

# Machine-Learning-Assisted Determination of the Global Zero-Temperature Phase Diagram of Materials

Jonathan Schmidt, Noah Hoffmann, Hai-Chen Wang, Pedro Borlido, Pedro J. M. A. Carriço, Tiago F. T. Cerqueira, Silvana Botti,\* and Miguel A. L. Marques

Crystal-graph attention neural networks have emerged recently as remarkable tools for the prediction of thermodynamic stability. The efficacy of their learning capabilities and their reliability is however subject to the quantity and quality of the data they are fed. Previous networks exhibit strong biases due to the inhomogeneity of the training data. Here a high-quality dataset is engineered to provide a better balance across chemical and crystal-symmetry space. Crystal-graph neural networks trained with this dataset show unprecedented generalization accuracy. Such networks are applied to perform machine-learning-assisted high-throughput searches of stable materials, spanning 1 billion candidates. In this way, the number of vertices of the global  $T = 0$  K phase diagram is increased by 30% and find more than  $\approx 150\,000$  compounds with a distance to the convex hull of stability of less than  $50\text{ meV atom}^{-1}$ . The discovered materials are then accessed for applications, identifying compounds with extreme values of a few properties, such as superconductivity, superhardness, and giant gap-deformation potentials.

systematic exploration of a chemical space spanning millions of materials, searching for compounds with tailored properties for specific technological applications.<sup>[1–4]</sup> Currently, the most efficient approach to predict stoichiometric, ordered compounds consists in scanning the composition space for a fixed crystal structure prototype.<sup>[5–7]</sup> In such approaches, the key material property that is used to estimate if a material can be experimentally synthesized is the total energy, or more specifically the energy distance to the convex hull of thermodynamic stability.<sup>[6,8–17]</sup> Typically, given a chemical composition and a crystal-structure prototype (i.e., the combination of a Bravais lattice and a set of occupied Wyckoff positions) one performs a geometry optimization, for example, using some flavor of density functional theory (DFT), and compares the resulting

DFT energy with all possible decomposition channels.<sup>[18,19]</sup> Compounds on the convex hull (or close to it) are then selected for characterization and, if they possess interesting physical or chemical properties, proposed for experimental synthesis. Nonetheless, synthesis reactions are extremely complex processes and while the distance to the convex-hull correlates with synthesizability, it is not enough for deciding whether or not a material is experimentally accessible. Several recent works address this problem by directly predicting optimal synthesis conditions or probability of synthesis.<sup>[20–25]</sup>

For binary compounds the construction of the convex hull of thermodynamic stability is relatively straightforward, and therefore the binary phase space has been comprehensively explored.<sup>[26]</sup> For a single ternary compound,  $A_xB_yC_z$  the number of different combinations of chemical elements A, B, and C amounts to roughly 500 000, a value still within reach of DFT calculations, at least for crystals with high symmetry and relatively few atoms in the unit cell.<sup>[9]</sup> However, there are thousands of known ternary structure types, making a brute-force approach to the problem unrealistic. Despite the resulting huge number of candidate ternary compounds, it is worth observing that the largest computational databases only contain overall about 4 million materials.<sup>[2,26,27]</sup>

Machine learning methods have made it possible to accelerate material searches considerably. These methods are some of the most useful instruments added to the toolbox of material science and solid-state physics in the last decade. They have enabled the efficient prediction of a wide range of


## 1. Introduction

One of the most tantalizing possibilities of modern computational materials science is the prediction and characterization of experimentally unknown compounds. In fact, developments in theory and algorithms in the past decades allowed for the

J. Schmidt, N. Hoffmann, H.-C. Wang, M. A. L. Marques  
Institut für Physik  
Martin-Luther-Universität Halle-Wittenberg  
D-06099, Halle, Germany

P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira  
CFisUC  
Department of Physics  
University of Coimbra  
Rua Larga 3004-516, Coimbra, Portugal

S. Botti  
Institut für Festkörperteorie und -optik  
Friedrich-Schiller-Universität Jena  
Max-Wien-Platz 1 07743, Jena, Germany  
E-mail: silvana.botti@uni-jena.de

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/adma.202210788>.

© 2023 The Authors. Advanced Materials published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution-NonCommercial License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

DOI: 10.1002/adma.202210788

material properties with near ab initio accuracy.<sup>[3,4,28]</sup> Early works in this direction achieved speedups by factors of about 5–30.<sup>[9,10,15]</sup> These works were generally based on relatively simple machine learning models, for example, decision trees or kernel ridge regression, that used hand-built features of the composition as input and had to be retrained for different crystal families. Significant progress toward more general models was made by Ward et al.<sup>[15]</sup> who included structural descriptors applicable to high-throughput searches in terms of Voronoi tessellations. This allowed Ward et al. to use training data from all compounds, resulting in improved performance for high-throughput searches. Other important steps forward were achieved by two other classes of models that were developed simultaneously: message-passing networks for crystal and molecular graphs, as well as deeper composition-based models.<sup>[29–32]</sup> We note that compositional models can be completely independent of the crystal structure. However, they are inadequate for large-scale high-throughput searches, as they cannot differentiate between polymorphs with the same chemical composition. Message passing networks, on the other hand, enabled unprecedented performance for the prediction of properties with ab initio accuracy<sup>[3,33,34]</sup> from crystal structures.

Until recently all message-passing networks for crystals used atomic positions, in some form, as input. However, this information is not available until calculations, for example using DFT structure optimization, are performed. The direct use of unrelaxed geometries as input features leads to significantly worse performance.<sup>[35,36]</sup> Recently, some of us<sup>[36]</sup> developed crystal-graph attention networks (CGATs) that circumvent the problem, replacing the precise bond distances with embedding of graph distances. In a similar spirit, Goodall et al.<sup>[37]</sup> proposed coarse-grained message-passing networks that use as input Wyckoff representations, that is, coordinate-free sets of symmetry-related positions in a crystal. In this work, we apply the former approach to explore a significantly enlarged space of crystalline compounds. Another very recent effort to improve predictions on unrelaxed structures is through data augmentation by Gibson et al.<sup>[38]</sup>

Currently, the largest issue concerning the accuracy of message-passing networks is no longer the topology of the networks, nor its complexity, but is related to the limitations of existing materials datasets. Some of us identified already in ref. [36] large biases stemming from the lack of structural and chemical diversity in the available data. These biases, ultimately of anthropogenic nature,<sup>[39,40]</sup> lead unfortunately to a poor generalization error. In fact, even if the error in test sets is of the order of 20–30 meV atom<sup>-1</sup>, the actual error during high-throughput searches can be easily one order of magnitude larger if the available training data is not representative of the actual material space.<sup>[36]</sup> A possible approach to this problem is active-learning based strategies, as presented in refs. [41–43], where a small number of calculations is used to frequently update the surrogate model.

In this work, we tackle this challenging problem using instead a larger scale stepwise approach. First, we perform a series of high-throughput searches with an extended set of chemical elements (including lanthanides and some actinide elements), applying the transfer learning approach presented in ref. [36], consisting in continuing the training of a pre-trained

general-purpose model on a specific family of compounds. Thanks to the additional data generated by these calculations, we expect to reduce the bias due to the representation of the chemical elements in the dataset. In a subsequent step, we retrain the CGAT and employ it to scan a material space of almost 1 billion compounds that comprises more than 2000 crystal-structure types. We obtain in this way a dataset of DFT calculations with a considerably larger structural diversity, that we then use to retrain a network. This CGAT is then shown to possess a massively improved generalization error and a strongly reduced chemical and structural bias. Finally, we offer a demonstration of the usefulness of our approach and inspect this dataset to search for materials with extreme values of some interesting physical properties.

## 2. Construction of Datasets and Networks

### 2.1. Enlarging the Chemical Space

Our starting point is the dataset used by some of us for training in ref. [36]. We will refer to this dataset as “DCGAT-1” and to the crystal-graph network of ref. [36] as “CGAT-1,” respectively.

As discussed previously, the training data in DCGAT-1 is biased with respect to the distribution of chemical elements and crystal symmetries. To circumvent the first problem we performed a series of high-throughput calculations for specific structure types. We used a larger chemical space than previous works, considering 84 chemical elements, including all elements up to Pu (with the exception of Po and At, for which we do not have pseudopotentials, Yb whose pseudopotential exhibits numerical problems, and rare gases). This results in 6972 possible permutations per binary, 571 704 permutations per ternary, and 46 308 024 permutations per quaternary system. For all these compositions we considered a (largely arbitrary) selection of crystal structures, including ternary garnets, Ruddlesden–Popper layered perovskites, cubic Laves phases, ternary and quaternary Heuslers, auricuprides, etc. In total, we included 11 binary, 8 ternary, and 1 quaternary compound families (a complete list and more details can be found in the Supporting Information).

For each structure type included in the selection, we performed a high-throughput study using the transfer learning approach of ref. [36]: i) The machine-learning model is used to predict the distance to the convex hull of stability for all possible chemical compositions. At the start we use the pre-trained CGAT-1 machine; ii) We perform DFT geometry optimizations to validate all compounds predicted stable, or unstable with a distance of less than 200 meV atom<sup>-1</sup>, from the convex hull; iii) We add these calculations to a dataset containing all DFT calculations for the corresponding structure type; iv) We use transfer learning to train a new model on the basis of this dataset with a training/validation/testing split of 80%/10%/10%; v) The cycle is restarted one to three times until the mean absolute error (MAE) of the model is smaller than 30 meV atom<sup>-1</sup>.

This procedure resulted in 397 438 additional DFT calculations, yielding 4382 compounds below the convex hull of DCGAT-1 (and therefore already increasing the size of the known convex hull by approximately ten percent). Moreover,

we added a large dataset of mixed perovskites<sup>[36]</sup> plus data concerning oxynitride, oxyfluoride, and nitrofluoride perovskites from ref. [44], amounting to around 381 000 DFT calculations. Finally, we recalculated and added 1343 compounds that were possibly unconverged outliers from AFLOW<sup>[26]</sup> according to the criteria in ref. [45]. The final dataset resulting from all these changes and additions contains ≈780 000 compounds more than DCGAT-1 and will be denoted as DCGAT-2.

In **Figure 1** we plot the element distribution in both datasets DCGAT-1 and DCGAT-2. As expected, the original dataset is quite biased with a drastic undersampling of most lanthanides and actinides. Despite its smaller size, the new dataset includes between three and twenty times more compounds containing undersampled elements, and it, therefore, counteracts the unbalanced distribution of chemical elements of DCGAT-1. Note that, in particular, metallic elements appear in very similar quantities in the revised dataset, with exception of the heavier actinides that are still somewhat under-represented.

