Structural Constraint Integration in Generative Model for Discovery of Quantum Material Candidates

Ryotaro Okabe[1,2,*], Mouyang Cheng[1,3,4], Abhijatmedhi Chotrattanapituk[1,5], Manasi Mandal[1,6], Kiran Mak[1,3], Denisse Crdova Carrizales[1,6], Nguyen Tuan Hung[1,6,7], Xiang Fu[5], Bowen Han[8], Yao Wang[9], Weiwei Xie[10,], Robert J. Cava[11], Tommi S. Jaakkola[5], Yongqiang Cheng[8,***], and Mingda Li[1,6,****]**

[1]Quantum Measurement Group, Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA
[3]Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[4]Center for Computational Science & Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[5]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, USA
[6]Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA
[7]Frontier Research Institute for Interdisciplinary Sciences, Tohoku University, Sendai 980-8578, Japan
[8]Chemical Spectroscopy Group, Spectroscopy Section, Neutron Scattering Division Oak Ridge National Laboratory, Oak Ridge, TN, USA
[9]Department of Chemistry, Emory University, Atlanta, Georgia, USA
[10]Department of Chemistry, Michigan State University, East Lansing, MI, USA
[11]Department of Chemistry, Princeton University, Princeton, NJ, USA
[*]e-mail: rokabe@mit.edu
[**]e-mail: xieweiwe@msu.edu
[***]e-mail: chengy@ornl.gov
[****]e-mail: mingda@mit.edu

ABSTRACT

Billions of organic molecules have been computationally discovered, but relatively few functional inorganic materials with practical applications have been found, given the lack of training databases and the relative complexity of inorganic materials. Recent advancements in machine-learning-based generative models, particularly diffusion models, show great promise for generating new, stable materials. However, integrating geometric patterns–which often exhibit quantum phenomena–into materials generation remains challenging. Here, we propose a generative diffusion model that allows us to discover materials with distinct physical properties from structural motifs, such as honeycomb or kagome. Our approach, Structural Constraint Integration in the GENerative model (SCIGEN), can modify any trained generative diffusion model by introducing the desired constraints while maintaining the validity of the outputs. We mathematically prove that SCIGEN performs conditional sampling from the original distribution, which is crucial for generating constrained materials with enough stability. This framework efficiently explores quantum material candidates by increasing the likelihood of discovering magnetic ordering and unique electronic behaviors. We generate ten million compounds using Archimedean lattices and Lieb lattices as geometric constraints, and over 10% of the compounds survive multi-staged stability pre-screening. Our high-throughput density functional theory (DFT) calculations on 26,000 down-selected compounds reveal that over 95% achieve DFT convergence. Among these compounds, more than 53% are stable under structural optimization, and a separate graph neural network-based classifier identifies magnetism in 41% of the optimized structures. Furthermore, we synthesize and characterize two predicted materials, $TiPd_{0.22}Bi_{0.88}$ and $Ti_{0.5}Pd_{1.5}Sb$ inspired by SCIGEN's prediction, which exhibit paramagnetic and diamagnetic behavior, respectively. Since the properties of quantum materials are closely related to geometric lattices, our results indicate that SCIGEN provides a general framework to generate quantum materials candidates.

Introduction

Structure-property material relationships are instrumental in understanding functional quantum phenomena. Symmetry and geometry are critical indicators that profoundly influence materials' properties based on the structures. For example, materials with inversion symmetry can lead to topological crystalline insulators[1], whereas breaking inversion symmetry can result in a variety of phenomena such as Rashba spin-orbit coupling[2], ferroelectricity[3], second harmonic generation[4], and topological

Weyl semimetals[5]. A material's geometric pattern is closely linked to its electronic states and magnetic orderings. The square lattice serves as a platform for high-temperature cuprate superconductors[6], while triangular, honeycomb, and kagome lattices can host exotic magnetic states like quantum spin liquids[7,8]. Additionally, kagome and Lieb lattices can support electronic flat bands[9,10] with the technological implication of these lattices replacing rare-earth elements[11]. Also, porous structures like zeolite lattices are used in catalysis[12]. However, designing stable materials with the desired quantum properties can be nontrivial. For example, only a dozen quantum spin liquid candidates have been identified after a decade of research[13].

Machine-learning (ML) based material generators have led to a paradigm shift in material design. Diffusion models, like CDVAE, UniMat, and DiffCSP[14–16], and graph neural network models like GNoME[17] have shown great promise in identifying stable structures to generate millions of materials. However, most ML-based generators sample from the statistical distribution of the training dataset, generating materials that reflect the structures and properties commonly found in the database. This approach makes it challenging to design materials with specific constraints that are underrepresented or absent in the dataset. Although there have been some developments to incorporate crystallographic space groups in materials generation[18–20], integrating geometric patterns into generation algorithms for functional materials remains challenging. Space group symmetry can provide a material's structural framework, but does not always capture the detailed local geometric patterns critical for certain quantum phenomena. For example, monoclinic stacking variants of hexagonal symmetry layers may not be fully described by the space group symmetry alone. In systems like frustrated magnets with kagome lattices, the geometric pattern is more critical in supporting exotic quantum phenomena than the overall space group. Therefore, the key to exploring new quantum phenomena is the development of an ML-based generator capable of producing materials constrained by particular geometric patterns.

This work presents SCIGEN: Structural Constraint Integration in the GENerative model. SCIGEN is a scheme that any pre-trained generative diffusion model can utilize to incorporate geometric patterns and symmetry constraints to generate materials without retraining or fine-tuning. Starting from the target constraints, SCIGEN diffuses a random constrained structure over multiple time steps. The constrained structures mask the denoised structure before each diffusion step, creating an inductive bias that directs the generation process toward producing outputs that adhere to the constraints. The iterative usage of mask at every denoising step converges all atoms belonging to the mask into the sites of the imposed geometry deterministically, making SCIGEN distinct from the conventional approach of conditional generation. SCIGEN effectively performs conditional generation with respect to the probability distribution of the base model. This indicates that the constraint imposed by SCIGEN will preserve the integrity of the base generative model, including but not limited to the stability of generated materials and chemical element distribution. We apply SCIGEN as a conditioning scheme to DiffCSP[16] to validate the preservation of the base generative model's capabilities. We generate materials constrained by Archimedean lattices (ALs)[21,22], which are a collection of 2D lattice tilings of square, triangular, honeycomb, kagome, and a few other geometric patterns–all rich harbors of exotic quantum phenomena. We generate a total of 10.07 million materials with ALs. After a four-stage stability pre-screening, over 1.01 million materials survived. We compute structure relaxation on a subset of 26,000 materials with high-throughput density functional theory (DFT), determining that 95% of the materials complete the calculation and more than 53% reach the energy minimum within 150 steps of structural optimization. Based on a graph neural network-based classifier, we identify magnetism in 41% of the DFT-optimized structures. From the down-selected dataset of materials, we synthesize two new materials, TiPdBi and TiPbSb–which, to our knowledge, are previously unreported compounds. Although the competing phases during synthesis can cause a slight deviation from the predicted products, the experimental results remain closely aligned with our calculations. Magnetization measurements reveal distinct magnetic behaviors in the compounds: TiPdBi and TiPbSb show paramagnetic and diamagnetic properties, respectively. Since SCIGEN requires no extra training apart from the underlying generative model, it offers a flexible and generally applicable conditioning scheme of materials generation with constraints from symmetry and geometric patterns.

Results

Structural Constraint Integration in the Generative Model

Figure 1 presents the schematic overview of SCIGEN. The goal of crystal structure generation is to find periodic crystals $\boldsymbol{M}$, which can be represented by the three components: the lattice matrix containing three basis vectors $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3] \in \mathbb{R}^{3\times3}$, the fractional coordinates $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N] \in [0,1)^{3\times N}$, and one-hot representations of atom types $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N] \in [0,1]^{h\times N}$. Our methods impose geometric constraints on $\boldsymbol{L}$, $\boldsymbol{F}$, and $\boldsymbol{A}$ respectively in diffusion-based material generation. Figure 1a illustrates notable geometries, including triangular, honeycomb, and kagome lattices. Following the guideline in Figure 1b, we initiate the constrained structures with an AL composed of magnetic atoms. Figure 1c explains the algorithm of the generative process integrating the constrained components. The initialized structure is subjected to a diffusion process by adding noise over T-steps denoted as $\boldsymbol{M}_t^c$ where $t \in [1..T]$ ($T = 1000$ by choice), providing the pre-defined pathway of denoising process for the constrained components. An unconstrained structure is initiated as a completely noisy structure $\boldsymbol{M}_T^u$. Both $\boldsymbol{M}_T^c$ and $\boldsymbol{M}_T^u$ are integrated to form $\boldsymbol{M}_T$. This composite structure $\boldsymbol{M}_T$ is then denoised to retrieve $\boldsymbol{M}_{T-1}^u$. SCIGEN repeats this

process through all steps; it merges $\boldsymbol{M}_t^c$ and $\boldsymbol{M}_t^u$ to get $\boldsymbol{M}_t$ and then predicts $\boldsymbol{M}_{t-1}^u$. This iteration optimizes the final material structure $\boldsymbol{M}_0$ by guiding a subset of atoms to form AL planar structures. A more detailed explanation of the implementations is shown in Supplementary Information 3, and the detailed training process is shown in Supplementary Information 5. Additionally, Supplementary Information 4 provides a mathematical proof that the constraint imposed by SCIGEN does not alter the distribution of the underlying unconstrained generative model. This is further verified computationally in Supplementary Information 6 through analyses of element histograms, atomic co-occurrence matrices, and the oxygen-cell size relationship in the generated crystals. Together, these results indicate that SCIGEN can faithfully impose constraints without disrupting the original material distribution. Following the generation of a large set of material candidates, we evaluate their stability through a four-stage pre-screening process. The pre-screening involves applying chemical rules such as charge neutrality and the volume of atoms occupying the lattice unit cell, along with auxiliary neural networks that predict stability based on the energy above the convex hull (E_{hull}) values. After that, we employ high-throughput DFT to relax structures and identify potentially stable candidates.

