

Discovery of 2D Materials using Transformer Network-Based Generative Design

Rongzhi Dong, Yuqi Song, Edirisuriya M. D. Siriwardane, and Jianjun Hu*

Two-dimensional (2D) materials offer great potential in various fields like superconductivity, quantum systems, and topological materials. However, designing them systematically remains challenging due to the limited pool of fewer than 100 experimentally synthesized 2D materials. Recent advancements in deep learning, data mining, and density functional theory (DFT) calculations have paved the way for exploring new 2D material candidates. Herein, a generative material design pipeline known as the material transformer generator (MTG) is proposed. MTG leverages two distinct 2D material composition generators, both trained using self-learning neural language models rooted in transformers, with and without transfer learning. These models generate numerous potential 2D compositions, which are plugged into established templates for known 2D materials to predict their crystal structures. To ensure stability, DFT computations assess their thermodynamic stability based on energy-above-hull and formation energy metrics. MTG has found four new DFT-validated stable 2D materials: NiCl_4 , IrSBr , CuBr_3 , and CoBrCl , all with zero energy-above-hull values that indicate thermodynamic stability. Additionally, GaBrO and NbBrCl_3 are found with energy-above-hull values below 0.05 eV. CuBr_3 and GaBrO exhibit dynamic stability, confirmed by phonon dispersion analysis. In summary, the MTG pipeline shows significant potential for discovering new 2D and functional materials.


been successfully synthesized.^[5] The initial isolation of individual graphene sheets, proving the existence of 2D systems, paved the way for the identification of numerous 2D materials exhibiting remarkable superconducting,^[12] electronic,^[13] magnetic,^[14] and topological properties.^[15] Beyond serving as platforms for investigating reduced-dimensional system behaviors, 2D materials hold immense potential for diverse applications in optoelectronics,^[16] catalysis,^[17] and the energy sector.^[18] The research endeavor has predominantly focused on systems that possess bulk counterparts, manifesting as anisotropic crystals with layers interconnected by van der Waals forces. Notably, graphene and graphite stand out as prominent instances. The feeble interlayer interactions in these systems facilitate the inherent structural segregation of 2D subunits in the crystals, enabling mechanical or liquid-phase exfoliation.

Currently, there are three commonly used computational approaches for generating 2D materials: the top-down exfoliation method starts with bulk material and

1. Introduction

Two-dimensional (2D) materials have garnered substantial interest as promising functional materials with a broad range of applications, owing to the intriguing fundamental physics that emerges in reduced dimensions.^[1] Numerous studies have been dedicated to the systematic exploration and synthesis of functional 2D materials.^[2–8] Endowed with exceptional and customizable properties, 2D materials hold significant promise across semiconductor, energy, and health-related applications.^[9,10] Following the Nobel Prize-winning discovery of graphene in 2010,^[11] a simple 2D carbon structure with intricate and appealing physics, only a limited number of distinct 2D materials have

exfoliates to make it thinner and peels the layers to obtain 2D materials; the bottom-up approach instead starts with existing 2D materials and uses element substitution to generate new materials. The third one is the de novo structure generation approach^[3] based on deep learning generative models such as crystal diffusion variational autoencoder (CDVAE).^[19] To get new 2D materials through the exfoliation method, we need to judge whether a 3D bulk material is layered so that it can be exfoliated. The layer screening process first checks the distance between atoms to identify whether these atom pairs are bonded. It then calculates the bonded atom clusters both in a $3 \times 3 \times 3$ supercell and the unit cell. If these $3 \times 3 \times 3$ supercells can be conceptualized as three layers of 2D supercells, with dimensions

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202300141>.

© 2023 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202300141

R. Dong, Y. Song, J. Hu
Department of Computer Science and Engineering
University of South Carolina
Columbia, SC 29201, USA
E-mail: jianjunh@cse.sc.edu

E. M. D. Siriwardane
Department of Physics
University of Colombo
Colombo 00300, Sri Lanka

of $3 \times 3 \times 1$, the structure is designated as layered.^[20] The exfoliation of 2D materials is theoretically achieved by extracting a single cluster from the conventional unit cell of the screened layered bulk structures. In the element substitution method, the elements of the periodic table are systematically grouped based on their column numbers. Elements sharing the same column (group) number share the same number of electrons in their outermost orbit, while elements within the same row (period) have a common number of electronic layers. Consequently, elements within the same group or neighboring groups exhibit similar chemical properties. The substitution method starts with the structure of a known 2D material and replaces one or more elements in this material with other elements either in the same group or its neighbor elements.

Both the element substitution method and the de novo generation method start with known 2D crystal structures. Currently, there are several open-source 2D material databases generated through exfoliation, substitution, or de novo generation methods. The computational 2D materials database (C2DB)^[5,21] uses both exfoliation and substitution methods to organize a wealth of computed properties for 4038 (checked in October 2022) atomically thin 2D materials. Notably, the materials encompass both experimentally validated and previously unexplored structures. These have been methodically generated by systematically decorating diverse 2D crystal lattices. Derived exclusively from the exfoliation method, MC2D^[6] selects 5619 compounds with layered characteristics out of 108 423 unique 3D compounds with established experimental records (as of October 2022). These selections are based on robust geometric and bonding criteria. Through high-throughput computations leveraging van der Waals density functional theory (DFT), validated against experimental structural data, and random phase approximation binding energy calculations, 1825 compounds primed for exfoliation have been identified. The 2D materials encyclopedia (2DMatPedia) database^[2,22] systematically examines bulk materials from the Materials Project database, pinpointing layered structures through a topology-based detection algorithm, and virtually exfoliating them into monolayers. Novel 2D materials emerge by introducing chemical substitutions of elements within established 2D materials, replacing them with elements from the same group in the periodic table. The current iteration of the 2DMatPedia database (as of December 2022) comprises 6351 materials. Among these, 2940 are derived via exfoliation of existing layered materials (top-down approach), 3409 through chemical substitutions in 2D materials (bottom-up approach), and 2 through avenues not conforming to either top-down or bottom-up methodologies. The bottom-up approach commences with the 35 unary and 755 binary compounds obtained from the top-down approach, limiting substitutions to elements within the same column. Leveraging 22 distinct 2D crystal prototypes and 52 chemical elements from the periodic table, the virtual 2D materials database (V2DB)^[4] employs a brute-force substitution technique to construct a systematic library featuring over 72 million 2D compounds. Subsequent filtering steps involving symmetry, neutrality, and stability assessments narrow down the selection to 316 505 presumably stable 2D materials.

Materials cloud^[23] offers a practical and straightforward approach for evaluating the potential of exfoliating 3D compounds into 2D layers. This multistep procedure begins with

a preliminary screening of layered structures, using geometric criteria that only consider the atomic positions within the structure. After that, a random forest classifier is employed to determine whether the resulting structures can be exfoliated or possess high binding energy. Friedrich et al.^[24] introduced a novel collection of non-van der Waals 2D materials by leveraging data-driven concepts and extensive computations. Through careful filtering of the AFLOW-ICSD database based on the structural prototypes of the experimentally realized Fe₂O₃ and FeTiO₃ systems, they identified 8 binary and 20 ternary 2D material candidates. The most recent advancement in 2D material generation involves the utilization of deep learning generative models. Lyngby et al.^[3] harnessed a CDVAE^[19] to produce novel 2D structures characterized by high chemical and structural diversity. These newly generated structures exhibit formation energies that mirror those of the training set. The researchers also employed an element substitution method to derive additional potential 2D materials based on the newly generated structures. In total, they generated 11 630 predicted new 2D materials, comprising 3073 structures from CDVAE and 8599 structures resulting from element substitutions applied to the CDVAE-generated ones. Notably, 2004 of these newly generated 2D candidates are within a 50 meV range of the convex hull, implying their potential for synthesis. To comprehensively capture the structural characteristics of 2D and quasi-2D materials, Wang et al.^[25] developed an innovative 2D structure search module within the CALYPSO code. This module is based on the 2D particle swarm optimization algorithm, which enables atomic coordinates relaxation in the perpendicular direction. Wang et al. successfully predicted a new family of layered structures, B_xN_y, with varying chemical compositions.

