Enhancing Materials Discovery with Valence Constrained Design in Generative Modeling: Supplementary Information

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

[1]Quantum Measurement Group, MIT, Cambridge, MA 02139, USA
[2]Center for Computational Science & Engineering, MIT, Cambridge, MA 02139, USA
[3]Department of Materials Science and Engineering, MIT, Cambridge, MA 02139, USA
[4]Department of Chemistry, MIT, Cambridge, MA 02139, USA
[5]Department of Physics, MIT, Cambridge, MA 02139, USA
[6]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[7]Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA
[8]Department of Nuclear Science and Engineering, MIT, Cambridge, MA 02139, USA
[9]Department of Chemical Engineering, MIT, Cambridge, MA 02139, USA
[†]These authors contributed equally.
[*]e-mail: vipandyc@mit.edu
[**]e-mail: mingda@mit.edu

Contents

1 Preparation of dataset

1.1 Statistics of crystal structure and composition

In this section, we present key statistics regarding the structural prototypes and chemical compositions in the dataset to provide a comprehensive understanding of the data distribution used for CrysVCD.

Regarding chemical diversity, Fig. S1 illustrates the dataset's elemental distribution, covering a total of 84 distinct elements. Fig. S1(a) and (b) highlight the frequency of each element within the training and validation datasets, respectively. Common elements, including O, Cu, Li, F, and S, exhibit substantial representation, while rare-earth and heavy elements appear comparatively infrequently. Notably, noble gases are completely absent from the dataset due to their limited propensity for crystal formation.

Fig. S2 provides an overview of crystal structure diversity. Panel (a) depicts the distribution of crystal compositions across the entire dataset–including both training and validation sets. The majority of crystals fall within the

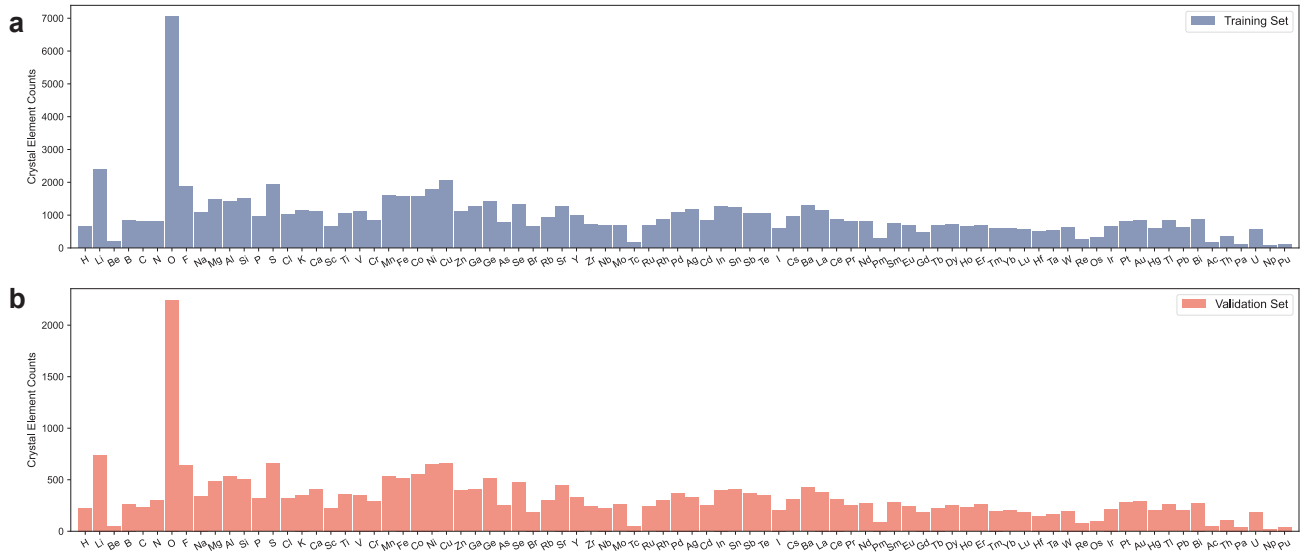

Figure S1. Elemental distribution statistics among the training and validation datasets. **a.** Elemental composition of crystals in the training dataset. **b.** Elemental composition of crystals in the validation dataset.

binary to quinary range, with only a minimal number of crystals consisting of more than five elements. Among all crystals in the dataset, ternary crystals are notably predominant, with more than 20,000 occurrences. Panel (b) details the distribution of atoms per unit cell, ranging from 1 to 20 atoms.