We used DCGAT-2 to retrain a CGAT with the same hyperparameters used in ref. [36] (the resulting network will be denoted as CGAT-2). The CGAT-2 network has a mean absolute test error of 21 meV atom<sup>-1</sup> for the distance to the convex hull using a training/validation/test split of 80%/10%/10%. Although the test error is of the same order of magnitude as CGAT-1, we will see that the generalization error is drastically reduced. We also trained a network to predict the volume per atom of the crystals, obtaining a test error of 0.25 Å<sup>3</sup> atom<sup>-1</sup>.

## 2.2. Enlarging the Structural Space

After having successfully removed the bias in our dataset in the distribution of chemical elements, we now tackle the lack of structural variety. Our strategy consists in adding calculations of under-represented structural types, keeping in mind that we are mainly interested in phases that are thermodynamically stable, or close to stability. We start by querying our database using the pymatgen<sup>[46]</sup> structure matcher to identify all distinct structural models present in DCGAT-1. This is performed by normalizing the primitive unit cell to unit volume and comparing lattice constants and angles, as well as atomic positions up to a certain threshold. This makes our practical definition different from the usual one of prototypes as, for example, different *c/a* ratios of a wurtzite crystal can lead to distinct structural models. It is nevertheless important to keep track of all these models in order to increase the precision of the crystal-graph network predictions.<sup>[36]</sup> We found a total of ≈58 000 structural models, the large majority of them appearing only once or twice in the dataset. We then selected all models with less than 21 atoms in the unit cell, a space-group number larger than nine and that appeared at least ten times in our dataset. The first two criteria are chosen to limit the run-time of the DFT calculations. Following these criteria, we end up with 639 binary and 1829 ternary crystal-structure models, spanning a space of 1 050 101 724 possible compounds. These models also densely cover the composition space, as depicted in the generic phase diagram of **Figure 2**.

In **Figure 3** we plot the distributions of the number of atoms in the unit cell (**Figure 3a**) and of crystal systems (**Figure 3b**)

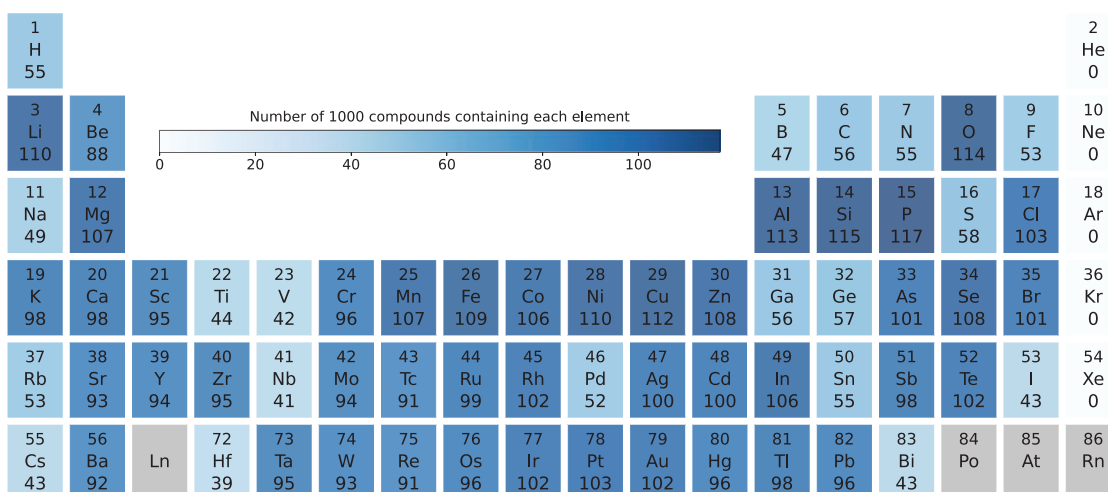
in the set of selected structural models. The distribution of the number of models displays a maximum at six atoms per unit cell, decreasing then slowly for a larger number of atoms. It is also clear that models with an even number of atoms are far more common than those with an odd number of atoms. The most represented crystal system is orthorhombic, followed by monoclinic and tetragonal, while cubic structures are rare. Note that the number of monoclinic structures is reduced by the imposed restriction on the space group number, as monoclinic structures have space groups between 3 and 15. Also due to this restriction, no triclinic structures are present in the dataset. All these conclusions apply to both binary and ternary crystals.

We use our CGAT-2 network to predict the distance to the convex hull for these structural models, after grouping them according to their general composition A<sub>x</sub>B<sub>y</sub>C<sub>z</sub>. For every composition, we occupy the lattice sites of each structural model with all permutations of the A, B, and C chemical elements, and let the machine predict the ones that are at a distance of less than 50 meV atom<sup>-1</sup> above the convex hull. In case several geometries are below this threshold we just keep the one with the lowest energy. We also remove duplicates, and materials with Eu, Gd, Yb, and Lu due to converge issues with the DFT calculations. In total, we obtain 530 937 materials satisfying our cutoff criteria.

We note the geometries have not been optimized yet. We can obtain a good estimate of the unit cell volume using a CGAT network that we have trained for this quantity. We use this information to build the starting point for DFT geometry optimizations, as described in Section 5. After removing unconverged calculations we are left with DFT calculations for 515 653 new compounds. Combining the new data with DCGAT-2 we arrive at our final dataset “DCGAT-3”. From the new data, we separate a test set composed of materials that correspond to eight randomly chosen ternary compositions, encompassing 93 crystal structure models and 57 252 entries. The remaining data from DCGAT-3 is then used to train our last network, called CGAT-3.

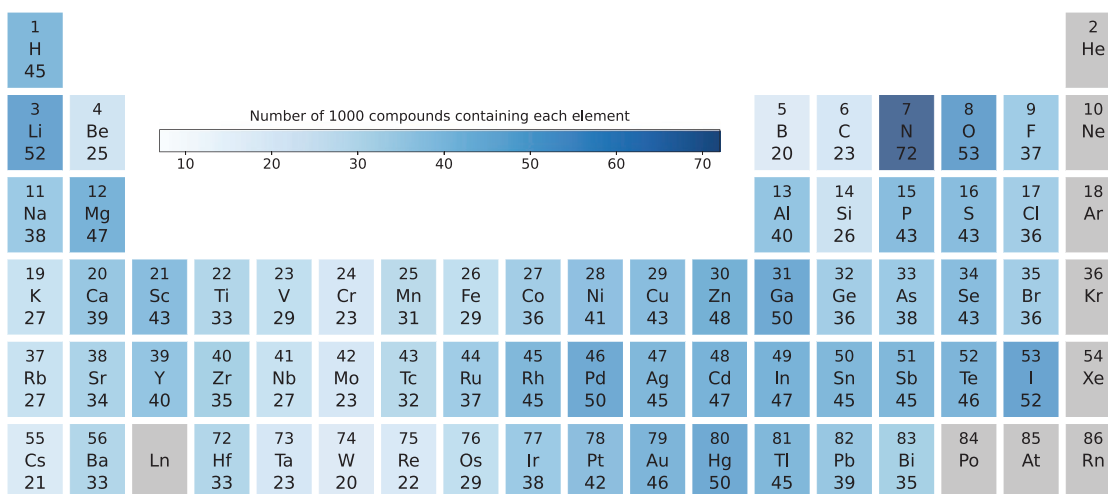
By separating a test set of compositions and structures sparsely represented in our training sets we expect to have a proper statistical estimation of the generalization error of the networks. The MAE improves from 92 meV atom<sup>-1</sup> for CGAT-1, to 87 meV atom<sup>-1</sup> and 57 meV atom<sup>-1</sup>, for CGAT-2 and CGAT-3, respectively. As we can see in **Figure 1** the element coverage in the training set already significantly improved from CGAT-1 to CGAT-2 while the structural diversity only improved marginally. As a result, the decrease in MAE is rather small at 5%. On the other hand, the increase in structural diversity from CGAT-2 to CGAT-3 results in a major 33% improvement. Consequently, we can conjecture that the majority of future improvements with respect to data will come from the sampling of additional geometrical arrangements.

In **Figure 4** we see the element-resolved MAEs for CGAT-1 and CGAT-3. For CGAT-1 we observe a strong dependence of the MAE on the chemical element, with a significantly higher MAE for the first-row elements, most likely due to the first-row anomaly that has been observed in multiple studies.<sup>[9,10,36]</sup> This effect is strongly reduced for CGAT-3, with an MAE that is much more uniform across the periodic table. Indeed, the maximum MAE for CGAT-1 is 258 meV atom<sup>-1</sup> for boron, while this value is reduced to 181 meV atom<sup>-1</sup> for CGAT-3, proving that we could essentially eliminate the chemical element bias from our dataset.



|                |               |               |               |               |               |               |               |               |               |               |               |               |               |               |
|----------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| 57<br>La<br>94 | 58<br>Ce<br>9 | 59<br>Pr<br>9 | 60<br>Nd<br>9 | 61<br>Pm<br>1 | 62<br>Sm<br>9 | 63<br>Eu<br>6 | 64<br>Gd<br>5 | 65<br>Tb<br>9 | 66<br>Dy<br>9 | 67<br>Ho<br>8 | 68<br>Er<br>8 | 69<br>Tm<br>8 | 70<br>Yb<br>9 | 71<br>Lu<br>6 |
| 89<br>Ac<br>1  | 90<br>Th<br>7 | 91<br>Pa<br>2 | 92<br>U<br>6  | 93<br>Np<br>5 | 94<br>Pu<br>6 | 95<br>Am      | 96<br>Cm      | 97<br>Bk      | 98<br>Cf      | 99<br>Es      | 100<br>Fm     | 101<br>Md     | 102<br>No     | 103<br>Lr     |

(a)



|                |                |                |                |                |                |                |                |                |                |                |                |                |           |                |
|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----------|----------------|
| 57<br>La<br>42 | 58<br>Ce<br>25 | 59<br>Pr<br>23 | 60<br>Nd<br>24 | 61<br>Pm<br>32 | 62<br>Sm<br>24 | 63<br>Eu<br>20 | 64<br>Gd<br>25 | 65<br>Tb<br>24 | 66<br>Dy<br>24 | 67<br>Ho<br>25 | 68<br>Er<br>25 | 69<br>Tm<br>25 | 70<br>Yb  | 71<br>Lu<br>24 |
| 89<br>Ac<br>29 | 90<br>Th<br>21 | 91<br>Pa<br>20 | 92<br>U<br>7   | 93<br>Np<br>7  | 94<br>Pu<br>24 | 95<br>Am       | 96<br>Cm       | 97<br>Bk       | 98<br>Cf       | 99<br>Es       | 100<br>Fm      | 101<br>Md      | 102<br>No | 103<br>Lr      |

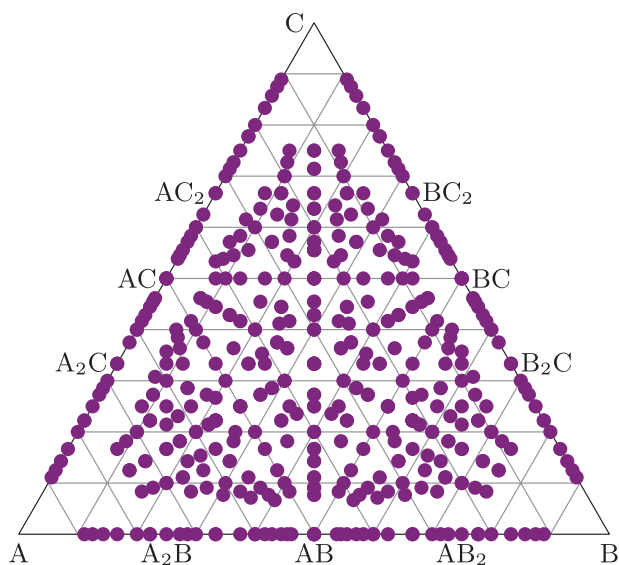
(b)

**Figure 1.** Number of materials in a) DCGAT-1 and b) added to DCGAT-1 to obtain DCGAT-2 containing a specific chemical element of the periodic table.

