# scientific **data**

OPEN

DATA DESCRIPTOR

# A dataset of 175k stable and metastable materials calculated with the PBEsol and SCAN functionals

Jonathan Schmidt[1], Hai-Chen Wang [1], Tiago F. T. Cerqueira[2], Silvana Botti [3] & Miguel A. L. Marques [1] ✉

In the past decade we have witnessed the appearance of large databases of calculated material properties. These are most often obtained with the Perdew-Burke-Ernzerhof (PBE) functional of density-functional theory, a well established and reliable technique that is by now the standard in materials science. However, there have been recent theoretical developments that allow for increased accuracy in the calculations. Here, we present a dataset of calculations for 175k crystalline materials obtained with two functionals: geometry optimizations are performed with PBE for solids (PBEsol) that yields consistently better geometries than the PBE functional, and energies are obtained from PBEsol and from SCAN single-point calculations at the PBEsol geometry. Our results provide an accurate overview of the landscape of stable (and nearly stable) materials, and as such can be used for reliable predictions of novel compounds. They can also be used for training machine learning models, or even for the comparison and benchmark of PBE, PBEsol, and SCAN.

## Background & Summary

The search for new materials remains one of the most important quests but, unfortunately, also one of the great challenges of materials science. Nowadays, data-driven searching strategies have become the most cost-effective methods to tackle this problem, and the fastest way of finding new materials or study their properties are computational high-throughput searches. After years of data accumulation, there are millions of calculations of materials available in open databases[1,2] that are used as an invaluable reservoir to select and filter promising candidates for further experimental synthesis and characterization.

These high-throughput studies in solid-state material science[3–6] have broadened the exploration of the vast chemical space, while plenty of works have successfully found and predicted promising materials for technological applications. However, nearly all high throughput searches rely on the use of density functional theory (DFT) within the Perdew-Burke-Ernzerhof (PBE) approximation to the exchange-correlation functional[7]. This is a well established and reliable approach that earned its place as the standard technique in solid-state research. However, the PBE functional is now over 25 years old, and more recent (and accurate) functional have by now been proposed in the literature. For example, the Armiento-Mattson 2005[8] or the PBE for solids[9] functionals consistently lead to superior geometries[10,11], while the SCAN meta-generalized gradient approximation[12] yields formation energies that are on average better by a factor of two than the PBE[13]. Unfortunately, and in stark contrast with the abundance of PBE data, there are no available comparable large scale datasets calculated with these improved functionals.

There are a number of other databases that use either higher accuracy methods, like $G_0 W_0$, or apply density-functionals different from PBE. For example, we can mention the Computational 2D Materials Database[14,15] that provides a dataset of 4000 2D materials calculated with HSE, $G_0 W_0$, RPA and the

[1]Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, 06120, Halle (Saale), Germany. [2]CFisUC, Department of Physics, University of Coimbra, Rua Larga, 3004-516, Coimbra, Portugal. [3]Institut für Festkörpertheorie und -optik and European Theoretical Spectroscopy Facility, Friedrich-Schiller-Universität Jena, D-07743, Jena, Germany. ✉e-mail: miguel.marques@physik.uni-halle.de

Bethe-Salpeter equation, while JARVIS contains a large number of calculations with vdW corrections using OptB88vdW[16,17] and performed with the modified Becke-Johnson potential[18–20].

In a previous work[21] we combined data from the AFLOW database[1], the Materials Project[2] and from our own group to create a rather complete convex hull of thermodynamic stability at the PBE level. The details of the selection of the dataset can be found in ref. [21]. Specifically, we selected all materials that were calculated with the same functional, pseudopotential, as well as $U$-parameters used in the Materials Project. We removed duplicates, i.e. entries with the same space group, composition and total energy (rounded to the 4th digit). Here we determined the space group using pymatgen with the "symprec" keyword set to 0.1. From the AFLOW database we further removed all prototypes labeled "_DEVIL_PROTOTYPES_" and all other combinations of prototypes and pseudopotentials that are noted as ill-converged in the code of ref. [22]. As AFLOW is still known to contain outliers[22], we also removed them from the calculation of the hull following a strategy similar to the one explained in ref. [22]. We used the total energies of all the remaining structures to calculate the convex hulls applying the corrections from the materials project workflow to the energies.

