Symmetry-based computational search for novel binary and ternary 2D materials

To cite this article: Hai-Chen Wang *et al* 2023 *2D Mater.* **10** 035007

View the article online for updates and enhancements.

2D Materials

PAPER

Symmetry-based computational search for novel binary and ternary 2D materials

Hai-Chen Wang[1] , Jonathan Schmidt[1] , Miguel A L Marques[1,*] , Ludger Wirtz[2]
and Aldo H Romero[2,3,*]

[1] Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, D-06099 Halle, Germany
[2] Department of Physics and Materials Science, University of Luxembourg, 162a Avenue de la Faïencerie, L-1511 Luxembourg, Luxembourg
[3] Department of Physics, West Virginia University, Morgantown, WV 26506, United States of America
* Authors to whom any correspondence should be addressed.

E-mail: miguel.marques@physik.uni-halle.de and Aldo.Romero@mail.wvu.edu

Abstract

We present a symmetry-based systematic approach to explore the structural and compositional richness of two-dimensional materials. We use a 'combinatorial engine' that constructs candidate compounds by occupying all possible Wyckoff positions for a certain space group with combinations of chemical elements. These combinations are restricted by imposing charge neutrality and the Pauling test for electronegativities. The structures are then pre-optimized with a specially crafted universal neural-network force-field, before a final step of geometry optimization using density-functional theory is performed. In this way we unveil an unprecedented variety of two-dimensional materials, covering the whole periodic table in more than 30 different stoichiometries of form A_nB_m or $A_nB_mC_k$. Among the discovered structures, we find examples that can be built by decorating nearly all Platonic and Archimedean tessellations as well as their dual Laves or Catalan tilings. We also obtain a rich, and unexpected, polymorphism for some specific compounds. We further accelerate the exploration of the chemical space of two-dimensional materials by employing machine-learning-accelerated prototype search, based on the structural types discovered in the systematic search. In total, we obtain around 6500 compounds, not present in previous available databases of 2D materials, with a distance to the convex hull of thermodynamic stability smaller than 250 meV/atom.

1. Introduction

Since the synthesis of single graphene layers [1], two-dimensional (2D) materials have attracted significant interest from the community. Their relevance extends to different research fields, such as catalysis, electronic transport, optical properties, and topological properties. However, the chemical space for 2D materials is still relatively unexplored, even though great effort has been spent on investigating the vast chemical space for bulk, three-dimensional (3D) compounds. In fact, experimental synthesis efforts have focused on a few structures, mostly obtained by exfoliation of known 3D layered materials [2].

On the computational side, we can find a few online databases of 2D materials, such as Materials Cloud two-dimensional crystals database (MC2D) [3], V2DB [4], 2DMatpedia [5], and the Computational 2D Materials Database (C2DB) [6–8]. These databases were built starting from 3D databases, by exfoliating single-layers from layered, van der Waals compounds. At the moment the vast majority of known 2D materials correspond to binaries [4, 9]. An exception is the very recent addition of materials discovered via a crystal diffusion variational autoencoder in [8].

These 2D databases are newer, and considerably smaller, than their three-dimensional counterparts,

e.g. Materials Project [10], the crystallographic open database [11], the Cambridge structural database [12], the NIST crystallographic database [13], OQMD [14], AFLOW [15], Materials Cloud [16], and many others. All these initiatives were seeded by experimental crystal structures stored in the inorganic crystal structure database (ICSD) and other experimental databases. In fact, the creation of the ICSD in 1912 [17–19] paved the way to the systematic study of the relationship between crystal structure and materials properties. To complement experimental data, many databases (both 2D and 3D) also contain results from high-throughput studies (often accelerated by machine learning).

High-throughput searches are responsible for a majority of the calculations in the large theoretical databases like AFLOW [15], OQMD [14] and DCGAT [20]. Traditional high-throughput searches rely on simple empirical rules to select candidate materials for evaluation with density functional theory (DFT). Consequently, they contain a large number of highly unstable systems. A particularly popular approach is prototype search, where new materials are hypothesized by changing the chemical elements in a known crystal structure (often stemming from ICSD). In some cases, all combinations of chemical elements are taken into account, while in other cases arguments based on charge neutrality, atomic or ionic radii, etc are used to circumvent the combinatorial nature of the problem.

In comparison to these rule-based selections, machine learning algorithms generally allow us to consider all combinations of the chemical elements due to their computational efficiency [21–24]. In fact, recent progress has enabled us to speed up the scanning of crystal prototypes by a factor of up to ~2000 [22] with respect to traditional DFT high-throughput studies. A second research direction are generative models that do not rely on existing prototypes. Here, generative adversarial networks [25, 26], variational auto encoders [27, 28] and, more recently, diffusion models [8, 29] are the most successful approaches. While these generative models have made great progress over the last year and improved with respect to their bias toward stable structures, the stability of the structure still has to be evaluated with a secondary machine learning model. No matter the generation or selection algorithm, the next step consists in a local structural optimization of each compound, invariantly using DFT as the workhorse method [30, 31]. Analysis of thermodynamic stability can then be achieved by computing the formation energy or the distance to the convex hull. In this way, databases have grown considerably and can now sometimes reach millions of crystal structures.

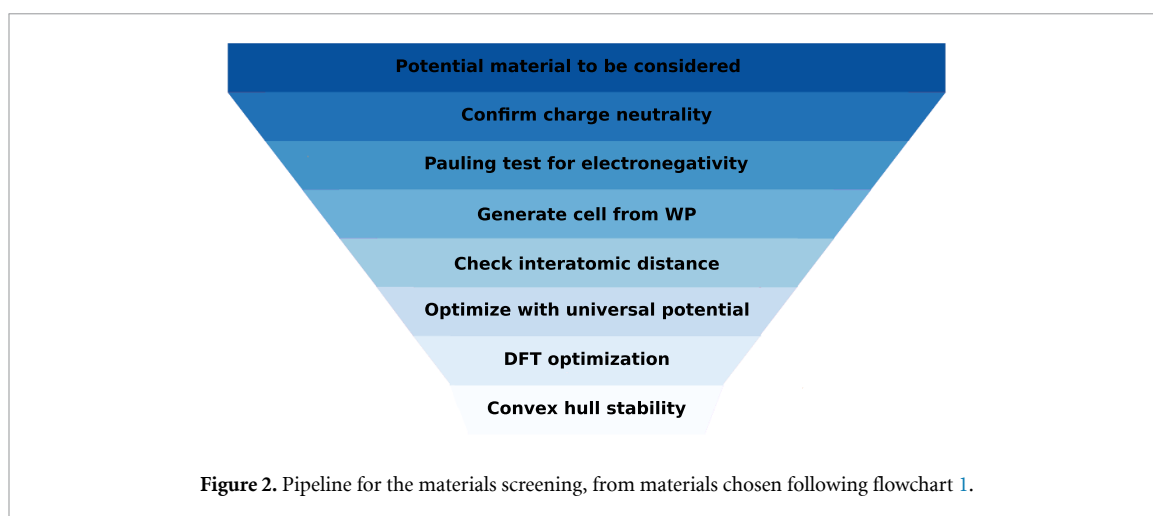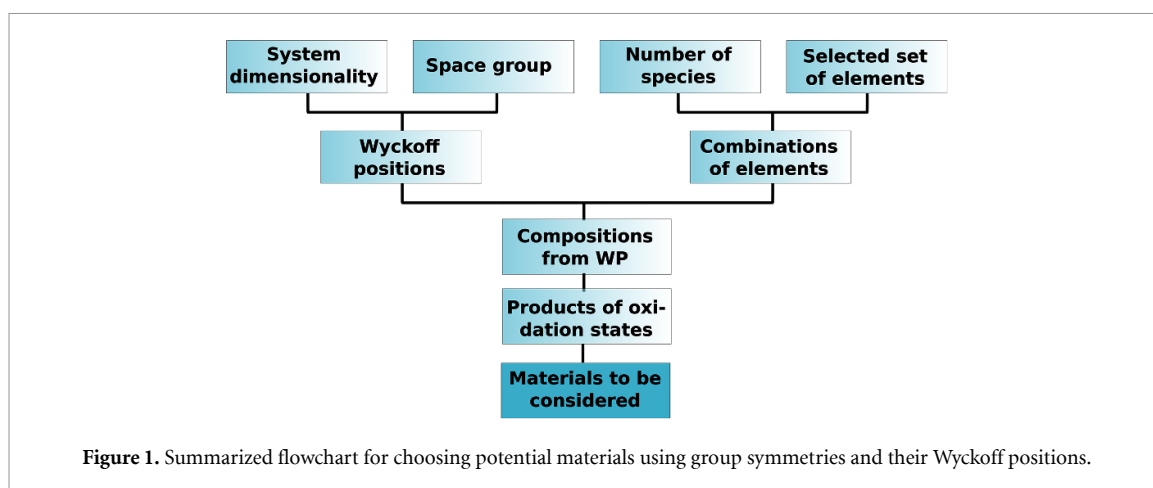While the success of using chemical combinatorics is recognized for 3D materials, it