In **Figure 5** we plot the distance to the convex hull for the DCGAT-1 dataset and the data that was added in DCGAT-2 and DCGAT-3. We see that the 1.89M materials of the original

dataset still exhibit a wide distribution with a median of 420 meV atom<sup>-1</sup> and a standard deviation of 570 meV atom<sup>-1</sup>. This is easy to understand as the data mostly originates from





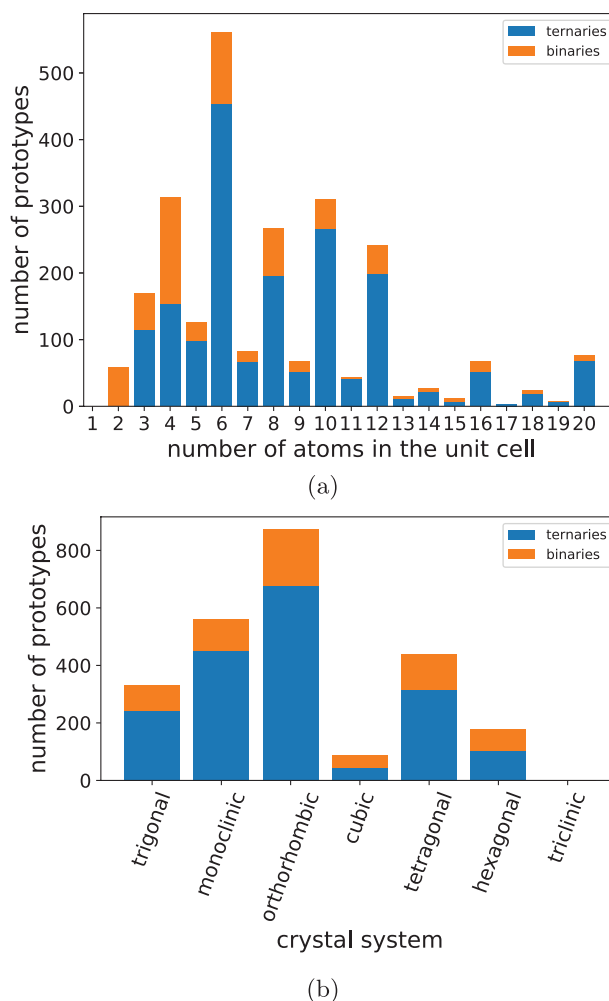
**Figure 2.** Ternary phase diagram showing the stoichiometries covered in this work.

traditional high-throughput searches. The peak close to zero is due to the experimentally known stable materials as well as to the data from some studies using machine learning or chemical substitution strategies.<sup>[10,47]</sup> The perovskite data added to DCGAT-2 has one peak at roughly 200 meV atom<sup>-1</sup>, due to the compounds generated using machine learning with a cutoff of 200 meV atom<sup>-1</sup> from the convex hull using the neural network of ref. [36]. The wide distribution comes from the random perovskites generated in the same study.<sup>[36]</sup> The remainder of DCGAT-2 was also generated using a similar approach and therefore, is centered at 200 meV atom<sup>-1</sup>. The total DCGAT-2 dataset contained 2.67M materials. Finally, we can see that the distribution of the 515k materials we added to arrive at the 3.18M DCGAT-3 entries has a median of 117 meV atom<sup>-1</sup> with a standard deviation of 154 meV atom<sup>-1</sup>. Compared to the usual range of distances to the convex hull of a high-throughput search, this distribution is extremely narrow, showing the remarkable accuracy and generalization error of our machine-learning models.

The flowchart shown in Figure S1, Supporting Information, summarizes how the final model has been generated. We can observe that the dataset increases from 2.1M to 2.8M and finally to 3.1M calculations going from DCGAT-1 to DCGAT-2 and then DCGAT-3. We chose to increase substantially the dataset, performing a large amount of DFT calculations between two successive training steps, because training our complex neural network is particularly expensive and we wanted to be sure that the added data would have the potential to decrease the prediction error.

### 3. Material Properties

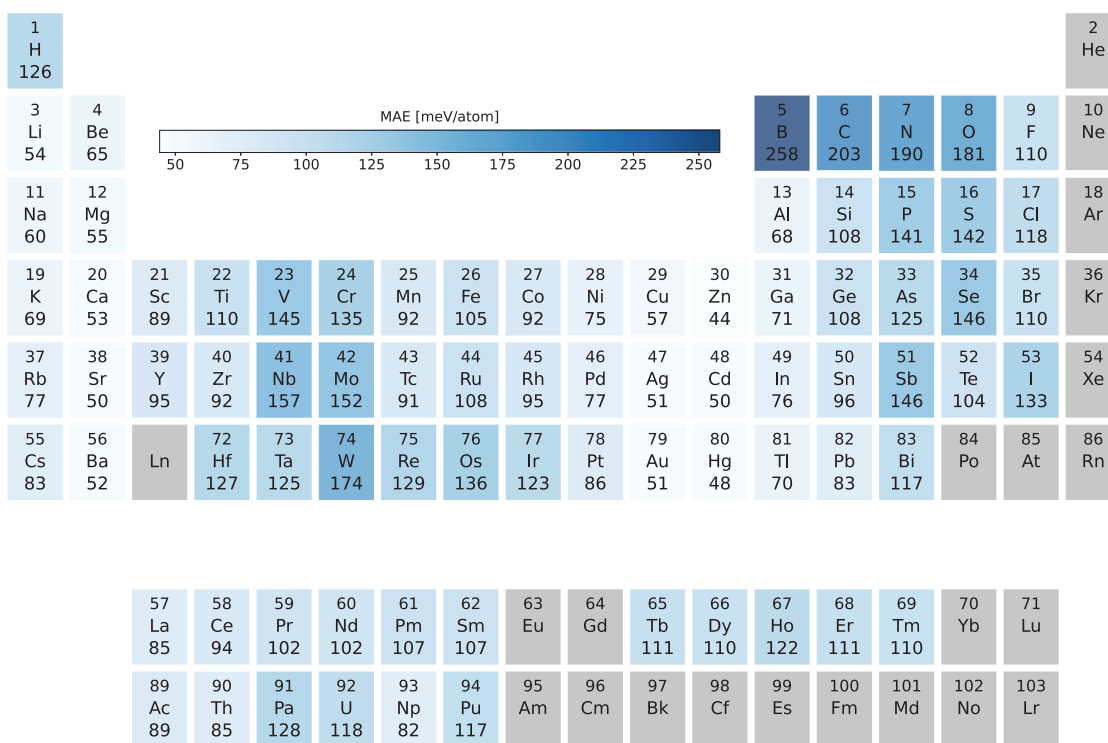
In the process of constructing our unbiased datasets, we have discovered 19 512 compounds on the convex hull, 168 340 unstable compounds with a distance of less than



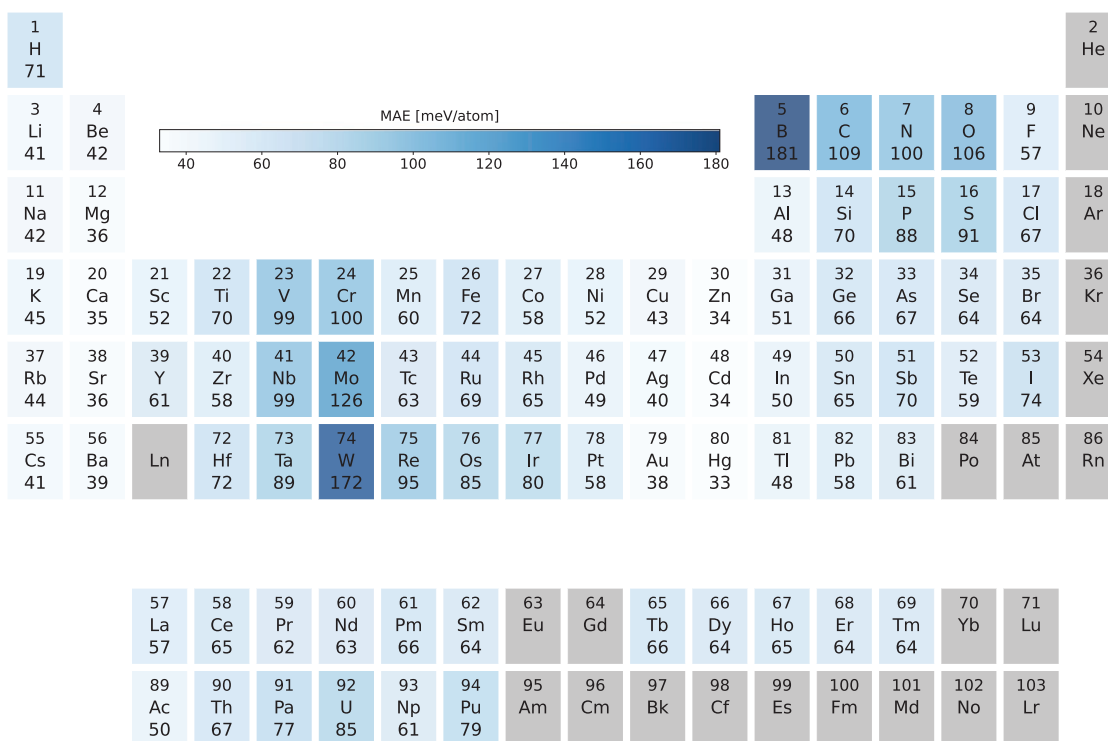
**Figure 3.** The histograms show the distribution of a) the number of atoms per unit cell and b) crystal systems in the set of structural models scanned in the high-throughput search. The counts for the binary and ternary models are stacked on top of each other in orange and blue, respectively.

