npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Data-driven discovery of 2D materials by deep generative models

Peder Lyngby[1] ✉ and Kristian Sommer Thygesen[1]

Efficient algorithms to generate candidate crystal structures with good stability properties can play a key role in data-driven materials discovery. Here, we show that a crystal diffusion variational autoencoder (CDVAE) is capable of generating two-dimensional (2D) materials of high chemical and structural diversity and formation energies mirroring the training structures. Specifically, we train the CDVAE on 2615 2D materials with energy above the convex hull $\Delta H_{\mathrm{hull}} < 0.3$ eV/atom, and generate 5003 materials that we relax using density functional theory (DFT). We also generate 14192 new crystals by systematic element substitution of the training structures. We find that the generative model and lattice decoration approach are complementary and yield materials with similar stability properties but very different crystal structures and chemical compositions. In total we find 11630 predicted new 2D materials, where 8599 of these have $\Delta H_{\mathrm{hull}} < 0.3$ eV/atom as the seed structures, while 2004 are within 50 meV of the convex hull and could potentially be synthesised. The relaxed atomic structures of all the materials are available in the open Computational 2D Materials Database (C2DB). Our work establishes the CDVAE as an efficient and reliable crystal generation machine, and significantly expands the space of 2D materials.

npj Computational Materials (2022)8:232 ; https://doi.org/10.1038/s41524-022-00923-3

INTRODUCTION

The discovery of new materials that meet specific requirements e.g., in terms stability, compatibility, or physical properties, is an exciting scientific challenge of great relevance for our society. First-principles quantum mechanical calculations, e.g., based on density functional theory (DFT)[1], can predict the structure and properties of materials with high accuracy even before they are made in the lab. However, a DFT code by itself is insufficient for realising the paradigm of inverse materials design, where instead of mapping from structure to property using ab initio methods, the goal is to do the inverse map: from target property to atomic structure.

Considering the vast number of possible materials and the complexity of general structure-property relations, it becomes clear that successful inverse design relies on the following critical components: (i) automated execution and management of large numbers of atomistic calculations, (ii) access to large amounts of relevant high quality materials data, and (iii) efficient algorithms that can propose new candidate materials from data. In addition, synthesis and characterisation experiments must be included in the loop as well, but this aspect will not be considered here.

Components (i) and (ii) are largely in place. Indeed, the advent of workflow management engines for computational materials science[2–6] have made it possible to perform high-throughput (HT) computations for thousands of materials with minimal human intervention[7–22]. Atomic structures and basic materials properties from such HT studies have been stored in computational databases[2,23–31], which together contain results of millions of DFT calculations. Complemented by experimental crystal structure databases, this makes a rich and rapidly growing data source for materials science.

The main challenge concerns component (iii). In previous HT studies, the candidate materials to be explored were mostly produced by lattice decoration of known reference materials. An obvious limitation of this approach is that the resulting materials by construction will be similar to the reference materials. In particular, the 3-tuple: (space group, occupied Wyckoff positions, stoichiometry) is invariant under element substitution.

Generative machine learning algorithms could potentially broaden the diversity of candidate materials beyond the lattice decoration paradigm. However, designing a successful generative model for periodic materials has proved challenging due the problem of creating representations of the lattice, atomic coordinates and elemental composition that are both invariant to translations and rotations and is invertible[32]. The vast chemical space of elements that can be present in inorganic crystals further complicates the design of representations. Therefore, previous implementations of generative models for periodic materials have either been limited to a fixed subset of chemical elements[33–35] and/or a subset of possible crystal structures[36,37]. Recently, a general invertible representation has been proposed[38], which encodes the material as a matrix of both real and reciprocal space features, but is not invariant under translations and rotations. Xie et al. developed a crystal diffusion variational autoencoder (CDVAE) model[39], which uses a generative diffusion model to circumvent the need for an invertible representation by working directly on the atomic coordinates of the structures and employs an equivariant graph neural network to ensure invariance (in fact, equivariance).

In this work, we train a CDVAE[39] on 2615 2D materials with formation energy up to 0.3 eV/atom from the convex hull, and generate 10000 two-dimensional (2D) crystals. We compare these structures to a set of 14192 2D crystals obtained by systematic lattice decoration of the training structures. While ref. [39] assessed validity and diversity of the generated crystals by means of qualitative measures, such as charge neutrality and minimum bond distance, we here conduct a systematic, unbiased quantitative analysis by performing full DFT-based relaxations and stability analysis of the generated structures. Compared to the crystals in the training set, the structures generated by the CDVAE

[1] Computational Atomic-scale Materials Design (CAMD), Department of Physics, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark. ✉email: pmely@dtu.dk

(after DFT relaxation) show similar formation energies but significant differences in both composition and crystal structure. In general, CDVAE seems able to produce more complex materials without compromising stability.

As a direct test of the CDVAE model's capacity to learn the stability properties of the training structures, we also train a CDVAE on materials lying at least 0.4 eV/atom above the hull. We find that the structures generated by this model have significantly higher formation energies than those produced by the CDVAE trained on the more stable materials.

In addition to providing a quantitative assessment of the CDVAE, our work identifies no less than 8599 new unique 2D materials with an energy above the convex hull below 0.3 eV/atom many of which could potentially be synthesised. The generated crystal structures are freely available as part of the C2DB[29].

RESULTS

Crystal diffusion variational autoencoder

The CDVAE combines a variational autoencoder[40] and a diffusion model to generate new periodic materials. The crystal is represented by a tuple consisting of the atomic number of the N atoms, their respective coordinates, and the unit cell basis vectors. CDVAE consist of three networks: the encoder, a property predictor, and the decoder which all are trained concurrently. The encoder is a SE(3) equivariant periodic graph neural network (PGNN), which encodes the material onto a lower dimensional latent space from which the property predictor predicts the number of atoms N, the lattice vectors, and the composition, which is the fraction present of each element. The decoder is a noise conditional score network diffusion model[41] that takes a structure with noise added to the atom types and coordinates and learns to denoise it into the original stable structure. Noise added to the atom types changes type of element for each atom into another element within the predicted composition with a certain probability given by the noise-level. Coordinate noise on the other hand is simply Gaussian noise added to the coordinates of each atom of the structure. The score of the conditional score network diffusion model is an estimate of the gradient of the underlying probability distribution of the materials and is predicted by another SE(3) equivariant PGNN. The use of a equivariant diffusion model as the decoder makes it possible to work directly with the atomic positions without the need for any intermediate representations like descriptors or graphs. This in turn makes the CDVAE framework quite general and agnostic to the kind chemical elements and structure which it is used for, which allows CDVAE to generate 2D materials even though it was designed for 3D bulk materials.

