npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Predicting stable crystalline compounds using chemical similarity

Hai-Chen Wang[1], Silvana Botti[2] and Miguel A. L. Marques[1]✉

We propose an efficient high-throughput scheme for the discovery of stable crystalline phases. Our approach is based on the transmutation of known compounds, through the substitution of atoms in the crystal structure with chemically similar ones. The concept of similarity is defined quantitatively using a measure of chemical replaceability, extracted by data-mining experimental databases. In this way we build 189,981 possible crystal phases, including 18,479 that are on the convex hull of stability. The resulting success rate of 9.72% is at least one order of magnitude better than the usual success rate of systematic high-throughput calculations for a specific family of materials, and comparable with speed-up factors of machine learning filtering procedures. As a characterization of the set of 18,479 stable compounds, we calculate their electronic band gaps, magnetic moments, and hardness. Our approach, that can be used as a filter on top of any high-throughput scheme, enables us to efficiently extract stable compounds from tremendously large initial sets, without any initial assumption on their crystal structures or chemical compositions.

npj Computational Materials (2021)7:12 ; https://doi.org/10.1038/s41524-020-00481-6

INTRODUCTION

The quest for new materials is one of the most important endeavors of materials science[1,2]. The discovery of materials with tailored properties hold the promise of improving existing technologies, but also of enabling new disruptive applications[3]. Unfortunately, there exist many examples of technologies that remain in the realm of science fiction due to the unavailability of adequate materials[4,5]. This may happen because known compounds are toxic, rare, or too expensive for industrial, large scale use, or simply because no material is known with good enough properties[6–8].

It is clear that the number of imaginable materials is extremely large, as it derives from the combinatorial problem of arranging chemical elements of the periodic table in all possible stoichiometries and dynamically stable crystal structures[9]. This number is, however, reduced as most combinations are not prone to experimental synthesis[2]. There are several reasons for this: the crystal structure may describe a high-energy polymorph that can not be stabilized, the stoichiometry itself may be highly unstable to decomposition to other compounds, or it may simply be that there is no easy thermodynamically favored reaction path for experimental synthesis. In spite of these problems, there remains a very large number of experimentally reachable materials, of which we know only a small fraction[10].

For the past decades, we have witnessed spectacular advances in computational materials science. One of the main reasons for this was the progression of density functional theory (DFT)[11,12] that, thanks to its excellent accuracy combined with remarkable computational efficiency, has become the workhorse method for the theoretical study of materials[13]. Favored by the advent of faster supercomputers and better software, DFT opened the way for extensive numerical studies of large datasets of compounds[14]. These so-called high-throughput studies[15], whose results are conveniently stored in online databases, have greatly extended our knowledge of materials and have already lead to the discovery of a variety of compounds with improved properties[15–18].

There are several strategies that can be used for the theoretical search of materials[18,19]. One of the most prominent approaches for inorganic solids is "component prediction", following the definition of ref. [19], meaning that one scans the composition space of a prototype structure searching for stable materials, instead of scanning the space of possible crystal structures for a given composition[19–21].

In this context, we use the word "stable” to denote thermodynamical stability, i.e., compounds that do not transform or decompose (even in infinite time) to other different phases or stoichiometrically compatible compounds[9]. It is true that metastable materials, like diamond, are also synthesizable and advances in chemistry have made them more accessible[22,23]. Nevertheless, thermodynamically stable compounds are in general easier to produce and handle. The usual criterion for thermodynamic stability is based on the energetic distance to the convex hull[24]: the energy distance of a compound to the convex hull is hence a measure of its instability.

Using high-throughput approaches, the whole periodic table has already been scanned for a series of prototypes of relevant crystal structures. The most extensive studies of this kind can be found in the aflowlib database[25] that, at present, includes more than 2 million compounds. Unfortunately, this number is dwarfed by the total number of possibilities. Just for ternary intermetallics, there are 1391 structure-types known experimentally[26] and there are ~500,000 possibilities of combining three metallic elements for each of these prototypes. Moreover, ternary structures can be rather complex: the average number of atoms in the unit cell turns out to be 14, but the majority of intermetallic ternary prototypes is considerably larger[26]. The situation is obviously even worse for quaternary or multinary systems. Considering that a DFT calculation scales with the cube of the number of atoms in the unit cell,

[1]Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, 06120 Halle (Saale), Germany. [2]Institut für Festkörpertheorie und -Optik, Friedrich-Schiller-Universität Jena and European Theoretical Spectroscopy Facility, Max-Wien-Platz 1, 07743 Jena, Germany. ✉email: miguel.marques@physik.uni-halle.de

we are quickly led to conclude that an exhaustive search of the composition space will be out of reach for the foreseeable future.

To mitigate the combinatorial curse, chemical constrains have been successfully applied to filter out compounds that are unlikely to be formed[27]. Alternatively, machine learning can be used to predict compounds and their properties[14,28–31]. In view of the scarcity of experimental data, the machine is usually trained on DFT calculations and then used to predict which compositions and/or crystal structures are more likely to be stable[14,19,28,29]. Already in 2010, in the seminal work by Hautier et al.[32], machine learning was used to predict the probability that a chemical substitution of an existing compound can give another stable compound. Predictions are then validated *a posteriori* performing DFT calculations of the candidate systems.

In this article, we propose an approach to scan efficiently the space of all possible stable materials that relies on data mining rather than empirical rules or chemical intuition, inspired from ref. [32]. We borrow the idea of component prediction[19–21] and combine it with the concept of chemical similarity. This means that the compositions to be tested are selected using a measure of the likelihood that a chemical element A can be replaced by another B in a given structure. Such a scale of similarity was obtained by statistical analysis through data mining in ref. [33]. To some extent, the concept of similarity can be intuitively understood from the graphical representation of the periodic table. Elements that are neighbors in the periodic table are known to be similar chemically, a fact has been used by chemists to create materials for more than 100 years. However, statistical analysis goes beyond pure chemical intuition and can identify unexpected correspondences.

Any approach based on chemical similarity can be applied immediately to any crystal structure, and even to systems of reduced dimensionality, such as two-dimensional materials and nano-structures.

RESULTS

Thermodynamic stability

The number of substituted materials in each iteration that were not in the database, and hence were calculated is shown in Table 1. Our initial set was composed by elemental, binary, ternary, quaternary, and quintenary compounds. The first iteration is strongly biased by the distribution of materials in the database, which is mainly composed of binary and ternary compounds.

