Materials Today Physics 48 (2024) 101560

Contents lists available at ScienceDirect

Materials Today Physics

journal homepage: www.journals.elsevier.com/materials-today-physics

Improving machine-learning models in materials science through large datasets

Jonathan Schmidt [a,1], Tiago F.T. Cerqueira [b,1], Aldo H. Romero [c], Antoine Loew [d], Fabian Jäger [a], Hai-Chen Wang [d], Silvana Botti [d,*], Miguel A.L. Marques [d,**]

[a] Department of Materials, ETH Zürich, Zürich, CH-8093, Switzerland
[b] CFisUC, Department of Physics, University of Coimbra, Rua Larga, 3004-516, Coimbra, Portugal
[c] Department of Physics, West Virginia University, Morgantown, WV, 26506, USA
[d] Research Center Future Energy Materials and Systems of the University Alliance Ruhr and ICAMS, Ruhr University Bochum, Universitätsstraße 150, D-44801, Bochum, Germany

ABSTRACT

The accuracy of a machine learning model is limited by the quality and quantity of the data available for its training and validation. This problem is particularly challenging in materials science, where large, high-quality, and consistent datasets are scarce. Here we present ALEXANDRIA, an open database of more than 5 million density-functional theory calculations for periodic three-, two-, and one-dimensional compounds. We use this data to train machine learning models to reproduce seven different properties using both composition-based models and crystal-graph neural networks. In the majority of cases, the error of the models decreases monotonically with the training data, although some graph networks seem to saturate for large training set sizes. Differences in the training can be correlated with the statistical distribution of the different properties. We also observe that graph-networks, that have access to detailed geometrical information, yield in general more accurate models than simple composition-based methods. Finally, we assess several universal machine learning interatomic potentials. Crystal geometries optimised with these force fields are very high quality, but unfortunately the accuracy of the energies is still lacking. Furthermore, we observe some instabilities for regions of chemical space that are undersampled in the training sets used for these models. This study highlights the potential of large-scale, high-quality datasets to improve machine learning models in materials science.

1. Introduction

The advent of machine learning has revolutionized numerous fields, allowing for the development of models with unprecedented predictive capabilities. In fields like image recognition and generation, natural language processing, and even finance, the ability to process and learn from vast amounts of data has led to remarkable advancements and practical applications [1–5]. The effectiveness of machine learning models is intrinsically linked to the quality and quantity of data used for their training and validation [6–8]. This limits the complexity and predictive power of models in fields where data is scarce and difficult to obtain, such as materials science [9–12]. In fact, experimental data collection can be slow and costly, involving sophisticated equipment and procedures. A popular workaround is to use computational data, which eliminates several of the problems: (i) it is relatively easy to obtain large amounts of computational data due to the ever increasing power of modern computers; (ii) it can be made systematic, allowing for fully consistent datasets; (iii) it can cover the whole periodic table and structural diversity, regardless of stability of the phases, difficulty in the synthesis, presence of radioactive or toxic chemical elements, etc. In most cases, density-functional theory, the workhorse theory of materials science, is used to produce the computational data. This theory provides an unmatched compromise between accuracy and computational cost, so it is not surprising that it underlies nearly all large material databases currently available.

The advent of high-throughput computational methods and increased computing power in the 2000s led to the creation of several theoretical inorganic crystal databases, such as the Materials Project [13], AFLOWLIB [14], the Open Crystallographic Database [15], the NOMAD archive [16], JARVIS [17], the Materials Commons [18], and the Open Quantum Materials Database [19,20]. From the experimental side, there are also some available databases, such as the open High Throughput Experimental Materials Database [21], the ASM alloy phase diagrams database [22], the Pearson's Crystal Data [23], the Cambridge

* Corresponding author.

** Corresponding author.

E-mail addresses: silvana.botti@rub.de (S. Botti), miguel.marques@rub.de (M.A.L. Marques).

[1] These authors contributed equally.

https://doi.org/10.1016/j.mtphys.2024.101560
Received 10 July 2024; Received in revised form 12 September 2024; Accepted 20 September 2024
Available online 25 September 2024

Structural Database [24], and the Inorganic Crystal Structure Database [25,26]. Most theoretical platforms not only store experimental data and the reoptimised structure obtained by density-functional theory, but also contain computed properties of the materials, significantly aiding the prediction and design of new materials. In this context, the integration of artificial intelligence in material synthesis has been a game-changer [11,27,28]. Materials Project [13], NIST [29] and AFLOWLIB [14], for instance, now provide extensive datasets for training machine-learning models, showcasing the synergy between computational materials science and artificial intelligence.

The generalization ability and effectiveness of various machine-learning models are deeply intertwined with the availability and richness of materials databases. This abundance and diversity of data are crucial for several reasons. Structural and compositional diversity within the dataset are required for the machine-learning models to capture the vast material space, enhancing their predictive accuracy and applicability across different scenarios and applications. Large datasets also allow us to scale to larger models with new emerging capabilities.

In this work, we showcase and make available a large database of density-functional calculations of inorganic materials that we call ALEXANDRIA. The purpose of this collection of datasets is fourfold. (i) They can be used in the discovery of new compounds, especially thermodynamically stable compounds, as these are the most relevant for materials science. With this in mind, we design workflows that steer the construction of the datasets to include compounds close to the convex hull of stability. (ii) They provide, on the same footing, calculations for one- (1D), two- (2D), and three-dimensional (3D) materials. As lower-dimensional compounds often contain atoms with lower coordination, the inclusion of these materials is key to provide structural diversity and varied local geometrical environments; (iii) They allow to combine calculations at different levels of theory, to enhance the quality of the predictions. This also enables the use of transfer learning [30] and related techniques, to train more accurate models from established lower-accuracy models. (iv) They are easily accessible both for materials science tasks as well as machine learning applications.

The ALEXANDRIA database is available under an open-source license and is also accessible through the OPTIMADE API [31], a recent effort to unify the interfaces of materials databases. Additionally, we provide an interface for machine learning applications through the Open MatSci ML Toolkit [32]. This comprehensive dataset addresses the critical need for large-scale, high-quality data in materials science. To show how this enables the development and refinement of machine-learning models, we use our large datasets to train a series of models of different complexities to predict a set of material properties. In this way we not only generate rather accurate and reliable models for the community, but also measure how the error of the model varies with increasing data, and how the final error correlates with the statistical distribution of the predicted property.

The remainder of this work is structured as follows. We begin with a brief overview of ALEXANDRIA, including the rationale and the many choices behind it. We also show how machine learning has accelerated the creation of the database and enabled the discovery of numerous thermodynamically stable compounds. We then present compositional models and crystal graph networks trained to predict structure-property relationships. We further discuss universal machine learning interatomic potentials trained and/or validated on ALEXANDRIA. Finally, we present our conclusions and an outlook on the future impact on materials science of combining machine learning with large materials databases.

2. Results

2.1. Database

Initially, ALEXANDRIA was created through various prototype-based high-throughput searches aimed at identifying new thermodynamically stable materials [33–38]. Recently, the database has been significantly expanded to include a larger number of compounds, as well as 1D and 2D systems [39], and by adding calculations performed using different approximations to the exchange correlation functional of density functional theory. The extension to new dimensions has been made possible by recent advances in machine learning models, which have greatly accelerated our workflow. Specifically, we use crystal graph attention networks (CGAT) [34] to predict compounds near the convex hull and universal machine learning interatomic potentials (MLIPs) for preliminary geometry optimisations of the crystal structures.

For convenience, we separate ALEXANDRIA in different datasets, specifically (i) 3D compounds computed with the Perdew-Burke-Ernzerhof (PBE) approximation to the exchange-correlation functional [40]; (ii) 2D compounds with PBE; (iii) 1D compounds with PBE; (iv) 3D compounds with PBE for solids (PBEsol) [41]; (v) 3D compounds with the strongly constrained and appropriately normed (SCAN) exchange-correlation functional [42,43]. These datasets can be accessed through OPTIMADE [31] at https://alexandria.icams.rub.de/, or downloaded in a convenient JSON format suitable for large-scale machine learning studies. Furthermore, we offer information on the complete geometry optimisation paths that can be used to train universal MLIPs.

While ALEXANDRIA represents a significant advancement in the compilation and accessibility of DFT data for inorganic materials, it is important to acknowledge its limitations, which are mostly shared by similar databases developed in recent years. One important issue is the inherent error associated with the use of the PBE functional [44–46]. Although alternative functionals, such as PBEsol [41] and SCAN [42] (used in ALEXANDRIA) or R2SCAN [47] (introduced in the Materials Project [48]) offer improvements in certain properties, their accuracy must still be taken into account when interpreting results, as discrepancies between calculated and experimental values remain.

Another critical limitation is the neglect of defects and non-stoichiometric compounds in our calculations. These factors can stabilize (or destabilize) certain phases. While density functional theory has the potential to account for these phenomena, their incorporation in our workflow is computationally prohibitive. MLIPs may however offer a practical solution in the near future by reducing computational costs, allowing for more comprehensive modeling that includes these effects.

Entropy effects are also neglected in the current datasets. There are two main sources of entropy that are relevant: vibrational entropy (including anharmonic effects [49]) and configurational entropy [50], which is crucial for the formation of alloyed phases. Ignoring these effects can lead to inaccuracies [51], such as the prediction of stable ordered phases when the actual experimental system is a disordered alloy, or the incorrect identification of stable stoichiometric low-symmetry phases when the true ground state is a high-symmetry phase with partial occupations. These issues have been highlighted in a recent critique [52] in response to the release of the GNoME convex hull of stability [53]. However, by closely examining the database in regions of materials space with many related crystal structures, differing only by the substitution of similar chemical elements, we believe that these problems can be identified and corrected. This process can be automated, improving the accuracy of phase diagrams derived from the database.