New materials can be generated after training by using the property predictor to sample the latent space. A unit cell with the predicted basis vectors is then initialised with the predicted atoms placed at random positions. Using the decoder, the atom types and coordinates of the initial random placed atoms are then gradually denoised into a material that is similar to the data distribution of the training data. CDVAE utilizes that adding noise to a stable material will likely decrease its stability and, thus, by learning to denoise the noisy stable structure, the decoder learns to increase the stability of the structure. Therefore CDVAE should be trained only on stable materials. An in-depth description of CDVAE can be found in Xie et al.[39].

The set of materials used as training data for the CDVAE and seed structures for the lattice decoration protocol (LDP), respectively, consists of 2615 unique 2D materials from the C2DB[29,31]. As our aim is to discover new stable materials we limited the initial set of materials to the subset of C2DB with energy above the convex hull $\Delta H_{\mathrm{hull}} < 0.3$ eV/atom. This was done because both the CDVAE (LDP) are more likely to generate stable materials when trained on (seeded by) stable materials. We did not exclude dynamically unstable materials.

After training the CDVAE model, 10.000 structures were generated of which 1106 failed CDVAE's basic validity check (charge neutrality and bond lengths above 0.5 Å). Of the remaining 8894 structures, 3891 are duplicate structures which are sorted out (see "Method" for more details) and the rest are relaxed using DFT.

Lattice decoration protocol

The lattice decoration protocol (LDP) substitutes the atoms in the seed structures by atoms of similar chemical nature. As a measure of chemical similarity we use the probability matrix P_{AB} introduced by ref. [42], which describes the likelihood that a stable material containing a chemical element A remains stable after the substitution $A \rightarrow B$. Glawe et al. constructed this probability matrix based on an analysis of materials in the Inorganic Crystal Structure Database[43]. We choose a substitution probability of 10% ($P_{AB} >$ 0.1), which generates the substitutions shown in Fig. 1. Based on these substitution relations, we perform all possible single and double substitutions for all seed structures. For example, the seed structure MoS_2 generates six MX_2 structures with M = Mo,W and X = O, S, Se (the seed structure itself included). The total set of resulting materials are analysed for structures that share the same reduced formula and space group. Such structures are considered

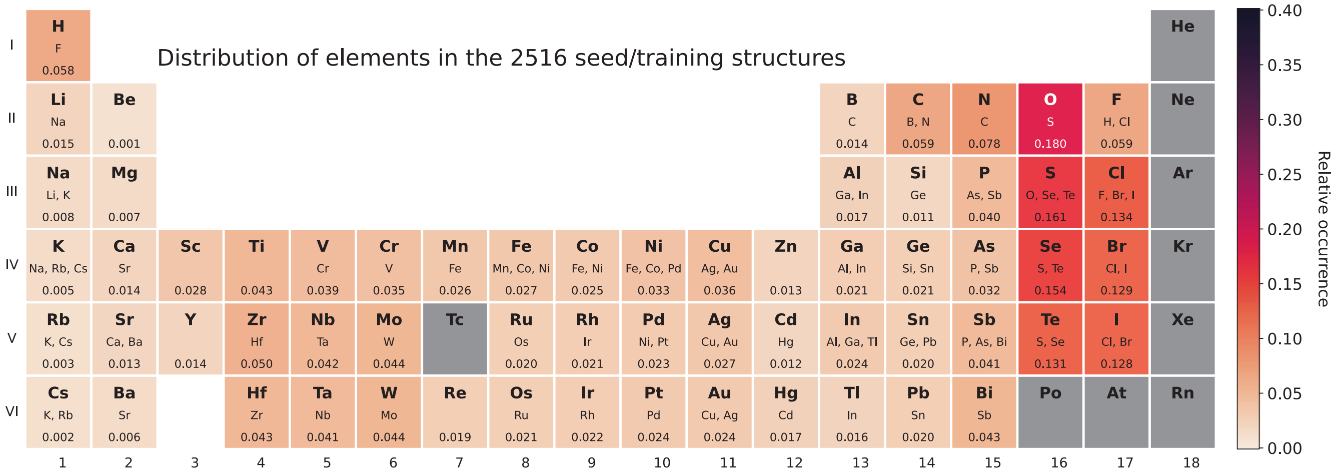

Fig. 1 Heat map of the relative occurrence of each element in the 2D materials used to train (seed) the CDVAE (LDP). The middle row shows the element substitutions for the LDP corresponding to $P_{AB} > 0.1$. The relative occurrence is shown in the last row.

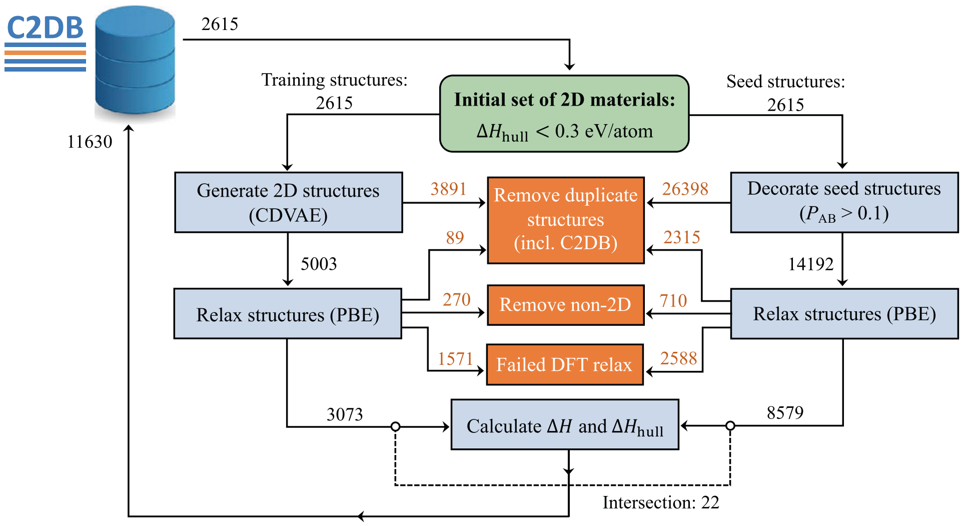
C2DB
2615
11630
Training structures: 2615
Initial set of 2D materials:
$\Delta H_{\mathrm{hull}} < 0.3$ eV/atom
Seed structures: 2615
Generate 2D structures (CDVAE)
3891
Remove duplicate structures (incl. C2DB)
26398
Decorate seed structures ($P_{\mathrm{AB}} > 0.1$)
89
2315
5003
14192
Relax structures (PBE)
270
Remove non-2D
710
Relax structures (PBE)
1571
Failed DFT relax
2588
3073
Calculate ΔH and ΔH_{hull}
8579
Intersection: 22

