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], 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: chengy@ornl.gov
[***]e-mail: mingda@mit.edu

ABSTRACT

Billions of organic molecules are known, but only a tiny fraction of the functional inorganic materials have been discovered, a particularly relevant problem to the community searching for new quantum materials. Recent advancements in machine-learning-based generative models, particularly diffusion models, show great promise for generating new, stable materials. However, integrating geometric patterns into materials generation remains a challenge. Here, we introduce Structural Constraint Integration in the GENerative model (SCIGEN). Our approach can modify any trained generative diffusion model by strategic masking of the denoised structure with a diffused constrained structure prior to each diffusion step to steer the generation toward constrained outputs. Furthermore, we mathematically prove that SCIGEN effectively performs conditional sampling from the original distribution, which is crucial for generating stable constrained materials. We generate eight million compounds using Archimedean lattices as prototype constraints, with over 10% surviving a multi-staged stability pre-screening. High-throughput density functional theory (DFT) on 26,000 survived compounds shows that over 50% passed structural optimization at the DFT level. Since the properties of quantum materials are closely related to geometric patterns, our results indicate that SCIGEN provides a general framework for generating quantum materials candidates.

Introduction

The structure-property relationships are instrumental in understanding quantum and functional materials, and are a fundamental part of any materials science curriculum. Two key structural indicators, symmetry, and geometric pattern, profoundly influence materials properties. 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]. Meanwhile, a material's geometric pattern is closely linked to its electronic states and magnetic orderings. The square lattice serves as a prototype 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 importance of replacing rare-earth elements[11]. Also, porous structures like zeolite lattices find applications in catalysis[12]. However, designing stable materials with desired properties can be nontrivial. For example, only a dozen quantum spin liquid candidates have been identified after a decade of research[13], and even fewer are known for the Lieb lattice.

Machine-learning (ML) based materials 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 create new materials with respect to the distribution of the database, making it challenging to generate materials with specific constraints. Although there have been some developments in the incorporation of crystallographic space groups in the materials generation[18–20], the integration of geometric patterns into generation algorithms for functional materials remains challenging. In quantum materials, space group symmetry and geometric patterns can be independent; including the space group symmetry alone will sometimes not allow for proper screening, as is often encountered in monoclinic stacking variants of hexagonal symmetry layers. Moreover, in frustrated magnets, the geometric pattern such as a kagome lattice filled with magnetic atoms plays more important roles than the overall space group in supporting the exotic magnetic structures. Therefore, there is a pressing need to develop an ML-based generator capable of producing new materials constrained by particular geometric patterns.

To answer this need, in this work, we present SCIGEN: Structural Constraint Integration in the GENerative model. SCIGEN is a scheme that can be utilized by any pre-trained generative diffusion model for the incorporation of geometric pattern and symmetry constraints during the generation, without the need for retraining or fine-tuning. Starting from the target constraints, SCIGEN diffuses a random constrained structure over multiple time steps. The constrained structures are used to 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. It turns out that, as we have proven, SCIGEN effectively performs conditional generation with respect to the distribution of the base model. This indicates that the constraint set by SCIGEN would preserve the integrity of the base generative model, including but not limited to the stability of generated materials. To demonstrate, we apply SCIGEN to DiffCSP[16] for generating materials constrained by Archimedean lattices (ALs)[21,22], which are a collection of 2D lattice tiling with square, triangular, honeycomb, kagome, and a few other geometric patterns, and rich harbors for exotic quantum materials. We generate a total of 7.87 million materials belonging to ALs. After a four-staged stability pre-screening, over 790,000 materials survived. Structure relaxation on a subset containing 26,000 materials is computed with high-throughput density functional theory (DFT), showing that 95% completed the calculation, and more than 53% can reach the energy minimum within 150 steps of structural optimization. Since SCIGEN requires no extra training apart from the underlying generative model, it also offers a flexible and generically applicable conditioning scheme of materials generation with constraints from both 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 Fig. 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. More details are shown in Supplementary Information 3. Additionally, as proven in Supplementary Information 4, SCIGEN can take the structural constraints and fill the remaining unconstrained components, while preserving the integrity of the unconstrained optimization. Following the generation of a large set of material candidates, we evaluate their stability through a four-staged 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 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 display the results of materials generation constrained by three primary AL types: (a) triangular, (b) honeycomb, and (c) kagome. The AL structures are formed in the generated structures, as SCIGEN algorithm has guided them to be formed as pre-defined. The positions of the other unconstrained atoms are not specified rigorously but are often found to reside on the sites that bridge the AL atoms. For triangular lattices (Fig. 2a), each of the unconstrained atoms is placed on the sites connecting with three magnetic atoms forming an equilateral triangle. For honeycomb lattices (Fig. 2b), we often observe

materials with one unconstrained atom at the center of the hexagon formed with magnetic atoms within the same plane. As to kagome materials (Fig. 2c), unconstrained atoms bridge the equilateral triangles and the hexagons of kagome lattice layers. If the space inside the polygons is too small compared to the atomic radii, the filler atoms will be pushed outside the AL plane. On the other hand, large polygons like hexagons can accommodate the filler atoms to fit 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 for initializing the large-scale generative process. Figure 2d shows the sampling profile of N, which covers all of the 10 common 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 choose to sample atom types for AL vertices with equal probabilities for large-scale materials generation and database construction. Methods section and Supplementary Information 2 describe in detail the sampling schemes for the initialization conditions.

