High-Throughput Discovery and Characterization of Inorganic Materials Using Ab-initio Methods

Dissertation

zur Erlangung des Doktorgrades der Naturwissenschaften
Dr. rer. nat.

Der

Naturwissenschaftlichen Fakultä II

Chemie, Physik und Mathematik

der Martin-Luther-Universität

Halle-Wittenberg

vorgelegt von

Herrn Wang, Hai-Chen

geb. am 25.Apr.1988 in Kunming, China

Erstgutachter: Prof. Dr. Miguel Marques
Zweitgutachter: Prof. Dr. Nektarios N. Lathiotakis
Drittgutachter: Prof. Dr. Ingrid Mertig
Verteidigung: 22.Aug.2023

Abstract

Functional materials are one of the foundations of modern technology. However, often they contain toxic and unsustainable elements. Searching for more efficient, cheaper, and eco-friendly alternatives is one of the most urgent tasks of solid-state physics. Luckily, the vast majority of the chemical space is unknown, which provides enormous opportunities for finding new materials. However, sadly, the total number of possible compositions and structures in the chemical space exceeds the capability of brute force experimental or theoretical searches.

In this cumulative thesis, we perform prototype-based high-throughput searches in the chemical space of inorganic materials, i.e., we confine our exploration of the structural subspace to prototype structures and only scan the compositional subspace.

We use the density functional theory (DFT) calculations to optimize the crystal structures of the candidates and evaluate their thermodynamic stability according to their distances to the convex hull.

We first show that for one specific prototype, simply applying intuition from domain knowledge to pre-select compositions with non-negligible chances to be stable will achieve successful search outcomes. Then, we use a less-human-intervened pre-selection strategy based on a data-mining scale of similarities between elements. We generate new candidates by replacing elements with similar ones in known crystals. In this way, we are able to perform high-throughput searches for all of the prototypes simultaneously, rather than just on a few ones. The resulting stable compounds amount to about 19,000, more than 50% of the theoretically known stable compounds at the time. Moreover, the success rate, i.e., the percentage of stable compounds among all computed ones, is almost 10%, which is one or two orders of magnitude higher than the usual systematic brute force high-throughput search.

We push the limits of high-throughput even further by using the trained machine learning models in the pre-selection phase. It should be noted that training such models also benefited from the large datasets we obtained in previous high-throughput searches. We exhaustively and systematically generate possible two-dimensional motifs (prototypes), and with the help of the models, we filter out around 6,500 new 2D systems, more than doubling the known amount of (meta-)stable 2D systems. We also show that using machine learning models in pre-filtering can dramatically increase the success rate of the high-throughput search.

Abstract

Funktionswerkstoffe bilden die Grundlage vieler moderner Technologien. Allerdings enthalten sie oft giftige und nicht nachhaltige Elemente. Die Suche nach effizienteren, billigeren und umweltfreundlicheren Alternativen ist eine der dringendsten Aufgaben der Materialwissenschaften. Glücklicherweise, ist der größte Teil des chemischen Raums noch unbekannt, was enorme Möglichkeiten für die Suche nach neuen Materialien bietet. Leider übersteigt jedoch die Gesamtzahl der möglichen Zusammensetzungen und Strukturen im chemischen Raum die Möglichkeiten der experimentellen oder theoretischen brute-force Hochdurchsatzstudien.

In dieser kumulativen Arbeit führen wir eine prototypbasierte Hochdurchsatzsuche im chemischen Raum anorganischer Materialien durch, d.h. wir beschränken unsere Erkundung des strukturellen Unterraums auf Prototypstrukturen und scannen alle möglichen chemischen Zusammensetzungen.

Mit Hilfe der Dichtefunktionaltheorie (DFT) optimierten wir die Kristallstrukturen der Materialskandidaten und bewerteten ihre thermodynamische Stabilität anhand ihrer Abstände zur konvexen Hülle.

Wir zeigen zunächst, dass bei einem bestimmten Prototypen die Anwendung der chemischen Intuition zur Vorauswahl von Zusammensetzungen, zu erfolgreichen Suchergebnissen führt. Anschließend verwenden wir eine Vorauswahlstrategie, die auf einer Data-Mining-Skala für Ähnlichkeiten zwischen Elementen basiert und weniger auf menschliches Wissen setzt. Wir generieren neue Kandidaten, indem wir Elemente durch ähnliche Elemente in bekannten Kristallen ersetzten. Auf diese Weise sind wir in der Lage, eine Hochdurchsatzsuche für alle Prototypen gleichzeitig durchzuführen, anstatt nur für einige wenige. Als Resultat wurde die Anzahl theoretisch bekannter stabiler Verbindungen um 19.000 Materialien oder etwa 50% erhöht. Darüber hinaus liegt die Erfolgsquote, d. h. der Prozentsatz der stabilen Verbindungen unter allen berechneten Verbindungen, bei fast 10% und damit um eine oder zwei Größenordnungen höher als bei den üblichen systematischen brute-force Hochdurchsatzsuche.

In der letzten Studie schieben wir die Grenzen von Hochdurchsatzstudien noch weiter hinaus, indem wir maschinellen Lernmodelle in der Vorselektionsphase eingesetzt haben. Es ist anzumerken, dass das Training solcher Modelle massiv von den großen Datensätzen profitiert, die wir bei den vorherigen Hochdurchsatzsuche produziert

haben. Wir haben alle mögliche zweidimensionale Motive (Prototypen) systematisch generiert und mit Hilfe der Modelle rund 6.500 neue 2D-Systeme herausgefiltert, was mehr als eine Verdoppelung der bekannten Menge an (meta-)stabilen 2D-Systemen bedeutet. Wir zeigen ebenfalls, dass maschinelles Lernen bei der Vorfilterung die Erfolgsrate der Hochdurchsatzsuche drastisch erhöhen kann.

Contents

Introduction

One of modern technology's most essential and urgent tasks is to design new functional materials that improve existing applications or even unlock new ones. In the past century, pursuing this goal has indeed brought humankind numerous breakthroughs, such as the giant magneto resistance [1, 2], blue light-emitting diode [3, 4], the lithium-ion battery [5–7], semiconductor hetero-structures [8], optical fibers [9], etc. Nearly all of these achievements are dominantly done through experiments, which, to a large extent, are mainly based on human ingenuity, or in other words, intuition from chemistry and physics. Moreover, one successful and pioneering experimental attempt is the tip of the iceberg formed by randomized try-and-error fails. Limited by efficiency, the experimental exploration is unfortunately confined. For decades, the synthesis focused on islands of known stable compounds and their vicinity in the chemical space. The total amount of inorganic crystal structures known experimentally is around tens of thousands. For example, the Inorganic Crystal Structure Database (ICSD) contains 272,260 entries of information on about 40 thousand crystal structures [10]. On the other hand, the chemical space for one ternary structure prototype, e.g., the perovskite ABC_3, contains around 500 thousand systems, On top of that, the number of ternary prototypes is at the level of hundreds. The vast majority of the chemical space for inorganic materials remains unexplored. However, we often lack a sketch map for the unknown part, searching for new materials is more like finding a route through a labyrinth.

In the past decades, exponential improvements in the performance of the central processing unit (CPU) have enabled powerful computational approaches to explore this labyrinth. Applying the long-established and wide-utilized methods, for example, density functional theory (DFT), we can get the extrapolated energy surface in ternary chemical coordinates nowadays [WPhD12]. With more than a decade of theoretical high-throughput searches, the size of modern computational inorganic material databases is one or two orders of magnitude larger than experimental ones, e.g., the AFLOW database [11] contains 3,528,653 entries on 188,343 systems. More importantly, compatible and transferable computational data allows for convenient and efficient high-throughput filtering of new functional materials [WPhD1, 12–17].

The work of this Thesis tries to humbly contribute to this topic. Firstly, we performed a high-throughput search on double perovskites for promising p-type transparent

semiconductors [WPhD1]. In Chapter 3, we aim to find wide band gaps with high hole mobilities (small hole effective masses), two components unlikely to co-exist from the prediction of $k \cdot p$ theory and the band edge characters of current popular oxides candidates. As a result, we find 17 halide double perovskites with promising properties. These halides are chemically and structurally compatible with the inorganic-organic hybrid perovskites widely experimented with in photovoltaics. Furthermore, 10 of them do not contain toxic or rare chemical elements.

We further searched stable mixed anion perovskites for optoelectronic applications [WPhD2]. In Chapter 4, we consider the stoichiometry ABX_2Y for the mixed anion (X, Y = N, F, O) perovskites. Our results are consistent with the experimental literature on synthesized systems. Moreover, we predicted a series of novel oxynitrides and oxyfluorides. The nitrofluorides, on the other hand, could only stabilize in the $LaMgF_2N$ composition. We also show that the disorder of anions thermodynamically stabilizes the structure, even without considering the entropy effect. Moreover, many of our predicted systems are semiconducting or insulating, with electronic band gaps going from less than 1 eV to several eV, confirming that anion alloying is an effective way of bandgap engineering.

Although these high-throughput searches based on *ab initio* methods accelerate the exploration of the labyrinth, it is still too computationally expensive to scan it "inch by inch" by performing a brute force systematic search. One alternative is to avoid wasting computational resources on highly unstable areas. Unfortunately, we lack enough knowledge to avoid them *a priori* without scanning the majority of the labyrinth. To solve this, we borrow the intuition from experimental chemistry, where the similarity of elements has been extensively discussed. The similarity between a pair of elements **A**–**B** correlates with the probability of obtaining a stable compound of **B** (Comp(**B**)) from a known compound of **A** (Comp(**A**)) via substituting **A** with **B**. Such a similarity scale is available from data-mining the experimental results [18]. In Chapter 5, by using this scale, we rule out replacements between non-similar elements which should have a negligible chance of giving stable compounds. In this way, we can sketch a map of the labyrinth *a priori*, and our "guided" high-throughput search has a significantly higher success rate (the ratio of finding stable materials out of all calculations) than usual systematic scans [WPhD3].