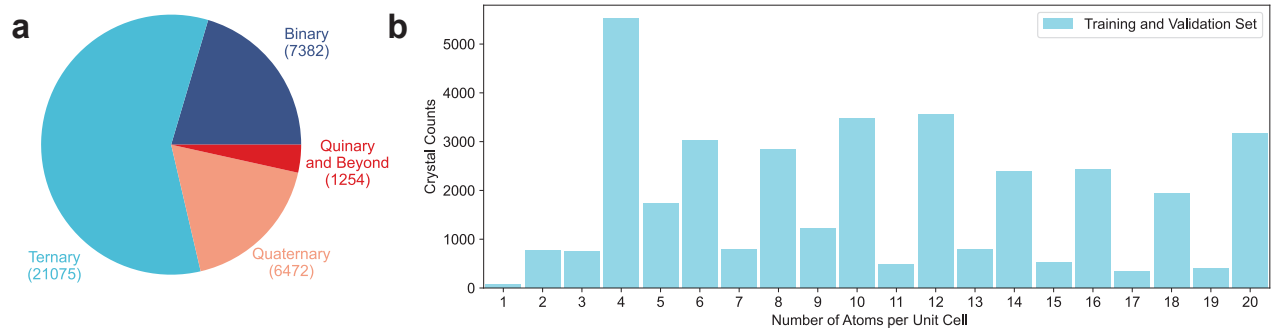

Figure S2. Structural statistics of the entire dataset, including both training and validation ones. **a.** Distribution of crystal compositions, most of which are among binary to quinary systems. **b.** Distribution of the number of atoms per unit cell.

1.2 List of chemical valence

Below, we list the electronic configuration embeddings for the chemical formula transformer in Fig. S3-Fig. S6 for ions and Fig. S7 for elements in alloys. The physical meaning of each entry is labeled at the y-axis, where Z is the atomic number (nuclear charge), n_{1s} is the number of 1s electrons in its ground-state electronic configuration, and n_{vs} is the number of s electrons in valence shells. The numerical value of each entry is normalized by the maximum value in our scope of the periodic table.