Our exploration of materials with geometrical constraints does not end with the three primary types of ALs but can 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, Great rhombitrihexagonal, is not presented due to the challenge to generate 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, even when not explicitly constrained to form specific lattice structures, these unconstrained atoms frequently organize into recognizable Archimedean patterns. 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 AL that consists of a square lattice with additional atoms located at the centers of each edge of the squares, as visualized in Fig. 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, as we expect for AL structures. This can result in complex magnetic states, which are of significant interest for studying quantum magnetism. Beyond that, Lieb lattices are studied to possess characteristic electronic properties. One key feature of the Lieb lattice is the presence of a flat electronic band[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 Lieb lattice are extremely rare.

In this work, we also focus on the Lieb-like lattice where magnetic atoms sit on the Lieb lattice. Figures 4b,c show the generated materials with Lieb-lattice-based 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 there are very few known materials that fit the desired geometric pattern.

Database of the materials with Archimedean lattice

As detailed in Supplementary Information 7, we generate an AL materials database using SCIGEN. The database contains three components: the total 7.87 million materials generated by the SCIGEN model, the 790 thousand materials that survived four

stages of stability pre-screening processes, and 24,743 out of 26,000 sub-sampled materials in which DFT calculations are successfully converged. By systematically extending our exploration to encompass a wider range of ALs, we can investigate new exotic magnetic orderings, discover porous structures beyond zeolites, and explore new electronic flatband structures, among other possibilities.

Conclusions

In this work, we present SCIGEN, a new generative model aimed at discovering quantum material candidates that adhere to geometric constraints. Our method leverages an AL layer within the crystal structure to identify potential quantum materials. These materials have been validated through DFT to ensure that their relaxed structures are consistent with the machine-learning generations.

To further enhance the validity of SCIGEN, it is crucial to conduct experimental verification through the synthesis of these machine-generated materials. Computation-aided synthesizability check may involve additional analysis in binary, ternary, or other more complex phase diagrams. Setting aside experimental validation, SCIGEN paves an avenue for a few future directions. By focusing on atomic arrangements, we can explore additional geometry-related constraints, such as bonding types, coordination numbers, short-range orderings, and point-group and space group symmetries, during materials generation. Additionally, integrating other diffusion channels, like through the virtual node approach, allows us to incorporate more complex constraints such as defect constraints and magnetic interaction constraints, broadening the scope of the SCIGEN model. Moreover, conditioning the generation process with targeted functionalities, such as specific electrical and optoelectronic properties, or sustainability or environmental impact of materials, can lead to the direct creation of materials with tailored performance. Our SCIGEN model represents a general machine learning-based framework for discovering new quantum materials. It leverages information typically absent from crystal structure databases, offering deeper insights into the structure-property relationships of emerging quantum materials.

Methods

Initialization of the Archimedean Lattices

We describe the workflow to initialize the materials generation process related to the both AL and the entire structure of 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. Here we provide a summary of the initialization process in Fig. 2b, where 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 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. To generate candidate materials which may host geometrically frustrated quantum magnetism, we specify 10 types of common magnetic atoms (Mn, Fe, Co, Ni, Ru, Nd, Gd, Tb, Dy, Yb) on the vertices. The atom types are chosen independently from the AL choice above.

Third, we sample the constrained magnetic bond lengths d^c, aka the distances between the nearest-neighbor magnetic atoms forming the ALs. For each magnetic atom type $\mathscr{A}^c$, we generated the profile of the bond lengths by sampling the nearest-neighbor distances between the corresponding atoms in the MP-20 dataset[29,30] using CrystalNN[31]. To ensure the nearest-neighbor distances do not become significantly close, we cut the minimum lengths by the metallic radii[32] for each atom type. 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 preferable 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 distribution of with each magnetic atom type $\mathscr{A}^c$, are 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 them to be placed 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 for any diffusion model as a base model. Without loss of generality, 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 injecting noise to the input structure, $\boldsymbol{M}_0$. The inference probability of most diffusion models 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 denoising generative model p that can inverse 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. Note that, since q is normally chosen to be a simple probabilistic function, we cannot easily find its inverse. Hence, the training for the denoising generative model is required.