2.1.1. 3D compounds with PBE

The collection of PBE calculations of 3D crystalline materials is the largest dataset in ALEXANDRIA containing approximately 4.5 million entries for 2.6 million unique chemical compositions. Currently we count 1.7 thousand elementary substances, 241 thousand binaries, 2.96 million ternaries, 1.27 million quaternaries and 15 thousand higher multinary compounds. Together, these entries form a convex hull of thermodynamic stability with 116 thousand vertices. The figures given represent calculations up to the date of submission of this paper, but we will continue to expand the database.

In ALEXANDRIA we are mostly interested in thermodynamically stable

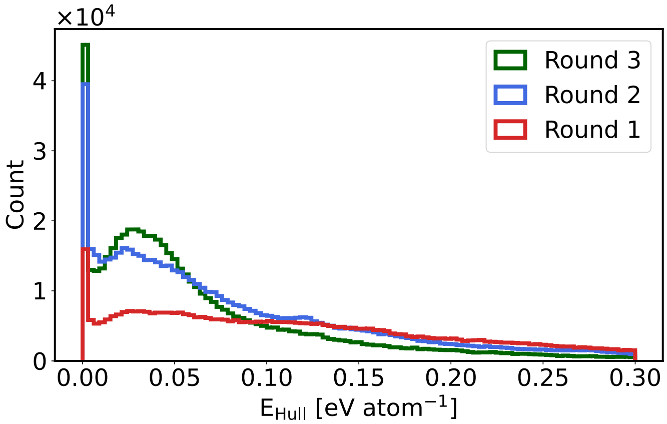

Fig. 1. Distribution of the distances to the convex hull (E_{hull}) for each round. After each round the CGAT model was retrained. Omitting entries with $E_{hull} >$ 0.3 eVatom^{-1}.

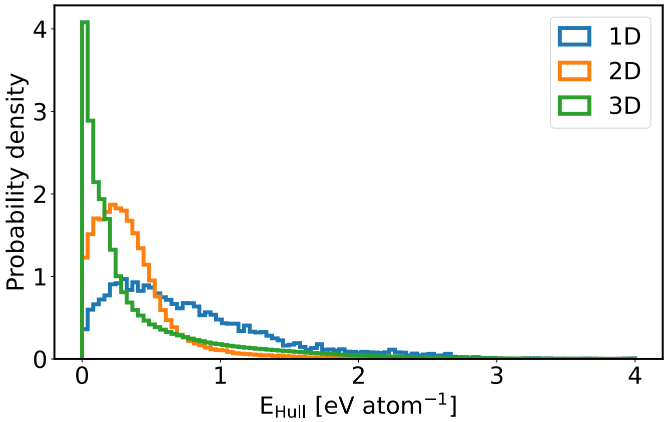

Fig. 2. Distribution of the distances to the convex hull (E_{hull}) for the different databases. The area of each histogram is normalized to one. The 1D, 2D and 3D datasets contain, respectively, 13 thousand, 138 thousand and 4.5 million entries. Omitting entries with $E_{hull} >$ 4 eVatom^{-1}.

or weakly unstable materials, so our workflow for selecting crystal structures is geared towards this goal. We note, however, that unstable compounds are also relevant for training machine learning models, as they inform the model about this part of the material space (particularly important for thermodynamic stability models, for example).

The current workflow largely follows that described in Ref. [35]: (i) For each chemical composition (e.g. AB_2C_7), we select from the database all structural prototypes up to a certain number of atoms in the primitive unit cell (usually up to 32 atoms) (ii) We perform an exhaustive enumeration of all possible combinations of chemical elements in these structural prototypes. We use 83 chemical elements, including most of the periodic table with the exception of the noble gases. (iii) In the next step, the CGAT model is used to predict the distance to the convex hull and to select the most stable crystal structure for each composition. We emphasise that our CGAT model does not use exact bond lengths, but is based only on graph distances of neighbours, and is therefore adequate for this workflow. The accuracy of the model is comparable to the best universal MLIPs [54,55], while being considerably cheaper to evaluate, as we discuss in Table 5. (iv) In contrast to Ref. [35], all compounds below a given cutoff distance to the convex hull (usually between 0.025 and 0.2 eVatom^{-1}, depending on the complexity of the structure) are pre-relaxed with a universal force field. (v) The final geometry relaxation with VASP leads to the result that is added to the database. (vi) After accumulating about half a million new entries, the CGAT model is retrained to increase its accuracy. We find that this value is a good compromise between the numerical effort required to retrain the model and the current size of ALEXANDRIA. We performed three rounds of the workflow, amounting to scanning approximately 19 billion materials. Round #1 was already presented in our previous work [35], and consisted of the scan of ~1 billion binary and ternary materials and subsequent validation of 510 thousand compounds. Round #2 scanned ~13.1 billion quaternaries (80 chemical elements, 347 prototypes, 665 thousand validation calculations), and round #3 searched through ~ 150 million binaries and 4.5 billion ternaries (584 thousand validation calculations).

A histogram illustrating the distribution of distances to the convex hull in each round can be observed in Fig. 1. The peak at zero is attributed to the thermodynamically stable compounds. It can be observed that this peak and the distribution nearby increase with each subsequent round, while the tails decay more rapidly. This is a consequence of the improvement in accuracy of the CGAT model, which becomes more effective at predicting thermodynamic stability. Indeed, the proportion of compounds found below 0.1 eVatom^{-1} increased from 42.6 % in round #1 to 63.4 % in round #2 and to 74.8 % in round #3. For hull materials, the fractions were 2.0 %, 3.6 % and 5.4 %. This can be compared, for example, with Mattergen [56], which obtained 3 % of the compounds on the convex hull. The distribution of the distance to the hull of all compounds in ALEXANDRIA can be seen in Fig. 2. It is highly peaked at zero and decays rapidly with increasing E_{hull}, demonstrating the effectiveness of the workflow.

In Fig. 3 we depict the most common structural types observed within the convex hull of stability. It is to be expected that, due to the combinatorial nature of the chemical composition, quaternary and ternary compounds will occupy the top places. A significant proportion of families exhibit highly symmetric crystal structures, such as cubic or trigonal, although a diverse range of geometric motifs can be observed. The most prevalent is a quaternary family with a layered structure analogous to the trigonal delafossites with composition ABC_2. These, in turn, constitute the third most common family in our dataset, accounting for more than 1 % of the compounds. As might be expected, the second most common structural type is the cubic Heusler compounds. The distribution of structural types in the convex hull exhibits a long, right-skewed tail. From the almost 14 000 different structural prototypes present in the hull, a mere 200 are responsible for approximately half of the compounds, and 1000 for 75 %.

2.1.2. 2D compounds with PBE

Due to their reduced dimensionality, 2D materials exhibit a number of intriguing properties that are uncommon in bulk materials. They are often regarded as a significant potential for post-silicon electronics [57,58], spintronics [59] and numerous other applications [60–62]. As in the case of 3D materials, the necessity for extensive characterization of 2D crystal structures has resulted in the creation of databases containing thousands of entries, as evidenced by Refs. [63–65].

It is unfortunate that, in the two-dimensional context, there is not yet a plethora of experimental and theoretical structural prototypes available, as is the case in the three-dimensional world. To circumvent this problem, Ref. [39] employed an approach based on the systematic generation of crystal structures based on their space group and Wyckoff positions. The aforementioned positions were occupied by chemical species that respected the principle of charge neutrality and the Pauling test for neutral materials. A universal MLIP was also employed for preliminary geometry optimisation (see section II D.) Once a sufficient number of compounds had been generated, a CGAT model was retrained using transfer learning, and a similar workflow as described above for 3D compounds was applied.

The strategy has been continued in order to expand the 2D dataset, which now contains 138 000 entries. The distribution of distances to the hull for the 2D materials can be observed in the orange histogram in Fig. 2. This distribution is not monotonically decreasing in the same manner as its 3D counterpart; rather, it exhibits a peak at approximately

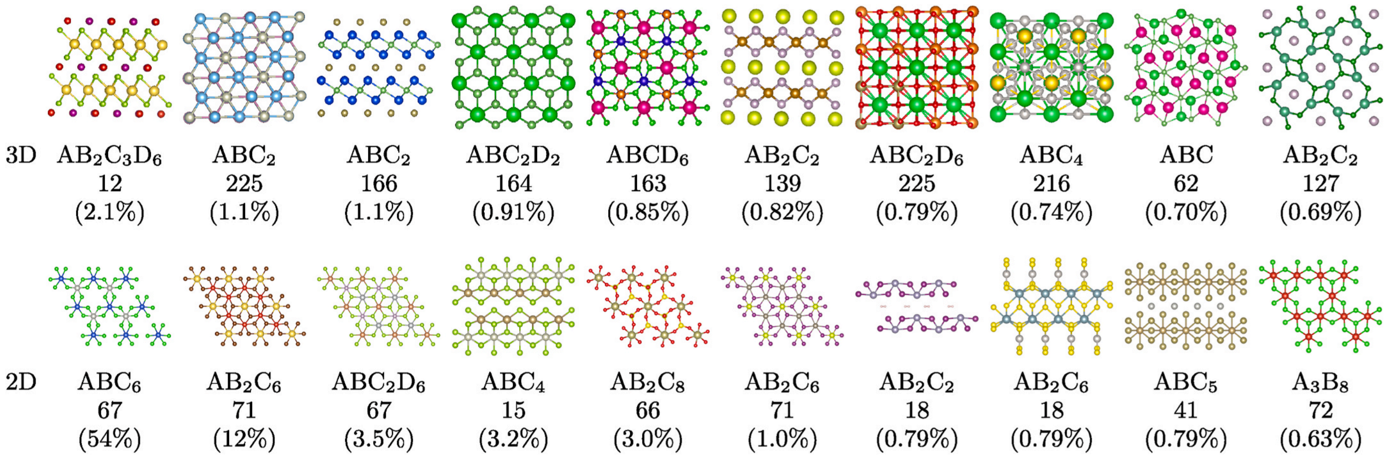

Fig. 3. Structural prototypes most common on the convex hull of thermodynamic stability in the 3D (top) and 2D (bottom) PBE datasets.