Before discussing in detail the results, we can better motivate the choice to set the threshold value of the element replaceabilty at 5%. We verified that higher values of the threshold would lead to a higher percentage of stable materials. In particular, our results indicate that a threshold of around 20% would maximize the fraction of stable compounds found in each iteration. However, the total number of stable compounds would be reduced by a factor of three. We believe therefore that setting the threshold value to around 5% is a more convenient compromise.

There are a total of 713 different prototypes in the first generation, and the most common one is the cubic full-Heusler compound, with a total of 10,653 systems. These are very simple ternary cubic compounds (from the crystallographic point of view) with composition ABC_2, and that can be stable for a large variety of elements in the periodic table. This family has already been subject to extensive and systematic studies using either high-throughput or machine learning techniques, and the optimized crystal structures for most compounds can be found, e.g., in the Aflowlib database[25]. In the second generation, Heusler compounds remain the most common prototype, but with only 4238 systems. The situation changes in the third generation, where the most common prototype becomes the hexagonal ZrNiAl–Fe_2P structure, with 5009 compounds.

It is interesting to analyze the distance to the convex hull (E_{hull}) of stability for all 189,981 materials. A histogram with this information can be found in Fig. 1. Note that we plot E_{hull} with respect to the hull composed of compounds in the materials database solely. This means that stable structures not included in the database will appear with negative E_{hull}. Of course, in this case, the hull has to be redefined to include these compounds. This will be further discussed in the following.

The first impression we get from the figure is that the distribution of E_{hull} is very different from a skewed Gaussian we know for DFT calculations of families of materials (e.g., perovskites[30] or tI10 materials[31]). In fact, we believe that the distribution displayed in Fig. 1 is a demonstration of the validity of our approach. In comparison with the distributions shown in refs. [30,31], obtained by performing systematic substitutions, we observe an enhanced percentage of materials with a negative distance to the hull, while the histogram decays rapidly for positive distances. The large peak at zero is due to substitutions leading to materials already present in the database. We did check whether the transmuted material is already in the database, i.e., if an entry with the same composition and space group exists before running the calculation. However, often the geometry optimization procedure relaxes structures into other space groups (usually to more symmetric ones), and these final structures can sometimes be found in the database.

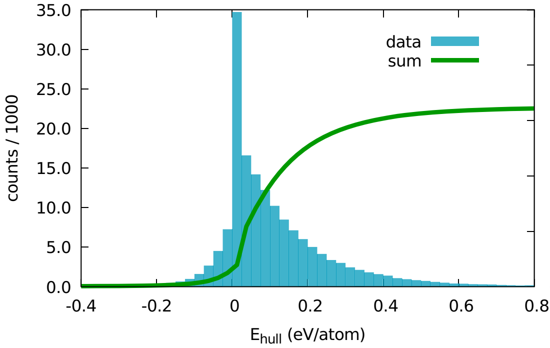

Fig. 1 Thermodynamic stability. Distribution of the distances to the convex hull of all 189,981 compounds.

Table 1. The number of new structures (not in the database) at each iteration.

Loop	Structure	Elementary	Binary	Ternary	Quaternary	Quinternary
1	59,853	370 (0.62%)	14,309 (23.9%)	40,455 (67.6%)	4432 (7.4%)	287 (0.48%)
2	50,917	44 (0.09%)	5708 (11.2%)	38,959 (76.5%)	6077 (11.9%)	129 (0.25%)
3	79,211	45 (0.06%)	6554 (8.3%)	60,136 (75.9%)	12,216 (15.4%)	260 (0.33%)
Total	189,981	459 (0.24%)	26571 (14.0%)	139,550 (73.5%)	22,725 (12.0%)	676 (0.36%)

Compounds for which the calculations failed to converge were excluded.

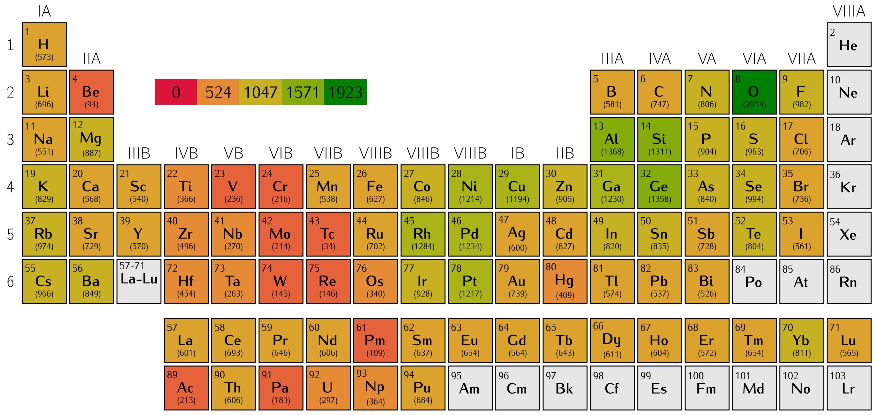

Fig. 2 Distribution of stable materials. The number of stable materials containing a given element through the periodic table.

There are a total of 31,602 structures with a negative distance to the convex hull, but not all of these can be counted as stable structures. Firstly, the procedure we follow could find more than one structures with negative E_{hull} with the same composition. And secondly, we have to redefine the convex hull including all our structures. After taking these two points into consideration, we found a total of 18,479 systems on the redefined convex hull. The structures of these materials are available in our website (see Section "Data Availability"). We crosschecked our list against the Aflowlib database[25], and found that only 417 out of 18,479 (2.3%) stable structures are overlapping with entries of this database. Thus, almost all stable compounds reported in the present work are not included in materials databases.

We have to stress that our calculations are approximate (after all, we are using DFT with the PBE approximation to the exchange-correlation functional), and that we are working at zero temperature, neglecting entropy effects. Systematic analysis reported that the error in DFT estimated stabilities are around several tens meV per atom, e.g. 24 meV per atom[34], and 70 meV per atom[35]. Therefore, one can still expect that a large majority of these 18,479 structures may indeed be stable thermodynamically, and are therefore promising candidates for experimental synthesis.

In this work we decided not to take into account all systems that are "technically" unstable (having positive E_{hull}). In our opinion those structures that have a small positive distance to the theoretical convex hull should however not be completely discarded for two reasons: (i) Some might actually be stable, and only appear above the hull due to the Perdew–Burke–Ernzerhof (PBE) approximation; (ii) Some might be stabilized by temperature, pressure, defects, etc. and thus could be experimentally synthesizable. Nevertheless, due to the large number of structures, we decided, for the time being, to concentrate only on the theoretically stable materials and leave the rest for future investigations.