Fig. 2 Workflow diagram. Workflow to generate candidate 2D materials using the CDVAE generative model (left branch) and lattice decoration (right branch). The same set of 2615 materials is used to train the CDVAE model and as seed structures for lattice decoration, respectively. Black numbers indicate the number of materials present at a given step of the workflow while orange numbers indicate number of materials discarded.

Table 1. Summary statistics for the DFT relaxation of the two methods for generating initial structures.

	LDP	CDVAE
Success rate	82%	69%
Avg. number of steps	40.1	55.5
Avg. energy decrease [eV/atom]	0.62	0.51

as duplicate structures and are filtered out. After removal of duplicates, we are left with 14,192 unique 2D crystals (the seed structures excluded) which are relaxed using DFT.

Workflow

Our workflow is illustrated in Fig. 2. Starting with the initial set of 2D materials, we generate two new sets of crystal structures using CDVAE and LDP, respectively. Duplicate structures within each set are removed (see “Method” for more details). The now unique crystal structures are relaxed using DFT calculations employing the PBE xc-functional (see “Method” for more details). After the relaxation, any new duplicate structures are removed again and as are materials that have relaxed into non 2D structures (we refer to ref. [31] for details on the dimensionality analysis). Finally the heat of formation, ΔH, and the energy above convex hull, ΔH_{hull}, are calculated.

In Table 1, we report the success rates for the DFT relaxations of the structures generated by CDVAE and LDP, respectively, together with the average number of relaxation steps and the average energy decrease from the initial to the relaxed structure. All three parameters are assumed to describe how close the initial structures are to the final DFT relaxed structures - e.g., a structure from a perfect generative method would only need one relaxation step and the energy decrease would be zero. As expected, neither LDP or CDVAE generate stable relaxed structures. However, while the LDP on average requires less steps to relax, the CDVAE structures are closer in energy to the relaxed structure. The fact that the number of relaxation steps and reduction in energy upon relaxation is comparable for LDP and CDVAE, suggest that the CDVAE-generated crystals are as close to relaxed structures as the LPD-generated structures.

We observe that the DFT relaxation fails for about 18% of the LDP-generated and about 31% of the CDVAE-generated structures. The vast majority of these failures are due to problems in converging the Kohn–Sham SCF cycle. We suspect that a large fraction of the convergence problems occur for materials with magnetic ground state (all calculations are performed with spin polarisation). This is supported by the fact that 30% of the materials containing one or more of the magnetic 3d-metals (V, Cr, Mn, Fe, Co, Ni), fails due to convergence errors, while this is only happens for 10% of other materials. Moreover, 38% of the CDVAE-generated structures contains at least one of the the magnetic 3d-metals, while this is only the case for 30% of the LDP-generated structures. This difference is consistent with the difference in the observed success rate.

Thermodynamic stability

A histogram of the heat of formation and the energy above the convex hull for the (DFT-relaxed) structures resulting from the CDVAE and LDP are shown in Fig. 3. The distributions of both ΔH and ΔH_{hull} obtained for the two structure generation methods are remarkably similar. For example, 73.8% of the CDVAE materials have ΔH_{hull} below 0.3 eV/atom (as the training data) while this is the case for 74.0% of the LDP materials. It should, however, be noted that the smaller success rate of the DFT relaxation of the CDVAE generated materials could influence these statistics as it likely that many of the structures which could not be converged would have resulted in unstable structures. The inset of Fig. 3 shows how the energy above the convex hull is distributed depending on the number of different elements in the structure. First of all it is evident that CDVAE is able to create structures with a larger number of unique elements than is present in the training data (5 unique elements is the maximum in the seed structures), while LDP is limited to the stoichiometries present in the seed materials. However, generally the thermodynamic stability is lower for the materials with larger number of unique elements. Examples of some of the most stable CDVAE generated structures is shown in Fig. 4. The material Zr_2CCl_2 shown in c) is one of the 22 materials which are found both by the CDVAE and LDP method.

To predict whether a given 2D material can be synthesised is a complex problem that involves many factors. Often the size of ΔH_{hull} is used a soft criterion for synthesizability as it determines the material’s thermodynamic stability relative to other competing