From this dataset[23] we selected around 175k compounds that were either stable (i.e., on the convex hull), or close to stable (within 100 meV/atom of the hull). These were then reoptimized with PBEsol[9]. Finally, the total energies were reevaluated with SCAN[12] to create highly accurate formation energies and convex hulls.

## Methods

Our starting point was the dataset used in the machine learning study of ref. [21]. This included PBE calculations stemming from the Materials Project database[2], AFLOW[1], and our own calculations. These were then filtered to obtain a homogeneous set in what regards the calculation parameters, leading to a dataset containing more than two million compounds. We then constructed the convex hull of thermodynamic stability and extracted entries that were either on the hull or within 0.1 eV/atom[24,25]. The reason for the choice of this cutoff was twofold. First, its value is still below the average error in the formation energies calculated with PBE[26,27], but is larger than the estimated error in the distances to the convex hull[28,29]. As such, the compounds that were misidentified by the PBE as thermodynamically unstable are likely to be included in the set. Second, there are a number of materials that are metastable, but experimentally accessible (for example for compositions that have more than one polymorph). We can reasonably expect that the cutoff allows for the inclusion of the majority of these cases. We also eliminated materials with unit cells that were too large for our computational resources, leading to a final amount of ~175k compounds.

All calculations were performed using density-functional theory, within the projector augmented wave method (PAW)[30] as implemented in the Vienna *ab initio* simulation package (VASP)[31,32]. We used the PAW set-ups shipped with version 5.4 of vasp that include information on the kinetic energy density of the core electrons. We mostly followed the recommendations of the Materials Project for the choice of the pseudopotentials. The exception was Cs, for which we used an improved pseudopotential generated by the vasp developers, as the stock PAW setup often led to negative densities during the self-consistent cycle, crashing the calculation. All calculations were performed taking into account spin-polarization, and started from a ferromagnetic configuration (as in the large majority of high-throughput studies). This most likely leads to an incorrect spin configuration for antiferromagnetic systems, resulting in an energy higher than the true groud-state. However, it is well known that in most cases magnetic exchange energies are rather small, so the error in the total energy is limited to a few tens of meV/atom. Methfessel-Paxton order one smearing with a width of 0.2 eV was applied in the integration of the Brillouin zone.

The structures were optimized using the PBEsol[9] approximation, the "High" precision keyword of vasp, and a $\Gamma$-centered k-point grid with 2000 k-points per reciprocal atom, until the forces on the atoms were below 5 meV/Å. To detect the structures that required re-optimization we checked that the the absolute value of the forces (in meV/Å) and the individual components of the stress tensor (in meV/Å$^3$), calculated with PBEsol with stricter convergence criteria (520 eV for the wave-function cutoff and 8000 k-points per reciprocal atom), remained smaller than 0.05. If it was not the case, we increased both the energy cutoff and the number of k-points and performed again the geometry optimization. Around 3% of all structures required this further calculation step. We note that as we had already a reasonable starting point, specifically the PBE geometry, the geometry optimization required a relatively small number of steps in most cases.

A final energy evaluation with the SCAN meta-GGA functional was performed with a cutoff of 520 eV, 8000 k-points per reciprocal atom, and including the non-spherical contributions from the gradient corrections inside the PAW spheres. As expected, SCAN calculations were much more unstable than PBEsol, due to the well known numerical instabilities of this functional[33,34], which led to a much lower average convergence rate. In any case, we succeeded in converging nearly all calculations. In total, the calculations presented here required around 10 million CPU hours.

## Data Records

The output files of vasp were collected and processed with the pymatgen library[35]. Each final data record consisted of a *ComputedStructureEntry* that included the chemical composition, the total energy, and the detailed crystal structure of the entry. Two files, containing all entries for each functional, can be freely downloaded from the Materials Cloud repository[36] and can be loaded trivially using the json module of python. For convenience, we also provide a summary of the data in tabulated form, that include the fields listed in Table 1.