Comparing the number of stable structures (18,479) with the total amount of systems tried (189,981), we find a success rate of 9.72%. This result is encouraging if we compare it with the success rates of systematic high-throughput and machine learning studies. With a threshold set at 25 meV above the convex hull, Sarmiento-Perez et al.[36] have a success rate of 1%, while Körbel et al. in ref. [37] consider a much larger set of compositions and achieve only 0.25% unreported stable compounds. We should also consider that the success rate of a random search is already biased by restricting calculations to a specific family of compounds. In fact, one usually selects a family of systems that looks intuitively promising to start a materials search. In ref. [30], the success rate of systematic calculations of the whole dataset of around 250,000 perovskites is 0.25%, while the proposed machine learning procedure allows to increase the rate by a factor of 4–5.

Indeed, by combining chemistry intuition with a high-throughput approach, our method provides a remarkably efficient overview of large portions of the phase space of stable compounds, at a strongly reduced computational effort. Furthermore, we should not forget that most of the "unstable" transmuted compounds are rather close to the hull, and might therefore be interesting for further research.

To further characterize our set of stable systems, we plot, in Fig. 2, the number of materials that contain one specific chemical element. We see that most stable materials include oxygen. One reason is probably the large number of oxides in our starting set, although other elements are also present in large numbers. We would also like to emphasize the abundance of predicted materials with lanthanide and actinide atoms. These elements are often overlooked in systematic studies, but of great importance in many areas of science. For example, they are often components of permanent magnets[38], or are relevant to understand which materials are formed upon nuclear decay of radioactive waste[39]. In our work, we found 8970 and 2437 stable compounds including lanthanides and actinides, respectively, and the corresponding success rates were 11.6% and 12.2%, respectively. If we exclude entirely these chemical elements, we have 96,543 transmuted structures and a total of 7421 stable compounds. This gives a success rate of 7.7% for compounds that do not contain either lanthanides or actinides. Thus, replacements involving lanthanides and actinides are more likely to yield stable compounds, but 7.7% is still a rather high success rate. In contrast, we note the relatively small number of stable materials containing Be, and transition metals of the groups IVB–VIIB. These elements seem therefore to be harder to combine and form stable compounds.

Now we turn to how the distribution of stable structures and how the success rate changes across the periodic table. The number of stable (N_{new}) and initial structures (N_{ini}) that contain a certain chemical element are showed in Supplementary Fig. 1. We also show in that figure the success rates for substitutions that involved that element.

It can be seen that the distribution of compounds follows to some extent the distribution of the initial structures. For example, there are many oxides in both sets, and N_{new} is approximately proportional to N_{ini} for most 3d transition metals. However, there are several exceptions, e.g. for Al, Si, K, Ga, As, Rb, Cs, lanthanides, and actinides. In some of these cases, there are many more compounds than expected. In contrast, for Mo and W, there are much fewer than expected. This shows that the distribution is not completely biased by the initial database. Furthermore, there is some variation of the success rate through the periodic table, but most elements have a success rate around 10%. However, there are indeed some elements that yield very high success rates, especially some lanthanides or actinides like Pm or Pa.

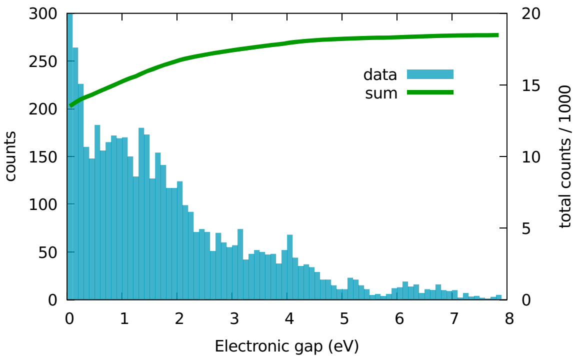

Fig. 3 **Band gaps.** Histogram of the electronic band-gap for all new stable compounds.

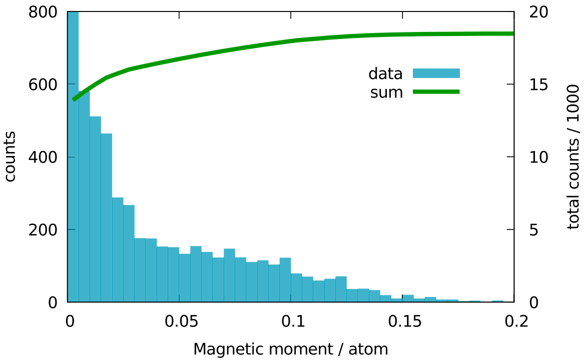

Fig. 5 **Total magnetization.** Histogram of the total magnetization per unit volume (in $\mu_B \cdot \text{Å}^{-3}$) for all new stable compounds.

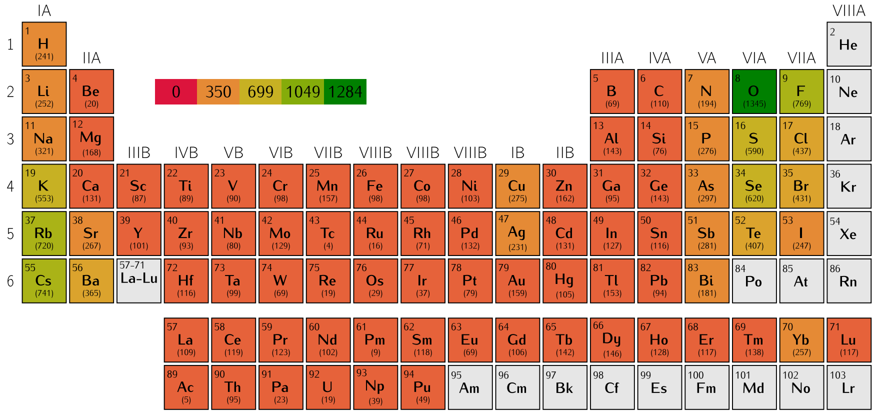

Fig. 4 **Distribution of semiconductors and insulators.** The number of stable semiconductors and insulators containing the given chemical element of the periodic table.