Our approach to material design is summarized in Algorithm 1, where integration of geometrical constraints plays a pivotal role in ensuring that certain structural elements, like lattice configurations or specific atomic distributions, adhere closely to predefined parameters. This method effectively blends prescribed structural characteristics with the creative latitude allowed in other aspects of the material's architecture. Previously, a generative model with unmasked areas as constraints has been employed in image generation for image inpainting, such as RePaint method[33]. However, the geometrical pattern constraint for crystal generation is still challenging and differs in a few ways. The generation of crystal generative model with geometrical constraint involves several key steps:

1. **Adding noise to the constraint structures:** Initially, we introduce noise to a structure which is randomly selected from structures that satisfy the target constraints (This constrained structure can be unstable, or unrealistic as long as it contains the target constraints.) with the diffusion inference model, q, to get diffused constrained structures for each time step $t \in [1, T]$. This operation is aimed at creating a predefined pathway for denoising the constrained structures. The unconstrained components of the crystals are guided by this known denoising pathway, which results in the presence of constrained components in the final outputs.

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

3. **Integration of the constrained and unconstrained structures:** For each denoising step, after processing both parts independently, they are carefully recombined. This combination is critical as it ensures the integrity of the predefined constraints is maintained while integrating seamlessly with the freely generated segments. This method preserves essential structural features and fosters innovation in material design.

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, indicated with superscript c, $\boldsymbol{M}_0^c$, along with the corresponding masks $\boldsymbol{m}$, which indicate the constrained components in $\boldsymbol{M}_0^c$ with binary masking, i.e., assigns value of 1 to constrained, and 0 to the unconstrained components. The final-time-step unconstrained structure, indicated with superscript u, $\boldsymbol{M}_T^u$, and constrained structure $\boldsymbol{M}_T^c$ are sampled from the probability prior P_T of the base model. Through masking, 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 $\boldsymbol{M}_T^u$ formulated 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 at time step $t-1$, $\boldsymbol{M}^u_{t-1}$, as $p_{t,t-1}(\boldsymbol{M}^u_{t-1}|\boldsymbol{M}_t)$. Concurrently, the diffusion process gives the distribution of the constrained structure $\boldsymbol{M}^c_{t-1}$ as $q_{0,t-1}(\boldsymbol{M}^c_{t-1}|\boldsymbol{M}^c_0)$. Then, the unconstrained structure $\boldsymbol{M}^u_{t-1}$, and constrained structure $\boldsymbol{M}^c_{t-1}$ 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. Using masks to combine constrained and unconstrained parts ensures that the constraints are maintained throughout the iterative refinement process, resulting in high-quality material structures suitable for practical applications.

Algorithm 1 Structural Constraint Integration in Material Generation Procedure

Input: constrained structure $\boldsymbol{M}^c_0$, constraint mask $\boldsymbol{m}$, diffusion inference model q, denoising generative model p, number of steps T, probability prior P_T
Sample $\boldsymbol{M}^u_T \sim P_T$, $\boldsymbol{M}^c_T \sim P_T$
$\boldsymbol{M}_T \leftarrow \boldsymbol{m} \odot \boldsymbol{M}^c_T + (1-\boldsymbol{m}) \odot \boldsymbol{M}^u_T$
for $t = T, \ldots, 1$ **do**
    Sample $\boldsymbol{M}^c_{t-1} \sim q_{0,t-1}(\boldsymbol{M}^c_{t-1}|\boldsymbol{M}^c_0)$
    Sample $\boldsymbol{M}^u_{t-1} \sim p_{t,t-1}(\boldsymbol{M}^u_{t-1}|\boldsymbol{M}_t)$
    $\boldsymbol{M}_{t-1} \leftarrow \boldsymbol{m} \odot \boldsymbol{M}^c_{t-1} + (1-\boldsymbol{m}) \odot \boldsymbol{M}^u_{t-1}$
return $\boldsymbol{M}_0$.

To demonstrate the algorithm for the generation of AL materials, we chose DiffCSP[16] as the base model of SCIGEN before applying geometric constraint. 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

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

Since the denoising generative process of DiffCSP utilizes Predictor-Corrector sampling[34] mechanism on the fractional coordinate components only, the model need 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

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

Basically, 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.

Pre-screening procedure to retrieve stable materials structures

Following the generation of materials constrained by AL structures, it becomes essential to evaluate their stability. Due to the high volume of generated candidates—often reaching into the millions—a rapid yet reliable method is necessary to assess stability. Here, we describe our four-staged pre-screening of materials based on a series of stability criteria.

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 $\boldsymbol{M}_0 = (\boldsymbol{L}_0, \boldsymbol{F}_0, \boldsymbol{A}_0)$.

Charge Neutrality

Materials must be electrically neutral to ensure stability and real-world applicability. We employed the SMACT approach[35] to evaluate the charge neutrality of generated materials. This process, inspired by methodologies described in the CDVAE approach[14], ensures that only chemically feasible materials are considered in subsequent steps.