Here, we propose a computational pipeline, material transformer generator (MTG), for the generative discovery of new 2D materials (and other crystal materials). Our method is based on combining a 2D composition generator trained with known 2D material compositions, two template-based crystal structure predictors, two machine learning potential-based structure relaxers, and DFT relaxation. Extensive experiments show that our MTG pipeline can be used to discover a large number of hypothetical 2D materials.

2. Experimental Section

Figure 1 shows the framework of our MTG pipeline for 2D material generation. We collected known 2D formulas and their structures from open datasets C2DB, MC2D, 2DMatPedia, and V2DB. We then train a set of blank language models for materials (BLMM) composition generators with known 2D formulas to generate new 2D formulas. Next, we use known 2D structures as templates for structure prediction of these candidate 2D formulas using two crystal structure prediction algorithms TCSP and CSPML. TCSP is a template-based crystal structure prediction algorithm based on oxidation state patterns. CSPML is a machine learning-based crystal structure prediction method using a machine learning model to select templates. For a given new 2D formula such as SrTiO₃, both models will first select all template structures with prototype ABC₃, but they are very different when sorting all these templates. TCSP calculates the

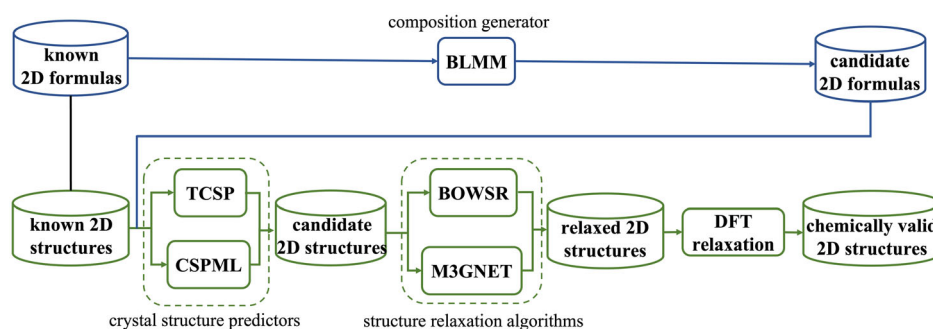


Figure 1. Architecture of our material transformer generator (MTG) pipeline. BLMM^[28] is a transformer neural network-based composition generator. TCSP^[55] and CSPML^[37] are template-based crystal structure prediction algorithms, and BOWSR^[38] and M3GNET^[39] are machine learning potential-based structure relaxing algorithms. DFT relaxation is a first-principles calculation method.

element mover distance score and element oxidation states, which focus on element distance. However, CSPML selects candidates using structural similarity. This structural similarity measure exclusively relies only on the topological attributes of atomic coordinates and without any data regarding the elemental composition. After choosing the appropriate templates and generating new 2D structures, we use two machine learning potential-based relaxation algorithms to optimize the structures. The first one is BOWSR, a Bayesian optimization algorithm with symmetry relaxation. The second one is M3GNET, which uses materials graph neural networks with 3-body interactions as an energy estimation model for structure relaxation. The BOWSR algorithm relaxes each structure by changing the independent lattice parameters and atomic coordinates to obtain lower potential energy. During relaxation, the M3GNET algorithm takes all atom coordinates and the 3×3 lattice matrix into consideration. The attributes of the bond, atom, and state are updated in order. For each attribute update, all previous attributes of these three parameters are considered. After all these operations, for all generated 2D formulas, we obtain near-equilibrium relaxed structures. After the fast machine learning potential-based relaxation, we further apply the DFT-based relaxation procedure to optimize the structures. Finally, we calculate the formation energy and e-above-hull energy of top structures to evaluate the final performance.

2.1. BLMM: Transformer-Based 2D Material Composition Generation

The material composition can be mapped into a sequence generation problem as a composition such as SrTiO_3 can be conveniently expanded into a specific sequence (e.g., Sr Ti O O O) sorted by the electronegativities of the elements. The BLMM model is a composition generator built on the latest transformer deep neural network models, shown to be excellent in sequence learning and sequence generation. By adopting the self-attention mechanism to weigh the significance of all tokens in the input sequence, the transformer model^[26] has been proved as state-of-the-art in the fields of natural language processing and computer vision. Based on the traditional transformer, Shen et al.^[27] proposed a blank language model (BLM) that could generate sequences by dynamically creating and filling in blanks. Our

BLMM composition generator^[28] is developed based on the BLM blank-filling model. All material formulas can be rewritten as sequences (e.g., SrTiO_3 to Sr Ti O O O) composed of a vocabulary with 118 or fewer elements. We then train a BLMM-based 2D composition generator using our 2D materials dataset. The architecture of the BLMM algorithm is shown in Figure 2. The process of generation initiates with an initial empty space and concludes when no further empty spaces remain. At each iterative step, the model identifies an empty space, forecasts an element, and subsequently substitutes the empty space with predicted element, along with any adjacent empty spaces as required. Through the repetition of this cycle of selecting and populating empty spaces, a single empty space can be extended to encompass any quantity of elements. Then we use this well-trained BLMM model to generate new 2D compositions. After getting the generated compositions, we first remove duplicate compositions that are already included in known 2D datasets and then take the nonredundant formulas as our new 2D material candidates to be fed to the step of template-based 2D material structure prediction.

2.2. Template-Based 2D Material Structure Prediction

At present, the challenge of predicting generic crystal structures remains unresolved, even though global optimization-based algorithms like USPEX and CALYPSO have demonstrated success in solving structures for smaller systems. Simultaneously, it is noteworthy that similar to bulk materials,^[29–31] the majority of existing 2D material structures can be classified into a limited set of structural prototypes. This observation suggests that their structures can be derived through template-based elemental substitution.

After composition generation and duplicate checking, we obtained a 2D material composition candidates dataset. To gain the probable structures of all candidates, we use two different template-based element substitution methods to select the most similar structure template and then use element substitution to generate target structures. The crystal structure generated by these two methods has the same lattice parameters and atomic coordinates as the template structure and needs to be further relaxed.

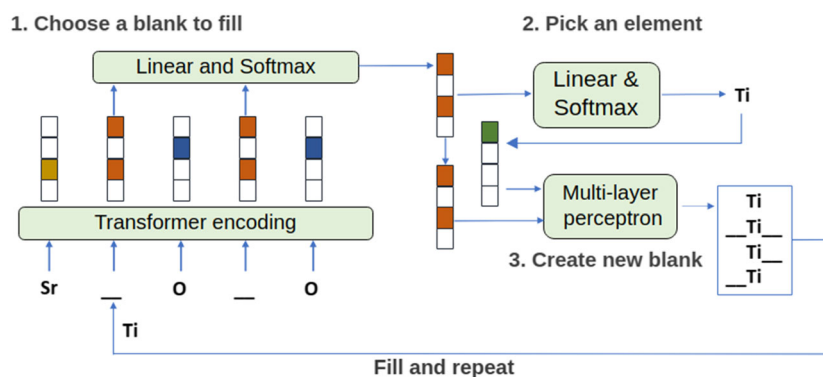


Figure 2. The architecture of BLMM.^[28] For a given formula sequence with blanks, we first use transformer encoding to get the embedding for every element and the blank, and then choose one single blank to predict the most suitable element for this blank based on the context information. After that, we use this predicted element and possibly adjoining blanks to fill the original blank. These choosing and filling process are repeated until there is no blank left.