0.04 eV. The reason for this discrepancy is that the systematic generation of structures is considerably less efficient than CGAT in discovering stable compounds. It is essential to note that this 2D database pertains solely to single (or double) layer 2D materials and does not encompass any information pertaining to heterostructures.

At the present time, ALEXANDRIA contains 631 2D materials on the convex hull and 8663 below 0.05 eVatom^{-1}. It should be noted that the concept of the convex hull, which is employed as a measure of synthesizability for bulk materials, is less well defined for low-dimensional systems, given that these materials are often synthesized on substrates. The strength of the substrate interaction may be sufficient to alter the relative energy ordering of different free-standing structures or even to stabilize dynamically or thermodynamically unstable phases. An alternative approach to the synthesis of single-layer 2D materials is micromechanical cleavage. However, this necessitates an understanding of the three-dimensional counterpart, particularly with regard to the stacking order. Nevertheless, the concept of a convex hull remains a valuable tool for analysing 2D systems, particularly when constructed using bulk phases as a reference.

In Fig. 3 we present the most prevalent two-dimensional layers on the convex hull of stability. It should be noted that the group number in this case corresponds to the 2D layer group. It can be observed that the list is largely comprised of hexagonal structures. The top family belongs to the space group number 68 and has chemical composition ABC_6, where the two metallic sites (A and B) and sixfold coordinated to the non-metallic site C. This amounts to more than 300 compounds on the convex hull and a more than 1700 compounds below 0.05 eVatom^{-1}.

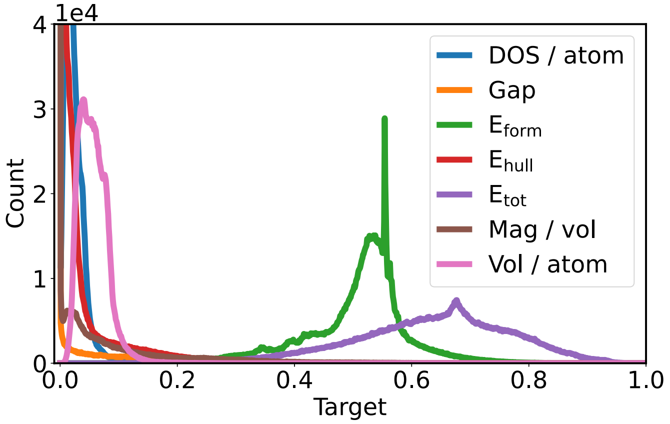

Fig. 4. Distributions of the different target properties, scaled to the interval [0, 1].

Subsequently, we find another hexagonal structure with space group 71 and chemical composition AB_2C_6, where again the two metals A and B are sixfold coordinated to the non-metallic site C. It follows then a series of families with a much lower representation in the convex hull.

2.1.3. 1D compounds with PBE

The most recent addition to ALEXANDRIA is a dataset of 1D compounds. This includes what we can describe as small-diameter nanowires and nanotubes (also known as rod geometries). As for 2D, we have very limited information on the structural diversity in 1D. Therefore, we have used the systematic approach described above to generate possible 1D compounds. At the moment the dataset is still relatively small, with 13 thousand entries, but we expect this number to grow considerably in the near future. The distribution of E_{hull} for these compounds is represented by the blue histogram in Fig. 2. Unsurprisingly, this has the widest distribution of all the datasets, with a peak just before 0.05 eVatom^{-1} and a very fat tail extending beyond 2 eVatom^{-1}. However, we expect that this data will already be useful for training universal force fields to correctly capture the unique geometric features of 1D systems.

2.1.4. 3D compounds with PBEsol and SCAN

One of the problems with current materials databases is that they are mostly constructed using a single code —VASP– and a single approximation to the exchange correlation functional —PBE. While this is the workhorse method in materials science, it increases the likelihood of introducing unwanted biases into the data. To mitigate this problem, in Refs. [30,66] we gave the first steps in building a dataset calculated with PBEsol [41] and SCAN [42]. The former has the same form as the standard PBE, but with adjusted parameters to increase its performance for solids. In fact, crystal structures optimised with PBEsol are much closer to experimental values, significantly exceeding the accuracy of PBE, which tends to overestimate lattice constants [67]. SCAN, on the other hand, has been shown to give better results than PBE in many cases [43]. However, since SCAN is a meta-GGA functional, it is computationally more demanding and, due to its construction, presents numerical instabilities [47,68]. To avoid these problems, we decided to perform single shot calculations on the already excellent PBEsol geometry.

Initially [66,69], these datasets contained almost exclusively compounds on or close to the convex hull of stability. We have supplemented these with some unstable compounds so that they can be used to train machine learning models to predict stability. Currently, the PBEsol dataset contains 415 thousand 3D compounds, defining a convex hull of stability with 47 thousand compounds. The SCAN dataset contains the same compounds, of which 50 thousand are on the convex hull. We note that by using multi-fidelity or transfer learning approaches, certain

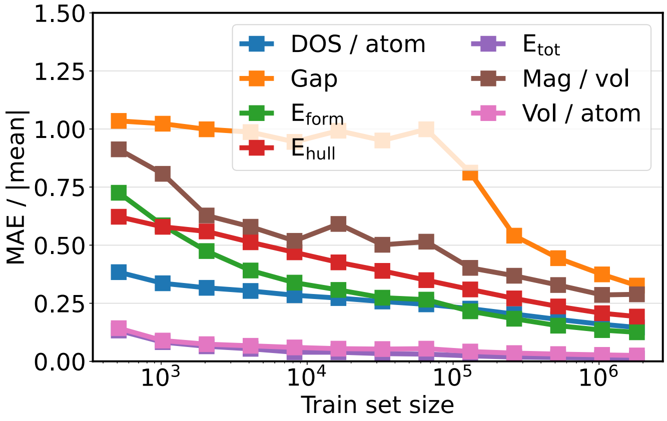

Fig. 5. Mean absolute error of the crabnet models as a function of size of training set. Errors are divided by the mean value of the corresponding distribution in order for easy comparison between the different properties.

PBEsol or SCAN models can be trained with comparable accuracy to those trained on the larger PBE dataset [30].

2.2. Compositional models

Perhaps the simplest way of modelling material properties are compositional models. In these, the crystal is represented purely by its chemical composition (the chemical formula). Of course, this has a severe limitation, namely that different allotropes cannot be distinguished by the model. Here, we chose the compositionally restricted attention-based network (crabnet) [70], which uses a self-attention mechanism to dynamically learn and update individual element representations based on their chemical environment. We trained seven different networks for the total energy, the formation energy, the distance to the convex hull of stability, the volume per atom, the absolute value of the magnetisation per unit volume, the (indirect) band gap and the density of electronic states per atom at the Fermi level. The latter will be referred to simply as the density of electronic states.

The distribution of the compositional datasets is shown in Fig. 4. Note that the data has been scaled to the interval [0, 1] in order to better compare the different quantities. It is clear from the figure that none of the distributions are normal, and that the different properties lead to very different curves. The total energy per atom has a very broad distribution with relatively short tails, while the formation energy is more peaked, with a clear maximum. Note that the very narrow peak is due to the many systems with $E_{\rm form} \sim 0$ eV (all elemental substances and many alloys), and that the majority of compositions in alexandria have a phase with negative formation energy. The distance to the convex envelope, on the other hand, has a maximum at 0 eV and decays rapidly with energy. This is partly due to the fact that alexandria is biased by construction towards compounds on or near the shell. As the vast majority of compounds in the dataset are metals, it is not surprising that the distribution of band gaps is also highly peaked at 0 eV, and decays with a fat tail for larger values [46]. Note that the band gaps are strongly underestimated as they were obtained with PBE [46]. The distribution of the density of electronic states per atom is somewhat complementary to that of the band gaps. It also has a very large peak at zero, due to semiconducting and insulating materials, and a very long tail. Most low energy compounds are non-magnetic, giving a maximum at zero magnetisation per unit volume. The distribution then shows a small peak and finally decays slowly for larger values. Finally, the volume per atom has a clear maximum and then decays slowly for larger volumes. This very long tail is unfortunately unphysical as it is due to the absence of van der Waals interactions in our workflow, leading to excessively large unit cells for some layered compounds.

Table 1
Mean and standard deviation (std) for the datasets used to train crabnet, together with the absolute and percentage error (with respect to the mean) of the models trained.

Property	Units	mean	std	MAE	err (%)
Gap	eV	0.17	0.65	0.054	33
$E_{\rm form}$	eVatom^{-1}	−0.48	0.96	0.060	13
$E_{\rm hull}$	eVatom^{-1}	0.30	0.48	0.058	19
$E_{\rm tot}$/atom	eVatom^{-1}	−5.2	1.9	0.063	1.2
Mag/vol	$\mu_{\rm B}$Å^{-3}	0.0083	0.020	0.0024	29
Vol/atom	Å^{3}atom^{-1}	24	9.1	0.62	2.5
DOS/atom	states (eV atom)$^{-1}$	0.79	0.63	0.12	15

From Fig. 5 it is evident that crabnet benefited from the large size of the dataset, without saturating even for the largest training sets used here (of 1.8 million entries). (Note that we performed only one training run for each training set size, so the curves are expected to show some erratic behaviour). Obviously, the different distributions lead to different training behaviour for each model. The machine is already able to recover 80 % of the values of $E_{\rm tot}$ and volume per atom by training on a few hundred compounds. This is to be expected, as these quantities can be approximated to some extent by a linear model with one parameter per chemical species. The most difficult property to predict with the compositional model is the band gap, the property with the most unbalanced distribution. In fact, crabnet was only able to train for band gaps when the number of training compounds was at least 10^5. Another difficult property to predict seems to be the density of electronic states per atom, which can be understood by its complicated distribution, especially close to zero. The other properties showed an intermediate behaviour, with the error usually decreasing monotonically with the size of the training set.