Furthermore, machine learning techniques are used to pre-explore the labyrinth. Within the past several years, the development of machine learning models has

allowed extraordinarily efficient and successful predictions for the stability of inorganic materials [19, 20]. In the last part of this Thesis (Chapter 6), we apply a pre-trained model [21] for three-dimensional (3D) structures to accelerate the high-throughput search for novel two-dimensional (2D) materials [WPhD4]. By starting with an exhaustive systematic approach to generate candidate structures, we tried to go beyond the available 2D prototypes from the exfoliation of 3D systems. We applied the machine model on thermodynamic stability [21] to pre-filter out stable structures. Additionally, we used a machine learning force field [22] to pre-optimize the geometry to speed up the DFT validation procedure. The DFT results are then used to re-train the model during the process. As a result, this work discovers thousands of unexpected 2D phases with no layered three-dimensional counterpart. Moreover, utilizing machine learning is self-accelerating because the accumulation of DFT validation data gradually improves the accuracy of models. This virtuous cycle will pave the way for a complete exploration of the two-dimensional part of the labyrinth in the near future.

Density Functional Theory

The first-principles, also known as the *ab initio*, refers to the method of obtaining the electronic structures, atomic interactions, and other characteristics of the object of study through self-consistent calculations without relying on any experimental parameters or fitting manners. Its core and basis is the solution of the time independent Schrödinger equation for a many-body system

$$\hat{H}(\mathbf{r},\mathbf{R})\Psi(\mathbf{r},\mathbf{R}) = E\Psi(\mathbf{r},\mathbf{R}), \tag{1}$$

$$\hat{H} = \hat{T} + \hat{V}. \tag{2}$$

The kinetic operator contains both nuclei and electron parts

$$\hat{T} = \hat{T}_{\mathrm{N}} + \hat{T}_{\mathrm{e}} = -\frac{1}{2m_I}\sum_{I=1}^{M}\nabla_I^2 + -\frac{1}{2}\sum_{i=1}^{N}\nabla_i^2 \tag{3}$$

The potential operator contains the nuclei-nuclei, nuclei-electron, and electron-electron interactions

$$\hat{V} = \hat{V}_{\mathrm{NN}} + \hat{V}_{\mathrm{Ne}} + \hat{V}_{\mathrm{ee}} = \sum_{I<J}^{M}\frac{Z_I Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} - \sum_{i,I=1}^{\mathrm{N,M}}\frac{Z_I}{|\mathbf{r}_i - \mathbf{R}_I|} + \sum_{i<j}^{N}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{4}$$

The (ground state) structure of any system and its properties can be deduced from the solution of the above Schrödinger equation. However, the interaction potentials inside the many-body system are extremely complex, and the number of particles in the real system is so large that it is impossible to solve analytically. Since the twentieth century, technological advances and development of algorithms have provided the conditions for numerically approach. However, the process still requires expensive computational resources, thus is still unfeasible for most cases. Therefore, a series of approximations are necessary. Of course, the reliability and practicality of approximations are always tested in practice.

Adiabatic Approximation

The mass of the nucleus is much higher than the mass of the electron, the former being at least 10^3 times larger than the latter, which means the motion of electrons is significantly faster than that of the nuclei, the electron would "instantaneous" adjust

their positions after the nuclei moved so that the nuclei and electron motions can be separated. This is the basis for the adiabatic or Bonn-Oppenheimer approximation.

Now, the nuclei are independent of the electron motion and the nuclear wavefunction $\chi_i(\mathbf{R})$ is only associated with the following nuclei Shrödinger equation

$$[\hat{T}_{\mathrm{N}}(\mathbf{R}) + \hat{V}_{\mathrm{NN}}(\mathbf{R})]\chi_i(\mathbf{R}) = E\chi_i(\mathbf{R}) \tag{5}$$

And the Schrödinger equation for the electrons takes the nuclear coordinates $\mathbf{R}$ as parameters

$$[\hat{T}_{\mathrm{e}}(\mathbf{r}) + \hat{V}_{\mathrm{Ne}}(\mathbf{r}, \mathbf{R}) + \hat{V}_{\mathrm{ee}}(\mathbf{r})]\Psi_i(\mathbf{r} : \mathbf{R}) = E_i\Psi_i(\mathbf{r} : \mathbf{R}) \tag{6}$$

The overall wavefunction of the many-body system is just

$$\Psi(\mathbf{r}, \mathbf{R}) = \sum_i \chi_i(\mathbf{R})\Psi_i(\mathbf{r} : \mathbf{R}). \tag{7}$$

Single-Electron Approximation

Even with the adiabatic approximation, the total number of electrons in the general system is so large that the description of the interaction between electrons and electrons becomes extremely complicated as the number of electrons increases. Nevertheless, the direct solution is still very difficult and time-consuming. Hartree–Fock single-electron approximation (H–F approximation) can simplify the electron-electron interaction further. In Hartree–Fock the potential for each electron is split into two parts: the Coulomb potential from all nuclei and the total Coulomb potential from all electrons except the electron under consideration. In this way, the Hamiltonian can be formulated as the sum of one electron (Fock) operators $\hat{F}[\{\phi_j\}](1)$, and

$$\hat{F}[\{\phi_j\}](1) = \hat{H}^{\mathrm{core}}(1) + \sum_{j=1}^{N/2}[2\hat{J}_j(1) - \hat{K}_j(1)] \tag{8}$$

$$\hat{H}^{\mathrm{core}}(1) = -\frac{1}{2}\nabla_1^2 - \sum_\alpha \frac{Z_\alpha}{r_{1\alpha}} \tag{9}$$

$\hat{H}^{\mathrm{core}}(1)$ is the one-electron core Hamiltonian, $\hat{J}_j(1)$ is the Coulomb operator defining the electron–electron repulsion in the j-th orbital ϕ_j, and $\hat{K}_j(1)$ is the exchange operator describing the electron exchange energy due to the antisymmetry of the total N-electron wavefunction. Such a single-electron treatment obviously reduces the calculation of a large number of electron interactions, but the Hartree–Fock method still scales poorly with the number of electrons.

Hohenberg–Kohn Theorem and Kohn–Sham Method

Thomas [23] and Fermi [24] proposed the Thomas-Fermi model for the electron distribution in the many-body system based on statistical mechanics in the 1920s. Although this model is semi-classical, it provides an alternative, "density functional" perspective to solve the many-body Schrödinger equation.

From quantum mechanics, the expectation value of the electronic kinetic energy operator in 3-dimensional Euclidean space can be approximated by

$$T = C_{\mathrm{k}} \int [n(\mathbf{r})]^{5/3} \, d\mathbf{r}, \tag{10}$$

where $C_{\mathrm{k}} = \frac{3}{10}(3\pi^2)^{\frac{2}{3}}$. Then assuming electrons are classical particles moving in the external field $V(\mathbf{r})$, then the total energy is

$$E[n] = T \; + \; U_{eN} \; + \; U_{ee} \tag{11}$$

$$= C_{\mathrm{k}} \int [n(\mathbf{r})]^{5/3} \, d\mathbf{r} \; + \int n(\mathbf{r}) V_N(\mathbf{r}) d\mathbf{r} \; + \; \frac{1}{2} \int \frac{n(\mathbf{r}) \; n(\mathbf{r}\,')}{|\mathbf{r} - \mathbf{r}\,'|} \, d\mathbf{r} \; d\mathbf{r}' \tag{12}$$

where $C_{\mathrm{k}} = \frac{3}{10}(3\pi^2)^{\frac{2}{3}}$. Solving for a density $n(\mathbf{r})$ minimizing the total energy gives the ground state electron density.

The Thomas–Fermi model connects the total energy as an electronic density functional. This reduces the many-body problem of N electrons with $3N$ spatial coordinates to three spatial coordinates of the density, which is the basic motivation of the development of density functional theory (DFT). But, of course, since the Thomas–Fermi model applied the uniform electronic density approximation in the kinetic energy term as well as the neglect of exchange and correlation in the potential term, Thomas–Fermi approximation is very inaccurate and the calculation results usually deviate significantly from the real system.

To obtain a better formalism about the energy functional of the electron density, Hohenberg and Kohn proposed two theorems[25, 26]: First, the external potential is the unique functional of the ground state electron density apart from a trivial additive constant, i.e., for a given density the potential is fixed and vice visa, in other words, that the mapping between density and the external potential is bijective. Further, the Hamiltonian and the wavefunctions are fully determined by giving the knowledge of the ground-state density, so the expectation value of any observable as functionals of the ground-state density

$$O_0 = O[n_0] = \langle \Psi_0[n_0] | \, \hat{O} \, | \Psi_0[n_0] \rangle \, . \tag{13}$$

Second, there exists a universal functional of the density $F[n(\mathbf{r})]$ for any number of particles and external potential. Given the external potential $v(\mathbf{r})$, minimizing the energy as functional of density

$$E[n] = \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r} + F[n] \tag{14}$$

with respect to the constraint of the total electron number

$$\int n(\mathbf{r})d\mathbf{r} = N \tag{15}$$

gives the ground state energy. This is the variational principle widely used in DFT. Hohenberg and Kohn overlook two problems. Firstly, it is not clear *a priori* that every well-behaved density is derivable from a well-behaved wavefunction, this is the so-called n-representability problem, which has been solved by Gilbert [27] and Harriman [28].

The second problem is called the v-representability problem, i.e., it is not clear *a priori* that every well-behaved density can be derived from a wavefunction which is the properly ground-state wavefunction for a many-body system under given external potential V. Unfortunately, there *exists* well-behaved $n(\mathbf{r})$ which fail to be ground-state densities for *any* many-body system in an external potential[25, 29–31]. To circumvent this problem Levy [32, 33] and Lieb [34] proposed a definition of the universal functional as

$$F[n] = \min_{\Psi \to n(\mathbf{r})} \langle \Psi | \, \hat{T} + \hat{V} \, | \Psi \rangle \,, \tag{16}$$

with the minimization taken over all well-behaved wavefunctions giving the same density.

The Hohenberg–Kohn theorems guarantee the existence of a universal functional. However, they did not provide the explicit form of this unknown universal functional. Following the spirit of Hohenberg–Kohn theorem, Kohn and Sham provided a powerful method which has been applied with great success, the Kohn–Sham formalism. The key idea in Kohn–Sham is to use an auxiliary non-interacting system whose ground-state density represents the ground-state density of the considered interacting system.