Materials Generation with Archimedean Lattice Constraints

Figures 2a-c present the results of materials generation constrained by three primary AL types: (a) triangular, (b) honeycomb, and (c) kagome. The AL structures are formed as intended by the SCIGEN algorithm, which guides the arrangement of the constrained atoms to result in pre-defined geometric patterns. The constraints do not explicitly dictate the positions of the unconstrained atoms. However, their placement emerges during the generation process, influenced by spatial and bonding considerations. For triangular lattices (Figure 2a), unconstrained atoms tend to occupy sites that bridge three magnetic atoms, forming an equilateral triangle. For honeycomb lattices (Figure 2b), unconstrained atoms frequently emerge at the center of the hexagons formed by magnetic atoms in the same plane. In kagome materials (Figure 2c), unconstrained atoms bridge the equilateral triangles and the hexagons of kagome lattice layers. If the space within the polygons is too narrow relative to the atomic radii, these unconstrained atoms may shift outside the AL plane. On the other hand, large polygons, such as hexagons, can accommodate the unconstrained atoms within the same plane.

To generate constrained material structures with a higher likelihood of stability, we develop a scheme for sampling initial conditions. We analyze the ratio of stable outputs, defined as the survival ratio after the multi-staged pre-screening processes. First, we sample the number of atoms per unit cell (N) from a uniform distribution to identify which N values are more likely to pass the stability pre-screening. This results in a probability distribution p_N values based on their pre-screened stability. We then use this probability distribution p_N to sample N to initialize the large-scale generative process. Figure 2d shows the sampling profile of N, which covers all of the ten ordinary magnetic atoms as the vertices of AL structures. For triangular lattice materials, smaller N values show higher success rates, whereas larger N values are favored for honeycomb and kagome. This result is reasonable since the AL type is directly linked to the unit cell size, which is a linear function of bond lengths for each class of AL types. For triangular lattice, the lattice parameters l_1 and l_2 are the same as the bond length of the neighbor node, while for honeycomb and kagome lattices, the lattice parameters are $\sqrt{3}$ and 2 times of the bond length, respectively. In the case where many atoms are packed into the unit cell of a small cross-section of the AL, the cell needs to be "tall", i.e., l_3 needs to be larger with respect to l_1 and l_2. Next, we survey which magnetic atoms are suitable as the vertices of ALs. Figure 2e presents the number of stable materials after the prescreening with respect to magnetic atom types, which we analyze from the set of 3000 generated materials for each lattice type and each magnetic atom. Despite variations, all magnetic atoms are shown to be able to form AL structures. Therefore, we sample atom types for AL vertices with equal probabilities for large-scale materials generation and database construction. Methods section and Supplementary Information 2 describe the sampling schemes for the initialization conditions in detail.

Our exploration of materials with geometrical constraints does not end with the three primary types of ALs but can also apply to other geometrical patterns. In contrast to the common types of triangular, honeycomb, and kagome lattices, magnetic systems known to fit in other ALs are extremely rare. Figure 3 showcases $3\times3\times1$ supercells of the generated materials with seven other types of Archimedean lattices: Square, Elongated triangular, Snub square, Truncated square, Small rhombitrihexagonal, Snub hexagonal, and Truncated hexagonal. One type of AL lattice, the Great rhombitrihexagonal, is not presented due to the challenge of generating stable materials. The unconstrained atoms within these materials often play a critical role in the overall stability of the structures. They tend to bridge gaps between structured lattice layers, either by sitting at the center of polygons on the same plane or contacting all vertices of the polygon structures, effectively stabilizing the AL layers. This bridging is not just a passive consequence of the material generation process but actively contributes to the mechanical and thermal stability of materials[23]. Interestingly, these unconstrained atoms frequently organize into recognizable Archimedean patterns even when not explicitly constrained to form specific lattice structures. This trend could suggest an inherent preference or stability in the configurations of AL whose vertices are equivalent with respect to the local coordinates.

Materials Generation with Lieb-like Lattice Structures

The Lieb lattice is a variation of the square AL with additional atoms located at the centers of each edge of the squares, as visualized in Figure 4a. Each unit cell of the Lieb lattice contains three atoms. The geometry of the Lieb lattice can lead to magnetic frustration when interacting spins are placed at each lattice site. In addition to magnetic frustration, which is potentially a feature among AL structures, the Lieb lattice stands out due to the possibility of possessing unique electronic properties represented by flat bands[10]. Contrary to localized atomic orbitals, the flat bands formed from the Lieb lattice originate from the destructive quantum interference effect, which quenches the kinetic energy. This may lead to interesting physical phenomena such as enhanced electron correlation and high-temperature superconductivity[24,25]. Also, recent research has shown that the Lieb lattice can exhibit non-trivial topological properties when subjected to various perturbations[26]. However, the Lieb lattice has mainly been achieved in artificial systems like photonic crystals[27,28], and atomic solids that can host the Lieb lattice are extremely rare.

In this work, we also focus on the Lieb-like lattice where magnetic atoms sit on the Lieb lattice. Lieb-like crystals include variants of the Lieb lattice (Figure 4b), such as additional atoms positioned within the lattice framework. Figure 4c shows the generated materials with Lieb lattice crystal structures and their calculated band structures. In these generated materials, magnetic atoms such as terbium (Tb) and dysprosium (Dy) are strategically positioned at the nodes of the Lieb lattice. Following structural relaxation through DFT calculations, the integrity of the Lieb lattice architecture is maintained, and the structures exhibit the anticipated flat-band characteristics close to the Fermi level. These outcomes demonstrate SCIGEN's ability to generate new, stable materials with exotic geometric patterns, even when few known materials fit the desired geometric pattern. To push SCIGEN toward realistic oxide analogues of cuprate and nickelate superconductors[29,30], we introduce an adaptive-unmasking (AU) strategy. The goal of AU is to partially relax the full structural constraint of SCIGEN while enhancing the flexibility and stability of generated crystals. In AU, the rigid in-painting mask that fixes the Lieb motif is gradually relaxed after an early fraction of the denoising schedule, allowing oxygen and transition-metal sites to seek chemically reasonable bond lengths while the overall Lieb topology is retained (Figure 4b). Figure 4d displays representative AU-generated heteroatomic Lieb lattices such as $Ti_3CoCu_2O_6$ and Li_3MnO_4. Supplementary Information 11 details the AU protocol, which covers the time-dependent release of constraints and the assignment of multi-species sites.

Database of the Materials with Archimedean Lattice

As detailed in Supplementary Information 12, we generate an AL materials database using SCIGEN. The database has three components: the total 10.07 million materials generated by the SCIGEN model, the 1.01 million materials that survived four stages of stability pre-screening processes, and 24,743 out of 26,000 sub-sampled materials in which DFT calculations successfully converge. To identify magnetic structures, we train a GNN-based classification model based on Merker et al.[31], achieving an accuracy of 92%. 41% of the converged generated structures are identified as magnetic materials. For specific elements (Fe, Gd, Nd), the predicted ratios–representing the fraction of magnetic materials among structures containing each element–align well with their ground truth (GT) values. 98.53% of the materials containing Fe atoms are predicted to be magnetic, closely aligning with the ground truth ratio of 93.63%. Similarly, 100% of the materials containing Gd atoms are predicted as magnetic, matching the ground truth. For Nd-containing materials, 19.10% are predicted as magnetic, compared to a ground truth ratio of 26.61%. This agreement supports the intuition that Gd consistently shows high magnetic properties with its strongly magnetized orbitals, while Nd, with weaker magnetization, presents lower ratios. Further details can be found in Supplementary Information 8.

By systematically extending our exploration to encompass a broader range of geometric patterns–including ALs such as triangular, honeycomb, and kagome lattices, as well as Lieb lattices–we explore new exotic magnetic orderings, discover porous structures beyond zeolites, and investigate electronic flat band structures, among other possibilities.

Experimental Efforts Guided by SCIGEN-Generated Magnetic Materials

Our SCIGEN framework shows promise in generating stable quantum material candidates, particularly with magnetic properties. Magnetism remains a central focus in quantum material discovery because many generated lattice structures, such as Archimedean and Lieb lattices, can host exotic magnetic states including frustration and emergent phenomena. Experimental exploration is critical to assess whether ML-predicted structures or their competing phases[32,33] can be synthesized. Successful synthesis may reveal behaviors beyond current theoretical understanding. Given the importance of Ti atoms for both frustrated magnetism and functional devices, we select two previously unreported SCIGEN-predicted compounds, TiPdBi and TiPdSb, for synthesis and structural and magnetic characterization.