50 meV atom<sup>-1</sup> from the hull, and 326 433 compounds above the hull, at a distance of less than 100 meV atom<sup>-1</sup>. These crystalline materials are, to our knowledge, not yet included in available databases. An overview of the chemical nature of the new compounds on the convex hull is summarized in Figure 6, where we plot the number of newly discovered stable compounds containing each chemical element. We see that these materials cover the entire periodic table, but with a maximum for compounds including Li, Mg, transition metals around Pd and Ga, and lanthanides and actinides. Concerning the latter, we see that compounds including Eu, Gd, U, and Np are relatively under represented. This is not due to a lesser ability of these chemical elements to form stable compounds, but to a technical reason: the available pseudopotentials for these elements often lead to numerical problems making calculations hard to converge.

In the following, we want to analyze these materials in more detail. To this end, we perform machine-learning-assisted data-mining of several non-trivial physical properties to reveal



(a)

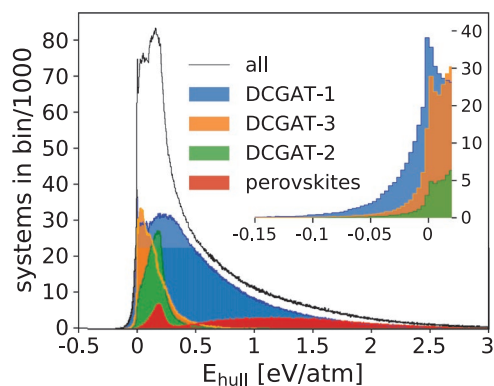


(b)

**Figure 4.** MAE in the test set separated from DCGAT-3 for compounds containing each element of the periodic table, when the predictions are obtained with a) CGAT-1 and b) CGAT-3.

compounds with extreme behavior. We decided to restrict our search to (quasi-)stable materials, defining a threshold of

50 meV atom<sup>-1</sup> above the convex hull of stability. For these systems, we evaluate elastic constants, superconductivity, and



**Figure 5.** Distance to the convex hull for DCGAT-1 and for the added data contained in DCGAT-2 and DCGAT-3. The mixed perovskites studied in ref. [36] are separated from DCGAT-2 and the rest of the data. In the inset plot, we zoom into the range of stable compounds.

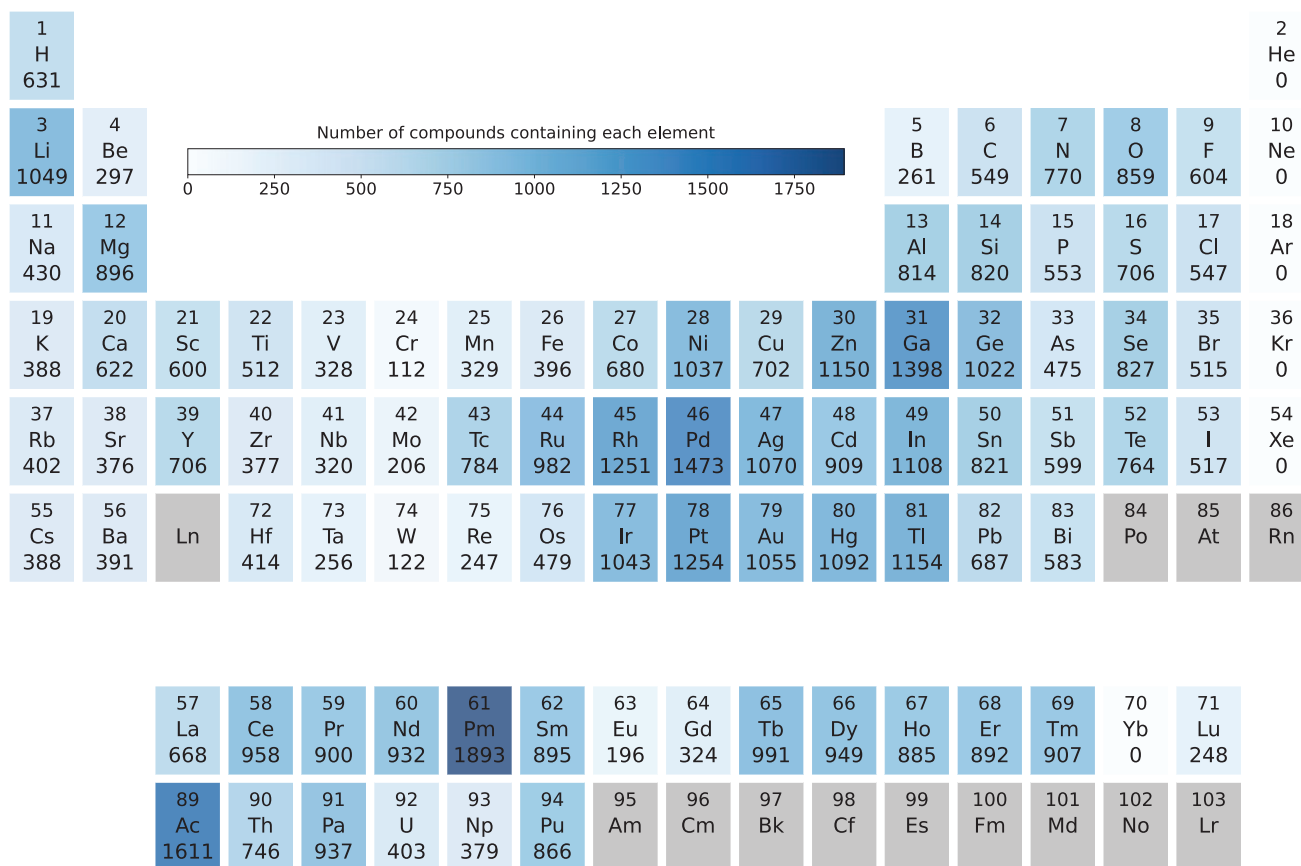
gap deformation potentials. The strategy in the three cases is similar: we train machine learning models based on crystal graph convolutional networks (CGCNN)<sup>[48]</sup> to provide an efficient prediction of the specific property. Promising materials are then investigated in more detail using DFT or density-functional perturbation theory to validate machine learning predictions and to provide further insights into the physics and the mechanism behind the extreme values of a certain property.

### 3.1. Ultra-Hard and Incompressible Materials

Describing the elastic response of a material requires knowledge of its stiffness tensor, composed of (at most) 21 independent elastic constants. Direct analysis of the whole tensor is rather cumbersome, so most studies concentrate on two derived properties: the Voigt–Reuss–Hill averaged<sup>[49]</sup> bulk and shear moduli  $G_{VRH}$  and  $K_{VRH}$ , respectively, that describe the compressibility of the material. Besides considering the average elastic response of the material, we can use  $G_{VRH}$  and  $K_{VRH}$  to estimate the Vicker’s hardness<sup>[50,51]</sup>  $H_V$  and use this quantity to identify ultra-hard materials.

A recent example of this approach is ref. [52], where the authors used Bayesian optimization with symmetry relaxation to obtain optimized structures, followed by materials graph network<sup>[53]</sup> models to predict formation energies as well as  $G_{VRH}$  and  $K_{VRH}$ . This methodology was applied to search for ultra-hard transition metal borides and carbides, exploring a comparatively small space of circa 400 000 compounds.

In this work we perform a screening of the values of  $G_{VRH}$ ,  $K_{VRH}$ , and  $H_V$  on our much larger dataset. We predict the values of the first two quantities using CGCNN models trained on the dataset of Matbench,<sup>[33]</sup> and use these predicted values to estimate the Vicker’s hardness for each material using the model of ref. [51]. Details of the training can be found in the Supporting Information. The compounds within 50 meV atom<sup>-1</sup>



**Figure 6.** Number of stable compounds containing each chemical element discovered in this work.

**Table 1.** Chemical formula, distance to the convex hull ( $E_{\text{hull}}$ ), space group number (Spg.), bulk modulus ( $K_{\text{VRH}}$  in GPa), shear modulus ( $G_{\text{VRH}}$  in GPa), and Vicker's hardness ( $H_V$  in GPa) for the materials with the highest calculated  $H_V$  (top section),  $K_{\text{VRH}}$  (middle section), and  $G_{\text{VRH}}$  (bottom section). All indicated materials satisfy the Born–Huang elastic stability criteria.<sup>[54,55]</sup>

| Formula                          | $E_{\text{hull}}$ | Spg. | $N$ | $K_{\text{VRH}}$ | $G_{\text{VRH}}$ | $H_V$ |
|----------------------------------|-------------------|------|-----|------------------|------------------|-------|
| TiVB <sub>3</sub>                | 16                | 63   | 10  | 262              | 245              | 42    |
| TaTiB <sub>3</sub>               | 14                | 63   | 10  | 268              | 230              | 36    |
| Ta <sub>2</sub> BeB <sub>3</sub> | 0                 | 69   | 12  | 272              | 232              | 36    |
| BeNb <sub>2</sub> B <sub>3</sub> | 0                 | 69   | 12  | 255              | 222              | 35    |
| TiB <sub>3</sub> W               | 0                 | 63   | 10  | 293              | 235              | 33    |
| Ir <sub>3</sub> Os <sub>5</sub>  | 28                | 25   | 8   | 378              | 196              | 25    |
| Os <sub>2</sub> Ru               | 14                | 15   | 6   | 369              | 233              | 27    |
| Os <sub>5</sub> Ru <sub>3</sub>  | 18                | 25   | 8   | 365              | 236              | 27    |
| MoOs <sub>4</sub> Ru             | 21                | 13   | 12  | 357              | 220              | 25    |
| Os <sub>2</sub> RuW              | 49                | 51   | 8   | 350              | 189              | 23    |
| TiVB <sub>3</sub>                | 16                | 63   | 10  | 262              | 245              | 42    |
| Os <sub>5</sub> Ru <sub>3</sub>  | 18                | 25   | 8   | 365              | 236              | 27    |
| TiB <sub>3</sub> W               | 0                 | 63   | 10  | 293              | 235              | 33    |
| Os <sub>2</sub> Ru               | 14                | 15   | 6   | 369              | 233              | 27    |
| Ta <sub>2</sub> BeB <sub>3</sub> | 0                 | 69   | 12  | 272              | 232              | 36    |

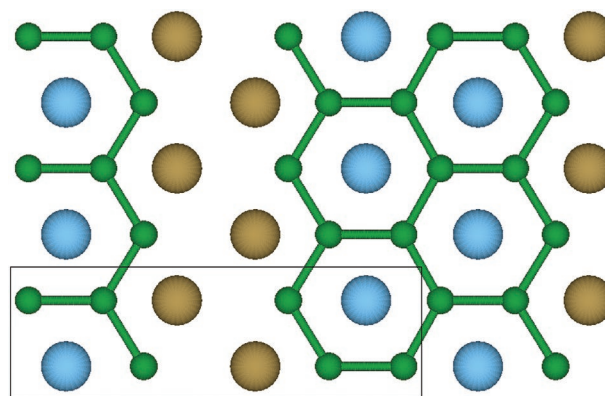
from the convex hull and with the 25 highest  $G_{\text{VRH}}$  and  $K_{\text{VRH}}$  were selected for subsequent analysis, that is, their stiffness tensors were calculated using DFT, as described in Section 5. The top-five materials for each quantity are presented in **Table 1** (while a complete list is given in the Supporting Information).

