nature computational science

Article

https://doi.org/10.1038/s43588-026-01037-2

Enhancing materials discovery with valence-constrained design in generative modeling

Received: 27 July 2025

Accepted: 16 July 2026

Published online: 26 August 2026

Check for updates

Mouyang Cheng[1,2,3,10] ✉, Weiliang Luo[4,10], Hao Tang[3,10], Bowen Yu[5,10], Yongqiang Cheng[6], Weiwei Xie[7], Ju Li[2,3,8], Heather J. Kulik[2,4,9] & Mingda Li[1,2,8] ✉

Diffusion-based deep generative models have emerged as powerful tools for inverse materials design. Yet many existing approaches overlook essential chemical constraints, such as oxidation-state balance, which can lead to chemically invalid structures. Here we introduce 'crystal generator with valence-constrained design' (CrysVCD), a modular framework that integrates chemical rules directly into the generative process. CrysVCD first uses a transformer-based elemental language model to generate valence-balanced compositions, followed by a diffusion model to generate crystal structures. The valence constraint enables orders-of-magnitude more efficient chemical valence checking compared with pure data-driven approaches with post-screening. When fine-tuned on stability metrics, CrysVCD achieves 85% metastability (E_{hull} < 0.1 eV per atom) and 68% phonon stability. Moreover, CrysVCD supports conditional generation of functional materials, enabling discovery of candidates such as high thermal conductivity semiconductors and high dielectric constant (high-κ) materials. Designed as a general-purpose plugin, CrysVCD can be integrated into diverse generative pipelines to promote chemical validity, offering a reliable, scientifically grounded path for materials discovery.

Materials design has long been a cornerstone of modern technology, driving advancements across diverse fields, from structural and functional materials to energy and quantum technologies. The integration of computational methods has substantially enhanced this process, enabling the automated discovery of promising crystal candidates[1–4], with recent advances in artificial intelligence accelerating these efforts further. In particular, inverse design has emerged as a transformative paradigm, facilitating the targeted generation of functional materials with specific properties while reducing the computational burden of exhaustive forward screening[5]. Among inverse design techniques, deep generative models have achieved remarkable success by capturing intricate structure–property relationships to enable conditional materials generation[6,7]. Notable examples include CDVAE[8], DiffCSP[9] and MatterGen[10], which demonstrate the capability of diffusion-based models to generate chemically diverse materials with novel crystal structures. Another approach incorporates large language models in the materials design process, utilizing inverse design prompted by human natural language[11–13]. Together, these advances highlight the growing role of deep generative architectures in automated materials discovery.

Despite substantial progress, current generative approaches still encounter challenges in practical solid-state material discovery scenarios. One key issue is to ensure that generated materials adhere to fundamental chemical principles, such as common oxidation states and stoichiometric balance. While these chemical rules can be learned

[1]Quantum Measurement Group, MIT, Cambridge, MA, USA. [2]Center for Computational Science and Engineering, MIT, Cambridge, MA, USA. [3]Department of Materials Science and Engineering, MIT, Cambridge, MA, USA. [4]Department of Chemistry, MIT, Cambridge, MA, USA. [5]Department of Physics, MIT, Cambridge, MA, USA. [6]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [7]Department of Chemistry, Michigan State University, East Lansing, MI, USA. [8]Department of Nuclear Science and Engineering, MIT, Cambridge, MA, USA. [9]Department of Chemical Engineering, MIT, Cambridge, MA, USA. [10]These authors contributed equally: Mouyang Cheng, Weiliang Luo, Hao Tang, Bowen Yu. ✉e-mail: vipandyc@mit.edu; mingda@mit.edu

from sufficiently large datasets, chemically implausible candidates can still be proposed by current state-of-the-art models. Another issue is the thermodynamic stability of the generated materials. Methods such as guided generation[14–16] and post-processing with density functional theory (DFT) help address this, although the DFT's computational cost, particularly for dynamical stability when evaluated with phonon calculations, limit scalability. In addition, while models have shown success in generating materials with specific target properties such as bulk modulus and bandgap[10], incorporating a wider range of functional characteristics, such as thermal and electrical conductivity, or optical behavior, could enhance their real-world utility.

In this work, we present 'crystal generator with valence-constrained design' (CrysVCD), a deep generative architecture that combines diffusion and transformer models to generate ab initio stable crystal structures while respecting fundamental chemical constraints. Built upon recent advances such as MatterGen and DiffCSP, CrysVCD introduces a dedicated transformer module that explicitly ensures charge-balanced compositions. The generation process follows two steps: in stage 1, a transformer-based elemental language model outputs chemically valid formulas where each autoregressive token corresponds to a charged ion, explicitly enforcing oxidation state balance; in stage 2, a conditional diffusion model generates the corresponding atomic structure on the basis of the formula. This design enables rapid screening of chemically plausible compositions during inference, ensuring that all generated crystals adhere to fundamental chemical valence constraints, an essential stability condition that has been underexplored. Within a comparable inference time, CrysVCD enables rapid screening of total chemical valence during stage 1, which guarantees balanced charge valence for all end-product crystals. The stability of generated crystals are evaluated in tandem with a foundation model of the machine learning interatomic potential (MLIP), which serves as a fast surrogate to DFT for estimating both formation energy above the convex hull (E_{hull}) and phonon imaginary frequency as the stability indicators. This enables efficient large-scale screening of generated candidates under physical constraints. By further conditioning the model on these MLIP-labeled stability metrics and fine-tuning, the conditional CrysVCD model achieves >85% stability rate of materials according to the criteria of E_{hull} ($E_{\text{hull}} < 0.1$ eV per atom), and 68% showing no substantial imaginary phonon modes. We further apply CrysVCD to search for materials with high thermal conductivity and high dielectric constant (high-κ) materials, both relevant for microelectronics applications. Overall, CrysVCD provides a general-purpose, physics-constrained framework that improves the reliability and throughput of materials discovery.

Results

Model architecture

We start by briefly reviewing recent diffusion-based generative modeling for materials design. A periodic crystal structure can generally be represented as $\mathbf{M} = (\mathbf{A}, \mathbf{X}, \mathbf{L})$, comprising atom types ($\mathbf{A}$), fractional atomic coordinates ($\mathbf{X}$) and lattice matrices ($\mathbf{L}$). Diffusion models progressively add Gaussian noise to these components over a sequence of time steps. A neural network, often instantiated as a graph neural network (GNN) representing the geometric structure[17,18], is trained to learn the reverse denoising process by approximating the score function[19]. This learned score function enables the generation of diverse crystal structures by sampling from a noise distribution and iteratively denoising it back toward physically plausible configurations. The current crystal structure generative models, such as DiffCSP and MatterGen, perform joint equivariant diffusion of ($\mathbf{A}$, $\mathbf{X}$, $\mathbf{L}$). More specifically, the probability of generating ($\mathbf{A}_0$, $\mathbf{X}_0$, $\mathbf{L}_0$) can be expressed as

$$p(\mathbf{A}_0, \mathbf{X}_0, \mathbf{L}_0) = p_T(\mathbf{A}_T, \mathbf{X}_T, \mathbf{L}_T) \prod_{t=1}^{T} \int \mathrm{d}\mathbf{M}_{T-t+1}\, q_\theta (\mathbf{A}_{T-t}, \mathbf{X}_{T-t}, \mathbf{L}_{T-t} | \mathbf{A}_{T-t+1}, \mathbf{X}_{T-t+1}, \mathbf{L}_{T-t+1}), \tag{1}$$

where $\mathbf{M}_t = (\mathbf{A}_t, \mathbf{X}_t, \mathbf{L}_t)$ is the crystal structure at diffusion time step t, q_θ is a learnable function parametrized by an equivariant GNN and p_T is the prior noise distribution at the final time step T. This enables flexible structure generation and materials design.