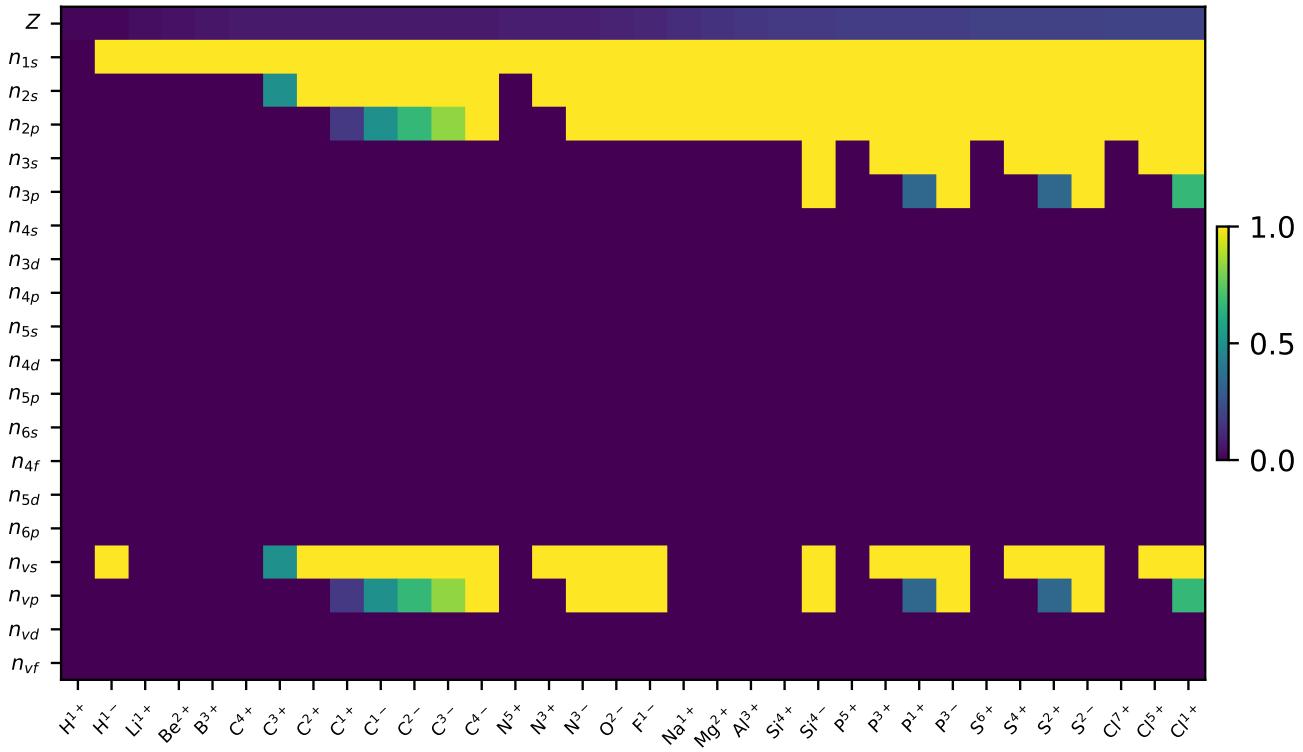

Figure S3. Electronic configuration embeddings of ions in the first–third rows of the periodic table (H-Cl). The *x* axis lists the atom types and the *y* axis stands for each entry in the embedding vector including the nuclear charge and the electronic shell occupancy under the aufbau principle.

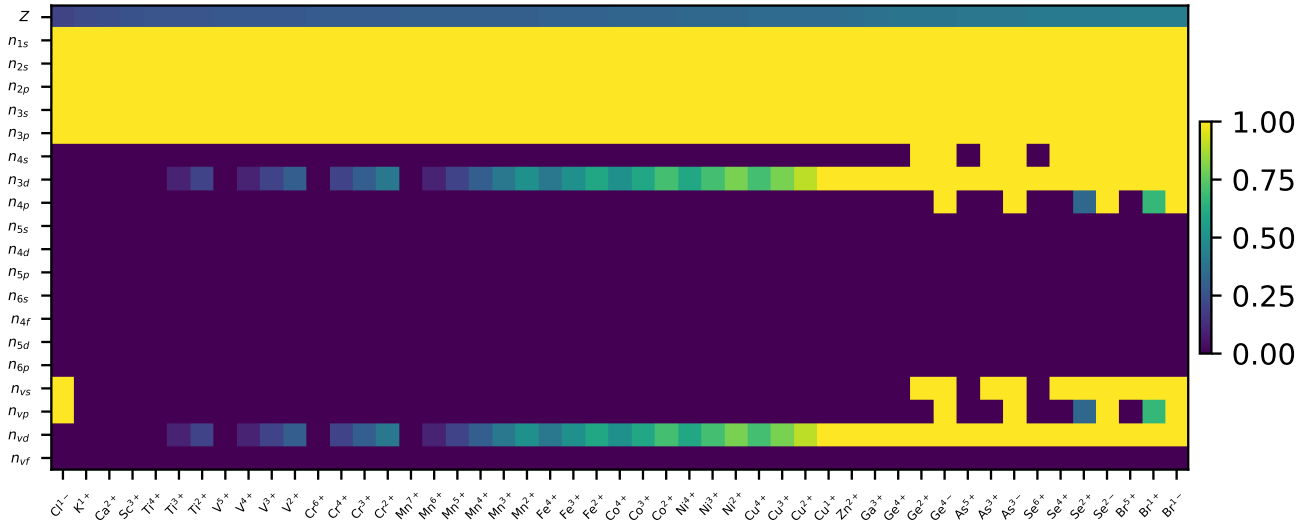

Figure S4. Electronic configuration embeddings of ions in the third-fourth rows of the periodic table (Cl-Br). The *x* axis lists the atom types and the *y* axis stands for each entry in the embedding vector including the nuclear charge and the electronic shell occupancy under the aufbau principle.