Band gap

The electronic band-gap is certainly one of the most important properties of materials, and it can be used to determine the suitability of a given compound for opto-electronic applications. We plot a histogram of the electronic (indirect) band-gap in Fig. 3 for our stable materials. These were calculated with the PBE approximation to the exchange-correlation functional and are, therefore, underestimated by around 45% on average[40]. We find a total of 4840 systems with a gap larger than 0.1 eV, which is 26.1% of the total number of our stable systems. We should also expect a number of false negatives of around 5–10%, i.e., around 250–500 systems are likely misidentified as metals due to the PBE approximation.

Not surprisingly, the histogram decays with a fat tail as a function of the band-gap. We also show the distribution of semiconductors and insulators through the periodic table in Fig. 4. The most common non-metallic elements in the list of stable semiconductors and insulators is O and F, followed by halogens and other chalcogens. As expected from the electronegativity scale[41], the largest gaps are obtained for fluorides, followed by oxides and chlorides. There are fewer, and still thousands, systems with narrower gaps that include pnictogens and hydrogen. For metallic elements, the most common one found in semiconductors and insulators are the heavy alkali metals Cs, Rb, and K. In all these systems, the largest PBE gap we found was around 7.8 eV for a series of tetragonal ternary fluorides, namely $LiLnF_4$, where Ln is a lanthanide (Tm, Dy, Ho, Tb, Er, Sm, Nd, Pr in order of decreasing band gap).

Magnetic properties

Another property we analyzed is the magnetic moment. In Fig. 5 we plot a histogram of the total magnetization per unit volume (in $\mu_B \cdot \text{Å}^{-3}$) for all our compounds in the convex hull. Before analyzing the results, we would like to stress that each calculations started from an initial ferromagnetic configuration of the spin moments, as common in other high-throughput studies[15,37]. Thus we very likely obtain ferromagnetic states for most magnetic compounds after optimization. However, the correct identification of the ferromagnetic, antiferromagnetic, or ferrimagnetic ground states is crucial for understand the spin interactions in each system. Unfortunately, this would require accurate energy calculations for large supercells, drastically increasing the computational effort. Therefore, in present work we adopt the usual setup of high-throughput studies, and leave the precise identification of the correct ground-states magnetic phases for future research. In any case, from the energetic point of view, this problem is harmless, because the differences of total energy between different magnetic phases are often of the order of the meV per atom[42], while the stability of the composition is evaluated on an energy scale that is one or two orders of magnitude larger.

As expected, from Fig. 5 one can see that a large majority of the systems is not spin-polarized (note that the y-axis is truncated). In fact, the probability of finding a magnetic compound is only 22.6% (4187 systems out of 18,479), and with the number of systems decreasing rapidly with the total magnetization. We show the number of magnetic systems containing each given element of the periodic table in Fig. 6. The ten most represented metallic elements in these magnetic compounds are, in decreasing order, Pu, Eu, Gd, Mn, Fe, Np, Ge, Ce, Ni, and Co. These include, evidently, the actinides (Pu and Np), the lanthanides (Eu, Gd, and Ce), and the 3d transition metals (Mn, Fe, Ni, and Co).

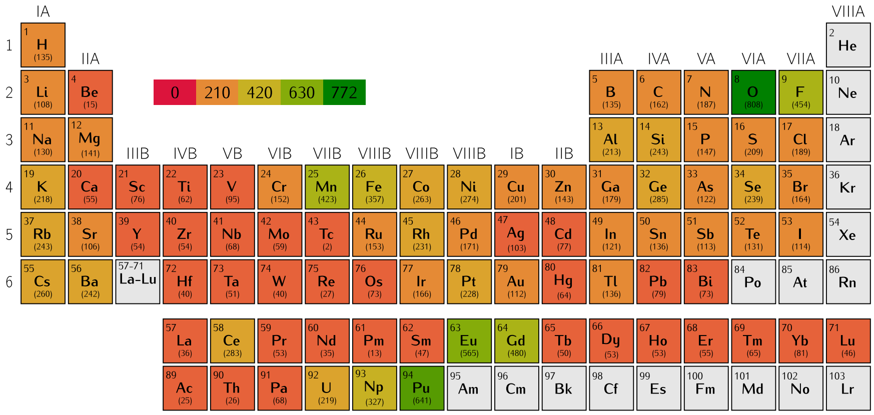

Fig. 6 Distribution of magnetic systems. The number of magnetic systems containing a given element of the periodic table.

The fact that Ge appears in this list is actually interesting. By looking closer at the composition of the magnetic compounds containing this chemical element, we found that 91% of Ge-containing magnetic compounds include at least one other element included in the top-10 list. Moreover, the remaining 9% compounds also contain other rare-earth or transition metals. A quick look at some specific materials in our list reveals that the magnetic moments are not localized on Ge, but on the other (magnetic) atoms. Therefore, the reason why Ge appears in the list is that Ge is likely to form stable compounds together with magnetic elements. This also implies that Ge compounds could be a promising search ground for experimentalists aiming at the synthesis of magnetic compounds.

The most common non-metallic elements found in this set are O, and F. Among all systems, the highest magnetization is around 0.2 $\mu_B \cdot \text{Å}^{-3}$ for a cubic structure of $SnGd_3$, followed by several other Gd and Eu compounds, often in the inverted perovskite structure (such as $NAlGd_3$, $CGeGd_3$, $CGaGd_3$, and $CSnGd_3$). Finally, the most common crystal phase is the cubic double-perovskite structure with 215 compounds, while magnetic systems were found in a total of 253 different prototypes.

Having looked at magnetic systems and semiconductors, it is natural to ask how many magnetic semiconductors are found in our dataset. If the two properties are completely uncorrelated, the probability of finding a system exhibiting both is given by the product of the individual probabilities, yielding 22.6% × 26.1% = 5.9%. The actual number of systems that we found was 884, yielding a probability of 4.8%. This is consistent with the two properties being uncorrelated. We also performed a similar analysis on the Materials Project database. The fractions of stable systems with a gap above 0.1 eV and of magnetic systems are 45.7% and 31.5%, respectively. This yields a combined probability of 14.4% to find magnetic semiconductors if the two properties are uncorrelated. The actual percentage of stable magnetic semiconductors in the database is 12.1%, which also supports the hypothesis of absence of correlations.

Among all semiconducting magnetic systems, the most common prototype that we found was again the cubic double-perovskite (75 systems). We note that most magnetic semiconductors could be, in fact, antiferromagnetic. Moreover, usually the antiferromagnetic state has a larger gap than the ferromagnetic one. Therefore, those band gaps could be "doubly" underestimated—due to the PBE approximation and the misidentification of magnetic phases. This subset of 884 materials is however, quite interesting, as it can serve, e.g., as a starting basis for the discovery of unreported transparent ferromagnets or anti-ferromagnets with high critical temperatures.