Not surprisingly, the materials with the highest  $H_V$  are mostly metal borides. These materials are known to emulate the hardness of diamond thanks to a mixture of high valence (provided by the metal) and short bonds (provided by boron).<sup>[56–58]</sup> Most of the borides seen here belong to the same prototype, with the anonymous formula  $\text{MNB}_3$ , with M and N a metal, and space group 63. This crystal structure, depicted in **Figure 7**, consists of hexagonal boron nanoribbons intercalated with layers of transition metals. Overall, this arrangement of atoms is reminiscent of other ultra-hard materials such as WB, WB<sub>2</sub>, and TiB<sub>2</sub>. We remark on the prediction of the superhard ternary compound TiVB<sub>3</sub> with a hardness of 42 GPa.

For what concerns incompressible materials, the presence of osmium compounds (Ir<sub>3</sub>Os<sub>5</sub> and Os<sub>2</sub>Ru, etc.) is also not too surprising, as osmium and osmium compounds are known to have extremely large bulk modulus, although they are not necessarily hard, due to the metallic nature of their bonds.<sup>[60]</sup>

### 3.2. Superconductors

Searching for new conventional superconductors with a high critical temperature ( $T_c$ ) is always a tempting application for large material datasets. This turns out to be a complex task due to the interplay between the different ingredients that determine  $T_c$ , as well as the lack of reliable simple indicators of superconductivity.<sup>[61]</sup> McMillan's formula<sup>[62]</sup> suggests to use



**Figure 7.** Crystal structure of TiTaB<sub>3</sub>. The blue atoms represent Ti, the gold Ta, and the green B. We also depict the primitive unit cell. Picture produced with VESTA.<sup>[59]</sup>

Debye's temperature ( $\Theta_D$ ) and the density of states at the Fermi level ( $\text{DOS}(E_F)$ ) as estimators for high- $T_c$ , a connection that has recently been used with some success.<sup>[63]</sup> Within the context of the present work, we can easily estimate  $\Theta_D$  from the bulk and shear moduli<sup>[64]</sup> obtained in the previous section.

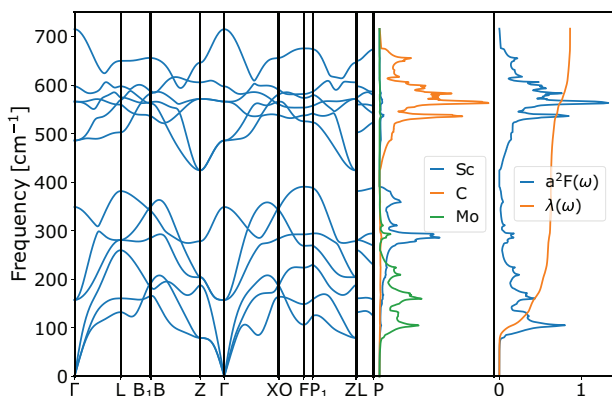
Following this line of thought, we selected non-magnetic materials with a predicted Debye's temperature above 300 K and with a  $\text{DOS}(E_F)$  larger than 0.5 states per eV. Furthermore, we restricted our search to space-group numbers greater or equal to 160 (highly symmetric compounds), and to cells with a maximum of eight atoms. We ordered the resulting 2717 materials by  $\Theta_D$ , and performed electron–phonon calculations for the first 50. The large majority of these are dynamically stable (two were found to have imaginary frequencies), and we found 19 systems with  $T_c$  above 1 K, as calculated using McMillan's formula. For the compounds with the largest calculated  $T_c$ , we performed better converged electron–phonon calculations by increasing the density of the  $q$ - and  $k$ -grids. The calculated superconducting properties for these compounds can be seen in **Table 2**, while a complete table for the 50 screened materials is available in the Supporting Information.

The compound with the highest transition temperature in **Table 2** is ScMoC<sub>2</sub>, with a  $T_c = 15.97$  K. This is a very interesting material with a rhombohedral lattice exhibiting alternating layers of Sc–C–Mo–C. In **Figure 8** we see that the lowest frequency phonon modes have Mo-character, while the optical phonon modes until around 400  $\text{cm}^{-1}$  have mostly Sc-character. These are separated by a gap from a manifold of optical modes

**Table 2.** Formula, distance to the convex hull ( $E_{\text{hull}}$ ), space group (Spg.), and calculated superconducting properties for the screened materials.

| Formula                          | $E_{\text{hull}}$ | Spg. | $\lambda$ | $\omega_{\text{log}}$ [K] | $T_c$ [K] |
|----------------------------------|-------------------|------|-----------|---------------------------|-----------|
| ScMoC <sub>2</sub>               | 49                | 166  | 0.86      | 287.94                    | 15.66     |
| NbRhBe <sub>4</sub>              | 47                | 216  | 0.70      | 311.69                    | 10.91     |
| YZr <sub>3</sub> N <sub>4</sub>  | 35                | 221  | 0.54      | 398.16                    | 6.56      |
| Sc <sub>4</sub> NO <sub>3</sub>  | 8                 | 221  | 0.50      | 428.04                    | 5.12      |
| Zr <sub>4</sub> CN <sub>3</sub>  | 0                 | 221  | 0.50      | 419.02                    | 5.05      |
| ScZr <sub>3</sub> N <sub>4</sub> | 0                 | 221  | 0.48      | 440.18                    | 4.50      |





**Figure 8.** Phonon dispersion, the density of states, and the electron–phonon coupling  $\alpha^2F(\omega)$  of  $\text{ScMoC}_2$ .

exclusively due to the C-atoms. A large part of the electron–phonon coupling constant  $\lambda$  comes from a softening of an acoustic branch in the  $\Gamma \rightarrow X$  direction, and that is ultimately responsible for the large value of  $T_c$ .

The only intermetallic compound in the top five list is  $\text{NbRhBe}_4$ . This is actually a ternary generalization of the cubic Laves (C15) phase.<sup>[65]</sup> Interestingly, both  $\text{NbBe}_2$  and  $\text{RhBe}_2$  have been synthesized,<sup>[66]</sup> and are superconducting with  $T_c$  of 2.14 and 1.37 K, respectively. Our prediction of 11.61 K is considerably higher than for each of the individual binaries but in line with the related A15 compound  $\text{Nb}_3\text{Be}$  that has  $T_c = 10$  K.<sup>[67]</sup>

Interestingly, the list in Table 2 also includes several nitrides. As an example, we take a closer look at  $\text{ScZr}_3\text{N}_4$  (that is isostructural to  $\text{YZr}_3\text{N}_4$  also on the list). This material has a simple cubic structure (space group  $\text{Pm}\bar{3}\text{m}$  #221) that can be derived from the NaCl-type structure, with N occupying one site (Wyckoff positions 1b and 3d), and the cations occupying the other site (Sc in the 1a and Zr in the 3c Wyckoff positions). The phonon dispersion and the density of states, and the electron–phonon coupling  $\alpha^2F(\omega)$  of  $\text{ScZr}_3\text{N}_4$  can be found in the Supporting Information. Consistently with the difference in atomic masses of the composing atoms, the acoustic and lower optical branches have Zr character, the following manifold just below  $300\text{ cm}^{-1}$  has mostly Sc character and the highest manifold at around  $400\text{--}500\text{ cm}^{-1}$  is related to N. Finally, all modes contribute to the coupling constant  $\lambda$  that reaches a value of around 0.8.

### 3.3. Deformation Potentials

Finally, we take a look at the hydrostatic deformation potentials  $\Xi$ , which measure the variation of the band gap ( $E_g$ ) with respect to hydrostatic variations of the structure. This quantity is defined as

$$\begin{aligned} \Xi &= \frac{dE_g}{d\ln(V)} \\ &= \frac{\partial E_g}{\partial \ln(V)} + \sum_i \frac{\partial E_g}{\partial u_i} \frac{\partial u_i}{\partial \ln(V)} \end{aligned} \quad (1)$$

**Table 3.** Materials with the largest (in absolute value) hydrostatic deformation potentials. Shown here are the chemical formula, distance to the hull ( $E_{\text{hull}}$  in  $\text{meV atom}^{-1}$ ), space group number (Spg.), band gap ( $E_g$  in eV), deformation potential ( $\Xi$  in eV), and predicted deformation potential ( $\tilde{\Xi}_{\text{pred}}$ , in eV).

| Material                          | $E_{\text{hull}}$ | Spg. | $E_g$ | $\Xi$ | $\tilde{\Xi}_{\text{pred}}$ |
|-----------------------------------|-------------------|------|-------|-------|-----------------------------|
| $\text{InGaO}_3$                  | 49                | 148  | 1.48  | −7.76 | −8.02                       |
| $\text{NaLi}_4\text{F}_5$         | 42                | 139  | 7.64  | −7.30 | −7.40                       |
| $\text{AcAlF}_6$                  | 0                 | 166  | 7.31  | −7.22 | −7.30                       |
| $\text{NaLi}_3\text{F}_4$         | 43                | 65   | 7.64  | −6.88 | −7.03                       |
| $\text{Na}_2\text{SiN}_2$         | 14                | 72   | 2.13  | −6.30 | −8.53                       |
| $\text{PaO}_6$                    | 0                 | 148  | 2.01  | −5.95 | −8.31                       |
| $\text{Li}_3\text{ClF}_2$         | 27                | 71   | 6.35  | −5.82 | −6.92                       |
| $\text{KNdF}_4$                   | 9                 | 123  | 6.56  | −5.75 | −6.28                       |
| $\text{AcGaF}_6$                  | 0                 | 166  | 5.83  | −5.68 | −7.77                       |
| $\text{LiTlSe}$                   | 30                | 11   | 1.01  | −5.38 | 3.64                        |
| $\text{Pb}_2\text{SeS}_4$         | 7                 | 139  | 0.42  | 3.20  | 3.35                        |
| $\text{TlIn}_4\text{Cl}_5$        | 23                | 87   | 1.69  | 4.04  | 4.14                        |
| $\text{TlIn}_4\text{Br}_5$        | 11                | 87   | 1.49  | 4.08  | 3.61                        |
| $\text{In}_4\text{GaBr}_5$        | 31                | 79   | 1.43  | 4.08  | 3.21                        |
| $\text{LiTi}_4\text{I}_5$         | 30                | 166  | 1.94  | 4.12  | 3.18                        |
| $\text{In}_5\text{Br}_4\text{Cl}$ | 18                | 166  | 1.39  | 4.23  | 3.45                        |
| $\text{In}_3\text{Br}_2\text{Cl}$ | 22                | 44   | 1.43  | 4.61  | 3.63                        |
| $\text{InHg}_2\text{F}$           | 36                | 11   | 0.82  | 5.26  | 4.58                        |
| $\text{LiGeF}_3$                  | 0                 | 148  | 4.14  | 5.80  | 3.24                        |
| $\text{TlHg}_2\text{F}$           | 17                | 59   | 1.23  | 6.81  | 3.82                        |