2.2.1. TCSP Is a Template-Based Crystal Structure Prediction Algorithm

The architecture of the TCSP algorithm is shown in **Figure 3a**. For a given 2D material formula candidate, the TCSP algorithm first searches all known 2D material structure templates that share the same composition prototype as this formula (e.g., SrTiO_3 has prototype ABC_3). The element mover distance (ElMD)^[32] serves as a metric for gauging the compositional resemblance between the query formula and various compositions derived from potential template structures. Subsequently, the five structures with the smallest compositional disparities are deemed as prospective candidate templates. Among these candidates, we employ the Pymatgen^[33] toolkit to verify if their oxidation states match those of the query formula. Templates exhibiting identical oxidation states are incorporated into the ultimate template roster. If such templates are absent, all five top structures are retained as final templates. To eliminate redundant template structures, we employ Pymatgen's StructureMatcher module. Within each structural cluster, only one representative structure is retained, significantly curbing the number of akin structure templates. Following the assignment of templates for the query formula, the algorithm proceeds to systematically enumerate potential element substitution pairs between the query and template formulas. It is plausible for one template to require multiple element pair substitutions to yield

the target formula. A score for replacement quality is subsequently computed by summing the ElMD values for all possible element pair substitution configurations. This score reflects the similarity of the substitution element pairs. Lower scores denote greater similarity and, consequently, higher quality.

2.2.2. CSPML Is a Machine Learning-Based Crystal Structure Prediction Algorithm

CSPML relies on metric learning^[34] for crystal structure prediction, which can select template structures from known structure databases with high similarity to the given composition. Metric learning uses a binary classifier to distinguish whether two given compositions have similar structures as defined by a similarity threshold of local structure order parameters (LoStOps).^[35] The architecture of the CSPML algorithm is shown in **Figure 3b**. For a given 2D formula, CSPML first restricts the candidates to structures with the same compositional ratio (e.g., SrTiO_3 has a composition ratio of 1:1:3). The compositional descriptor of the query formula and templates is then calculated by XenonPy.^[36] XenonPy provides 58 physicochemical features for each element. For a given composition, by calculating the weighted mean, weighted sum, weighted variance, min-pooling, and max-pooling of all elements, XenonPy generates a 290-dimensional (58×5) descriptor vector. A traditional multilayer perceptron is used to figure out how similar the template structure and the query

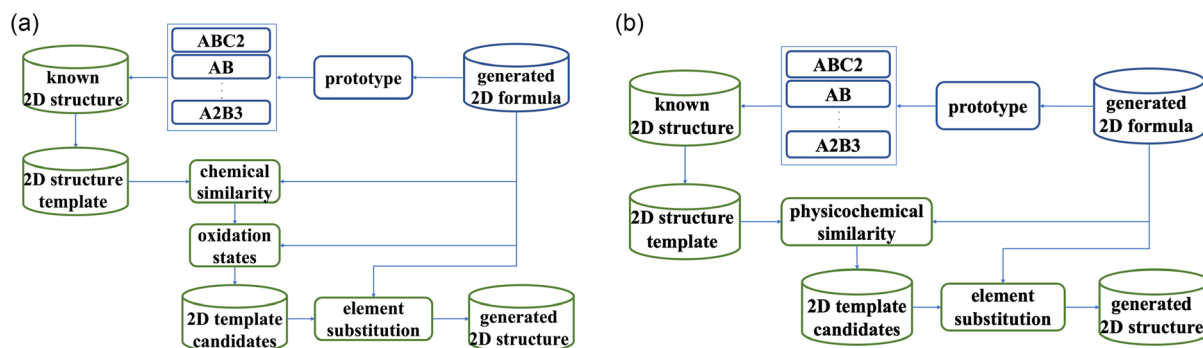


Figure 3. Template-based CSP algorithms. a) TCSP architecture.^[55] b) CSPML architecture.^[37]

formula are. The absolute difference between two compositional descriptors is used as the input. We pick the top five template structures with the biggest similarity scores with the query formula as the template candidates. The structure of the query formula is then generated by replacing the atoms in the templates with atoms in the query composition. When two or more elements have the same composition ratio, the substitution element pairs are not uniquely determined. In such cases, we substitute a pair of elements with the most similar physicochemical properties, as described in the study of Kusaba et al.^[37]

2.3. Structure Relaxation

Predicting novel stable crystal structures and their properties with precision is a pivotal objective in the realm of computation-guided materials discovery. Although *ab initio* methods like DFT have achieved remarkable success in this domain, their considerable computational overhead and limited scalability have curtailed their widespread utilization across diverse chemical and structural spaces. To surmount this constraint, machine learning has emerged as an innovative paradigm for crafting efficient surrogate models capable of predicting material properties on a larger scale. In this article, subsequent to acquiring foundational structures via template-based element substitution techniques, we employ and contrast two distinct machine-learning potential-based methods for the relaxation of structures.

2.3.1. BOWSR: Bayesian Optimization with Symmetry Relaxation Algorithm

BOWSR^[38] is a graph neural network-based structure relaxation algorithm that uses Bayesian optimization as an optimizer. Bayesian optimization serves as an adaptive technique for globally optimizing functions. In the context of crystal structure relaxation, the objective function to be optimized corresponds to the potential energy surface, characterizing the energy of the crystal structure. Throughout the relaxation procedure in the BOWSR algorithm, constraints are imposed on the lattice symmetry and the atomic Wyckoff positions. The algorithm permits modifications solely to the independent lattice parameters and atomic coordinates. The BOWSR algorithm sets parameters for each structure based on these changeable, independent lattice parameters and atomic coordinates. The potential energy surface of each training observation is calculated by a graph neural network energy model MEGNET, which is trained with 12 277 stable structures with DFT-calculated formation energies. Bayesian optimization is then used to relax structures iteratively toward states with lower energies.

The geometry relaxation of a structure of N atoms requires optimizing $3N + 6$ variables, 3 fractional coordinates for each atom, and 6 lattice parameters in total. By keeping the symmetry the same during the relaxation process, it can reduce the number of independent variables. New structures are generated by Bayesian optimization, which minimizes the formation energy, and the changed variables are then used as inputs to predict new energy. The previous step is then repeated multiple times until the formation energy reaches the lowest point or reaches the maximum number of iterations. The final structure is the BOWSR relaxation results.

2.3.2. M3GNET: Materials Graph Neural Networks with 3-Body Interactions

M3GNET^[39] presents a novel architecture for materials graph neural networks, incorporating 3-body interactions to predict formation energy. This approach amalgamates graph-based deep learning interatomic potential (IAP) with traditional IAPs' many-body features, coupled with the versatility of flexible graph material representations. The model takes position-included graphs as inputs, embedding atomic numbers and pair bond distances as graph features. Computation of three-body and many-body interaction atom indices and angles ensues through the many-body computation module. The graph convolution module then updates bond and atom information.