has been a substantial handicap for predicting new 2D materials. The number of known 2D materials prototypes is unfortunately very small. Various research groups have considered different strategies to address this issue, often resorting to machine learning methods. This paper presents an entirely different approach which is not based on motifs or chemical substitutions. Instead, we create all possible combinations of chemical elements for binary (and ternary) systems for specific two-dimensional space groups. Therefore, based on symmetries and chemical criteria, we can arrive at a sizeable two-dimensional crystal structure database and a diverse set of structural prototypes. The crystal shapes show a variety of bondings and forms absent in existing databases. As our search is systematic, crystal structures contain a large number of different chemical formulae as well as almost all possible Wyckoff positions (WPs) allowed by the space groups. Here we focus on two-dimensional materials but our approach is general, and it can also be applied to three-, one-, or even zero-dimensional structures.

This article is structured as follows. We start by discussing in detail the systematic approach to discover 2D crystal structures and our strategies to accelerate the search. We then present an overview of the materials we discover, giving a few examples of structural diversity and polymorphism. In the following, we discuss machine-learning accelerated prototype search based on the wealth of prototypes obtained. In the appendix following the conclusions, we give details on our methodology.

2. Strategy

Figures 1 and 2 summarize our approach for the generation of materials. The first step corresponds to a combinatorial workflow that creates hypothetical compounds. The initial input parameter is the number of desired chemical species in the particular material. We consider most elements of the periodic table, from Li to Bi, including first-row rare earth elements. We exclude, however, radioactive elements and rare gases, namely At, Tc, Pr, Pm, He, Ne, Ar, Kr, Xe, and Rn. We then generate all possible combinations of the different elements selected from the periodic table. For example, a total of 2701 combinations are obtained with no repeated elements for a binary compound.

The second parameter is the two-dimensional space group. We use the table of layer group symmetries, created by considering the wallpaper group and adding reflections in the perpendicular direction. A full description of the possible groups is given in [32–34]. To generate the atomic positions provided by the layer space group for all possible WPs, we used the package PyxTal [35]. A list of the possible WPs can

Figure 1. Summarized flowchart for choosing potential materials using group symmetries and their Wyckoff positions.

Figure 2. Pipeline for the materials screening, from materials chosen following flowchart 1.

be found in the Bilbao Crystallographic Server. From the 80 layered groups, we studied the 18 that have the smallest number of different WPs and therefore the smallest number of combinations.

The next step is the creation of all possible combinations of the WPs for each chemical element in the combined list. For example, if the number of WPs is four, we get 4! possibilities. This strategy allows different WPs for the same species and therefore broadens the number of possible stoichiometries (i.e. a chemical species can occupy more than one WP). We then create a product of this list associated with the number of selected species. This selection leads to the definition of a chemical composition based on the occupation of the different WPs for each species in the compound. As certain WPs have free parameters, our approach is not exhaustive. For example, for the *p*1 space group we only occupy the (single) position 1*a* once for each atomic species. This leads to a single possibility for both the binary compound AB, and the ternary compound ABC. The number of possibilities increases, however, very rapidly with the number of different WPs available within the space group.

In parallel, we create a list of possible oxidation states of the considered species. We make all possible

combinations without replacement for each element from this list, and we create a product of the different list elements. We used the experimentally most common oxidation states, as they will have considerably larger potential to be synthesized (the selected oxidation states are included in the supplementary information). Finally, a compound is created from the provided number of species, the combination of WPs, and the oxidation state.

We have not imposed any explicit limit on the number of atoms in the unit cell. However, the procedure we use to generate the compounds does lead to an *implicit* constraint which, however, depends on the number and multiplicity of the WPs for each space group. For the space groups studied here, the maximum number of atoms in the unit cell is 32, although the majority of the compounds has less than 16 atoms in the unit cell.

After the material is obtained from the previous step, and before we perform a complete electronic structure calculation, we conduct a screening, which allows us to reduce the number of compounds to be fully considered. For the screening, we used rules implemented in the open-source material-screening Python package SMACT [36]. In this package, decisions are made based on stoichiometry. The

Table 1. Crystallographic summary of the layer groups considered in this work: the space group symbol (and number in parenthesis), the Wyckoff positions (and site symmetries in parenthesis). We also show the number of binary ($N_{tot}^{(2)}$) and ternary systems ($N_{tot}^{(3)}$) generated by our combinatorial engine and the number of entries that were found below 250 meV/atom from the convex hull of stability ($N_{<0.25}^{(2)}$ and $N_{<0.25}^{(3)}$).

Space group	Wyckoff positions	$N_{tot}^{(2)}$	$N_{<0.25}^{(2)}$	$N_{tot}^{(3)}$	$N_{<0.25}^{(3)}$
p1 (01)	1a (1)	225	5		
p11 m (04)	2b (1), 1a (..m)	1321	339		
p11a (05)	2a (1)	225	47	1944	307
p211 (08)	2c(1), 1b(2..), 1a(2..)	6645	623		
$p2_1$11 (09)	2a (1)	225	40	1944	273
c211 (10)	4b (1), 2a (2..)	1321	153		
pb11 (12)	2a (1)	225	15	1944	448
cm11 (13)	4b (1), 2a (m..)	1321	268		
$p2_1$/b11 (17)	4c (1), 2b (−1), 2a (−1)	3129	220		
$p2_12_1$2 (21)	4c(1), 2b(..2),2a(..2)	6645	398		
pb2_1m (29)	4b (1), 2a (..m)	1321	140		
pb2b (30)	4c (1), 2b (.2.), 2a (.2.)	6645	228		
pm2a (31)	4c (1), 2b (m..), 2a (.2.)	6645	478		
pm2_1n (32)	4b (1), 2a (m..)	1321	148		
pb2_1a (33)	4a (1)	225	19	1944	139
pb2n (34)	4b (1), 2a (.2.)	1321	74		
cm2e (36)	8c (1), 4b (m..), 4a (.2.)	6645	383		
p31m (70)	6d (1), 3c (..m), 2b (3..) 1a (3.m)	15 728	783		

first rule is to have only charge neutral compounds, which can be easily computed from the stoichiometry and the oxidation states. The second rule is the so-called Pauling test for materials which requires that positive ions have lower electronegativity than negative ions.