The experimentally realized compounds, $TiPd_{0.22}Bi_{0.88}$ and $Ti_{0.5}Pd_{1.5}Sb$, differ in stoichiometry and crystal symmetry from their SCIGEN-predicted counterparts. These discrepancies likely result from disorder, site occupancy effects, and the influence of thermodynamically preferred phases under experimental conditionsfactors not explicitly accounted for by the generative model, which is trained on zero-Kelvin DFT data. Rather than serving as direct validations, these experimental efforts

demonstrate the capacity of SCIGEN to propose chemically plausible motifs that present evidence for further exploration, even if not realized in their idealized forms. Details are provided in Supplementary Information 10.

Conclusions

This work presents SCIGEN, a generative model that discovers quantum material candidates that adhere to geometric constraints. SCIGEN introduces a masking strategy that integrates geometric patterns directly into the diffusion process, guiding material generation toward structures consistent with specified constraints. This approach allows SCIGEN to perform conditional generation without retraining the base diffusion model, enabling efficient exploration of diverse lattice symmetries and atomic arrangements. Our method incorporates Archimedean lattice (AL) constraints into the material generation process, guiding SCIGEN to produce crystal structures with specific geometric patterns. This conditional scheme increases the likelihood of discovering quantum materials with unique properties, including frustrated magnetism and flat band structures. These materials have been validated through DFT to ensure that their relaxed structures are consistent with the machine-learning generations. 41% of the DFT-optimized structures are identified as magnetic materials utilizing a GNN model. Our preliminary experiments suggest that SCIGEN possesses a level of potential to contribute to the discovery of quantum materials.

To further enhance SCIGENs capabilities, future work could explore further geometry-related constraints, such as bonding types, coordination numbers, short-range orderings, and symmetries like point groups and space groups. A preliminary analysis on the space group distribution of SCIGEN-generated materials is shown in Supplementary Information 9. Integrating more complex constraints, such as defects and magnetic interactions, will expand SCIGEN's versatility. Conditioning the generation process with targeted functionalities–such as specific thermal, electronic, and optoelectronic properties or environmental impacts–could further facilitate the material design with tailored performance.

SCIGEN represents a general machine learning-based framework for discovering quantum material candidates. This innovation opens previously inaccessible regions of the materials design space, laying the foundation for uncovering structure-property relationships. It marks the beginning of a broader exploration into geometry-driven materials discovery, paving the way for quantum material breakthroughs.

Methods

Initialization of the Archimedean Lattices

Here, we describe the workflow to initialize the materials generation process related to both AL and the entire structure of the crystal for the diffusion model. This initialization process involves a few steps: the choice of AL, the atom types, and the total number of atoms per unit cell. We summarize the initialization process in Figure 2b, where a more detailed scheme can be found in Supplementary Information 2.

First, we assign the required geometric domain condition to the crystals. In SCIGEN, we specify one of the AL structures, such as triangular, honeycomb, or kagome lattice, as a geometric domain condition. Each type of AL requires the number of vertices per unit cell and the size of the unit cell. Supplementary Information 1 presents the geometric patterns and the preliminary profiles of all AL and Lieb lattices.

Second, we choose the constrained atom type $\mathscr{A}^c$ placed on the vertices of the AL structure assigned above. We specify ten types of ordinary magnetic atoms (Mn, Fe, Co, Ni, Ru, Nd, Gd, Tb, Dy, Yb) to generate candidate materials that may host geometrically frustrated quantum magnetism on the vertices. The atom types are chosen independently from the AL choice above.

Third, we sample the constrained magnetic bond lengths d^c, or the distances between the nearest-neighbor magnetic atoms, forming the ALs. For each magnetic atom type $\mathscr{A}^c$, we generate the profile of the bond lengths by sampling the nearest-neighbor distances between the corresponding atoms in the MP-20 dataset[34,35] using CrystalNN[36]. We impose a lower bound on the bond lengths based on the metallic radii[37] for each atom type, ensuring the nearest-neighbor distances do not become unrealistically short. The bond length distribution for each magnetic atom type $\mathscr{A}^c$, $p_{d^c}(\mathscr{A}^c)$, is presented in Supplementary Information 2.

Finally, we sample the total number of atoms per unit cell N. Each of the ALs has the distribution of the preferred N values with better stability. We generate p_N, the stable materials probability distribution of N. We can sample N from p_N as the sampling profile of N for each AL type. The sampling profile of both p_N and $p_N(\mathscr{A}^c)$, which is the distribution of each magnetic atom type $\mathscr{A}^c$, is displayed in Supplementary Information 2.

To impose Archimedean lattice as the constraints, we organize masks $\boldsymbol{m} = (\boldsymbol{m^L}, \boldsymbol{m^F}, \boldsymbol{m^A})$ which give constraints to the lattice $\boldsymbol{L}$, fractional coordinates $\boldsymbol{F}$, and atom types $\boldsymbol{A}$ respectively. $\boldsymbol{m^L}$ is equal to 1 for the two lattice basis vectors $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$, defining the unit cell of AL layer plane. $\boldsymbol{m^L}$ is equal to 0 for $\boldsymbol{l}_3$, as we let the diffusion model generate $\boldsymbol{l}_3$ without explicit constraints. We assign $\boldsymbol{m^F}$ is equal to 1 for the i-th atoms ($i \in [1, N^c]$) to guide their placement at the vertex positions of AL layers. The same rule applies for $\boldsymbol{m^A}$ so that the atoms at AL vertices result in the magnetic atom types $\mathscr{A}^c$.

Integration of constrained and unconstrained components to guide materials generation

We design SCIGEN as a generic framework applicable to any diffusion model as a base model. Without loss of generality, we let the pre-trained base model represent a periodic structure as $\boldsymbol{M}_0$ with T diffusion steps; a sampling probability prior P_T; diffusion inference model q, which is normally chosen to map the materials distribution of the training dataset to P_T; and denoising generative model p which needs to be trained. The diffusion inference model q works by iteratively imposing noise to the input structure, $\boldsymbol{M}_0$. The inference probability of most diffusion models, including the base model used in SCIGEN, is a Markov process, i.e., the probability of diffusing $\boldsymbol{M}_0$ for t steps to $\boldsymbol{M}_t$ can be written as

$$P(\boldsymbol{M}_t|\boldsymbol{M}_0) = q_{0,t}(\boldsymbol{M}_t|\boldsymbol{M}_0) = \prod_{s=1}^{t} q_{s-1,s}(\boldsymbol{M}_s|\boldsymbol{M}_{s-1}) \tag{1}$$

with

$$P(\boldsymbol{M}_T|\boldsymbol{M}_0) = q_{0,T}(\boldsymbol{M}_T|\boldsymbol{M}_0) \approx P_T. \tag{2}$$

A well-trained diffusion model should have a denoising generative model p that can invert the diffusion, i.e., for a denoising step from $\boldsymbol{M}_t$ to $\boldsymbol{M}_{t-1}$,

$$p_{t,t-1}(\boldsymbol{M}_{t-1}|\boldsymbol{M}_t) \approx q_{t,t-1}(\boldsymbol{M}_{t-1}|\boldsymbol{M}_t). \tag{3}$$

Here, the subscripts of p and q indicate the initial and final time steps that the models are applied to, e.g., q_{t_1,t_2} is the diffusion inference from time step t_1 to t_2. Since q is normally chosen as a simple probabilistic function, we cannot easily find its inverse. Hence, training is required for the denoising generative model. In SCIGEN, we performed this training to ensure the model could effectively reconstruct valid materials from noisy structures using DiffCSP as the base model (Supplementary Information 5).

Our approach to material design is summarized in Algorithm 1. This method integrates geometrical constraints to ensure that specific structural elements, such as lattice configurations or atomic distributions, adhere closely to predefined constraints. At the same time, the model allows flexibility in unconstrained regions, enabling exploration of other structural components. This balance between constraint and freedom is central to SCIGENs ability to generate materials with tailored characteristics. While similar masking-based approaches have been employed in image generation tasks like inpainting (e.g., RePaint method[38]), applying geometrical constraints to crystal generation introduces unique challenges. For instance, crystal structures must account for periodic boundary conditions and interatomic interactions, which require additional care when applying constraints. SCIGEN addresses these complexities by incorporating constraints directly into the diffusion process, ensuring that generated materials adhere to desired geometric patterns without compromising stability or diversity. The generation of a crystal generative model with geometrical constraints involves several key steps:

1. **Adding noise to the constraint structures:** First, we introduce noise to a target structure that is randomly initialized to satisfy the target constraints (This constrained structure can be unstable or unrealistic depending on the imposed target constraints.) We use the diffusion inference model, q, to get diffused constrained structures for each time step $t \in [1, T]$. This operation aims to create a predefined pathway for denoising the constrained structures. The unconstrained components of the crystals are guided by this predefined denoising pathway so that the unconstrained components fill the spaces given by the constrained components in the final outputs.

2. **Denoising the unconstrained structures with a base diffusion model:** Concurrently, the unconstrained parts of the structure undergo a normal denoising process with p. This process, facilitated by the base model, iteratively refines these regions by reducing the introduced noise, thereby nudging them toward physically realistic configurations.

3. **Integration of the constrained and unconstrained structures:** Each denoising step recombines the constrained and unconstrained components after processing them independently. This combination is critical as it ensures the predefined constraints are maintained while integrating seamlessly with the freely generated segments. This method preserves essential structural features while enabling the discovery of novel material configurations with tailored properties.

