

The continuous shape measure (CShM) scores the deviation of one structure from another after optimal superposition, on a scale from 0, for an identical match, to 100; it normalizes the squared displacement by the spread of the point set, so that structures of different size are compared on one scale (Supplementary Information S4.2). Coordination

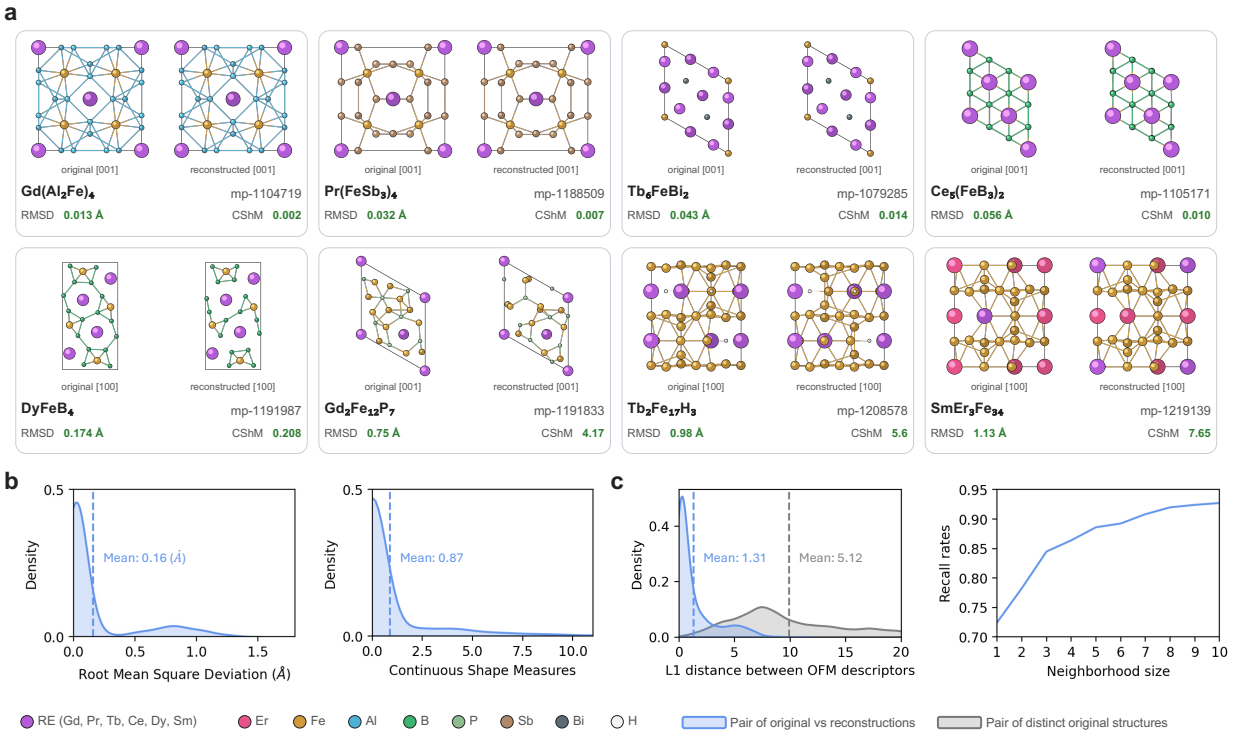


Figure 3: Fidelity of the translation from crystal to motif profile and back. (a) Eight representative reconstructions spanning the range of outcomes, each viewed along [001] or [100] as labeled. In each pair, the left denotes the original Materials Project structure, and the right denotes its reconstruction by Motif Translate from the motif profile. RMSD (Å) is the root-mean-square displacement between corresponding atoms after optimal superposition, and CShM the dimensionless continuous shape measure (0–100) of the reconstruction relative to the original [48, 49]. (b) Distributions of the RMSD (left) and CShM (right) over the matched reconstructions; dashed lines mark the means. (c) Fidelity of the local atomic environments, measured with the OFM [50, 51]. Left: distribution of the L_1 distance between the OFM descriptors of each original and its reconstruction (blue; all original–reconstruction pairs of $\mathcal{D}_{\text{seed}}$), against the distribution of L_1 distances from each original to its 100 nearest neighbors among the other originals of $\mathcal{D}_{\text{seed}}$ (gray); dashed lines mark the means. Right: fraction of originals whose own reconstruction is among their k nearest OFM neighbors (recall), as a function of k; neighbors are sought among all other structures, that is, the remaining originals and all reconstructions.

chemistry reads it in bands: essentially ideal below approximately 0.1, slightly distorted up to 3, notably distorted from 3 to 7.5, and a breakdown of the target geometry is considered beyond 7.5 [48, 55, 56, 57]. Those bands were calibrated on single coordination polyhedra. Because a unit cell is more extended than a polyhedron, a given displacement yields a smaller CShM here; therefore, we use the bands to describe the distribution rather than to judge it. Across the matched reconstructions, 82% fell in the essentially ideal band and a further 7% in the slightly distorted band; 8% showed notable distortion and 3% lay beyond the breakdown boundary. The median CShM was 0.0054 and the mean was 0.87 (Fig. 3b, right), the mean was set by the tail. Over the reconstructions of $\mathcal{D}_{\text{seed}}$ as a whole, not only the matched ones, the means rise to 0.43 Å for the RMSD and 3.94 for the CShM. The eight examples of Fig. 3a place the bands on real structures: 0.002–0.014 for the four rebuilt almost exactly, 0.21 for $DyFeB_4$, and 4.17, 5.60, and 7.65 for the three largest departures, the last lying just beyond the breakdown boundary.

The rebuilt structures also recover the local atomic environments of each crystal—the coordination of each atom, which governs stability, magnetic moment, and anisotropy. We compared them with the OFM [50, 51], which encodes this coordination, needs no atom correspondence, and applies to all 948 original–reconstruction pairs. The median L_1 distance between the OFM descriptors of an original and its reconstruction was 0.44 (mean 1.31), against a median of 8.66 (mean 9.95) from each original to its 100 nearest neighbors among the other originals of $\mathcal{D}_{\text{seed}}$ (Fig. 3c, left). The reconstructions are also specifically close: for 72.5% of the originals, the single nearest neighbor among all other structures—the remaining originals and every reconstruction—was their own reconstruction, and for 92.7% it lay among the 10 nearest (Fig. 3c, right). Because the OFM also encodes chemical identity, part of this specificity reflects the shared composition rather than structure alone.

This experiment is the proof of concept on which the rest of the framework rests, in both directions: a crystal can be written as a motif profile, and that profile alone suffices

Table 1
Benchmark of MatEvolve against generative models on magnetic-material design. Each method received 400 converged DFT relaxations, spent on the candidates its own ranking selected: for each baseline, the top 400 of 1,000 samples by CHGNet formation energy and magnetic moment; for MatEvolve, the top 100 of 200 proposals in each of four generations, ranked by the ERS. The second row holds the 400 MatEvolve proposals the ERS did not select, relaxed with the same DFT protocol. S.U.N.: stable ($E_{\text{hull}} \leq 0.1$ eV/atom), unique within the set, and novel with respect to the Materials Project and Alexandria. Magnetic: S.U.N. candidates with $\mu_0 M_s > 1$ T. New prototype: S.U.N. candidates whose prototype is absent from both databases. Both: S.U.N. candidates that are at the same time new prototypes and magnetic, the two requirements of the design target taken together. Bold text indicates the highest rate in each column.

Method	S.U.N. candidates	S.U.N. candidates meeting the design target		
		Magnetic	New prototype	Both
MatEvolve (selected by the ERS)	**115 (28.8%)**	**104 (26.0%)**	**89 (22.3%)**	**80 (20.0%)**
MatEvolve (not selected by the ERS)	97 (24.3%)	80 (20.0%)	53 (13.3%)	48 (12.0%)
Chemeleon2 (unconditioned)	82 (20.5%)	5 (1.3%)	11 (2.8%)	3 (0.8%)
MatterGen (unconditioned)	112 (28.0%)	8 (2.0%)	22 (5.5%)	2 (0.5%)
MatterGen (magnetization-conditioned)	94 (23.5%)	88 (22.0%)	5 (1.3%)	5 (1.3%)

Note: Each cell gives the number of candidates that meet the criterion, followed in parentheses by that number as a percentage of the method's 400 relaxations, that is, the rate. Novelty is decided by structure matching against all Materials Project and Alexandria entries (pymatgen `StructureMatcher`, database de-duplication tolerances); prototype status by the AFLOW label [53, 54] of the symmetrized structure, matched against the labels of all such entries at a common symmetry tolerance (0.1 Å).

to build the crystal back. Carrying no coordinates, lattice parameters, space group, or structure-type label, the profile still captures what the crystal is built from—its motifs, their arrangement and their coordination. Thus, the profile is not merely a description of a crystal but a specification from which one can be built: editing it specifies an offspring's motifs rather than its coordinates, and the placement is settled downstream by M3GNet pre-relaxation (Supplementary Information S2) and first-principles relaxation.

2.4. Benchmarking and discovery of new stable magnets

MatEvolve was benchmarked against two generative models, MatterGen [17] and Chemeleon2 [52], under the protocol of Section 2.2. The ranking score of each method—the ERS for MatEvolve, CHGNet formation energy, and magnetic moment for the baselines—decided only which candidates were validated by DFT. Every rate below was determined from the DFT-relaxed structures and is given as a percentage of the 400 relaxations per method (Table 1). Four rates were compared. The S.U.N. rate counts the candidates that are stable ($E_{\text{hull}} \leq 0.1$ eV/atom), unique, and novel with respect to the Materials Project and Alexandria. The magnetic rate counts the S.U.N. candidates with $\mu_0 M_s > 1$ T, and the new-prototype rate those whose prototype is absent from both databases; these two express the two requirements of the design target, and the joint rate counts the candidates that meet both at once.

Table 1 shows the four rates of each method: on the S.U.N. and magnetic rates MatEvolve is matched by one of the baselines, whereas on the new-prototype and joint rates it leads all of them clearly. On the S.U.N. rate, MatEvolve's 115 candidates (29%) and unconditioned MatterGen's 112 (28%) are statistically indistinguishable,[1] with magnetization-conditioned MatterGen at 24% and Chemeleon2 at 21%. Among these S.U.N. candidates, MatEvolve has 104 magnet compounds (26% of its 400 relaxations) and conditioned MatterGen has 88 (22%), a gap not resolved at this sample size,[2] so the two are comparable on magnet compounds; unconditioned MatterGen and Chemeleon2, trained for stability alone, have only 8 (2%) and 5 (1%). By contrast, 89 of MatEvolve's S.U.N. candidates (22%) are new prototypes, against 22 (6%), 11 (3%), and 5 (1%) for unconditioned MatterGen, Chemeleon2, and conditioned MatterGen, respectively; the baselines' S.U.N. candidates are mostly new compounds on known prototypes, as expected of generators trained to reproduce known crystals. Requiring both properties at once, as the design target does, separates the methods most sharply: 80 of MatEvolve's candidates (20%) are both magnetic and new prototypes, against 5 (1.3%) for conditioned MatterGen, 3 (0.8%) for Chemeleon2, and 2 (0.5%) for unconditioned MatterGen—sixteen times the best baseline.