Table 2. Vicker's Hardness (H_V), as well as bulk (B) an shear (G) moduli of some hard and superhard materials.

Formula	H_V (GPa)	B (GPa)	G (GPa)
NiO_4	28.7	60.9	35.7
AsB_3O_6	29.7	53.2	33.3
CuO_4	31.7	65.4	25.0
CoO_4	36.6	107.6	72.5
$BeCrFe_2$	25.2	241.9	126.4
RuN_2	30.2	163.8	93.3
IrN_2	30.7	177.8	99.0
CoH	34.8	217.2	116.9
VRu_2Sn	41.5	210.8	85.8
$CrGeRu_2$	58.3	235.3	117.3
MnH_2	64.4	133.6	49.6

Mechanical properties

Finally, we performed a preliminary analysis of the mechanical properties by evaluating the hardness. The calculation of the Vicker's hardness for the predicted structures was based on the simple model by Zhang et al.[43] This model extends the work of Šimůnek and Vackář[44,45] and improves the earlier hardness models[46] based on bond strength by applying the Laplacian matrix[47] to account for highly anisotropic and molecular systems. It turns out that laminar systems are correctly described as having low hardness, but this model still fails for some molecular crystals that are incorrectly assigned large values for the hardness. This is, however, not a big problem as these false-positive cases can be easily identified and discarded.

Most systems are found to be extremely soft, with only a handfull of materials being hard or superhard (hardness > 40 GPa). These, usually a combination of light covalent elements with transition metals, are shown in Table 2, together with their bulk and shear moduli (calculated with the PBE). We found that the oxides in this list have low elastic moduli, which implies that the simple model has likely overestimated their hardness. This anomalous behavior can be explained by the unusual oxidation states and bonding patterns present in these structures. One should keep in mind that the stability of these oxides is likely overestimated, as it has been shown in several references[48,49]

The remaining systems do exhibit large values of the hardness and of the bulk and shear moduli, indicating that they are probably hard or even super-hard. This is particularly true for three compounds, namely VRu_2Sn, $CrGeRu_2$, and MnH_2.

DISCUSSION

In this work, we combined fundamental knowledge of chemistry with high-throughput calculations to efficiently search for stable crystals. To this end, we replaced chemical elements in known stable substances by choosing substitutions with "similar" chemical elements. The elusive concept of similarity was quantified by a similarity scale obtained by data-mining experimental databases of crystal structures. The transmuted compounds were then studied with DFT, and their stability was evaluated with respect to the convex hull of stability. The stable compounds were in their turn transmuted, and this cycle was repeated three times.

We obtained in total 18,479 stable crystal structures out of 189,981 substitutions, resulting in a success rate of about 10%, one order of magnitude larger than the usual one of high-throughput methods. This success rate shows not only the validity of our approach, but also its high efficiency, leading to a significant reduction of the computational costs. Our set of stable materials include elements from across the periodic table, from main group elements to transition metals, to lanthanides and actinides.

We also performed a preliminary analysis of the physical properties of these crystals. We obtained 4840 semiconductors, with band gaps (calculated with the PBE approximation) extending almost to 8 eV. These include not only many oxides and fluorides, but also semiconductors with other halogens, chalcogens, pnictogens, etc. We also identified 4187 magnetic systems with magnetizations extending up to 0.2 $\mu_B \cdot \text{Å}^{-3}$. As expected, these mostly include some actinides, some lanthanides, and some 3d transition metals. Combining both properties, we filtered out 884 structures having non-zero gap and magnetic moments. Finally, we evaluated the hardness of our materials, and found few possible hard and super-hard systems that deserve further attention.

All in all, this work shows that with a systematic help of common chemistry knowledge, one can greatly improve the output of high-throughput calculations for material prediction. Thanks to this iterative procedure of transmutation, we efficiently gain access to large unknown portions of the phase space of stable materials, that may be hiding key materials for future technologies.

METHODS

Prediction strategy

The starting point of our search is a set of stable compounds, i.e. the (experimental or theoretical) crystal structures and compositions of a series of materials on the convex hull of stability. We obtained these structures from the materials project database[50]. For computational affordability, we limited crystal structures to a maximum of 12 atoms in the unit cell. The starting set is composed of 9524 compounds in 713 different prototype crystal structures. For each material in this set, we mutate the composition by replacing each chemical element by another "similar" element, if the probability of a successful replacement is higher than a certain threshold. We will see below how we define this probability. Note that only one element is replaced at a time, and that we do not perform partial substitutions, i.e. all atoms of a given element in the crystal structure are replaced simultaneously.

The outcome of this procedure is a set of hypothetical materials. We observe that it is impossible to perform systematic substitutions of all elements in known stable crystal structures, employing all other atoms of the periodic table. Assuming 84 atomic species, from H up to Bi, excluding noble gases and including Ac, Th, Pa, U, Np, and Pu, and considering 713 prototype crystal structures, we can build 59,892 elementary crystals,