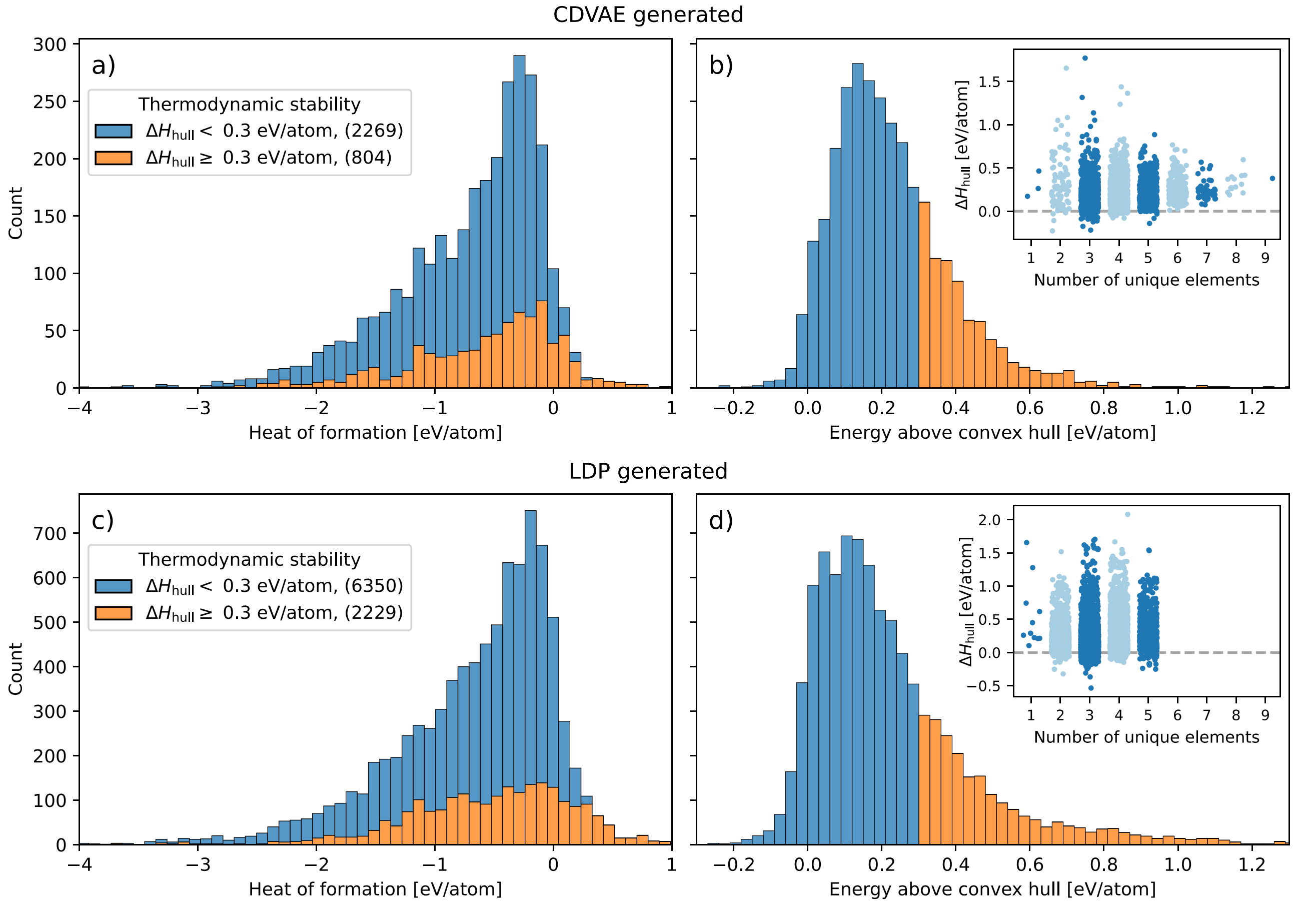

Fig. 3 Histogram of the heat of formation and energy above convex hull for the DFT-relaxed structures resulting from the CDVAE and LDP methods. a, **b** CDVAE generated structures (**c**, **d**) LDP generated structures. The inset shows the energy above convex hull with respect to the number of unique elements in the structure.

phases (this criterion neglects growth kinetics and substrate interactions both of which can be important for 2D materials). A previous study of 700 polymorphs in 41 common inorganic bulk material systems showed that a threshold of $\Delta H_{\mathrm{hull}} < 0.1$ eV/atom will exclude 26% of the known synthesised polymorphs[44]. We also note that the T-phase of MoS_2 was synthesised both as a monolayer[45] and a layered bulk[46], despite having $\Delta H_{\mathrm{hull}} = 0.18$ eV/atom[47]. These examples demonstrate that many of the predicted 2D materials with $\Delta H_{\mathrm{hull}} < 50$ meV/atom (2004) or even $\Delta H_{\mathrm{hull}} < 100$ meV/atom (3400), are likely to be synthesizable.

While the ΔH_{hull}-distributions in Fig. 3 are clearly peaked close to zero they also have a tail of less stable materials. In particular, about 26% of the materials have $\Delta H_{\mathrm{hull}} > 0.3$ eV/atom (the threshold to select the training structures). A natural question to ask is then to what extent the structures produced by the CDVAE are in fact biased towards high stability structures? To answer this question, we trained a CDVAE model on 988 2D materials with $\Delta H_{\mathrm{hull}} > 0.4$ eV/atom and used it to generate another 10.000 structures from which we randomly selected 1000 non-duplicate structures, which we relaxed following the same workflow as described before. The distribution of the energy above the convex hull of the relaxed structures for both the stable and unstable CDVAE models are shown in Fig. 5 together with the distribution of their respective training sets. We clearly see that the CDVAE model trained to generate unstable materials produces structures that are significantly further from the convex hull than the stable model. This illustrates that CDVAE successfully learns the chemistry of the materials in the training data.

Structural diversity

Having established the capability of the CDVAE to produce materials with good stability properties, we now turn to its ability to generate crystals of high chemical and structural diversity. While the LDP is restricted to stoichiometries and crystal structures already present in the seed structures, the CDVAE (in principle) has no such limitations. Figure 1 shows the relative occurrence of each element in the seed/training structures. The corresponding plots for the materials generated by LDP and CDVAE (after relaxation) are shown in Supplementary Fig. 1. Both LDP and CDVAE produces diverse compositions with elements covering most of the periodic table. However, CDVAE has a significantly higher occurrence of oxygen and chalcogens (S and Se) as well as halogens (Cl, Br and I). This trend is also present for the materials prior to relaxation and, thus does not originate from a potential higher DFT convergence rate for these elements. Instead, the six elements are also more prevalent, albeit slightly, in the seed structures which could indicate an overfitting of the model.