MatEvolve's rates combine two contributions: the motif editing that proposes candidates and ERS that selects which of them to validate. To separate the two, the 400 proposals the ERS did not select were relaxed after the run with the same DFT protocol (Table 1, second row). Of these, 97 (24%) are S.U.N., 53 (13%) are new prototypes, and 80 (20%) are magnetic, against 115, 89, and 104 for the selected set, respectively; the largest difference is in new prototypes (13% against 22%). The ERS ranks on stability alone; thus, the gain in stable candidates (115 against 97) is its expected effect. Because the other two rates count subsets of the stable candidates, their gain may follow from it rather than from

[1] MatEvolve produced 3 more S.U.N. candidates than unconditioned MatterGen (115 against 112, out of 400 each). At this sample size, a difference of 3 is not statistically significant (two-sided Fisher's exact test, $p = 0.88$), so neither method is shown to be higher.

[2] MatEvolve produced 16 more magnetic candidates than conditioned MatterGen (104 against 88, out of 400 each). At this sample size, a difference of 16 is not statistically significant (two-sided Fisher's exact test, $p = 0.21$), so MatEvolve is not shown to be higher.

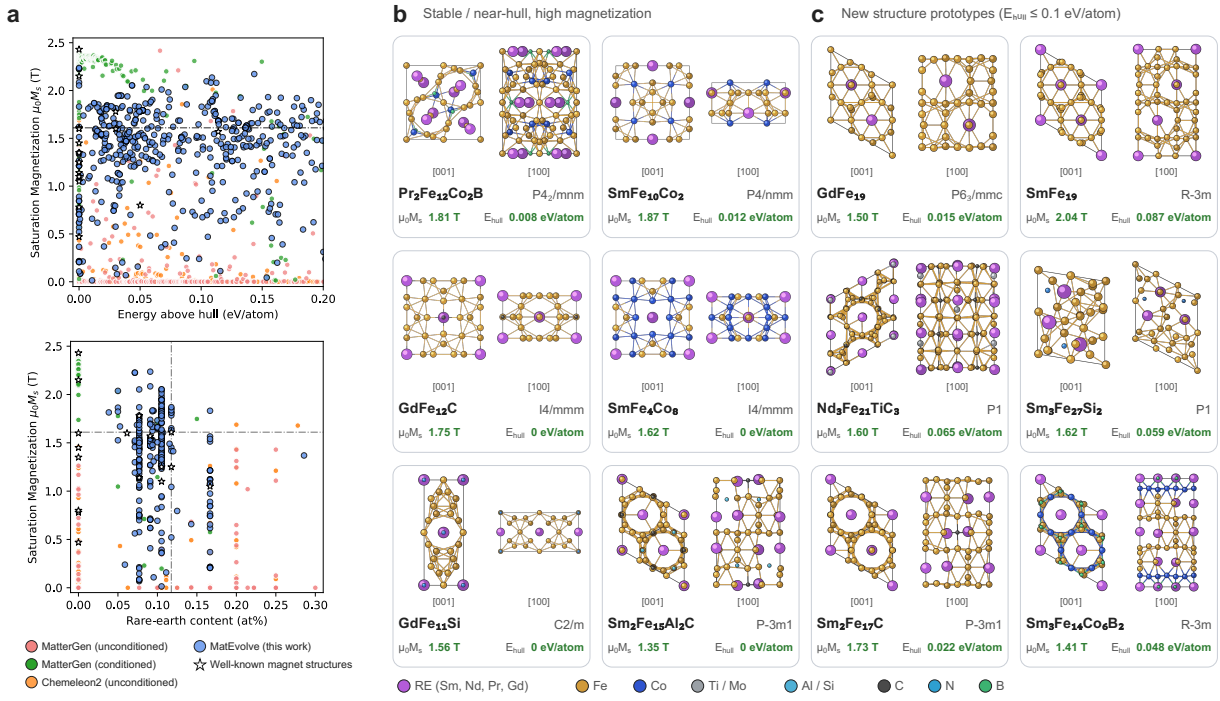


Figure 4: Magnetization, stability, and rare-earth content of the validated candidates, with representative MatEvolve structures. (a) Saturation magnetization $\mu_0 M_s$ versus energy above the convex hull for the DFT-validated candidates of each method (top) and versus the rare-earth atomic fraction for the (meta)stable ones only, with $E_{\text{hull}} \leq 0.1$ eV/atom (bottom): MatEvolve (blue), unconditioned MatterGen (pink), magnetization-conditioned MatterGen (green), and Chemeleon2 (orange). Gray stars: well-known magnet compounds, taken from the Materials Project. Dash-dot lines: $\mu_0 M_s$ and rare-earth fraction of $Nd_2Fe_{14}B$. (b) Six structures of known prototypes generated by MatEvolve, each with $\mu_0 M_s > 1$ T and lying on or near the hull. (c) Six structures generated by MatEvolve that are both new prototypes and magnetic ($\mu_0 M_s > 1$ T; $0 < E_{\text{hull}} \leq 0.1$ eV/atom). In (b) and (c), each DFT-relaxed structure is viewed along [001] and [100] and labeled with its formula, space group, $\mu_0 M_s$, and E_{hull}.

any preference of the ERS for new prototypes or magnet compounds. What the nonselected proposals show is that motif editing alone, before any selection, already yields new prototypes at 13%, more than twice the best baseline, and magnet compounds at 20%, level with conditioned MatterGen; the ERS then raises both yields further.

Beyond these rates, Figure 4a examines the validated candidates further, by magnetization, stability, and rare-earth content. Only MatEvolve and conditioned MatterGen produce high-magnetization candidates in appreciable numbers (upper panel): the unconditioned baselines lie mostly below 0.5 T, and the few Chemeleon2 candidates near or above 1 T are rare-earth-rich (fractions of 0.17–0.30). Conditioned MatterGen reaches the highest values, up to about 2.4 T near the hull; MatEvolve follows, with many of its candidates above the 1.6 T of $Nd_2Fe_{14}B$ and the highest at 2.2 T. The lower panel keeps only the (meta)stable candidates ($E_{\text{hull}} \leq 0.1$ eV/atom) and plots them against the rare-earth fraction, which matters because a permanent magnet needs some rare earth—its $4f$ electrons supply the magnetocrystalline anisotropy that makes the magnet hard—but as little as possible because rare-earth materials are scarce and costly. Here, the two methods part: conditioned MatterGen's high magnetizations lie almost entirely at zero rare-earth content, in Fe–Co-rich compositions similar to the soft magnetic materials α-Fe and $Fe_{65}Co_{35}$; however, all of MatEvolve's (meta)stable candidates above 1 T contain rare earth, and 88% of them lie at fractions up to 0.12, the range of the well-known magnet compounds $SmFe_{11}Ti$ and $Sm(Fe,Co)_{12}$ (0.08), $Sm_2Fe_{17}N_3$ (0.09), and $Nd_2Fe_{14}B$ (0.12).

Figure 4b shows six candidates that exceed 1 T, lie on or near the hull, and are new to $\mathcal{D}_{\text{ref}}$ as ordered compounds but not as prototypes. All belong to the principal Fe-rich families studied for rare-earth permanent magnets [58], each represented in $\mathcal{D}_{\text{seed}}$, so they test physical plausibility rather than demonstrate discovery. Four are $ThMn_{12}$-type 1:12 derivatives: $GdFe_{12}C$, with interstitial carbon (2.22 T); $GdFe_{11}Si$, with Si substitution (2.05 T); and $SmFe_{10}Co_2$ and $SmFe_4Co_8$ (1.87 and 1.62 T), ordered variants of the known $Sm(Fe,Co)_{12}$ system with Fe and Co on distinct sites. The other two are $Pr_2Fe_{12}Co_2B$, a Co-substituted $Nd_2Fe_{14}B$-type 2:14:1 phase (1.81 T), and $Sm_2Fe_{15}Al_2C$, a Th_2Zn_{17}-related 2:17 carbide with Al substitution (1.35 T). The interstitial and substituent elements (C, Co, Si, Al) are those used in practice to stabilize or tune these families.

Figure 4c shows six candidates that are metastable ($0 < E_{\text{hull}} \leq 0.1$ eV/atom), exceed 1 T, and realize a prototype absent from $\mathcal{D}_{\text{ref}}$. They reach their new prototypes in two ways. The first is *dumbbell substitution*, the mechanism

behind the known Fe-rich families: replacing a fraction s of the R atoms of $CaCu_5$-type RT_5 by Fe–Fe dumbbells gives the $R_{1-s}Fe_{5+2s}$ series, of which R_2Fe_{17} ($s = 1/3$) and $ThMn_{12}$-type RFe_{12} ($s = 1/2$) are established members [58]. Three candidates extend this series to members absent from $\mathcal{D}_{\mathrm{ref}}$. $GdFe_{19}$ (1.82 T) and $SmFe_{19}$ (2.04 T) lie at $s = 2/3$, beyond RFe_{12}; they adopt the space groups of the two R_2Fe_{17} polytypes ($P6_3/mmc$ and $R\bar{3}m$) and contain isolated Fe–Fe pairs at 2.36–2.39 Å, the dumbbell distance of the 2:17 phases. $Sm_3Fe_{27}Si_2$ (1.62 T), triclinic, has the 3:29 stoichiometry ($s = 2/5$) with Si on transition-metal sites and likewise contains isolated pairs. The second approach is *motif substitution*: a motif of a known structure is replaced by, or supplemented with, a different one. $Sm_2Fe_{17}C$ (1.73 T) fills every third 2:17 block with carbon, as $Sm_2Fe_{17}N_3$ is filled with nitrogen [59], the carbon occupying the same octahedral sites of four Fe and two Sm. $Nd_3Fe_{21}TiC_3$ (1.60 T) places carbon in similar octahedral interstices within a Nd–Fe–Ti framework. $Sm_3Fe_{14}Co_6B_2$ (1.41 T) inserts boron-centered Co_6 trigonal prisms, the Co analog of the Fe_6B prisms of $Nd_2Fe_{14}B$ [60], into an Fe–Co framework. In all six, the local environments—dumbbells, C- and B-centered sites—are known motifs or close variants of them; the novelty lies in their arrangement.