The wavefunction of the auxiliary non-interacting N electrons can be easily written down as the Slater determinant of single-particle orbitals

$$\psi(\mathbf{x_1}, \mathbf{x_2}..., \mathbf{x_N}) = \frac{1}{\sqrt{N!}} \det(\varphi_1, \varphi_2, ..., \varphi_N). \tag{17}$$

And the energy functional of this system is

$$E_{\mathrm{s}}[n] = T_{\mathrm{s}}[n] + \int v_{\mathrm{s}}(\mathbf{r})n(\mathbf{r})d\mathbf{r}. \tag{18}$$

Now for the interacting system, one can re-write Eq. 14 in terms of the auxiliary systems and obtain

$$\begin{aligned} E[n] &= T[n] + V[n] + \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r} \\ &= T[n] - T_{\mathrm{s}}[n] + T_{\mathrm{s}}[n] + V[n] - E_{\mathrm{H}}[n] + E_{\mathrm{H}}[n] + \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r} \\ &= T_{\mathrm{s}}[n] + E_{\mathrm{H}}[n] + E_{\mathrm{xc}}[n] + \int v(\mathbf{r})n(\mathbf{r})d\mathbf{r}. \end{aligned} \tag{19}$$

Here E_{H} is the Hartree (or Coulomb) energy

$$E_{\mathrm{H}}[n] = \frac{1}{2}\int d\mathbf{r}\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}, \tag{20}$$

and E_{xc} the exchange and correlation energy functional

$$E_{\mathrm{xc}}[n] = T[n] - T_{\mathrm{s}}[n] + V[n] - E_{\mathrm{H}}[n]. \tag{21}$$

Applying the variational principle to the energy functional (Eq. 19) with respect to the density of the auxiliary system, we get the well-known Kohn and Sham equation

$$\Big[-\frac{\nabla^2}{2} + v(\mathbf{r}) + v_{\mathrm{H}}[n](\mathbf{r}) + v_{\mathrm{xc}}\left[n\right](\mathbf{r})\Big]\psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}) \tag{22}$$

where v is the external potential, v_{H} is the Hartree potential, and $\psi_i(\mathbf{r})$ are the Kohn–Sham orbitals. The exchange-correlation potential v_{xc} is the the functional derivative of E_{xc}

$$v_{\mathrm{xc}}\left[n\right](\mathbf{r}) = \frac{\delta E_{\mathrm{xc}}\left[n\right]}{\delta n(\mathbf{r})}. \tag{23}$$

Eq 22 is the backbone of DFT calculations, in which this equation can be solved self-consistently if the exchange-correlation potential is known, which is, again, unfortunately, not the case.

Finding a proper approximation of the exchange-correlation potential is indeed one of the most important tasks the DFT community faces. One may follow a pure mathematical path, starting from an exploration of the abstract properties that the universal functional $F[n]$ must have. This path provides practical guidance on explicitly constructing better approximations. However, it requires fundamental

progress in the mathematical understanding of the Hohenberg-Kohn theorems. Moreover, the true universal functional is most likely transcendental, i.e., might be unable to write down exactly in closed form.

Another option to construct functionals relies or numerical algorithms, one often starts at defining an *ansatz* form of v_{xc} with some flexible parameters. Then the parameters are fitted to either experimental data or exact solutions of simple systems (e.g., homogeneous electron gas). Due to the usage of experimental data, some would argue that the fitted functional is not *ab initio* anymore. Of course, even those holding the most restrict standard would agree that a functional only fitted to quantum Monte Carlo (QMC) simulations is "first principle". However sadly, nowadays computational resource is barely enough to produce the required amount of QMC data in a limited time scale.

Exchange-Correlation Functionals

The first and yet the still being commonly used *ansatz* is the local density approximation (LDA), where the exchange-correlation energy is approximated to the exchange-correlation energy $\varepsilon_{\mathrm{xc}}$ (per electron) of homogeneous electrons gas

$$E_{\mathrm{xc}}^{\mathrm{LDA}}[n] = \int \varepsilon_{\mathrm{xc}}[n(\mathbf{r})]n(\mathbf{r})d\mathbf{r}. \tag{24}$$

The exchange part of is

$$E_{\mathrm{x}}^{\mathrm{LDA}}[n] = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}\int \rho(\mathbf{r})^{4/3}\, d\mathbf{r} \tag{25}$$

The correlation part is accurately known from Ceperley and Alder's quantum Monte Carlo calculation [35]. The LDA functional works well (by design) for systems with slow varying (more homogeneous) densities, but poorly for finite systems with abrupt density gradients. To correct this behaviour, it is natural to consider not only the density but also the density gradient (or even higher order of derivatives) in the exchange-correlation functional, and reach the generalized gradient approximation (GGA) *ansatz*

$$E_{\mathrm{xc}}^{\mathrm{GGA}}[n] = \int \varepsilon_{\mathrm{xc}}(n, \nabla n)n(\mathbf{r})d\mathbf{r}. \tag{26}$$

Benefiting from the inclusion of gradient, GGA functionals are semi-local and work fairly in the majority of the real systems. Moreover, to include more of the non-locality

nature of the exchange-correlation, one can also include the derivative of the Kohn–Sham orbitals and construct the meta-GGA functionals depending on the Laplacian of the density ($\nabla^2 n$) and kinetic density (τ)

$$\begin{aligned} E_{\mathrm{xc}}^{\text{meta-GGA}}[n] &= \int \varepsilon_{\mathrm{xc}}(n, \nabla n, \tau, \nabla^2 n) n(\mathbf{r}) d\mathbf{r} \\ \tau(\mathbf{r}) &= \frac{1}{2} \sum_{i}^{\mathrm{occ}} \left| \nabla \psi_i(\mathbf{r}) \right|^2 . \end{aligned} \tag{27}$$

Although the (meta-)GGA functionals provide a practical route to correct the LDA and produce better results, the physics underlying these gradient corrections is not yet fully understood, the consequence is the lack of a systematic procedure for improving. The progress of (meta-)GGAs relies heavily on physical intuition, choice of constraining relationships, and simple trial and error. To partially solve this problem, one of the ideas is to treat exchange exactly, due to the fact that exchange is the dominant part of the exchange-correlation energy. The small correlation part could then be approximated. This idea leads to the so-called hybrid functionals, in which the total exchange energy is expressed as a linear combination of exact Hartree–Fock exchange and exchange of other (semi-)local DFT functionals

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{hyb}} &= \alpha E_{\mathrm{x}}^{\mathrm{HF}} + (1-\alpha) E_{\mathrm{x}}^{\mathrm{DFT}}[n] + E_{\mathrm{c}}^{\mathrm{DFT}}[n] \\ E_{\mathrm{x}}^{\mathrm{HF}} &= -\frac{1}{2} \int \int \psi_i^*(\mathbf{r}_1) \psi_j^*(\mathbf{r}_2) \frac{1}{r_{12}} \psi_j(\mathbf{r}_1) \psi_i(\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2, \end{aligned} \tag{28}$$

where α determines the relative amount of Hartree–Fock and semi-local exchange, for example, the PBE0 [36] functional use 1/4 of the exact and 3/4 of the Perdew–Burke-Ernzerhof (PBE) GGA exchange. Furthermore, one can also define different α according to the range (called screened or range-separated hybrids)

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{hyb}} &= \alpha E_{\mathrm{x}}^{\mathrm{HF,SR}}[\mu] + (1-\alpha) E_{\mathrm{x}}^{\mathrm{DFT,SR}}[\mu, n] \\ &\quad + E_{\mathrm{x}}^{\mathrm{DFT,LR}}[\mu, n] + E_{\mathrm{c}}^{\mathrm{DFT}}[n] \\ E_{\mathrm{x}}^{\mathrm{HF,SR}}[\mu] &= -\frac{1}{2} \int \int \psi_i^*(\mathbf{r}_1) \psi_j^*(\mathbf{r}_2) \frac{\mathrm{erfc}(\mu r_{12})}{r_{12}} \psi_j(\mathbf{r}_1) \psi_i(\mathbf{r}_2)\, d\mathbf{r}_1\, d\mathbf{r}_2, \end{aligned} \tag{29}$$

where erfc is the complementary error function, and μ is the parameter that defines the range separation

$$\frac{1}{r} = S_\mu(r) + L_\mu(r) = \frac{\mathrm{erfc}(\mu r)}{r} + \frac{\mathrm{erf}(\mu r)}{r}. \tag{30}$$

The hybrid functionals give more accurate results than density-only semi-local functionals. However, there is a trade off for this improvement to computational costs, because the HF exchange integral scales very poorly with the system sizes. Further improvement can be achieved by including perturbative second-order (PT2) correction to the correlation, leading to even more accurate double-hybrid functionals

$$E_{\mathrm{xc}}^{\mathrm{hybrid}} = \alpha_{\mathrm{x}} E_{\mathrm{x}}^{\mathrm{HF}} + (1-\alpha_{\mathrm{x}}) E_{\mathrm{x}}^{\mathrm{DFT}} + \alpha_{\mathrm{c}} E_{\mathrm{c}}^{\mathrm{PT2}} + (1-\alpha_{\mathrm{c}}) E_{\mathrm{c}}^{\mathrm{DFT}}. \tag{31}$$

Choice of Functionals

The majority of the DFT calculations in this Thesis use the Perdew–Burke–Ernzerhof [37], a GGA functional, with the exchange part as

$$\begin{aligned} E_{\mathrm{x}}^{\mathrm{PBE}}[n_\uparrow, n_\downarrow] &= \frac{1}{2}\left(E_{\mathrm{x}}^{\mathrm{PBE}}[2n_\uparrow] + E_{\mathrm{x}}^{\mathrm{PBE}}[2n_\downarrow]\right) \\ E_{\mathrm{x}}^{\mathrm{PBE}}[n] &= \int n \varepsilon_{\mathrm{x}}^{\mathrm{HEG}}(n) F_x(s) d\mathbf{r} \\ F_x(s) &= 1 + \kappa(1 - \frac{1}{1+\mu s^2/\kappa}), \end{aligned} \tag{32}$$

where $\varepsilon_{\mathrm{x}}^{\mathrm{HEG}}(n)$ is the exchange energy for the homogeneous electron gas (HEG) with density n

$$\begin{aligned} \varepsilon_{\mathrm{x}}^{\mathrm{HEG}}(n) &= -\frac{3k_{\mathrm{F}}}{4\pi} \\ k_{\mathrm{F}} &= (3\pi^2 n)^{1/3} \\ s &= \frac{|\nabla n|}{2k_{\mathrm{F}} n}. \end{aligned} \tag{33}$$