where the second term of the last equation comes from the dependency on the internal parameters (i.e., atomic positions and cell vectors). For large scale studies, the use of the complete Equation (1) is rather cumbersome, and it is preferable to resort to the fixed shape hydrostatic deformation potential,  $\tilde{\Xi} \equiv \frac{\partial E_g}{\partial \ln(V)}$  to detect large deformation potentials. Using the

dataset published by some of us in ref. [68], which provides  $\tilde{\Xi}$  (calculated by scaling the unit cell volumes, that is, without optimizing the internal coordinates) for a series of semiconductors, we trained a CGCNN model to predict this quantity (which we define by  $\tilde{\Xi}_{\text{pred}}$ ) for the materials in the present dataset. We considered compounds within  $50\text{ meV atom}^{-1}$  from the hull, with a maximum of ten atoms in the unit cell and electronic band gaps larger than  $0.1\text{ eV}$ . From the predicted values of  $\tilde{\Xi}$ , we identified a set of 338 extreme materials for which we calculated the full deformation potential  $\Xi$ . A summary of the system with the largest absolute values of the deformation potentials can be found in Table 3.

Materials with extreme negative gap deformation potentials are very diverse, both in terms of their chemistry and the size of their band gap. In fact, we find oxides, fluorides, nitrides, etc., with band gaps ranging from  $2.0\text{ eV}$  to more than  $7.5\text{ eV}$ . We also note the appearance of alloyed systems, with two closely related cations, such as  $\text{InGaO}_3$  or  $\text{NaLi}_4\text{F}_5$ . The latter material is a very interesting example:  $\text{NaLi}_4\text{F}_5$  and  $\text{NaLi}_3\text{F}_4$  are

ordered alloys of NaF and LiF. Their end components are wide gap materials (6.1 eV for NaF and 8.7 eV for LiF) with suitable refractive indices to be used in ultraviolet optics and Cherenkov radiators,<sup>[69]</sup> but are perhaps best known for their use in molten salt reactors.<sup>[70]</sup> Owing to the isoelectronic substitution of Na by Li, the electronic structure of these ordered alloys is qualitatively similar to that of NaF and LiF: F's p-orbitals contribute heavily to the valence band while the conduction band shows primarily s-orbitals with contributions from all elements. As expected, their band gap also lies in between that of NaF and LiF, at 7.6 eV for both entries. The values of the deformation potentials are also very similar between the two, −7.30 eV for NaLi<sub>4</sub>F<sub>5</sub> and −6.88 eV NaLi<sub>3</sub>F<sub>4</sub>.

Similar reasoning follows for InGaO<sub>3</sub>, which crystallizes in the ilmenite structure (space group  $R\bar{3}$ ), a derivative of the corundum structure family. This material is essentially an ordered alloy of Ga<sub>2</sub>O<sub>3</sub> and In<sub>2</sub>O<sub>3</sub>, and along with Al<sub>2</sub>O<sub>3</sub>, it forms an isoelectronic set of corundum phases. These materials have already been observed to possess high deformation potentials, with aluminum corundum being commercially used to measure the pressure inside diamond anvil cells. As Ga and In are neighbors in the periodic table, alloying Ga and In in this structure leads to a band structure that is qualitatively very similar to that of the end components. The valence band is dominated by contributions of oxygen's p-orbitals, making it very “flat,” while the almost parabolic conduction band is more complex, showing an admixture of s-orbitals from In, O, and Ga, as well as p-orbitals from O. If we consider the isoelectronic sequence {Al, Ga, In}<sub>2</sub>O<sub>3</sub>, we observe a decrease of the band gap with increasing atomic number (5.85, 2.40 and 0.96 eV,<sup>[2]</sup> respectively) and a decrease of the absolute deformation potential (−12, −10 and −8 eV,<sup>[68]</sup> respectively). The band gap and deformation potential of InGaO<sub>3</sub> lies in the middle of the range of values.

These encouraging results point to the possibility of engineering the band gap deformation potentials of (Li,Na)F or (In,Ga)O<sub>3</sub> alloys by controlling the ratio Li/Na or In/Ga in the aforementioned phases. Mixing these compounds is not energetically favorable and therefore we expect the formation of random alloys at adequate temperatures. A detailed study of the thermodynamics of these alloys would be necessary to make quantitative predictions.

Finally, we note the presence of a false positive, LiTlSe, on the list in Table 3. In this compound the relaxation of the internal coordinates leads to an enormous correction of the fixed shape deformation potential, even leading to a change of sign. In general, our approach leads to systematically larger errors in materials where the modification of the band gap with pressure is strongly dependent on the variation of the internal atomic coordinates.

LiTlSe belongs to the matlockite family (e.g., refs. [71, 72]) and this crystal is almost layered. The structure is comprised of sheets of LiSe, where the Li atoms are the center of flat tetrahedra with Se at the vertices. These are arranged in a square lattice, such that both vertices and edges are shared between adjacent tetrahedra. The Tl atoms are placed in the concavities of the tetrahedra, thus separating the LiSe layers. The valence band shows a predominance of Se-p-orbitals followed by Tl-s-orbitals, while the conduction is primarily owed to Tl- and Se-p-orbitals.

Also, the compounds with the highest positive gap deformation potentials display a large variety of chemistry and band gaps. The latter range from 0.4 eV to more than 4 eV. The material with the largest values of  $\Xi$  is TiH<sub>2</sub>F.

## 4. Conclusions

We propose a universal crystal-graph attention neural network that predicts the phase diagram at zero temperature of the whole materials space with unprecedented accuracy, from the sole knowledge of chemical composition and prototype crystal structures. To obtain this result, we removed biases originating from under-represented chemical elements and structural arrangements in the training dataset of materials calculations.

Applying our neural network we scrutinized nearly a billion materials and were able to expand the known theoretical convex hull by roughly 30%, revealing tens of thousands of realistic targets for experimental synthesis. To exemplify how to take advantage of the uncovered opportunities for materials discovery, we further predicted a selection of material properties using a combination of machine learning and standard approaches. In this way, we discovered a number of ultra-hard and superconducting materials, as well as materials with extreme gap-deformation potentials. We suggest with the highest priority as interesting synthesis targets, for example, ultra-hard TiVB<sub>3</sub> or superconducting ScMoC<sub>2</sub> with a predicted critical temperature of 16 K.

Our results point to the importance of the quality of the training data and demonstrate that creating additional and diverse data is the key to improve large-scale machine learning models in material science so that they perform with a consistently small error across the structure and composition space. As an extension of this work, we are currently looking over quaternary systems and these new calculations will soon further enlarge and diversify our materials dataset. As a perspective, with our data-driven approach, we aspire in the near future to reduce the false negative rate to such an extent that machine learning predictions will largely replace DFT-based high-throughput searches.

## 5. Experimental Section

**Geometry Relaxations:** All geometry optimizations and total energy calculations were performed with the code VASP.<sup>[73,74]</sup> All parameters for the calculations were chosen to be compatible with the materials project database.<sup>[2]</sup> The Brillouin zones were sampled by uniform  $\Gamma$ -centered  $k$ -point grids with a density of 1000  $k$ -points per reciprocal atom. The projector augmented wave parameters<sup>[75,76]</sup> of VASP version 5.2 with a cutoff of 520 eV were applied. The calculations were converged to forces smaller than 0.005 eV Å<sup>−1</sup>. As exchange-correlation functional the Perdew–Burke–Ernzerhof<sup>[77]</sup> functional with on-site corrections for oxides, fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V, and W was used. The repulsive on-site corrections to the d-states were respectively 3.32, 3.7, 5.3, 3.9, 4.38, 6.2, 3.25, and 6.2 eV. The authors encountered convergence issues with heavy elements, like Pu for which the calculations often did not converge within their time limits, and several Lanthanides, for example, Gd and Eu for which the self-consistent cycles sometimes did not converge. Furthermore, Cs has a problematic pseudopotential which leads to additional unconverged calculations. Unconverged calculations were eliminated from the datasets.

**Elastic Constants:** Calculation of the stiffness tensors was performed using DFT with VASP<sup>[73,74]</sup> via atomate<sup>[78]</sup> workflows, using the corresponding default input parameters. In a nutshell,<sup>[79]</sup> the calculation was done by straining the cell with six deformation gradients, in four different magnitudes, for a total of 24 distorted cells. From the results of these calculations, the components of the stiffness tensor were fitted and suitably symmetrized. Once the stiffness tensor is known, all derived quantities ( $K_{\text{VRH}}$ ,  $G_{\text{VRH}}$ ,  $H_V$ ) could be trivially obtained.<sup>[49,51]</sup>

**Electron-Phonon Coupling:** Electron-phonon calculations were performed using version 7.0 of QUANTUM ESPRESSO<sup>[80]</sup> with the Perdew-Burke-Ernzerhof functional for solids (PBEsol)<sup>[81]</sup> generalized gradient approximation. Pseudopotentials from the PSEUDODOJO project,<sup>[82]</sup> specifically the PBEsol stringent norm-conserving set were used. This pseudopotential table had been systematically constructed and validated in a series of seven tests in crystalline environments, specifically the  $\Delta$ -Gauge,<sup>[83]</sup>  $\Delta'$ -Gauge,<sup>[84]</sup> GBRV-FCC, GBRV-BCC, GBRV-compound,<sup>[85]</sup> ghost-state detection, and phonons at the  $\Gamma$ -point.