In panel a of Fig. 1 we plot the histogram of the number of different chemical elements in our materials. We can clearly see that the dataset is dominated by ternary compounds, followed by binary and quaternary. Relatively few multinary materials with more than 4 different chemical elements are present. We can understand this distribution by considering that the number of permutations increases very rapidly with the number of chemical elements, explaining, e.g., why we have many more binaries than elementary substances or ternary

| | |
|---|---|
| Composition | Chemical composition of each material. |
| Number of sites | Number of atoms in the unit cell. |
| $E_{PBE}$ | Total energy per atom calculated with PBE (extracted from the primary dataset). |
| $E_{PBEsol}$ | Total energy per atom calculated with PBEsol. |
| $E_{SCAN}$ | Total energy per atom calculated with SCAN. |
| $V_{PBE}$ | Volume per atom of the PBE unit cell (extracted from the primary dataset). |
| $V_{PBEsol}$ | Volume per atom of the PBEsol unit cell. |
| $Gap_{PBEsol}$ | Band gap calculated with PBEsol. |
| $Gap_{SCAN}$ | Band gap calculated with SCAN. |
| $M_{PBEsol}$ | Total magnetic moment per atom calculated with PBEsol. |
| $M_{SCAN}$ | Total magnetic moment per atom calculated with SCAN. |
| $S_{SCAN}$ | The diagonal elements of the stress tensor calculated with SCAN. |

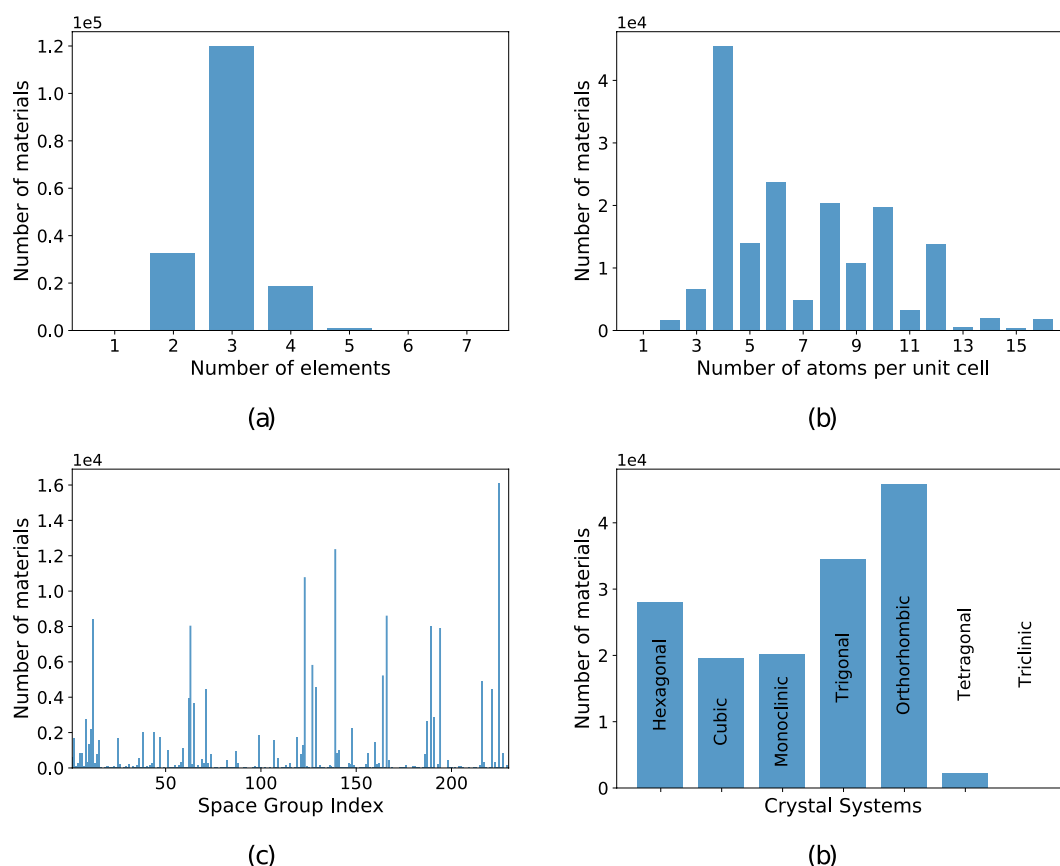**Table 1.** Fields included in the summary of the data in tabulated form.