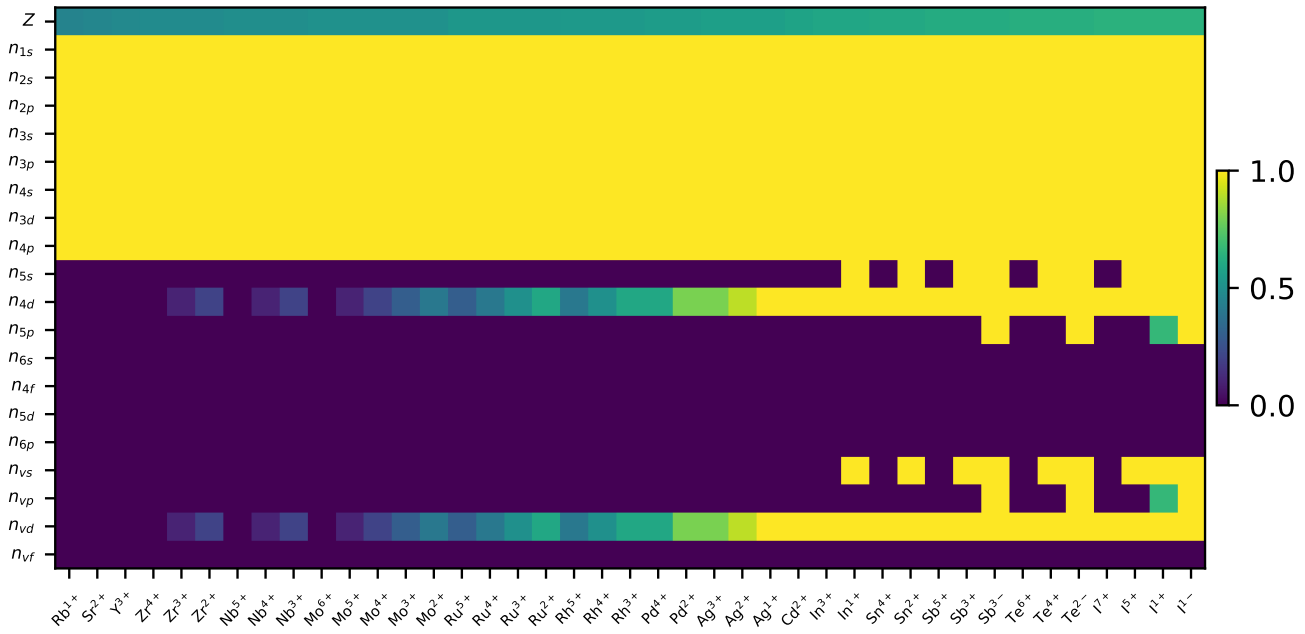

Figure S5. Electronic configuration embeddings of ions in the fifth row of the periodic table (Rb-I). The *x* axis lists the atom types and the *y* axis stands for each entry in the embedding vector including the nuclear charge and the electronic shell occupancy under the aufbau principle.