Algorithm 1 presents the SCIGEN sampling procedure designed to generate material structures with structural constraints. The algorithm utilizes a diffusion model to iteratively refine structures, ensuring the generated structures contain specific geometry as constraints. The procedure begins with the initialization of constrained structures $\boldsymbol{M}_0^c$, indicated with superscript c, along with the corresponding masks $\boldsymbol{m}$. The mask $\boldsymbol{m}$ indicates the constrained components in $\boldsymbol{M}_0^c$ with binary masking, i.e., assigns the value of 1 to the constrained and 0 to the unconstrained components. The final-time-step unconstrained structure $\boldsymbol{M}_T^u$, indicated with superscript u, and constrained structure $\boldsymbol{M}_T^c$ are sampled from the probability prior P_T of the base model.

We obtain the final-time-step structure $\boldsymbol{M}_T$ that contains the constrained components from $\boldsymbol{M}_T^c$, and the remaining parts from $\boldsymbol{M}_T^u$. We formulate this operation as $\boldsymbol{M}_T \leftarrow \boldsymbol{m} \odot \boldsymbol{M}_T^c + (1-\boldsymbol{m}) \odot \boldsymbol{M}_T^u$ where $\odot$ represents a component-wise multiplication. Basically, the components of $\boldsymbol{M}_T$ that got masked (values in $\boldsymbol{m}$ equal to 1) come from $\boldsymbol{M}_T^c$ while the remaining components (values in $\boldsymbol{m}$ equal to 0) come from $\boldsymbol{M}_T^u$

The iterative process begins from the final time step T and proceeds backward to 0. For each time step t, the structure $\boldsymbol{M}_t$ undergoes the denoising process, giving the distribution of the unconstrained structure $\boldsymbol{M}_{t-1}^u$ at time step $t-1$, as $p_{t,t-1}(\boldsymbol{M}_{t-1}^u|\boldsymbol{M}_t)$. Concurrently, the diffusion process gives the distribution of the constrained structure $\boldsymbol{M}_{t-1}^c$ as $q_{0,t-1}(\boldsymbol{M}_{t-1}^c|\boldsymbol{M}_0^c)$. Then, the unconstrained structure $\boldsymbol{M}_{t-1}^u$, and constrained structure $\boldsymbol{M}_{t-1}^c$ are sampled from their corresponding probability distributions. The structure $\boldsymbol{M}_t$ is updated by combining the sampled constrained and unconstrained parts using the mask $\boldsymbol{m}$ similar to the final-time-step case. The process continues iteratively until the initial time step is reached, at which point the refined structure $\boldsymbol{M}_0$ is returned. This ensures that the generated material structures respect the given constraints and exhibit realistic and viable configurations. Supplementary Information 3 provides the schematic explanation of the denoising process, as well as the integration of constrained and unconstrained components of material structures. This approach effectively integrates structural constraints into the diffusion model, enabling the generation of novel material structures that align with the AL structures as the predefined requirements. The mask $\boldsymbol{m}$ plays a crucial role in ensuring constrained and unconstrained parts are seamlessly combined during the iterative refinement process.

Algorithm 1 Structural Constraint Integration in Material Generation Procedure

Input: constrained structure $\boldsymbol{M}_0^c$, constraint mask $\boldsymbol{m}$, diffusion inference model q, denoising generative model p, number of steps T, probability prior P_T
Sample $\boldsymbol{M}_T^u \sim P_T$, $\boldsymbol{M}_T^c \sim P_T$
$\boldsymbol{M}_T \leftarrow \boldsymbol{m} \odot \boldsymbol{M}_T^c + (1-\boldsymbol{m}) \odot \boldsymbol{M}_T^u$
for $t = T, \ldots, 1$ **do**
    Sample $\boldsymbol{M}_{t-1}^c \sim q_{0,t-1}(\boldsymbol{M}_{t-1}^c|\boldsymbol{M}_0^c)$
    Sample $\boldsymbol{M}_{t-1}^u \sim p_{t,t-1}(\boldsymbol{M}_{t-1}^u|\boldsymbol{M}_t)$
    $\boldsymbol{M}_{t-1} \leftarrow \boldsymbol{m} \odot \boldsymbol{M}_{t-1}^c + (1-\boldsymbol{m}) \odot \boldsymbol{M}_{t-1}^u$
return $\boldsymbol{M}_0$.

To demonstrate the algorithm for generating AL materials, we choose DiffCSP[16] as the base model of SCIGEN without applying geometric constraints. In DiffCSP, the structure representation got divided into three components $\boldsymbol{M} = (\boldsymbol{L}, \boldsymbol{F}, \boldsymbol{A})$: the lattice matrix containing three basis vectors $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3] \in \mathbb{R}^{3\times3}$, the fractional coordinates $\boldsymbol{F} = [\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N] \in [0,1)^{3\times N}$, and one-hot representations of atom types $\boldsymbol{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N] \in [0,1]^{h\times N}$. Using these components, the infinite periodic crystal can be described as

$$\left\{(\boldsymbol{a}_i, \boldsymbol{x}_i) \mid \boldsymbol{x}_i = \boldsymbol{L} \cdot (\boldsymbol{f}_i + \boldsymbol{k}), \forall \boldsymbol{k} \in \mathbb{Z}^{3\times1}, \forall i \in [1..N]\right\} \tag{4}$$

which tell all atomic types $\boldsymbol{a}$, and Cartesian coordinates $\boldsymbol{x}$ of every atoms in the structure. DiffCSP uses normalized Gaussian diffusion in its diffusion inference model q which have standard normal distribution as probability prior, i.e., $P_T = \mathcal{N}(0, I)$. Furthermore, the diffusion is applied independently between components, making it possible to split the model as $q = (q^{\boldsymbol{L}}, q^{\boldsymbol{F}}, q^{\boldsymbol{A}})$ where the superscripts indicate the components that the diffusion acts on. We can write the split diffusion inference model of DiffCSP as

$$q_{t,t+1}(\boldsymbol{M}_{t+1}|\boldsymbol{M}_t) = q_{t,t+1}^{\boldsymbol{L}}(\boldsymbol{L}_{t+1}|\boldsymbol{L}_t) \cdot q_{t,t+1}^{\boldsymbol{F}}(\boldsymbol{F}_{t+1}|\boldsymbol{F}_t) \cdot q_{t,t+1}^{\boldsymbol{A}}(\boldsymbol{A}_{t+1}|\boldsymbol{A}_t). \tag{5}$$

Since the denoising generative process of DiffCSP only utilizes the Predictor-Corrector sampling[39] mechanism on the fractional coordinate components, the model needs to be split into $p = (p^{\boldsymbol{L}}, p^{\boldsymbol{F},p}, p^{\boldsymbol{F},c}, p^{\boldsymbol{A}})$ where the boldface superscripts indicates the components that the diffusion gives while p and c superscripts indicate predictor and corrector sub-models, respectively. We can write the split denoising generative model of DiffCSP as

$$p_{t,t-1}(\boldsymbol{M}_{t-1}|\boldsymbol{M}_t) = p_{t,t-1}^{\boldsymbol{L}}(\boldsymbol{L}_{t-1}|\boldsymbol{M}_t) \cdot p_{t,t-1}^{\boldsymbol{F},p}(\boldsymbol{F}_{t-\frac{1}{2}}|\boldsymbol{M}_t) \cdot p_{t,t-1}^{\boldsymbol{F},c}(\boldsymbol{F}_{t-1}|\boldsymbol{L}_{t-1}, \boldsymbol{F}_{t-\frac{1}{2}}, \boldsymbol{A}_{t-1}) \cdot p_{t,t-1}^{\boldsymbol{A}}(\boldsymbol{A}_{t-1}|\boldsymbol{M}_t). \tag{6}$$

The lattice and atomic type components got denoised by their respective model to get $\boldsymbol{L}_{t-1}$, and $\boldsymbol{A}_{t-1}$, respectively. For the fractional coordinate components, the predictor denoises them to the half-time-step point $\boldsymbol{F}_{t-\frac{1}{2}}$, and the corrector uses $\boldsymbol{F}_{t-\frac{1}{2}}$, $\boldsymbol{L}_{t-1}$, and $\boldsymbol{A}_{t-1}$ to predict $\boldsymbol{F}_{t-1}$. Because of this additional prediction of $\boldsymbol{F}_{t-\frac{1}{2}}$, algorithm 1 need to be slightly modified to accommodate the constraints that are also imposed on the $\boldsymbol{F}_{t-\frac{1}{2}}$ as shown in algorithm 2.