**Fig. 1** Distribution of (**a**) number of different chemical elements per unit cell, (**b**) number of atoms per unit cell, (**c**) index of space groups, and (**d**) crystal systems for all the materials in our dataset.

than binary compounds. However, we have to keep in mind that the complexity of the unit cells also increases and that most of the systematic high-throughput DFT studies were performed for ternary systems[37–42]. This can also be seen in panel (b) of Fig. 1 where we show a histogram of the number of atoms in the primitive unit cell. The distribution is dominated by a peak centered around five atoms per unit cell, arising from the materials stemming from the high-throughput studies of AFLOW[1] and from our group[3,37,38]. Obviously, we do not expect that the true distribution of all possible thermodynamically stable and metastable materials follows the behavior depicted in Fig. 1. Finally, in panels (c) and (d) of the same figure, we plot an histogram of the number of materials as a function of the space group index and of the crystal system. We see that the most represented systems are the trigonal and orthorhombic, while the tetragonal and triclinic are the least represented.

In Fig. 2 we show the distribution of chemical elements for the calculated structures. We considered all elements except noble gases up to bismuth as well as most lanthanides, and actinides up to plutonium. We can observe some obvious trends. Not surprisingly, the most common element is oxygen due to the abundance of
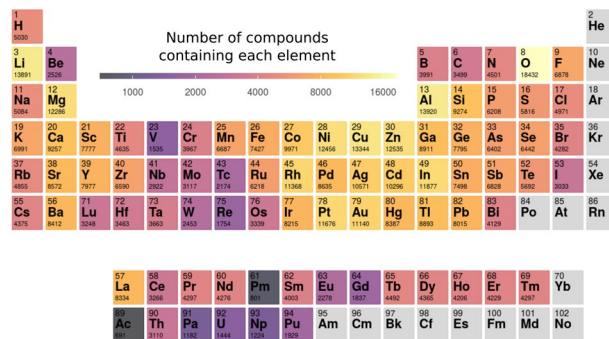
**Fig. 2** Periodic table depicting the chemical elements present in our dataset. The number beneath the chemical symbol is the number of materials present in the dataset that contain the given element.

oxides in our planet and their stability in our oxygen-rich atmosphere. The remaining chalcogens (S, Se, Te) are all equally represented, which is probably an indication of their chemical similarity. For the halogen family we see a decreasing number of materials following the decrease of the electronegativity, with iodine compounds being half as abundant as fluorides. For the pnictogens we witness exactly the opposite behavior, with much fewer nitrides than compounds containing antimony. This can be understood from the high chemical stability of the $N_2$ molecule and the rather high nominal oxidation state of nitrogen ($-3$) that hinders the synthesis of nitrides. Finally carbon-containing materials are rather scarce due to the absence of organic compounds in our dataset.

From the metals, the two with the highest number of compounds are aluminum and lithium. The former is within a cluster of highly represented chemical elements (such as nickel, copper, zinc, or indium). The exception in this region of the periodic table is gallium, in spite of its importance in many semiconductors used in electronics and optoelectronics. Lithium compounds, on the other hand, are much more common in our dataset than materials containing any other alkali element, which might be explained by the popularity of studies in lithium compounds in view of their application in battery technologies. Interestingly, in contrast to the other alkali earth elements, beryllium appears in relatively few compounds, probably due to its high toxicity. Another region that exhibits relatively few compounds is the one centered in vanadium and that includes rhenium, hafnium, niobium, etc. We can observe a continuity in the values in this region that might indicate that these chemical elements, often exhibiting the very high oxidation numbers of $+4$, $+5$ or $+6$, have more difficulty producing stable compounds than other metals with lower oxidation numbers. Finally, the least represented groups are unsurprisingly the lanthanides and the actinides, showing how little we know the chemistry of these chemical elements that are essential in a multitude of technologies, such as in hard magnets or in the storage of nuclear waste.

It is also interesting to look at the structural diversity in our dataset. With that objective, we used pymatgen to divide our structures into groups according to structural similarity. In total, our data turned out to contain 24706 prototypes of 4557 different generic compositions. However, and as expected, the distribution of compounds among these prototypes was rather unbalanced. For example, 15399 prototypes only appeared once in the dataset and 2412 only twice. On the other hand, most materials belong to just a few prototypes. The most common was by far the Heusler family of compounds, with almost 12 000 compositions, followed by the double perovskite family with more than 5000 elements. Of course, these numbers reflect not only the chemical stability and size of each one of the families, but also the interest of the community for these compounds.