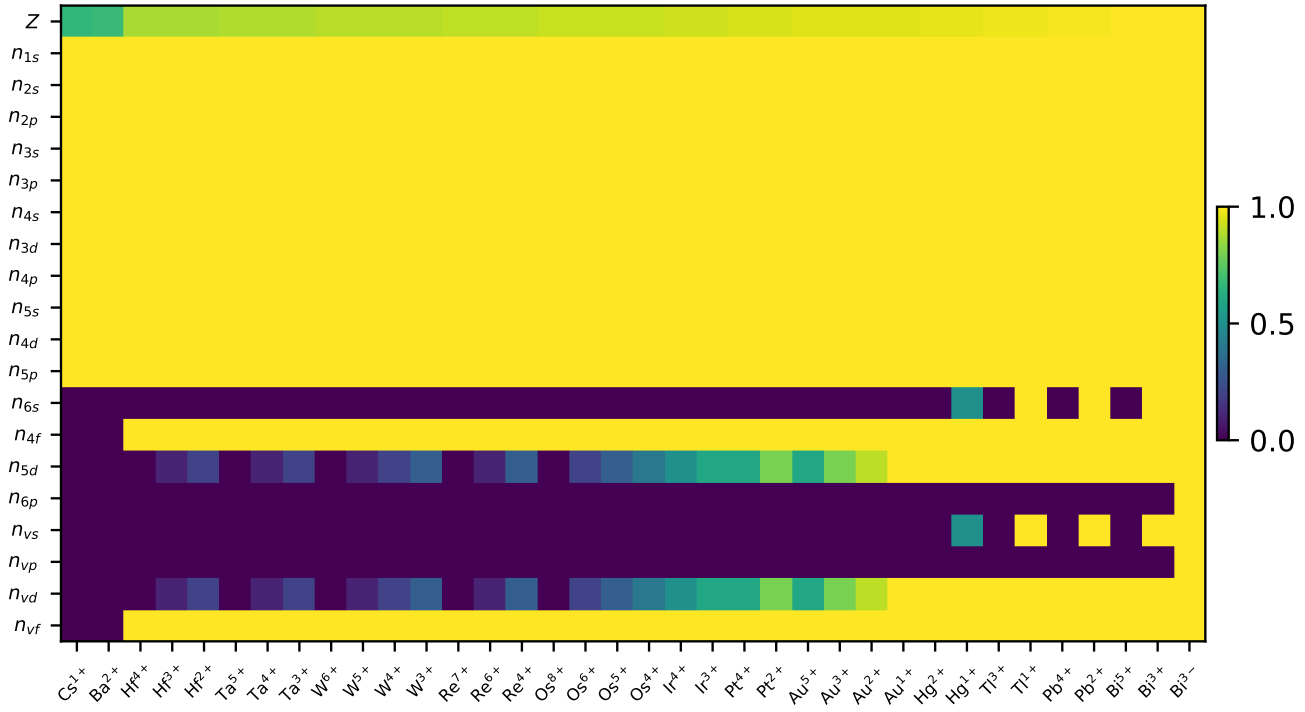

Figure S6. Electronic configuration embeddings of ions in the sixth row of the periodic table (Cs-Bi). The *x* axis lists the atom types and the *y* axis stands for each entry in the embedding vector including the nuclear charge and the electronic shell occupancy under the aufbau principle.

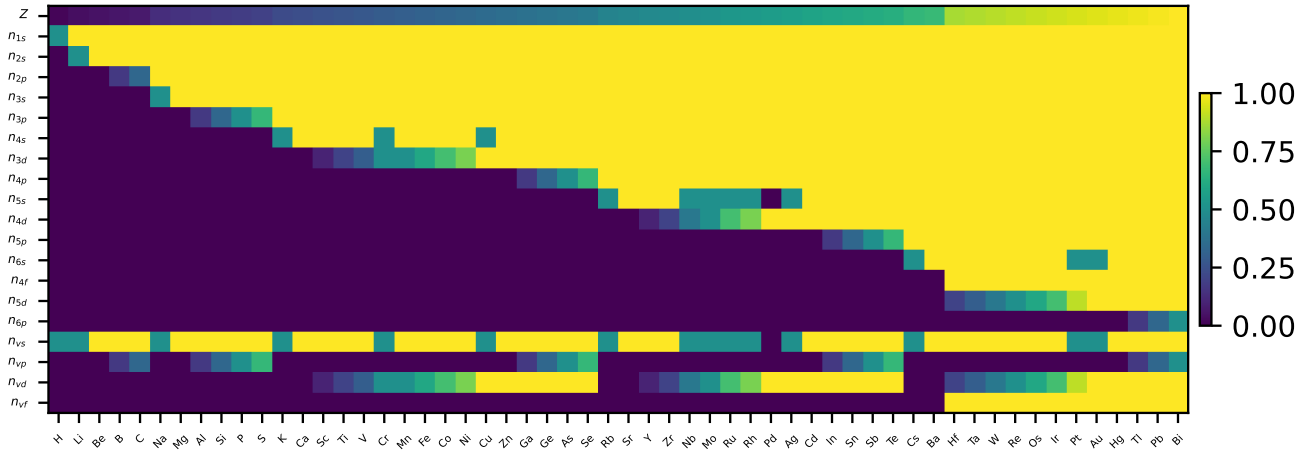

Figure S7. Electronic configuration embeddings of elements in alloys. Including nonmetallic elements like H, B, C and common metallic elements. The x axis lists the atom types and the y axis stands for each entry in the embedding vector including the nuclear charge and the electronic shell occupancy under the aufbau principle.