Algorithm 2 Structural Constraint Integration in Material Generation with DiffCSP

Input: constrained structure $\boldsymbol{M}_0^c = (\boldsymbol{L}_0^c, \boldsymbol{F}_0^c, \boldsymbol{A}_0^c)$, constraint mask $\boldsymbol{m} = (\boldsymbol{m}^{\boldsymbol{L}}, \boldsymbol{m}^{\boldsymbol{F}}, \boldsymbol{m}^{\boldsymbol{A}})$, diffusion inference model $q = (q^{\boldsymbol{L}}, q^{\boldsymbol{F}}, q^{\boldsymbol{A}})$, denoising generative model $p = (p^{\boldsymbol{L}}, p^{\boldsymbol{F},p}, p^{\boldsymbol{F},c}, p^{\boldsymbol{A}})$, number of steps T
Sample $\boldsymbol{M}_T^u = (\boldsymbol{L}_T^u, \boldsymbol{F}_T^u, \boldsymbol{A}_T^u) \sim \mathcal{N}(0, I)$, $\boldsymbol{M}_T^c = (\boldsymbol{L}_T^c, \boldsymbol{F}_T^c, \boldsymbol{A}_T^c) \sim \mathcal{N}(0, I)$
$\boldsymbol{M}_T = (\boldsymbol{L}_T, \boldsymbol{F}_T, \boldsymbol{A}_T) \leftarrow \boldsymbol{m} \odot \boldsymbol{M}_T^c + (1 - \boldsymbol{m}) \odot \boldsymbol{M}_T^u$
for $t = T, \ldots, 1$ **do**
    Sample $\boldsymbol{M}_{t-1}^c = (\boldsymbol{L}_{t-1}^c, \boldsymbol{F}_{t-1}^c, \boldsymbol{A}_{t-1}^c) \sim q_{0,t-1}(\boldsymbol{M}_{t-1}^c | \boldsymbol{M}_0^c)$
    Sample $\boldsymbol{L}_{t-1}^u \sim p_{t,t-1}^{\boldsymbol{L}}(\boldsymbol{L}_{t-1}^u | \boldsymbol{M}_t)$, $\boldsymbol{A}_{t-1}^u \sim p_{t,t-1}^{\boldsymbol{A}}(\boldsymbol{A}_{t-1}^u | \boldsymbol{M}_t)$
    $\boldsymbol{L}_{t-1} \leftarrow \boldsymbol{m}^{\boldsymbol{L}} \odot \boldsymbol{L}_{t-1}^c + (1 - \boldsymbol{m}^{\boldsymbol{L}}) \odot \boldsymbol{L}_{t-1}^u$
    $\boldsymbol{A}_{t-1} \leftarrow \boldsymbol{m}^{\boldsymbol{A}} \odot \boldsymbol{A}_{t-1}^c + (1 - \boldsymbol{m}^{\boldsymbol{A}}) \odot \boldsymbol{A}_{t-1}^u$
    Sample $\boldsymbol{F}_{t-\frac{1}{2}}^c \sim q_{0,t-1}^{\boldsymbol{F}}(\boldsymbol{F}_{t-1}^c | \boldsymbol{F}_0^c)$
    Sample $\boldsymbol{F}_{t-\frac{1}{2}}^u \sim p_{t,t-1}^{\boldsymbol{F},p}(\boldsymbol{F}_{t-\frac{1}{2}}^u | \boldsymbol{M}_t)$
    $\boldsymbol{F}_{t-\frac{1}{2}} \leftarrow \boldsymbol{m}^{\boldsymbol{F}} \odot \boldsymbol{F}_{t-\frac{1}{2}}^c + (1 - \boldsymbol{m}^{\boldsymbol{F}}) \odot \boldsymbol{F}_{t-\frac{1}{2}}^u$
    Sample $\boldsymbol{F}_{t-1}^u \sim p_{t,t-1}^{\boldsymbol{F},c}(\boldsymbol{F}_{t-1}^u | \boldsymbol{L}_{t-1}, \boldsymbol{F}_{t-\frac{1}{2}}, \boldsymbol{A}_{t-1})$
    $\boldsymbol{F}_{t-1} \leftarrow \boldsymbol{m}^{\boldsymbol{F}} \odot \boldsymbol{F}_{t-1}^c + (1 - \boldsymbol{m}^{\boldsymbol{F}}) \odot \boldsymbol{F}_{t-1}^u$
return $\boldsymbol{M}_0 = (\boldsymbol{L}_0, \boldsymbol{F}_0, \boldsymbol{A}_0)$.

Pre-screening procedure to retrieve stable materials structures

After generating materials constrained by AL structures, evaluating their stability becomes critical. Given the large volume of generated candidates–often in the millions–a fast and reliable method is essential to identify viable structures. To address this, we implement a four-staged pre-screening filter that systematically evaluates materials against a series of stability criteria, ensuring only the most promising candidates are advanced for further analysis.

Stage 1. Charge Neutrality

Ensuring electrical neutrality is essential for material stability and practical applications. We applied the SMACT approach[40] to evaluate the charge neutrality of generated materials. This method, inspired by the DiffCSP framework[16], screens out chemically infeasible candidates, allowing only neutral and viable structures to proceed to subsequent evaluation stages.

Stage 2. Space Occupancy Ratio

Generated materials with excessively dense atomic configurations are often unrealistic in crystalline phases. To address this, we compare the atomic density of generated structures to a reference dataset (MP-20) and eliminate those with overly high densities. The space occupancy ratio, R_{occ}, is calculated as

$$R_{occ} = \frac{\sum_{i=1}^{N} \frac{4\pi r_i^3}{3}}{V_{cell}} \tag{7}$$

where r_i is the radius of the i-th atom and V_{cell} is the volume of the unit cell. As shown in Supplementary Information 7, the R_{occ} distribution in the MP-20 dataset remains consistent across different N. We exclude materials with $R_{occ} > 1.7$, a threshold based on the distribution of R_{occ} in the MP-20 dataset.

Graph Neural Networks for Stability Evaluation

To rapidly assess the stability of the remaining material candidates, we utilize graph neural networks (GNNs) based on the E3NN[41,42] framework, designed for their efficiency in handling crystallographic data. We develop two GNN models, serviing as Stage 3 and 4 in the stability filer, respectively:

- **Stage 3. GNN classifier for energy above hull (high/low):** This model predicts whether a material's energy above the convex hull (E_{hull}) is below a threshold of 0.1 eV or not, which is indicative of thermodynamic stability. The model is trained using data from the Matbench-discovery[43].
- **Stage 4. GNN for stable structure distinction:** Recognizing the potential for our model to generate materials that diverge from known stable structures, this classifier distinguishes between pristine and diffused structures. Training data includes original structures from the MP-20 dataset and those with added Gaussian noise on unit cell matrix $\boldsymbol{L}$ and fractional coordinates $\boldsymbol{F}$, simulating potential inaccuracies in atomic positions during generation. As the training dataset

of at this stage, we diffuse $\boldsymbol{L}$ and $\boldsymbol{F}$ as $\boldsymbol{L}' = \boldsymbol{L} + (l_1, l_2, l_3)^\intercal \cdot \mathcal{N}(0, \sigma_d^2 I)$ and $\boldsymbol{F}' = w(\boldsymbol{F} + \mathcal{N}(0, \sigma_d^2 I))$, respectively. Here, σ_d regulates the diffusion level and $w(\cdot)$ is a wrapping function that adjusts fractional coordinates to fit within $[0, 1)$. We decide that 1% diffusion materials ($\sigma_d = 0.01$) are stable, but 5% diffusion materials ($\sigma_d = 0.05$) are unstable. This GNN model helps ensure that even stable materials are not overly distorted or unrealistic.

The two GNN-based classifiers have relatively simple architectures but present great accuracy. We present the confusion matrices of the two models in Supplementary Information 7. The prediction accuracy for the test dataset is 0.83 and 0.99, respectively.

The four-staged pre-screening filter provides an efficient screening tool to evaluate the stability of the generated materials and down-select potential stable materials. We argue this stability evaluation is valid, as we observe the materials gain stability as it goes through the denoising process from the noisy structures ($\boldsymbol{M}_T$) to the pristine ones ($\boldsymbol{M}_0$). Moreover, more than 50% of pre-screened materials after the four-staged filter survive DFT relaxation, indicating the efficacy of the pre-screening process. The details of the stability evaluation are shown in Supplementary Information 7.

DFT for stability evaluation and structural relaxation

The candidate models are further evaluated with DFT for potential structural stability. Due to the high cost of DFT calculations, it is necessary to balance the accuracy and throughput. In the first stage of DFT screening, we choose to use a relatively coarse treatment of the electronic structure to evaluate as many candidates as possible (up to 26,000). Plane-wave DFT calculations are performed using the Vienna Ab initio Simulation Package (VASP)[44]. The calculation used Projector Augmented Wave (PAW) method[45,46] to describe the effects of core electrons and Perdew-Burke-Ernzerhof (PBE)[47] implementation of the Generalized Gradient Approximation (GGA) for the exchange-correlation functional. The energy cutoff is $1.2 \cdot \max(\text{ENMAX})$ for the plane-wave basis of the valence electrons. The electronic structure is calculated on Γ-centered mesh for the unit cell (the grid length density is 5 k-points per nm^{-1}). The total energy tolerance for electronic energy minimization is 10^{-6} eV, and the energy criterion for structure optimization is 10^{-5} eV. The maximum number of steps is 60 for electronic self-consistent calculation and 150 for structural optimization. During the structural relaxation, the symmetry of the crystal is maintained while the cell shape/size and all atomic coordinates are allowed to relax. Non-spin-polarized calculation is employed for this initial screening. A small fraction of models failed the electronic structure calculation. We terminate the calculation of the failure cases, and we regard the corresponding candidates as unstable. For the candidates with completed VASP calculation (either because the energy criterion is reached or the maximum number of relaxation steps is reached), the following quantities are extracted/calculated as indicators of potential stability: (1) maximum interatomic force after relaxation, (2) initial and final total energy, (3) average changes in lattice constants, (4) average changes in atomic coordinates. These quantities are then analyzed and compared to identify potentially stable candidates worth further and more rigorous evaluations.