## Technical Validation

In the left panel of Fig. 3 we plot the distribution of the volumes per atom with the PBE (obtained from the primary data) and PBEsol. Both curves look similar, with a peak at around 15 Å³ and skewed towards larger volumes. We can also observe that the PBE data is shifted toward larger volumes with respect to PBEsol. This is due to the well known underbinding of the PBE[43] that leads to lattice constants that are larger than their experimental values by 2–3% (and consequently volumes that are overestimated by ~10%). This underbinding is almost totally corrected by PBEsol, yielding smaller volumes in much better agreement with experiment. In the right panel of the same figure we plot a histogram of the three diagonal components of the stress tensor calculated with SCAN at the PBEsol geometry. We can see that the calculations yield rather small stresses, showing that the structures are close to mechanical equilibrium. This is expected because SCAN, like PBEsol, yields high quality geometries in good agreement with experiment[10,44]. This also validates our pragmatic approach of using a single point SCAN calculation at the PBEsol geometries.

In Fig. 4 we show the distribution of (indirect) band gaps calculated with both PBEsol and SCAN. As expected, the curves decay monotonically with the value of the gap, and exhibit a fat tail that extends beyond 10 eV. It is known that PBEsol, as well as PBE, strongly underestimate the value of the band gaps by essentially a factor of two, leading to mean absolute percentage errors bordering the 50%[45,46] with respect to experiment. This is to some extent corrected by SCAN, that increases consistently the band gaps leading to a mean absolute percentage error of around 40%[45,46]. This increase is evident from Fig. 4, where the SCAN band gap distribution is shifted to the right with respect to PBEsol.
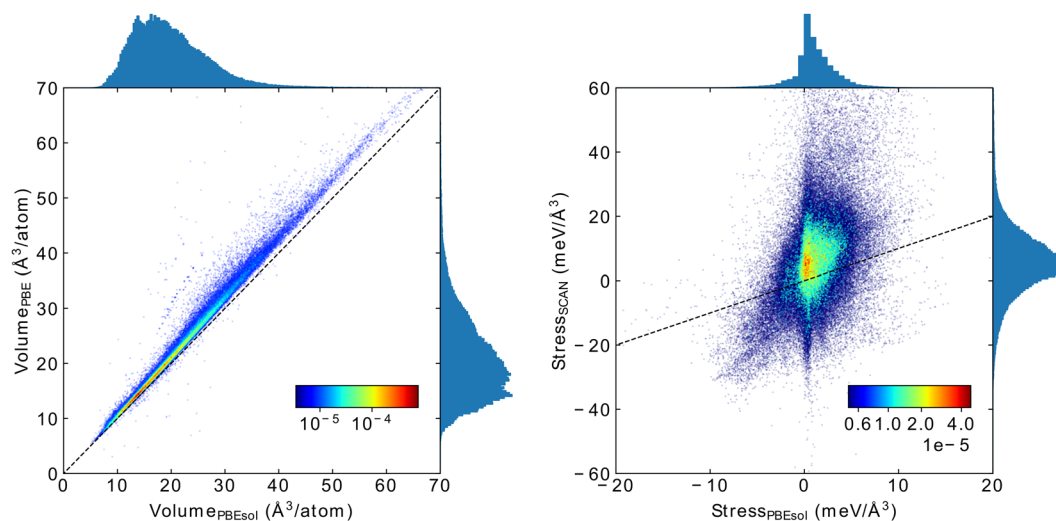
**Fig. 3** Left: Distribution and scatter plots of volumes per atom calculated with PBE (from the primary data) and PBEsol functionals. The width of the bins is 0.35 Å$^3$/atom. Right: distribution and scatter plots of the diagonal components of the stress tensor calculated with PBEsol and SCAN at PBEsol geometries. The width of the bins is 0.6 meV/Å$^3$.
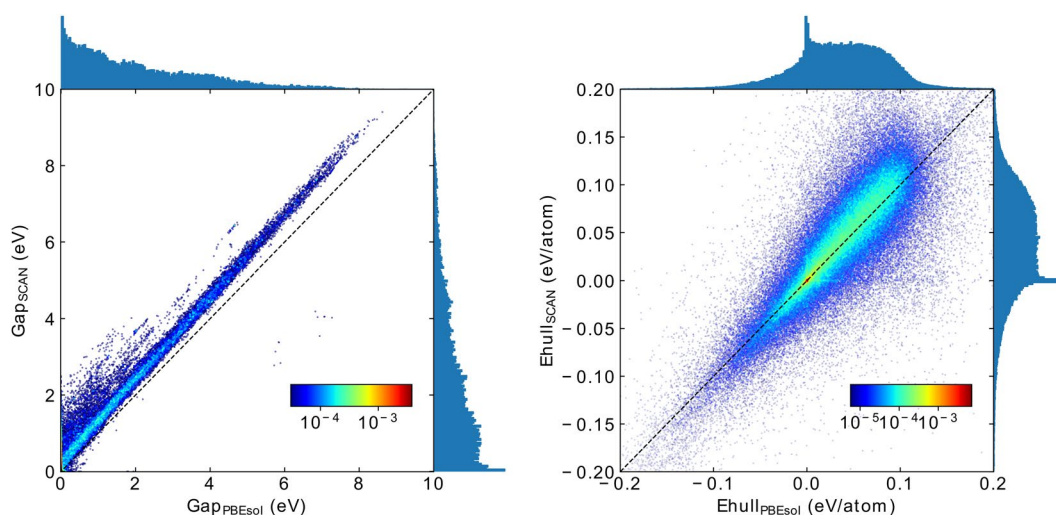