The CDVAE generates significantly different chemical compositions and crystal structures as compared to the seed structures and those generated by the LDP. Figure 6 shows the relative frequencies of stoichiometry, space group number and occupied Wyckoff positions, respectively. Only the most common classes of the seed structures are shown. We find 239 unique stoichiometries among the CDVAE-generated materials, while there is only 87 and 103 unique stoichiometries in the seed structures and LDP-generated structures, respectively. The higher number of unique stoichiometries in the LDP-generated structures than in the seed structures is due to new stoichiometries being created when two

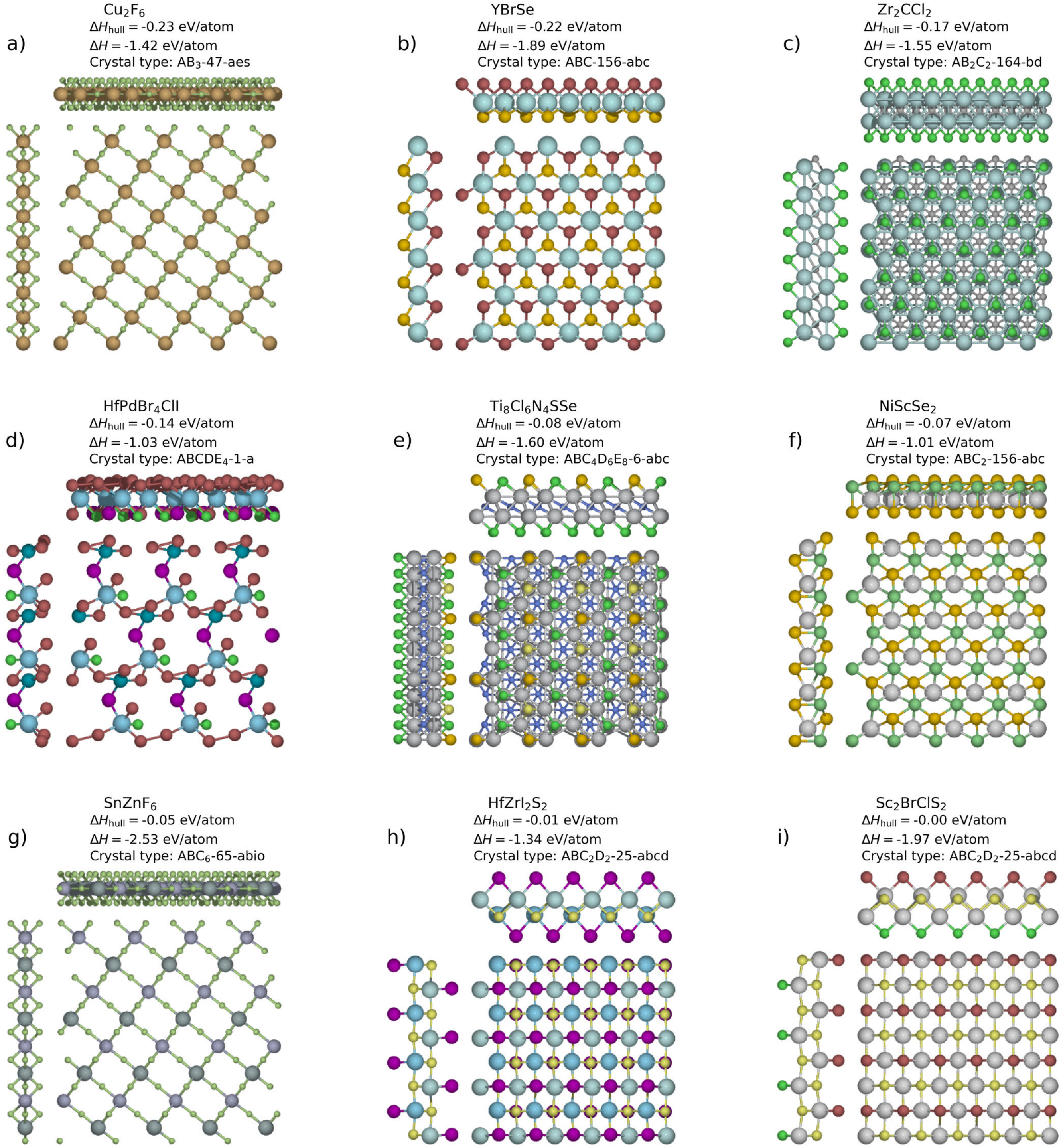
a)
Cu_2F_6
ΔH_{hull} = -0.23 eV/atom
ΔH = -1.42 eV/atom
Crystal type: AB_3-47-aes
b)
YBrSe
ΔH_{hull} = -0.22 eV/atom
ΔH = -1.89 eV/atom
Crystal type: ABC-156-abc
c)
Zr_2CCl_2
ΔH_{hull} = -0.17 eV/atom
ΔH = -1.55 eV/atom
Crystal type: AB_2C_2-164-bd
d)
$HfPdBr_4ClI$
ΔH_{hull} = -0.14 eV/atom
ΔH = -1.03 eV/atom
Crystal type: $ABCDE_4$-1-a
e)
$Ti_8Cl_6N_4SSe$
ΔH_{hull} = -0.08 eV/atom
ΔH = -1.60 eV/atom
Crystal type: $ABC_4D_6E_8$-6-abc
f)
$NiScSe_2$
ΔH_{hull} = -0.07 eV/atom
ΔH = -1.01 eV/atom
Crystal type: ABC_2-156-abc
g)
$SnZnF_6$
ΔH_{hull} = -0.05 eV/atom
ΔH = -2.53 eV/atom
Crystal type: ABC_6-65-abio
h)
$HfZrI_2S_2$
ΔH_{hull} = -0.01 eV/atom
ΔH = -1.34 eV/atom
Crystal type: ABC_2D_2-25-abcd
i)
Sc_2BrClS_2
ΔH_{hull} = -0.00 eV/atom
ΔH = -1.97 eV/atom
Crystal type: ABC_2D_2-25-abcd

Fig. 4 Examples of CDVAE generated structures. a–g Examples of CDVAE generated materials with negative convex hull energies. **h**, **i** Examples of CDVAE generated stable materials with the new discovered combination of stoichiometry ABC_2D_2, space group number 25 and occupied Wyckoff positions a, b, c, d.

different elements are substituted by the same element, or when an element is being substituted with an element already present in the seed material. For example, the seed materials $Te_2Cu_4O_{12}$ (stoichiometry AB_2C_6) becomes Cu_4S_{14} (stoichiometry A_2B_7) under the double substitution O → S and Te → S. The significantly larger number of unique stoichiometries generated by CDVAE compared to the LDP shows that the former is able to produce new classes of structures that are not present in the training data. Another indication of new structural prototypes being created is the occurrence of new occupied Wyckoff positions within each space group when comparing to the training data. These new combinations of space group and occupied Wyckoff position are shown in Supplementary Tables 1 and 2 for both the CDVAE dataset and the LDP dataset. In total there are 130 new combinations and 357 materials with these new combinations in the CDVAE-generated materials, while there are only 76 new combinations in the LDP-generated materials and only 339 materials with the combinations - even though the LDP dataset is almost three times as large as the CDVAE dataset. It might seem strange that LDP generates new combinations of space group and Wyckoff position as simple element substitution should preserve the space group and occupied Wyckoff position. However during the DFT relaxation the crystal symmetry can change and thus so can the space group and occupied Wyckoff positions.