2 Architecture for CrysVCD

2.1 Elemental language model

The detailed architecture of our chemical formula transformer is shown in Fig. S8, whose implementation is provided by the HuggingFace version of GPT-2. Due to the small scale of training data, we cut down the size of the default options to reduce the computational cost and avoid potential overfitting. The dimension of the token embedding `n_emb` is set to 128. The `n_position`, corresponding to the maximum sequence length, is set to 10 based on the statistics of our chemical formula dataset. The number of transformer layers `n_layer` is set to 3, and the multi-head self-attention mechanism uses `n_head`=4 heads. We adopt HuggingFace's default values for all other architectural hyperparameters. Both pretraining and fine-tuning use a learning rate of 10^{-3} in the Adam optimizer, with a fixed epoch number of 100.

2.2 Diffusion module for crystal structure prediction

Our workflow allows for any crystal structure prediction algorithm, including genetic algorithms, particle-swarm algorithms, deep generative models like variational auto-encoders, large language models, as well as the diffusion model, to build the structure from a chemically valid chemical formula generated by the chemical formula transformer. As a showcase, we utilize the code from the official implementation of the DiffCSP model on Github as the crystal structure prediction module in our material design workflow, and set the same architecture and training hyperparameters as the default of the MP-20 task in the DiffCSP codebase. We keep the pretrained model with the best loss on the validation set in 1000 epochs at a learning rate of 10^{-3} in the Adam optimizer. When fine-tuning on phonon stability or energy above the hull, the learning rate is reduced to 5×10^{-4}, and further reduced to 10^{-4} during thermal conductivity or dielectric constant fine-tuning. Our evaluation is also based on the fine-tuned model with the best loss on the validation set in 1000 epochs.

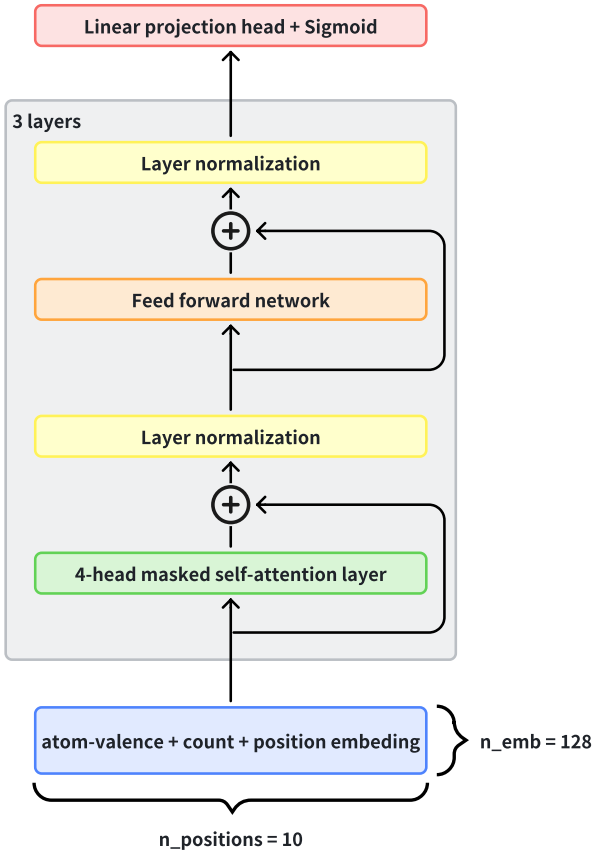

Figure S8. The model architecture of the chemical formula transformer used in CrysVCD.

3 Additional results for CrysVCD

Below, we show additional results related to CrysVCD in the conditional generation tasks, including computational overhead, stability evaluation, and surrogate model prediction on the dielectric constant.

3.1 Computational overhead of the chemical formula transformer on the material generation

In Fig. S9, we test the speed of chemical formula generation with the chemical formula transformer and the crystal structure generation with the crystal diffusion model. For the training stage, models are trained on the training partition of the MP-20-valence dataset. Because a unified crystal diffusion model is trained for both alloys and ionic compounds, they share the same training speed in the figure. For the inference stage, the chemical formula transformer samples chemical formulas one by one until 100 distinct species are obtained. They are then fed into the unconditioned crystal transformer model to generate one structure for each species. The results indicate that the chemical formula generation as an additional prefix of the material generation task in our framework doesn't significantly increase the computation time.