The workflow consisted of the following steps: i) The energy cutoff was set to the maximum of PSEUDODOJO's high precision hint of the elements in a given material. ii) The lattice constant was optimized using uniform  $\Gamma$ -centered  $k$ -point grids with a density of 1500  $k$ -points per reciprocal atom. If this resulted in an odd number of  $k$ -points in a given direction, the next even number was used instead. Convergence thresholds for energies, forces, and stresses were set to  $1 \times 10^{-8}$  a.u.,  $1 \times 10^{-6}$  a.u., and  $5 \times 10^{-2}$  kbar, respectively. For the electron-phonon coupling a double grid technique, with the same  $k$ -grid used in the lattice optimization as the coarse grid, and a  $k$ -grid quadrupled in each direction as the fine grid was used. iv) For the  $q$ -sampling of the phonons half of the  $k$ -point grid described above was used. v) The double  $\delta$ -integration to obtain the Eliashberg function was performed with a Methfessel-Paxton smearing of 0.03 Ry. vi) The values of  $\lambda$  and  $\omega_{\text{log}}$  were then used to calculate the superconducting transition temperature using the Allen-Dynes modification<sup>[86]</sup> to the McMillan formula<sup>[62]</sup>

$$T_c = \frac{\omega_{\text{log}}}{1.20} \exp \left[ -1.04 \frac{1 + \lambda}{\lambda - \mu^* (1 + 0.62\lambda)} \right] \quad (2)$$

The value of  $\mu^*$  was arbitrarily taken as,  $\mu^* = 0.10$  for all materials studied. For the higher accuracy calculations, the previous steps were repeated by changing: i) the initial  $k$ -point grid density used for the geometry optimization was set to 3000  $k$ -points per reciprocal atom; ii) the  $k$ -grid used as the coarse grid was set to the double of the  $k$ -grid used for the geometry optimization.

**Deformation Potentials:** The calculation of the deformation potentials was done within DFT with VASP<sup>[73,74]</sup> and the PBE approximation<sup>[77]</sup> as the exchange-correlation functional. Geometry optimizations were performed using  $\Gamma$ -centered grids with 1500  $k$ -points per reciprocal atom until the forces were smaller than  $5 \text{ meV}\text{\AA}^{-1}$ . Densities of states were calculated using grids with 2000  $k$ -points per reciprocal atom and band-structure using a line density along the high-symmetry path of  $60 \times 2\pi \text{\AA}^{-1}$ . The non-spherical contributions from the gradient corrections inside the augmentation spheres were included as well. Apart from this, the remaining inputs (e.g., pseudopotential choice and Hubbard parameters) were chosen to be the same as recommended by the Materials Project. For the band structures the notation of ref. [87] was used to build the paths in reciprocal space, with the conversion to the standard representation being handled by the pymatgen package.<sup>[46]</sup> The deformation potentials were computed from the band structures calculated at three different optimized cell volumes: the optimized volume, a volume compressed by 3%, and a volume expanded by 3%. For the distorted cells, a geometry optimization at a fixed volume was performed. Finally, the deformation potentials were obtained by fitting a first-order polynomial to reproduce the resulting  $E_g$  and  $\ln(V)$  data.

**Machine Learning:** CGATs were used for the prediction of the distance to the convex hull and the volume of the crystal structures. CGATs are message-passing networks on crystal graphs relying on the attention mechanism<sup>[88]</sup> to construct the messages and updates. The vector representing the  $i$ th node, that is atom, at time step  $t$  of the message-passing process was denoted as  $h_i^t$  and the corresponding

edge to the atom  $j$  as  $e_{ij}^t$ . In general, the message passing and update equation could be summarized as

$$h_i^{t+1} = U \left( h_i^t, \{ h_j^t, e_{ji}^t \}, j \in \mathcal{N}(i) \right) \quad (3)$$

where  $\mathcal{N}(i)$  is the neighborhood of the  $i$ th node determined by a cutoff radius and a maximum number of neighbors within that cutoff radius. The messages  $m_{ij}^n$  and attention vectors  $a_{ij}^n$  were calculated by a fully connected network from a concatenation of the previous node and edge embeddings.  $n$  networks were run for messages and attention coefficients in parallel, each representing one so-called attention head. Here, FCNN $_a^{t,n}$  is the network of the  $n$ th attention head at timestep  $t$ .

$$s_{ij}^{t,n} = \text{FCNN}_a^{t,n}(h_i^t \parallel h_j^t \parallel e_{ij}^t) \quad (4)$$

$$a_{ij}^{t,n} = \frac{\exp(s_{ij}^{t,n})}{\sum_j \exp(s_{ij}^{t,n})} \quad (5)$$

$$m_{ij}^{t,n} = \text{FCNN}_m^{t,n}(h_i^t \parallel h_j^t \parallel e_{ij}^t) \quad (6)$$

$$h_i^{t+1} = h_i^t + \text{HFCNN}_{\theta_t} \left( \left\| \sum_j a_{ij}^n m_{ij}^n \right\| \right) \quad (7)$$

In Equation (7) the messages weighted by the attention coefficients through a sum were combined and then the attention heads were averaged. The resulting vector entered a hyper-network that was calculated from the difference between the starting node representation and the node representation at timestep  $t$ . The edges were updated in a similar manner

$$s_{ij}^{e,n} = \text{FCNN}_a^{e,n}(h_i^t \parallel h_j^t \parallel e_{ij}^t) \quad (8)$$

$$a_{ij}^{e,n} = \frac{\exp(s_{ij}^{e,n})}{\sum_n \exp(s_{ij}^{e,n})} \quad (9)$$

$$m_{ij}^{e,n} = \text{FCNN}_m^{e,n}(h_i^t \parallel h_j^t \parallel e_{ij}^t) \quad (10)$$

$$e_{ij}^{t+1} = e_{ij}^t + \text{FCNN}_{\theta_t} \left( \left\| \sum_n a_{ij}^{e,n} m_{ij}^{e,n} \right\| \right) \quad (11)$$

After the last message passing step, the atomic representations were concatenated with a global context vector calculated with a ROOST<sup>[30]</sup> model and then combined through another attention layer. Finally, the target quantity was calculated with a residual neural network.

The networks CGAT-2 and CGAT-3 trained for this publication both used the same hyperparameters, specifically, optimizer, AdamW; learningrate, 0.000125; starting embedding, matscholar-embedding; nbr-embedding-size, 512; msg-heads, 6; batch-size, 512; max-nbr, 24; epochs, 390; loss, L1-loss; momentum, 0.9; weight-decay,  $1 \times 10^{-6}$ ; atom-fea-len, 128; message passing steps, 5; roost message passing steps, 3; other roost parameters, default; vector-attention, True; edges, updated; learning rate, cyclical; learning rate schedule, (0.1, 0.05); learning rate period, 130; hyper network, three hidden layers, size 128; hyper network activ. funct., tanh; FCNN, one hidden layer, size 512; FCNN activ. funct., leaky RELU.<sup>[89]</sup>

Due to the size of DCGAT-3 it was decided to only use validation and test set sizes of 5% which still encompassed 156 483 materials. The training of each network cost  $\approx 7$  days on 8 NVIDIA V100 GPUs, that is,  $\approx 1000$  GPU hours.

**Transfer Learning:** The high-throughput searches with transfer learning were started with the CGAT-1 network and one round of predictions was performed for the selected crystal structures. Using a cutoff of 200 meV atom<sup>-1</sup> validation calculations were performed with DFT and the resulting data was used to transfer learn a separate network for each prototype. Here the learning rate was reduced by a factor of ten and the batch-size by a factor of 8 in comparison to the normal training and the network was optimized until the validation error converged.

Depending on the number of stable compounds that were found during the next cycle of predictions and the error of the network up to three cycles of transfer learning were performed. As can be seen in Table S1, Supporting Information, for all except two prototypes, the MAE was already sufficiently small after one round of transfer learning. Only the garnets and the Ruddlesden–Popper layered perovskites required a second round of data accumulation and training to reach such a small MAE.

## Supporting Information

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

## Acknowledgements

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer SUPERMUC-NG at Leibniz Supercomputing Centre (www.lrz.de) under the project pn25co. T.F.T.C., P.J.M.A.C., and P.B. acknowledge financial support from FCT - Fundação para a Ciência e Tecnologia, Portugal (projects UIDB/04564/2020 and UIDP/04564/2020 and contract 2020.04225. CEECIND) and computational resources provided by the Laboratory for Advanced Computing at University of Coimbra.

Open access funding enabled and organized by Projekt DEAL.

## Conflict of Interest

The authors declare no conflict of interest.

## Author Contributions

J.S. and N.H. performed the training of the machines and the machine learning predictions of the distance to the hull; M.A.L.M. and H.C.W. performed the DFT high-throughput calculations; P.J.M.A.C. performed the training and the machine learning predictions for the material properties; P.B. and T.F.T.C. performed the calculations of the material properties; and M.A.L.M. and S.B. directed the research. All authors contributed to the analysis of the results and the writing of the manuscript.

## Data Availability Statement

The data that support the findings of this study are openly available in Materials Cloud at <https://doi.org/10.24435/materialscloud:m7-50>, reference number 126.

## Keywords

high-throughput density functional theory calculations, machine learning material science, material discovery, superconductivity, superhard materials

Received: November 20, 2022

Revised: February 28, 2023

Published online: April 7, 2023

- [1] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **2013**, *12*, 191.
- [2] 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.
- [3] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *Npj Comput. Mater.* **2019**, *5*, 83.
- [4] H. J. Kulik, T. Hammerschmidt, J. Schmidt, S. Botti, M. A. L. Marques, M. Boley, M. Scheffler, M. Todorović, P. Rinke, C. Oses, A. Smolyanyuk, S. Curtarolo, A. Tkatchenko, A. P. Bartók, S. Manzhos, M. Ihara, T. Carrington, J. Behler, O. Isayev, M. Veit, A. Grisafi, J. Nigam, M. Ceriotti, K. T. Schütt, J. Westermayr, M. Gastegger, R. J. Maurer, B. Kalita, K. Burke, R. Nagai, et al., *Electron. Struct.* **2022**, *4*, 023004.
- [5] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Phys. Rev. Lett.* **2016**, *117*, 135502.
- [6] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, C. Wolverton, *Phys. Rev. Mater.* **2018**, *2*, 123801.
- [7] F. Legrain, J. Carrete, A. van Roekeghem, G. K. Madsen, N. Mingo, *J. Phys. Chem. B* **2017**, *122*, 625.
- [8] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Phys. Rev. Lett.* **2016**, *117*, 135502.
- [9] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M. A. L. Marques, *Chem. Mater.* **2017**, *29*, 5090.
- [10] J. Schmidt, L. Chen, S. Botti, M. A. L. Marques, *J. Chem. Phys.* **2018**, *148*, 241728.
- [11] F. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Int. J. Quantum Chem.* **2015**, *115*, 1094.
- [12] W. Li, R. Jacobs, D. Morgan, *Comput. Mater. Sci.* **2018**, *150*, 454.
- [13] J. Carrete, N. Mingo, S. Wang, S. Curtarolo, *Adv. Funct. Mater.* **2014**, *24*, 7427.
- [14] G. Hautier, C. Fischer, V. Ehrlicher, A. Jain, G. Ceder, *Inorg. Chem.* **2011**, *50*, 656.
- [15] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2017**, *96*, 024104.
- [16] G. R. Schleder, C. M. Acosta, A. Fazzio, *ACS Appl. Mater. Interfaces* **2020**, *12*, 20149.
- [17] Z. Li, Q. Xu, Q. Sun, Z. Hou, W.-J. Yin, *Adv. Funct. Mater.* **2019**, *29*, 1807280.
- [18] S. P. Ong, L. Wang, B. Kang, G. Ceder, *Chem. Mater.* **2008**, *20*, 1798.
- [19] S. P. Ong, A. Jain, G. Hautier, B. Kang, G. Ceder, *Electrochem. Commun.* **2010**, *12*, 427.
- [20] R. Zhu, S. I. P. Tian, Z. Ren, J. Li, T. Buonassisi, K. Hippalgaonkar, *ACS Omega* **2023**, *8*, 8210.
- [21] H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain, G. Ceder, *Chem. Mater.* **2022**, *34*, 7323.
- [22] A. Lee, S. Sarker, J. E. Saal, L. Ward, C. Borg, A. Mehta, C. Wolverton, *Commun. Mater.* **2022**, *3*, 73.
- [23] J. Jang, G. H. Gu, J. Noh, J. Kim, Y. Jung, *J. Am. Chem. Soc.* **2020**, *142*, 18836.
- [24] M. Aykol, V. I. Hegde, L. Hung, S. Suram, P. Herring, C. Wolverton, J. S. Hummelshøj, *Nat. Commun.* **2019**, *10*, 2018.