The CDVAE tends to generate rather complex, low-symmetry structures, which is illustrated by the large fraction of materials with space group number 1 and occupied Wyckoff position a. Moreover, the average number of different elements in the unit cell is 4.0 for the CDVAE generated materials while it is only 2.6 for the C2DB seed structures. The larger number of different elements is part of the reason for the higher fraction of materials with low symmetry. This tendency of CDVAE to generate structures with more complex composition is also noted by Xie et al., who attributes this to a non-Gaussian distribution of the underlying structure of the materials. Thus, when CDVAE generates new materials it samples from a Gaussian distribution $\mathcal{N}(0,1)$ from which it predicts the number of atoms and composition. However if $\mathcal{N}(0,1)$ is not representative of the latent space, out of distribution materials can be generated. For materials discovery this could, however, be advantageous as this makes CDVAE able to generate new crystal types which are not present in the training data.

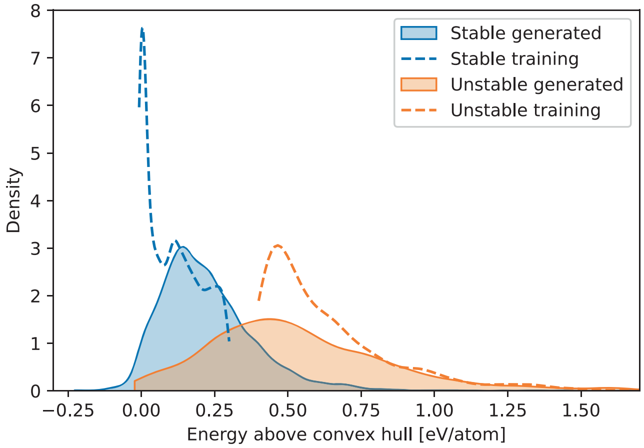

Fig. 5 Kernel density estimate showing the distribution of the convex hull energies for the stable and unstable CDVAE generated dataset. The dashed line shows the distribution of the corresponding training data.

To give a global overview of the structural distribution of the three data sets, a t-SNE embedding is shown in Fig. 7. The t-SNE analysis is made for 2500 materials sampled randomly from each data set. Here the structure is represented as a tuple given by the space group, occupied Wyckoff positions, and stoichiometry, where each is one-hot encoded before the t-SNE embedding. We see that most of the training data form clear clusters, which represent the most common stoichiometries, space group and Wyckoff positions. The LDP generated materials mostly follow the same pattern as the seed structures. However, the CDVAE generated structures are more spread out, which is partly due to the large variation in their stoichiometries, while a few clusters appear due to the large fraction of low symmetry materials with space group number 1. One noteworthy example is the cluster of CDVAE generated materials with stoichiometry ABC_2D_2, space group number 25 and occupied Wyckoff positions a, b, c, d. For this specific combination, CDVAE discovered 123 new materials of which 30 lies within 50 meV of the convex hull, while there is no examples of such materials in the training set nor in the LDP generated structures. Two of the most stable discovered materials of this type can be seen in Fig. 4h, i. The new class of structures have broken out-of-plane symmetry either due to the outermost atoms (i) or the innermost atoms (h). The fact that the CDVAE is able to generate new classes of stable materials, which are not present in the training data, is very promising and a clear advantage of deep generative models compared to lattice decoration protocols.

DISCUSSION

In conclusion, we have successfully employed a deep generative model in combination with a systematic lattice decoration protocol (LDP) to generate more than 8500 unique 2D crystals with formation energies (ΔH) within 0.3 eV/atom of the convex hull. Out of these, more than 2000 have ΔH within 50 meV/atom of the convex hull, and could potentially be synthesised. This represents at least a doubling of the known stable 2D materials.

In addition to the very significant expansion of the known space of 2D materials, our work provides a quantitative assessment of the crystal diffusion variational autoencoder (CDVAE)[39], and establishes its excellent performance with respect to the two key criteria: ability to learn the stability properties of the training structures, and ability to generate crystals with high chemical and structural diversity. In fact, only 25% of the generated materials had ΔH_{hull} above the 0.3 eV/atom threshold used to select the training structures, and the stoichiometries of the generated materials span 239 types versus 87 present in the training structures.

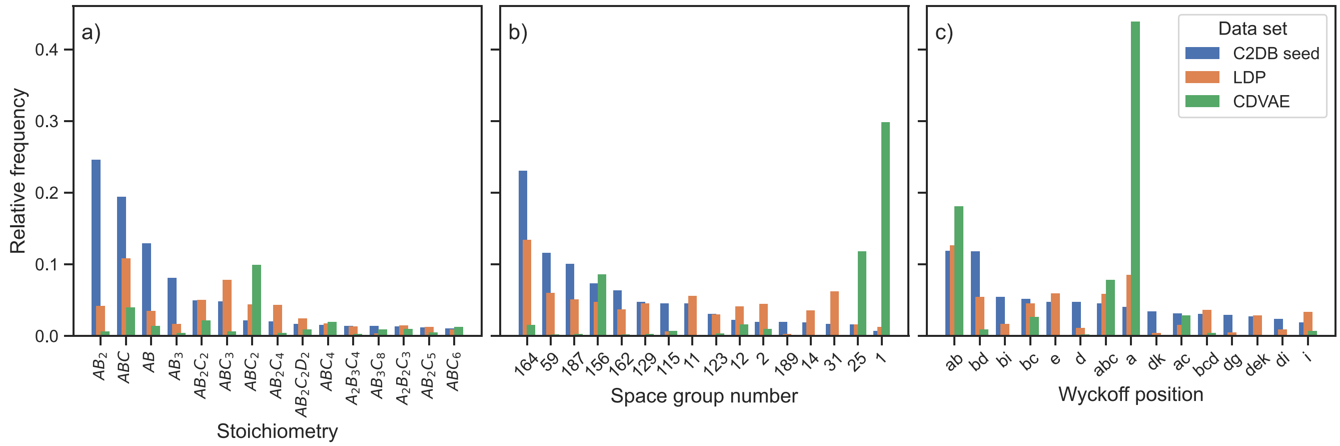

Fig. 6 Histograms of structural parameters. Relative frequency of the stoichiometry, space group number and occupied Wyckoff position for each of the data set.