**Fig. 4** Left: Distribution and scatter plots of the (in)direct band gaps calculated with PBEsol and SCAN at the PBEsol geometry. The width of the bins is 0.1 eV. Right: Distribution and scatter plots of the distances to the convex hull calculated with PBEsol and SCAN at PBEsol geometries. The corresponding hulls contain 40246 and 38692 materials. The width of the bins is 2 meV/atom.

Finally, we performed an analysis of the convex hulls obtained with PBE, PBEsol, and SCAN. The hulls contained 36985, 40246 and 38692 materials respectively. These differences are expected, and mostly stem from relatively small changes in the formation energy for compounds that were close to the hull. The distributions of the distances to the convex hull of thermodynamic stability are plotted in the right panel of Fig. 4. For this plot, we always removed the compound under study from the hull, allowing therefore for negative distances. Although this allows for a better interpretation of the results, we should remember that technically speaking all compounds with negative distances should be placed strictly at zero. Both distributions grow fast for negative distances to the hull, having a marked peak at zero. The main contributors to this behavior are the experimental compounds. The number of materials with positive distance to the hull is relatively constant, until reaching our cutoff at $\pm 0.2$ eV/atom. This cutoff is obviously smeared for the PBEsol and SCAN calculations.

As an illustration, we plot in Fig. 5 the ternary phase diagrams of Li–Al–Cu and Mg–Sc–Zn. We see that the three functionals agree to a large extent on which are the stable materials. However, some differences are also clear. For example, MgZn is stable with the PBE functional but not with PBEsol or SCAN, or LiAl3 is unstable with SCAN but stable with the other two functionals. In any case, compounds that are stable with one functional do appear stable, or at least metastable with the other functionals. Of course, SCAN is the most accurate of the

**Fig. 5** Ternary phase diagrams of the Li–Al–Cu (upper panel) and Mg–Sc–Zn (lower panel) systems, calculated with the PBE (left), PBEsol (middle) and SCAN (right). The blue points indicate compositions on the convex hull, while red points denote materials that are within 50 meV/atom from the hull.

three functionals in what concerns formation energies and distances to the convex hull, so the SCAN diagrams should on average have the highest accuracy.

## Usage Notes

The data can be downloaded from the Materials Cloud repository[36]. The energies, compositions, and structures for each material are formatted as *ComputedStructureEntries* and stored as compressed json files. They can therefore be trivially loaded in python and analyzed with pymatgen. We note that we used version v2019.10.2 of pymatgen, but the data should be compatible with other versions.

## Code availability

All data can be easily processed with publicly available tools such as json and pymatgen[35]. An example usage is provided with the data. The dataset was generated with VASP, the bash and python scripts to generate input files or manage the output files can be downloaded from github repository: https://github.com/hyllios/utils/tree/main/ht_pd_scan.