After screening, we use the main properties of a given material, such as oxidation states, stoichiometry, and WPs to generate the potential structures. In this step, we use the PyxTal utility [35] to create a 2D unit cell with the given number of species. When WPs have internal degrees of freedom, PyxTal tries to create a unit cell with the provided symmetry constraints. First, the cell directions are selected according to the space group. Then, the WPs are generated from the symmetry operations, and, if there are internal degrees of freedom, they are set randomly. Next, the cell parameters and the volume are determined, assuming that each atom has a radius equal to its covalent bond radius. Finally, a density is obtained from the cell volume and atomic masses, which is compared with a threshold density. If the cell density is smaller than 0.75 (in scaled units), the package attempts first to re-define the atomic positions randomly (setting the number of attempts to 40), and, in case this fails, it tries to change the cell parameters (up to ten times) and repeat the generation of the cell. If a cell cannot be defined in this way, the structure generation is considered unsuccessful, and the next candidate is considered. A summary of the pipeline is represented in figure 2. We generate systematically two dimensional structures for the space groups shown in table 1. In this table we also include the corresponding WPs, the site symmetry and the

number of different compounds generated for each space group.

The next step is the geometry optimization. Unfortunately, the initial structures are usually very far away from equilibrium, making structural optimization with DFT cumbersome. To increase the efficiency of our workflow we perform an intermediate geometry optimization step using a universal neural-network force-field [23]. In contrast to standard force fields that are usually trained to reproduce the potential energy surface of a specific system, universal neural network force fields describe all possible compounds. Of course, the objective of the latter is not to replace the former, that will be more precise but with a more limited applicability. Instead, they provide a reasonable description for all geometrical arrangements and chemical elements. Our model, trained using a transfer learning approach, has a median absolute error of 96 meV/atom for geometry optimizations. This is already a competitive value, suitable for describing 2D materials in this intermediate screening step.

At this point we remove from our dataset the materials that are too thick (using a threshold of 7.5 Å) or that are predicted to be too unstable by the machine learning model (more than 600 meV/atom from the hull, corresponding approximately to twice the mean absolute error (MAE) of the original model). We also remove structures that were already included in C2DB [7] (excluding the very recent structures of [8]).

The use of machine learning force fields resolves several technical problems: the pre-converged geometries are, in most cases, already quite good, only

requiring a few steps of geometry optimization using DFT. They also allow us to discard many repeated and very high-energy structures. After the DFT geometry optimization we evaluate the distance to the convex hull of stability. We use the convex hull of [20, 22] that is considerably larger than the one of the Materials Project [10], in particular in what concerns the ternary (and quaternary) sector. Consequently, our distances to the hull are sometimes larger than in other 2D databases.

Note that besides thermodynamic stability, the issue of dynamical stability is a crucial factor for 2D materials, and should always be verified before a specific material is proposed for synthesis. A material is dynamically stable when it exhibits no imaginary phonon frequencies across the Brillouin zone. Unfortunately, the calculation of the phonon dispersion is extremely time-consuming, and even more so for 2D systems due to issues related to the vacuum required to treat the long-range part of the Coulomb interaction. [37] We also note that imaginary phonon frequencies could be an indication of a charge-density wave phase (at even lower formation energy) which we might be overlooking due to the use of unit cells with a limited number of atoms.

3. 2D materials

It turns out that our workflow was able to arrive at the large majority of systems already present in C2DB. This is particularly true for binary systems, as these were more extensively investigated than ternaries (see table 1). This, in our opinion, fully validates our workflow.

Figure 3 presents a comparison of the binary materials present in our database (excluding the ones found in C2DB) compared to C2DB. For the discussion, we only took into account the materials that are within a distance of 250 meV/atom from the convex hull of stability, that corresponds loosely to the definition of 'high-stability' in C2DB [7]. Note that for consistency we have reoptimized the C2DB structures using our convergence criteria and our selected set of pseudopotentials.

We find 2D compounds across the whole periodic table, including some with lanthanides that have been up to now excluded from previous works. Not surprisingly, the majority of compounds includes a non-metal element (due to the requirement of charge neutrality), leading to the pronounced peaks for O, S, Se, Te, F, Cl, Br, I, etc. The figure also reveals some differences in the prevalence of certain elements between our dataset and C2DB. For example, we find considerably more compounds with F than with O, while in C2DB we observe the opposite behavior. As our approach is to a large extent systematic in what concerns chemical compositions and geometries, we

believe that the differences are explained by a bias already present in ICSD and other databases that were used to seed the 2D databases. For example, it is well known that oxides are over-represented in experimental works as they can be more easily synthesized and are often stable in air.

Other conclusions can be drawn from figure 3. For example, it can be seen clearly that the non-metals in the second row have more difficulty in forming low-energy compounds than other non-metals in the same group. This is a consequence of the Singularity Principle [38], i.e. that the chemistry of the these elements is often different to the later members of their respective groups. Furthermore, elements like N, O, C, and F form very strong directional covalent bonds that leave comparatively little room for distortions that would be required to form different structures. As for metallic elements, it is in particular the transition elements in the fourth row from Ti to Cu (and in particularly this last one), together with late group III-A (In and Tl) seem to form easily 2D compounds.

The diversity of stoichiometries is illustrated in figure 4. As our emphasis has been on binary compounds, it is not surprising that most represented stoichiometries are binary. Among these, the simple AB_2, AB_3, A_2B_3, etc dominate the low-energy structures. This fact can be easily understood by the requirement of charge neutrality and the fact that most non-metals have oxidation states of -I, -II, or -III. As such, the same situation can be found for bulk, 3D semiconductors and insulators. However, we do find a long list of other stoichiometries (more than 30), and these often reveal very interesting and unexpected structures.

In figure 5 we give a glimpse of the diversity of structural motifs found by our method. Note that this is far from a complete list of all 2D structures found. We concentrate on unusual arrangements that go beyond the most common square and hexagonal lattices. We emphasize that these motifs appeared naturally in our workflow and were not constructed by hand. Interestingly, we easily found examples that can be derived from the majority of the different Euclidean uniform tilings, both Platonic and Archimedean as well as their dual Laves or Catalan tilings. Moreover, many of these tilings seem to be unique to the two-dimensional world, as no layered 3D material is known to possess them.