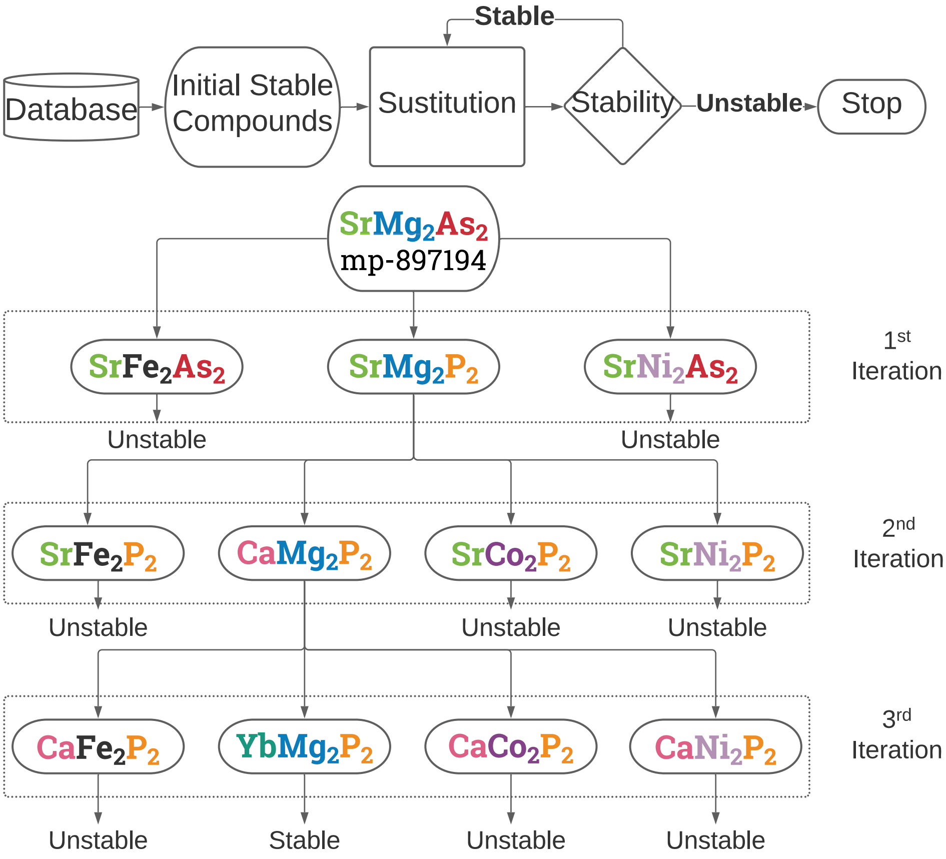

Fig. 7 Work flow. An illustration of a work flow for predicting stable materials based on substitution.

Table 3. The number of substitution pairs (N_{pairs}) and the quantity of resulting compounds ($N_{compounds}$) as a function of the threshold (t), starting from the initial set of 9524 compounds.

t	N_{pairs}	$N_{compounds}$
70%	7	214
60%	16	214
50%	47	824
40%	118	1469
30%	200	1957
20%	346	12,007
10%	626	35,579
5%	992	73,375
4%	1111	87,738
3%	1281	104,508
2%	1556	142,617
1%	2235	234,385
0.75%	2554	277,111
0.5%	3008	341,180

almost 5 million binaries, 400 million ternaries, and 33 billion quaternary compositions. We can clearly see that we need to filter out the most unlikely substitutions and focus on the most promising ones.

At any iteration, we validate the set by performing a geometry optimization of the resulting structure with DFT, and calculating its distance to the convex hull of stability. This step is performed with PYMATGEN[51], using all materials present in the Materials Project database[50] as reservoirs. All stable phases (with negative distances to the materials project convex hull) are then collected, and the construction of the convex hull is repeated including our structures. A new cycle of substitutions starts then for the stable compounds identified in the previous iteration. In total we performed three iterations of this kind, replacing always one chemical species per iteration. Thus, the prediction procedure is illustrated in Fig. 7.

Of course, the crucial part of this approach is the knowledge of the probability that replacing an element by another will yield a stable compound. We could just take advantage of the periodic table, and define this probability as the (geometrical) distance between the two elements in its usual two-dimensional representation. A couple of counter-examples show, however, that this is clearly not the ideal approach. For example, it turns out that H can be much more easily replaced by F and not by Li, or Ba can be replaced by Eu more often than by Cs.

One can certainly use for filtering empirical rules based, e.g., on ionic radii and oxidation states[27]. However, in the age of data-driven research, we have the option to let computer algorithms transform empirical chemical knowledge into a similarity scale between the chemical elements. Recently, by performing a statistical analysis of stable crystal phases

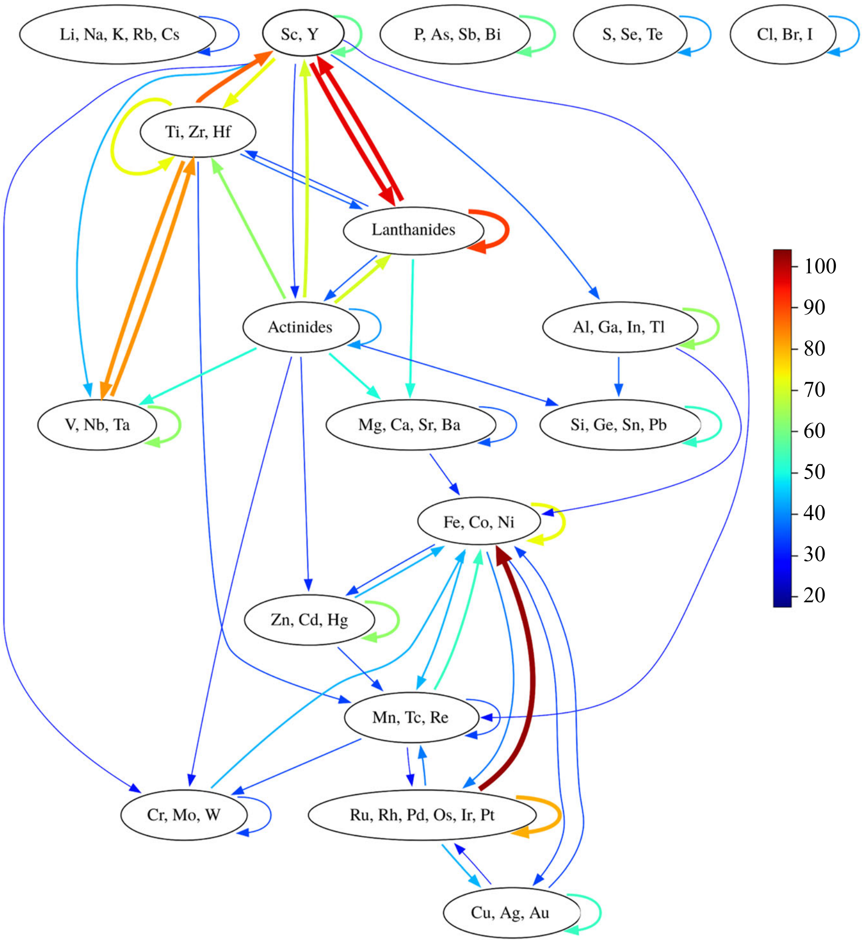

Fig. 8 Substitution schema. Replacements are shown by arrows that start from the elements being replaced. Substitutions between elements within a group are indicated by arrows starting from and pointing to the same box. The thickness of the arrow and the color scale are proportional to the number of substitutions between the groups, with the thick red line between the Ru-group and the Fe-group corresponding to 100% replaceability between the two groups. For example, we can immediately see that most lanthanides can be replaced by Sc or Y, but the elements of the group IIIA can only sometimes be replaced by Fe, Co, or Ni.

present in the inorganic crystal structure database[52,53], some of us determined such a scale[33]. The first step was the calculation of the likelihood that an element A can be replaced by another B in a given structure. This information was then used to construct a matrix where each entry (A, B) is a measure of this likelihood. To obtain a probability, every entry of this matrix has to be normalized in some way. This is a rather non-trivial step that is complicated by the fact that our knowledge of materials is unfortunately rather incomplete. Here, we used the quantity[33]

$$S_{AB} := \frac{1}{N_A}\sum\nolimits_{I,J\neq I}\delta_{AB}^{IJ} \qquad (1)$$

where $\delta_{AB}^{IJ} = 1$ if materials I and J are both in the experimental database and are connected by the substitution of the chemical element A by B, and is 0 otherwise. The normalization factor (N_A) is the total number of materials including the given chemical element that are present in the database.