## References

1. Curtarolo, S. *et al*. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226, https://doi.org/10.1016/j.commatsci.2012.02.005 (2012).
2. Jain, A. *et al*. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002, https://doi.org/10.1063/1.4812323 (2013).
3. Körbel, S., Marques, M. A. L. & Botti, S. Stable hybrid organic–inorganic halide perovskites for photovoltaics from ab initio high-throughput calculations. *J. Mater. Chem. A* **6**, 6463–6475, https://doi.org/10.1039/c7ta08992a (2018).
4. Graužinytė, M., Botti, S., Marques, M. A. L., Goedecker, S. & Flores-Livas, J. A. Computational acceleration of prospective dopant discovery in cuprous iodide. *Phys. Chem. Chem. Phys.* **21**, 18839–18849, https://doi.org/10.1039/c9cp02711d (2019).
5. Flores-Livas, J. A., Sarmiento-Pérez, R., Botti, S., Goedecker, S. & Marques, M. A. L. Rare-earth magnetic nitride perovskites. *JPhys Mater.* **2**, 025003, https://doi.org/10.1088/2515-7639/ab083e (2019).
6. Wang, H.-C., Pistor, P., Marques, M. A. L. & Botti, S. Double perovskites as p-type conducting transparent semiconductors: a high-throughput search. *J. Mater. Chem. A* **7**, 14705–14711, https://doi.org/10.1039/c9ta01456j (2019).
7. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865, https://doi.org/10.1103/PhysRevLett.77.3865 (1996).
8. Armiento, R. & Mattsson, A. E. Functional designed to include surface effects in self-consistent density functional theory. *Phys. Rev. B* **72**, 085108, https://doi.org/10.1103/PhysRevB.72.085108 (2005).
9. Perdew, J. P. *et al*. Restoring the density-gradient expansion for exchange in solids and surfaces. *Phys. Rev. Lett.* **100**, 136406, https://doi.org/10.1103/PhysRevLett.100.136406 (2008).
10. Csonka, G. I. *et al*. Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B* **79**, 155107, https://doi.org/10.1103/PhysRevB.79.155107 (2009).
11. Tran, F., Stelzl, J. & Blaha, P. 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**, 204120, https://doi.org/10.1063/1.4948636 (2016).

12. Sun, J., Ruzsinszky, A. & Perdew, J. P. Strongly constrained and appropriately normed semilocal density functional. *Phys. Rev. Lett.* **115**, 036402, https://doi.org/10.1103/PhysRevLett.115.036402 (2015).

13. Zhang, Y. *et al.* Efficient first-principles prediction of solid stability: Towards chemical accuracy. *Npj Comput. Mater.* **4**, https://doi.org/10.1038/s41524-018-0065-z (2018).

14. Haastrup, S. *et al.* The computational 2D materials database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials* **5**, 042002, https://doi.org/10.1088/2053-1583/aacfc1 (2018).

15. Gjerding, M. N. *et al.* Recent progress of the computational 2d materials database (c2db). *2D Materials* **8**, 044002, https://doi.org/10.1088/2053-1583/ac1059 (2021).

16. Thonhauser, T. *et al.* Van der waals density functional: Self-consistent potential and the nature of the van der waals bond. *Phys. Rev. B* **76**, 125112, https://doi.org/10.1103/PhysRevB.76.125112 (2007).

17. Klimeš, J. C. V., Bowler, D. R. & Michaelides, A. Van der waals density functionals applied to solids. *Phys. Rev. B* **83**, 195131, https://doi.org/10.1103/PhysRevB.83.195131 (2011).

18. Choudhary, K. *et al.* The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials* **6**, https://doi.org/10.1038/s41524-020-00440-1 (2020).

19. Choudhary, K. *et al.* Computational screening of high-performance optoelectronic materials using OptB88vdw and TB-mBJ formalisms. *Scientific Data* **5**, https://doi.org/10.1038/sdata.2018.82 (2018).

20. Choudhary, K., Cheon, G., Reed, E. & Tavazza, F. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. *Phys. Rev. B* **98**, 014107, https://doi.org/10.1103/PhysRevB.98.014107 (2018).

21. Schmidt, J., Pettersson, L., Verdozzi, C., Botti, S. & Marques, M. A. L. Crystal graph attention networks for the prediction of stable materials. *Science Advances* **7**, eabi7948, https://doi.org/10.1126/sciadv.abi7948 (2021).

22. Oses, C. *et al.* Aflow-chull: Cloud-oriented platform for autonomous phase stability analysis. *J. Chem. Inf. Model.* **58**(12), 2477–2490, https://doi.org/10.1021/acs.jcim.8b00393 (2018).

23. Schmidt, J. & Pettersson, L. Crystal-graph attention networks for the prediction of stable materials. *Materials Cloud* https://archive.materialscloud.org/record/2021.222 (2021).