The first two structures can be derived from a truncated square and a rhombic tiling. In the first case, Cs_2Br_2 squares are connected, forming regular empty octahedra, leading to a rather open lattice. In the second, Se_3O_3 rectangular units form bonds along the corners, leading to flattened octahedra. We then present an example of a Pythagorean tiling, a motif that is composed of two different squares
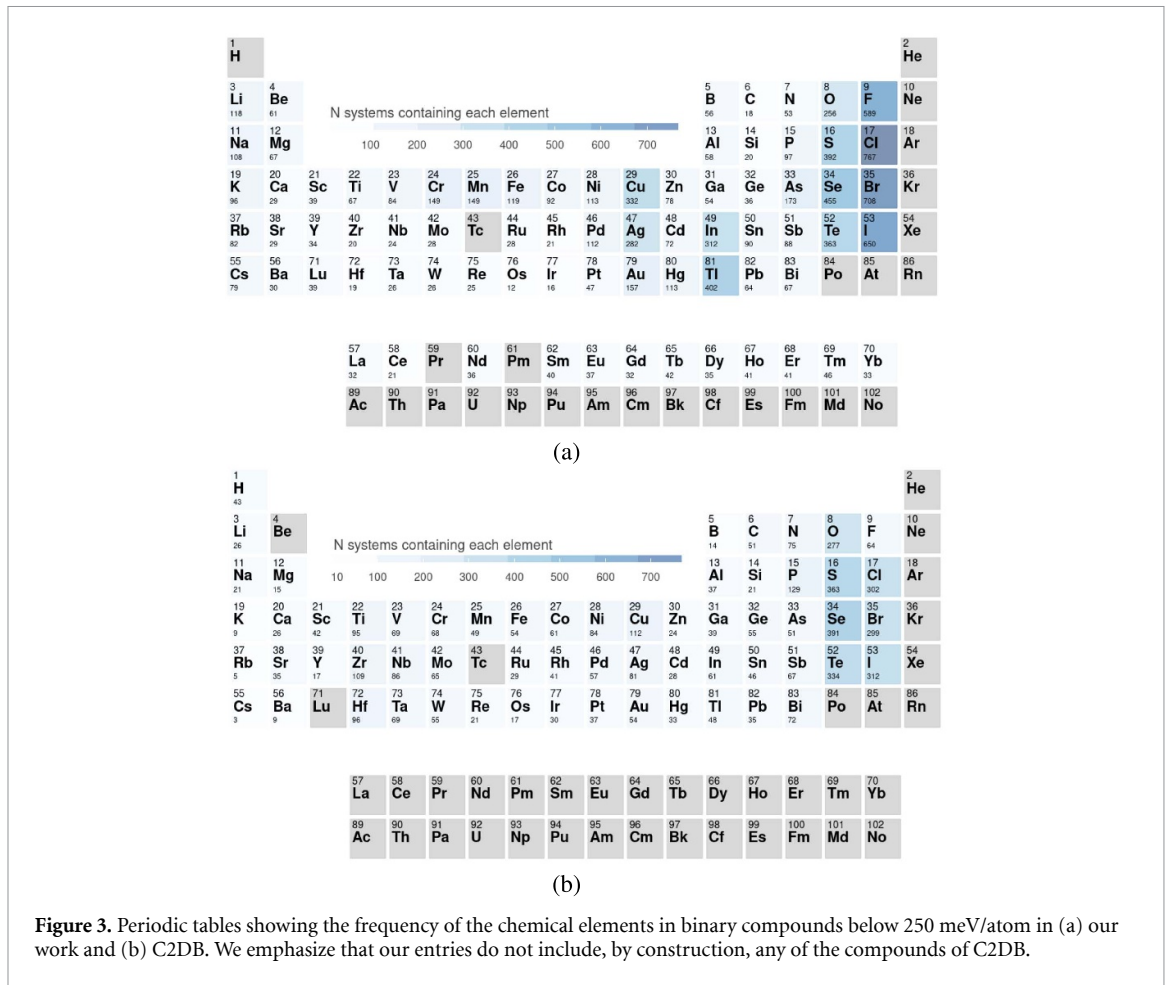
Figure 3. Periodic tables showing the frequency of the chemical elements in binary compounds below 250 meV/atom in (a) our work and (b) C2DB. We emphasize that our entries do not include, by construction, any of the compounds of C2DB.
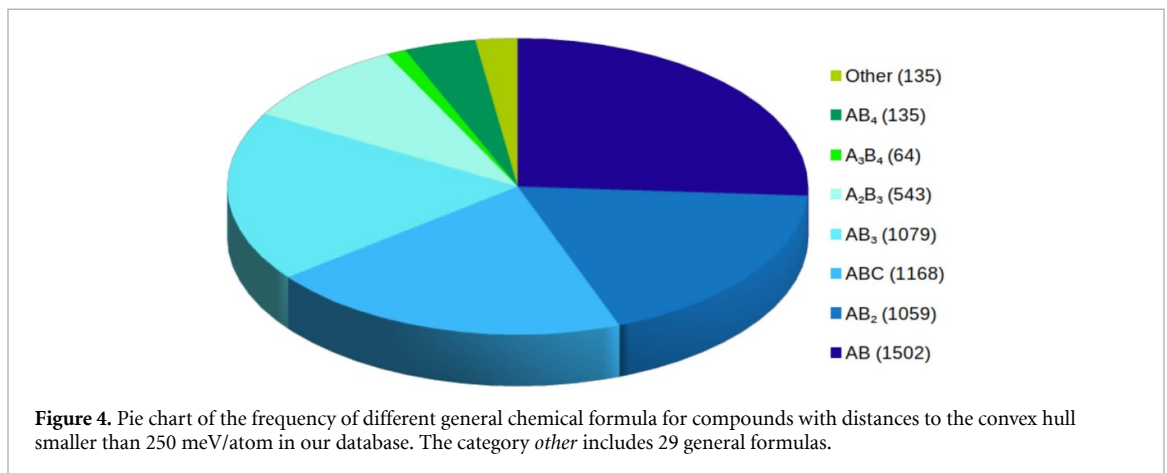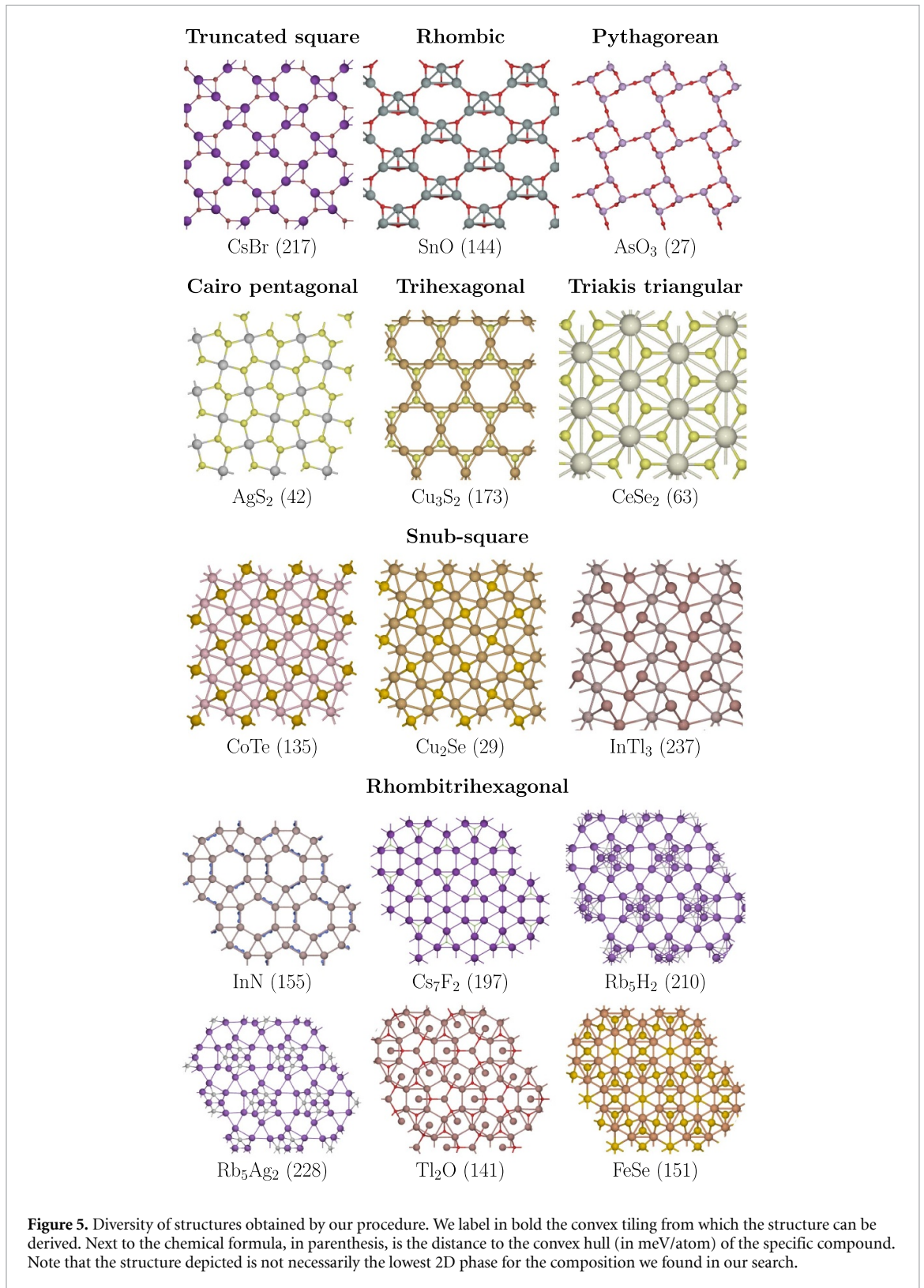