$Sm_2Fe_{17}C$ also illustrates the degree to which two arrangements of one composition can differ. $Sm_2Fe_{17}C_x$ carbides are experimentally established [61], and the composition is already in $\mathcal{D}_{\mathrm{ref}}$ as mp-1219267, a monoclinic $C2/m$ structure with a different carbon arrangement, 0.28 eV/atom above the hull at 0.064 T; the generated $P\bar{3}m1$ structure lies 0.02 eV/atom above the hull at 1.73 T. Whether the difference stems from the carbon arrangement or from the magnetic states to which the two calculations converged cannot be ascertained from this pair alone.

Taken together, these results show that editing crystals in the language of motifs reaches both requirements of the target simultaneously, a new prototype that is also a magnet compound, at a rate that generators sampling the distribution of known crystals do not approach. It did so with the target written into the agents' prompts rather than trained into a model, which suggests that the same framework could be redirected to other design targets in materials discovery by restating the objective; however, this remains to be tested. In the new prototypes examined above, the novelty lies not in new motifs but in a new arrangement of known ones; the question we next address is whether that holds across all new-prototype candidates.

2.5. Structural insights into novelty and stability through motif analysis

The interpretable motif profile allows us to return to the question $Sm_2Fe_{17}C$ posed and to characterize the discoveries, not merely count them: for the structures the run produced—all 800, novel or not—we can ask where the novelty of the new prototypes resides and what underlies their stability. Both questions turn on the local motifs from which the structures are built; therefore, we first establish a common vocabulary of geometric groups from the seed alone. A consensus procedure—five language-model voters casting categorical votes, with a deterministic routine performing all merging, splitting, and counting, so that the number of groups is discovered rather than fixed (Methods 4)—partitions the motifs of the seed structures into 17 groups. Each motif in the generated structures is then assigned to one of these groups using the same categorical voting procedure, or, if it matches none, it is left unassigned. This shared labeling allows the recommender to score each candidate against the stability statistics derived from the seed.

Across the study, the 11,478 motifs from all 1,748 structures were distributed across the 17 groups, apart from a small unassigned tail noted below, each of which was summarized as a word cloud in Fig. 5a by its most representative motif keywords (construction detailed in Supplementary Information S6). The groups were defined along the axes used by a crystal chemist—unit class, coordination geometry, and coordination-number band—rather than in a latent coordinate. Most groups represented a specific geometry, spanning much of the standard vocabulary of intermetallic chemistry—tetrahedral, octahedral, and dimer units; triangular-planar and square-planar frameworks; chain, square-net, kagome, honeycomb, hexagonal, and ring motifs; and icosahedral, Frank–Kasper, prismatic, and generic cages. The emergence of such a small, recognizable set was not an artefact of clustering: intermetallic coordination is known to collapse onto approximately 20 recurring environments—with the icosahedron and $Z12$–$Z16$ Frank–Kasper cages among the most common [62, 63, 64, 65]—which our clustering recovered.

Every group was broadly compatible with stability: across all 17 groups, the stable-or-metastable fraction fell within a similar range (55–68%; Fig. 5c); thus, no single geometry was uniquely stabilizing or destabilizing. Whether a structure was stable thus appeared to depend more on how its groups *combined* than on which groups it contained, which we analyzed further below. By contrast, saturation magnetization separated the groups (Fig. 5d): the highest values belonged to the kagome layer, dimer/dumbbell unit, Frank–Kasper cage, and honeycomb layer, while the dense metallic frameworks and hexagonal layer also averaged above the 1 T target. These motifs marked iron-rich environments, and the ordering tracked iron content: the transition-metal nets and dense metallic frameworks carried the itinerant $3d$ moment [66, 58], while the dimer/dumbbell and Frank–Kasper-cage groups showed high values because they marked iron-rich rare-earth–iron intermetallics—the cage housing the rare-earth atom that supplies the $4f$ anisotropy [59], with the moment coming from the surrounding iron framework. Recovering this magnetically meaningful vocabulary from geometry alone made the partition interpretable.

The vocabulary of the seeds accounted for the great majority of the motifs in the generated structures. Assigning each generated motif to one of the 17 seed groups—using the same categorical vote, which could also return *none*—mapped most of them onto an existing group, so that every

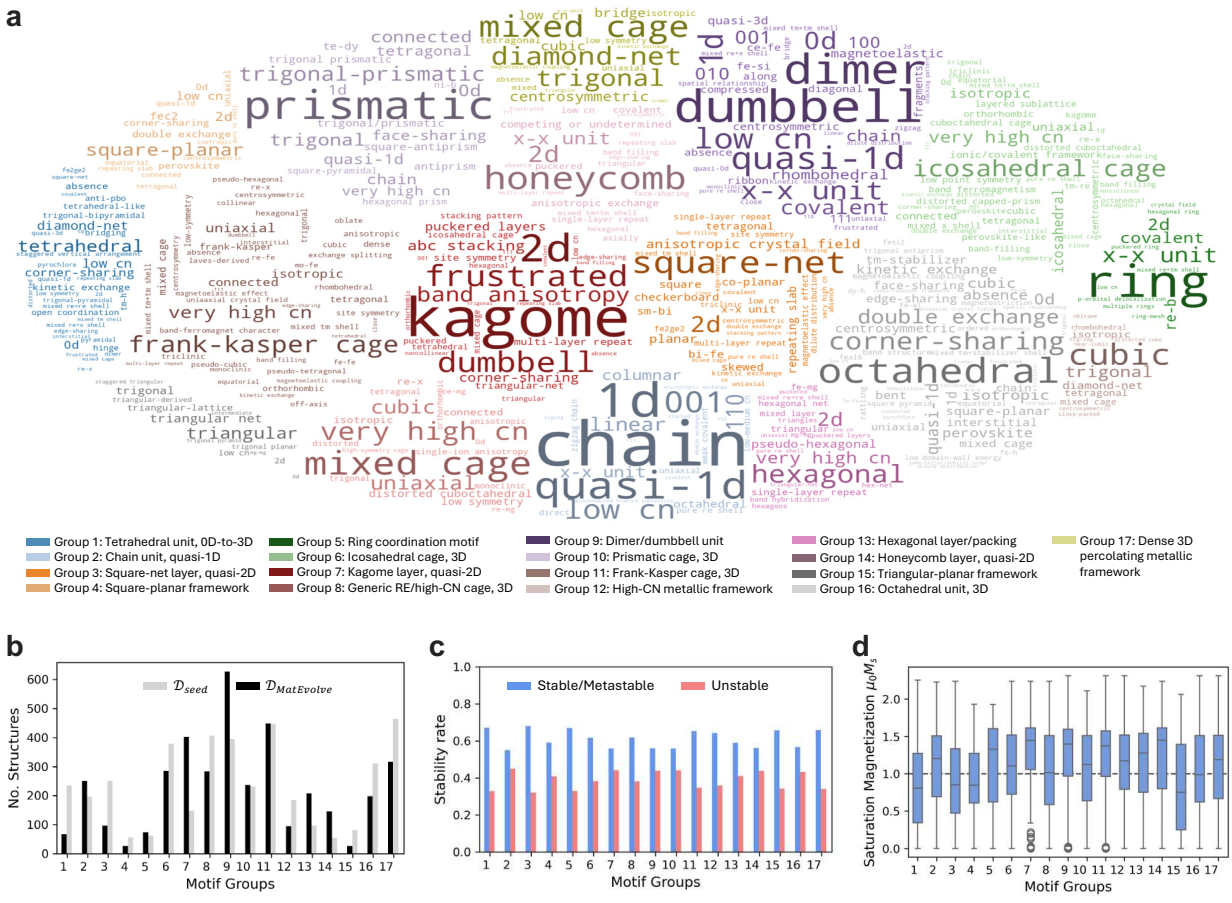


Figure 5: Motif groups discovered from the seed and assigned across the study. The groups are discovered from the seed structures' motifs, and the motifs of the generated structures are then assigned to them (11,478 motifs across all 1,748 structures). (a) Word clouds of the 17 motif groups; each group is summarized by its most representative keywords, ranked by TF–IDF over the group's motif descriptions from candidate terms proposed by the language-model agents. (b) For each group, the number of contributing structures from the Materials Project seed set ($\mathcal{D}_{\text{seed}}$, gray) and from MatEvolve ($\mathcal{D}_{\text{MatEvolve}}$, black). (c) For each group, the fraction of contributing structures that are stable or metastable (blue) versus unstable (red). (d) For each group, the distribution of saturation magnetization $\mu_0 M_s$ across the structures that contain a motif from that group; the dashed line marks the 1 T design target.

group was used by both the seed and generated set (Fig. 5b). A small fraction was instead returned unassigned—local geometries with no counterpart among the seed's groups—and were set aside from the combinatorial analysis below. The small size of this tail was important: MatEvolve's structural novelty was predominantly combinatorial, arising far more from new arrangements of the shared vocabulary than from new building blocks, as the next paragraphs quantified.

We extended this analysis to the whole-structure level by representing each structure as the set of motif groups to which its motifs were assigned. Fig. 6a maps the 1,748 structures under this motif-group representation—structures sharing groups, or mutually substitutable ones [44, 51], were placed close together (details of the map construction are given in Supplementary Information S5). The seed structures formed two dense regions (gray kernel-density contours), and the generated structures largely gathered along the border between them rather than in a region isolated from the seed. A structure combining groups from both regions was placed between them; therefore, this bridging position reflected new combinations of the existing groups. Since almost every generated motif still mapped onto a seed group rather than remaining unassigned, the generated structures recombined existing groups rather than introducing new ones.