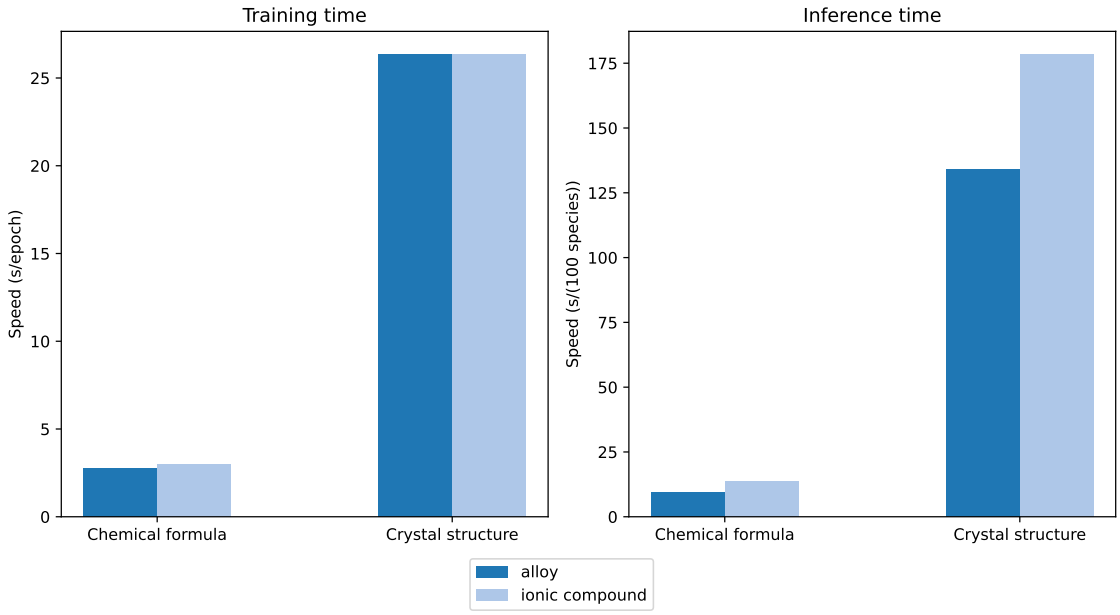

Figure S9. Comparison between the speed of chemical formula and crystal structure generation.

3.2 Validation of energy above hull using MLIP

To assess the reliability of the MatterSim MLIP in evaluating the quality of generated crystals, we first benchmark its predictions on known structures from the MP-20 training dataset. Since all entries in this dataset are known to have energy above hull values below 0.1 eV (as computed via DFT), this serves as a consistency check for the MLIP. As shown in Fig. S11, the MatterSim-evaluated E_{hull} values cluster tightly below the 0.1 eV threshold, demonstrating strong agreement with DFT results. This validates MatterSim's ability to serve as a proxy for rapid stability assessment and supports its use in evaluating newly generated crystals.

3.3 Validation of thermal conductivity using MLIP

In this section, we benchmark the fast thermal conductivity evaluation in our CrysVCD inverse design workflow using a MatterSim MLIP model. While MatterSim MLIP has been validated on the Matbench-Discovery benchmark[1], the present test targets a biased subset: candidate materials with human-guided high thermal conductivity, typically containing light elements. This bias naturally shifts the distribution towards higher κ, where errors could tend to grow. For this dataset, the symmetric relative mean error (κ_{RMSE}) in total thermal conductivity predictions is 0.86,