Diverging from previous materials graph implementations like MEGNET, M3GNET uniquely incorporates atomic coordinates and the 3×3 lattice matrix in crystals. These additions are indispensable for obtaining tensorial quantities, including forces and stresses, through autodifferentiation. Notably, M3GNET stands apart from BOWSR's graph neural network potential by being trained on both stable and unstable structures. The relaxation algorithm based on M3GNET diverges from BOWSR's Bayesian optimization. It employs the FIRE algorithm, a modified form of molecular dynamics with added velocity adjustments, adaptive time steps, and inertia, resulting in a rapid inertial relaxation engine. The exceptional capacity of the M3GNET-based relaxation algorithm to swiftly and precisely relax diverse crystal structures while predicting their energies renders it well-suited for expansive materials discovery endeavors.

2.4. DFT Calculations

We conducted first-principles calculations based on DFT using the Vienna *ab initio* simulation package (VASP)^[40–43] to optimize the candidate structures suggested by the machine learning models. We utilized projected augmented wave pseudo-potentials^[44,45] to handle electron-ion interactions, employing a 520 eV plane-wave cutoff energy. For the exchange-correlation functions, we employed the generalized gradient approximation-based Perdew–Burke–Ernzerhof method.^[46,47] Throughout the DFT calculations, an energy convergence criterion of 10^{-5} eV and a force convergence criterion of 10^{-2} eV Å⁻¹ were maintained. The Brillouin zone integration for unit cells was performed using Γ -centered Monkhorst–Pack k -meshes. Formation energies (in eV/atom) of the materials were determined using the formula in Equation (1), where $E[\text{Material}]$ represents the total energy per unit formula of the target structure, $E[A_i]$ is the energy of the i^{th} element in the material, x_i denotes the number of A_i atoms in a unit formula, and n represents the total number of atoms in a unit formula ($n = \sum_i x_i$).

$$E_{\text{form}} = \frac{1}{n} \left(E[\text{Material}] - \sum_i x_i E[A_i] \right) \quad (1)$$

The Pymatgen code^[33] was used to compute the energy above hull values of the materials with negative formation energies. The Phonopy code^[48] was used to study the phonon dispersion

of the materials, while density functional perturbation theory^[49] was employed to study the mechanical stability.

2.5. Evaluation Criteria

We use a series of performance metrics to evaluate our 2D material generation pipeline. To evaluate the BLMM 2D material composition generator, we calculate the validity, uniqueness, recovery rate, and novelty. Formation energy is used as an indicator to evaluate the template-based 2D structure generator and relaxer. To further verify the structures, we use VASP to calculate the energy above the hull.

2.5.1. Validity

For all formulas generated by the BLMM algorithm, we use semi-conducting materials by analogy and chemical theory^[50] to check whether they obey the charge neutrality and electronegativity (CNEN) rules.

2.5.2. Uniqueness

Uniqueness percentage is calculated by using the number of unique samples divided by the total generated samples. The uniqueness indicator shows the BLMM model's ability to generate diverse samples.

2.5.3. Recovery Rate and Novelty

To check the BLMM model's capability to generate novel materials, we calculate the recovery rate and novelty of generated formulas. The recovery rate shows the percentage of training samples that have been rediscovered. Novelty shows how many new samples have been generated.

2.5.4. Formation Energy

The way to evaluate the structure generation models is to check the stability of generated structures. For structures generated and relaxed through our pipeline, we calculate their formation energy using M3GNET.

2.5.5. Energy above the Convex Hull

The energy convex hull is formed based on established stable structures, as explored by Liu et al.^[51] Structures whose energies reside on the convex hull are considered thermodynamically stable, while those positioned above it tend to be either metastable or unstable. Among all structures exhibiting negative formation energies, we leverage the energy above the convex hull as an additional criterion for selecting more stable candidates.

2.6. Hyperparameters and Training

For 2D formula generation, each BLMM model trained on 2D datasets generates 100 000 samples. After generation, we use the TCSP and CSPML methods separately to generate structure

Table 1. Hyperparameters used in models.

TCSP		CSPML		BLMM	
Candidate	10	Candidate	10	Data workers	32
Top	5	Top	5	Max steps	200 000
Sort	ELMD	Sort	XenonPy	Max token	40 000
Filter	Ratio	Filter	Ratio	Vocab size	130
Filter	Oxidation			Max len	205
BOWSR		M3GNET			
Optimizer	Bayesian optimizer	Optimizer	FIRE		
Initial points	100	Force tolerance	0.1		
Iteration steps	100	Relax steps	500		
Seed	42				

candidates for these samples. BOWSR and M3GNET are then used to relax generated structures. **Table 1** shows the hyperparameters used in the BLMM, TCSP, CSPML, BOWSR, and M3GNET models. In the BLMM model, we use an element vocabulary with a size of 130, and the generated formula sequence length is limited to 205. The maximum number of tokens per batch is set to 40 000, and the number of training steps is set to 200 000. The candidate template structure numbers of both the TCSP and CSPML models are set to 10. Only the top five candidates, sorted by ELMD and XenonPy, respectively, can be used as real templates. Relax method BOWSR uses a Bayesian optimizer with initial points 1000, iteration steps 1000, and seed number 42. The M3GNET relax method uses FIRE^[52] optimizer with a 0.1 total force tolerance for relaxation convergence and 500 relax steps.

3. Results

3.1. Datasets

As shown in **Table 2**, our template-based 2D materials generation models are trained using the materials downloaded from the C2DB,^[5,21] MC2D,^[6] 2DMatPedia,^[2,22] and V2DB^[4] databases with a total of 328 719 formula samples and 12 214 structures. The C2DB dataset was initially generated by decorating an experimentally known crystal structure prototype with atoms chosen from a (chemically reasonable) subset of the periodic table. The MC2D dataset starts from experimentally known 3D compounds

Table 2. Open source datasets used in 2D material discovery.

Dataset	Formula	Structure	From	Exfoliation	Substitution
C2DB	4038	4038	Known 2D crystal structure prototype	N/A	All
MC2D	1825	1825	Known 3D compounds	All	N/A
2DMatPedia	6351	6351	Materials Project	2940	3409
V2DB	316 505	N/A	22 known 2D crystal prototypes	N/A	All

and finds 1825 compounds that are either easily or potentially exfoliated. The 2DMatPedia dataset is searched from the Materials Project database^[53] and uses exfoliate first to find possible 2D structures and new structures generated by exfoliation are then used as templates of element substitution. The generation of the V2DB dataset employs the brute-force element substitution method. This method generated 72 522 240 possible combinations of 2D materials and only 0.4% of these passed the symmetry, neutrality, and stability validation.

In this work, we separated these 2D samples into two datasets. The first one is an experimental dataset (exp2d for short) with 4023 formulas and corresponding structures from the C2DB and MC2D datasets. The second dataset (all2d for short) contains all the samples in the above-mentioned known 2D databases with a total of 302 174 unique formulas and 8019 structures.