Fig. 6b quantified this pattern by classifying the binary, ternary, and quaternary combinations of the 17 groups—the sets of two, three, and four groups—based on whether they appeared in the seed, the generated set, both, or neither. Of the binary and ternary combinations that occurred at all, almost all already appeared in the seed—only 1 of the 136 observed binary combinations and 30 of the 624 observed ternary combinations (5%) were unique to the generated set. Therefore, at low orders, the generated set introduced few new combinations. The novelty appeared at the whole-structure level, as the quaternary analysis below shows. Quaternary combinations were sampled far more sparsely: of the 2,380 possible combinations, 702 occurred in both

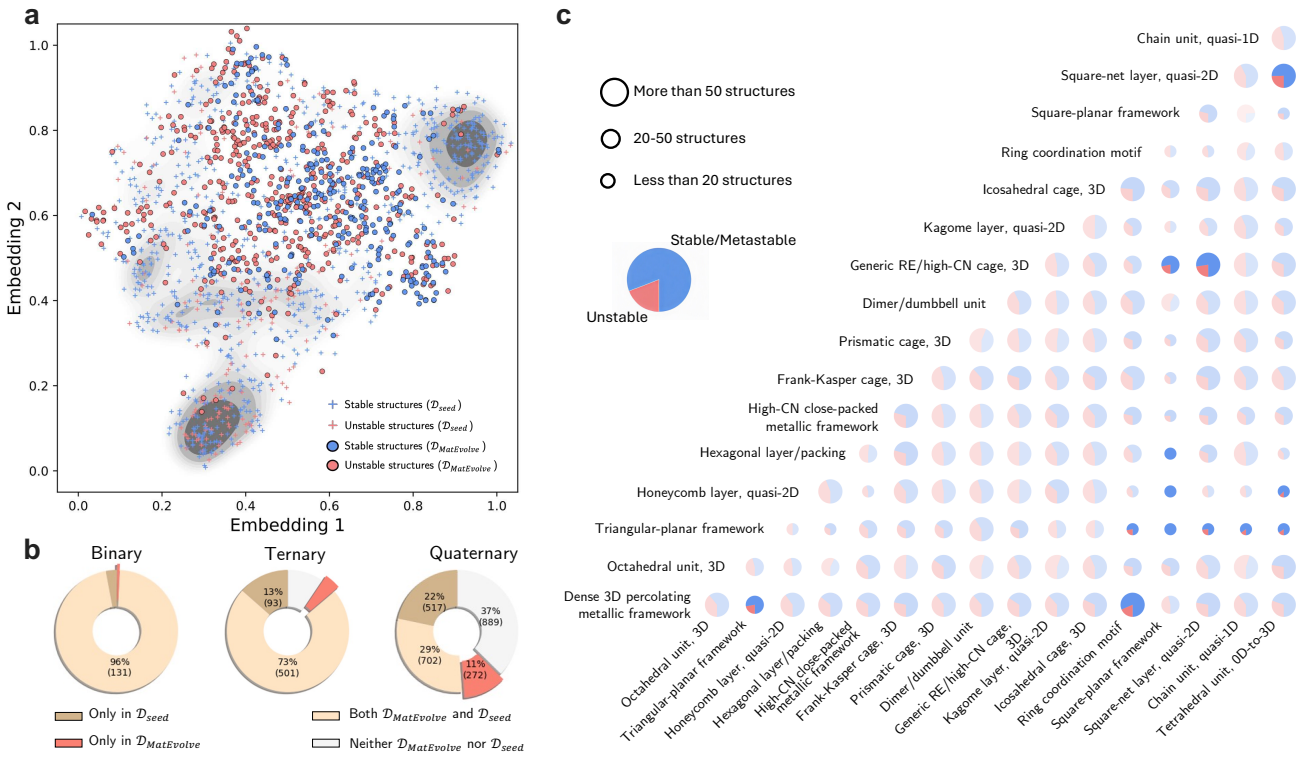


Figure 6: Motif-group composition, combination coverage and pairwise stability across the 1,748 structures. (a) Structure map: each of the 1,748 structures is represented by the set of 17 motif groups it contains and embedded so that structures sharing groups—or built from mutually substitutable groups—lie close together (t-SNE on a substitutability-weighted set-overlap distance; axes are arbitrary embedding coordinates). Marker shape encodes origin—Materials Project seed structures (crosses) versus MatEvolve-generated structures (circles). The origin is denoted by marker shape, not by shade as in Fig. 5b, and the color encodes stability: blue for stable or metastable and red for unstable; gray contours are a kernel-density estimate of the seed structures' distribution in the embedding. (b) Coverage of motif-group combinations by order: for the binary, ternary, and quaternary combinations of the 17 groups, the fraction of all possible combinations observed only in the seed set ($\mathcal{D}_{\text{seed}}$), only among the generated structures ($\mathcal{D}_{\text{MatEvolve}}$), in both, or in neither (counts in parentheses). (c) Pairwise motif-group stability: for every pair of the 17 groups, a pie denotes the fraction of the structures containing the pair that is stable or metastable (blue) versus unstable (red), and pie size is binned by the number of contributing structures (more than 50, 20–50, or fewer than 20); pairs with a stable-or-metastable fraction of at least 75% are shown in full color and the others faded.

sets, 517 in the seed only, 272 in the generated set only, and 889 in neither. Of the 800 generated structures, 182 were built from combinations of motif groups that appeared in no seed structure, suggesting that—at the resolution of this grouping—MatEvolve's novelty was largely combinatorial, favoring new combinations of the seed's existing groups over new groups.

The stability of those combinations is a separate issue. For each of the 136 group pairs, we computed the fraction of structures containing that pair that is stable or metastable across the seed and generated sets (Fig. 6c). The highest fractions, up to 100%, belong to pairs found in fewer than 20 structures and are therefore not robust. Among the pairs supported by more than 50 structures, three stand out, at 76–82%. Each corresponds to the local-motif signature of a recognizable intermetallic family—an analogy we draw after ranking rather than a prediction: a ring coordination motif within a dense percolating metallic framework (82%), as in $CaCu_5$-type RT_5 compounds and their dumbbell-substituted derivatives, the RFe_{12} and R_2Fe_{17} phases, in which the rare-earth atom sits in a hexagonal ring of transition-metal atoms [58]; and a square-net layer paired with either a generic rare-earth high-coordination cage (78%) or tetrahedral units (76%), as in $ThCr_2Si_2$-type RT_2X_2 compounds (R = rare earth, T = transition metal, X = Si, Ge, P, or As), whose layers of edge-sharing TX_4 tetrahedra place the transition-metal atoms on a square net between rare-earth atoms in high-coordination cages.

Taken together, these analyses support two conclusions at this group resolution. First, MatEvolve advances mainly by recombination rather than invention: where its structures depart from the seed, they form new arrangements of the existing motif vocabulary rather than new building blocks, and the most stable of these fall into recognizable intermetallic families. This is not in contrast with the new prototypes reported above: a prototype is a whole-structure

arrangement; thus, a combination of familiar local motifs realized in an arrangement absent from the reference can be combinatorially familiar in its parts while being new as a prototype—the advance is new global arrangements of a shared local vocabulary, not new local building blocks. Second, that vocabulary is human-readable; therefore, the representation that generates a structure also allows us to read it back—a step toward an interpretable design language. Turning these descriptive regularities into predictive design rules and probing novelty below the group resolution may be explored in future work.

Discussion

MatEvolve shows that the motif profile serves not merely as a description of a material but as the medium for material design: the agent edits the profile and constructs a crystal from the modified one, and the most promising candidates are validated by first-principles calculation. On the rare-earth-lean permanent-magnet task it identified novel (meta)stable crystals that are both new prototypes and magnet compounds an order of magnitude more often than generative models, given the same validation budget. It also exceeded every baseline several-fold on new prototypes and matched a magnetization-conditioned generator on magnet compounds without any training on the property. In this benchmark, an agent designing at the level of motifs is thus competitive with generators trained for the task, and exceeds them where the task calls for new structure types.

A second strength is its potential transferability. The agent is a commercial language model (Claude Fable 5) that was used as delivered and never fine-tuned on magnetic-material data; however, it designs competitive magnet compounds. The magnetism-specific ingredients are a few swappable components—the objective, property surrogate, seed set, and first-principles settings—while the motif-editing loop is not tailored to magnetism. Therefore, retargeting should require no retraining, only a restated objective and revised task procedures, carried out largely in natural language so that a scientist from the target field can direct the search. However, this remains to be demonstrated on a second target.

Considered more broadly, MatEvolve is a small instance of a larger shift—AI becoming an engine of scientific discovery, now argued to mark a new, fifth paradigm of science [2, 3, 4]. What a materials example adds is a template built on discipline rather than autonomy: an agent earns trust when its proposals are readable, admitted only once they reconstruct a real structure, and adjudicated by first principles rather than by the model itself. Its deeper promise is an iterative cycle between the two directions of this paradigm. One—*AI for science*—is realized here: the agent accelerates discovery, proposing and verifying candidates at a scale beyond manual reach. The other—*science for AI*—is only motivated by this work: because the agent reasons in readable terms, the insights it exposes may be read back to sharpen the scientist's understanding and inform the design of a subsequent, more capable agent—a loop we do not close here. This mutual reinforcement, rather than autonomy, is seen as the way forward: not a black box that supplants the scientist, but AI and science advancing each other.

Several limitations bound these claims. The demonstration is computational, and its property scope is narrow. DFT evaluates stability and saturation magnetization, but not magnetocrystalline anisotropy, the Curie temperature, or synthesizability—each decisive for a working magnet. Thus, the candidates are precursors to functional magnets rather than instances of them. The magnetization values are moreover 0 K collinear upper bounds (for the heavy rare earths, the antiparallel $4f$–$3d$ coupling can partly compensate the transition-metal moment), and stability is assessed thermodynamically, from pure generalized gradient approximation (GGA) energies and without a phonon screen. Establishing whether any candidate becomes a usable magnet will require these further figures of merit and, ultimately, synthesis.

The search is also loosely guided, and its analysis inherits assumptions from the seed set. The agent itself sees no first-principles feedback within a design step—DFT labels reenter only through the parent pool and ERS; the motif vocabulary is clustered once from the seed structures, so novelty is read categorically, as previously unobserved *combinations* of fixed groups; and the recommender is calibrated on crystals only by analogy with the alloy setting in which it was validated. The natural next steps are feeding the surrogate and DFT signals back into the agent's reasoning, re-deriving the groups as data accrue, grading novelty continuously, and validating the recommender directly on crystals. A comparison of the motif profile with earlier text-based treatments of crystals is provided in Supplementary Information S7.

Methods

1. Evolutionary search over motif profiles

MatEvolve combines the components above into an evolutionary search over motif profiles (Algorithm 1). The population is initialized from the seed set $\mathcal{D}_{\text{seed}}$ and evolved for four generations. In each generation, parents are selected from the current pool; Structure Translate writes each selected parent as a motif profile (Methods 2); Motif editing edits these parent profiles into offspring profiles, which Motif Translate realizes as candidate structures (Methods 3); the ERS ranks the candidates (Methods 5); and DFT is used to validate them in that priority order (Methods 6). Every validated candidate—both stable and unstable—is added to the pool, updating both the parent pool and ERS evidence set for the next generation. Thus, the search is informed by all of its own first-principles labels.