Figure 4. Pie chart of the frequency of different general chemical formula for compounds with distances to the convex hull smaller than 250 meV/atom in our database. The category *other* includes 29 general formulas.
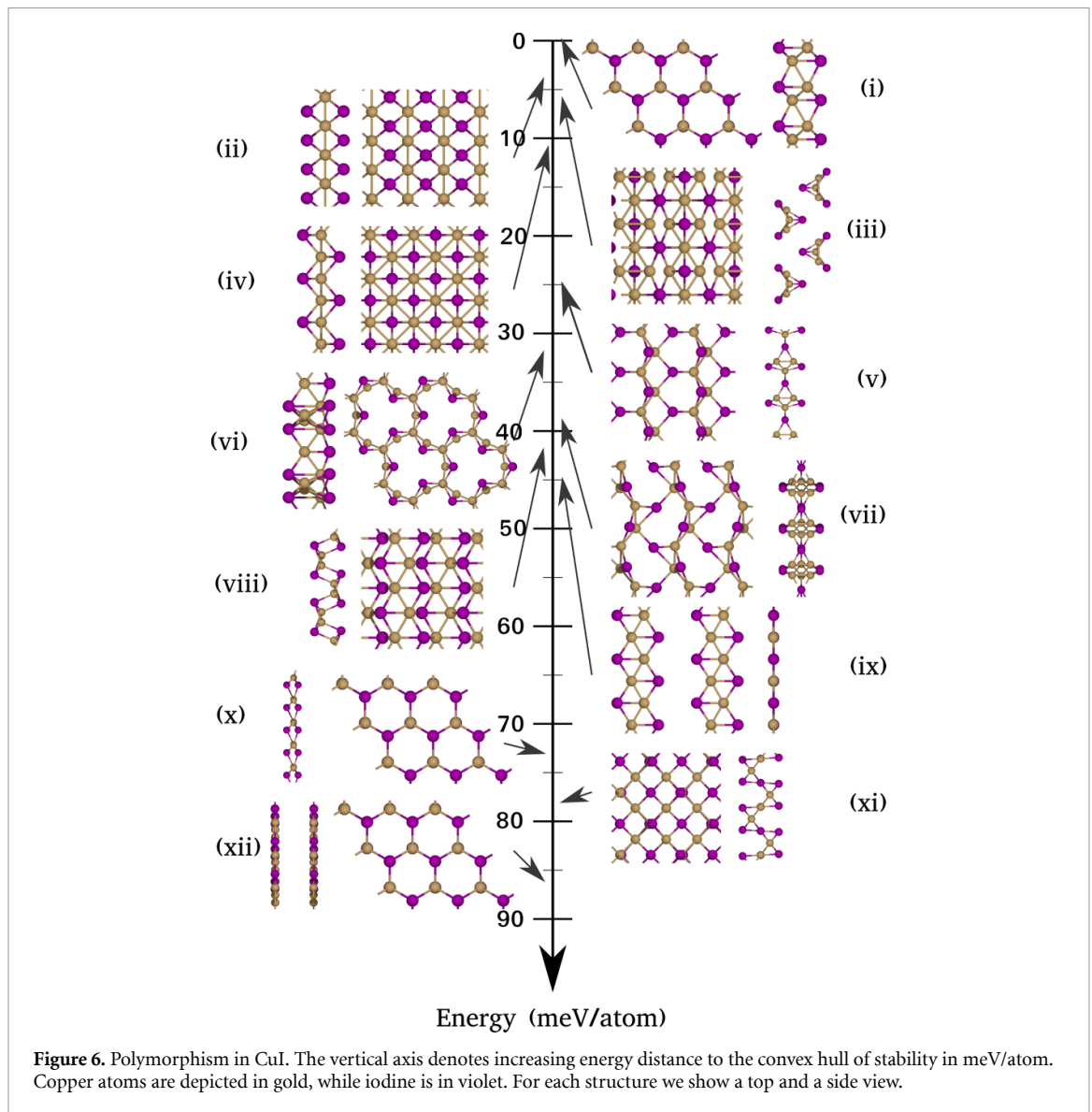
that share one side, and can be found all over the world in kitchen or garage floors. Interestingly, it was proposed recently that elementary, two-dimensional Cl, Br, and I might be able to adopt this arrangement [39]. We also find a series of Cairo pentagonal tilings. In this example, AgS_2 forms two overlaying tessellations of the plane by irregular hexagons, where each of the hexagons is formed by four identical pentagons. At the center of the hexagons we find a Se–Se bond. Note that this is the same Cairo pentagonal tiling that was found for $PdSe_2$ [40–42]. One of the possible structures of Cu_3S_2 consists on a trihexagonal tiling (that is often called the Kagome lattice due to its use in traditional Japanese basketry) of the plane by Cu atoms, decorated by a S atom in the middle of the triangles. Triakis triangular lattices appear quite commonly in our data. The example in figure 5 can be seen as composed of Ce equilateral triangles decorated with a Se atom at its center.

Figure 5. Diversity of structures obtained by our procedure. We label in bold the convex tiling from which the structure can be derived. Next to the chemical formula, in parenthesis, is the distance to the convex hull (in meV/atom) of the specific compound. Note that the structure depicted is not necessarily the lowest 2D phase for the composition we found in our search.

The following three examples are derived from snub-square lattices. In the first two, the metal forms this interesting square-triangle lattice and the non-metal decorates the squares. In the case of CoTe, Te-atoms can be found above and below the plane of the Co atoms, while in Cu_2Se, Se-atoms alternate above and below the plane of the Cu-atoms. We note that

this specific lattice was recently proposed for some noble metal chalcogenide monolayers [43] and for certain Ba and Ti oxides [44]. The third example is more complex, as both In and Tl form a distorted version of the snub-square lattice, with further Tl-atoms alternating above and below the plane. Note, however, that this curious structure is almost at the

Figure 6. Polymorphism in CuI. The vertical axis denotes increasing energy distance to the convex hull of stability in meV/atom. Copper atoms are depicted in gold, while iodine is in violet. For each structure we show a top and a side view.

limit of our energy threshold. Finally, we present six examples of rhombitrihexagonal lattices, where the metal atoms form the triangle-square-hexagon lattice that is then decorated (mostly) by the non-metals. We found a very diverse number of different decorations, allowing for many different stoichiometries, ranging from the simple AB and AB_2 to the more unconventional A_2B_5 and A_2B_7. A very interesting possibility raised by the finding of all these snub-square and rhombitrihexagonal structures is that these can be easily inflated by a recursive approach to generate quasi-crystalline systems [45, 46]. This is, however, only possible for structures not including out-of-plane alternating atoms, as this induces frustration in the system reducing its stability [44].