We also need a threshold value of the element replaceability, below which we do not consider as likely the corresponding element mutation. We set the threshold to a value that is a good compromise to keep affordable the total number of substituted compounds and to have at the same time a sufficient variety of substitution pairs. A threshold lower than 20% is necessary to include all substitutions within each group of the periodic table. This means that fixing this threshold to 20% would lead to include only "obvious" cases, while we would miss other less intuitive and less common substitutions. We therefore decided to favor a practical approach and include as many substitutions as possible, selecting the lowest threshold that our computational resources could reasonably support. We have to keep in mind that the number of substitution increases rapidly with the number of substitution pairs, because we have a large initial set of materials (see Table 3 and discussion in Section "Thermodynamic stability"). We chose a threshold value equal to 5% that gives 992 pairs (see List 1 in Supplementary Notes), a number that is approximately twice as large as the number of in-group substitution pairs.

A schema depicting the result of this procedure can be found in Fig. 8. To improve readability, we gathered the chemical elements in groups. There are a series of immediate conclusions we can draw from the figure. First of all, with the chosen threshold, almost no first-row element can be replaced by any other element. In chemistry this is known as the first-row anomaly[54], i.e., the small-core elements of the first row are in some sense special and are only vaguely similar to second-row elements. Second, many elements only accept replacements with elements within the same group of the periodic table. This is in particular true for the alkali metals, the halogens, etc. Third, we identify two main groups of metals in Fig. 8, one centered around the lanthanides and the other around Fe, Co, and Ni.

It is rather interesting that our threshold roughly divides the metals in two families. The subdivision is simply related to the geometry of the periodic table, namely family I includes the left side of the periodic table (groups 2–5, as well as the lanthanides and actinides), while family II contains the remaining groups (6–15). Furthermore, we find no substitutions between group 5 and 6. This would indeed indicate that there seems to be a significant discontinuity in the periodic table. In fact, we can see some indications of this discontinuity by looking, for example, at the typical oxidation states that show from a monotonous increase from +2 (group 2), +3 (group 3), +4 (group 4), +5 (group 5) back to +3 and +4 in group 6. However, we emphasize that this analysis depends on our choice for the threshold, and that a more detailed investigation, using more powerful statistical tools, is required to achieve general conclusions.

DFT calculations

We used the code VASP[55,56], where all parameters were set to guarantee compatibility with the data available in the Materials Project database[50]. We used the PAW[57] datasets of version 5.2 with a cutoff of 520 eV. The Brillouin zone was sampled by Γ-centered k-point grids with a uniform density calculated to yield 1000 points per reciprocal atom, i.e. the same k-point density used by the Materials Project[58]. All energies were converged to better than 2 meV per atom and the geometry optimization was stopped when forces were smaller than 0.005 eV per Å. We used a denser k-point mesh of 5000 points per atom to calculate band structures. All calculations were performed with spin-polarization using the PBE[59] exchange-correlation functional, with the exception of oxides and fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V, W, where an on-site Coulomb repulsive interaction U with a value of 3.32, 3.7, 5.3, 3.9, 4,38, 6.2, 3.25, and 6.2 eV, respectively, was added to correct the d-state (https://docs.materialsproject.org/methodology/gga-plus-u/#calibration-of-u-values). A correction scheme which allows to mix GGA and GGA + U calculations to obtain the correct formation energy and distance to the convex hull is applied[60].

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request. All stable structures found in this work are available on https://tddft.org/bmg/data.php.

CODE AVAILABILITY

The related codes are available from the corresponding authors upon reasonable request.

Received: 27 February 2020; Accepted: 1 December 2020;
Published online: 26 January 2021

REFERENCES

1. Wood, J. The top ten advances in materials science. *Mater. Today* **11**, 40–45 (2008).
2. Chamorro, J. R. & McQueen, T. M. Progress toward solid state synthesis by design. *Acc. Chem. Res.* **51**, 2918–2925 (2018).
3. Jose, R. & Ramakrishna, S. Materials 4.0: materials big data enabled materials discovery. *Appl. Mater. Today* **10**, 127–132 (2018).
4. Arsenault, A. et al. Towards the synthetic all-optical computer: science fiction or reality? *J. Mater. Chem.* **14**, 781–794 (2004).
5. Fortunato, E. & Martins, R. Where science fiction meets reality? with oxide semiconductors! *Phys. Status Solidi RRL* **5**, 336–339 (2011).
6. Atwater, H. A. et al. Materials challenges for the starshot lightsail. *Nat. Mater.* **17**, 861–867 (2018).
7. Marzari, N. Materials modelling: the frontiers and the challenges. *Nat. Mater.* **15**, 381–382 (2016).
8. Gielen, D., Boshell, F. & Saygin, D. Climate and energy challenges for materials science. *Nat. Mater.* **15**, 117–120 (2016).
9. Walsh, A. Inorganic materials: the quest for new functionality. *Nat. Chem.* **7**, 274–275 (2015).
10. Butler, K. T., Frost, J. M., Skelton, J. M., Svane, K. L. & Walsh, A. Computational materials design of crystalline solids. *Chem. Soc. Rev.* **45**, 6138–6146 (2016).
11. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
12. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
13. Burke, K. Perspective on density functional theory. *J. Chem. Phys.* **136**, 150901-1–150901-9 (2012).
14. Tanaka, I., Rajan, K. & Wolverton, C. Data-centric science for materials innovation. *MRS Bull.* **43**, 659–663 (2018).
15. Potyrailo, R. et al. Combinatorial and high-throughput screening of materials libraries: review of state of the art. *ACS Comb. Sci.* **13**, 579–633 (2011).
16. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
17. Ward, L. & Wolverton, C. Atomistic calculations and materials informatics: a review. *Curr. Opin. Solid State Mater. Sci.* **21**, 167–176 (2017).
18. Cerqueira, T. F. T. et al. Materials design on-the-fly. *J. Chem. Theory Comput.* **11**, 3955–3960 (2015).
19. Liu, Y., Zhao, T., Ju, W. & Shi, S. Materials discovery and design using machine learning. *J. Materiomics* **3**, 159–177 (2017).
20. 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 (2010).
21. Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104-1–094104-7 (2014).
22. Zakutayev, A. et al. Experimental synthesis and properties of metastable $CuNbN_2$ and theoretical extension to other ternary copper nitrides. *Chem. Mater.* **26**, 4970–4977 (2014).
23. Shoemaker, D. P. et al. In situ studies of a platform for metastable inorganic crystal growth and materials discovery. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 10922–10927 (2014).
24. Blum, V. & Zunger, A. Prediction of ordered structures in the bcc binary systems of Mo, Nb, Ta, and W from first-principles search of approximately 3,000,000 possible configurations. *Phys. Rev. B* **72**, 020104-1–020104-4 (2005).