Parent selection. In the magnet run reported herein, each design step draws two parents from the current pool—one from the 200 structures with the lowest energy above the convex hull and one from the 200 structures with the highest saturation magnetization—and combines them by crossover into a single offspring. Thus, each candidate inherits from a stability-favored and magnetization-favored parent. The

Algorithm 1 MatEvolve generation loop

Input: seed population Π_0; objective $\mathcal{P}$
Output: the DFT-validated candidates accumulated over all generations

for $g = 1$ **to** 4 **do**
  select a parent set from Π_{g-1}
  $\mathcal{T} \leftarrow \{\textsc{StructureTranslate}(S) : S \text{ a selected parent}\}$
  $\mathcal{T}_{\text{off}} \leftarrow \{\textsc{MotifEditing}(\text{parents} \subseteq \mathcal{T}, \mathcal{P})\}$
  $\mathcal{O} \leftarrow \{\textsc{MotifTranslate}(T_{\text{off}}) : T_{\text{off}} \in \mathcal{T}_{\text{off}}\}$
  rank $\mathcal{O}$ by the ERS ▷ *Details are shown in Methods 5*
  $\mathcal{V} \leftarrow$ DFT-VALIDATE(top-100 of $\mathcal{O}$ by ERS rank)
  $\Pi_g \leftarrow \Pi_{g-1} \cup \mathcal{V}$
end for

Algorithm 2 StructureTranslate: forward translation with bidirectional-consistency enforcement

Input: crystal S
Output: validated motif profile T, or **reject**

feedback $\leftarrow \varnothing$
for $i = 1$ **to** 5 **do**
  $T \leftarrow \textsc{CIF2Text}(S, \textit{feedback})$
  $S' \leftarrow \textsc{Text2CIF}(T)$
  $T' \leftarrow \textsc{CIF2Text}(S')$
  if $\textsc{Coverage}(T, T')$ **and** $\textsc{StructureMatcher}(S, S')$ **then return** motif profile T
  feedback $\leftarrow$ missing/spurious motifs; composition or symmetry mismatch
end for
return reject

element- and geometry-mutation operators (Methods 3; Supplementary Information S3) are available in the framework but were not applied in this run. In each generation, 200 offspring are proposed and the ERS selects the top 100 for DFT validation, giving 400 validated structures across the four generations; the 400 unselected proposals are additionally relaxed as diagnostics, outside the search loop.

2. Structure Translate — forward translation with bidirectional-consistency enforcement

Structure Translate turns a crystal S into a *validated* motif profile T. A profile that merely sounds plausible can still be incomplete as it may omit details needed to specify the structure; therefore, a profile is retained only if the original crystal can be rebuilt from it. We call this round-trip test *bidirectional consistency*. The translation is performed through the accept-or-revise loop of Algorithm 2 using the two prompt-based basis functions, CIF2Text and Text2CIF, whose implementations are given in Supplementary Information S1 and Supplementary Information S2. CIF2Text drafts a profile T from S; Text2CIF reconstructs a crystal's S' from T; and S' is passed through CIF2Text once more to produce a profile T' for checking. When the checks fail, the agent revises the profile rather than discarding it: deterministic, machine-checkable feedback—the target motifs that were missing or spurious, and any composition or symmetry mismatch—is returned, and a revised profile T is drafted, with a maximum of five attempts.

Acceptance — round-trip fidelity. A draft profile T is accepted—and the loop stops—only when the crystal rebuilt from it agrees with the original in two complementary aspects: it must be built from the same motifs and must assemble them into the same structure. The two checks in Algorithm 2 test these requirements in turn. The first, *motif coverage* $\textsc{Coverage}(T, T')$, asks whether the reconstruction retains the same building blocks; it compares the target motif set $T = \{t_1, \ldots, t_{|T|}\}$ with the set T' obtained by re-translating the reconstruction. By using $[\, t \preceq U \,]$ to denote that a motif t is present in a set U (matched at the archetype level, with a motif split into finer pieces counted as present), the recall r and precision p are as follows:

$$r = \frac{1}{|T|} \sum_{t \in T} \big[\, t \preceq T' \,\big], \qquad p = \frac{1}{|T'|} \sum_{t' \in T'} \big[\, t' \preceq T \,\big], \quad (1)$$

Requiring $r = 1$ ensures that every target motif reappears; thus, no building block is lost or misdescribed in the round trip; the looser $p \geq 0.7$ allows a few additional motifs, since a faithful reconstruction may describe the same structure at slightly finer granularity. An excess of detail is therefore considered benign, whereas a missing motif represents lost information. The second check, $\textsc{StructureMatcher}(S, S')$, then asks whether those building blocks were assembled correctly, accepting the profile only when the rebuilt crystal S' matches the original S geometrically (pymatgen, using the acceptance tolerances `ltol` 0.4, `stol` 0.7, `angle_tol` 15°). The two checks are complementary as a profile can name the correct motifs but arrange them incorrectly, or reproduce the coordinates while relying on different motifs. Therefore, passing both certifies that T encodes S faithfully and sufficiently. Coverage is applied first because it directly tests the defining building blocks, providing a semantic agreement that a global coordinate metric can mask. The recall and precision defined here are reused wherever two motif profiles are compared (Methods 3).

3. Motif editing and Motif Translate — offspring proposal and realization

Motif editing and Motif Translate together create a new candidate crystal: the first edits the motif profiles of existing structures—recombining or mutating their building blocks rather than sampling from a distribution of known structures—and the second realizes the edited profile as a crystal. Rebuilding a crystal from an edited profile is underdetermined because the same profile can be realized by more than one arrangement of atoms. Therefore, a generated crystal is retained only if it actually contains the motifs that were designed.

Motif editing. A design agent, the evolutionary operator EvoOperator (Supplementary Information S3), reads one or two parent profiles from the current pool and proposes an offspring profile T_{off} by crossing over motifs from two parents or by applying element or geometry mutations to a parent's motifs.

Algorithm 3 MOTIFTRANSLATE: realization of an offspring profile

Input: offspring profile T_{off}
Output: candidate structure S_{off}, or **reject**

feedback ← ∅
for $j = 1$ **to** 5 **do**
  S_{off} ← TEXT2CIF(T_{off}, *feedback*)
  T'_{off} ← CIF2TEXT(S_{off})
  if COVERAGE($T_{\text{off}}, T'_{\text{off}}$) **then return** structure S_{off}
  feedback ← uncovered/spurious motifs
end for
return reject

Motif Translate. Motif Translate follows the loop in Algorithm 3: the basis agent TEXT2CIF (Supplementary Information S2) realizes T_{off} as a candidate crystal S_{off}, which is accepted or retried as follows.

Acceptance — realization consistency. A generated crystal is accepted only if it realizes the profile from which it was designed. Here, the target is a *profile*, not an existing crystal. Therefore, no reference structure exists to confirm a geometric match, and the check reduces to the motif-coverage criterion of Methods 2: CIF2TEXT re-derives the motifs of S_{off} as a profile T'_{off}, and S_{off} is accepted when COVERAGE($T_{\text{off}}, T'_{\text{off}}$) passes—every designed motif is present and few spurious motifs are introduced. Otherwise, the missing or spurious motifs are returned as feedback, and the realization is retried, up to a budget of five attempts. This is a consistency check, not a measure of quality: it certifies that a candidate is built from the motifs it was designed to contain, not that it is a good magnet compound. Whether a candidate is stable, novel, and magnetic is decided downstream, where the ERS ranks it (Methods 5) and, if selected, first-principles calculations validate it (Methods 6), within the run described in Methods 1.

4. Motif clustering

We grouped the motifs into a small number of recurring geometric archetypes. Because a motif is text rather than a vector, ordinary distance-based clustering does not apply. Instead, we used an agentic version of K-means (Algorithm 4) [67], in which each group is represented by a *prototype*—a short paragraph summarizing the motifs it contains, analogous to a centroid. Language-model workers voted on the group to which each motif was best matched (assignment), and the update step resummarized each prototype from its members and adjusted the number of groups. Only the categorical votes and prototype summaries came from the model; all counting and every merge, split, and stopping decision were deterministic. Thus, no numerical judgment rested with the model, and K was discovered rather than fixed. *Assignment (E-step).* Each motif, shown alongside the current group prototypes, was voted on by $R = 5$ independent workers (Claude Fable 5 at temperature 0.5; five workers trade off robustness against cost), with each worker choosing the best-matching group or *none*. If n_{ig} workers chose group g for motif i, its consensus was the majority share,

$$s_i = \frac{1}{R} \max_g n_{ig}. \tag{2}$$

A high s_i value indicated that the workers agreed on motif i; repeated *none* votes flagged a geometry not yet covered by any group.

Update (M-step). Each group's prototype was first resummarized from the motifs assigned to it—analogous to moving a K-means centroid. Thus, the next round of voting used an up-to-date description. The routine then *merged* two groups that were frequently voted together, *split* a group whose members disagreed, and *birthed* a group from motifs repeatedly voted *none*. Two ratios drove these operations—the confusion c_{gh} shared by groups g and h, and the contested fraction ρ_g of low-consensus members (consensus below 0.6) in a group:

$$c_{gh} = \frac{|g \sqcap h|}{\min(|g|, |h|)}, \qquad \rho_g = \frac{|\{i \in g \,:\, s_i < 0.6\}|}{|g|}, \tag{3}$$

where $g \sqcap h$ is the set of motifs voted into both groups and $|g|$ is the size of group g. Groups were merged when $c_{gh} \geq 0.5$ (over at least three motifs) and split when $\rho_g > 0.3$, unless the contested votes were concentrated on one rival group, in which case the two were merged instead. Motifs that workers repeatedly voted *none* matched no current group. Once at least 15 such motifs had accumulated, the model analyzed them and summarized them into one or more new groups, which entered the next round of voting. Thus, K changed only when the votes warranted it. Groups were defined at the archetype level—unit class, coordination geometry, and coordination-number band—so that each represented a recognizable building block.

Convergence. The steps alternated until the groups stopped changing, as judged by the mean consensus $\bar{s} = |\mathcal{M}|^{-1} \sum_i s_i$ and the agreement between assignments in successive rounds, measured by the Adjusted Rand Index (ARI) [68],

$$\text{ARI} = \frac{\sum_{gh} \binom{n_{gh}}{2} - \left[\sum_g \binom{a_g}{2} \sum_h \binom{b_h}{2}\right] \Big/ \binom{n}{2}}{\frac{1}{2}\left[\sum_g \binom{a_g}{2} + \sum_h \binom{b_h}{2}\right] - \left[\sum_g \binom{a_g}{2} \sum_h \binom{b_h}{2}\right] \Big/ \binom{n}{2}}, \tag{4}$$

with n_{gh} denoting the motifs shared by group g of one round and h of the next, a_g and b_h denote their sizes, and n is the total number of motifs. Clustering stopped when $\bar{s} \geq 0.8$ and ARI ≥ 0.95 for two rounds, or after eight rounds.