We have given several examples of the different stoichiometries in our dataset and of the structural variety stemming from them. Now, we look at the issue of polymorphism, i.e. the different phases possible for a specific chemical composition. Not surprisingly, polymorphism depends strongly on the

chemical elements present in the compound. For example, for BN we found a single structure below 250 meV/atom, the well-known honeycomb lattice, while for other compounds we have an extraordinary variety in the same energy range.

As an example, we show in figure 6(a) selection of the crystal structures that we found for CuI. We recall that zincblende CuI is at the moment the most promising *p*-type transparent conducting semiconductor [47]. However, CuI has a number of polymorphs, including a couple of trigonal phases [48–51] that are layered, with a bonding pattern rather different from the γ-phase. Figure 6 shows that also in the 2D case, we find a large variety of structures and of bonding patterns.

As the lowest-energy 2D layer we find a covalently bound hexagonal double-layer (i) that is essentially on the convex hull of thermodynamic stability. The (buckled) single layer (x) and the van-der-Waals bound double flat-layer also appear in the energy spectrum but considerably higher, at more than

70 meV/atom. The second most stable structure is, surprisingly, a rectangular lattice of Cu–I (ii), with the I-atoms alternating above and below the plane of the Cu-atoms. A related lattice (iv) appears just a few meV/atom above. Structure (iii), which is only 6 meV/atom above the hull, and structure (ix) are arrangements of one-dimensional objects. The first exhibits nanowires with a triangular section arranged in an alternating fashion as depicted in figure 6. The latter (ix) is a periodic arrangements of nanostripes. (Incidentally, higher in energy, at 161 meV/atom, we even find a molecular crystal of Cu_4I_4 pyramidal clusters.) All these systems turn out to be semiconducting, with calculated (PBE) band gaps ranging from around 0.5 eV to more than 2.1 eV.

4. Prototype search

The biggest advantage of the workflow presented above is that it is (i) systematic and (ii) unbiased in what concerns the structural variety. Unfortunately, the price to pay for these advantages is efficiency, in the sense that it is computationally expensive to go through all possible compositions and space groups and that many of the possibilities turn out to be highly unstable or lead to thick slabs. It is, however, possible to accelerate considerably the exploration of the 2D material space by using the structural prototypes discovered by our approach, and combining them with a machine-learning model appropriate for prototype search [52, 53]. Of course, in this way we will not discover new structural motifs, but we can explore the whole compositional space very efficiently.

Our approach follows the same basic principles as V2DB [4], but goes beyond it in a number of different directions. First, we perform transfer learning from a 3D machine, which allows to transfer many of the chemical principles that govern atomic bonding. Second, we use a much larger training set, increasing the accuracy of the machine. Third, we lift several constrains (like charge neutrality or electronegativity rules) used in V2DB and in our systematic search, and we expand the possible chemical elements to the whole periodic table. This allows us to discover a variety of intermetallics and compounds combining elements with unusual oxidation states. We furthermore perform machine-learning predictions for all two-dimensional prototypes, either already present in C2DB or stemming from our systematic search. Finally, we perform validation DFT calculations for some of the predictions, specifically for the binary stoichiometries A_2B_5, A_2B_7, and the ternary ABC_2, ABC_3, AB_2C_2, $A_2B_2C_3$.

Note that, in contrast to the systematic generation of structures based on the space groups, in the machine-learning assisted prototype search, we do not impose any constraint on the possible oxidation states. As such, the machine can, and does, propose

2D systems including chemical elements in other, less common oxidation states.

To keep the number of structures manageable, we asked the machine to output all structures that it found below 200 meV/atom for the binaries and 50 meV/atom for the ternaries. In total, we obtained 1023 candidates that were pre-optimized with our neural-network force-field and then optimized with DFT. From these 638 were found to be below 250 meV/atom from the hull, yielding an exceptional success rate of around 62%. The lowest success rate, of only 9%, was found for the A_2B_7 stoichiometry: as these compounds were sparsely present in the training set, the machine could not learn the specificity of those structures. The problem can, of course, be solved by adding further samples to the dataset, in order to remove the structural (and compositional) bias, as previously shown for bulk systems in [22]. We are currently performing DFT calculations for $\sim$ 40000 more materials, resulting from 238 million machine-learning predictions, that will be available in the next release of our dataset.

5. Conclusions

We have presented a systematic approach to explore the structural and compositional diversity that is possible in the chemical space of 2D materials. The main advantage of this approach is that it is not based on a specific number of structural prototypes. This is particularly important for 2D materials, as the space of possible structures is still rather unexplored and only few prototype structures, mainly from exfoliation of layered 3D materials, are known so far. In this way, we have discovered thousands of unexpected phases that have no counterpart in the world of layered three-dimensional materials. We expect that such unusual bonding and geometrical patterns will also lead to unique mechanical, electronic, optical, and magnetic properties.

Our method relies heavily on the use of machine learning. The extensive use of neural networks in several parts of our workflow is self-accelerating. In fact, the faster we generate more data for two-dimensional systems, the larger will be our training sets, resulting in even more accurate machine learning models. This leads to a virtuous cycle that, in our opinion, will pave the way for a rather complete exploration of the binary, ternary, and eventually also the quaternary two-dimensional phases in the near future.

Finally, an important question is how many of the phases in our dataset can be synthesized. We chose to filter our results to include only compounds with an energy less than 250 meV/atom above the convex hull, as these have higher stability and therefore higher probability to be synthesized. (For comparison, silicene, that has been experimentally synthesized [54],

is more than 600 meV/atom above the hull in its free-standing form.) However, besides these thermodynamic arguments, a key factor will be the choice of suitable substrates that stabilize the two-dimensional layers, and, of course, the ingenuity of experimental physicists and chemists to design targeted synthesis strategies.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.24435/materialscloud:sb-cy. The database, including structures, distances to the hull, and other basic properties, can be accessed at https://tddft.org/bmg/physics/2D/ through a simple web-based interface.

Author contributions

A H R performed the systematic structure generation; M A L M and H C W performed the DFT high-throughput calculations; J S performed the training of the machines and the machine learning predictions of the distance to the hull; A H R, L W, and M A L M directed the research; all authors contributed to the analysis of the results and to the writing of the manuscript.

Conflict of interest

The authors declare that they have no competing interests.

Appendix. Methods

DFT calculations

We performed all geometry optimizations and total energy calculations with the code VASP [55, 56]. The 2D Brillouin zones were sampled by uniform Γ-centered k-point grids with a density of 6 k-points Å^{-2}. We performed spin-polarized calculations starting from a ferromagnetic state, and used the projector augmented wave setups [57, 58] of VASP version 5.2 with a cutoff of 520 eV. We converged the calculations to forces smaller than 0.005 eV Å^{-1}. As exchange-correlation functional we used the Perdew–Burke–Ernzerhof [59] functional with on-site corrections for oxides and fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V, or W. The repulsive on-site corrections to the d-states were 3.32, 3.7, 5.3, 3.9, 4.38, 6.2, 3.25, and 6.2 eV, respectively. These parameters were chosen to be compatible with the Materials Project database [10]. We imposed a vacuum region of 15 Å, and systems that resulted in structures with a thickness greater than 7.5 Å were automatically discarded. Finally, as it is common in this kind of approaches, some of the calculations did not converge due to a multitude of reasons. The corresponding phases were then simply eliminated from the dataset.