However, the above diffusion-based process is entirely data-driven and lacks explicit physical or chemical constraints. Current approaches typically first generate materials and down-select on the basis of constraints, which reduces efficiency. It would be desirable to impose fundamental constraints earlier in the process, for instance, before data generation. For instance, fundamental rules such as stoichiometry should apply to the atom types $\mathbf{A}$. This raises a key question: can we eliminate non-physical candidates before initiating the generative diffusion process? If so, this would enable a lightweight and scalable approach to incorporating domain-specific constraints early on. In light of this, we introduce CrysVCD, a two-step, property-aware crystal generation framework that explicitly incorporates physical and chemical constraints, as shown in Fig. 1.

CrysVCD decouples the generation of crystal compositions and geometries into two integrated and separately trained modules: (1) an elemental language model guided by target properties and chemical valence rules, and (2) a geometric diffusion model that generates 3D atomic structures consistent with the predicted composition and target properties. More specifically, CrysVCD generates atomic species $\mathbf{A}_0$ first using the elemental language model, and then generates the corresponding crystal structures $\mathbf{X}$ and $\mathbf{L}$. Thus, the probability of generating ($\mathbf{A}_0$, $\mathbf{X}_0$, $\mathbf{L}_0$) for CrysVCD is expressed as

$$p(\mathbf{A}_0, \mathbf{X}_0, \mathbf{L}_0) = p(\mathbf{A}_0) \cdot p_T(\mathbf{X}_T, \mathbf{L}_T | \mathbf{A}_0) \prod_{t=1}^{T} \int \mathrm{d}\mathbf{X}_{T-t+1} \mathrm{d}\mathbf{L}_{T-t+1}\, p_\theta(\mathbf{X}_{T-t}, \mathbf{L}_{T-t} | \mathbf{A}_0, \mathbf{X}_{T-t+1}, \mathbf{L}_{T-t+1}), \tag{2}$$

$$\text{where } p(\mathbf{A}_0) = p(a_1) \prod_{i=1}^{n-1} p_\phi(a_{i+1} | a_1, \cdots, a_i). \tag{3}$$

Here, $p_\phi(\cdot)$ is a parametrized transformer model that performs autoregression for each elemental token a_{i+1}, which first generates the target atomic composition $\mathbf{A}_0$ ($\mathbf{A}_0$ is represented by the elemental tokens (a_1, a_2,$\cdots$, a_n)). Then, the second step of CrysVCD is equivalent to sampling $\mathbf{X}$ and $\mathbf{L}$ conditioned on $\mathbf{A}_0$, thus sampling the joint distribution $p(\mathbf{A}_0, \mathbf{X}_0, \mathbf{L}_0) = p(\mathbf{A}_0)p(\mathbf{X}_0, \mathbf{L}_0 | \mathbf{A}_0)$. Thus, $p_\theta(\cdot)$ is a parametrized $E(3)$-equivariant GNN learning the score function with one fewer degree of freedom, effectively performing crystal structure prediction (CSP) during the denoising process. This CSP task is trained and implemented using the architecture of DiffCSP as a demonstration, but it can be achieved with other generative models as well. More details of the elemental language model in stage 1 of CrysVCD are discussed in the 'Chemical formula generation' section in the Methods.

Building on the two-step architecture, CrysVCD brings both efficiency and flexibility to crystal generation. Compared with stage 2, the autoregressive elemental language model in stage 1 requires only a few inference steps (typically four or five), orders of magnitude faster than the diffusion process (typically 1,000 denoising steps), thus imposing chemical constraints at negligible cost (more details shown in Supplementary Section 3). These constraints avoid expensive diffusion calculations during the CSP step on chemically invalid compositions to be dropped by post-generation filters. Therefore, our method improves the overall discovery efficiency on valid materials. Another core feature is the property guidance, which plays a key role in both steps. Target properties, such as E_{hull}, phonon stability or functional attributes such as thermal conductivity and dielectric constants, are embedded and used to condition both the composition generator and the denoising process. This enables CrysVCD to align material generation with desired physical or functional criteria, while remaining modular and extensible for future design goals.

https://doi.org/10.1038/s43588-026-01037-2

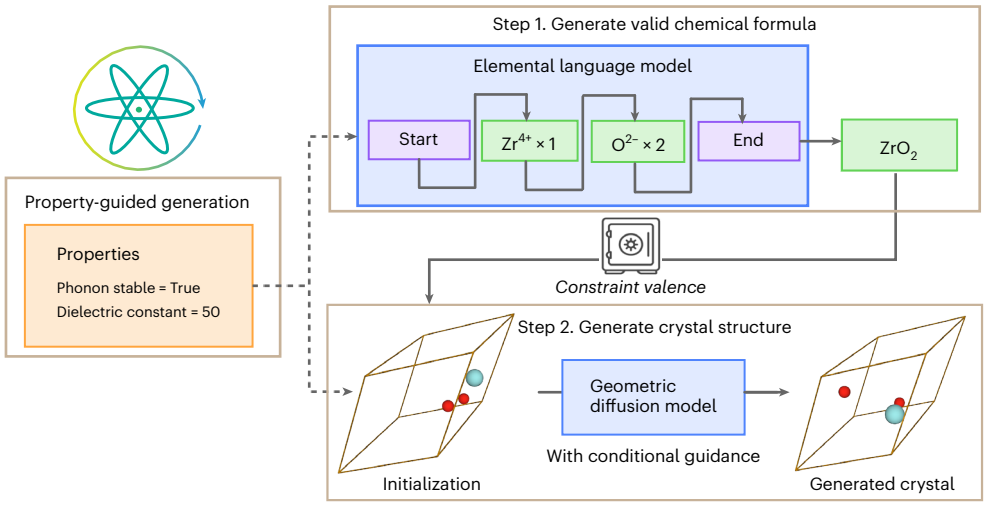

Fig. 1 | Overview of CrysVCD. Rather than performing pure data-driven denoising diffusion generation, CrysVCD introduces a two-step, property-guided pipeline that incorporates physical and chemical constraints. A property-conditioned elemental language model first generates a composition with balanced chemical valence (for instance, ZrO_2), which is then refined by a geometric diffusion model under classifier-free guidance, which interpolates conditional and unconditional outputs. Properties can encode thermodynamic stability as well as target functional performance.

Physics-inspired elemental language model embedding in CrysVCD

As discussed above, a key innovation in CrysVCD is the two-step design that decouples composition generation from structural generation (Fig. 1). In this section, we detail stage 1 for CrysVCD on the elemental language model. Mapping discrete vocabulary to continuous vectors, namely word2vec, is the foundation of the connection between languages and high-dimensional representation spaces in deep learning. In natural language processing, each word is typically assigned a randomly initialized embedding vector, which is learned by optimization through backpropagation across large corpora. However, for our elemental language model, where data scarcity is a substantial challenge, we seek to incorporate prior chemical knowledge into the embedding design. Instead of random initialization, we engineer embeddings that reflect the periodic trends observed in materials science. For example, oxygen and sulfur exhibit similar valence behavior as chalcogens, yet differ in their metal affinities. The embedding also captures relationships and distinctions between multiple oxidation states of the same element, such as the valence diversity of manganese (Mn) from +2 to +7. Inspired by the SpookyNet machine learning potential[20], we use the Aufbau principle to encode the electronic configurations of atoms and ions as the initial embeddings.