5. Evidence-based Recommender System (ERS)

Within each generation, the ERS *ordered* the candidates produced by the agents so that costly first-principles validation was performed first on the most promising candidates. It did not determine stability—that was the role of DFT (Methods 6)—but supplied an inexpensive, uncertainty-aware prior over which candidates were worth relaxing. The ERS scored each candidate based on its motif-group composition using Dempster–Shafer evidence theory, following the descriptor-free recommender introduced for alloy

Algorithm 4 MOTIFCLUSTER: agentic consensus clustering

Input: a set of motifs $\mathcal{M}$
Output: groups $\mathcal{G}$ (K discovered) and an assignment of every motif in $\mathcal{M}$

$\mathcal{G} \leftarrow$ initial prototypes proposed from $\mathcal{M}$ (approximately a dozen)
repeat
  E-step. 5 workers vote every motif of $\mathcal{M}$ → a group or *none*; compute consensus s_i
  M-step. resummarize each group's prototype from its members;
    merge groups with confusion $c_{gh} \geq 0.5$;
    split a group with contested fraction $\rho_g > 0.3$;
    birth new groups from ≥ 15 motifs voted *none*
until (mean consensus $\bar{s} \geq 0.8$ **and** ARI ≥ 0.95, two rounds) **or** 8 rounds
return $\mathcal{G}$ and the converged assignment of $\mathcal{M}$

discovery [44, 34]. The 17 motif groups defined in Methods 4 serve as chemical elements in that framework. The groups were fixed after clustering, and each candidate's motifs were assigned to them before scoring, ensuring that the vocabulary over which the ERS ranked candidates remained consistent throughout the run. Because each motif group represented a class of geometrically similar building blocks, substituting one group for another was a physically reasonable operation—exchanging one structural building block for another—and the ERS learned from the observed structures which such substitutions preserved the target property.

Each observed structure was represented by the *set* $S = \{g_1, \ldots, g_k\}$ of the motif groups it contained, with each $g_i \in \mathcal{G}$ belonging to the 17-group vocabulary of Methods 4. Multiplicity, connectivity, composition, and lattice information below this resolution were deliberately not encoded, so the belief returned by the ERS was a prior over motif-set plausibility rather than a thermodynamic prediction. Because composition was among the omitted variables, the group co-occurrences it learned partly reflected shared chemistry rather than structural interchangeability alone. Each structure carried a binary target label $y_S \in \{\top, \neg\top\}$ with $y_S = \top$ when S met the design objective. For the stability target used to prioritize the run, $y_S = \top$ when $E_{\text{hull}}(S) \leq \varepsilon$ at the same $\varepsilon = 0.1$ eV/atom used throughout—a metastability window rather than strict on-hull stability (Supplementary Information S4.1). The evidence set $\mathcal{D}$ consisted of the labeled seed structures and was augmented each generation by the DFT-validated candidates, thereby allowing the recommender to refresh as the search progressed. Evidence was accumulated in two stages—the substitutability of motif-group combinations, followed by the target belief of a candidate—with each represented by a Dempster–Shafer mass function fused by Dempster's rule.

Substitutability of motif-group combinations. Any two observed structures sharing at least one motif group, $S_i \cap S_j \neq \emptyset$, provide one piece of evidence as to whether the group combinations by which they differ are interchangeable. By using $C_t = S_i \setminus S_j$ and $C_v = S_j \setminus S_i$ to refer to the differing combinations against the shared context $S_i \cap S_j$, and taking a frame of discernment $\Omega_{\text{sub}} = \{\text{sub}, \neg\text{sub}\}$, the pair contributes

$$m^{C_t,C_v}_{S_i,S_j}(\{\text{sub}\}) = \alpha\,[\,y_{S_i} = y_{S_j}\,], \tag{5}$$

$$m^{C_t,C_v}_{S_i,S_j}(\{\neg\text{sub}\}) = \alpha\,[\,y_{S_i} \neq y_{S_j}\,], \tag{6}$$

$$m^{C_t,C_v}_{S_i,S_j}(\{\text{sub}, \neg\text{sub}\}) = 1 - \alpha, \tag{7}$$

where $[\,\cdot\,]$ is the Iverson bracket and the discount $\alpha \in (0, 1)$ leaves a residual mass $1 - \alpha$ on the full frame, encoding the uncertainty of a single observation. Combining all pieces of evidence bearing on a given pair of combinations over the data set $\mathcal{D}$ by Dempster's rule (below) gives a substitutability belief $M(C_t, C_v) = m^{C_t,C_v}_{\mathcal{D}}(\{\text{sub}\})$, symmetric in its two arguments.

Target belief of a candidate. A candidate S_{new} is scored over every observed host $S_{\text{host}} \in \mathcal{D}$ and every single substitution $C_t \leftarrow C_v$, with $C_t \subseteq S_{\text{host}}$, that yields $S_{\text{new}} = (S_{\text{host}} \setminus C_t) \cup C_v$ and whose differing pair carries nonzero substitutability belief. Over the target frame $\Omega = \{\top, \neg\top\}$, each such host contributes

$$m^{S_{\text{new}}}_{S_{\text{host}},C_t\leftarrow C_v}(\{\top\}) = M(C_t, C_v)\,[y_{S_{\text{host}}}=\top], \tag{8}$$

$$m^{S_{\text{new}}}_{S_{\text{host}},C_t\leftarrow C_v}(\{\neg\top\}) = M(C_t, C_v)\,[y_{S_{\text{host}}}=\neg\top], \tag{9}$$

$$m^{S_{\text{new}}}_{S_{\text{host}},C_t\leftarrow C_v}(\{\top, \neg\top\}) = 1 - M(C_t, C_v), \tag{10}$$

a Shafer discount in which the host transfers its own label to the candidate with a strength equal to the substitutability belief, with the residual $1 - M(C_t, C_v)$ left as ignorance.

Combination and ranking. Pieces of evidence are fused by Dempster's rule of combination,

$$(m_a \oplus m_b)(\omega) = \frac{1}{1-K} \sum_{\omega_k \cap \omega_h = \omega} m_a(\omega_k)\, m_b(\omega_h),$$
$$K = \sum_{\omega_k \cap \omega_h = \emptyset} m_a(\omega_k)\, m_b(\omega_h), \tag{11}$$

which is commutative and associative. Thus, the fusion is independent of the order in which hosts are combined. Fusing all hosts determines a candidate's final mass $m^{S_{\text{new}}}_{\mathcal{D}}$, and candidates are ranked in descending order of the likelihood that they meet the target, $m^{S_{\text{new}}}_{\mathcal{D}}(\{\top\})$. The discount α is fixed by a grid search that maximizes the cross-validated reproduction of the observed labels ($\alpha = 0.1$), with the ranking being largely insensitive to it, as in the alloy setting [44].

6. Property evaluation and first-principles validation

The validated quantities—the relaxed structure, total energy, energy above the convex hull, and target property—are recorded, and the structure is returned to the database.

DFT Relaxation. Final structural optimization is performed using the Vienna Ab initio Simulation Package (VASP, version 5.4.4) [69, 70, 71] with spin-polarized calculations (ISPIN = 2) to capture the magnetic nature of the rare-earth–transition-metal compounds studied. The

GGA with the Perdew–Burke–Ernzerhof functional [72] is employed to describe exchange-correlation effects. Initial magnetic moments are assigned per element: 5.0 μ_B for Fe, Mn, Ni, Cr, V, Mo, W and Ce, 10.0 μ_B for Eu, and 0.6 μ_B for every other element for which a moment was defined (Supplementary Information S8). The projector augmented-wave (PAW) method [73, 74] is employed with Sm_3 and Fe_pv potentials to properly describe the valence-electron configurations, using a plane-wave energy cutoff of 520 eV. Nonspherical contributions to the PAW spheres (LASPH = .TRUE.) and extended angular-momentum mixing (LMAXMIX = 6) are enabled for accurate charge-density mixing in these open-core rare-earth systems (the Sm $4f$ electrons being treated in the core by the Sm_3 potential, with their moment restored as described below). Brillouin zone sampling employs Γ-centered k-point meshes with KSPACING = 0.25 Å^{-1}. These settings follow established high-throughput practice for intermetallic convex-hull construction and match the Materials Project workflow used to compute the reference energies. Each candidate's energy above the convex hull is measured against a consistently computed hull, and the 0.1 eV/atom stability cutoff has the same meaning for generated candidates as for the reference database. The potentials are sourced from the POTCAR library version 5.4 of VASP [75, 74, 73]. The potential and initial magnetic moment of every element, and the versions of the Materials Project and Alexandria databases used as references, are listed in Supplementary Information S8.

Electronic self-consistency is achieved with an energy-convergence criterion of EDIFF = 10^{-4} eV, while ionic relaxation stops when the total energy changes by less than 10^{-3} eV between ionic steps (the VASP default, EDIFFG = 10×EDIFF). Full structural relaxation including ionic positions, cell shape, and cell volume (ISIF = 3) is performed using the conjugate gradient algorithm (IBRION = 2) for up to 50 ionic steps (NSW = 50) with a maximum of 100 electronic self-consistency iterations per ionic step (NELM = 100). Electronic structure calculations use the tetrahedron method with Blöchl corrections (ISMEAR = −5). The FAST algorithm is employed for electronic minimization, and orbital-decomposed density of states are computed (LORBIT = 11) for the postanalysis of magnetic properties.

Formation-energy Calculation. Thermodynamic stability is assessed through formation-energy calculations. The formation energy per atom, ΔH_f, is computed as follows:

$$\Delta H_f = \frac{1}{N}\left(E_{\text{total}} - \sum_i n_i \mu_i\right), \tag{12}$$

where E_{total} is the DFT total energy of the relaxed structure, N is the total number of atoms, n_i is the number of atoms of element i, and μ_i is the chemical potential of element i referenced to its standard elemental ground state. The formation energy serves as a key indicator of the stability of candidate compounds [76].

Energy Above Hull. Thermodynamic stability relative to competing phases is quantified by the energy above the convex hull. For each generated structure, a convex hull is constructed for its chemical system from Materials Project entries spanning the constituent elements. The energy above the hull is then evaluated as the difference between the structure's DFT formation energy per atom (relative to elemental references calculated with the same settings) and hull energy at its composition, the latter is taken as the proportion-weighted formation energies of the structure's decomposition products on the hull. A value of zero denotes an on-hull (thermodynamically stable) structure, and positive values denote metastability.