We summarise our results in Table 1. In the end, the three models for total energy, formation energy and distance to the hull (all related quantities) give relatively similar absolute errors (although of course with very different relative errors). In this case, and to avoid propagation of errors, one should use the model that gives the quantity of interest directly, and not take the extra step of calculating derived quantities. Errors for the magnetisation per volume and the band gap are still large, although in the latter case they are comparable to the accuracy of existing density functionals [46]. Finally, we can see that compositional models, when trained on large datasets, are good enough for many purposes, especially for properties such as total energy or volume per atom. This is especially true when considering the far superior computational efficiency of these methods compared to the more complex graph networks.

2.3. Crystal-graph networks

Crystal graph networks solve many of the problems of the composition-based models discussed above. In these models, the crystal structure is described as a graph, with nodes representing atoms and edges representing neighbours. Message-passing layers, sometimes including attention mechanisms, update the embeddings of the nodes and (sometimes) the edges, which are then combined into a vector that is used to predict the final property. These models can distinguish between most different allotropes for the same chemical composition. It is possible to train these models as force fields, but here we are interested in structure-property relationships, i.e. the prediction of a material property based on the knowledge of the minimum energy structure. We chose the Atomistic Line Graph Neural Network (alignn) [71], which performs message passing on both the interatomic bond graph and its line graph corresponding to the bond angles. This is a fairly standard and well-tested package that is sufficiently efficient (from a numerical point of view) to allow training on the millions of data points available in alexandria.

The general shape of the distribution of these traits is similar to that plotted in Fig. 4, so we do not include their plots. The results are

Table 2
Mean and standard deviation (std) for the datasets used to train ALIGNN, together with the absolute and percentage error (with respect to the mean) of the models trained.

Property	Units	mean	std	MAE	err (%)
Gap	eV	0.13	0.58	0.030	23
E_{form}	$eVatom^{-1}$	−0.25	0.93	0.016	6.5
E_{hull}	$eVatom^{-1}$	0.41	0.54	0.018	4.4
E_{tot}/atom	$eVatom^{-1}$	−5.1	2.0	0.039	0.76
Mag/vol	$\mu_B Å^{-3}$	0.0090	0.021	0.0027	30
Vol/atom	$Å^3$ $atom^{-1}$	24	9.2	0.039	0.16
DOS/atom	states $(eV\ atom)^{-1}$	0.81	0.62	0.14	17

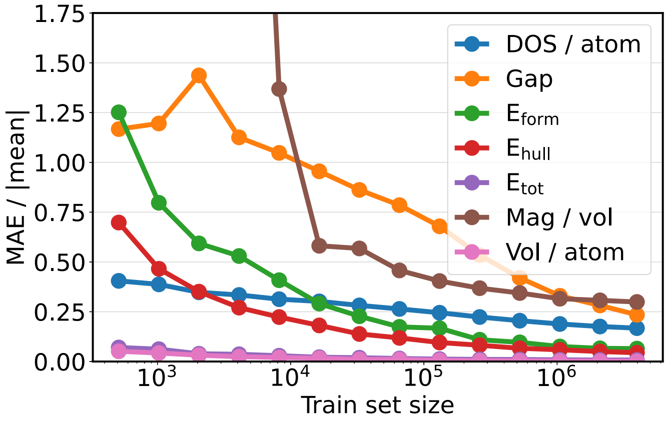

Fig. 6. Mean absolute error of the ALIGNN models as a function of size of training set. Errors are divided by the mean value of the corresponding distribution in order for easy comparison between the different properties.

summarised in Table 2 and Fig. 6.

Not surprisingly, we see that the detailed information of the crystal structure as encoded in the crystal graph is important for modelling the properties. In fact, the error of ALIGNN is significantly lower than the error of compositional models for the band gap, the formation and total energies, the distance to the hull and the volume per atom. The two exceptions are the magnetisation per unit volume and the density of electronic states, where the error is very similar between CRABNET and ALIGNN. As discussed above, these two properties follow very complicated distributions, making them particularly difficult to predict directly.

Once again, the models trained with ALIGNN generally benefited from the large dataset, although some of the models started to plateau at training set sizes above one million. This is particularly true for the total and formation energies, suggesting that improvements can be expected from using a more complex model. In any case, we emphasise the high quality of the models: for example, the absolute errors of the total and formation energies and of the distance to the shell are well below the usual threshold of chemical accuracy (1 kcal/mol = 0.04336 $eVatom^{-1}$), while 30 meV is an excellent error for the band gap. However, we should keep in mind that these models are based on the optimised geometry and the errors may increase when using less accurate crystal structures.

2.4. Universal machine-learning interatomic potentials

One of the most exciting developments in machine learning applications in materials science in recent years has been the development of universal MLIPs. Universal in this context means that the potential should be able to describe all chemical elements in all possible geometrical configurations. This is in contrast to normal MLIPs, which are usually trained for a specific chemical system and for rather restricted atomic positions. Here we consider two different MLIPs, namely M3GNET [54] and MACE [72], which were originally trained on the Materials Project [13] geometry optimisation paths.

We used the data from ALEXANDRIA in two different ways. First, we optimise a model for two-dimensional systems (model $MACE^{2D}$) and for the more accurate PBEsol data (model $M3GNET^{PBEsol}$ and $MACE^{PBEsol}$), and second, we provide an independent evaluation of universal force fields previously proposed in the literature [54,72]. In addition, we compare with the transfer-learned $M3GNET^{2D}$ from Ref. [39], which was used to create the 2D database and was trained on eight times fewer structures than the $MACE^{2D}$ model.

In Fig. 7 we plot the errors of the MLIPS studied here, calculated for the PBE datasets. We are not only interested in the absolute error of these potentials, but also in their transferability, so we separate the plots into 1D, 2D and 3D data. In each case, we relaxed the PBE structure present in ALEXANDRIA with the corresponding MLIP and computed the error in geometry and energy of the optimised structure. For the energy error, we decided to plot the absolute error in energy per atom. Unfortunately, it is more difficult to find a suitable metric for the geometry due to the different dimensionalities. In the end, we decided to calculate the effective volume of the systems by removing the vacuum from the 1D and 2D systems (assuming that the atoms have an effective size of 3 Å) and plot the relative error in the effective volume per atom. We would like to emphasise that there is a certain amount of data contamination, as some of the systems in ALEXANDRIA were used in the training of the MLIPs, so the errors should be seen as a lower bound. For the $MACE^{2D}$ model, we performed the same relaxations on the test set only to remove any data leakage, and obtained an MAE of 0.051 $eVatom^{-1}$ and a MAPE of 1.3 %.

We see that both the original M3GNET and MACE give similar and very high quality geometries for all dimensionalities, with an error that is usually smaller than that of the original PBE data. We recall that PBE gives a mean absolute error in lattice constants greater than 0.05 Å [73], leading to an overestimation of the value by more than 5 % [67]. In 3D the best performing MLIP is the original MACE, slightly better than the M3GNET, while in the reduced dimensions the reoptimised $MACE^{2D}$ gives the best geometries. The performance of the 2D models decreases significantly for the 3D system. However, it is interesting to note that the $MACE^{2D}$ model, trained on significantly more 2D data, shows a comparable performance to the M3GNET model in the 3D case.

It should be emphasised that all these errors are highly dependent on the removal of outliers (which we define as cases with volume differences exceeding ± 100 %). When outliers are included in the statistics, all errors increase significantly. We also note the relatively large number of compounds for which geometry optimisation failed, usually due to unphysical energies and forces returned by the force field. While M3GNET has the largest number of outliers, MACE has the most problems with failed optimisations.

In terms of energy, for 3D systems the original MACE gives the best results, closely followed by M3GNET, while for 2D and 1D $MACE^{2D}$ again has the smallest error. It is important to note that M3GNET gives the worst energies for systems with reduced dimensionality (despite the reasonable geometries), with a large systematic underestimation of the energy. Again we see the same problem with outliers and failed geometry optimisations.

In summary, we observe that the main difference between the models is the energy, while the geometries are relatively consistent. The best performing models are based on MACE, so these are the clear choice when high accuracy in energy is required. However, the MACE model has a number of shortcomings. Firstly, due to its more complex nature, it is more than an order of magnitude slower than M3GNET. Second, there are many regions of material space where MACE gives unphysical results, with forces and energies diverging. Therefore, and despite the degradation in energy and geometry quality, M3GNET may be the pragmatic choice for some workflows.