Specifically, we construct a vocabulary containing common oxidation states of elements and decompose each material's composition into element–valence pairs and their corresponding counts. Our decomposition algorithm successfully handles mixed-valence cases, such as the coexistence of Fe^{2+} and Fe^{3+} in magnetite (Fe_3O_4), and explains the majority of compositions in the dataset. Compared with the element-wise treatment that assigns the same oxidation state for all atoms of the same element, our algorithm achieves more than threefold improvement of the valence assignment failure rate (from 8.5% to 2.4%) on our relevant chemical space in the MP-20 dataset. In this way, each word corresponds to a token that combines an element's oxidation state with its atom count. As illustrated in Fig. 2a, the model queries an electronic configuration table for element–valence pairs. This table encodes the atomic number, the occupancy of each electronic energy level and the occupancy of the valence shell for each atom or ion. A learnable linear transformation then maps the electronic configurations to a continuous embedding vector, which is passed into a transformer model. During autoregressive generation, the model produces sequences of tokens specifying elements and their counts, which are combined into a chemically valid formula. This formula serves as input to the geometric diffusion model in stage 2. More details on the implementation of our elemental language model can be found in the 'Chemical formula generation' section in the Methods and Supplementary Section 2.1.

Stability-guided generation and fine-tuning in CrysVCD

Building on the CrysVCD architecture, we present examples of generated crystals in Fig. 3a, including alloys and ionic compounds with varying chemical complexity. The oxidation states for all elements are all labeled in Fig. 3a, and all generated crystals are guaranteed to be valence-balanced by construction. To evaluate the structural quality of the generated crystals, we assess two stability metrics: (1) the energy above hull (E_{hull}) and (2) the phonon imaginary frequencies of the generated structures. Structures with low E_{hull} and an absence of imaginary phonon modes are generally regarded as indicators of thermodynamic and dynamic stability, respectively. A key feature of CrysVCD compared with other generative approaches lies in its explicit incorporation of stability-guided learning. Other material generative models often operate in a purely data-driven manner, learning directly from known structure databases composed of stable compounds. This introduces an intrinsic bias into the generative process that limits their ability to generalize beyond the training distribution.

To overcome this limitation, we adopt a concurrent learning framework that integrates unconditional generation with stability-conditioned fine-tuning, as shown in Fig. 2b. The process begins with training an unconditional CrysVCD model on the MP-20 dataset to learn general crystal representations. This model is then used to generate a diverse set of hypothetical compounds, which are explicitly evaluated using high-throughput stability screening. These results are labeled as either stable positive (green) or unstable negative (yellow) and used to construct a new dataset. These newly labeled samples are then aggregated into a training dataset (dotted blue line) for a second-step conditional CrysVCD model, which is trained to generate structures conditioned on stability labels. This closed-loop feedback pipeline enables active learning in CrysVCD, where the model not only generates but also learns from its own outputs. By incorporating both stable and unstable generations, CrysVCD gains insight into structural features that lead to instability and can adapt accordingly.

We benchmark the stability of materials generated by CrysVCD in Fig. 3 in terms of stability metrics. Even without conditional fine-tuning, CrysCVD exhibits a lower E_{hull} distribution compared with the existing

https://doi.org/10.1038/s43588-026-01037-2

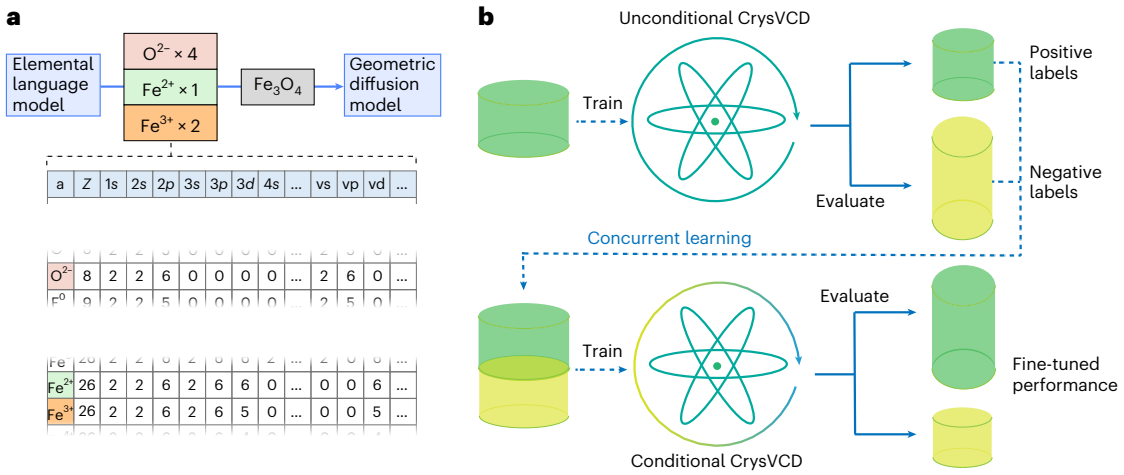

Fig. 2 | Strategies of CrysVCD architecture and training. a, The tokenizer for the elemental language model as stage 1 of CrysVCD. Chemical formulas are first parsed into valence-specific elemental tokens with electronic configurations of oxidation states. These tokenized vectors are then passed into the geometric diffusion model (stage 2) for the structure generation of CrysVCD. **b**, The concurrent learning scheme of CrysVCD for optimizing the generative quality. An unconditional CrysVCD model is initially trained on materials with positive labels, and generated structures are evaluated for target properties (for example, stability metrics). On the basis of these evaluations, structures are labeled as positive or negative and used to fine-tune a conditional CrysVCD model. This concurrent loop enables targeted generation with improved quality in terms of stability or other desired properties.

DiffCSP model, as shown in Fig. 3b, despite the fact that the diffusion model of CrysVCD is DiffCSP. This is attributed to stage 1 of CrysVCD, which ensures balanced chemical valence for generated crystals and demonstrates the generalizability of CrysVCD to improve other generative models. To further evaluate the stability of the generated crystals, we use MatterSim to evaluate the E_{hull} and phonon frequencies, and use these labels to fine-tune the CrysVCD model. The stability fine-tuned CrysVCD model generates substantially lower E_{hull} distributions than the unconditional CrysVCD model. Moreover, the fine-tuned CrysVCD generates fewer phonon imaginary frequencies than the unconditional CrysVCD (Fig. 3c). The portion of phonon-stable configurations (without any imaginary frequency) in the generated crystals increases from 49% (unconditional) to 68% (fine-tuned). These results demonstrate that CrysVCD can effectively generate phonon-stable structures. We further conducted systematic benchmarking of computational speed and generated-structure stability against MatterGen and CDVAE, which are implemented using architectures distinct from the DiffCSP backbone (Supplementary Section 3.1). Under identical training conditions on the MP-20 dataset, our model outperforms or matches these methods in both efficiency and quality.

Generating high-performance functional materials with CrysVCD

Having established the architecture and stability-guided learning framework of CrysVCD, we now demonstrate its capability to generate functional materials tailored for real-world applications. Specifically, we focus on two technologically important categories: materials with intrinsically high thermal conductivity and those exhibiting high dielectric constants (high-κ), both of which are crucial for next-generation microelectronic systems. Materials with high phonon thermal conductivity (k_{ph}), as well as those that are electrically insulating, are vital for thermal management in microelectronics and energy systems[21]. While high-throughput screening have identified promising candidates[22–24], finding stable, intrinsically high k_{ph} remains challenging. Diamond leads in k_{ph} but has a high cost, and other recent generative efforts are limited to carbon allotropes[25].