Distances to the convex hull were evaluated using PYMATGEN [60] using the large complex hull of [22] corresponding to the dataset available in the Materials Cloud repository [20].

M3GNET

We employed the universal neural-network force-field M3GNET [23] that was developed to reproduce the energies and the forces of bulk structures with remarkable results. As a starting point we used the pretrained network distributed with M3GNET. We tested this model on 1300 of our systems by measuring the difference between the energy calculated with M3GNET (at the M3GNET relaxed structure) and the energy calculated with DFT (at the DFT relaxed structure). We arrived at a MAE of 320 meV/atom and a median absolute error of 223 meV/atom. These numbers are already rather small, especially when we consider that the training set of M3GNET did not include 2D systems that can exhibit very different bonding patterns compared to bulk structures. As soon as enough data was available from our own simulations, we used transfer learning techniques to retrain
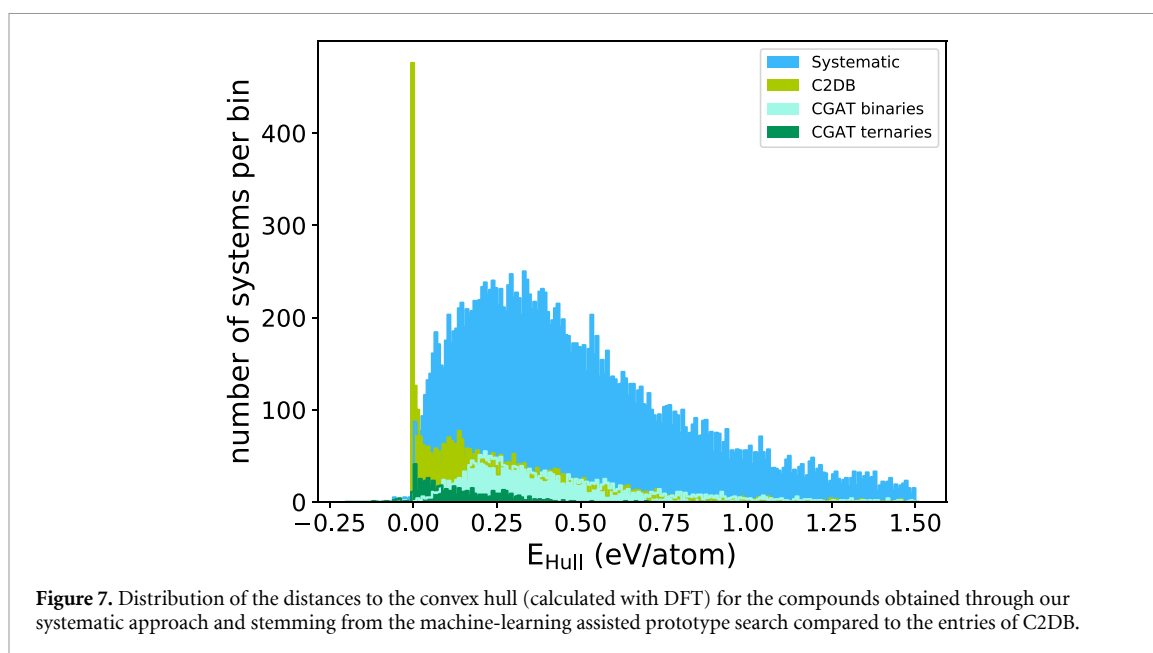
M3GNET for 2D materials (see appendix). Specifically, we build a dataset comprising energies, forces, and stresses from structures calculated during the geometry optimization steps. Structures with extremely high forces above 50 eVÅ^{-1} were removed from the data as were structures with no neighboring atoms inside the cutoff radius to avoid errors during training. To balance the training set, for systems with more than 4 recorded geometry optimization steps only the first, last and $N_{steps}/3$ step were used. The final training set for M3GNET contained 11 612 geometry relaxations corresponding to 34 944 energies and structures. The resulting network had a validation MAE of 61 meV/atom for direct energy predictions after training. The test errors for geometry optimizations on the same dataset as the pretrained model were 198 meV/atom for the MAE and 96 meV/atom for the median absolute error proving the efficiency of our transfer learning strategy. Of course, we expect these errors to decrease further simply by adding more data to the training set. The models were trained with the base hyperparameters from M3GNET and by setting the loss function of the stress in the non-periodic direction to zero.

Crystal-graph attention networks
We used the crystal-graph attention neural networks developed in [61] as they were specifically crafted for prototype searches. In particular, they require as input only the (unrelaxed) structural prototype and not accurate relaxed structures. Of course, this model was trained on bulk 3D structures, so we do not expect it to perform accurately in our case. However, many of the bonding patterns present in our 2D materials can already be found in the 3D world. To take advantage of this, we performed transfer learning of the 3D model, using the 2D structures in our dataset as training data. We used a dataset of DFT calculations with 22 007 entries, 80% of which were used for training 10% for validation and 10% for testing. Evaluating both models on the test set, we arrive at an MAE of 222 meV/atom for the original model and 86 meV/atom for the model transferred to the 2D data.

In figure 7 we present a histogram with the distances to the convex hull of stability calculated with DFT. The C2DB data is in light green, and is highly peaked at zero, decaying slowly for larger energies. This is expected, as C2DB was seeded with stable 3D, van der Waals bonded structures from ICSD. In blue we depict the structures obtained through our systematic approach. These form a continuous distribution with a peak at around 300 meV/atom, and extending beyond 1.5 eV. Knowing that such distribution for random compounds can extend beyond 4 eV, we see how the charge and electronegativity constraints lead to relatively stable compounds (at the price of overlooking intermetallics or compounds with unusual oxidation states. In light blue we show the machine-learning binaries predicted to be within 200 meV/atom from the hull. This shows a peak at around that value, as expected form the cutoff, then decaying similarly to the C2DB data. The ternary entries, displayed in green, are shifted to much lower energy, as expected by the smaller cutoff of 50 meV/atom. This results are consistent with the MAE of 86 meV/atom for the 2D model. The CGAT-hyperparameters are listed in the supplementary material and the code can be found at https://github.com/hyllios/CGAT.git.

Figure 7. Distribution of the distances to the convex hull (calculated with DFT) for the compounds obtained through our systematic approach and stemming from the machine-learning assisted prototype search compared to the entries of C2DB.