Magnetization Calculation. The total magnetic moment $\mu[x]$ of a structure is the DFT moment corrected for the lanthanides whose $4f$ electrons are treated in the core (Supplementary Information S8) [59]:

$$\mu[x] = M[x] + \sum_{k\in \mathrm{Ln_c}} \sigma_k \, g_{J_k} J_k, \tag{13}$$

where $M[x]$ is the total moment of the cell from DFT, which already contains the $4f$ moments of Ce, Eu, and Gd, whose $4f$ electrons are in the valence; the sum runs over the lanthanides $\mathrm{Ln_c}$ whose $4f$ electrons are in the core and restores each one's $4f$ moment analytically: $g_{J_k} J_k$ is its Hund's-rule ground-state moment (g_{J_k} is the Landé factor and J_k is the total angular momentum), and $\sigma_k = \pm 1$ denotes its coupling to the transition-metal sublattice. The $4f$–$3d$ exchange holds the rare-earth spin antiparallel to the transition-metal moment. Therefore, by Hund's third rule, the net $4f$ moment *adds* for the light lanthanides ($\sigma_k = +1$; Pr, Nd, Sm, whose $4f$ shell is less than half-filled) and *opposes* it for the heavier ones ($\sigma_k = -1$; Tb, Dy, Ho, Er): $g_J J = 0.714\ \mu_B$ for Sm^{3+} is added, whereas the larger heavy-lanthanide moments ($\sim$9–10 μ_B) are subtracted; Yb, treated as divalent, has a filled $4f$ shell and contributes none. We neglect the J-mixing enhancement of the Sm^{3+} moment, which is small on the scale of the 1 T threshold [59]. The saturation magnetization is the magnitude of this net moment, $\mu_0 M_s[x] = \mu_0 \, |\mu[x]|/V$, where V is the unit cell volume and μ_0 is the vacuum permeability.

Author Contributions

D.-K.L.: Conceptualization, Methodology, Software, Validation, Investigation, Writing – review & editing. **M.-Q.H.**: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **H.-P.V.-D.**: Investigation, Validation. **V.-C.N.**: Software, Validation. **H.T.**: Conceptualization, Writing – review & editing. **T.T.**: Conceptualization, Writing – review & editing. **T.M.**: Conceptualization, Investigation, Validation, Writing – review & editing. **H.K.**: Conceptualization, Investigation, Validation, Writing – review & editing. **H.-C.D.**: Conceptualization, Supervision, Resources, Methodology, Writing – review & editing.

Conflicts of Interest

The authors declare no competing interests.

Data Availability

All data generated and analysed in this study, including the DFT-relaxed structures, are deposited in Zenodo [77].

Code Availability

The code implementing MatEvolve, including the complete prompt templates and the scripts that compute the benchmark rates, is available on GitHub [78] under an MIT license.

Acknowledgements

This work is supported by the JST-CREST Program: Innovative Measurement and Analysis, under grant number JPMJCR2235; the MEXT Program: Data Creation and Utilization-Type Material Research and Development Project (Digital Transformation Initiative Center for Magnetic Materials), under grant number JPMXP1122715503; the JST "Adopting Sustainable Partnerships for Innovative Research Ecosystem (ASPIRE)" Project, under grant number JPMJAP2321; and JSPS KAKENHI grant numbers 20K05301, JP19H05815, 23KJ1035, 23K03950, and JP23H05403.

References

[1] T. Hey, S. Tansley, K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research, Redmond, Washington, 2009.

[2] P. Berens, K. Cranmer, N. D. Lawrence, U. von Luxburg, J. Montgomery, AI for science: An emerging agenda, 2023. doi:10.48550/arXiv.2303.04217. arXiv:2303.04217.

[3] C. Bishop, AI4Science to empower the fifth paradigm of scientific discovery, Microsoft Research Blog, 2022.

[4] Y. Ioannidis, The fifth paradigm: AI-driven scientific discovery, Communications of the ACM 67 (2024). doi:10.1145/3702970.

[5] 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, The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials 1 (2013) 011002. URL: http://link.aip.org/link/AMPADS/v1/i1/p011002/s1&Agg=doi. doi:10.1063/1.4812323.

[6] L. Pauling, The principles determining the structure of complex ionic crystals, Journal of the American Chemical Society 51 (1929) 1010–1026. doi:10.1021/ja01379a006.

[7] A. R. Oganov, C. W. Glass, Crystal structure prediction using ab initio evolutionary techniques: Principles and applications, The Journal of Chemical Physics 124 (2006) 244704. doi:10.1063/1.2210932.

[8] Y. Wang, J. Lv, L. Zhu, Y. Ma, Calypso: A method for crystal structure prediction, Computer Physics Communications 183 (2012) 2063–2070. doi:10.1016/j.cpc.2012.05.008.

[9] H.-C. Dam, H. Kino, T. Miyake, From data to discovery: Challenges and solutions in AI-driven materials science, Journal of the Physical Society of Japan 95 (2026) 042001. doi:10.7566/JPSJ.95.042001.

[10] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, npj Computational Materials 7 (2021) 185. URL: https://doi.org/10.1038/s41524-021-00650-1. doi:10.1038/s41524-021-00650-1.

[11] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chemistry of Materials 31 (2019) 3564–3572. URL: https://doi.org/10.1021/acs.chemmater.9b01294. doi:10.1021/acs.chemmater.9b01294.

[12] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, G. Ceder, Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nature Machine Intelligence 5 (2023) 1031–1041. URL: https://doi.org/10.1038/s42256-023-00716-3. doi:10.1038/s42256-023-00716-3.

[13] C. Chen, S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nature Computational Science 2 (2022) 718–728. URL: https://doi.org/10.1038/s43588-022-00349-3. doi:10.1038/s43588-022-00349-3.

[14] T.-S. Vu, M.-Q. Ha, D.-N. Nguyen, V.-C. Nguyen, Y. Abe, T. Tran, H. Tran, H. Kino, T. Miyake, K. Tsuda, H.-C. Dam, Towards understanding structure–property relations in materials with interpretable deep learning, npj Computational Materials 9 (2023) 215. URL: https://doi.org/10.1038/s41524-023-01163-9. doi:10.1038/s41524-023-01163-9.

[15] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, T. Jaakkola, Crystal diffusion variational autoencoder for periodic material generation, in: International Conference on Learning Representations (ICLR), 2022. ArXiv:2110.06197.

[16] R. Jiao, W. Huang, P. Lin, J. Han, P. Chen, Y. Lu, Y. Liu, Crystal structure prediction by joint equivariant diffusion, in: Advances in Neural Information Processing Systems (NeurIPS), volume 36, 2023.

[17] C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, C. Yang, W. Li, R. Tomioka, T. Xie, A generative model for inorganic materials design, Nature 639 (2025) 624–632. doi:10.1038/s41586-025-08628-5.

[18] B. K. Miller, R. T. Q. Chen, A. Sriram, B. M. Wood, FlowMM: Generating materials with riemannian flow matching, in: Proceedings of the 41st International Conference on Machine Learning, volume 235 of *PMLR*, 2024, pp. 35664–35686. doi:10.48550/arXiv.2406.04713.

[19] N. Gruver, A. Sriram, A. Madotto, A. G. Wilson, C. L. Zitnick, Z. Ulissi, Fine-tuned language models generate stable inorganic materials as text, in: International Conference on Learning Representations (ICLR), 2024. ArXiv:2402.04379.

[20] L. M. Antunes, K. T. Butler, R. Grau-Crespo, Crystal structure generation with autoregressive large language modeling, Nature Communications 15 (2024) 10570. doi:10.1038/s41467-024-54639-7.

[21] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, E. D. Cubuk, Scaling deep learning for materials discovery, Nature 624 (2023) 80–85. doi:10.1038/s41586-023-06735-9.

[22] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design, npj Computational Materials 5 (2019) 21. URL: https://doi.org/10.1038/s41524-019-0153-8. doi:10.1038/s41524-019-0153-8.

[23] N. J. Szymanski, B. Rendy, Y. Fei, R. E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, G. Ceder, An autonomous laboratory for the accelerated synthesis of inorganic materials, Nature 624 (2023) 86–91. doi:10.1038/s41586-023-06734-w.

[24] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, L. Ward, Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery, Molecular Systems Design & Engineering 3 (2018) 819–825. doi:10.1039/C8ME00012C.

[25] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates, Physical Review Materials 2 (2018) 083802. doi:10.1103/PhysRevMaterials.2.083802.

[26] A. K. Cheetham, R. Seshadri, Artificial intelligence driving materials discovery? perspective on the article: Scaling deep learning for materials discovery, Chemistry of Materials 36 (2024) 3490–3495.

doi:10.1021/acs.chemmater.4c00643.

[27] H. Wang, T. Fu, Y. Du, W. Gao, K. Huang, Z. Liu, P. Chandak, S. Liu, P. Van Katwyk, A. Deac, A. Anandkumar, K. Bergen, C. P. Gomes, S. Ho, P. Kohli, J. Lasenby, J. Leskovec, T.-Y. Liu, A. Manrai, D. Marks, B. Ramsundar, L. Song, J. Sun, J. Tang, P. Veličković, M. Welling, L. Zhang, C. W. Coley, Y. Bengio, M. Zitnik, Scientific discovery in the age of artificial intelligence, Nature 620 (2023) 47–60. doi:10.1038/s41586-023-06221-2.

[28] Anthropic, Claude Fable 5 and claude Mythos 5, Anthropic, 2026.

[29] V. Natarajan, et al., Accelerating scientific discovery with Co-Scientist, Nature (2026). doi:10.1038/s41586-026-10644-y.

[30] Y. Kang, J. Kim, ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models, Nature Communications 15 (2024) 4998. doi:10.1038/s41467-024-48998-4.

[31] M. Raissi, P. Perdikaris, G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686–707. doi:10.1016/j.jcp.2018.10.045.

[32] T. H. Trinh, Y. Wu, Q. V. Le, H. He, T. Luong, Solving olympiad geometry without human demonstrations, Nature 625 (2024) 476–482. doi:10.1038/s41586-023-06747-5.

[33] D.-K. Le, M.-Q. Ha, T.-D. Nguyen, V.-H. Nguyen, H.-C. Dam, Crystallography as language: transforming structural data into textual description for LLM-aided materials discovery, Molecular Physics (2026). doi:10.1080/00268976.2025.2610374.

[34] M.-Q. Ha, D.-K. Le, V.-C. Nguyen, H. Kino, S. Curtarolo, H.-C. Dam, Beyond interpolation: integration of data and ai-extracted knowledge for high-entropy alloy discovery, Digital Discovery 5 (2026) 1079–1097. URL: http://dx.doi.org/10.1039/D5DD00400D. doi:10.1039/D5DD00400D.