The parameters $\kappa = 0.804$ and $\mu = 0.2195$ are determined by ensuring 1) exchange gradient correction cancels correlation in low gradient limit $s \to 0$ and 2) the local Lieb-Oxford bound[38]

$$\begin{aligned} &1)\ F_{\mathrm{x}} \to \quad 1 + \mu s^2, \text{ as } s \to 0 \\ &2)\ E_{\mathrm{x}} \geq \quad -1.679 \int d\mathbf{r} n(\mathbf{r})^{4/3}. \end{aligned} \tag{34}$$

And the correlation part is

$$
\begin{aligned}
E_{\mathrm{c}}^{\mathrm{PBE}}[n] &= \int n(\mathbf{r})\{\varepsilon_{\mathrm{c}}^{\mathrm{HEG}}(r_s,\zeta)+H(t,r_s,\zeta)\}d\mathbf{r} \\
H(t,r_s,\zeta) &= \gamma\phi^3 \ln\left\{1+\frac{\beta}{\gamma}t^2\left[\frac{1+At^2}{1+At^2+A^2t^4}\right]\right\} \\
A &= \frac{\beta}{\gamma}\left[e^{-\varepsilon_{\mathrm{c}}^{\mathrm{HEG}}/(\gamma\phi^3)}-1\right]^{-1}
\end{aligned}
\tag{35}
$$

where

$$
\begin{aligned}
r_s &= (\frac{3}{4\pi n})^{1/3} \\
\zeta &= \frac{(n_\uparrow - n_\downarrow)}{n} \\
t &= \frac{|\nabla n|}{2k_s\phi n} \\
\phi &= \frac{1}{2}[(1+\zeta)^{2/3}+(1-\zeta)^{2/3}] \\
k_s &= (\frac{4k_{\mathrm{F}}}{\pi})^{1/2},
\end{aligned}
\tag{36}
$$

with the parameters $\gamma = 0.031\ 091$ and $\beta = 0.066\ 725$. The gradient correlation part H fulfill the bounding conditions: 1) in the slowly varying limit H is given by its second-order gradient expansion [39], 2) in rapid varying limit correlation vanishes [37], and 3) under uniform scaling to high-density limit the correlation energy scale to a constant [40]

$$
\begin{aligned}
&1)\quad H \to \beta\phi^3 t^2, \text{ as } t \to 0 \\
&2)\quad H \to -\varepsilon_{\mathrm{c}}^{\mathrm{HEG}}, \text{ as } t \to \infty \\
&3)\quad E_{\mathrm{c}} \to \text{const}, \text{ with } n(\mathbf{r}) = \lambda^3 n(\lambda\mathbf{r}) \text{ as } \lambda \to \infty.
\end{aligned}
\tag{37}
$$

The construction and parameterization of PBE functional do not depend on fitting to experimental data, which makes PBE *ab initio*, and the computational cost is relatively cheap. Unfortunately, the PBE errors in formation enthalpy are usually around 200 meV/atom [41], comparatively larger than the chemical accuracy in formation enthalpy (40 meV/atom). The PBE, as a GGA functional, also suffers from the failure of exact cancellation of self-interaction and not properly describing the derivative discontinuity [42], so the calculated band gaps, one of the most important properties of materials, are around 60 % of the experiment values [43, 44].

Hybrid functionals can give better results on the gaps [44]. We chose the Heyd–Scuseria–Ernzerhof–06 (HSE06) version, which belongs to the range-separated hybrid functionals (Eq. 29). The range separation parameter μ for HSE06 is 0.2 [45],

and the portion of HF exact exchange (α) is 1/4 [45], the same as PBE0 [36]. The HSE06 functional gives excellent results on the band gaps, with mean absolute percentage error (MAPE) around 29 %, improved from 40 % for PBE. However, the calculation cost is too high for HT screen (quartic scaling with electron number), so it is only selectively used to produce better band structures in this Thesis.

Another similar accurate functional to calculate band gaps is the modified-Becke-Johnson (mBJ) functional [46, 47]. The mBJ belongs to the meta-GGA family, so is much cheaper than the hybrid functionals. However, unlike PBE and HSE06, mBJ is a potential-only functional, i.e., the functional is constructed from defining the exchange-correlation potential

$$\begin{aligned} v_{\mathrm{x}}^{\mathrm{MBJ}}(\mathbf{r}) &= c v_{\mathrm{x}}^{\mathrm{BR}}(\mathbf{r}) + (3c-2)\frac{1}{\pi}\sqrt{\frac{5}{6}}\sqrt{\frac{\tau(\mathbf{r})}{n(\mathbf{r})}} \\ v_{\mathrm{x}}^{\mathrm{BR}}(\mathbf{r}) &= -\frac{1}{b(\mathbf{r})}\left(1 - e^{-x(\mathbf{r})} - \frac{1}{2}x(\mathbf{r})e^{-x(\mathbf{r})}\right), \end{aligned} \tag{38}$$

where $n(\mathbf{r})$ is electron density and $\tau(\mathbf{r})$ is the kinetic density as defined in Eq. 27, and

$$\begin{aligned} b(\mathbf{r}) &= \left[\frac{x(\mathbf{r})^3 e^{-x(\mathbf{r})}}{8\pi n(\mathbf{r})}\right]^{1/3} \\ c &= \alpha + \beta\left(\frac{1}{V_{\mathrm{cell}}}\int_{\mathrm{cell}}\frac{|\nabla n(\mathbf{r}')|}{n(\mathbf{r}')}d\mathbf{r}'\right)^{1/2}. \end{aligned} \tag{39}$$

The parameters $\alpha = -0.012$ and $\beta = 1.023$ bohr$^{1/2}$ are determined by fitting to band gaps of a set of solids [47]. The $x(\mathbf{r})$ in the above equation satisfies the following equation [48]

$$\begin{aligned} \frac{xe^{-2x/3}}{x-2} &= \frac{2\pi^{2/3}}{3}\frac{n^{5/3}}{Q} \\ Q &= \frac{1}{6}(\nabla^2 n - 2\gamma D) \\ D &= \tau - \frac{1}{4}\frac{(\nabla n)^2}{n}, \end{aligned} \tag{40}$$

where γ is a parameter having a value of one from the Taylor expansion of exact spherically averaged LDA exchange hole potential. However, a value of 0.8 is better to recover the LDA [48].

Plane Wave Basis Set and Pseudopotentials

The Kohn–Sham equations are nonlinear eigenvalue problems, which are usually solved by the self-consistent method, in which it is convenient to expand the wavefunctions

in Eq. 22 in a basis set:

$$\psi_i(\mathbf{r}) = \sum_\alpha \mathrm{c}_\alpha \varphi_\alpha(\mathbf{r}). \tag{41}$$

For finite systems, it is natural to choose an atomic orbital basis set origin from the chemical intuition that molecular orbitals could be seen as linear combinations of atomic orbitals. The atomic orbitals take the form of

$$\varphi_\alpha(\mathbf{r}) = \varphi_\alpha(\mathrm{r}) Y_{lm}(\Theta, \phi), \tag{42}$$

where $\varphi_\alpha(\mathrm{r})$ could either be Gaussian type $e^{-\alpha r^2}$ or Slater type $e^{-\alpha r}$ functions. A *relatively* small atomic orbital basis set is able to provide fairly good results, but it is non-orthogonal and suffers from the basis set superposition error (BSSE) [49]. For solids with the periodic boundary condition, we have the Bloch's theorem for the periodic wave functions

$$\varphi_{n\mathbf{k}}(\mathbf{r} + \mathbf{R}) = \varphi_{n\mathbf{k}}(\mathbf{r}) e^{i\mathbf{k}\mathbf{R}}, \tag{43}$$

where n denotes the n-th one-electron band, $\mathbf{k}$ is the so-called Bloch wave vector, and $\mathbf{R}$ represents translational vectors keeps the Hamiltonian invariant. We can expand the periodic function $\varphi_{n\mathbf{k}}$ in plane waves

$$\varphi_{n\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} C_{\mathbf{G}n\mathbf{k}} e^{i(\mathbf{G}+\mathbf{k})\cdot\mathbf{r}}, \tag{44}$$

where $\mathbf{G}$ is the reciprocal lattice vector. Plane waves are naturally orthonormal and can efficiently transform into real space with fast Fourier transforms. In practice, the plane wave basis set is finite by setting cutoff energy and truncating the plane wave expansion to only include those having $|\mathbf{G} + \mathbf{k}|^2 < 2E_{\text{cutoff}}$. This also provides a convenient way for convergence test on the basis size, i.e., simply via systematically increasing E_{cutoff}. However, as the wavefunctions of the electrons in the *core* region of atoms oscillate rapidly, their expansion requires an enormous amount of plane waves. This substantially increases the computational cost and hinders the application of plane wave basis sets. One way to solve this problem is to expand the core electrons in atomic basis sets and the valence ones in plane waves. This is the general idea of the full-potential linearized augmented-plane-wave method (FLAPW). A further step is to replace the Coulomb potential between nuclei and core electrons with an effective pseudopotential. Then smoother core electron wavefunctions without nodes will require fewer plane waves to describe.