24. Emery, A. A. & Wolverton, C. High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO3 perovskites. *Sci. Data* **4**, https://doi.org/10.1038/sdata.2017.153 (2017).

25. Sun, W. *et al.* The thermodynamic scale of inorganic crystalline metastability. *Sci. Adv.* **2**, e1600225, https://doi.org/10.1126/sciadv.1600225 (2016).

26. Sarmiento-Pérez, R., Botti, S. & Marques, M. A. L. Optimized exchange and correlation semilocal functional for the calculation of energies of formation. *J. Chem. Theory Comput.* **11**, 3844–3850, https://doi.org/10.1021/acs.jctc.5b00529 (2015).

27. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies. *Phys. Rev. B* **85**, 115104, https://doi.org/10.1103/PhysRevB.85.115104 (2012).

28. Hautier, G., Ong, S. P., Jain, A., Moore, C. J. & Ceder, G. Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability. *Phys. Rev. B* **85**, 155208, https://doi.org/10.1103/PhysRevB.85.155208 (2012).

29. Bartel, C. J., Weimer, A. W., Lany, S., Musgrave, C. B. & Holder, A. M. The role of decomposition reactions in assessing first-principles predictions of solid stability. *Npj Comput. Mater.* **5**, 1–9, https://doi.org/10.1038/s41524-018-0143-2 (2019).

30. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979, https://doi.org/10.1103/PhysRevB.50.17953 (1994).

31. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15–50, https://doi.org/10.1016/0927-0256(96)00008-0 (1996).

32. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186, https://doi.org/10.1103/PhysRevB.54.11169 (1996).

33. Bartók, A. P. & Yates, J. R. Regularized SCAN functional. *J. Chem. Phys.* **150**, 161101, https://doi.org/10.1063/1.5094646 (2019).

34. Furness, J. W., Kaplan, A. D., Ning, J., Perdew, J. P. & Sun, J. Accurate and numerically efficient r2SCAN meta-generalized gradient approximation. *J. Phys. Chem. Lett.* **11**, 8208–8215, https://doi.org/10.1021/acs.jpclett.0c02405 (2020).

35. Ong, S. P. *et al.* Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319, https://doi.org/10.1016/j.commatsci.2012.10.028 (2013).

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

37. Schmidt, J. *et al.* Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103, https://doi.org/10.1021/acs.chemmater.7b00156 (2017).

38. Schmidt, J., Chen, L., Botti, S. & Marques, M. A. L. Predicting the stability of ternary intermetallics with density functional theory and machine learning. *J. Chem. Phys.* **148**, 241728, https://doi.org/10.1063/1.5020223 (2018).

39. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD). *JOM* **65**, 1501–1509, https://doi.org/10.1007/s11837-013-0755-4 (2013).

40. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**, 3762–3767, https://doi.org/10.1021/cm100795d (2010).

41. Shi, J. *et al.* High-throughput search of ternary chalcogenides for p-type transparent electrodes. *Sci. Rep.* **7**, https://doi.org/10.1038/srep43179 (2017).

42. Oliynyk, A. O. *et al.* High-throughput machine-learning-driven synthesis of full-heusler compounds. *Chem. Mater.* **28**, 7324–7331, https://doi.org/10.1021/acs.chemmater.6b02724 (2016).

43. Haas, P., Tran, F. & Blaha, P. Calculation of the lattice constant of solids with semilocal functionals. *Phys. Rev. B* **79**, https://doi.org/10.1103/physrevb.79.085104 (2009).

44. Isaacs, E. B. & Wolverton, C. Performance of the strongly constrained and appropriately normed density functional for solid-state materials. *Phys. Rev. Mater.* **2**, 063801, https://doi.org/10.1103/PhysRevMaterials.2.063801 (2018).

45. Borlido, P. *et al.* Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *Npj Comput. Mater.* **6**, https://doi.org/10.1038/s41524-020-00360-0 (2020).

46. Borlido, P. *et al.* Large-scale benchmark of exchange-correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **15**, 5069–5079, https://doi.org/10.1021/acs.jctc.9b00322 (2019).

## Acknowledgements

## Author contributions

J.S., H.W. and M.A.L.M. performed the calculations. All authors contributed to the analysis of the results and to the writing the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to M.A.L.M.

**Reprints and permissions information** is available at www.nature.com/reprints.

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