[35] A. M. Ganose, A. Jain, Robocrystallographer: automated crystal structure text descriptions and analysis, MRS Communications 9 (2019) 874–881. doi:10.1557/mrc.2019.94.

[36] T. Mohanty, M. Mehta, H. M. Sayeed, B.-E. Oded, I. Pitussi, A. Borenstein, V. Srikumar, T. D. Sparks, Crystext: A generative ai approach for text-conditioned crystal structure generation using llm, Integrating Materials and Manufacturing Innovation 15 (2026) 416–430. doi:10.1007/s40192-026-00451-8.

[37] A. Niyongabo Rubungo, C. Arnold, B. P. Rand, A. B. Dieng, Llm-prop: predicting the properties of crystalline materials using large language models, npj Computational Materials 11 (2025) 186. doi:10.1038/s41524-025-01536-2.

[38] M. Munjal, J. Lee, C. Park, S. Han, Lattice lingo: Effect of textual detail on multimodal learning for property prediction of crystals, arXiv:2412.04670, 2024. doi:10.48550/arXiv.2412.04670.

[39] S. Ozawa, I. Takahara, T. Mizoguchi, Scale-dependent input representation and confidence estimation for LLMs in materials property prediction, arXiv:2605.03515, 2026. doi:10.48550/arXiv.2605.03515.

[40] B. Romera-Paredes, M. Barekatain, A. Novikov, M. Balog, M. P. Kumar, E. Dupont, F. J. R. Ruiz, J. S. Ellenberg, P. Wang, O. Fawzi, P. Kohli, A. Fawzi, Mathematical discoveries from program search with large language models, Nature 625 (2024) 468–475. doi:10.1038/s41586-023-06924-6.

[41] S. Jia, C. Zhang, V. Fung, Llmatdesign: Autonomous materials discovery with large language models, 2024. arXiv:2406.13163, preprint at arXiv:2406.13163.

[42] I. Takahara, T. Mizoguchi, B. Liu, Accelerated inorganic materials design with generative AI agents, Cell Reports Physical Science 6 (2025) 103019. doi:10.1016/j.xcrp.2025.103019.

[43] J. Gan, P. Zhong, Y. Du, Y. Zhu, C. Duan, H. Wang, D. Schwalbe-Koda, C. P. Gomes, K. A. Persson, W. Wang, Matllmsearch: Crystal structure discovery with evolution-guided large language models, arXiv:2502.20933, 2025. doi:10.48550/arXiv.2502.20933.

[44] M.-Q. Ha, D.-N. Nguyen, V.-C. Nguyen, T. Nagata, T. Chikyow, H. Kino, T. Miyake, T. Denœux, V.-N. Huynh, H.-C. Dam, Evidence-based recommender system for high-entropy alloys, Nature Computational Science 1 (2021) 470–478. URL: https://doi.org/10.1038/s43588-021-00097-w. doi:10.1038/s43588-021-00097-w.

[45] A. P. Dempster, Upper and lower probabilities induced by a multivalued mapping, The Annals of Mathematical Statistics 38 (1967) 325–339. doi:10.1214/aoms/1177698950.

[46] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, New Jersey, 1976.

[47] J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti, M. A. L. Marques, Machine-learning-assisted determination of the global zero-temperature phase diagram of materials, Advanced Materials 35 (2023) 2210788. doi:10.1002/adma.202210788.

[48] M. Pinsky, D. Avnir, Continuous symmetry measures. 5. the classical polyhedra, Inorganic Chemistry 37 (1998) 5575–5582. doi:10.1021/ic9804925.

[49] H. Zabrodsky, S. Peleg, D. Avnir, Continuous symmetry measures, Journal of the American Chemical Society 114 (1992) 7843–7851. doi:10.1021/ja00046a033.

[50] T. L. Pham, H. Kino, K. Terakura, T. Miyake, K. Tsuda, I. Takigawa, H. C. Dam, Machine learning reveals orbital interaction in materials, Science and Technology of Advanced Materials 18 (2017) 756–765. doi:10.1080/14686996.2017.1378060.

[51] T.-L. Pham, D.-N. Nguyen, M.-Q. Ha, H. Kino, T. Miyake, H.-C. Dam, Explainable machine learning for materials discovery: predicting the potentially formable Nd–Fe–B crystal structures and extracting the structure–stability relationship, IUCrJ 7 (2020) 1036–1047. URL: https://doi.org/10.1107/S2052252520010088. doi:10.1107/S2052252520010088.

[52] H. Park, A. Onwuli, A. Walsh, Exploration of crystal chemical space using text-guided generative artificial intelligence, Nature Communications 16 (2025) 4379. doi:10.1038/s41467-025-59636-y.

[53] M. J. Mehl, D. Hicks, C. Toher, O. Levy, G. L. W. Hart, S. Curtarolo, et al., The AFLOW library of crystallographic prototypes: Part 1, Computational Materials Science 136 (2017) S1–S828. doi:10.1016/j.commatsci.2017.01.017.

[54] D. Hicks, M. J. Mehl, M. Esters, C. Oses, O. Levy, G. L. W. Hart, C. Toher, S. Curtarolo, The AFLOW library of crystallographic prototypes: Part 3, Computational Materials Science 199 (2021) 110450. doi:10.1016/j.commatsci.2021.110450.

[55] S. Alvarez, P. Alemany, D. Casanova, J. Cirera, M. Llunell, D. Avnir, Shape maps and polyhedral interconversion paths in transition metal chemistry, Coordination Chemistry Reviews 249 (2005) 1693–1708. doi:10.1016/j.ccr.2005.03.031.

[56] L. Shao, Y. Li, J. Yu, R. Xu, Divalent-metal-stabilized aluminophosphates exhibiting a new zeolite framework topology, Inorganic Chemistry 51 (2012) 225–229. doi:10.1021/ic201515z.

[57] A. V. Funes, L. Carrella, Y. Rechkemmer, J. van Slageren, E. Rentschler, P. Alborés, Synthesis, structural characterization and magnetic behaviour of a family of $[Co_2^{III}Ln_2^{III}]$ butterfly compounds, Dalton Transactions 46 (2017) 3400–3409. doi:10.1039/c6dt04713k.

[58] J. M. D. Coey, H. Sun, Improved magnetic properties by treatment of iron-based rare earth intermetallic compounds in ammonia, Journal of Magnetism and Magnetic Materials 87 (1990) L251–L254. doi:10.1016/0304-8853(90)90756-G.

[59] T. Miyake, H. Akai, Quantum theory of rare-earth magnets, Journal of the Physical Society of Japan 87 (2018) 041009. URL: https://doi.org/10.7566/JPSJ.87.041009. doi:10.7566/JPSJ.87.041009. arXiv:https://doi.org/10.7566/JPSJ.87.041009.

[60] J. F. Herbst, R2fe14b materials: intrinsic properties and technological aspects, Reviews of Modern Physics 63 (1991) 819–898. doi:10.1103/RevModPhys.63.819.

[61] D. B. de Mooij, K. H. J. Buschow, Some novel ternary Th_2Zn_{17}-derived carbides $R_2Fe_{17}C_x$, Journal of the Less-Common Metals 142 (1988) 349–357.

[62] J. L. C. Daams, J. H. N. van Vucht, P. Villars, Atomic-environment classification of the cubic "intermetallic" structure types, Journal of Alloys and Compounds 182 (1992) 1–33. doi:10.1016/0925-8388(92)90570-Y.

[63] P. Villars, J. L. C. Daams, J. H. N. van Vucht, Atlas of Crystal Structure Types for Intermetallic Phases, ASM International, Materials Park, OH, 1991.

[64] F. C. Frank, J. S. Kasper, Complex alloy structures regarded as sphere packings. I. definitions and basic principles, Acta Crystallographica 11 (1958) 184–190. doi:10.1107/S0365110X58000487.

[65] F. C. Frank, J. S. Kasper, Complex alloy structures regarded as sphere packings. II. analysis and classification of representative structures, Acta Crystallographica 12 (1959) 483–499. doi:10.1107/S0365110X59001499.

[66] J.-X. Yin, B. Lian, M. Z. Hasan, Topological kagome magnets and superconductors, Nature 612 (2022) 647–657. doi:10.1038/s41586-022-05516-0.

[67] S. Monti, P. Tamayo, J. Mesirov, T. Golub, Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data, Machine Learning 52 (2003) 91–118.

[68] L. Hubert, P. Arabie, Comparing partitions, Journal of Classification 2 (1985) 193–218.

[69] G. Kresse, J. Hafner, Ab initio molecular dynamics for liquid metals, Phys. Rev. B 47 (1993) 558–561.

[70] G. Kresse, J. Hafner, Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium, Phys. Rev. B 49 (1994) 14251–14269.

[71] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mat. Sci. 6 (1996) 15–50.

[72] J. P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Physical Review Letters 77 (1996) 3865–3868.

[73] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953–17979. URL: https://link.aps.org/doi/10.1103/PhysRevB.50.17953. doi:10.1103/PhysRevB.50.17953.

[74] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758–1775. URL: https://link.aps.org/doi/10.1103/PhysRevB.59.1758. doi:10.1103/PhysRevB.59.1758.

[75] VASP Software GmbH, Available PAW potentials, VASP Wiki, 2024. URL: https://www.vasp.at/wiki/index.php/Available_PAW_potentials, accessed: 2026-06-14.

[76] M. Matsumoto, T. Hawai, K. Ono, (Sm,Zr)$Fe_{12-x}M_x$ (M = Zr, Ti, Co) for permanent-magnet applications: Ab initio material design integrated with experimental characterization, Phys. Rev. Appl. 13 (2020) 064028. URL: https://link.aps.org/doi/10.1103/PhysRevApplied.13.064028. doi:10.1103/PhysRevApplied.13.064028.

[77] D.-K. Le, M.-Q. Ha, H.-P. Vu-Dinh, V.-C. Nguyen, H. Tran, T. Tran, T. Miyake, H. Kino, H.-C. Dam, Data for "crystal-structure design by agentic AI in a language of motifs", Zenodo, 2026. URL: https://doi.org/10.5281/zenodo.23480862. doi:10.5281/zenodo.23480862.

[78] D.-K. Le, M.-Q. Ha, H.-P. Vu-Dinh, V.-C. Nguyen, H. Tran, T. Tran, T. Miyake, H. Kino, H.-C. Dam, MatEvolve, GitHub, 2026. URL: https://github.com/minhquyet2308/MatEvolve.

Supplementary Files

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

- supplementary.pdf