3.2. Composition Generation Performance

We use the BLMM algorithm to generate new 2D material compositions based on two different datasets, the experimental 2D dataset, exp2d, and an all 2D dataset, all2d. For each dataset, we train a generation model using these formulas and then use these well-trained models to generate new formulas. We also use transfer learning to train BLMM models on the Materials Project database and then fine-tune these pretrained models using our two datasets, which are named all2d-transfer and exp2d-transfer, respectively. Four composition generation models are trained to generate 100 000 formulas separately. The generated results are shown in **Table 3**. Furthermore, to check whether generated formulas are chemically valid, we employ two filters to check their charge neutrality and electronegativity, this checking step is called CNEN for short. The results are shown in Table 3, 93.3%, 67.1%, 93.2%, and 67.0% generated compositions passed the CNEN check. Besides, we remove duplicate composition in each model, and they achieve 65.8%, 26.5%, 69.4%, and 21.0% uniqueness respectively. We also calculate the recovery rate and novelty of these generational models. As we can see in Table 3, their recovery rates are 0.6%, 9.2%, 0.6%, and 11.2% while the novelties are 63.7%, 24.6%, 67.5%, and 18.8%. This evaluation demonstrates that our methods have the ability to generate stable and innovative compositions that form stable 2D structures. Since the exp2d dataset is smaller than the all2d dataset, the BLMM model trained with the exp2d dataset has much fewer samples to learn from and use for interpolation. Thus, the BLMM model trained with the exp2d dataset has lower validity, uniqueness, and novelty percentages than the BLMM model trained with the all2d dataset. However, the recovery rate of the BLMM model trained with fewer samples is higher because

Table 3. Composition generation results.

	all2d	exp2d	all2d-transfer	exp2d-transfer
Generated results	100 000	100 000	100 000	100 000
Validity	93.3%	67.1%	93.2%	67.0%
Uniqueness	65.8%	26.5%	69.4%	21.0%
Recover rate	0.6%	9.2%	0.6%	11.2%
Novelty	63.7%	24.6%	67.5%	18.8%

the interpolation space is smaller and the interpolated results generated by the BLMM model are more likely to be the same as the training samples. The composition generator pretrained with the Materials Project database and fine-tuned with the all2d dataset achieves higher uniqueness and novelty compared with the generator only trained with the all2d dataset. However, due to the lack of sufficient transfer learning samples, the BLMM model fine-tuned with the exp2d dataset has lower uniqueness and novelty compared with the BLMM model trained with the exp2d dataset.

3.3. Distribution of Generated Candidate 2D Compositions

To evaluate the composition generation performance of BLMM, we depict the element distribution of compositions in the 2DMatPedia dataset alongside our generated samples, as shown in **Figure 4**, where (a) and (b) show the element frequency in compositions in the 2DMatPedia dataset and the BLMM model generated dataset, respectively. Here, we take the BLMM model trained with the exp2d dataset as an example. The top 5 most frequent elements in the 2DMatPedia dataset are O, S, F, Te, and Cl. Out of the total 6351 formulas, element O has appeared 1642 times or about 26% of the 2DmatPedia dataset. The occurrences of the elements S, F, Te, and Cl are 653, 598, 586, and 584, respectively. The top 5 most frequent elements in our generated dataset are Se, O, S, Te, and Cl. The element Se has shown 13 294 times out of the whole set of 67 103 formulas, which is about 20% of the whole generated dataset. The elements O, S, Te, and Cl are observed 12 591, 11 440, 9546, and 8829 times, respectively. Notably, it becomes evident that four out of the top five most frequently occurring elements in both datasets are identical. Furthermore, six of the top ten most common elements also align between the datasets. This observation strongly suggests that the BLMM generator has adeptly captured and learned the fundamental compositional preferences intrinsic to 2D materials.

We also analyze the distribution of element pairs in the known 2D dataset and our generation results. To count the frequencies of element pairs, we take each of the possible 2-element combinations from the element set and count the number of compositions that contain this pair (we ignore the order of the two elements in the pair). The distribution of the top 50 element pairs in the 2DMatPedia and our generation datasets are shown in **Figure 5a,b**, respectively. The top 5 most frequent element pairs in the 2DMatPedia dataset are H-O, P-O, Li-O, V-O, and Bi-O. However, only the H-O element pair is shown in the top 5 of our generation results. The other 4 element pairs in our generation results are C-O, N-O, Cl-O, and H-C. These two datasets only share 2 common element pairs in the top 10 most frequent ones.

To verify whether our newly generated compositions share a similar distribution with the known 2D compositions, we use the t-distributed stochastic neighbor embedding^[54] technique to map the one-hot matrix of compositions to their corresponding formation energy. Each point in **Figure 6** corresponds to one formula and the colors represent the formation energy levels. Figure 6a shows the formation energy distribution of the 2DMatPedia samples. It can be found that most samples have formation energy

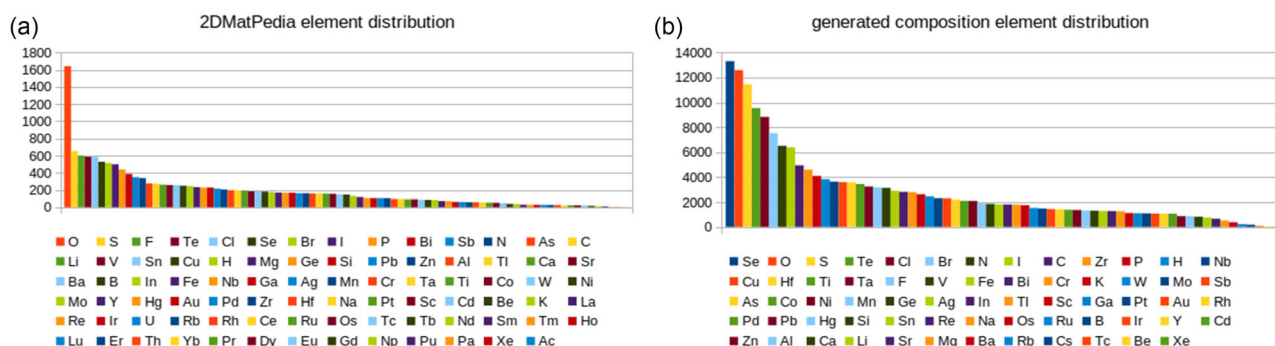


Figure 4. Elements distribution in training and generating samples. a) Element distribution in the 2DMatPedia dataset. b) Element distribution in BLMM model-generated samples.

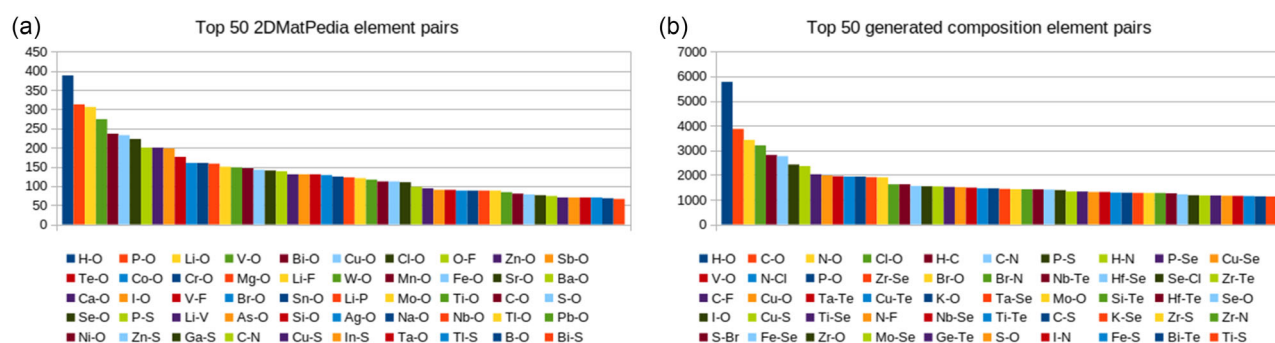


Figure 5. Top 50 element pairs distribution in training and generating samples. a) Element pairs distribution in 2DMatPedia dataset. b) Element pairs distribution in generated samples by BLMM model trained with exp2d dataset.