Band structure calculation for Lieb-like lattice materials

The DFT band structure calculations for Lieb-like lattice structures are performed using VASP. PAW method and PBE exchange-correlation functional are used for all DFT calculations. The initial electronic structure calculations are performed on a K-point mesh centered at Gamma point with resolved value $k_{mesh} = 0.03 \cdot 2\pi/\text{Å}$ for each structure. The band structure is subsequently calculated on a high symmetry path generated by the VASPKIT code[48].

Data visualization

We use VESTA[49] to visualize the materials structures presented in the main article. We utilize OVITO[50] to visualize the materials structures for the Supplementary Information.

Experimental synthesis of samples

The polycrystalline samples of the ternary compounds TiPdM (M = Bi and Sb) are synthesized via a standard arc-melting process. Further details on the experimental procedures, including synthesis, characterization, measurements, and discussions for TiPdM, are provided in Supplementary Information 10.

Data Availability Statement

We compile a comprehensive database of AL materials generated by SCIGEN. The dataset provides the folders of all generated materials (7.87 million), the materials that survived after the four-stage pre-screening process (790 thousand materials), and DFT-relaxed structures (24,743). The folder with DFT calculation contains materials structures before and after relaxation. The Supplementary dataset is available in Figshare repository[51].

Code Availability Statement

The source code is available at (https://github.com/RyotaroOKabe/SCIGEN).

Acknowledgements

RO and ML thank C Batista, A Christianson, F Frenkel, A May, R Moore, B Ortiz, and F Ronning for the helpful discussion. RO acknowledges the support from the U.S. Department of Energy (DOE), Office of Science (SC), Basic Energy Sciences (BES), Award No. DE-SC0021940 and Heiwa Nakajima Foundation. AC acknowledges support from National Science Foundation (NSF) Designing Materials to Revolutionize and Engineer our Future (DMREF) Program with Award No. DMR-2118448. BH and YC are partially supported by the Artificial Intelligence Initiative as part of the Laboratory Directed Research and Development (LDRD) program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC, for the US Department of Energy under Contract DE-AC05-00OR22725. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a Department of Energy Office of Science User Facility using NERSC award DDR-ERCAP0030758. Computing resources for a portion of the work were made available through the VirtuES project, funded by the LDRD Program and Compute and Data Environment for Science (CADES) at ORNL. Another portion of simulation results were obtained using the Frontera computing system at the Texas Advanced Computing Center. WX and RC were supported by the Department of Energy, grant DE-FG02-98ER45706. WX and RC thank G.J. Miller in Chemistry Department, Iowa State University, for offering clusters to perform LMTO calculations. ML acknowledges the support from NSF ITE-2345084, the Class of 1947 Career Development Chair, and the support from R. Wachnik.

References

1. Fu, L. Topological Crystalline Insulators. *Phys. Rev. Lett.* **106,** 106802. https://link.aps.org/doi/10.1103/PhysRevLett.106.106802 (10 Mar. 2011).
2. *Rashba-like physics in condensed matter - Nature Reviews Physics — nature.com* https://www.nature.com/articles/s42254-022-00490-y. [Accessed 11-06-2024].
3. Martin, L. & Rappe, A. Thin-film ferroelectric materials and their applications. *Nature Reviews Materials* **2,** 16087. https://www.nature.com/articles/natrevmats201687 (Nov. 2017).
4. Hung, N. T. *et al.* Symmetry breaking in 2D materials for optimizing second-harmonic generation. *Journal of Physics D: Applied Physics* **57,** 333002. https://dx.doi.org/10.1088/1361-6463/ad4a80 (May 2024).
5. Armitage, N. P., Mele, E. J. & Vishwanath, A. Weyl and Dirac semimetals in three-dimensional solids. *Rev. Mod. Phys.* **90,** 015001. https://link.aps.org/doi/10.1103/RevModPhys.90.015001 (1 Jan. 2018).
6. Hashimoto, M., Vishik, I. M., He, R.-H., Devereaux, T. P. & Shen, Z.-X. Energy gaps in high-transition-temperature cuprate superconductors. *Nature Physics* **10,** 483–495. ISSN: 1745-2481. https://doi.org/10.1038/nphys3009 (July 2014).
7. Savary, L. & Balents, L. Quantum spin liquids: a review. *Reports on Progress in Physics* **80,** 016502. https://dx.doi.org/10.1088/0034-4885/80/1/016502 (Nov. 2016).
8. Broholm, C. *et al.* Quantum spin liquids. *Science* **367,** eaay0668 (2020).
9. Kang, M. *et al.* Topological flat bands in frustrated kagome lattice CoSn. *Nature Communications* **11,** 4004. ISSN: 2041-1723. https://doi.org/10.1038/s41467-020-17465-1 (Aug. 2020).
10. Slot, M. R. *et al.* Experimental realization and characterization of an electronic Lieb lattice. *Nature physics* **13,** 672–676 (2017).
11. Checkelsky, J. G., Bernevig, B. A., Coleman, P., Si, Q. & Paschen, S. Flat bands, strange metals and the Kondo effect. *Nature Reviews Materials.* ISSN: 2058-8437. https://doi.org/10.1038/s41578-023-00644-z (Feb. 2024).
12. Van Speybroeck, V. *et al.* Advances in theory and their application within the field of zeolite chemistry. *Chem. Soc. Rev.* **44,** 7044–7111. http://dx.doi.org/10.1039/C5CS00029G (20 2015).
13. Chamorro, J. R., McQueen, T. M. & Tran, T. T. Chemistry of Quantum Spin Liquids. *Chemical Reviews* **121,** 2898–2934. ISSN: 0009-2665. https://doi.org/10.1021/acs.chemrev.0c00641 (Mar. 2021).
14. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. Crystal diffusion variational autoencoder for periodic material generation. *arXiv preprint arXiv:2110.06197* (2021).
15. Yang, M. *et al.* Scalable diffusion for materials generation. *arXiv preprint arXiv:2311.09235* (2023).
16. Jiao, R. *et al.* Crystal structure prediction by joint equivariant diffusion. *Advances in Neural Information Processing Systems* **36** (2024).

17. Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624,** 80–85 (2023).
18. Jiao, R., Huang, W., Liu, Y., Zhao, D. & Liu, Y. Space Group Constrained Crystal Generation. *arXiv preprint arXiv:2402.03992* (2024).
19. Zeni, C. *et al.* Mattergen: a generative model for inorganic materials design. *arXiv preprint arXiv:2312.03687* (2023).
20. Cao, Z., Luo, X., Lv, J. & Wang, L. Space Group Informed Transformer for Crystalline Materials Generation. *arXiv preprint arXiv:2403.15734* (2024).
21. Martinez, J. Archimedean lattices. *Algebra Universalis* **3,** 247–260 (1973).
22. Eddi, A., Decelle, A., Fort, E. & Couder, Y. Archimedean lattices in the bound states of wave interacting particles. *Europhysics Letters* **87,** 56002 (2009).
23. Zimmermann, N. E. & Jain, A. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC advances* **10,** 6063–6081 (2020).
24. Yin, J.-X., Lian, B. & Hasan, M. Z. Topological kagome magnets and superconductors. *Nature* **612,** 647–657 (2022).
25. Kang, M. *et al.* Topological flat bands in frustrated kagome lattice CoSn. *Nature communications* **11,** 4004 (2020).
26. Tsai, W.-F., Fang, C., Yao, H. & Hu, J. Interaction-driven topological and nematic phases on the Lieb lattice. *New Journal of Physics* **17,** 055016 (2015).
27. Mukherjee, S. *et al.* Observation of a localized flat-band state in a photonic Lieb lattice. *Physical review letters* **114,** 245504 (2015).
28. Vicencio, R. A. *et al.* Observation of localized states in Lieb photonic lattices. *Physical review letters* **114,** 245503 (2015).
29. Zhou, X. *et al.* High-temperature superconductivity. *Nature Reviews Physics* **3,** 462–465 (2021).
30. Pickett, W. E. The dawn of the nickel age of superconductivity. *Nature Reviews Physics* **3,** 7–8 (2021).
31. Merker, H. A. *et al.* Machine learning magnetism classifiers from atomic coordinates. *Iscience* **25** (2022).
32. Chang, J. *et al.* Direct observation of competition between superconductivity and charge density wave order in YBa2Cu3O6. 67. *Nature Physics* **8,** 871–876 (2012).
33. Tranquada, J. *et al.* Coexistence of, and Competition between, Superconductivity and Charge-Stripe Order in La 1. 6- x Nd 0.4 Sr x CuO 4. *Physical Review Letters* **78,** 338 (1997).
34. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **120,** 145301 (2018).
35. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **1,** 011002 (2013).
36. Pan, H. *et al.* Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic chemistry* **60,** 1590–1603 (2021).
37. Zachariasen, W. Metallic radii and electron configurations of the 5f- 6d metals. *Journal of Inorganic and Nuclear Chemistry* **35,** 3487–3497 (1973).
38. Lugmayr, A. *et al. Repaint: Inpainting using denoising diffusion probabilistic models* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022), 11461–11471.
39. Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).
40. Davies, D. W. *et al.* Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software* **4,** 1361 (2019).
41. Geiger, M. & Smidt, T. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453* (2022).
42. Chen, Z. *et al.* Direct prediction of phonon density of states with Euclidean neural networks. *Advanced Science* **8,** 2004214 (2021).
43. Riebesell, J. *et al.* Matbench Discovery–An evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920* (2023).
44. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical review B* **54,** 11169 (1996).