- [25] A. Davariashtiyani, Z. Kadkhodaie, S. Kadkhodaie, *Commun. Mater.* **2021**, 2, 115.
- [26] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, 58, 218.
- [27] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, 43, 676.
- [28] J. F. Rodrigues, L. Florea, M. C. F. de Oliveira, D. Diamond, O. N. Oliveira, *Discov. Mater.* **2021**, 1, 12.
- [29] D. Jha, L. Ward, A. Paul, W.-k. Liao, A. Choudhary, C. Wolverton, A. Agrawal, *Sci. Rep.* **2018**, 8, 17593.
- [30] R. E. A. Goodall, A. A. Lee, *Nat. Commun.* **2020**, 11, 6280.
- [31] X. Zheng, P. Zheng, R.-Z. Zhang, *Chem. Sci.* **2018**, 9, 8426.
- [32] X. Zheng, P. Zheng, L. Zheng, Y. Zhang, R.-Z. Zhang, *Comput. Mater. Sci.* **2020**, 173, 109436.
- [33] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, *Npj Comput. Mater.* **2020**, 6, 138.
- [34] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, G. Ceder, *Npj Comput. Mater.* **2020**, 6, 97.
- [35] C. W. Park, C. Wolverton, *Phys. Rev. Mater.* **2020**, 4, 063801.
- [36] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, M. A. L. Marques, *Sci. Adv.* **2021**, 7, eabi7948.
- [37] R. E. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, A. A. Lee, *Sci. Adv.* **2021**, 8, abn4117.
- [38] J. Gibson, A. Hire, R. G. Hennig, *npj Comput. Mater.* **2022**, 8, 211.
- [39] W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, *J. Am. Chem. Soc.* **2022**, 144, 4819.
- [40] G. Restrepo, *Digital Discovery* **2022**, 1, 568.
- [41] J. H. Montoya, K. T. Winthler, R. A. Flores, T. Bligaard, J. S. Hummelshøj, M. Aykol, *Chem. Sci.* **2020**, 11, 8517.
- [42] W. Ye, X. Lei, M. Aykol, J. H. Montoya, *Sci. Data* **2022**, 9, 302.
- [43] A. Dunn, J. Brenneke, A. Jain, *J. Phys.: Mater.* **2019**, 2, 034002.
- [44] H. Wang, J. Schmidt, S. Botti, M. A. L. Marques, *J. Mater. Chem. A* **2021**, 9, 8501.
- [45] C. Oses, E. Gossett, D. V. Hicks, F. Rose, M. J. Mehl, E. Perim, I. Takeuchi, S. Sanvito, M. Scheffler, Y. Lederer, O. Levy, C. Toher, S. Curtarolo, *J. Chem. Inf. Model.* **2018**, 58, 2477.
- [46] 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.
- [47] H. Wang, S. Botti, M. A. L. Marques, *npj Comput. Mater.* **2021**, 7, 12.
- [48] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, 120, 145301.
- [49] R. Hill, *Proc. Phys. Soc., London, Sect. A* **1952**, 65, 349.
- [50] R. L. Smith, G. E. Sandly, *Proc. - Inst. Mech. Eng.* **1922**, 102, 623.
- [51] E. Mazhnik, A. R. Oganov, *J. Appl. Phys.* **2019**, 126, 125109.
- [52] Y. Zuo, M. Qin, C. Chen, W. Ye, X. Li, J. Luo, S. P. Ong, *Mater. Today* **2021**, 51, 126.
- [53] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, 31, 3564.
- [54] F. Mouhat, F. m. c.-X. Coudert, *Phys. Rev. B* **2014**, 90, 224104.
- [55] M. Born, *Math. Proc. Cambridge Philos. Soc.* **1940**, 36, 160.
- [56] G. Akopov, L. E. Pangilinan, R. Mohammadi, R. B. Kaner, *APL Mater.* **2018**, 6, 070901.
- [57] L. E. Pangilinan, S. Hu, S. G. Hamilton, S. H. Tolbert, R. B. Kaner, *Acc. Mater. Res.* **2022**, 3, 100.
- [58] S. A. Tawfik, P. Nguyen, T. Tran, T. R. Walsh, S. Venkatesh, *J. Phys. Chem. C* **2022**, 126, 15952.
- [59] K. Momma, F. Izumi, *J. Appl. Cryst.* **2011**, 44, 1272.
- [60] M. T. Yeung, R. Mohammadi, R. B. Kaner, *Annu. Rev. Mater. Res.* **2016**, 46, 465.
- [61] B. Lilia, R. Hennig, P. Hirschfeld, G. Profeta, A. Sanna, E. Zurek, W. E. Pickett, M. Amsler, R. Dias, M. I. Eremets, C. Heil, R. J. Hemley, H. Liu, Y. Ma, C. Pierleoni, A. N. Kolmogorov, N. Rybin, D. Novoselov, V. Anisimov, A. R. Oganov, C. J. Pickard, T. Bi, R. Arita, I. Errea, C. Pellegrini, R. Requist, E. K. U. Gross, E. R. Margine, S. R. Xie, Y. Quan, et al., *J. Phys.: Condens. Matter.* **2022**, 34, 183002.
- [62] W. L. McMillan, *Phys. Rev.* **1968**, 167, 331.
- [63] K. Choudhary, K. Garrity, *npj Comput. Mater.* **2022**, 8, 244.
- [64] O. L. Anderson, *J. Phys. Chem. Solids* **1963**, 24, 909.
- [65] F. Stein, A. Leineweber, *J. Mater. Sci.* **2021**, 56, 5321.
- [66] H. Hosono, K. Tanabe, E. Takayama-Muromachi, H. Kageyama, S. Yamanaka, H. Kumakura, M. Nohara, H. Hiramatsu, S. Fujitsu, *Sci. Technol. Adv. Mater.* **2015**, 16, 033503.
- [67] A. Z. Tuleushev, V. N. Volodin, Y. Z. Tuleushev, *J. Exp. Theor. Phys.* **2003**, 78, 440.
- [68] P. Borlido, J. Schmidt, H.-C. Wang, S. Botti, M. A. L. Marques, *npj Comput. Mater.* **2022**, 8, 156.
- [69] R. Arnold, J. Guyonnet, Y. Giomataris, P. Pétrouff, J. Séguinot, J. Tocqueville, T. Ypsilantis, *Nucl. Instrum. Methods Phys. Res., Sect. A* **1988**, 273, 466.
- [70] B. A. Frandsen, S. D. Nickerson, A. D. Clark, A. Solano, R. Baral, J. Williams, J. Neufeind, M. Memmott, *J. Nucl. Mater.* **2020**, 537, 152219.
- [71] F. E. haj Hassan, H. Akbarzadeh, S. Hashemifar, A. Mokhtari, *J. Phys. Chem. Solids* **2004**, 65, 1871.
- [72] A. H. Reshak, Z. Charifi, H. Baaziz, *Phys. B: Condens. Matter.* **2008**, 403, 711.
- [73] G. Kresse, J. Furthmüller, *Comput. Mater. Sci.* **1996**, 6, 15.
- [74] G. Kresse, J. Furthmüller, *Phys. Rev. B* **1996**, 54, 11169.
- [75] P. E. Blöchl, *Phys. Rev. B* **1994**, 50, 17953.
- [76] G. Kresse, D. Joubert, *Phys. Rev. B* **1999**, 59, 1758.
- [77] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, 77, 3865.
- [78] K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. H. Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S. P. Ong, K. Persson, A. Jain, *Comput. Mater. Sci.* **2017**, 139, 140.
- [79] M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. K. Ande, S. van der Zwaag, J. J. Plata, C. Toher, S. Curtarolo, G. Ceder, K. A. Persson, M. Asta, *Sci. Data* **2015**, 2, 150009.
- [80] P. Giannozzi, O. Andreussi, T. Brumme, O. Bunau, M. B. Nardelli, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, M. Cococcioni, N. Colonna, I. Carnimeo, A. D. Corso, S. de Gironcoli, P. Delugas, R. A. D. Jr, A. Ferretti, A. Floris, G. Fratesi, G. Fugallo, R. Gebauer, U. Gerstmann, F. Giustino, T. Gorni, J. Jia, M. Kawamura, H.-Y. Ko, A. Kokalj, E. Küçükbenli, M. Lazzeri, et al., *J. Condens. Matter Phys.* **2017**, 29, 465901.
- [81] J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou, K. Burke, *Phys. Rev. Lett.* **2008**, 100, 136406.
- [82] M. van Setten, M. Giantomassi, E. Bousquet, M. Verstraete, D. Hamann, X. Gonze, G.-M. Rignanese, *Comput. Phys. Commun.* **2018**, 226, 39.
- [83] K. Lejaeghere, V. V. Speybroeck, G. V. Oost, S. Cottenier, *Crit. Rev. Solid State* **2014**, 39, 772503.
- [84] F. Jollet, M. Torrent, N. Holzwarth, *Comput. Phys. Commun.* **2014**, 185, 1246.
- [85] K. F. Garrity, J. W. Bennett, K. M. Rabe, D. Vanderbilt, *Comput. Mater. Sci.* **2014**, 81, 446.
- [86] P. B. Allen, R. C. Dynes, *Phys. Rev. B* **1975**, 12, 905.
- [87] W. Setyawan, S. Curtarolo, *Comput. Mater. Sci.* **2010**, 49, 299.
- [88] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Adv. Neural Inf. Process. Syst.* **2017**, 30, 5998.
- [89] S. S. Liew, M. Khalil-Hani, R. Bakhteri, *Neurocomputing* **2016**, 216, 718.