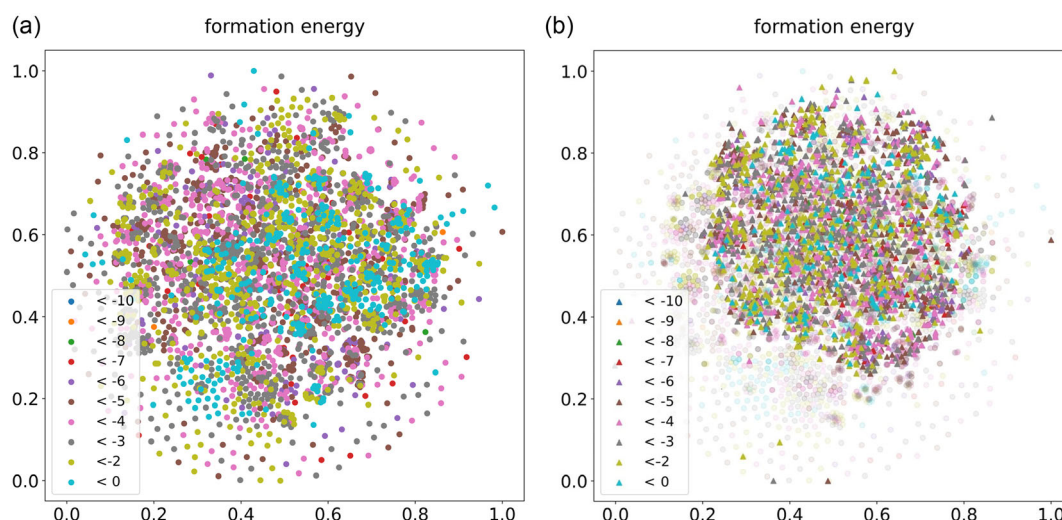


Figure 6. Formation energy distributions in the training (2DMatPedia) and generated samples (BLMM-exp2d). a) Formation energy distribution in the 2DMatPedia dataset. b) Formation energy distribution in generated samples.

between 0 and -3 eV atom^{-1} . Figure 6b displays the formation energy distribution of our generated compositions by BLMM-exp2d, which are developed through the following pipeline: first, we train the BLMM model using the exp2d dataset and generate candidate structures using the TCSP method and then relax these structures using the

M3GNET model; thirdly, the formation energies of these structures are predicted by the M3GNET method. As the BLMM model is trained by adding and filling in blanks in existing materials, it has a strong interpolation capability when generating new samples. Therefore, the newly generated samples are always located around known samples.

3.4. Stability Distribution of Generated Samples

Another way to check the quality of samples generated by our pipeline is to measure their formation energies and compare their distribution to that of the training set. We use formation energies predicted by the ML potentials M3GNET of both training samples and generated results.

We first check the formation energy distribution of a special material family AB_2 , which is the most frequent prototype in all existing 2D datasets: C2DB, MC2D, and 2DMatPedia. There are 1288 AB_2 samples in the exp2d dataset and 1928 AB_2 samples in our generated structures. The distributions of energies of these two datasets are shown in Figure 7a.

Next, we check the structure-based formation energy distribution of the whole exp2d dataset samples and compare it with those of our generated samples. Figure 7b shows that these two sets of structures have similar formation energy, which means that new structures generated through our MTG pipeline are of high quality.

Figure 7c compares the energy distribution of samples in the exp2d training set with samples generated throughout the MTG pipeline. These compositions are generated by the BLMM model trained with four different datasets, as introduced in Section 3.1. The energy distributions of formulas generated by BLMM trained with the all2d dataset and trained with the MP dataset but finetuned using the all2d dataset are very similar to each

other. The same situation in BLMM models trained and fine-tuned with the exp2d dataset.

3.5. Discovery Results

Our MTG pipeline generates 148 563 candidate 2D formulas. For each formula, we generate 10 structures using TCSP and CSPML and then we do M3GNET-based structure relaxation and pick the top 1 structure with the lowest energy. Then, we conduct DFT-based relaxation to generate final structures.

Figure 8 shows how we generated new structures based on specific template structures and how to relax newly generated structures to make them more stable. For formula $K_4Cr_2Ge_4Te_2$ generated by the BLMM algorithm, we first select the structure templates for predicting its crystal structure. As shown in Figure 8a, TCSP picked $Na_4Ti_2S_4O_2$, a layered material, as the template structure. Figure 8b is then created through the TCSP algorithm. After relaxing by M3GNET, we get a more stable structure, as shown in Figure 8c. This relaxed structure is then sent to VASP to do further DFT relaxation and energy calculations. We can find that Figure 8a,b are more similar as in (b) only elemental substitutions are applied in the structure (a) with no atomic coordinate fine-tuning. As shown in Figure 8c, adjusting the coordinates based on atom sizes and bond types leads to a more stable structure. A similar procedure is applied to discover

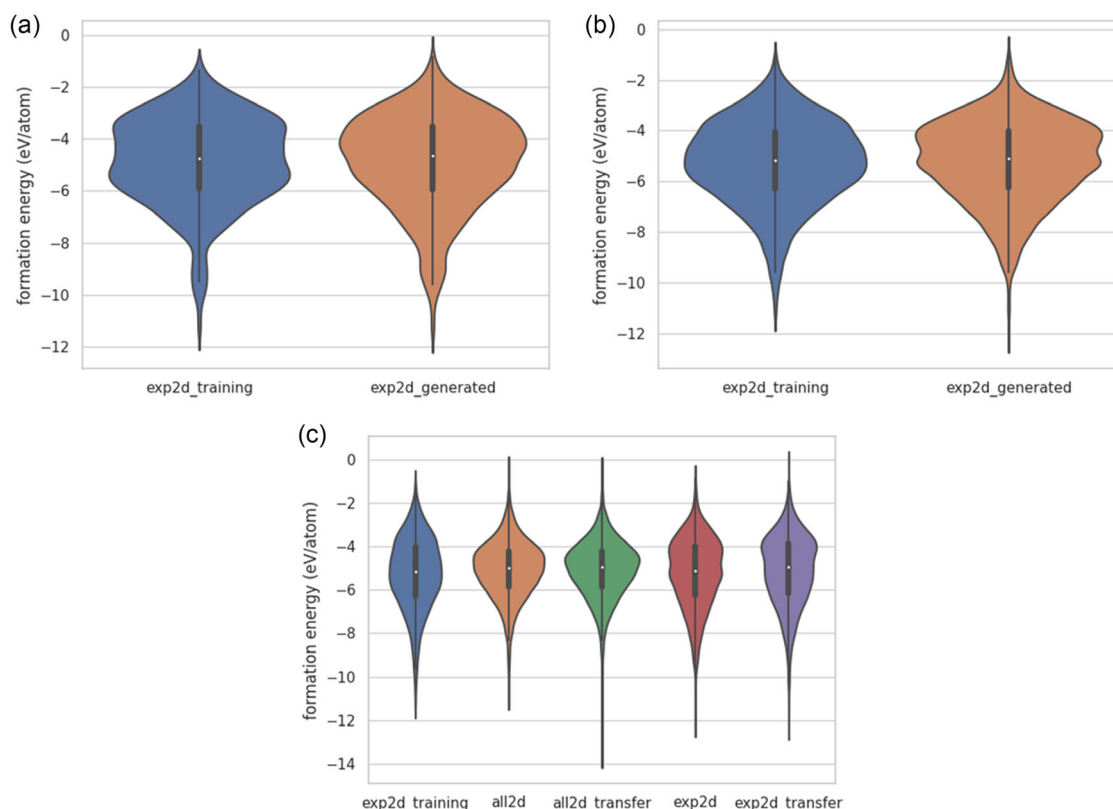


Figure 7. Formation energy per atom distribution. a) Formation energy distribution of AB_2 type structures in the exp2d dataset and generated through our MTG pipeline (predicted by M3GNET). b) Formation energy distribution of structures in the exp2d dataset and structures generated by our MTG-exp2d pipeline (formation energy predicted by M3GNET). c) Formation energy distribution of compositions in the exp2d dataset and structures generated by our MTG pipelines (BLMM models trained by all four datasets, formation energies predicted by M3GNET).