Vanderbilt pioneered the development of the ultrasoft pseudopotential in 1990[50], which has great advantages when dealing with $3d$ transition metals and rare earth

metals. The drawback of the method is that all information on the full wavefunction close to the nuclei is lost. The development of the projective augmented wave (PAW) method [51, 52] combines the idea of the pseudopotential approximation and the full-potential linearized augmented-plane-wave method. In PAW the one electron wavefunctions $\psi_{n\mathbf{k}}$ are derived from the pseudo wavefunction $\widetilde{\psi}_{n\mathbf{k}}$ by linear transformation

$$\begin{aligned}\psi_{n\mathbf{k}} &= \widetilde{\psi}_{n\mathbf{k}} + \sum_i \sum_a (\phi_i^a - \widetilde{\phi}_i^a)\langle \widetilde{p}_i^a | \widetilde{\psi}_{n\mathbf{k}} \rangle \\ &= \widetilde{\psi}_{n\mathbf{k}} + \sum_a \left(\psi_{n\mathbf{k}}^a(\mathbf{r} - \mathbf{R}^a) - \widetilde{\psi}_{n\mathbf{k}}^a(\mathbf{r} - \mathbf{R}^a) \right)\end{aligned} \tag{45}$$

The summation index i is the contracted index label containing the atomic site $\mathbf{R}$, the angular momentum numbers l, m, and an additional index k referring to the reference energy ϵ_{kl}, and the a is the index for atoms. The true partial waves ϕ_i are solutions of the radial Schrödinger equation for a non-spinpolarized reference atom at reference energy ϵ_{kl} for a specific angular momentum l. The pseudo partial waves $\widetilde{\phi}_i$ are equivalent to the true partial waves outside a cutoff radius r_c^l while continuously matching onto it in side r_c^l and expanded to plane waves

$$\langle \mathbf{r} | \widetilde{\psi}_{n\mathbf{k}} \rangle = \frac{1}{\sqrt{\Omega}} \sum_{\mathbf{G}} C_{n\mathbf{k}\mathbf{G}} e^{i(\mathbf{G}+\mathbf{k})\cdot\mathbf{r}}. \tag{46}$$

The $\widetilde{p}_i$ is the projector dual to partial waves $\langle \widetilde{p}_i | \widetilde{\phi}_j \rangle = \delta_{ij}$. Eq. 45 separates the true wavefunction into two parts, the first part is the smooth pseudo waves, and the second part is the rapid osculating part centered at each atom within a certain cutoff radius (PAW sphere). The two parts can be treated individually. The smooth pseudo wave can be easily treated using plane wave basis set (Eq. 46) and for the PAW sphere part of an atom-centered radial grid can be efficiently applied.

In practice, the PAW method can further be combined with the frozen core (FC) approximation. FC supposes that the core electrons do not participate in the bonding between atoms. Therefore, only the valence electrons are included in Eq 45, and the core electrons are fixed during calculations. Any expectation value of operator $\widehat{O}$ is then

$$\langle \widehat{O} \rangle = \sum_n^{\text{val}} f_n \langle \psi_n | \widehat{O} | \psi_n \rangle + \sum_a \sum_\alpha^{\text{core}} \langle \phi_\alpha^{a,\text{core}} | \widehat{O} | \phi_\alpha^{a,\text{core}} \rangle , \tag{47}$$

The second term is trivial, and for the first term for each index n using Eq 45, we have

$$\begin{aligned}\langle\psi|\,\widehat{O}\,|\psi\rangle &= \langle\widetilde{\psi}+\sum_a(\psi^a-\widetilde{\psi}^a)|\,\widehat{O}\,|\widetilde{\psi}+\sum_a(\psi^a-\widetilde{\psi}^a)\rangle \\ &= \langle\widetilde{\psi}|\,\widehat{O}\,|\widetilde{\psi}\rangle + \sum_a\Big(\langle\psi^a|\,\widehat{O}\,|\psi^a\rangle - \langle\widetilde{\psi}^a|\,\widehat{O}\,|\widetilde{\psi}^a\rangle\Big) \\ &\quad + \sum_a\Big(\langle\psi^a-\widetilde{\psi}^a|\,\widehat{O}\,|\widetilde{\psi}-\widetilde{\psi}^a\rangle + \langle\widetilde{\psi}-\widetilde{\psi}^a|\,\widehat{O}\,|\psi^a-\widetilde{\psi}^a\rangle\Big) \\ &\quad + \sum_{a\neq a'}\langle\psi^a-\widetilde{\psi}^a|\,\widehat{O}\,|\psi^{a'}-\widetilde{\psi}^{a'}\rangle\,.\end{aligned} \tag{48}$$

For local operators, the last summation is zero because the PAW spheres do not overlap for atoms on different sites, and the second last summation is also zero because $|\widetilde{\psi}-\widetilde{\psi}^a\rangle$ is only non-zero outside the PAW sphere while $\langle\psi^a-\widetilde{\psi}^a|$ is only non-zero inside. From the above equation, it is possible to define the pseudo operator for a local operator

$$\widehat{O}^{\rm ps} = \widehat{O} + \sum_{i,j}|\widetilde{p}_i\rangle\left(\langle\phi_i|\,\widehat{O}\,|\phi_j\rangle - \langle\widetilde{\phi}_i|\,\widehat{O}\,|\widetilde{\phi}_j\rangle\right)\langle\widetilde{p}_j|\,, \tag{49}$$

so that the expectation value can be evaluated as $\langle\psi|\,\widehat{O}\,|\psi\rangle = \langle\widetilde{\psi}|\,\widehat{O}^{\rm ps}\,|\widetilde{\psi}\rangle$. For example, the pseudo density operator gives the density

$$\langle\psi|\mathbf{r}\rangle\,\langle\mathbf{r}|\psi\rangle = |\widetilde{\psi}|^2 + \sum_{i,j}\langle\widetilde{\psi}|\widetilde{p}_i\rangle\left(\langle\phi_i|\mathbf{r}\rangle\,\langle\mathbf{r}|\phi_j\rangle - \langle\widetilde{\phi}_i|\mathbf{r}\rangle\,\langle\mathbf{r}|\widetilde{\phi}_j\rangle\right)\langle\widetilde{p}_j|\widetilde{\psi}\rangle \tag{50}$$

We can define the on-site density matrix $D^a_{ij} = \langle\widetilde{\psi}|\widetilde{p}_i\rangle\,\langle\widetilde{p}_j|\widetilde{\psi}\rangle$, then

$$\langle\psi|\mathbf{r}\rangle\,\langle\mathbf{r}|\psi\rangle = \widetilde{n}(\mathbf{r}) - \widetilde{n}^a(\mathbf{r}) + n^a(\mathbf{r}), \tag{51}$$

where

$$\begin{aligned}\widetilde{n}(\mathbf{r}) &= \langle\widetilde{\psi}|\mathbf{r}\rangle\,\langle\mathbf{r}|\widetilde{\psi}\rangle = |\widetilde{\psi}|^2 \\ \widetilde{n}^a(\mathbf{r}) &= \sum_{i,j}D^a_{ij}\widetilde{\phi}^a_i(\mathbf{r})\widetilde{\phi}^a_j(\mathbf{r}) \\ n^a(\mathbf{r}) &= \sum_{i,j}D^a_{ij}\phi^a_i(\mathbf{r})\phi^a_j(\mathbf{r})\end{aligned} \tag{52}$$

Here $\widetilde{n}$ is the pseudo density, $\widetilde{n}^a$ is the on-site charge density, and n^a is the true on-site density. The superscription a notes that the on-site density is only evaluated on radial grids centered at each atom.

Similarly, for kinetic energy operator, which is semi-local, we have the kinetic energy

$$\langle\psi|-\frac{\nabla^2}{2}|\psi\rangle = \widetilde{E}_{\rm kin} - \widetilde{E}^a_{\rm kin} + E^a_{\rm kin}, \tag{53}$$

where

$$\begin{aligned}
\widetilde{E}_{\text{kin}} &= \langle\widetilde{\psi}| - \frac{\nabla^2}{2}|\widetilde{\psi}\rangle \\
\widetilde{E}^a_{\text{kin}} &= \sum_{i,j} D^a_{ij} \langle\widetilde{\phi}^a_i| - \frac{\nabla^2}{2}|\widetilde{\phi}^a_j\rangle \\
E^a_{\text{kin}} &= \sum_{i,j} D^a_{ij} \langle\phi^a_i| - \frac{\nabla^2}{2}|\phi^a_j\rangle \,.
\end{aligned} \tag{54}$$

However, the pseudo-wavefunctions do not have the same norm as the true all electron wavefunctions inside the spheres, so for a non-local operator, it is necessary to introduce a compensation charge density (so-called augmentation density) $\hat{n}$. The augmentation density corrects the moments of the pseudo electron density within the paw sphere centered at atom position $\mathbf{R}_a$ to the true all electron density n^a. Thus the electrostatic potential from n^a is identical to that from $\widetilde{n}^a + \hat{n}$ outside the PAW sphere. With the compensation density included, one can finally write the total energy functional as

$$E = \widetilde{E} + \sum_a \Delta E^a, \tag{55}$$

where

$$\widetilde{E} = \widehat{T}^{\text{ps}}[\{\widetilde{\psi}_n\}] + V_H[\widetilde{n} + \sum_a \hat{n}] + E_{xc}[\widetilde{n}], \tag{56}$$

is the smooth part. And ΔE^a is the correction term computed inside each PAW sphere using a radial grid instead of a plane wave grid. This is the key reason that the PAW speeds up the calculations.

Chapter 2 High-Throughput and Machine Learning

The idea of high-throughput (HT) originated in the 1950s when experimental biologists attempted to automatize micro titration. Such technology eventually allows scientists to systematically perform thousands of hemagglutination inhibition tests in vaccination research[53]. HT experiments have become standard practice in biology laboratories worldwide, but in chemistry and material science, the experimental establishment and implementation of HT methods are still at the early stage [54–56].

On the other hand, the computational power of modern CPUs, robust computational codes, and efficient quantum-mechanical methods continuously improved, allowing researchers to calculate the thermodynamic and electronic structures of many systems at a reasonable efficiency-accuracy balance. Furthermore, with the parameters for these calculations chosen based on unified standards and convergence criteria, all calculations made by the research community can be accumulated to construct large computational databases. These databases often include not only existing experimental structures but also hypothetical systems. Moreover, we can filter out promising candidates for desired functionalities by mining from these databases. Such a process based on large datasets of theoretical calculations, i.e., the so-called high-throughput computational search, has gradually become one of the most efficient and fruitful schemes of discovering new functional materials [WPhD1, WPhD12, 12–17, 57, 58].

High-Throughput Strategies

Ideally, one should sample the structural and compositional spaces in a single high-throughput search. Although experimental realizations of this vision already exist [59], theoretically (or computationally), it is still impractical. Given the number of elements (N_e) and atoms N_a, for the structural space, the degree of freedom is $3N_\mathrm{a}-3$. Moreover, in the compositional space, the number of possible stoichiometries scales factorially with N_e. Therefore, the number of possible candidate crystal structures explodes combinatorially with increasing N_e and N_a. One feasible way to avoid the explosion is to "decouple" the search into structural and compositional parts.