To discover more high-k_{ph} materials, we fine-tuned CrysVCD using a screened subset of the MP-20 dataset guided by phonon transport knowledge. As shown in Fig. 4a, we prioritize materials that are likely to exhibit high phonon group velocities and long phonon lifetimes[26] by focusing on materials with lighter elements, strong covalent or mixed bonding and simple unit cells (<12 atoms), which can increase group velocities and reduce phonon scattering. Moreover, we exclude metallic compounds and gas-phase species to focus on phonon-dominated heat transport.

The resulting screened dataset serves both as the training set for conditional CrysVCD, as well as a benchmark for assessing its ability to emulate expert-driven design. After fine-tuning the conditional generative model for thermal conductivity targets, CrysVCD shifts its generative distribution toward high-k_{ph} candidates (Fig. 4b). Compared with the base model, the conditioned version generates structures with k_{ph} values aligned with expert-curated materials.

Among the generated compounds, we highlight hexagonal GeC with high thermal conductivity $k_{ph} = 183\ W\ m^{-1}\ K^{-1}$. Notably, this material does not appear in the MP-20 dataset. Its crystal structure and phonon band structure are shown in Fig. 4c. Combined with a wide bandgap of 2.34 eV, GeC is a promising candidate for thermal management applications. The large acoustic–optical phonon gap shown suppresses phonon scattering, contributing to its high thermal conductivity. Interestingly, given its chemical simplicity, hexagonal GeC is present in the Materials Project database, and prior DFT studies have predicted cubic GeC to exhibit very high thermal conductivity[27] ($270\ W\ m^{-1}\ K^{-1}$). However, so far, no crystalline phase of GeC, either cubic or hexagonal, has been experimentally synthesized, and the thermal transport properties of crystalline GeC therefore remain unexplored experimentally. CrysVCD's ability to identify hexagonal GeC with a similarly high predicted thermal conductivity highlights its potential for discovering physically consistent functional materials beyond current experimental coverage.

To further validate the efficacy of CrysVCD, we conducted another study on high-κ dielectrics, that is, materials with a high dielectric constant ε. These materials are of substantial technological interest for applications in capacitors, transistors and other microelectronic devices. Owing to the high computational cost of DFT-based optical properties, it is impractical to evaluate each generated structure using first-principle methods. To address this bottleneck, we leverage an existing database and trained another machine learning-based surrogate model that can provide rapid and reasonably accurate predictions of ε. As shown in Fig. 4d, we began by collecting 944 entries from the Materials Project within the MP-20 dataset, which contains DFT-computed dielectric data. This dataset is used for both fine-tuning

https://doi.org/10.1038/s43588-026-01037-2

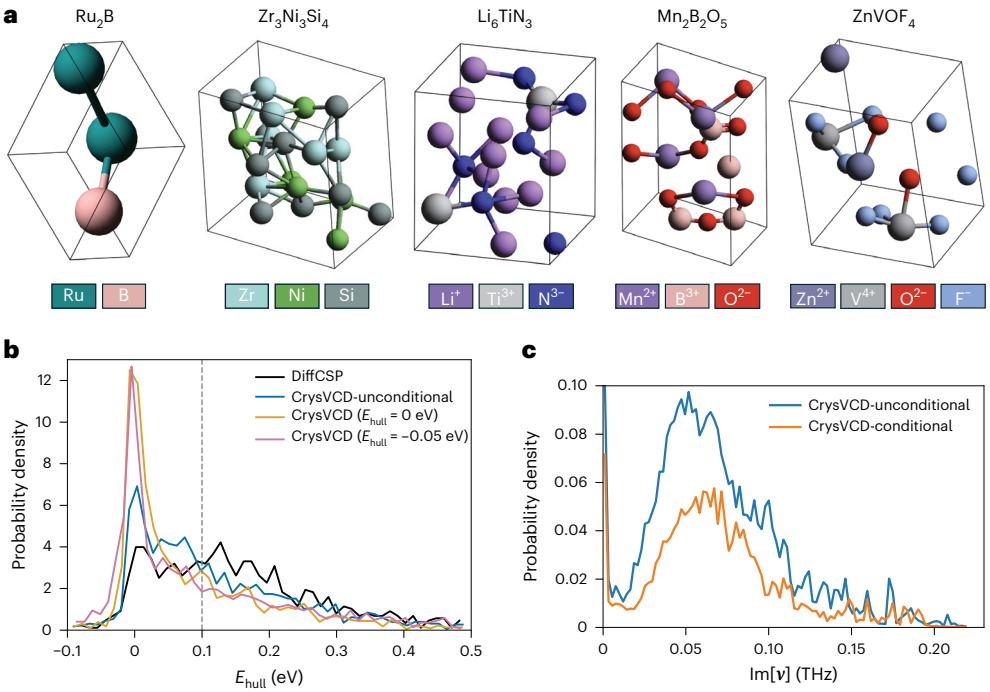

Fig. 3 | Thermodynamic and phonon stability of generated crystals by CrysVCD. a, Examples of generated crystals from CrysVCD, including alloys (binary and ternary) and ionic compounds with oxidation states (ternary and quaternary) from left to right. **b**, The normalized probability distribution of the energy above hull (E_{hull}) of 4,000 crystal structures generated by different versions of CrysVCD models and the previous generative model DiffCSP[9] without fine-tuning on E_{hull}. CrysVCD fine-tuned with an E_{hull} threshold of 0 and −0.05 eV per atom, and CrysVCD without conditional fine-tuning are included in the comparison. **c**, The normalized probability distribution of phonon imaginary frequencies of 4,000 generated crystal structures by CrysVCD before and after conditional fine-tuning with labels of phonon stability evaluated by the MatterSim-v1.0.0-5M model[10].

the CrysVCD model for conditional generation and training the surrogate model for predicting the dielectric constant. The surrogate model is implemented as an $E(3)$-equivariant GNN[18,28], adapted from a GNNOPT architecture for optical property prediction[29]. Once trained, the GNNOPT model is integrated into the conditional generation loop of fine-tuned CrysVCD to rapidly evaluate and guide the selection of candidate materials with high ε. The predictive performance of the $E(3)$-equivariant surrogate model is shown in Supplementary Section 3.

The resulting distribution of ε for generated materials, shown in Fig. 4e, exhibits a clear shift toward higher ε values compared with the original MP-20 dataset. During conditional generation, we explicitly set the target property to $\varepsilon = 50$, guiding CrysVCD to prioritize candidates above this threshold. As a result, a substantial number of generated materials achieve $\varepsilon > 50$, in contrast to the original dataset, where such high-dielectric entries are rare. This demonstrates the effectiveness of CrysVCD to steer the generative process toward chemically valid, high-ε candidates. Among the top-performing candidates, we identify $CoSnO_2$, which is not present in the Materials Project database. It exhibits a high dielectric constant, with $\varepsilon_\theta = 62.2$ predicted by the surrogate model, and a principal component value of $\varepsilon_{DFT,zz} = 70.5$ verified by DFT (Fig. 4f). However, it should be noted that these results rely on a surrogate model trained on DFT-computed dielectric data, which may still deviate from experimental values. Improving surrogate accuracy and bridging the gap between DFT and real-world behavior remain important directions for future work.

Discussion

Our work highlights the potential of chemical composition as a natural modality for generative modeling in materials science, which is widely supported in predictive models[30,31]. By first generating valid chemical compositions and then producing their crystal structures, our work helps yield stable candidates that are consistent with valence balance and bonding principles. Unlike purely data-driven approaches, our framework is explicitly guided by physical and chemical constraints. The chemical composition generation is explicitly guided by charge neutrality constraints, and element embeddings are initialized using electronic configurations. This design enables the model to capture periodic trends, accommodate multiple oxidation states and handle complex chemistry such as mixed-valence compounds.