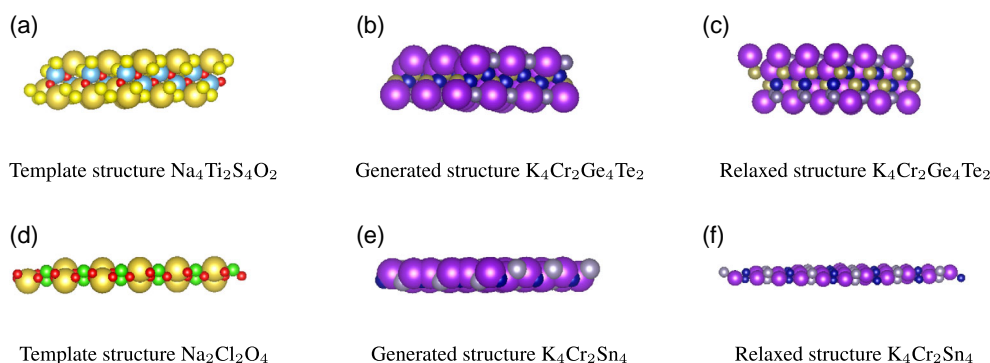


Figure 8. Illustration of the structure generation and relaxation process of our MTG pipeline. a–c) shows the template structure, structure upon element substitution, and the fine-tuned structure after ML potential-based relaxation for predicting the structure of $\text{K}_4\text{Cr}_2\text{Ge}_4\text{Te}_2$. d–f) shows the similar process for $\text{K}_4\text{Cr}_2\text{Sn}_4$.

the structure of $\text{K}_4\text{Cr}_2\text{Sn}_4$, via the three steps, as shown in Figure 8d,e,f.

Figure 9 shows four new 2D structures (which did not exist in any open source database) discovered through our MTG pipeline that have 0 e-above-hull energy (See cif information in Supplementary file). **Figure 10** shows two new 2D structures discovered through our MTG pipeline that have e-above-hull energy less than 0.05 eV (See cif information in Supplementary file). We have calculated the phonon dispersion for all structures in Figure 9 and 10. Our calculations show that two of the predicted materials, CuBr_3 and GaBrO , are thermodynamically, mechanically, and dynamically stable. **Figure 11** shows the phonon dispersion for new stable materials CuBr_3 and GaBrO . Both CuBr_3 and GaBrO show a layered structure with each layer forming a compact 2D structure, demonstrating that they have passed the dynamic stability check and the capability of our generative 2D materials design pipeline to find new 2D materials.

4. Conclusion

Two-dimensional materials have wide applications due to their unique properties. Here, we propose a generative design pipeline for 2D materials discovery by integrating a transformer-based 2D material composition generator, two template-based crystal structure predictors, and two graph neural network potential-based structure relaxation algorithms. It is found that the transformer composition generator can capture the composition preference which allows it to generate chemically valid potential 2D materials. We have applied our 2D generator pipeline to discover four hypothetical 2D materials with e-above-hull energy less than 0 and two materials with e-above-hull energy less than 0.05 eV. And two out of these six materials, CuBr_3 and GaBrO are

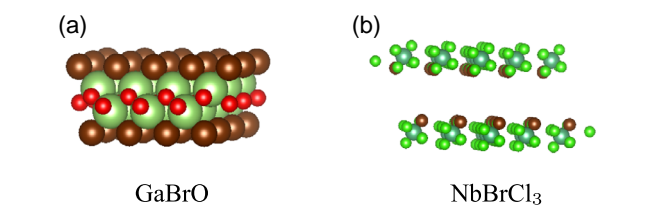


Figure 10. Two new 2D structures discovered by our MTG pipeline with E-above-hull energy less than 0.05 eV. a) GaBrO , and b) NbBrCl_3 .

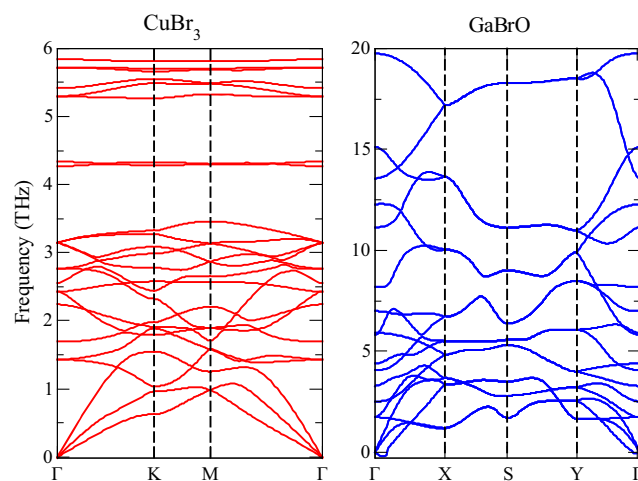


Figure 11. Phonon dispersion for new two stable materials CuBr_3 and GaBrO .

thermodynamically, mechanically, and dynamically stable. These results indicate our pipeline's capability to find new stable

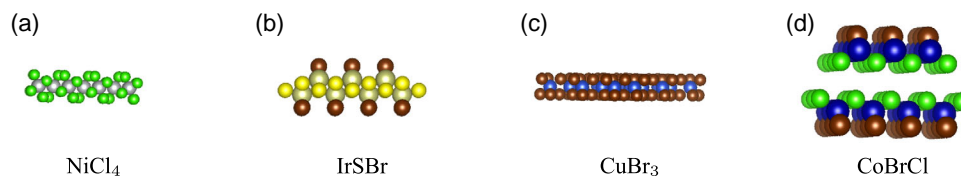


Figure 9. Four new 2D structures discovered by our MTG pipeline with 0 E-above-hull energy. a) NiCl_4 , b) IrSBr , c) CuBr_3 , and d) CoBrCl .

materials. Our pipeline is generic and can be used to train other types of materials' generative design models.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The research reported in this work was supported in part by the National Science Foundation under the grants 1940099 and 1905775. The views, perspectives, and content do not necessarily represent the official views of the NSF.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization, J.H.; methodology, R.D., J.H., Y.S., and E.S.; software, R.D. and Y.S.; resources, J.H.; writing—original draft preparation, R.D., E.S., and J.H.; writing—review and editing, J.H.; visualization, R.D. and E.S.; supervision, J.H.; funding acquisition, J.H.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

crystal structure prediction, deep learning, transformer neural networks, materials discovery, 2D materials, transformer-based materials generators