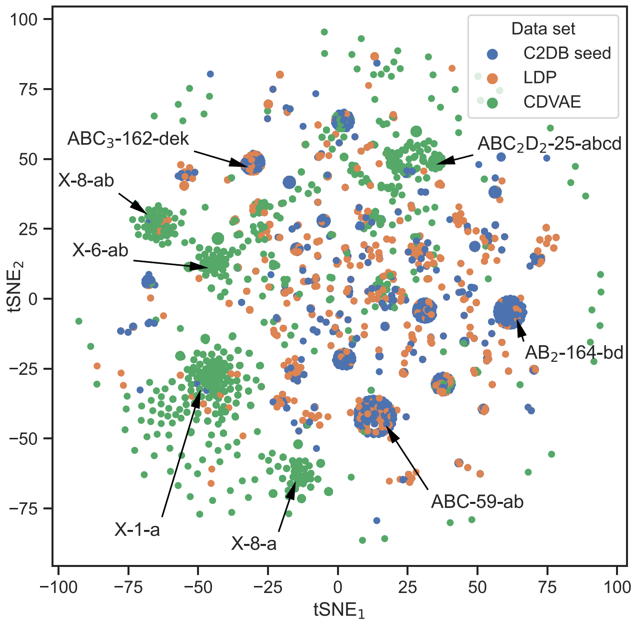

Fig. 7 t-SNE visualization of the structural diversity. t-SNE embedding of the structure represented as a individual one-hot encoding of the stoichiometry, space group, and occupied Wyckoff positions. Selected clusters are highlighted as “stoichiometry”-“space group”-“Wyckoff position”. “X” corresponds to an arbitrary stoichiometry.

Generally, the crystal structures generated by CDVAE have higher complexity and lower symmetries than the training structures. We found the method of lattice decoration to be complementary to the CDVAE generator with the two methods yielding only 22 common crystals out of the 11630 structures generated in total. While the LDP is limited to the structural blueprint of the seed materials, CDVAE is able to generate new classes of materials, which are not present in the initial data set. This is promising for an autonomous materials discovery method as it adds new genes to pool of trial materials and thus goes beyond the lattice decoration paradigm.

The fact that CDVAE is comparable to lattice decoration (with substitution by chemically similar elements) in terms of stability while producing new and diverse crystal structures, is a testimony to the prospect of using deep generative models in materials discovery.

All the structures are available in the C2DB database[47], and their basic properties will also be made available as the execution of the C2DB property workflow proceeds.

METHOD

Workflow

To set up and manage the workflow we use the Atomic Simulation Recipes[5], which has implemented tools for DFT relaxation, duplicate removal, dimensionality check, and for calculating the thermodynamic properties. The DFT calculations are performed using the GPAW code[48] with the PBE xc-functional, a plane wave cut-off energy of 800 eV and a *k*-point density of at least 4 Å. The relaxation is stopped when the maximum force is below 0.01 eV/Å and the maximum stress is below 0.002 eV/Å^3.

The duplicate removal recipe finds duplicate structures using the root mean square distance (RMSD) between the structures which is calculated using the Python library pymatgen[49]. We consider structures to be duplicate if RMSD < 0.3 Å and only keep the structure with the lowest heat of formation. See ref. [31] for more information. For the initial LDP generated materials (before the DFT relaxation) a more crude duplicate sorting of the structures is employed, where two materials with the same reduced formula and space group are considered duplicates.

To determine the convex hull we use as reference databases the C2DB as well as a database of reference structures comprising 9590 elementary, binary, and ternary crystals that all lie within 20 meV of the convex hull in the Open Quantum Materials Database (OQMD)[25]. These reference structures were relaxed using the VASP[50] code at the PBE level (PBE+U for selected transition metal oxides) as part of the OQMD project. Since we use the GPAW code to relax and evaluate the energy of the 2D materials, we have re-calculated the total energy of the reference structures (without re-optimisation) using the GPAW code.

CDVAE

CDVAE is designed to generate 3D bulk crystals, where the unit cell is periodic in all three directions. This introduces a problem when generating 2D materials, which are non-periodic in one direction. We solve this issue by introducing an artificial periodicity in the non-periodic direction with a lattice vector which is an order of magnitude larger than those in the periodic directions. This ensures that the graph networks only connect atoms in the 2D layer and thus CDVAE learns to generate 2D materials. When training the model, we used 70% of the materials in the training set, while 15% was used for validation and 15% for test. We used the same hyperparameters as employed by Xie et al. for their MP-20 data set. See ref. [39] for more information.

DATA AVAILABILITY

All the discovered crystal structures and their properties are available as a part of C2DB (https://cmr.fysik.dtu.dk/c2db/c2db.html). The discovered materials can be selected using the key `origin`, which have the values `Lyngby22_CDVAE` or `Lyngby22_LDP` for the CDVAE and LDP generated materials respectively.