Density and Space Occupancy Ratio

Some generated materials feature densely packed atomic configurations, which are unrealistic in actual crystalline phases. To address this, we compare these materials against a reference dataset (MP-20) to identify and eliminate those with excessively high atom 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. Materials in MP-20 observed a similar distribution of R_{occ} that is independent of N, as we can find in Supplementary Information 6. We discard materials with an R_{occ} value exceeding 1.7, a threshold based on the distribution of R_{occ} in the MP-20 dataset.

Graph Neural Network Classifiers for Stability Evaluation

To rapidly assess the stability of the remaining material candidates, we utilize graph neural networks (GNNs) based on the E3NN[36,37] framework, designed for their efficiency in handling crystallographic data. We develop two models:

- **GNN for Stability classification Ψ_1:** This model predicts whether a material's energy above the convex hull (E_{hull}) is below a threshold of 0.1 eV, which is indicative of thermodynamic stability. The model is trained using data from the Matbench-discovery[38].
- **GNN for Stability classification Ψ_2:** 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 the ones 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 Ψ_2, we diffuse $\boldsymbol{L}$ and $\boldsymbol{F}$ as $\boldsymbol{L}' = \boldsymbol{L} + (l_1, l_2, l_3)^T \cdot \mathcal{N}(0, \sigma_d^2 I)$ and $\boldsymbol{F}' = w(\boldsymbol{F} + \mathcal{N}(0, \sigma_d^2 I))$, respectively. Here, σ_d regulates the level of diffusion, and $w(\cdot)$ is a wrapping function that adjust 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 model helps ensure that even stable materials are not overly distorted or unrealistic.

The two GNN-based classifiers are in simple architecture, but presented high accuracy. We present the confusion matrices of the two models in Supplementary Information 6. The prediction accuracy for the test dataset is 0.83 and 0.99, respectively.

The four-staged pre-screening filters could provide us with an extremely efficient screening tool to evaluate the stability of the generated 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 survive DFT relaxation, indicating the efficacy of the pre-screening process. The detail of the stability evaluation is shown in Supplementary Information 6.

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). Planewave DFT calculations are performed using the Vienna Ab initio Simulation Package (VASP)[39]. The calculation used Projector Augmented Wave (PAW) method[40,41] to describe the effects of core electrons and Perdew-Burke-Ernzerhof (PBE)[42] 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, which are terminated and the corresponding candidates are considered 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[43].

Data visualization

We use VESTA[44] to visualize the materials structures presented in the main article. For Supplementary Information, we utilize OVITO[45] to visualize the materials structures.

Data Availability Statement

We compile a comprehensive database of AL materials generated by SCIGEN. The dataset provides the folders of the entire generated materials (7.87 million), the survived materials after the four-staged 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[46].

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. 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. 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. 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).
30. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials* **1,** 011002 (2013).
31. Pan, H. *et al.* Benchmarking coordination number prediction algorithms on inorganic crystal structures. *Inorganic chemistry* **60,** 1590–1603 (2021).
32. Zachariasen, W. Metallic radii and electron configurations of the 5f- 6d metals. *Journal of Inorganic and Nuclear Chemistry* **35,** 3487–3497 (1973).
33. 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.
34. Song, Y. *et al.* Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* (2020).
35. Davies, D. W. *et al.* Smact: Semiconducting materials by analogy and chemical theory. *Journal of Open Source Software* **4,** 1361 (2019).
36. Geiger, M. & Smidt, T. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453* (2022).
37. Chen, Z. *et al.* Direct prediction of phonon density of states with Euclidean neural networks. *Advanced Science* **8,** 2004214 (2021).
38. Riebesell, J. *et al.* Matbench Discovery–An evaluation framework for machine learning crystal stability prediction. *arXiv preprint arXiv:2308.14920* (2023).
39. 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).
40. Blöchl, P. E. Projector augmented-wave method. *Physical review B* **50,** 17953 (1994).
41. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical review b* **59,** 1758 (1999).
42. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical review letters* **77,** 3865 (1996).
43. 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).
44. Momma, K. & Izumi, F. VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data. *Journal of applied crystallography* **44,** 1272–1276 (2011).
45. 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).
46. 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.