The search in the former is mainly called (global) structure prediction, where one searches for energy minima for given compositions. The energy surface can be sampled by molecular dynamics (MD) (e.g., simulated annealing). Nowadays, modern CPUs

can afford *ab initio* MD simulation for small systems. However, one composition might require millions of MD steps to locate the global minimum finally. Therefore, applying MD to *every* composition is too expensive to find optimal candidates for real applications. Using an empirical force field for MD simulations would vastly accelerate the search, trading off the accuracy. Machine learning (ML) techniques can be applied to generate a force field nearly as efficiently as empirical force fields and as accurately as first-principle methods. However, a large amount of energy surface sampling is necessary to train the machine learning force field. Of course, to avoid the usage of a force field, one could also apply Monte Carlo (MC) statistic instead of MD to guide the sampling of the energy landscape. Unfortunately, no matter in MC or MD prediction algorithms, due to the Boltzmann distribution of energies, most samples are confined around local minima, and walking across high barriers to reach another minimum is very time intensive. To solve this problem, one could use sample algorithms that do not include the Boltzmann factor but do depend on the local structure of the energy landscape. Those algorithms are often called hopping methods (e.g., minimal hoping [60]). Furthermore, the efficiency of structure prediction can be improved by an evolutionary algorithm (e.g., genetic algorithm) which could simultaneously explore the compositional space. Although structure predictions are still a highly active field in recent years, and has been proven successful[12, 61, 62], this approach of fixing the composition and searching the structural space is intrinsically not efficient for a task dealing with a large number of compositions. For example, there will be $78 \times 77 \times 76 \approx 456,000$ total possible stoichiometries for a simple ternary formula ABC_3, when putting 78 chemical elements (from hydrogen to bismuth excluding noble gases) into A, B, and C sites. Searching the structural space for each of them *will eventually* provide a complete energy landscape for any ABC_3 ternary systems if the computational cost *is not* approaching infinite.

On the other hand, searching the entire structural space for all 456k combinations of the formula ABC_3 is unnecessary. Because "heuristically", we know many ABC_3 compositions crystalize in the perovskite structure (Fig. 1). Therefore, it is reasonable to prioritize the exploration of the perovskite structural sub-space, i.e. the perovskite prototype, for ABC_3 compositions. By limiting structural space in this way, one can scan the compositional space for each prototype. This method is known as the component prediction or prototype search [63]. Prototype search takes advantage of the results of a structural search from nature. One might argue that this method confines the exploration away from the unknown part of the structural space.

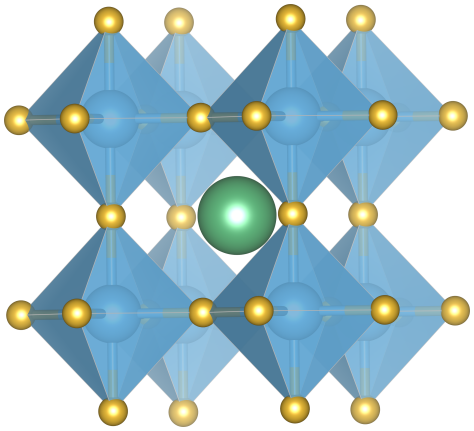


Figure 1: Crystal structure of perovskite prototype ABC_3, teal spheres are the A atoms, the BC_6 groups are shown as blue octahedra where the golden spheres on the corners of the octahedra are C atoms.

However, the limited structural space is still enormously large, considering that the number of available prototypes is at a magnitude of 10^3 [11], 10^4 [64], or even 2×10^4 [WPhD12] based on different classification standards. Although to the author's best knowledge, the most exhaustive prototype search has searched 2500 binary and ternary prototypes [WPhD12], exploration of the rest remains a challenge.

Pre-Filtering

Following the basic ideas from prototype search, i.e., taking advantage of the experimental (or natural) results, we can improve the efficiency of high-throughput searches.

The early attempts used well-known chemistry and physics intuitions to pre-select the more stable compositions. For example, one can consider the Goldschmidt tolerance factor for the perovskite (ABO_3) (Eq 57) [65].

$$t = \frac{r_A + r_O}{\sqrt{2}(r_B + r_O)} \tag{57}$$

However, studies have found that t is not a good descriptor for perovskite stability [66, 67]. Other empirical rules, such as Pauling's for ionic crystals [68] and the 18-electron rule for ABX compounds [69], could also be used. Another intuition is the oxidation state neutrality or charge neutrality, i.e., the summation of the oxidation state of the ions in a crystal should be zero to avoid an infinite electrostatic potential. With the possible oxidation state for a given element known from experiments and the octet rule

of the electron configuration, one can apply these criteria to reduce the number of total candidates. For example, the total number of compositions can be shrunk by at least an order of magnitude, increasing the efficiency of prototype-based high-throughput search [58].

There should also be a fundamental constraint, even if the oxidation states fulfill the charge neutrality rule, i.e., the electronegativity (χ) balance rule. The χ scale of an element represents the attraction of its atoms for electrons, so the order of oxidation states (from most negative to most positive) for the elements in the stoichiometry should follow the same of χ from largest to smallest. In the work of Davies *et al.* [58], using the Pauling χ scale, this constraint could significantly reduce the workload of DFT calculations, with only a fourth to a tenth of the compositions being electronegativity balanced.

Searching which composition could be potentially stable for a prototype is equivalent to searching the thermodynamically allowed replacements for elements in this prototype structure. This substitutional question is well investigated in metallurgy and leads to the heuristic Hume-Rothery rule [70]. Besides explicit chemistry and physics principles, one could also try to make a scatter point plot for all known crystal structures in all compositions. For example, the coordinates of the points for all composition AB crystallizing in a specific structural prototype (e.g., NaCl-type) are calculated from some physical properties of **A** and **B** supposed to crucially decide the stability of NaCl-structured-AB. Such a plot is called a structure map. In an ideal structure map, the region of each prototype is separated from the rest, leaving no overlap between distributions. The substitution of **A** by **C** would shift the coordinates of the hypothesis $A_{1-x}C_xB$ system in the AB structure map. By tracing which prototype region it lands in, one could predict the structures for a series of substitution concentrations x. From this point of view, we can also describe such structure maps as the empirical approach to the structure prediction mentioned above. Of course, the separation of prototypes highly depends on the choice of physical factors. One of the most sophisticated choices [71–74] is done by Villars [74], where a three-dimension map is plotted on axes of electronegativity difference ($\Delta\chi$), the atomic radius difference (ΔR), and the number of valence electrons per atom $\overline{N_e}$. However, high-dimensional structure maps are difficult to visualize, and the separation is somehow dissatisfying. In 1984 Pettifor [75] proposed an elegant solution to those problems. His solution does not come from theoretical consideration of the physics properties of the elements

but is entirely phenomenological. Pettifor's idea is that the elements crystallizing into similar structures are also chemically similar, so they should be grouped. The Pettifor's scale (χ^{P}) is constructed by reordering the periodic table one-dimensionally, only considering this similarity. Plotting the A_xB_y structure maps on axes of $\chi^{\mathrm{P}}_{\mathrm{A}}$ and $\chi^{\mathrm{P}}_{\mathrm{B}}$, Pettifor achieved near-perfect separation [76, 77].

The Pettifor scale is based on structures of hundreds of binary systems, and the similarity governing the formation of these crystals is extracted manually. It is possible, though, to construct mathematically such a scale [78]. Generally, for a set of M prototype structures $\{S^0, S^1, \dots S^M\}$ and a set of N compositions $\{c_0, c_1, \dots, c_N\}$, one can define the probability density of compositions crystallizing into prototypes $\{\mathbf{X} = S^i_{c_\alpha} | i \in M; \alpha \in N\}$ [78]

$$\begin{aligned} p(\mathbf{X}) &= \frac{1}{Z}\prod_{i=1}^{N} p(S^i) \prod_{\alpha<\beta;j<k} f(S^j_\alpha, S^k_\beta) \\ &= \frac{1}{Z}\prod_{i=1}^{N} p(S^i) \prod_{\alpha<\beta;j<k} \frac{p(S^j_{c_\alpha}, S^k_{c_\beta})}{p(S^j_{c_\alpha})p(S^k_{c_\beta})} \end{aligned} \tag{58}$$

The single probability $p(S^i)$ captures variations in $p(\mathbf{X})$ due to the independence of probability for prototype i appearing in a database, and the $f(S^j_\alpha, S^k_\beta)$ terms reproduce the correlation between pairs of composition (c_α, c_β) crystallizing into a pair of prototypes (j, k). Using an existing database to calculate $p(\mathbf{X})$ allows the prediction of the probability of the structure of hypothesis composition $\mathbf{X}'$. Furthermore, if we only consider two compositions α and β representing the stoichiometries before and after some substitution of element $\mathbf{A}$ by $\mathbf{B}$, the probability for α and β being stable in all possible prototypes can be calculated from $p(\mathbf{X})$. This probability represents how likely this substitution can happen, i.e., how similar elements $\mathbf{A}$ and $\mathbf{B}$ are similar. However, the similarity (Eq 58) is not straightforward because the probability density is based on prototype-wise vectors, not element-wise ones.