To address the scarcity and uneven distribution of experimental data, such as thermal conductivity or dielectric constants, we integrate the generation pipeline with machine-learning-accelerated predictive models. These models are not only applied for efficient post hoc evaluation of generated candidates, but also provide synthetic property labels that can be used to steer the generative model in subsequent rounds, or accommodate data augmentation from human expertise. This generation–prediction concurrence transforms the model into an active explorer of the chemical space, extrapolating beyond the support of the original dataset and proposing materials with novel compositions and desirable properties.

Nevertheless, the performance of CrysVCD is still constrained by the MP-20 training data, which excludes materials containing lanthanides, actinides, molecular crystals and other complex motifs such as zeolites and metal–organic frameworks. Within the remaining chemical space in the MP-20 dataset, although our valence decomposition covers 97.6% of the materials, there are potentially high-value materials with exotic oxidation states being overlooked by our model. For those successful valence decompositions, we emphasize that they provide a chemically consistent interpretation of composition, rather than a unique assignment of oxidation states. Multiple valence configurations may satisfy the same stoichiometry and represent distinct redox tendencies. Therefore, in our framework, valence serves primarily as a constraint to guide the generation toward chemically plausible compositions, while the definitive valence assignment must ultimately be assessed through downstream evaluations, such as electronic-structure analysis and crystal structure validation.

https://doi.org/10.1038/s43588-026-01037-2

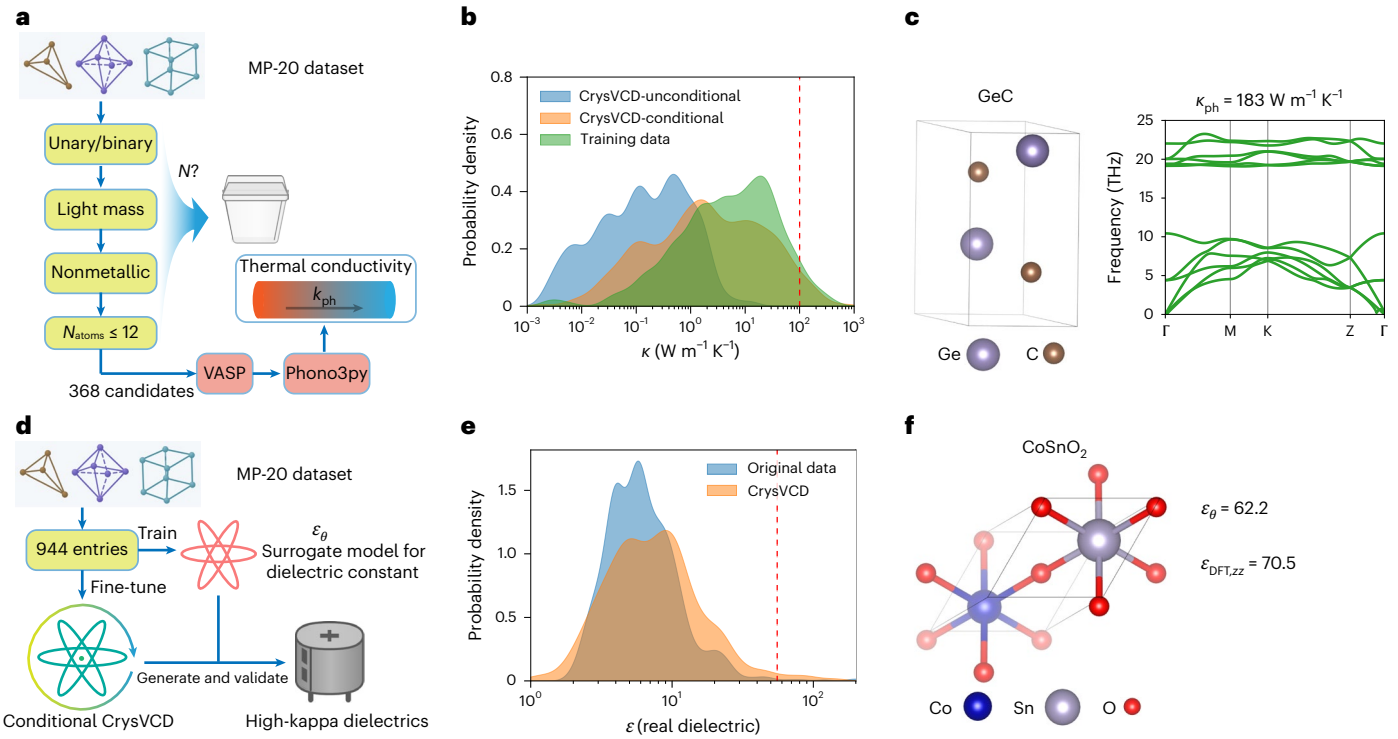

Fig. 4 | Generation of high-performance functional materials with CrysVCD. **a**, The workflow of first-principles calculations to generate a thermal conductivity dataset considering phonon anharmonicity. **b**, The probability distribution of the thermal conductivity of materials generated by the unconditional CrysVCD model and CrysVCD fine-tuned to maximize thermal conductivity. The distribution of high thermal conductivity materials from the training dataset selected by human intuition is shown as a reference. The dashed red line denotes $k_{ph} = 100$ W m^{-1} K^{-1} as the generative condition. **c**, First-principles calculations of the phonon band structure of a high thermal conductivity material, GeC, generated by the CrysVCD model fine-tuned to maximize thermal conductivity. **d**, The workflow for conditional CrysVCD generation on a dielectric constant. A surrogate model for a static dielectric constant (ε) is trained on 944 MP-20 entries, then used to guide CrysVCD toward generating structures with high ε. **e**, The probability distribution of ε for 1,800 generated materials (orange) versus original MP-20 dataset (blue), showing a rightward shift indicating successful targeting of high-ε candidates. The dashed red line denotes $\varepsilon = 50$ as the generative condition. **f**, A new high-κ candidate generated by CrysVCD $CoSnO_2$. The surrogate model predicts a dielectric constant of $\varepsilon_\theta = 62.2$, with DFT verification yielding $\varepsilon_{DFT,zz} = 70.5$.

Besides chemical diversity and validity, we acknowledge that our current framework is limited to the discovery of ideal crystalline materials, which assume integer stoichiometries and determined atomistic structures. However, mechanisms mediated by defects, doping and disordered structures are essential to many desired properties and relevant applications. Although those important characteristics have not been considered in our framework, we note that our valence-based compositional modeling remains promising with future extensions to defect-aware generation. For example, materials with point defects typically derive from perturbations of ideal valence-balanced compositions rather than complete violations of the valence rules. Compared with generative models without any constraints on chemical composition, our framework provides a chemically grounded starting point that could serve as a foundation for generating more realistic and stable defect-containing structures, with the development of modeling in nonideal crystal structure models.

In addition, we emphasize that verifying experimental synthesizability remains a fundamental and ongoing challenge in materials discovery. Standard computational criteria, such as phonon stability or E_{hull} at zero temperature, are neither sufficient nor necessary conditions for experimental realization. Many experimentally synthesized materials are known to exhibit relatively high DFT-predicted energies above the hull at zero temperature, while conversely, some thermodynamically and dynamically stable phases are difficult to synthesize owing to kinetic barriers, growth conditions or metastability effects. Bridging this gap between computational stability and experimental realizability will require a systematic incorporation of experimental data, finite-temperature effects and synthesis-aware constraints.