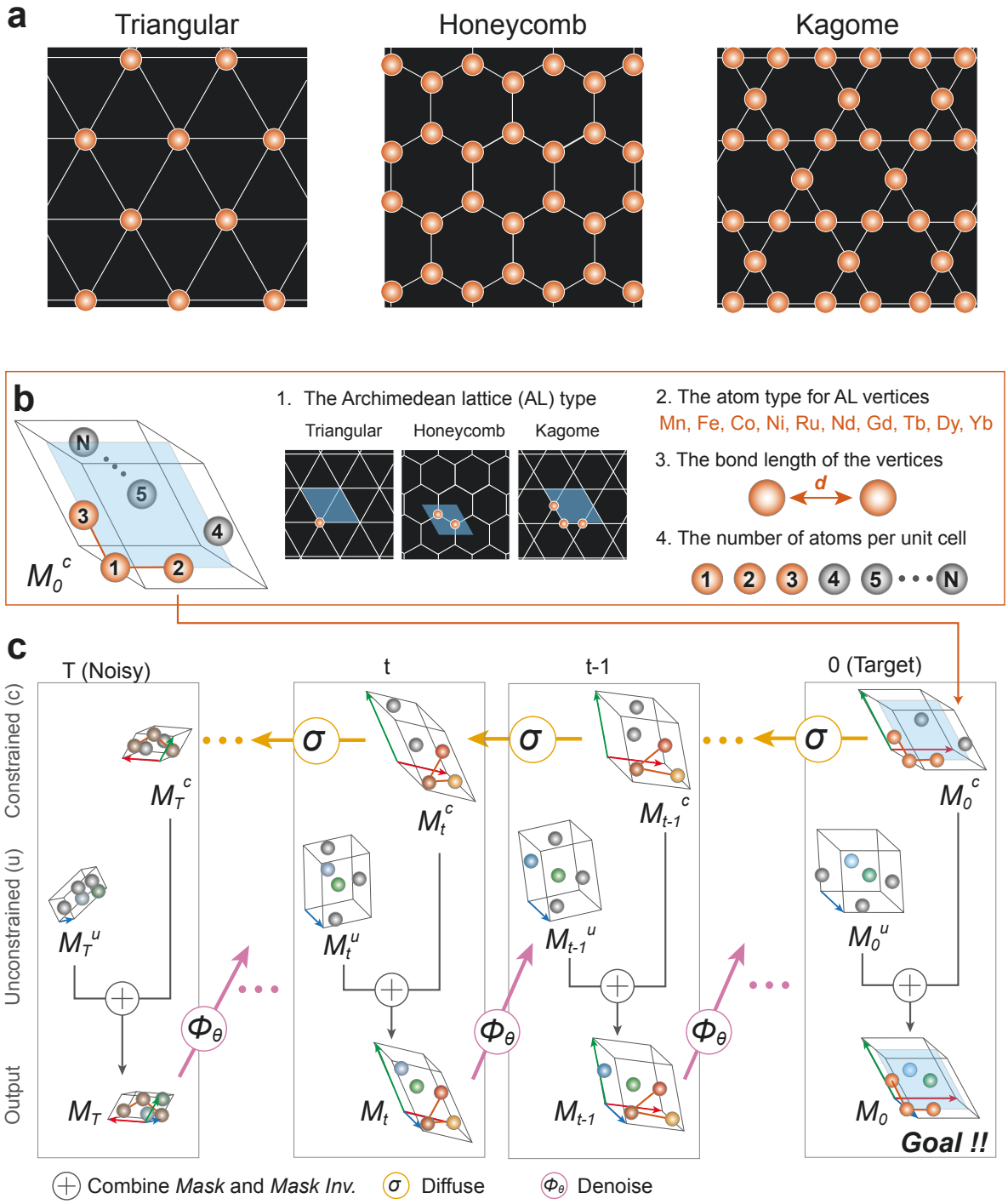

Figure 1. Schematic overview of material generation with geometric patterns as constraints. a. Three primary classes of Archimedean lattices with hexagonal unit cells: triangular, honeycomb, and kagome. **b.** Guideline for structure initialization for diffusion model, with magnetic atoms at Archimedean 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 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 with a fair chance to maintain stability.