25. Curtarolo, S. et al. AFLOWLIB.ORG: a distributed materials properties repository from high-throughput ab initio calculations. *Comput. Mater. Sci.* **58**, 227–235 (2012).
26. Dshemuchadse, J. & Steurer, W. More statistics on intermetallic compounds-ternary phases. *Acta Crystallogr.* **71**, 335–345 (2015).
27. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
28. Graser, J., Kauwe, S. K. & Sparks, T. D. Machine learning and energy minimization approaches for crystal structure predictions: a review and new horizons. *Chem. Mater.* **30**, 3601–3612 (2018).
29. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
30. Schmidt, J. et al. Predicting the thermodynamic stability of solids combining density functional theory and machine learning. *Chem. Mater.* **29**, 5090–5103 (2017).
31. 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**, 241728-1–241728-6 (2018).
32. Hautier, G., Fischer, C., Ehrlacher, V., Jain, A. & Ceder, G. Data mined ionic substitutions for the discovery of new compounds. *Inorg. Chem.* **50**, 656–663 (2011).
33. Glawe, H., Sanna, A., Gross, E. K. U. & Marques, M. A. L. The optimal one dimensional periodic table: a modified Pettifor chemical scale from data mining. *New J. Phys.* **18**, 093011-1–093011-8 (2016).
34. 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**, 155208-1–155208-18 (2012).
35. 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**, 4-1–4-9 (2019).
36. Sarmiento-Perez, R., Cerqueira, T. F. T., Körbel, S., Botti, S. & Marques, M. A. L. Prediction of stable nitride perovskites. *Chem. Mater.* **27**, 5957–5963 (2015).
37. Körbel, S., Marques, M. A. L. & Botti, S. Stability and electronic properties of new inorganic perovskites from high-throughput ab initio calculations. *J. Mater. Chem. C* **4**, 3157–3167 (2016).
38. Kirchmayr, H. R. Permanent magnets and hard magnetic materials. *J. Phys. D Appl. Phys.* **29**, 2763–2778 (1996).
39. McLellan, B., Corder, G., Ali, S., Golev, A. *Rare metals, unconventional resources, and sustainability* (Geological Society of America, 2016).
40. Tran, F. & Blaha, P. Importance of the kinetic energy density for band gap calculations in solids with density functional theory. *J. Phys. Chem. A* **121**, 3318–3325 (2017).
41. Pauling, L. *The Nature of the Chemical Bond...* (Cornell university press Ithaca, 1960).
42. Feng, X. & Harrison, N. M. Magnetic coupling constants from a hybrid density functional with 35% Hartree-Fock exchange. *Phys. Rev. B* **70**, 092402-1–092402-4 (2004).
43. Zhang, X. et al. First-principles structural design of superhard materials. *J. Chem. Phys.* **138**, 114101-1–114101-9 (2013).
44. Šimůnek, A. & Vackář, J. Hardness of covalent and ionic crystals: first-principle calculations. *Phys. Rev. Lett.* **96**, 085501-1–085501-4 (2006).
45. Šimůnek, A. How to estimate hardness of crystals on a pocket calculator. *Phys. Rev. B* **75**, 172108-1–172108-4 (2007).
46. Gao, F. M. & Gao, L. H. Microscopic models of hardness. *J. Superhard Mater.* **32**, 148–166 (2010).
47. Trinajstic, N. et al. The laplacian matrix in chemistry. *J. Chem. Inf. Model.* **34**, 368–376 (1994).
48. Wang, L., Maxisch, T. & Ceder, G. Oxidation energies of transition metal oxides within the GGA + U framework. *Phys. Rev. B* **73**, 195107-1–195107-6 (2006).
49. Kang, S., Mo, Y., Ong, S. P. & Ceder, G. A facile mechanism for recharging Li_2 O_2 in Li–O_2 batteries. *Chem. Mater.* **25**, 3328–3336 (2013).
50. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002-1–011002-11 (2013).
51. Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
52. Bergerhoff, G., Brown, I.D. *Crystallographic Databases* (International Union of Crystallography, 1987).
53. Belsky, A., Hellenbrandt, M., Karen, V. L. & Luksch, P. New developments in the inorganic crystal structure database (ICSD): accessibility in support of materials research and design. *Acta Crystallogr. B.* **58**, 364–369 (2002).
54. Miessler, G.L., Tarr, D.A. *Inorganic Chemistry 3rd edn* (Pearson Prentice Hall, 2004).
55. 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 (1996).
56. 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 (1996).
57. Blöchl, P. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
58. Jain, A. et al. A high-throughput infrastructure for density functional theory calculations. *Comput. Mater. Sci.* **50**, 2295–2310 (2011).
59. Perdew, J., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
60. Jain, A. et al. Formation enthalpies by mixing GGA and GGA + U calculations. *Phys. Rev. B* **84**, 045115-1–045115-10 (2011).

ACKNOWLEDGEMENTS

S.B. and M.A.L.M. acknowledge financial support from the DFG through Projects MA 6787/1-1, and BO 4280/8.

AUTHOR CONTRIBUTIONS

All three authors contributed equally to the calculations and writing of the paper.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-00481-6.

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