Finally, we compare the performance of M3GNET, MACE and CGAT for predicting stable structures in Table 5. To ensure zero data leakage, we run the evaluation on the new systems from round #3. The CGAT model was trained on data from ALEXANDRIA up to round #2. The MACE relaxations

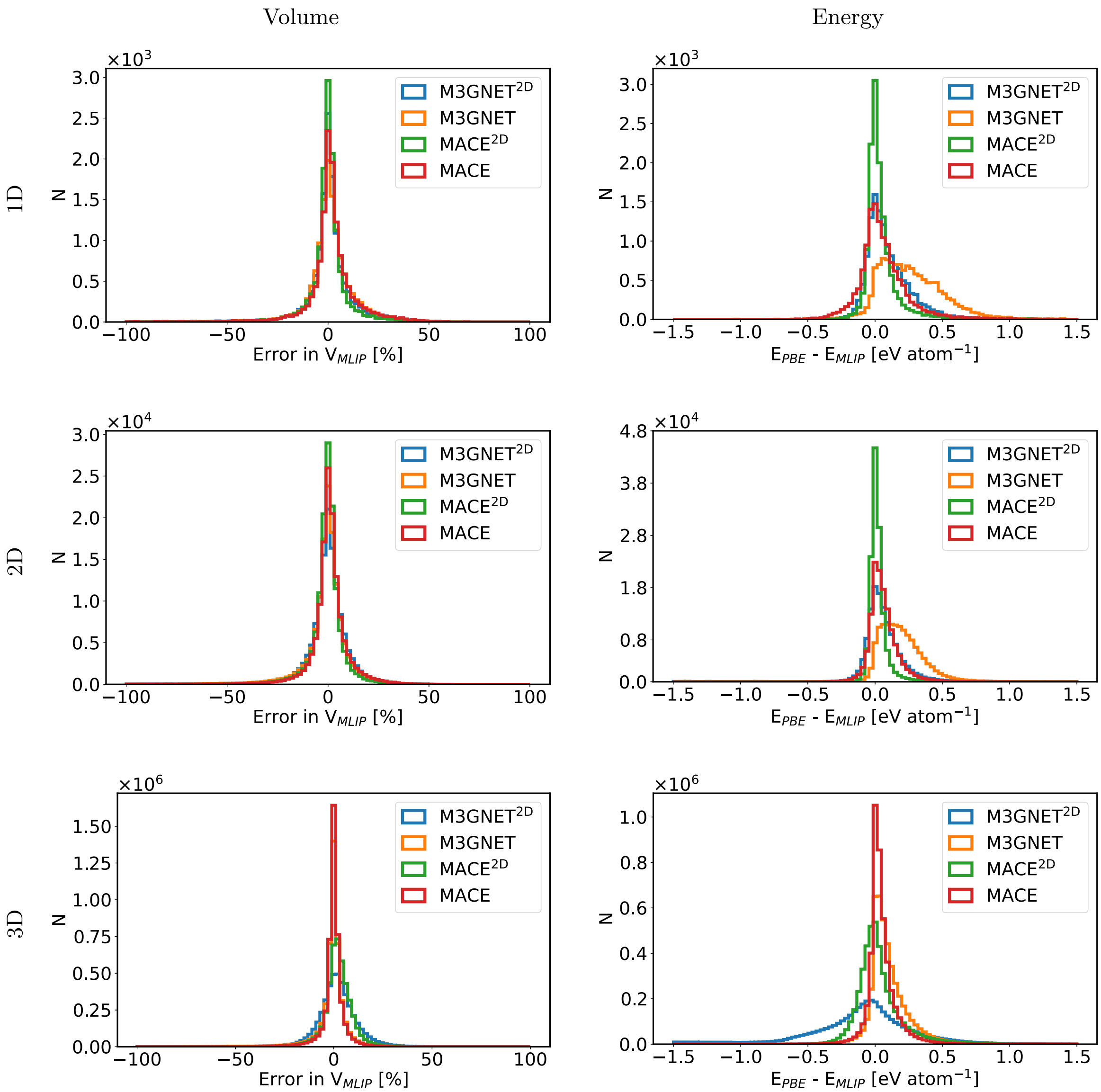

Fig. 7. Left: relative error of the effective volume per atom (see text) of the MLIP calculated on the PBE datasets of ALEXANDRIA. Right: error of the energy per atom of the MLIP calculated on the PBE datasets of ALEXANDRIA. The top panels are the results for the 1D dataset, the middle ones for 2D, and the bottom ones for 3D. For all plots the *y*-axis represents the number of test structures (*N*).

Table 3
Statistics and errors comparing the PBE and MLIP calculated equilibrium volumes. This includes the percentage of MLIP failures to the total number of PBE calculations (Failed), the proportion of outliers (Outliers, defined as cases with volume differences exceeding ± 100 %), the mean percentage error (MPE in %), the mean absolute percentage error (MAPE in %), and the mean absolute error (MAE in Å^{3}atom^{-1}). The averages are performed excluding the outliers.

		Failed	Outliers	MPE	MAPE	MAE
1D	M3GNET^{2D}	0.38	0.57	0.94	7.4	13
	M3GNET	0.44	0.38	1.5	7.8	14
	MACE^{2D}	0.33	0.21	0.018	5.9	10
	MACE	0.37	0.34	2.1	7.8	14
2D	M3GNET^{2D}	1.1	1.1	−0.72	7.3	2.5
	M3GNET	1.0	1.1	−1.3	6.9	2.4
	MACE^{2D}	1.0	0.31	−0.81	5.5	1.8
	MACE	1.4	0.58	0.46	6.2	2.1
3D	M3GNET^{2D}	0.098	0.24	1.6	7.6	2.0
	M3GNET	0.0057	0.47	−1.3	4.6	1.2
	MACE^{2D}	0.58	0.14	2.0	5.8	1.5
	MACE	0.12	0.13	−0.40	3.5	0.89

Table 4
Statistics and errors comparing the PBE and MLIP calculated energies per atom. This includes the percentage of MLIP failures to the total number of PBE calculations (Failed), the proportion of outliers (Outliers, defined as cases with energy differences exceeding ± 1.5eVatom^{-1}), the mean percentage error (MPE in %), the mean absolute percentage error (MAPE in %), and the mean absolute error (MAE in eVatom^{-1}). The averages are performed excluding the outliers.

		Failed	Outliers	MPE	MAPE	MAE
1D	M3GNET^{2D}	0.38	0.053	0.10	3.4	0.13
	M3GNET	0.44	0.12	0.27	7.2	0.28
	MACE^{2D}	0.33	0.045	0.042	1.8	0.075
	MACE	0.37	0.083	0.058	3.1	0.13
2D	M3GNET^{2D}	1.1	2.8	0.056	3.2	0.12
	M3GNET	1.0	0.056	0.20	5.4	0.20
	MACE^{2D}	1.0	0.036	0.025	1.2	0.045
	MACE	1.4	0.042	0.064	2.3	0.088
3D	M3GNET^{2D}	0.098	16	−0.10	6.5	0.29
	M3GNET	0.0057	0.10	0.10	2.8	0.12
	MACE^{2D}	0.58	0.16	0.036	2.7	0.12
	MACE	0.12	0.035	0.048	1.7	0.078

Table 5
Median absolute deviation (MAD, in eVatom^{-1}) and mean absolute error (MAE, in eVatom^{-1}) of the distance to the convex hull obtained with MACE, M3GNET and CGAT for the dataset of round #3. Here, the CGAT model was trained with the dataset that includes already round #2. We also list the average time for one relaxation (or prediction in the case of CGAT). Based on their architectures the models should all have a similar scaling with the number of atoms. In parenthesis we note the speed of MACE on one V100.

Model	MAD	MAE	Time (s)
M3GNET	0.038	0.060	1.96
MACE	0.022	0.039	12.1 (1.4)
CGAT	0.023	0.060	0.002

performed on a core or V100 GPU are about 5200 and 600 times slower than the CGATS evaluated on a V100 GPU. Note that the generation of the CGAT graphs is not included in this timing. However, as only one neighbour list is computed per prototype, this cost was negligible in our high-throughput studies. The Median Average Errors (MAD) are almost the same for MACE and CGAT, while the M3GNET model performs about 70 % worse. In terms of MAE, M3GNET and CGAT show similar performance with a 50 % increase in error compared to the MACE model (see Table 5).

3. Discussion

In this work we have presented one of the largest databases of theoretical crystal structures, distinguished as the only one that includes 3D, 2D and 1D materials. With this data, we increase the number of thermodynamically stable compounds published under a fully open licence by 69 %. Encouragingly, we find that the machine learning-accelerated high-throughput searches we perform to build the database improve with each iteration, and so far there is no limit in sight. Furthermore, by making the ALEXANDRIA database available through various task-specific interfaces, we hope to move one step closer to an environment of frictionless reproducibility in materials science [74]. The large number and diversity of stable and metastable materials in the ALEXANDRIA database is already enabling application specific high throughput searches such as for permanent magnets [75], high refractive index materials [76] and superconductors [77–79].

We also demonstrate the value of this additional data for the development of universal machine learning models, and make a variety of models - from property and stability predictions to new universal force fields - available to the community. The benefit of using larger datasets is also evident in Microsoft's recent work on generative material modelling [56]. While they initially improved their results by a factor of 1.8 (in terms of stability, uniqueness and novelty) and reduced the mean squared error by a factor of 3.1 by implementing a novel model architecture, they achieved an additional improvement by a factor of 1.6 and 5.5, respectively, simply by using a larger training dataset. This larger dataset was derived from an older version of ALEXANDRIA, which at the time consisted in approximately 600 thousand materials. Using the same criteria (distances to the convex hull less than 0.1 eVatom^{-1} and less than 20 atoms), ALEXANDRIA now contains more than 1.4 million materials, promising even greater improvements for future models.

We also expect to make significant progress by training universal force fields on the 3D geometry relaxation trajectories computed during the ALEXANDRIA database construction, considering that the dataset is more than an order of magnitude larger than the Materials Project trajectory dataset used to train MACE [72], and includes different dimensionalities, increasing its richness. With improved universal force fields, we believe we will be able to tackle the challenges of constructing temperature dependent phase diagrams, enabling high throughput studies of defects and off-stoichiometric compounds, or the systematic study of magnetic properties. However, to achieve this goal, open databases need to be systematically extended to these more complex materials.

4. Methods

4.1. DFT calculations

We performed geometry optimisations and total energy calculations using the VASP code [80,81]. All parameters, including pseudopotentials, were set to ensure compatibility with the data available in the Materials Project database [13]. To sample the Brillouin zones we used uniform Γ-centred k-point grids with a density of 1000 k-points per reciprocal atom, $(6 \times 2\pi)^2$ k-points/Å^{-2} and $(6 \times 2\pi)$ k-points/Å^{-1} for the 3D, 2D and 1D systems respectively. Spin-polarised calculations were started from a ferromagnetic configuration. We use the projector augmented wave (PAW) setup [82,83] within VASP version 5.2, applying a cutoff of 520 eV for all materials. We set the convergence criteria of the forces to be less than 0.005 eV/Å. We used the Perdew-Burke-Ernzerhof (PBE) exchange correlation functional [40] with on-site corrections for oxides and fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V or W for all 1D, 2D and 3D compounds. The U corrections for the d-states were 3.32, 3.7, 5.3, 3.9, 4.38, 6.2, 3.25 and 6.2 eV respectively. A vacuum region of 15 Å was implemented for the aperiodic directions. Some calculations did not converge due to various problems, and these entries were subsequently removed from the database.