A more intuitive approach [18] to compute the similarity between elements could start from considering directly the likelihood of the substitution between elements $\mathbf{A}$ and $\mathbf{B}$ preserving the crystal structure. Often elements of similar chemistry and physics properties can dope and replace each other in the host lattice. The doping, which usually substitutes a fraction of one specie with another, can sometimes be extended to completely replacing one with another. Consequently, we can define replaceability as analog to dopability. The replaceability can be seen approximated by the reaction

energy for **A** replacing **B** in the host lattice (labeled as L) under consideration,

$$R_{\mathrm{L}}^{\mathbf{A}\to\mathbf{B}} := E[\mathrm{Comp}(\mathbf{A})]-E[\mathrm{Comp}(\mathbf{B})]+E[\mathrm{Elementary}(\mathbf{B})]-E[\mathrm{Elementary}(\mathbf{A})], \quad (59)$$

and a probability distribution can be drawn,

$$P^{\mathbf{A}\leftrightarrow\mathbf{B}}(E) := \sum_{L} \delta(E - R_{\mathrm{L}}^{\mathbf{A}\to\mathbf{B}}), \quad (60)$$

if provided complete thermodynamic pictures for *all* L. We can then calculate the probability that a reaction can happen in infinite time to a threshold (t) to represent the similarity of elements **A** and **B**,

$$S(\mathbf{A}\leftrightarrow\mathbf{B}) := \frac{\int_{-\infty}^{t} P^{\mathbf{A}\to\mathbf{B}}(E)dE}{\int_{-\infty}^{+\infty} P^{\mathbf{A}\to\mathbf{B}}(E)dE}. \quad (61)$$

The numerator can count thermodynamically feasible reactions under threshold condition t. Nature has done a perfect job with t around room temperature. However, the experimental data is far less than complete but highly biased towards the more abundant elements and prefer some lattices over others due to research interests. Even so, if we still think the information is statistically significant, proper normalization can be introduced to solve this problem,

$$S(\mathbf{A}\leftrightarrow\mathbf{B}) = \sqrt{\frac{S(\mathbf{A}\leftrightarrow\mathbf{B})^2}{(\sum_{\mathbf{A}'} S(\mathbf{A}'\leftrightarrow\mathbf{B})(\sum_{\mathbf{B}'} S(\mathbf{A}\leftrightarrow\mathbf{B}')}}, \quad (62)$$

The obtained probability matrix $\mathbf{S}[A:B]$ can be further reordered to maximize the diagonal character, and the order of the elements (i.e., rows or columns) can be seen as an analog to the Pettifor scale to group the similar elements together [18].

If the structural information in the database is *complete*, the probability is then the actual similarity between elements. However, we must again emphasize that the sampling across the chemical space is biased. Unfortunately, the extent of bias is still unknown because of lacking systematic experimental high-throughput benchmark. Nevertheless, we can see the well-known chemistry knowledge in such a data-mining Pettifor scale [18]. For example, similarities within each periodic table group, within the lanthanides, and between the diagonal pairs(Be/Al, B/Si, etc.) are presented [18]. Therefore, the $S(\mathbf{A}\leftrightarrow\mathbf{B})$ deviates from the actual probability but is still a reasonable estimation if we assume the information is statistically significant. In Chapter 5, we will show that applying the similarity scale to pre-filtering the hypothesis systems could primarily increase the success rate of high-throughput search.

Machine Learning

Even with the help of these pre-filtering methods mentioned above, performing millions of DFT calculations to search for promising candidates is still akin to finding a needle in a haystack. Therefore we need more efficient pre-filtering methods, and one of the most successful tools to tackle this challenge is machine learning (ML).

The history of ML could be traced back to 1943 when logician Walter Pitts and neuroscientist Warren McCulloch published their mathematical model of neural networks responsible for the cognition of human brains [79]. After decades of development, with the advent of big data and high-performance computing, modern ML algorithms have achieved astonishing performance in numerous fields[19, 80, 81].

The general goal of ML is to train the model to recognize "patterns" in vast amounts of data and use the learned patterns to make predictions or decisions. An ideal machine that learned enough solid-state physics is expected to answer several questions, such as, what is the composition-pressure-temperature phase diagram of a multi-component system? What is the (free) energy of given crystal under given environmental conditions? Or, how about the universal density functional? Unfortunately and also fortunately, there is no such a machine yet. ML models may already achieve similar accuracy in prediction as physics theories. However, in many cases, ML models (especially neural networks) are often referred to as *black box*, being opaque to human understanding and unable to explain themselves. Nevertheless, with sufficient training, ML models perform surprisingly well. Speaking about solid-state physics, ML has been proven successful in many cases, including but not limited to: phase transition [82, 83], band topology [84, 85], ML density functionals [86, 87], free energy surface of reactions [88, 89], interatomic force fields [90–93], atomistic feature engineering [19, 94], etc. Among them, the last two are especially related to this Thesis.

Depending on whether the properties or labels of data are included for input, machine learning can be divided into unsupervised (un-labeled), supervised (labeled), and semi-/self-supervised learning. Based on the learning task, unsupervised learning can further be divided into subcategories including but not limited to: clustering models that are used to group input data that are closely related; dimension-reduction models that reduce the dimension of feature vectors; generative models that generate new data compatible with the input or optimal in a certain parameter space. These models (especially the first two) could be trained to separate and extract the similarities

between elements, i.e., acting like an ML version of Pettifor.

With properties as labels provided to training, the supervised models can accomplish tasks such as: Classifying input data into subsets according to their properties; predicting the properties from new input data; inversely generating input data with desired properties; etc. Connecting desired properties to the input data makes the supervised models popular for data-driven functional material design.

The self/semi-supervised learning could be seen as the mix or intermediate form of the above two. In semi-supervised learning, only a tiny portion of the input is labeled. For self-supervised learning, instead of using explicit labels, the correlations, metadata, or domain knowledge embedded in the data are implicitly and autonomously learned and extracted by the model.

There are three key ingredients to train a "intelligent" machine: first, the training data; second, the learning algorithm; and third, the representation of input data transforming the raw data into features. As discussed above, the HT results are a reliable and valuable data source.

Learning Algorithms

The second essential part, the algorithm, is the part on which most computer scientists focus. We have witnessed a rapid development of machine learning algorithms during the last decade, with breakthroughs and progress constantly made.

For classification or regression problems on smaller tabular datasets, decision tree-based algorithms like random forests [95], gradient boosting trees [96], and extremely randomized trees [97] perform well. The decision tree can be seen as a graph in tree form in which the nodes are logic conditions that divide input data into branches (classes). The decision tree tends to overfit, which can be avoided by combining an ensemble of randomized trees into, e.g., a random forest. Other classification algorithms like support vector classification and k-nearest neighbors can also be used.

For clustering the popular algorithms include k-means clustering, hierarchical clustering, and hidden Markov model, etc. k-means clustering partition n data points into k clusters with minimal in-cluster variances, the cluster center (mean for each cluster) is the cluster prototype. Hierarchical clustering builds a hierarchy of clusters. Hierarchy construction can be repetitively dividing upper-level clusters into sub-level

clusters or merging sub-clusters into higher-level ones. The hidden Markov model contains an unobservable Markov process ("hidden" X). The state of X is, however, influencing another observable process Y. By observing Y, the model deduces the probability distribution for X.

For regression, popular algorithms in material science are ridge regression , support vector regression, symbolic regression, and artificial neural networks. The ridge regression is a multi-dimensional least square linear fitting with regularization. However, in most cases, one must first transform the problem to proceed to linear fitting. The usual choice is to transform the original input into a high-dimensional feature vector, the transforming function is called the kernel function, and the corresponding regression is called kernel rigid regression. Support vector regression allows the least square linear fit to have an error tolerance (ϵ), i.e., errors less than ϵ are ignored, and instead of minimizing the residual, coefficients are minimized.

Neural networks (NNs) are probably the most widely applied algorithms, and based on the architecture, NNs can be divided further into subcategories. In the feed-forward NNs, the data pass through the networks without back-flow. The name "feed-forward" came from this character. Architectures that allow the data to loop backward, e.g., the recurrent NN, fall out of this Thesis's scope. Therefore, in the discussion below, we only focus on feed-forward NN and drop the "feed-forward" for simplicity.

Following the direction of data flowing, a NN is a layered structure that starts with an input layer, continues with several hidden ones, and ends with an output layer. A layer is formed by several nodes (neurons), and for example, the k-th layer, with N neurons, can be seen as a length N real vector $\mathbf{z}^{(k)}$. The neurons are connected between adjacent layers. If all neurons of every layer (except the output layer) are connected to all nodes on the next layer, the network is a fully connected NN (FCNN). The data flow from layer k to the layer $k+1$ can be done by a linear transformation:

$$\mathbf{z}_i^{(k+1)} = \mathbf{W}^{(k)}\mathbf{z}_i^{(k)}, \tag{63}$$

where $\mathbf{W}$ is the weight matrix that needs to be optimized. The nonlinearity is introduced by applying a non-linear activation function f

$$\mathbf{z}_i^{(k+1)} = f(\mathbf{W}^{(k)}\mathbf{z}_i^{(k)} + \mathbf{B}^{(k)}). \tag{64}$$

Here, $\mathbf{B}$ is the bias that shifts the activation functions and provides additional degrees of freedom.

The weight matrices $\mathbf{W}$ are usually randomly initialized. The training can be seen as an optimization procedure of $\mathbf{W}$ minimizing a suitable differentiable loss function L. Typically a mean-squared error loss function can be applied

$$L(x_i, y_i, \theta) = \sum_i (\mathrm{FCNN}_\theta(x_i) - y_i)^2, \tag{65}$$

where θ is the parameters (weights and bias), (x_i, y_i) is the i-th input–label pair and θ are the parameters in $\mathbf{W}$ and $\mathbf{B}$, the optimization is nearly always performed by gradient descent.

$$\theta_{\mathrm{new}} = \theta - r\frac{\partial L}{\partial \theta}, \tag{66}$$

where r is the step size or the so-called learning rate, more complex models, such as convolutional or graph neural networks, can be developed based on this concept of NNs.

Generally speaking, the choice of the machine learning algorithm is entirely problem dependent. For example, tree-based algorithms are more transparent and thus allow explanations for the predictions but are often less accurate. The other extreme is the neural networks which can be very accurate after adequately trained (usually costing significant amounts of data and computational resources). Although NNs are often described as a black box. the interpretatability can be improved to provide physics insights and understandings [19].

Representation of input

The third key ingredient, the representation, transforms human-interpretable data into machine-readable vectorized features, which may sound trivial. However, it is completely the contrary. Similar to the situation in building structure maps mentioned above, in ML solid state science, the early stage of representation relies on physics and chemistry intuition. Atomic property features, such as radius, electronegativities, number of valence electrons, etc., are naturally chosen and combined to represent a chemical system. A pure composition-based representation can be used to reduce human bias. For example, representation can be a vector $X = [c_0, .., c_i, ...c_N]$ where c_i is the normalized ratio of an element i in the composition [98]. This kind of composition vector can also be "folded" into a periodic table shape [99, 100], somehow implicitly including Mendeleev's periodic laws.