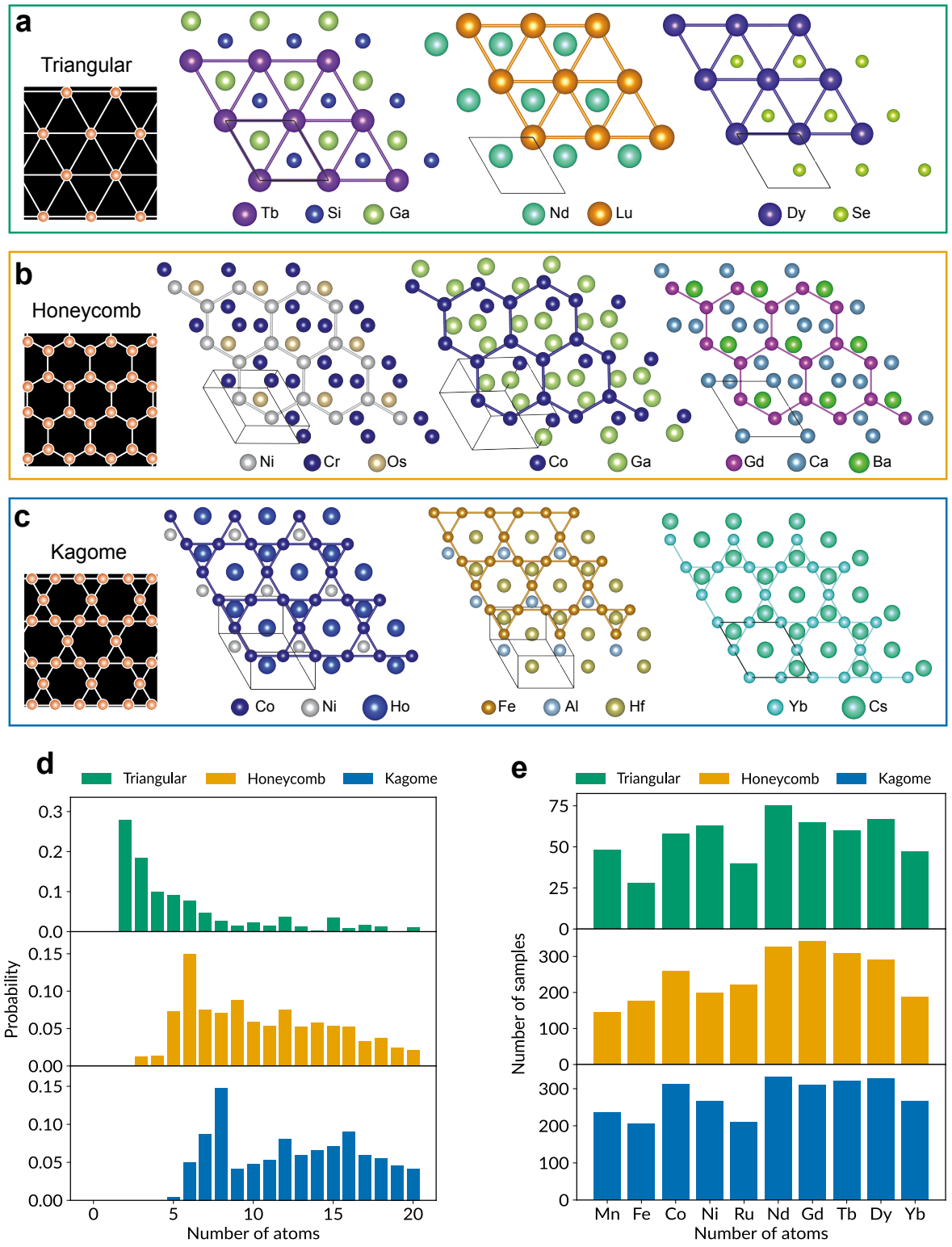

Figure 2. Generated materials with three primary types of archimedean lattices. Archimedean lattice patterns and generated material structures are displayed for **a.** Triangular, **b.** Honeycomb, and **c.** Kagome lattices. **d.** The sampling profile of the number of atoms per unit cell N, generated 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 of the primary geometrical patterns.