Received: 27 June 2022; Accepted: 23 October 2022;
Published online: 11 November 2022

REFERENCES

1. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
2. Curtarolo, S. et al. Aflow: An automatic framework for high-throughput materials discovery. *Computational Mater. Sci.* **58**, 218–226 (2012).
3. Jain, A. et al. Fireworks: a dynamic workflow system designed for high-throughput applications. *Concurrency Comput.* **27**, 5037–5059 (2015).
4. Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. Aiida: automated interactive infrastructure and database for computational science. *Computational Mater. Sci.* **111**, 218–230 (2016).
5. Gjerding, M. et al. Atomic simulation recipes: A python framework and library for automated workflows. *Computational Mater. Sci.* **199**, 110731 (2021).
6. Mortensen, J., Gjerding, M. & Thygesen, K. Myqueue: Task and workflow scheduling system. *J. Open Source Softw.* **5**, 1844 (2020).
7. Greeley, J., Jaramillo, T. F., Bonde, J., Chorkendorff, I. & Nørskov, J. K. Computational high-throughput screening of electrocatalytic materials for hydrogen evolution. *Nat. Mater.* **5**, 909–913 (2006).
8. Madsen, G. K. Automated search for new thermoelectric materials: the case of liznsb. *J. Am. Chem. Soc.* **128**, 12140–12146 (2006).
9. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
10. Kirklin, S., Meredig, B. & Wolverton, C. High-throughput computational screening of new li-ion battery anode materials. *Adv. Energy Mater.* **3**, 252–262 (2013).
11. Ørnsø, K. B., Garcia-Lastra, J. M. & Thygesen, K. S. Computational screening of functionalized zinc porphyrins for dye sensitized solar cells. *Phys. Chem. Chem. Phys.* **15**, 19478–19486 (2013).
12. Zhang, Z. et al. Computational screening of layered materials for multivalent ion batteries. *ACS Omega* **4**, 7822–7828 (2019).

13. Chen, W. et al. Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment. *J. Mater. Chem. C.* **4**, 4414–4426 (2016).
14. Hachmann, J. et al. The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).
15. Bhattacharya, S. & Madsen, G. K. High-throughput exploration of alloying as design strategy for thermoelectrics. *Phys. Rev. B* **92**, 085205 (2015).
16. Castelli, I. E. et al. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.* **5**, 5814–5819 (2012).
17. Hautier, G., Miglio, A., Ceder, G., Rignanese, G.-M. & Gonze, X. Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat. Commun.* **4**, 1–7 (2013).
18. Yu, L. & Zunger, A. Identification of potential photovoltaic absorbers based on first-principles spectroscopic screening of materials. *Phys. Rev. Lett.* **108**, 068701 (2012).
19. Kuhar, K., Pandey, M., Thygesen, K. S. & Jacobsen, K. W. High-throughput computational assessment of previously synthesized semiconductors for photovoltaic and photoelectrochemical devices. *ACS Energy Lett.* **3**, 436–446 (2018).
20. Aykol, M. et al. High-throughput computational design of cathode coatings for li-ion batteries. *Nat. Commun.* **7**, 1–12 (2016).
21. Mounet, N. et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nat. Nanotechnol.* **13**, 246–252 (2018).
22. Chen, L.-Q. et al. Design and discovery of materials guided by theory and computation. *npj Computational Mater.* **1**, 1–2 (2015).
23. Thygesen, K. S. & Jacobsen, K. W. Making the most of materials computations. *Science* **354**, 180–181 (2016).
24. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
25. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd). *JOM* **65**, 1501–1509 (2013).
26. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
27. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: An open-access online database for data mining. *PLoS ONE* **12**, e0171501 (2017).
28. Draxl, C. & Scheffler, M. The nomad laboratory: from data sharing to artificial intelligence. *J. Phys.: Mater.* **2**, 036001 (2019).
29. Haastrup, S. et al. The computational 2d materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).
30. Cheon, G. et al. Data mining for new two-and one-dimensional weakly bonded solids and lattice-commensurate heterostructures. *Nano Lett.* **17**, 1915–1923 (2017).
31. Gjerding, M. N. et al. Recent progress of the computational 2d materials database (c2db). *2D Mater.* **8**, 044002 (2021).
32. Noh, J., Gu, G. H., Kim, S. & Jung, Y. Machine-enabled inverse design of inorganic solid materials: promises and challenges. *Chem. Sci.* **11**, 4871–4881 (2020).
33. Noh, J. et al. Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
34. Kim, S., Noh, J., Gu, G. H., Aspuru-Guzik, A. & Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS Cent. Sci.* **6**, 1412–1420 (2020).
35. Long, T. et al. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *NPJ Comput. Mater.* **7**, 66 (2021).
36. Zhao, Y. et al. High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Adv. Sci.* **8**, 2100566 (2021).
37. Song, Y., Siriwardane, E. M. D., Zhao, Y. & Hu, J. Computational discovery of new 2d materials using deep learning generative models. *ACS Appl. Mater. Interfaces* **13**, 53303–53313 (2021).
38. Ren, Z. et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **5**, 314–335 (2022).
39. 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).
40. Kingma, D. P. & Welling, M.Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).
41. Song, Y. & Ermon, S. Generative modeling by estimating gradients of the data distribution. In Wallach, H. et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 32 (Curran Associates, Inc., 2019). https://proceedings.neurips.cc/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf.
42. Glawe, H., Sanna, A., Gross, E. K. U. & Marques, M. A. L. The optimal one dimensional periodic table: a modified pettifor chemical scale from data mining. *N. J. Phys.* **18**, 093011 (2016).
43. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. Sect. B* **58**, 364–369 (2002).
44. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
45. Kappera, R. et al. Phase-engineered low-resistance contacts for ultrathin mos2 transistors. *Nat. Mater.* **13**, 1128–1134 (2014).
46. Bell, R. E. & Herfert, R. E. Preparation and characterization of a new crystalline form of molybdenum disulfide. *J. Am. Chem. Soc.* **79**, 3351–3354 (1957).
47. https://cmr.fysik.dtu.dk/c2db/c2db.html.
48. Mortensen, J. J., Hansen, L. B. & Jacobsen, K. W. Real-space grid implementation of the projector augmented wave method. *Phys. Rev. B* **71**, 035109 (2005).
49. Ong, S. P. et al. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Mater. Sci.* **68**, 314–319 (2013).
50. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).

ACKNOWLEDGEMENTS

We thank Morten N. Gjerding for assistance with setting up the lattice decoration protocol. We acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme Grant No. 773122 (LIMA) and Grant agreement No. 951786 (NOMAD CoE). K.S.T. is a Villum Investigator supported by VILLUM FONDEN (grant no. 37789).

AUTHOR CONTRIBUTIONS

P.L. and K.S.T. developed the initial concept. P.L. ran the generative models, the DFT simulations and performed the data analysis. K.S.T. supervised the project and aided with the interpretation of the results. P.L. and K.S.T. wrote and discussed the paper together.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00923-3.

Correspondence and requests for materials should be addressed to Peder Lyngby.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.