ORCID iDs

Hai-Chen Wang ⬤ https://orcid.org/0000-0002-2892-5879

Jonathan Schmidt ⬤ https://orcid.org/0000-0001-5685-6404

Miguel A L Marques ⬤ https://orcid.org/0000-0003-0170-8222

Ludger Wirtz ⬤ https://orcid.org/0000-0001-5618-3465

Aldo H Romero ⬤ https://orcid.org/0000-0001-5968-0571

References

[1] Novoselov K S, Geim A K, Morozov S V, Jiang D, Zhang Y, Dubonos S V, Grigorieva I V and Firsov A A 2004 *Science* **306** 666

[2] Novoselov K S, Jiang D, Schedin F, Booth T J, Khotkevich V V, Morozov S V and Geim A K 2005 *Proc. Natl Acad. Sci. USA* **102** 10451

[3] Mounet N *et al* 2018 *Nat. Nanotechnol.* **13** 246

[4] Sorkun M C, Astruc S, Koelman J and Er S 2020 *npj Comput. Mater.* **6** 1

[5] Zhou J *et al* 2019 *Sci. Data* **6** 1

[6] Haastrup S *et al* 2018 *2D Mater.* **5** 042002

[7] Gjerding M N *et al* 2021 *2D Mater.* **8** 044002

[8] Lyngby P and Thygesen K S 2022 *npj Comput. Mater.* **8** 232

[9] Song Y, Siriwardane E M D, Zhao Y and Hu J 2021 *ACS Appl. Mater. Interfaces* **13** 53303

[10] Jain A *et al* 2013 *APL Mater.* **1** 011002

[11] Gražulis S, Chateigner D, Downs R T, Yokochi A, Quirós M, Lutterotti L, Manakova E, Butkus J, Moeck P and Le Bail A 2009 *J. Appl. Crystallogr.* **42** 726

[12] Allen F H 2002 *Acta Crystallogr.* B **58** 380

[13] Van Hove M A, Hermann K and Watson P 2002 *Acta Crystallogr.* B **58** 338

[14] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501

[15] Curtarolo S *et al* 2012 *Comput. Mater. Sci.* **58** 218

[16] Talirz L *et al* 2020 *Sci. Data* **7** 1

[17] Bergerhoff G and Brown I *et al* 1987 *Crystallographic Databases* vol 360, ed F Allen (Chester: International Union of Crystallography) pp 77–95

[18] Zagorac D, Müller H, Ruehl S, Zagorac J and Rehme S 2019 *J. Appl. Crystallogr.* **52** 918

[19] Belsky A, Hellenbrandt M, Karen V L and Luksch P 2002 *Acta Crystallogr.* B **58** 364

[20] Schmidt J, Hoffmann N, Wang H-C, Borlido P, Carriço P J M A, Cerqueira T F T, Botti S and Marques M A L 2022 Large-scale machine-learning-assisted exploration of the whole materials space Materials Cloud (https://doi.org/10.24435/materialscloud:m7-50)

[21] Park C W and Wolverton C 2020 *Phys. Rev. Mater.* **4** 063801

[22] Schmidt J, Hoffmann N, Wang H-C, Borlido P, Carriço P J M A, Cerqueira T F T, Botti S and Marques M A L 2023 Machine-learning-assisted determination of the global zero-temperature phase diagram of materials *Adv. Mater.* (https://doi.org/10.1002/adma.202210788)

[23] Chen C and Ong S P 2022 *Nat. Comput. Sci.* **2** 718

[24] Goodall R E, Parackal A S, Faber F A, Armiento R and Lee A A 2022 Rapid discovery of novel materials by coordinate-free coarse graining *Sci. Adv.* **8** eabn4117

[25] Zhao Y, Al-Fahdi M, Hu M, Siriwardane E M, Song Y, Nasiri A and Hu J 2021 *Adv. Sci.* **8** 2100566

[26] Long T, Fortunato N M, Opahle I, Zhang Y, Samathrakis I, Shen C, Gutfleisch O and Zhang H 2021 *npj Comput. Mater.* **7** 66

[27] Noh J, Kim J, Stein H S, Sanchez-Lengeling B, Gregoire J M, Aspuru-Guzik A and Jung Y 2019 *Matter* **1** 1370

[28] Ren Z *et al* 2022 *Matter* **5** 314

[29] Xie T, Fu X, Ganea O-E, Barzilay R and Jaakkola T 2021 arXiv:2110.06197

[30] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864

[31] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133

[32] Kopsky V and Litvin D eds 2002 *Volume E: Subperiodic Groups* International Tables for Crystallography (Dordrecht: Springer)

[33] Aroyo M I, Perez-Mato J M, Orobengoa D, Tasci E, de la Flor G and Kirov A 2011 *Bulg. Chem. Commun.* **43** 183

[34] Aroyo M I, Kirov A, Capillas C, Perez-Mato J and Wondratschek H 2006 *Acta Crystallogr.* A **62** 115

[35] Fredericks S, Parrish K, Sayre D and Zhu Q 2021 *Comput. Phys. Commun.* **261** 107810

[36] Davies D W, Butler K T, Jackson A J, Morris A, Frost J M, Skelton J M and Walsh A 2016 *Chem* **1** 617

[37] Sohier T, Calandra M and Mauri F 2017 *Phys. Rev.* B **96** 075448

[38] Restrepo G, Mesa H, Llanos E J and Villaveces J L 2004 *J. Chem. Inf. Model.* **44** 68

[39] Huran A W, Wang H-C, San-Miguel A and Marques M A L 2021 *J. Phys. Chem. Lett.* **12** 4972

[40] Lv P, Tang G, Liu Y, Lun Y, Wang X and Hong J 2021 *Extreme Mech. Lett.* **44** 101231

[41] Oyedele A D *et al* 2017 *J. Am. Chem. Soc.* **139** 14090

[42] Lei W, Zhang S, Heymann G, Tang X, Wen J, Zheng X, Hu G and Ming X 2019 *J. Mater. Chem.* C **7** 2096

[43] Gao L, Zhang Y-F and Du S 2021 *Nano Res.* **14** 2826

[44] Huran A W, Wang H-C and Marques M A L 2021 *2D Mater.* **8** 045002

[45] Förster S, Meinel K, Hammer R, Trautmann M and Widdra W 2013 *Nature* **502** 215

[46] Förster S *et al* 2020 *Phys. Status Solidi* b **257** 1900624

[47] Grundmann M, Schein F-L, Lorenz M, Böntgen T, Lenzner J and von Wenckstern H 2013 *Phys. Status Solidi* a **210** 1671

[48] Kurdyumova R and Baranova R 1961 *Kristallografiya* **6** 402

[49] Sakuma T 1988 *J. Phys. Soc. Japan* **57** 565

[50] Villars P and Cenzual K (eds) 2009 *Landolt-Börnstein— Group III Condensed Matter: Crystal Structures of Inorganic Compounds* (Berlin: Springer)

[51] Akopyan I K, Golubkov V V, Dyatlova O A, Mamaev A N, Novikov B V and Tsagan-Mandzhiev A N 2010 *Phys. Solid State* **52** 805

[52] Schmidt J, Marques M R G, Botti S and Marques M A L 2019 *npj Comput. Mater.* **5** 83

[53] Kulik H J *et al* 2022 *Electron. Struct.* **4** 023004

[54] Vogt P, De Padova P, Quaresima C, Avila J, Frantzeskakis E, Asensio M C, Resta A, Ealet B and Le Lay G 2012 *Phys. Rev. Lett.* **108** 155501

[55] Kresse G and Furthmüller J 1996 *Comput. Mater. Sci.* **6** 15

[56] Kresse G and Furthmüller J 1996 *Phys. Rev.* B **54** 11169

[57] Blöchl P E 1994 *Phys. Rev.* B **50** 17953

[58] Kresse G and Joubert D 1999 *Phys. Rev.* B **59** 1758

[59] Perdew J P, Burke K and Ernzerhof M 1996 *Phys. Rev. Lett.* **77** 3865

[60] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 *Comput. Mater. Sci.* **68** 314

[61] Schmidt J, Pettersson L, Verdozzi C, Botti S and Marques M A L 2021 *Sci. Adv.* **7** 49