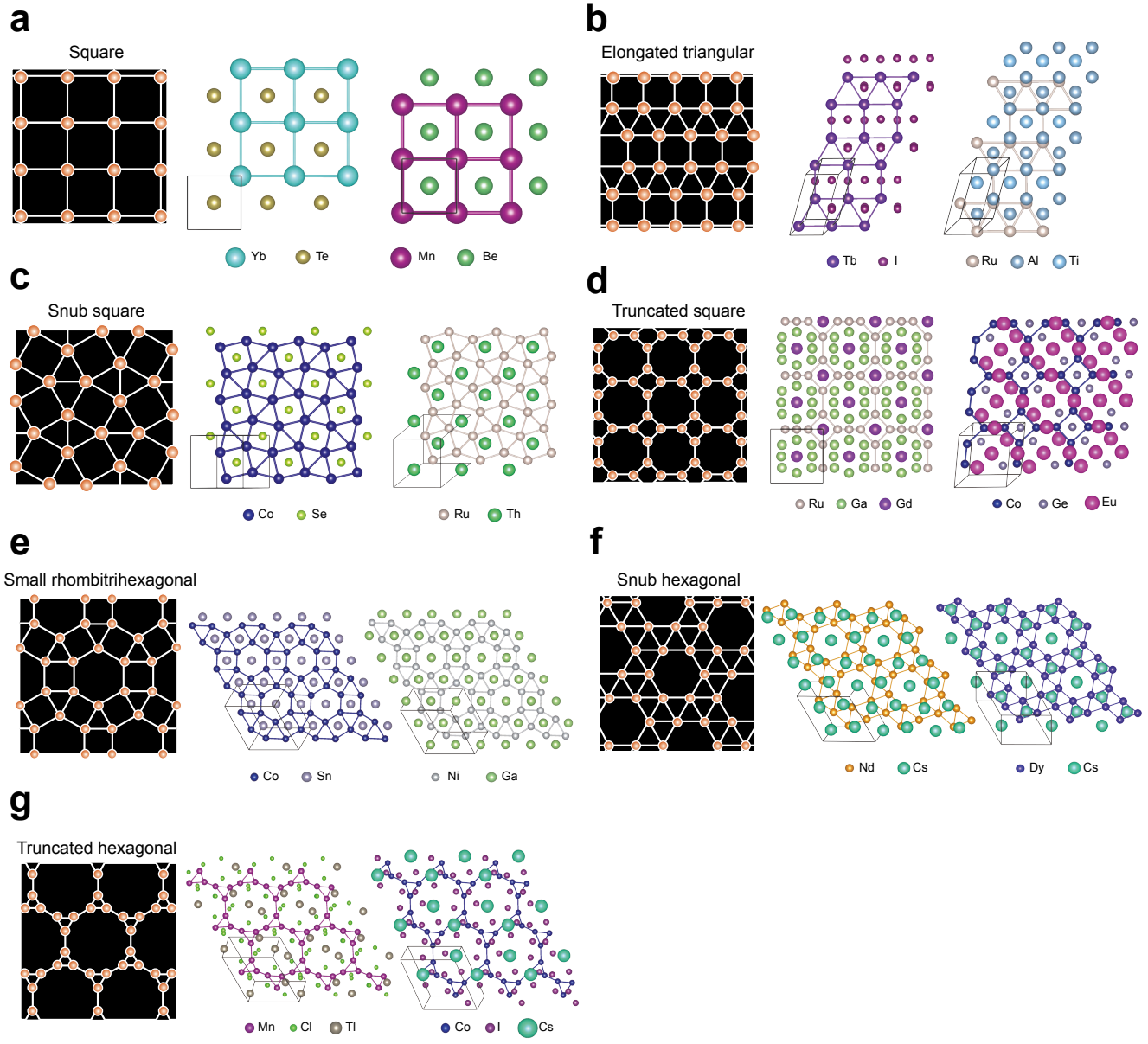

Figure 3. Generated materials with other Archimedean lattice structures. Materials examples covering the rest of Archimedean lattices are presented, with **a.** Square **b.** Elongated triangular **c.** Snub square **d.** Truncated square **e.** Small rhombitrihexagonal **f.** Snub hexagonal **g.** Truncated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