45. Blöchl, P. E. Projector augmented-wave method. *Physical review B* **50,** 17953 (1994).

46. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b* **59,** 1758 (1999).

47. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **77,** 3865 (1996).

48. 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. *Computer Physics Communications* **267,** 108033 (2021).

49. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of applied crystallography* **44,** 1272–1276 (2011).

50. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Modelling and simulation in materials science and engineering* **18,** 015012 (2009).

51. Okabe, R. *Structural Constraint Integration in Generative Model for Discovery of Quantum Material Candidates* 2024. https://doi.org/10.6084/m9.figshare.c.7283062.v1.

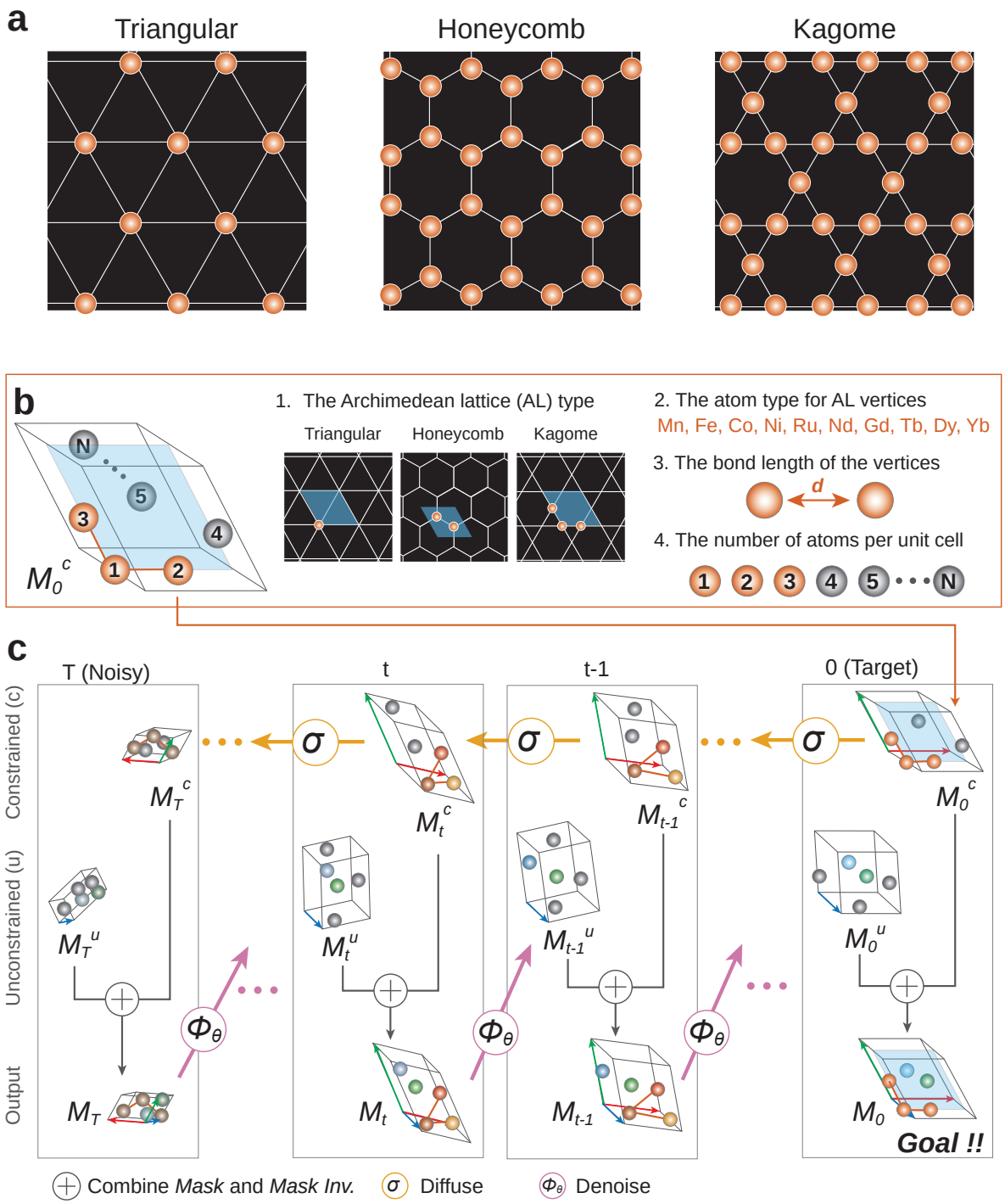

Figure 1. Schematic overview of material generation with geometric patterns as constraints. a. Three primary classes of Archimedean lattices with hexagonal unit cells: triangular, honeycomb, and kagome. **b.** Guideline for structure initialization for the diffusion model with magnetic atoms at Archimedean lattice vertices. Required components include (1) lattice types, (2) magnetic atom types, (3) nearest-neighbor distances, and (4) total number of atoms per unit cell. **c.** Methodology of crystal structure generation via diffusion denoising probabilistic model with geometrical pattern as constraints. The initialized structures are iteratively made noisy (σ) to prepare a predefined pathway of the constrained structure $\boldsymbol{M}_t^c, t \in [1, T]$. For each denoising step t, an unconstrained structure $\boldsymbol{M}_t^u$ is combined with constrained structure $\boldsymbol{M}_t^c$ to get an integrated structure $\boldsymbol{M}_t$. $\boldsymbol{M}_t$ is passed to the denoising model Φ_θ and denoised to become the unconstrained structure $\boldsymbol{M}_{t-1}^u$. By repeating this process, we obtain the final crystal structure $\boldsymbol{M}_0$, which is guided by the geometrical pattern constraints $\boldsymbol{M}_0^c$ but remains realistic. The output crystals are designed to increase the likelihood of stability in the subsequent stability evaluation steps.

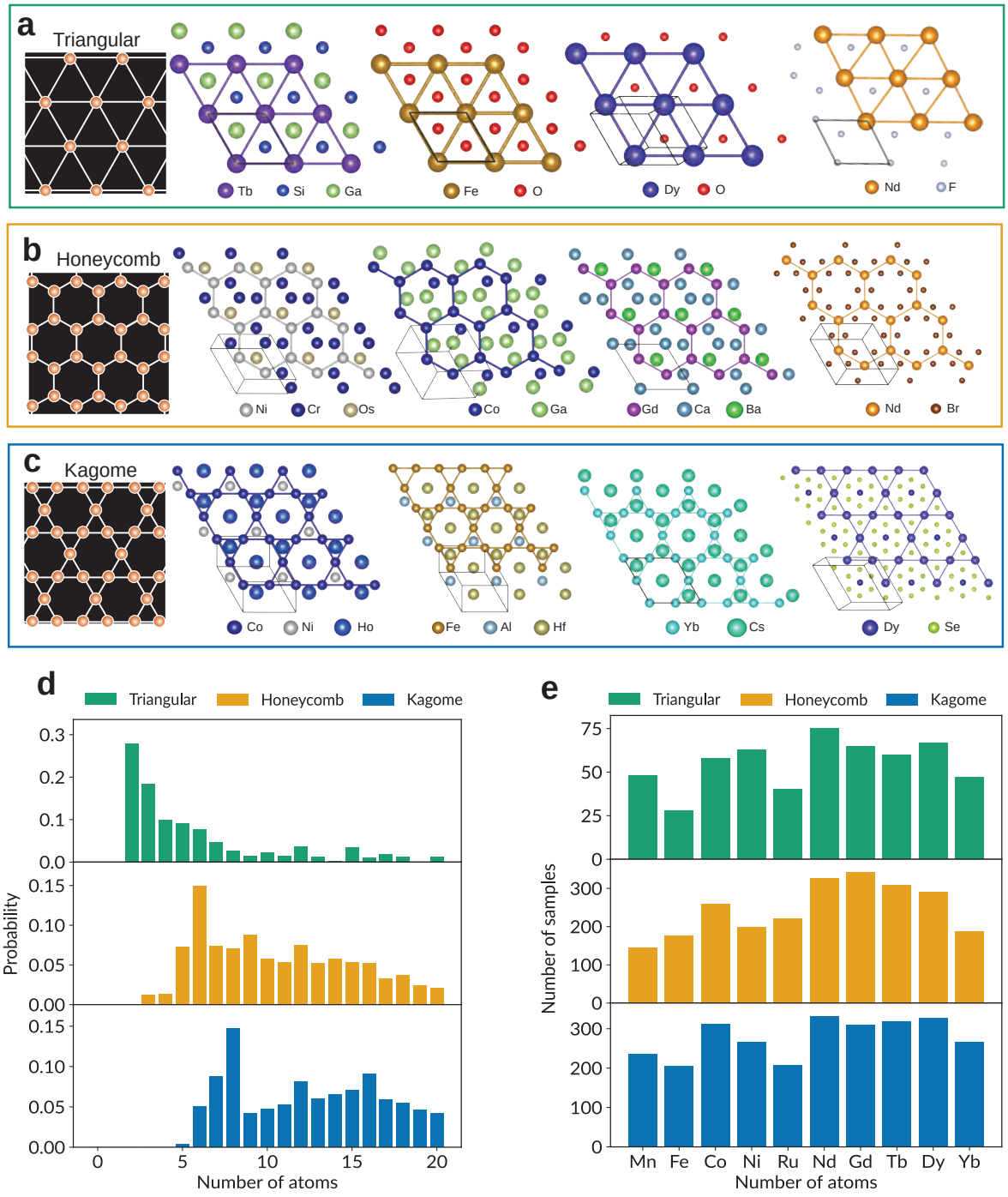

Figure 2. Generated materials with three primary types of Archimedean lattices. Archimedean lattice patterns and generated material structures are displayed for **a.** Triangular, **b.** Honeycomb, and **c.** Kagome lattices. **d.** The sampling profile of the number of atoms per unit cell N generated by measuring the survival ratio from a uniform sampling of N. **e.** The number of materials remaining after pre-screening is presented for the common magnetic atom types in each primary geometrical pattern.