Starting from the initial geometry, we performed a preliminary geometry optimisation using M3GNET (for 3D meshes) or M3GNET^{2D} (for 2D or 1D meshes). The choice of force field was mainly dictated by the superior numerical efficiency and stability of m3gnet. We then performed three successive relaxations with VASP and the final structure was inserted into ALEXANDRIA.

For the PBEsol calculations we used denser Γ-centred k-point grids with 2000 k-points per reciprocal atom. After converging the forces on the atoms below 0.005 eV/Å, we performed single point calculations using k-grids with 8000 k-points per reciprocal atom. In addition, if the forces in the single point calculation were found to be too large, the structure was re-optimised using stricter convergence criteria. In the end, about 3 % of all structures were reoptimised. After the PBEsol calculation, a final single point calculation was performed with the SCAN meta-GGA functional, using the same *k*-point grid as for PBEsol. As expected, the SCAN calculations were much more unstable than PBEsol, due to the known numerical instabilities of this functional [47,68], which led to a much lower average convergence rate. In any case, we were able to converge almost all calculations. Note that for the PBEsol and SCAN calculations we did not use any on-site correction.

4.2. CGAT

The CGAT model used in round #2 was trained on the data available after round #1 [35]. The CGAT models were all trained with the same hyperparameters: 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: 1e-06; 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; hypernetwork: 3 hidden layers, size 128; hypernetwork activ. funct.: tanh; FCNN: 1 hidden layer, size 512; FCNN activ. funct.: leaky RELU [84].

Outliers with a distance to the convex hull greater than 8 eVatom^{-1} were removed from the dataset. A 90/5/5 split was used for training/validation/testing, resulting in training sets of 2.8 million materials for the CGAT model used in round #2 and 3.5 million compounds for the CGAT model used in round #3. The models had respective test MAEs of 0.024 eVatom^{-1} and 0.021 eVatom^{-1} for their test sets. Unlike the ALIGNN models, they were still able to fully fit the training data, demonstrating the potential to scale to much larger datasets.

4.3. CRABNET

For all models we performed a random 70/15/15 split of the dataset into training, validation and test sets. The main PBE table of ALEXANDRIA contains about 2.6 million different chemical compositions. We decided to use the entry with the lowest formation energy of each composition for training with CRABNET, as we believe this is more scientifically relevant than, for example, taking the average of the values. This resulted in 1.8 million entries in the training set and 390 thousand each in the validation and test sets. In training the model, we mostly used the default values and did not try to optimise any hyper-parameters.

4.4. ALIGNN

We trained ALIGNN on exactly the same seven target quantities that we used with CRABNET for the 3D PBE dataset. Because of the ability to use allotropes for training, the total number of data points was now 4.6 million, of which we used 300 thousand for validation and another 300 thousand for testing. As for CRABNET, we did not try to optimise the hyperparameters of the network, and we left most parameters at their default values. The only exception was the introduction of an early stopping criterion, which stopped the training if five consecutive epochs did not improve the validation error. However, some tests showed that this early stopping did not significantly affect the accuracy of the model.

4.5. Force-fields

The dataset used to train the MACE^{2D} model consisted of relaxation trajectories from 88 thousand structures. Outliers were removed by (i) removing structures containing more than 20 sites to avoid GPU memory problems; (ii) removing isolated atom structures, i.e. structures containing atoms with no neighbouring atoms within the cut-off radius of 6 Å; and (iii) removing structures with $E_{\text{tot}} > 0$ eVatom^{-1} and $E_{\text{tot}} < -20$ eVatom^{-1}, forces greater than 20 eV/Å and stresses greater than 500 kbar. Since most of the geometry optimisation steps are performed very close to the local minimum with extremely small changes in energy and force, we imposed a minimum energy difference between steps. This means that if step n had an energy E_n, the next step added to the training/validation set would be the first one with an energy $E_n \pm$ cutoff. As a minimum, the first step, one third of the steps and the last step were included. The dataset was split with a 90/5/5 training, validation, test split, resulting in 230185 training, 12812 validation and 12838 test structures for an infinite cutoff.

For the MACE^{2D} model, cutoff energies of 0.01 eV, 0.02 eV and infinite cutoff and batch sizes of 128, 256 and 512 were tested. Other hyperparameters that were changed with respect to the default MACE version 0.2.0 are the highest order of spherical harmonics $L = 2$, node embeddings of size 192, and stochastic weight averaging with a weight of 10 for the energy after 50 epochs. We generally used single precision for training to reduce memory requirements. Based on the validation errors, we selected the MACE^{2D} model with a learning rate of 0.003 and a batch size of 128.

The PBEsol force field was trained on the relaxation paths from the PBEsol dataset with similar parameters and strategy. We kept all hyperparameters of the M3GNET model at default settings of version 0.2.4 except for an increased MLP size of 256 and a batchsize of 512. We used the same hyperparameters for the MACE$^{\text{PBEsol}}$ as for the 2D case. For both M3GNET$^{\text{PBEsol}}$ and MACE$^{\text{PBEsol}}$ we tested different energy cutoffs of 0.001 eVatom^{-1}, 0.005 eVatom^{-1} and 0.010 eVatom^{-1}, resulting in training set sizes of 1137794, 779982 and 698962 with a 90/5/5 training/val/test split. In addition, structures with $E_{\text{tot}} > 0$eVatom^{-1}, forces greater than 50 eV/Å and stresses greater than 800 kbar were discarded. For stresses and forces, the maximum component was used for outlier selection. For all force-fields we included in the cost function the error in the energies, forces, and stresses. For M3GNET the default weights are respectively 1.0, 1.0, 0.1 and for MACE 1.0, 100, 1.0 changing to 1000, 100, and 10 during the stochastic weight averaging stage.

Table 6
Statistics and errors for the universal force field trained on the PBEsol relaxation trajectories. "w.o" notes errors with the same outlier removal protocol as discussed before.

Metric	M3GNET$^{\text{PBESOL}}$	MACE$^{\text{PBESOL}}$
Failed percentage (%)	0.096	0.25
MAE (eVatom^{-1})	0.057	0.044
RMSE (eVatom^{-1})	0.086	0.083
Number of outliers	3	2
RMSE w.o. (eVatom^{-1})	0.082	0.072
90th quantile AE (eVatom^{-1})	0.122	0.090
MAD (eVatom^{-1})	0.042	0.029
MAE Volume (Å^{3}atom^{-1})	0.5	0.86

We tested the different networks again by relaxing the systems in the test set, starting from the initial PBEsol structures. We found that the MACE model with the smallest energy cutoff performed best, with an MAE of 0.044 eVatom^{-1} and a median absolute error of 0.029 eVatom^{-1}. As shown in Table 6 M3GNET$^{\text{PBEsol}}$ again showed better stability and convergence behaviour than MACE$^{\text{PBEsol}}$ with a smaller number of failed calculations, but at the cost of much higher mean errors for the energy. In terms of volume, M3GNET$^{\text{PBEsol}}$ performed 50 % better than MACE$^{\text{PBEsol}}$.

The errors here are not directly comparable with Tables 4 and 3 as the PBEsol data set is much less complete. For M3GNET$^{\text{PBEsol}}$ and MACE$^{\text{PBEsol}}$ relaxations with the FIRE optimizer with symmetry conservation and respective force convergence criteria of 0.01 eV/Å and 0.001 eV/Å were used. Stricter convergence criteria for M3GNET$^{\text{PBEsol}}$ did not result in any improvement.

CRediT authorship contribution statement

Jonathan Schmidt: Writing – review & editing, Writing – original draft, Investigation, Formal analysis. **Tiago F.T. Cerqueira:** Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis. **Aldo H. Romero:** Writing – review & editing, Writing – original draft, Resources, Data curation. **Antoine Loew:** Writing – review & editing, Writing – original draft, Data curation. **Fabian Jäger:** Formal analysis. **Hai-Chen Wang:** Writing – review & editing, Writing – original draft, Data curation. **Silvana Botti:** Writing – review & editing, Writing – original draft, Supervision. **Miguel A.L. Marques:** Writing – review & editing, Writing – original draft, Supervision, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

ALEXANDRIA can be accessed and/or downloaded from https://alexandria.icams.rub.de/ under the terms of the Creative Commons Attribution 4.0 License. All models developed in this work are freely available at https://github.com/hyllios/utils/tree/main/models/alexandria_v2.

Acknowledgements

F.J. and J.S. thank Nicola Spaldin for helpful discussions. The West Virginia University group would like to thank the Pittsburgh Supercomputer Center (Bridges2) and the San Diego Supercomputer Center (Expanse) for their support under grant DMR140031 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) programme, supported by National Science Foundation grants

#2138259, #2138286, #2138307, #2137603, and #2138296. They also acknowledge support from West Virginia Research under the Research Challenge Grand Program 2022 and NASA EPSCoR Award 80NSSC22M0173. T.F.T.C. acknowledges the financial support from FCT - Fundação para a Ciência e Tecnologia, I.P. through the projects UIDB/04564/2020, UIDP/04564/2020 and CEECINST/00152/2018/CP1570/CT0006, with DOI identifiers 10.54499/UIDB/04564/2020, 10.54499/UIDP/04564/2020 and 10.54499/CEECINST/00152/2018/CP1570/CT0006 respectively. M.A.L.M. and A.L. acknowledge partial funding from the Horizon Europe MSCA Doctoral network grant n.101073486, EUSpecLab, funded by the European Union. M.A.L.M. acknowledges the Keele Foundation through the SuperC collaboration. J.S. has also received support from the Joachim Herz Foundation. Part of the computing resources were provided by ETH Zurich and the Swiss National Supercomputing Centre (CSCS) under project no. s1128 and no. s1273. S.B. acknowledges financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the project SFB-1375 (project A02) and BO 4280/11-1. S.B. acknowledges funding from the Volkswagen Stiftung (Momentum) through the project “dandelion”. The authors thank 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 the Leibniz Supercomputing Centre (www.lrz.de) under the project pn25co. J.S. and M.A.L.M. would like to thank the NHR Centre Paderborn Center for Parallel Computing (PC2) for providing computing time on the Noctua 2 supercomputers through project hpc-prf-ecsm and hpc-prf-vibff.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv. Neural Inform. Process. Syst., 2017, pp. 5998–6008.