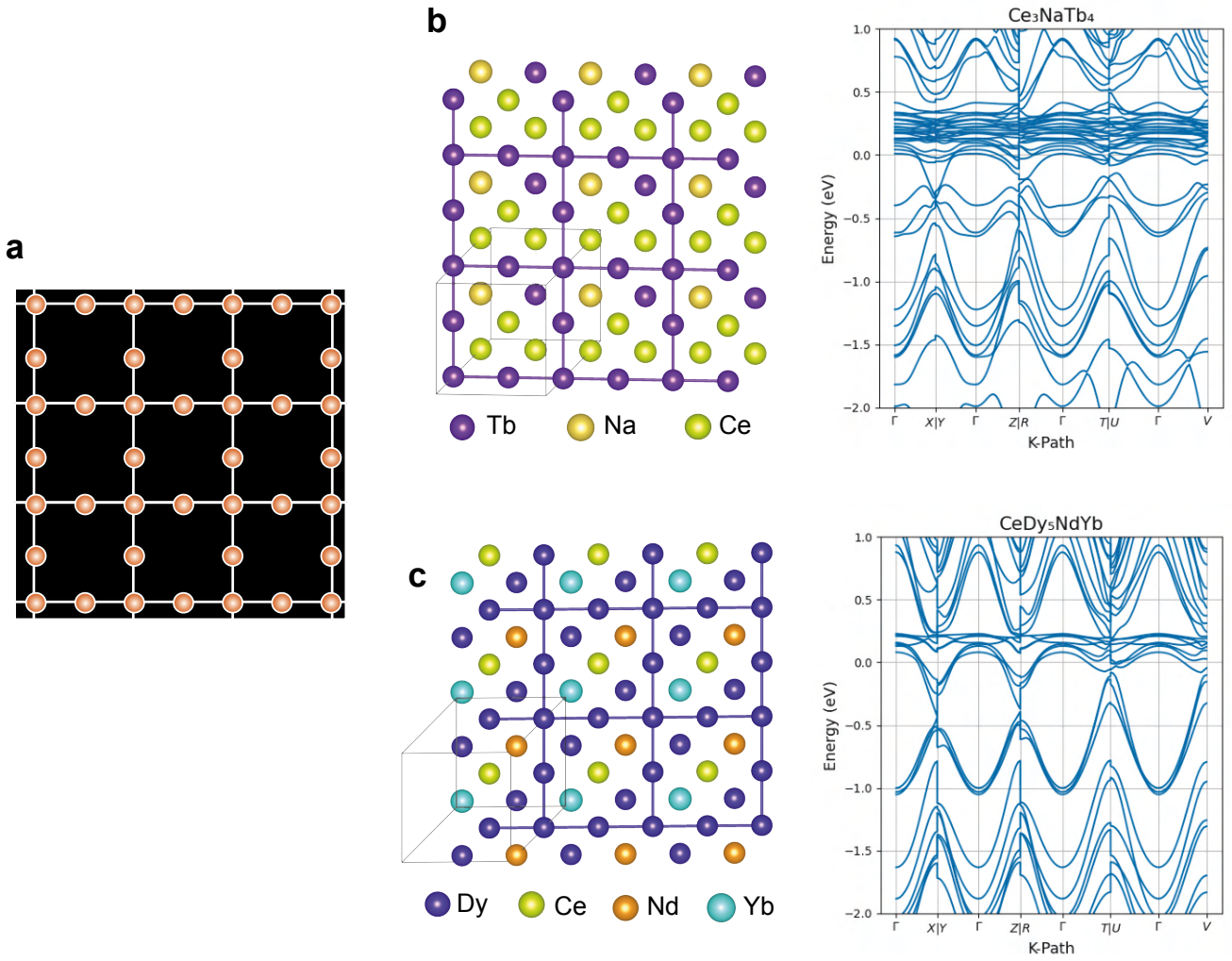

Figure 4. Generated materials of a Lieb-like lattice. a. The Lieb lattice pattern that we integrate into the generated structures. The supercell of the Lieb-like lattice materials and the flat band structures of **b.** Ce_3NaTb_4 and **c.** $CeDy_3NdYb$. We plot the band structures by setting the Fermi level E_F to 0 eV, and the flat bands in both examples are slightly (0.1 – 0.2 eV) above the Fermi level.

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

Ryotaro Okabe[1,2,*], Mouyang Cheng[1,3,4], Abhijatmedhi Chotrattanapituk[1,5], 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: chengy@ornl.gov
[***]e-mail: mingda@mit.edu

Contents

I Archimedean and Lieb lattices as geometrical pattern constraints

Archimedean lattices (ALs)[1,2], commonly referred to as Archimedean tilings, are distinctive for their planar, uniform tiling, where each vertex configuration is identical. Unlike regular tilings, which utilize only one type of regular polygon, Archimedean lattices incorporate multiple types of regular polygons but are arranged uniformly at each vertex. A key feature of ALs is their vertex-transitivity, which allows any vertex to be mapped to any other through a series of reflections, rotations, and translations, thus preserving the arrangement's overall symmetry.

There are exactly 11 types of ALs, each uniquely defined by the types and sequences of polygons that meet at each vertex. Each AL can be described both as a descriptive name and the numerical name called by the list of the polygons surrounding one vertex. These include the Triangular (3^6), Honeycomb (6^3), and Kagome $(3,6,3,6)$ lattices, which are composed of triangles and hexagons in different configurations. The Square lattice (4^4) consists solely of squares. The Elongated triangular ($3^3,4^2$) and Snub square ($3^2,4,3,4$) lattices mix triangles and squares in varied layouts. Other forms include the Truncated square ($4,8^2$), Small rhombitrihexagonal $(3,4,6,4)$, Snub hexagonal ($3^4,6$), Truncated hexagonal ($3,12^2$), and the Great rhombitrihexagonal $(4,6,12)$, which involve combinations of squares, hexagons, and dodecagons, each offering complex geometric arrangements.

The Lieb lattice is a unique two-dimensional lattice structure characterized by its three sites per unit cell configured in a square shape. The vertices of this lattice are positioned at each corner and at the midpoint of each edge, forming a bipartite lattice. This specific arrangement allows the lattice to be divided into two interpenetrating sublattices, where each site on one only interacts with sites on the other sublattice. The configuration of the Lieb lattice is particularly significant in research areas focusing on optical, magnetic, and transport properties due to its potential for facilitating unusual localized states. These attributes make the Lieb lattice a valuable model in theoretical physics and the practical development of materials with tailored electronic properties.

Table S1 lists the characteristics of Archimedean and Lieb-like lattices. It covers the header that we used for giving file names to each output material, the property of the unit cell, the number of nodes forming AL per unit cell N^c, and the fractional coordinates. Here, we write six lattice parameters as l_1, l_2, l_3, α, β, γ. SCIGEN imposes constraints for the lattice parameters that reflect the AL structures (l_1, l_2, γ), while there are no constraints on the other lattice parameters (l_3, α, β). We also included the constants $k_{latt} = l_1/d^c$ (or l_2/d^c) as the ratio of lattice vector length (l_1, l_2) to the bond length between neighboring vertices (d^c), indicating the relative size of the AL unit cell. Figure S1 visualizes all types of AL and a Lieb-like lattice. The unit cell area is highlighted with blue areas, and the red vertices represent the required positions within the unit cell.