Received: March 28, 2023

Revised: August 14, 2023

Published online: October 10, 2023

- [1] N. R. Glavin, R. Rao, V. Varshney, E. Bianco, A. Apte, A. Roy, *Adv. Mater.* **2020**, 32, 1904302.
- [2] L. Shen, J. Zhou, T. Yang, M. Yang, Y. P. Feng, *Acc. Mater. Res.* **2022**, 3, 572.
- [3] P. Lyngby, K. S. Thygesen, *npj Comput. Mater.* **2022**, 8, 232.
- [4] M. C. Sorkun, S. Astruc, J. M. V. A. Koelman, S. Er, *npj Comput. Mater.* **2020**, 6, 106.
- [5] M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen, K. S. Thygesen, *2D Mater.* **2021**, 8, 044002.
- [6] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, N. Marzari, *Nat. Nanotechnol.* **2018**, 13, 246.
- [7] K. M. Wyss, D. X. Luong, J. M. Tour, *Adv. Mater.* **2022**, 34, 2106970.
- [8] P. Ares, K. S. Novoselov, *Nano Mater. Sci.* **2022**, 4, 3.
- [9] N. Briggs, S. Subramanian, Z. Lin, X. Li, X. Zhang, K. Zhang, K. Xiao, D. Geohegan, R. Wallace, L.-Q. Chen, M. Terrones, A. Ebrahimi, S. Das, J. Redwing, C. Hinkle, K. Momeni, A. van Duin, V. Crespi, S. Kar, J. A. Robinson, *2D Mater.* **2019**, 6, 022001.
- [10] S. Li, L. Ma, M. Zhou, Y. Li, Y. Xia, X. Fan, C. Cheng, H. Luo, *Curr. Opin. Biomed. Eng.* **2020**, 13, 32.
- [11] K. S. Novoselov, A. K. Geim, S. V. Morozov, D.-E. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, A. A. Firsov, *Science* **2004**, 306, 666.
- [12] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, P. Jarillo-Herrero, *Nature* **2018**, 556, 43.
- [13] S. Manzeli, D. Ovchinnikov, D. Pasquier, O. V. Yazyev, A. Kis, *Nat. Rev. Mater.* **2017**, 2, 17033.
- [14] Y. L. Huang, W. Chen, A. T. S. Wee, *SmartMat* **2021**, 2, 139.
- [15] L. Kou, Y. Ma, Z. Sun, T. Heine, C. Chen, *J. Phys. Chem. Lett.* **2017**, 8, 1905.
- [16] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, M. S. Strano, *Nat. Nanotechnol.* **2012**, 7, 699.
- [17] D. Deng, K. S. Novoselov, Q. Fu, N. Zheng, Z. Tian, X. Bao, *Nat. Nanotechnol.* **2016**, 11, 218.
- [18] B. Anasori, M. R. Lukatskaya, Y. Gogotsi, *Nat. Rev. Mater.* **2017**, 2, 16098.
- [19] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, T. Jaakkola, arXiv preprint arXiv:2110.06197, **2021**.
- [20] P. M. Larsen, M. Pandey, M. Strange, K. W. Jacobsen, *Phys. Rev. Mater.* **2019**, 3, 034003.
- [21] S. Hastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen, K. S. Thygesen, *2D Mater.* **2018**, 5, 042002.
- [22] J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, Y. P. Feng, *Sci. Data* **2019**, 6, 86.
- [23] M. T. Vahdat, K. A. Varoon, G. Pizzi, *Mach. Learn.: Sci. Technol.* **2022**, 3, 045014.
- [24] R. Friedrich, M. Ghorbani-Asl, S. Curtarolo, A. V. Krashenninnikov, *Nano Lett.* **2022**, 22, 989.
- [25] Y. Wang, M. Miao, J. Lv, L. Zhu, K. Yin, H. Liu, Y. Ma, *J. Chem. Phys.* **2012**, 137, 224108.
- [26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *Adv. Neural Inf. Process. Syst.* **2017**, 30, 5998.
- [27] T. Shen, V. Quach, R. Barzilay, T. Jaakkola, in *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, November **2020**.
- [28] L. Wei, Q. Li, Y. Song, S. Stefanov, E. Siriwardane, F. Chen, J. Hu, arXiv preprint arXiv:2204.11953, **2022**.
- [29] M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. Hart, S. Curtarolo, *Comput. Mater. Sci.* **2017**, 136, S1.
- [30] C. Su, J. Lv, Q. Li, H. Wang, L. Zhang, Y. Wang, Y. Ma, *J. Phys.: Condens. Matter* **2017**, 29, 165901.
- [31] S. D. Griesemer, L. Ward, C. Wolverton, *Phys. Rev. Mater.* **2021**, 5, 105003.
- [32] C. J. Hargreaves, M. S. Dyer, M. W. Gaultois, V. A. Kurlin, M. J. Rosseinsky, *Chem. Mater.* **2020**, 32, 10610.
- [33] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, 68, 314.
- [34] B. Kulis, *Found. Trends Mach. Learn.* **2013**, 5, 287.
- [35] N. E. R. Zimmermann, A. Jain, *RSC Adv.* **2020**, 10, 6063.
- [36] C. Liu, E. Fujita, Y. Katsura, Y. Inada, A. Ishikawa, R. Tamura, K. Kimura, R. Yoshida, *Adv. Mater.* **2021**, 33, 2102507.
- [37] M. Kusaba, C. Liu, R. Yoshida, *Comput. Mater. Sci.* **2022**, 211, 111496.
- [38] Y. Zuo, M. Qin, C. Chen, W. Ye, X. Li, J. Luo, S. P. Ong, *Mater. Today* **2021**, 51, 126.
- [39] C. Chen, S. P. Ong, *Nat. Comput. Sci.* **2022**, 2, 718.
- [40] G. Kresse, J. Hafner, *Phys. Rev. B* **1993**, 47, 558.

- [41] G. Kresse, J. Hafner, *Phys. Rev. B* **1994**, 49, 14251.
- [42] J. Furthmüller, G. Kresse, *Comput. Mater. Sci.* **1996**, 6, 15.
- [43] G. Kresse, J. Furthmüller, *Phys. Rev. B* **1996**, 54, 11169.
- [44] P. E. Blöchl, *Phys. Rev. B* **1994**, 50, 17953.
- [45] G. Kresse, D. Joubert, *Phys. Rev. B* **1999**, 59, 1758.
- [46] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, 77, 3865.
- [47] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1997**, 78, 1396.
- [48] A. Togo, I. Tanaka, *Scr. Mater.* **2015**, 108, 1.
- [49] S. Baroni, S. de Gironcoli, A. D. Corso, P. Giannozzi, *Rev. Mod. Phys.* **2001**, 73, 515.
- [50] D. W. Davies, K. T. Butler, A. J. Jackson, J. M. Skelton, K. Morita, A. Walsh, *J. Open Source Software* **2019**, 4, 1361.
- [51] M. Liu, Z. Rong, R. Malik, P. Canepa, A. Jain, G. Ceder, K. A. Persson, *Energy Environ. Sci.* **2015**, 8, 964.
- [52] E. Bitzek, P. Koskinen, F. Gähler, M. Moseler, P. Gumbsch, *Phys. Rev. Lett.* **2006**, 97, 170201.
- [53] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, 1, 011002.
- [54] L. van der Maaten, G. Hinton, *J. Mach. Learn. Res.* **2008**, 9, 2579.
- [55] L. Wei, N. Fu, E. M. D. Siriwardane, W. Yang, S. S. Omeel, R. Dong, R. Xin, J. Hu, *Inorg. Chem.* **2022**, 61, 8431.