In the future, we hope the growth of material databases and new backbone algorithms for crystal generation can be incorporated into our model and extend it to a broader chemical and structural space. Ultimately, CrysVCD is designed as a modular and extendable plugin architecture that can be readily integrated into existing generative frameworks to enforce chemically valid discovery toward desired properties.

Methods

Chemical formula generation

We define a set of 'valid atom types' (A) by excluding noble gases, lanthanides, actinides and other radioactive species from the periodic table. We manually compiled a list of commonly observed nonzero oxidation states in ionic compounds and assigned an additional zero valence to those atom types that typically appear in alloys, forming a set of valid oxidation states $\mathbf{V} = \{V(a): a \in A\}$. Using this oxidation state list, we developed a valence labeling algorithm for the compounds containing only valid atom types in the MP-20 subset of the Materials Project[32]. Each material's valence assignment is represented as a set of tuples $\{(a_i^{v_i}, c_i)\}$, where a^v denotes the atom type a in valence state v, and c denotes the number of atoms of type a and valence v in the unit cell. A valid assignment must satisfy the charge neutrality condition $\sum c_i v_i = 0$.

For a material with a unit-cell stoichiometry $\mathbf{A}_x\mathbf{B}_y\cdots\mathbf{C}_z$, we classify it as an alloy if $0 \in V(a)$ for all $a \in \{\mathbf{A}, \mathbf{B}, \cdots, \mathbf{C}\}$, and label it as $\{(\mathbf{A}^0, x), (\mathbf{B}^0, y), \cdots, (\mathbf{C}^0, z)\}$. Otherwise, it is treated as an ionic compound. For ionic materials, we perform a breadth-first search over all nonzero valence combinations of the atoms in its stoichiometric unit from $\mathbf{V}$ to identify a valid assignment. This atom-wise treatment accommodates

mixed-valence states in complex compositions of ionic compounds. For example, Fe_3O_4 is labeled as {(Fe^{2+}, 1), (Fe^{3+}, 2), (O^{2-}, 4)}, and $Fe_4[Fe(CN)_6]_3$ (Prussian Blue) is labeled as {(Fe^{2+}, 3), (Fe^{3+}, 4), (C(II), 18), (N(-III),18)}. To improve efficiency, the search branch is pruned if remaining atoms with their plausible valence ranges in $\mathbf{V}$ cannot neutralize the searched atoms. In addition, we impose comproportionation constraints to avoid infeasible configurations: no element may simultaneously exhibit both positive and negative oxidation states, and if three oxidation states $u < v < w$ appear in $V(a)$, then coexistence of a^u, a^w is prohibited. Materials that fail to obtain a valid valence assignment or contain invalid elements are excluded. The remaining successfully labeled materials form the dataset 'MP-20-valence', which has 8,299 alloys and 7,491 ionic compounds in its training partition. We note that the comproportionation constraint introduces an inductive bias for computational tractability; in cases such as proton (H^+)/hydride (H^-) defect coexistence in semiconductors[33], elements may simultaneously exhibit positive and negative oxidation states, which are not captured under the present scheme.

We adopt the GPT-2[34] implementation on the Huggingface platform[35] for our chemical formula transformer, and design a domain-specific embedding scheme tailored to the valence assignment representation in MP-20-valence. For a count c, ranging from 0 to 20 in all MP-20-valence entries, we embed it into 21 learnable word embeddings $\mathbf{h}_c \in \mathbb{R}^d$. For a valence-labeled atom type a^v we define a vocabulary of 217 unique tokens in $\{a^v: a \in A, v \in V(a)\}$, augmented with two special tokens: '<START>' and '<END>'. They are first embedded into 219 learnable word embeddings $h^{(1)}_{a^v} \in \mathbb{R}^d$. Furthermore, inspired by the electronic embedding proposed in SpookyNet[20], we encode the electron counts of each orbitals in their full electronic configuration and the valence shell configuration for each a^v token by a linear transformation into a vector $\mathbf{h}^{(2)}_{a^v} \in \mathbb{R}^d$, while special tokens are mapped to zero vectors of the same dimensionality. As a result, each (a^v, c) pair is embedded as the combination $\mathbf{h}^{(1)}_{a^v} + \mathbf{h}^{(2)}_{a^v} + \mathbf{h}_c \in \mathbb{R}^d$. A valence assignment of l pairs will be converted to a sequence, padded with (<START>,0) and (<END>,0) at its head and tail, and transformed into an embedding sequence of $\mathbb{R}^{(L+2)\times 41}$, which serves as the word embedding of the transformer.

The transformer performs next-token prediction in an autoregressive way. The next-token probability is predicted as a concatenated logit $[\mathbf{y}_{a_v} || \mathbf{y}_c] \in \mathbb{R}^{240}$, where $\mathbf{y}_{a_v} \in \mathbb{R}^{219}$ corresponds to the logit of valence-labeled element tokens, and $\mathbf{y}_c \in \mathbb{R}^{21}$ corresponds to the logit of count tokens. In the training stage, we augment each training data entry by at most three times with random permutations of the valence assignment tuples. Two transformers for alloys and ionic compounds are separately trained with the total cross-entropy loss of the two logits in a teacher-forcing scheme[36]. In the inference stage, the sequence generation starts with (<START>,0), extends by sampling the next token (a^v, c) with the highest probability and ends when it reaches the maximum length of ten or generates <END> for a^v. We append a filter to the transformer for ionic compounds to ensure that only charge-balanced generation results are passed to the diffusion model. The initial atom-type condition $\mathbf{A}_0$ of the diffusion model in equation (2) is compiled from a_i and c_i of a valid generated sequence. The chemical formula transformer is thus trained using an autoregressive cross-entropy loss $\mathcal{L}_{\mathrm{CE}}$ over the valence-labeled elemental tokens and atom-count tokens. Specifically, the training objective of the chemical formula transformer is

$$\mathcal{L}_{\mathrm{CE}} = -\mathbb{E}_{(a_i^v, c_i)} \sum_i \left[\log p_\phi \left(a^v_{i+1} | a^v_{\le i}, c_{\le i}\right) + \log p_\phi \left(c_{i+1} | a^v_{\le i}, c_{\le i}\right)\right], \quad (4)$$

where p_ϕ denotes the autoregressive transformer, and a_i^v and c_i are the ith valence-labeled element token and atom-count token.

Property-guided conditional crystal generation

Our model supports two types of properties as the generation condition. The first type is binary properties, taking values in {0, 1}, such as phonon stability. The second type is continuous properties, represented as real-valued scalars, including properties such as E_{hull} and thermal conductivity. Each property has an adapter channel that transforms its value into an embedding vector $\mathbf{h}_p$ for the chemical formula transformer and the crystal diffusion model, respectively. A binary property is embedded by a word-embedding layer of size 2, while a continuous property is transformed by a linear mapping. When conditioned on multiple properties, all property embeddings are added together.

The chemical formula transformer adds its own $\mathbf{h}_p$ to the atom-valence and count embedding of the <START> during training on property-labeled datasets and during conditioned generation. The crystal diffusion model adds its own $\mathbf{h}_p$ to each initial atomic feature in its score estimator. The geometric diffusion model is first trained unconditionally using the standard denoising score-matching objective for diffusion models,

$$\mathcal{L}_{\mathrm{uncond}} = \mathbb{E}_{t, \mathbf{X}_0, \mathbf{L}_0, \epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{X}_t, \mathbf{L}_t, t)\|^2\right], \quad (5)$$

where $\epsilon \sim (\mathbf{0}, \mathbf{I})$ is the Gaussian noise added at diffusion step t, and $\boldsymbol{\epsilon}_\theta$ is the $E(3)$-equivariant network predicting the injected noise. More details regarding the crystal structure generation, including the incorporation of periodic boundary conditions, are shown in Supplementary Section 2.2.