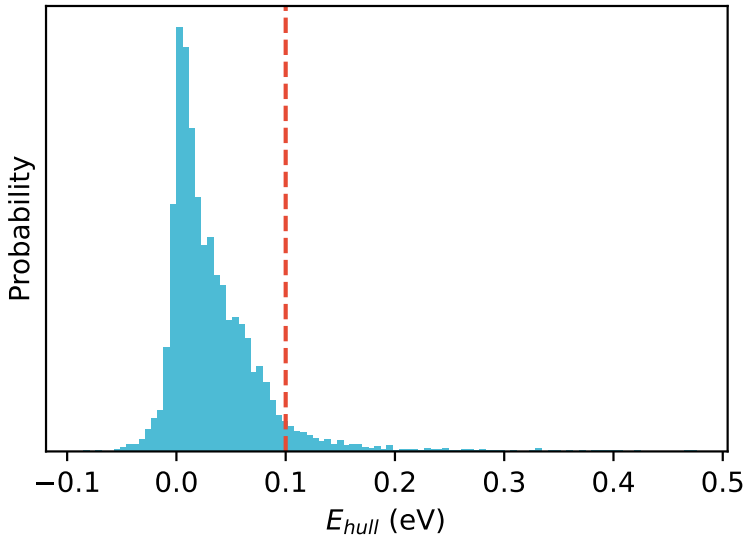

Figure S10. Distribution of energy above hull for crystals in MP-20 dataset evaluated using machine learning interatomic potential. MatterSim is used for on-the-fly evaluation, with the red dashed line marking the 0.1 eV threshold commonly used for assessing thermodynamic stability.

larger than the value (0.58) reported in MatterSim for general materials. Nevertheless, the model retains a reasonable correlation with DFT reference values and demonstrates practical reliability for rapid screening in high-κ design scenarios.

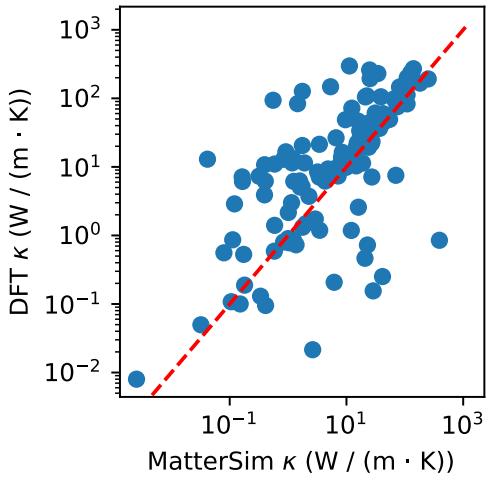

Figure S11. Comparison between the thermal conductivity κ computed by DFT and MatterSim MLIP. The difference lies in the supplement of force constants, either calculated by DFT or evaluated by MLIP, and both results are subsequently passed on to phono3py for thermal transport calculation.

3.4 Surrogate GNNOpt model for dielectric constant prediction

To enable efficient inverse design of high-κ dielectric materials, we employ an E(3)-equivariant graph neural network (e3nn) as a surrogate model to predict the static dielectric constant ε, based on the implementation of Hung et al.[2].

Given the high computational cost of direct DFT evaluation, this model allows for rapid screening of generated candidate materials. As shown in Fig. S12, the model demonstrates good agreement with ground-truth values on the training set and maintains reasonable accuracy on the testing set. The predictive performance justifies the suitability of the e3nn model for evaluating dielectric properties within our generative design framework.

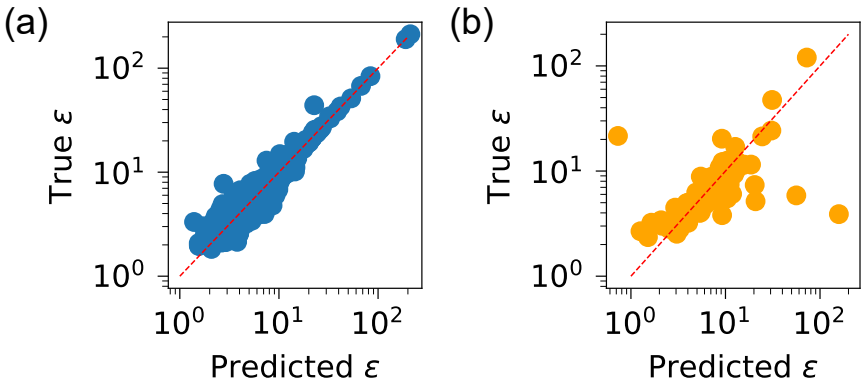

Figure S12. Predicted vs. true static dielectric constant (ε) using an E(3)-equivariant graph neural network (e3nn) model. **(a)** Training set and **(b)** testing set results are shown with a red dashed line representing the ideal parity line.

References

1. Riebesell, J. *et al.* Matbench Discovery–a framework to evaluate machine learning crystal stability predictions. *arXiv preprint arXiv:2308.14920* (2023).
2. 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).