[2] R. Bommasani, D.A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J.Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D.E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P.W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X.L. Li, X. Li, T. Ma, A. Malik, C.D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J.C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J.S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A.W. Thomas, F. Tramèr, R.E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S.M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the Opportunities and Risks of Foundation Models, 2021 arXiv:2108.07258.

[3] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P.S. Yu, L. Sun, A Comprehensive Survey on Pretrained Foundation Models: A History from Bert to Chatgpt, 2023: 09419 arXiv:2302.

[4] F.-A. Croitoru, V. Hondru, R.T. Ionescu, M. Shah, Diffusion models in vision: a survey, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2023) 10850–10869.

[5] B. Kelly, D. Xiu, Financial machine learning, Found, Trends Finance 13 (2023) 205–363.

[6] J.F. Rodrigues, L. Florea, M.C.F. de Oliveira, D. Diamond, O.N. Oliveira, Big data and machine learning for materials science, Discov. Mater. 1 (2021) 1.

[7] A. Ng, Machine learning yearning: technical strategy for ai engineers in the era of deep learning, Retrieved online at, https://www.mlyearning.org, 2019.

[8] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, 2017.

[9] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, npj Comput. Mater. 3 (2017) 54.

[10] K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science, Nature 559 (2018) 547–555.

[11] J. Schmidt, M.R.G. Marques, S. Botti, M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. 5 (2019) 83.

[12] D.T. Speckhard, T. Bechtel, L.M. Ghiringhelli, M. Kuban, S. Rigamonti, C. Draxl, How Big Is Big Data?, 2024 arXiv:2405.11404.

[13] 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, Commentary: the materials project: a materials genome approach to accelerating materials innovation, Apl. Mater. 1 (2013): 011002.

[14] 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, Aflow: an automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58 (2012) 218–226.

[15] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R.T. Downs, A. Le Bail, Crystallography open database (cod): an open-access collection of crystal structures and platform for world-wide collaboration, Nucleic Acids Res. 40 (2011) D420–D427.

[16] C. Draxl, M. Scheffler, Nomad: the fair concept for big data-driven materials science, MRS Bull. 43 (2018) 676–682.

[17] K. Choudhary, K.F. Garrity, A.C.E. Reid, B. DeCost, A.J. Biacchi, A.R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A.G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B.G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe, F. Tavazza, The joint automated repository for various integrated simulations (jarvis) for data-driven materials design, npj Comput. Mater. 6 (2020) 173.

[18] B. Puchala, G. Tarcea, E.A. Marquis, M. Hedstrom, H.V. Jagadish, J.E. Allison, The materials commons: a collaboration platform and information repository for the global materials community, JOM 68 (2016) 2035–2044.

[19] J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (oqmd), JOM 65 (2013) 1501–1509.

[20] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, The open quantum materials database (oqmd): assessing the accuracy of dft formation energies, npj Comput. Mater. 1 (2015): 15010.

[21] A. Zakutayev, N. Wunder, M. Schwarting, J.D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, An open experimental database for exploring inorganic materials, Sci. Data 5 (2018): 180053.

[22] P. Villars, H. Okamoto, K. Cenzual, ASM Alloy Phase Diagrams Database, ASM International, Materials Park, OH, USA, 2006.

[23] P. Villars, Pearson’s Crystal Data, Crystal Structure Database for Inorganic Compounds, ASM International, 2007.

[24] C.R. Groom, I.J. Bruno, M.P. Lightfoot, S.C. Ward, The cambridge structural database, Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater. 72 (2016) 171–179.

[25] G.G.F.H. Allen, R. Sievers (Eds.), *Crystallographic Databases* (International Union of Crystallography, 1987. Chester.

[26] M. Hellenbrandt, The inorganic crystal structure database (icsd)—present and future, Crystallogr. Rev. 10 (2004) 17–22.

[27] J.-P. Lai, Y.-M. Chang, C.-H. Chen, P.-F. Pai, A survey of machine learning models in renewable energy predictions, Appl. Sci. 10 (2020) 5975.

[28] 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, R. Akashi, O. Sugino, J. Hermann, F. Noé, S. Pilati, C. Draxl, M. Kuban, S. Rigamonti, M. Scheidgen, M. Esters, D. Hicks, C. Toher, P.V. Balachandran, I. Tamblyn, S. Whitelam, C. Bellinger, L.M. Ghiringhelli, Roadmap on machine learning in electronic structure, Electron. Struct. 4 (2022): 023004.

[29] D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, A. Agrawal, Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning, Nat. Commun. 10 (2019) 5316.

[30] N. Hoffmann, J. Schmidt, S. Botti, M.A.L. Marques, Transfer learning on large datasets for the accurate prediction of material properties, Digit. Discov. 2 (2023) 1368–1379.

[31] M. Evans, J. Bergsma, A. Merkys, C. Andersen, O.B. Andersson, D. Beltrán, E. Blokhin, T.M. Boland, R. Castañeda Balderas, K. Choudhary, A. Díaz Díaz, R. Domínguez García, H. Eckert, K. Eimre, M.E. Fuentes-Montero, A.M. Krajewski, J.J. Mortensen, J.M. Nápoles-Duarte, J. Pietryga, J. Qi, F.d.J. Trejo Carrillo, A. Vaitkus, J. Yu, A. Zettel, P.B. de Castro, J.M. Carlsson, T.F.T. Cerqueira, S. Divilov, H. Hajiyani, F. Hanke, K. Jose, C. Oses, J. Riebesell, J. Schmidt, D. Winston, C. Xie, X. Yang, S. Bonella, S. Botti, S. Curtarolo, C. Draxl, L.E. E. Fuentes-Cobas, A. Hospital, Z.-K. Liu, L. Marques, A. Miguel, N. Marzari, A. J. Morris, S.P. Ong, M. Orozco, K. Persson, K.S. Thygesen, C.M. Wolverton, M. Scheidgen, C. Toher, G. Conduit, G. Pizzi, S. Grazulis, G.-M. Rignanese, R. Armiento, Developments and applications of the optimade api for materials discovery, design, and data exchange, Dig. Dis. (2024), https://doi.org/10.1039/d4dd00039k.

[32] K.L.K. Lee, C. Gonzales, M. Nassar, M. Spellings, M. Galkin, S. Miret, Matsciml: A Broad, Multi-Task Benchmark for Solid-State Materials Modeling, 2023 arXiv:2309.05934.

[33] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, M.A.L. Marques, Predicting the thermodynamic stability of solids combining density functional theory and machine learning, Chem. Mater. 29 (2017) 5090–5103.

[34] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, M.A.L. Marques, Crystal graph attention networks for the prediction of stable materials, Sci. Adv. 7 (2021): eabi7948.

[35] J. Schmidt, N. Hoffmann, H. Wang, P. Borlido, P.J.M.A. Carriço, T.F.T. Cerqueira, S. Botti, M.A.L. Marques, Machine-learning-assisted determination of the global zero-temperature phase diagram of materials, Adv. Mater. 35 (2023): 2210788.

[36] J. Schmidt, L. Chen, S. Botti, M.A.L. Marques, Predicting the stability of ternary intermetallics with density functional theory and machine learning, J. Chem. Phys. 148 (2018): 241728.

[37] H.-C. Wang, S. Botti, M.A.L. Marques, Predicting stable crystalline compounds using chemical similarity, npj Comput. Mater. 7 (2021) 12.

[38] H.-C. Wang, J. Schmidt, S. Botti, M.A.L. Marques, A high-throughput study of oxynitride, oxyfluoride and nitrofluoride perovskites, J. Mater. Chem. A 9 (2021) 8501–8513.

[39] H.-C. Wang, J. Schmidt, M.A.L. Marques, L. Wirtz, A.H. Romero, Symmetry-based computational search for novel binary and ternary 2d materials, 2D Mater. 10 (2023): 035007.

[40] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865–3868.

[41] J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L. A. Constantin, X. Zhou, K. Burke, Restoring the density-gradient expansion for exchange in solids and surfaces, Phys. Rev. Lett. 100 (2008): 136406.

[42] J. Sun, A. Ruzsinszky, J.P. Perdew, Strongly constrained and appropriately normed semilocal density functional, Phys. Rev. Lett. 115 (2015): 036402.

[43] Y. Zhang, D.A. Kitchaev, J. Yang, T. Chen, S.T. Dacek, R.A. Sarmiento-Pérez, M.A. L. Marques, H. Peng, G. Ceder, J.P. Perdew, J. Sun, Efficient first-principles prediction of solid stability: towards chemical accuracy, npj Comput. Mater. 4 (2018) 9.

[44] R. Sarmiento-Pérez, S. Botti, M.A.L. Marques, Optimized exchange and correlation semilocal functional for the calculation of energies of formation, J. Chem. Theor. Comput. 11 (2015) 3844–3850.

[45] F. Tran, J. Stelzl, P. Blaha, Rungs 1 to 4 of dft jacob's ladder: extensive test on the lattice constant, bulk modulus, and cohesive energy of solids, J. Chem. Phys. 144 (2016): 204120.

[46] P. Borlido, T. Aull, A.W. Huran, F. Tran, M.A.L. Marques, S. Botti, Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids, J. Chem. Theor. Comput. 15 (2019) 5069–5079.