To enable conditional generation, we adopt classifier-free guidance[14], in which the same network is trained to also model the conditional distribution by providing the property embedding $\mathbf{h}_p$ as an additional input. During training, $\mathbf{h}_p$ is randomly dropped with a fixed probability, such that the conditional and unconditional objectives are jointly optimized using a single loss,

$$\mathcal{L}_{\mathrm{CFG}} = \mathbb{E}_{t, \mathbf{X}_0, \mathbf{L}_0, \epsilon, \mathbf{h}_p}\left[\left\|\epsilon - \epsilon_\theta(\mathbf{X}_t, \mathbf{L}_t, t | \mathbf{h}_p)\right\|^2\right]. \quad (6)$$

No explicit regression loss is imposed on the target properties; instead, property information enters solely through conditioning. We adopted a guidance weight $w = 2$ to scale the distribution conditioned on $\mathbf{h}_p$. It replaces $p(\mathbf{X}_t, \mathbf{L}_t)$ in equation (2) during the conditional denoising process.

$$\begin{aligned} p_w(\mathbf{X}_t, \mathbf{L}_t | \mathbf{h}_p) &\propto p(\mathbf{h}_p | \mathbf{X}_t, \mathbf{L}_t)^w p(\mathbf{X}_t, \mathbf{L}_t) \\ &\propto \left(\frac{p(\mathbf{X}_t, \mathbf{L}_t | \mathbf{h}_p)}{p(\mathbf{X}_t, \mathbf{L}_t)}\right)^w p(\mathbf{X}_t, \mathbf{L}_t) \\ &\propto p(\mathbf{X}_t, \mathbf{L}_t | \mathbf{h}_p)^w p(\mathbf{X}_t, \mathbf{L}_t)^{1-w} \end{aligned}. \quad (7)$$

During inference, that is, the generation of crystal structures, the conditional score corresponding to equation (2) is obtained by taking gradients of the logarithm with respect to $\mathbf{X}_t$ and $\mathbf{L}_t$.

$$\nabla \ln p_w(\mathbf{X}_t, \mathbf{L}_t | \mathbf{h}_p) = w \nabla \ln q_\theta(\mathbf{X}_t, \mathbf{L}_t | \mathbf{h}_p) + (1 - w) \nabla \ln q_\theta(\mathbf{X}_t, \mathbf{L}_t),$$

where the conditional score is estimated by the score model with property embeddings $\mathbf{h}_p$, and the unconditional score is obtained by replacing $\mathbf{h}_p$ with a zero vector.

MLIP and DFT calculations

Machine-learning interatomic potentials (MLIPs) provide a computationally efficient surrogate for first-principles electronic-structure methods such as DFT by learning the potential energy surfaces (PESs) from large sets of first-principles calculations. Once trained, MLIPs can predict energies, forces and Hessians on the PES with near-DFT accuracy at a fraction of the computational cost, enabling large-scale structural relaxation, formation energy calculations and phonon analyses that were previously computationally prohibitive with DFT alone.

This makes MLIPs particularly suitable for high-throughput evaluation of candidates generated by data-driven models, where both physical fidelity and computational efficiency are required.

In this work, we adopt the pretrained MatterSim models[37] as our MLIP backend to perform structural relaxation, E_{hull} evaluation and phonon stability analysis in tandem with our generative model. MatterSim models are a family of foundation models of MLIPs pretrained on 17 million DFT-calculated structures from a diverse chemical and configurational space of inorganic crystals, which sufficiently cover the MP-20 chemical space we generate the crystal from. They fuse the M3GNet[38] and GeoMFormer[39] GNN architecture. They encode the atom type and the local geometry (two-body and three-body combinations in a distance cutoff) of each atom with equivariant descriptors (atom-type embeddings, distances and spherical harmonics of bond angles), model their interactions by equivariant message-passing mechanisms and derive the forces with auto-differentiation of the energy, thus preserving the permutational, translational and rotational symmetry of the PES and energy conservation during the simulation in physical reality. Furthermore, previous benchmarks have shown that these models achieve near-DFT accuracy for formation energies, atomic forces and phonon-related quantities within their applicable domain[40]. Therefore, MatterSim models in our framework act as a reliable stability filter and thermal conductivity estimator, which enables scalable virtual screening and labeling mechanism while preserving a clear path toward first-principle verification.

MatterSim models are connected to the Atomic Simulation Environment (ASE) Python package in our simulation protocols[41]. The E_{hull} is evaluated using the MatterSim-v1.0.0-1M model with the convex hull information from the Materials Project database[32] using the pymatgen interface[42]. The atomic configuration is relaxed with the geometry optimization engine in ASE until the max atomic force is less than 0.01 eV Å^{-1} before the E_{hull} evaluation. For phonon stability, we use the MatterSim-v1.0.0-5M model to evaluate the phonon band structure. The atomic configuration is first relaxed until the maximum atomic force is less than 0.003 eV Å^{-1} using the BFGS algorithm[43]. Note that higher accuracy settings are adopted owing to the high requirements for phonon calculations. If the relaxation does not converge within 200 steps, the structure is considered unstable. The force constant matrix of the system is then calculated by the finite difference method with a differential step size of 0.1 Å. The phonon spectrum is evaluated using the phonon module of ASE with a supercell size of at least 20 Å in each direction and a k-point mesh of $8 \times 8 \times 8$. Phonon frequencies at all k points in each configuration are included in the statistics shown in Fig. 3b. A structure is considered dynamically unstable if substantial imaginary phonon modes (imaginary frequencies over 0.01 THz) are present, excluding numerical noise near zero, and the criteria remains consistent with the screening of thermal conductivity.

For the calculation of thermal conductivity, we use the phono3py package[44,45], which requires second-order and third-order force constants to perform anharmonic phonon scattering calculations. We generate 1,800 materials with a fine-tuned CrysVCD model on thermal conductivity, and supply these second-order and third-order force constants using MatterSim-v1.0.0-5M for screening, and DFT results for benchmarking on the 368 candidates (Fig. 4a), as well as verification of the high thermal conductivity candidate GeC (Fig. 4c). For both MLIP and DFT calculations of force constants, we construct supercells using the finite-displacement approach under the frozen phonon method. For the calculation of dielectric constants, we use the GNNOPT surrogate model to screen 4,000 candidates generated by a fine-tuned CrysVCD model, as discussed in the main text. Then we use density functional perturbation theory to verify the high-kappa dielectric $CoSnO_2$. All DFT calculations are conducted using the Vienna Ab initio Simulation Package (VASP, version 6.4.3)[46]. The projector-augmented wave method[47,48] is used, with exchange-correlation effects described by the Perdew–Burke–Ernzerhof formulation of the generalized gradient approximation[49]. The plane-wave cutoff energy is set as 1.2 × max(ENMAX) for sufficient convergence. Calculations are performed on a k-point mesh centered at the Γ point, with resolved value $k_{mesh} = 0.03 \times 2\pi$ Å^{-1} for each structure. In addition, our data analysis on DFT results utilizes the VASPKIT package[50].

Data availability

The training data and generated crystals with predicted functional properties are available via Zenodo at https://doi.org/10.5281/zenodo.20453419 (ref. 51). Source data are provided with this paper.

Code availability

The code used in this study is available via Zenodo at https://doi.org/10.5281/zenodo.20453419 (ref. 51) and via GitHub at https://github.com/vipandyc/CrysVCD.