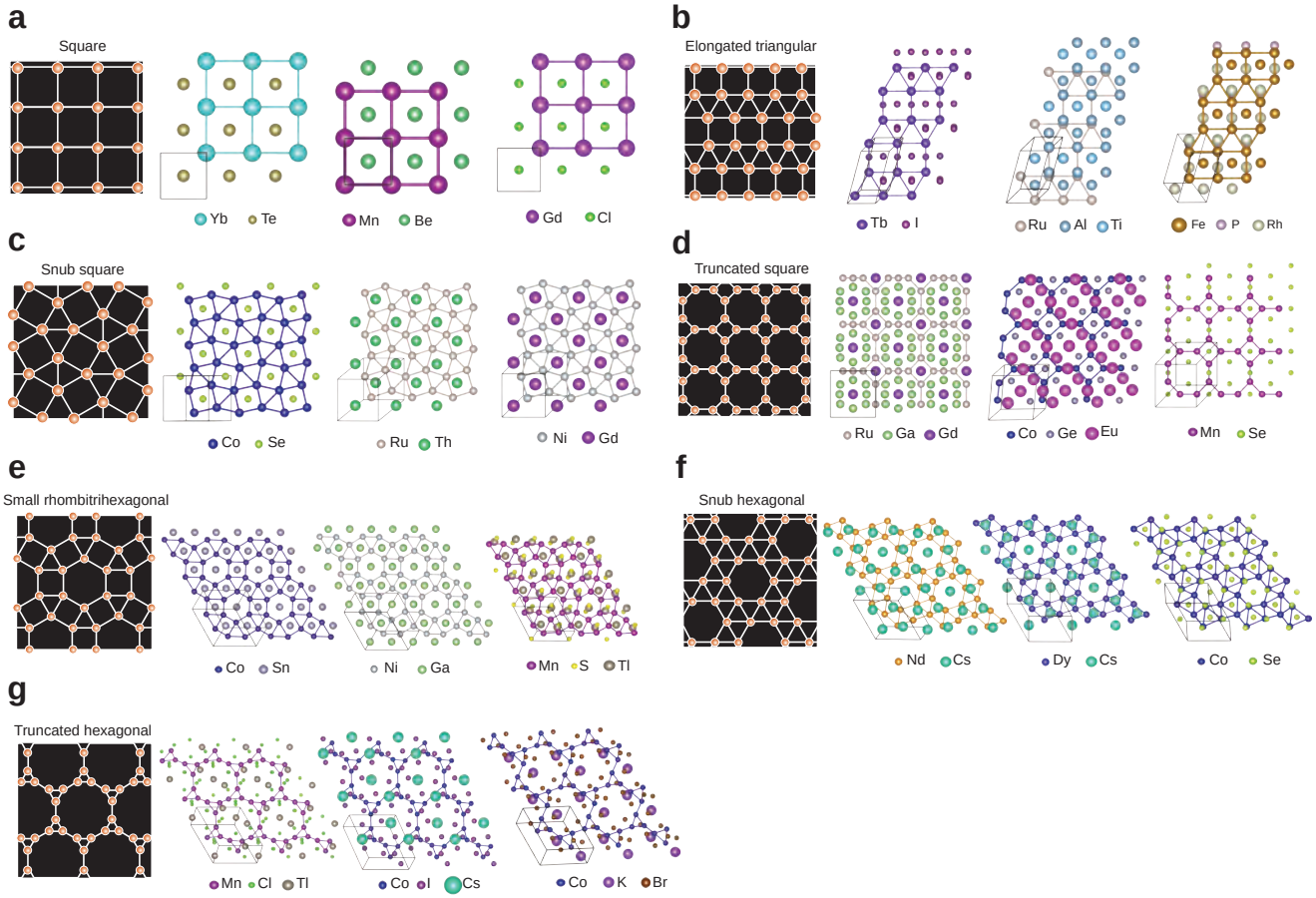

Figure 3. Generated materials with other Archimedean lattice structures. Materials examples covering the rest of Archimedean lattices are presented with **a.** Square **b.** Elongated triangular **c.** Snub square **d.** Truncated square **e.** Small rhombitrihexagonal **f.** Snub hexagonal **g.** Truncated hexagonal. In each subplot, the AL pattern and two examples of generated materials are displayed.

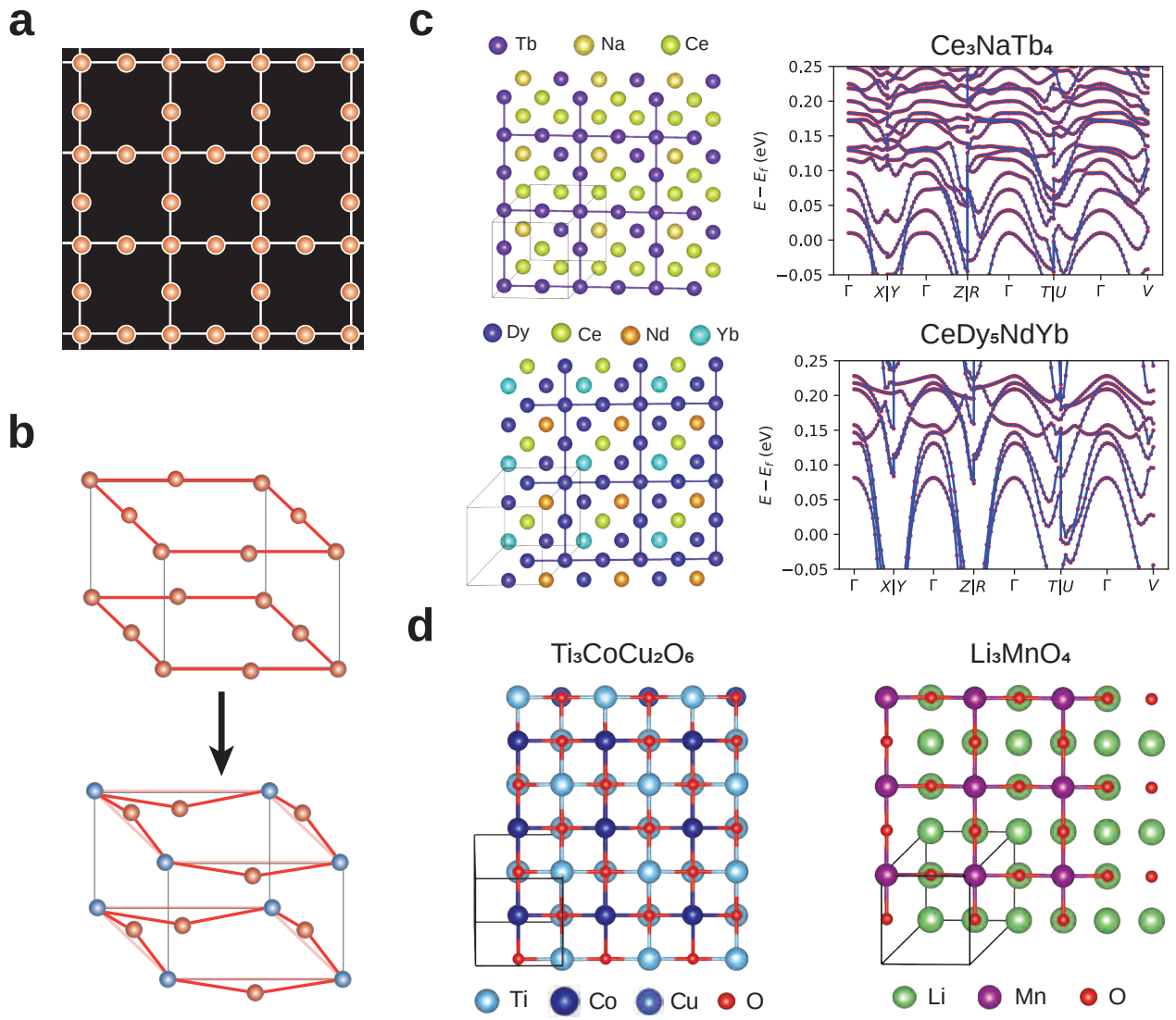

Figure 4. Generated materials of a Lieb-like lattice. a. The Lieb lattice pattern serves as the geometric template integrated into the generated material structures. **b.** Lieb-lattice structure from the side view. Concept of a heteroatomic Lieb lattice in which bridging sites are occupied by a second element (blue), allowing slight out-of-plane relaxation enabled by the Adaptive-unmasking strategy (see Supplementary Information 11). **c.** Generated supercell and flat band structures of the Lieb-like lattice materials for Ce_3NaTb_4 and $CeDy_3NdYb$, with the Fermi level E_F set to 0 eV. The flat bands in both examples are slightly (0.1 – 0.2 eV) above the Fermi level. **d.** $Ti_3CoCu_2O_6$ and Li_3MnO_4 as examples of the generated materials with heteroatomic Lieb lattices.