[47] J.W. Furness, A.D. Kaplan, J. Ning, J.P. Perdew, J. Sun, Accurate and numerically efficient r2scan meta-generalized gradient approximation, J. Phys. Chem. Lett. 11 (2020) 8208–8215.

[48] R. Kingsbury, A.S. Gupta, C.J. Bartel, J.M. Munro, S. Dwaraknath, M. Horton, K. A. Persson, Performance comparison of r^2SCAN and scan metagga density functionals for solid materials via an automated, high-throughput computational workflow, Phys. Rev. Mater. 6 (2022): 013801.

[49] L. Monacelli, R. Bianco, M. Cherubini, M. Calandra, I. Errea, F. Mauri, The stochastic self-consistent harmonic approximation: calculating vibrational properties of materials with full quantum and anharmonic effects, J. Phys. Condens. Matter 33 (2021): 363001.

[50] C. Sutton, S.V. Levchenko, First-principles atomistic thermodynamics and configurational entropy, Front. Chem. 8 (2020), https://doi.org/10.3389/fchem.2020.00757.

[51] J. Leeman, Y. Liu, J. Stiles, S.B. Lee, P. Bhatt, L.M. Schoop, R.G. Palgrave, Challenges in high-throughput inorganic materials prediction and autonomous synthesis, PRX Energy 3 (2024): 011002.

[52] A.K. Cheetham, R. Seshadri, Artificial intelligence driving materials discovery? perspective on the article: scaling deep learning for materials discovery, Chem. Mater. 36 (2024) 3490.

[53] A. Merchant, S. Batzner, S.S. Schoenholz, M. Aykol, G. Cheon, E.D. Cubuk, Scaling deep learning for materials discovery, Nature 624 (2023) 80–85.

[54] C. Chen, S.P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci. 2 (2022) 718–728.

[55] I. Batatia, D.P. Kovacs, G.N.C. Simm, C. Ortner, G. Csanyi, MACE: higher order equivariant message passing neural networks for fast and accurate force fields, in: A.H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Adv. Neural Inf. Process. Syst., 2022.

[56] C. Zeni, R. Pinsler, D. Zügner, A. Fowler, M. Horton, X. Fu, S. Shysheya, J. Crabbé, L. Sun, J. Smith, B. Nguyen, H. Schulz, S. Lewis, C.-W. Huang, Z. Lu, Y. Zhou, H. Yang, H. Hao, J. Li, R. Tomioka, T. Xie, Mattergen: a Generative Model for Inorganic Materials Design, 2023 arXiv:2312.03687.

[57] S. Das, A. Sebastian, E. Pop, C.J. McClellan, A.D. Franklin, T. Grasser, T. Knobloch, Y. Illarionov, A.V. Penumatcha, J. Appenzeller, Z. Chen, W. Zhu, I. Asselberghs, L.-J. Li, U.E. Avci, N. Bhat, T.D. Anthopoulos, R. Singh, Transistors based on two-dimensional materials for future integrated circuits, Nat. Electron. 4 (2021) 786–799.

[58] A. Avsar, H. Ochoa, F. Guinea, B. Özyilmaz, B.J. van Wees, I.J. Vera-Marun, Colloquium: spintronics in graphene and other two-dimensional materials, Rev. Mod. Phys. 92 (2020): 021003.

[59] Y. Liu, C. Zeng, J. Zhong, J. Ding, Z.M. Wang, Z. Liu, Spintronics in two-dimensional materials, Nano-Micro Lett. 12 (2020) 1.

[60] A. Bordoloi, A.C. Garcia-Castro, Z. Romestan, A.H. Romero, S. Singh, Promises and Technological Prospects of Two-Dimensional Rashba Materials, 2024 arXiv:2404.15071.

[61] K. Khan, A.K. Tareen, M. Aslam, R. Wang, Y. Zhang, A. Mahmood, Z. Ouyang, H. Zhang, Z. Guo, Recent developments in emerging two-dimensional materials and their applications, J. Mater. Chem. C 8 (2020) 387.

[62] C. Chang, W. Chen, Y. Chen, Y. Chen, Y. Chen, F. Ding, C. Fan, H. Jin Fan, Z. Fan, C. Gong, Y. Gong, Q. He, X. Hong, S. Hu, W. Hu, W. Huang, Y. Huang, W. Ji, D. Li, L.-J. Li, Q. Li, L. Lin, C. Ling, M. Liu, N. Liu, Z. Liu, K. Ping Loh, J. Ma, F. Miao, H. Peng, M. Shao, L. Song, S. Su, S. Sun, C. Tan, Z. Tang, D. Wang, H. Wang, J. Wang, X. Wang, X. Wang, A.T.S. Wee, Z. Wei, Y. Wu, Z.-S. Wu, J. Xiong, Q. Xiong, W. Xu, P. Yin, H. Zeng, Z. Zeng, T. Zhai, H. Zhang, H. Zhang, Q. Zhang, T. Zhang, X. Zhang, L.-D. Zhao, M. Zhao, W. Zhao, Y. Zhao, K.-G. Zhou, X. Zhou, Y. Zhou, H. Zhu, H. Zhang, Z. Liu, Recent progress on two-dimensional materials, Acta Phys. Sin. 4 (2021) 2108017.

[63] J. Zhou, L. Shen, M.D. Costa, K.A. Persson, S.P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, Y.P. Feng, 2dmatpedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches, Sci. Data 6 (2019) 86.

[64] M.N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A.H. Larsen, S. Manti, T.G. Pedersen, U. Petralanda, T. Skovhus, M.K. Svendsen, J.J. Mortensen, T. Olsen, K.S. Thygesen, Recent progress of the computational 2d materials database (c2db), 2D Mater. 8 (2021): 044002.

[65] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I.E. Castelli, A. Cepellotti, G. Pizzi, N. Marzari, Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds, Nat. Nanotechnol. 13 (2018) 246–252.

[66] J. Schmidt, H.-C. Wang, T.F.T. Cerqueira, S. Botti, M.A.L. Marques, A dataset of 175k stable and metastable materials calculated with the pbesol and scan functionals, Sci. Data 9 (2022) 64.

[67] R. Hussein, J. Schmidt, T. Barros, M.A.L. Marques, S. Botti, Machine-learning correction to density-functional crystal structure optimization, MRS Bull. 47 (2022) 765–771.

[68] A.P. Bartók, J.R. Yates, Regularized scan functional, J. Chem. Phys. 150 (2019): 161101.

[69] J. Schmidt, H.-C. Wang, T.F.T. Cerqueira, S. Botti, M.A.L. Marques, A new dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals, Materials Cloud (2021). https://archive.materialscloud.org/record/2021.164.

[70] A.Y.-T. Wang, S.K. Kauwe, R.J. Murdock, T.D. Sparks, Compositionally restricted attention-based network for materials property predictions, npj Comput. Mater. 7 (2021) 77.

[71] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, npj Comput. Mater. 7 (2021) 185.

[72] I. Batatia, P. Benner, Y. Chiang, A.M. Elena, D.P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W.J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, S.M. Blau, V. Cărare, J.P. Darby, S. De, F. Della Pia, V.L. Deringer, R. Elijošius, Z. El-Machachi, F. Falcioni, E. Fako, A.C. Ferrari, A. Genreith-Schriever, J. George, R.E.A. Goodall, C.P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. Holm, J. Jaafar, S. Hofmann, K.S. Jakob, H. Jung, V. Kapil, A.D. Kaplan, N. Karimitari, J.R. Kermode, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, J.T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A.A. Naik, S.P. Niblett, S.W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A.S. Rosen, L.L. Schaaf, C. Schran, B.X. Shi, E. Sivonxay, T. K. Stenczel, V. Svahn, C. Sutton, T.D. Swinburne, J. Tilly, C. van der Oord, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W.C. Witt, F. Zills, G. Csányi, A Foundation Model for Atomistic Materials Chemistry, 2024 arXiv:2401.00096.

[73] G.-X. Zhang, A.M. Reilly, A. Tkatchenko, M. Scheffler, Performance of various density-functional approximations for cohesive properties of 64 bulk solids, New J. Phys. 20 (2018): 063020.

[74] D. Donoho, Data science at the singularity, Harvard Data Sci. Rev. 6 (2024), https://doi.org/10.1162/99608f92.b91339ef.

[75] A. Vishina, D. Hedlund, V. Shtender, E.K. Delczeg-Czirjak, S.R. Larsen, O. Y. Vekilova, S. Huang, L. Vitos, P. Svedlindh, M. Sahlberg, O. Eriksson, H. C. Herper, Data-driven design of a new class of rare-earth free permanent magnets, Acta Mater. 212 (2021): 116913.

[76] V. Trinquet, M.L. Evans, C.J. Hargreaves, P.-P. De Breuck, G.-M. Rignanese, Optical Materials Discovery and Design with Federated Databases and Machine Learning, 2024 arXiv:2405.11393.

[77] N. Hoffmann, T.F.T. Cerqueira, J. Schmidt, M.A.L. Marques, Superconductivity in antiperovskites, npj Comput. Mater. 8 (2022) 150.

[78] N. Hoffmann, T.F.T. Cerqueira, P. Borlido, A. Sanna, J. Schmidt, M.A.L. Marques, Searching for ductile superconducting heusler X_2YZ compounds, npj Comput. Mater. 9 (2023) 138.

[79] T.F.T. Cerqueira, A. Sanna, M.A.L. Marques, Sampling the materials space for conventional superconducting compounds, Adv. Mater. 36 (2024): 2307085.

[80] G. Kresse, J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6 (1996) 15–50.

[81] G. Kresse, J. Furthmüller, Efficient iterative schemes forab initiototal-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169–11186.

[82] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953–17979.

[83] G. Kresse, D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59 (1999) 1758–1775.

[84] S.S. Liew, M. Khalil-Hani, R. Bakhteri, Bounded activation functions for enhanced training stability of deep neural networks on visual pattern recognition problems, Neurocomputing 216 (2016) 718–734.