References

1. Li, J. Universal interatomic potentials shine in finding crystal structures. *Nat. Mach. Intell.* **7**, 985–986 (2025).
2. Olson, G. B. Computational design of hierarchically structured materials. *Science* **277**, 1237–1242 (1997).
3. Olson, G. B. Designing a new material world. *Science* **288**, 993–998 (2000).
4. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **4**, 331–348 (2019).
5. Zunger, A. Inverse design in search of materials with target functionalities. *Nat. Rev. Chem.* **2**, 0121 (2018).
6. Han, X.-Q. et al. AI-driven inverse design of materials: past, present and future. *Chin. Phys. Lett.* **42**, 027403 (2025).
7. Cheng, M. et al. Artificial intelligence-driven approaches for materials design and discovery. *Nat. Mater.* **25**, 174–190 (2026).
8. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations* (2022).
9. Jiao, R. et al. Crystal structure prediction by joint equivariant diffusion. *Adv. Neural Inf. Process. Syst.* **36**, 17464–17497 (2023).
10. Zeni, C. et al. A generative model for inorganic materials design. *Nature* **639**, 624–632 (2025).
11. Gruver, N. et al. Fine-tuned language models generate stable inorganic materials as text. In *International Conference on Learning Representations* (2024).
12. Antunes, L. M., Butler, K. T. & Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nat. Commun.* **15**, 10570 (2024).
13. Yang, S. et al. Generative hierarchical materials search. *Adv. Neural Inf. Process. Syst.* **37**, 38799–38819 (2025).
14. Ho, J. & Salimans, T. Classifier-free diffusion guidance. Preprint at https://arxiv.org/abs/2207.12598 (2022).
15. Dhariwal, P. & Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **34**, 8780–8794 (2021).
16. Karras, T. et al. Guiding a diffusion model with a bad version of itself. *Adv. Neural Inf. Process. Syst.* **37**, 52996–53021 (2024).
17. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
18. Reiser, P. et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **3**, 93 (2022).
19. Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **33**, 6840–6851 (2020).

20. Unke, O. T. et al. SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **12**, 7273 (2021).
21. Zheng, Q., Hao, M., Miao, R., Schaadt, J. & Dames, C. Advances in thermal conductivity for energy applications: a review. *Prog. Energy* **3**, 012002 (2021).
22. Plata, J. J. et al. An efficient and accurate framework for calculating lattice thermal conductivity of solids: AFLOW-AAPL automatic anharmonic phonon library. *npj Comput. Mater.* **3**, 45 (2017).
23. Póta, B., Ahlawat, P., Csányi, G. & Simoncelli, M. Thermal conductivity predictions with foundation atomistic models. Preprint at https://arxiv.org/abs/2408.00755 (2024).
24. Li, J. et al. Probing the limit of heat transfer in inorganic crystals with deep learning. Preprint at https://arxiv.org/abs/2503.11568 (2025).
25. Guo, L. et al. Generative deep learning for predicting ultrahigh lattice thermal conductivity materials. *npj Comput. Mater.* **11**, 97 (2025).
26. Togo, A., Chaput, L. & Tanaka, I. Distributions of phonon lifetimes in Brillouin zones. *Phys. Rev. B* **91**, 094306 (2015).
27. Lindsay, L., Broido, D. & Reinecke, T. Phonon-isotope scattering and thermal conductivity in materials with a large isotope effect: a first-principles study. *Phys. Rev. B* **88**, 144306 (2013).
28. Geiger, M. & Smidt, T. e3nn: euclidean neural networks. Preprint at https://arxiv.org/abs/2207.09453 (2022).
29. Hung, N. T., Okabe, R., Chotrattanapituk, A. & Li, M. Universal ensemble-embedding graph neural network for direct prediction of optical spectra from crystal structures. *Adv. Mater.* **36**, 2409175 (2024).
30. Ma, A. et al. Topogivity: a machine-learned chemical rule for discovering topological materials. *Nano Lett.* **23**, 772–778 (2023).
31. Ma, A. & Soljačić, M. Learning simple heuristic rules for classifying materials based on chemical composition. Preprint at https://arxiv.org/abs/2505.02361 (2025).
32. Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
33. Van de Walle, C. G. & Neugebauer, J. Universal alignment of hydrogen levels in semiconductors, insulators and solutions. *Nature* **423**, 626–628 (2003).
34. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
35. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (eds Liu, Q. & Schlangen, D.) 38–45 (2020).
36. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 5998–6008 (2017).
37. Yang, H. et al. MatterSim: a deep learning atomistic model across elements, temperatures and pressures. Preprint at https://arxiv.org/abs/2405.04967 (2024).
38. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
39. Chen, T. et al. Geomformer: a general architecture for geometric molecular representation learning. In *Proceedings of the 41st International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) Vol. 235, 7118–7142 (2024).
40. Riebesell, J. et al. A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intell.* **7**, 836–847 (2025).
41. Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
42. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
43. Dai, Y.-H. Convergence properties of the BFGS algoritm. *SIAM J. Optim.* **13**, 693–701 (2002).
44. Togo, A., Chaput, L., Tadano, T. & Tanaka, I. Implementation strategies in phonopy and phono3py. *J. Phys. Condens. Matter* **35**, 353001 (2023).
45. Togo, A. First-principles phonon calculations with phonopy and phono3py. *J. Phys. Soc. Jpn.* **92**, 012001 (2023).
46. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169 (1996).
47. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953 (1994).
48. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758 (1999).
49. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
50. Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. VASPKIT: a user-friendly interface facilitating high-throughput computing and analysis using VASP code. *Comput. Phys. Commun.* **267**, 108033 (2021).
51. Cheng, M. et al. Enhancing materials discovery with valence constrained design in generative modeling. *Zenodo* https://doi.org/10.5281/zenodo.20453419 (2026).

Acknowledgements

We thank G. Daras, B. R. Ortiz, Q. Yan and R. J. Cava for helpful discussions.

Author contributions

M.C. and M.L. conceived the project. M.C. and W.L. developed the generative modeling workflow and performed model training. W.L. and M.C. designed the chemical valence assignment and constraint strategy. M.C. and H.T. carried out the first-principles calculations. M.C., B.Y. and W.L. conducted the high-throughput evaluation and concurrent learning, and prepared the figures. M.L. supervised the project. All authors contributed to preparing and revising the manuscript.

Funding Statement

M.C. acknowledges the support of the US Department of Energy, Office of Science, Basic Energy Sciences (award no. DE-SC0021940). W.L. and H.J.K. acknowledge the support of the US Department of Energy, Office of Science, Office of Advanced Scientific Computing, Office of Basic Energy Sciences, via the Scientific Discovery through Advanced Computing program. H.T. acknowledges the support of the Mathworks Engineering Fellowship. Y.C. is partly supported by the Scientific User Facilities Division, Office of Basic Energy Sciences, US Department of Energy (contract no. DE-AC0500OR22725 with UT Battelle, LLC). A portion of the DFT calculations used computing resources made available through the VirtuES project, funded by the Laboratory Directed Research and Development program and Compute and Data Environment for Science at Oak Ridge National Laboratory. M.L. acknowledges the support of National Science Foundation (award no. ITE-2345084) and the support from R. Wachnik. J.L. acknowledges support by DTRA (award no. HDTRA1-20-2-0002) Interaction of Ionizing Radiation with Matter University Research Alliance.

Competing interests

M.L. and M.C. have filed a patent application (US Patent application no. 64/105,647) on technology related to the processes described in this article.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-026-01037-2.

Correspondence and requests for materials should be addressed to Mouyang Cheng or Mingda Li.

Peer review information *Nature Computational Science* thanks Wei Chen, Zhong-Yi Lu, Alberto Roldan and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Kaitlin McCardle, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.