Despite the ignorance of structural information, when the problem is limited to a fixed prototype, these kinds of prototype-wise models can be quite helpful, for example, in searching stable perovskites[101], full Heuslers [99], and elpasolites [100, 102], etc. However, lacking structural information will inevitably introduce flaws in the model because the knowledge of the local chemistry environment, which fundamentally affects the crystal's properties, is lost in the representation. Also, without structure information, the machine will have a problem recognizing the difference between permutations of the same composition. For example, in the perovskite prototype ABC_3, the **A** and **B** elements occupy nonequivalent sites, but these models are often unable to distinguish ABC_3 from BAC_3. Moreover, the most natural structural representation, the atomic coordinates, cannot be used in machine learning for crystals because the choice of the lattice is not unique. Making ML models aware of different atomic and crystal environments and retaining invariance to symmetry operations is thus much more challenging than it seems.

Generally speaking, there are several requirements for an ideal representation of crystal structure: First, it should be invariant to the translation and rotational operations of the coordinate system as well as under the permutation of atomic indices. Second, it should be a continuous function of the structure input. Moreover, it should be bijective to crystal structures. One of the choices is to use a pair-wise, two-body matrix in the form of Coulomb potential to encode the element and the distances [103]

$$M^C_{nm} = \begin{cases} \frac{1}{2} Z_n^{2.4} & n = m \\ \frac{Z_n Z_m}{|\mathbf{R}_n - \mathbf{R}_m|} & n \neq m, \end{cases} \tag{67}$$

where Z is the atomic number, **R** is the atomic position and M^C_{nm} is the so called Coulomb matrix. In periodic systems, the matrix elements become

$$\begin{aligned} \phi_{nm} &= \sum_T \frac{Z_n Z_m}{|\mathbf{R}_n - \mathbf{R}_m + \mathbf{T}|} \\ \mathbf{T} &= h\mathbf{a} + k\mathbf{b} + l\mathbf{c}, \end{aligned} \tag{68}$$

where summation runs for all lattice vectors **T**. The summation goes to infinity if the system is not charge neutral, but a neutralizing background charge can be applied to force the convergence [104]. The resulting matrix is called Ewald sum matrix [105], which correctly captures the periodicity (translation invariant). However, the matrix is not unique under the atom index permutation [106]. To further encode the distance

(bond length) and the bond angles, one can use other geometry functions (g_k), which transform structural information into a scalar value. For example the atomic number and distance functions, respectively $g_1(Z_n)$ and $g_2(\mathbf{R}_n, \mathbf{R}_m)$ as in equation 67 above, furthermore the bond angle function $g_3(\mathbf{R}_n, \mathbf{R}_m, \mathbf{R}_l)$. The scalar output of these functions discretized on atomic positions are then broadened to distribution functions p_k through a density estimation kernel, e.g., a Gaussian

$$p_k = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{(x-g_k)^2}{2\sigma_k^2}} \tag{69}$$

Then a sum of each distribution p_k is made for each possible combination of k elements to give the so-called many-body tensor representation (MBTR) [107] of the structure, for example, for $k = 3$

$$\mathrm{MBTR}_3^{Z_1 Z_2 Z_3}(x) = \sum_n^{Z1} \sum_m^{Z2} \sum_l^{Z3} w_3^{nml} p_3^{nml}(x), \tag{70}$$

where the sums for n, m, and l run over all atoms with the respectively the atomic number Z_1, Z_2, and Z_3 and w_3 is a weighting function. For the periodic system, the summation extends to the periodic copies of atoms in neighboring cells, and an exponentially decaying w_k can be used to converge the summation. Other than using the straightforward geometry functions combined with probability density kernel to describe the "distribution" of structural information, one can adopt the atom-centered symmetry functions (ACSFs) G_k defined as follow [108]

$$\begin{aligned} G_1^{n,Z_1} &= \sum_m^{n,Z_1} f_c(R_{nm}) \\ G_2^{n,Z_1} &= \sum_m^{n,Z_1} e^{-\eta(R_{nm}-R_s)^2} f_c(R_{nm}) \\ G_3^{n,Z_1} &= \sum_m^{n,Z_1} \cos(\kappa R_{nm} f_c(R_{nm}) \\ f_c(r) &= \frac{1}{2}\left[\cos\left(\pi \frac{r}{r_{\mathrm{cut}}}\right)\right], \end{aligned} \tag{71}$$

where summation runs through atoms n with atomic number Z_1; η, R_s, and κ are control parameters; f_c is the smooth cutoff function and r_{cut} is the cutoff radius. Note here that all G_k are pair-wise functions, but three-body functions can also be defined in similar ways.[108] Furthermore, the atomic neighbor density $\rho^Z(\mathbf{r})$ which represents the local environment centered on atom n at position R_n, can be expanded in spherical

harmonics Y_{lm} and radial basis functions g_n [109]

$$\begin{aligned}\rho^Z(\mathbf{r}) &= \sum_n^Z e^{-\frac{1}{2\sigma^2}|\mathbf{r}-\mathbf{R}_n|^2} \\ &= \sum_{nml} c^Z_{nml} g_n(r) Y_{lm}(\theta,\varphi) \\ c^Z_{nml} &= \int\int\int_{\mathbf{R}}^3 \mathrm{d}V g_n(r) Y_{lm}(\theta,\varphi)\rho^Z(\mathbf{r}).\end{aligned} \tag{72}$$

The key part of structural information is the change in the local environments across the crystal. The similarity of the two local environments around atom n and m can be defined as the overlapping of ρ_n and ρ_m

$$\begin{aligned}S(\rho_n,\rho_m) &= \int \rho_n(\mathbf{r})\rho_m(\mathbf{r})d\mathbf{r} \\ k(\rho_n,\rho_m) &= \int |S(\rho_n,\hat{R}\rho_m)|^N d\hat{R} \\ &= \int d\hat{R}\Big|\int \rho_n(\mathbf{r})\rho_m(\hat{R}\mathbf{r})d\mathbf{r}\Big|^N \\ K(\rho_n,\rho_m) &= \frac{k(\rho_n,\rho_m)}{\sqrt{(k(\rho_n,\rho_n)k(\rho_m,\rho_m))}}.\end{aligned} \tag{73}$$

Here $\hat{R}$ is the rotation, and the integral is over all N possible rotations. Thus the representation is rotational invariant. The normalized overlapping kernel $K(n,m)$ is the so-called smooth overlap of atomic positions (SOAP) kernel [109].

The above representations are closely related and can be derived as projections of the atomic neighbor density onto variously chosen basis functions [110, 111]. Furthermore, several works have been devoted to optimizing the basis functions and/or the parameters to deliver better machine models [112–115].

However, these representations do not fulfill the requirement of being bijective. Generally, they cannot distinguish the structures having degenerate n-body correlations (distances, bond angles, dihedral angles, etc.) [116]. In practice, this issue can be mitigated by writing global properties as a sum of atom-centered contributions, but improving the performance of models to higher accuracy is hindered fundamentally by the incompleteness of representations [116].

Another route of extracting structural information is the graph-based representations [21, 94, 117–120], which, unlike previous representations that

are based on physical intuition, follow the principles of deep learning [121] and let the model learn the representation by itself.

These representations simplify the crystal structure into a graph, where the vertices (nodes) of the graph represent the atoms and the graph's edge represents the bonds. The crystal graph is undirected, so it is permutational invariant. It allows multiple edges between the same pair of vertices to represent bonding to periodic images of atoms. Thus it is translational invariant. To ensure the rotational invariance, one can use the Euclidean distance between atoms (nodes) to feature the edge [118, 119], corresponding neural network is called (real-)distance graph NNs (dGNN). The atomic information can be encoded (embedded) into the nodes feature vectors $\mathbf{v}_i$ and bond information between atom i and j can be coded as edge embedding $\mathbf{u}_{ij}$, the initial feature vectors are not the optimal representation, but during training, the nodes and edges information are passed and updated from time step t to $t+1$ by convolution (Conv)

$$\mathbf{v}_i^{(t+1)} = \mathrm{Conv}(\mathbf{v}_i^{(t)}, \mathbf{v}_j^{(t)}, \mathbf{u}_{ij}^{(t)}), \tag{74}$$

and makes the model gradually "learn" the local atomic environments through training. Of course, the design of the convolution function is as essential and non-trivial as the feature extraction [19].

One limitation of the dGNNs is the absence of information about the bond angles, which makes the graph representation also incomplete [122]. As mentioned above, the completeness (bijection) of the representation could crucially affect the predictive power of the model [122–124]. Increasing the cutoff radius for bonds works in specific cases, but does *not* guarantee bijective graph representation [122].

To solve this issue, one could include higher-order correlations (e.g., bond angles, dihedral angles, etc.) in the graph. However, it is unclear whether a higher order of correlation is sufficient to construct a bijective graph representation [122].

Nevertheless, higher-order correlations can be easily incorporated, benefiting from the high flexibility of graph representations. For example, using angles between dipoles of the surface and the absorbed molecule can improve the performance of predicting surface-molecule interactions [125]. The goal of achieving bijective graph-based representation is of great research interest and attention, and the family of

better representations is rapidly growing. New architectures, e.g the E(3) equivariant networks have been continuously developed during the recent years [126–131].

It is also worth noting that the bijection is not the only factor affecting the model's predictive power. Another issue of dGNNs or any GNNs requiring precise values for higher order correlations is that the actual positions of atoms are unknown for the hypothetical candidates *a priori*. Replacing the Euclidean n-body correlations with their graph counterparts could circumvent this problem [21, 120].

CGAT and M3GNET

The crystal graph attention networks (CGAT) architecture [21] used by one of the works of this Thesis [WPhD4] is an example of extending the above graph representation. Instead of using Euclidean distances between atoms as edge embedding, CGAT uses graph distance, i.e., first neighbor, second, third, etc. In CGAT, the convolution is first through fully connected neural networks (FCNN), which output the message vectors ($\mathbf{m}_{ij}$) and the attention coefficients ($\mathbf{a}_{ij}^n$) [21]

$$\begin{aligned} \mathbf{m}_{ij} &= \mathbf{FCNN}_m^n(\mathbf{v}_i^{(t)} \oplus \mathbf{v}_j^{(t)} \oplus \mathbf{u}_{ij}^{(t)}) \\ \mathbf{s}_{ij}^n &= \mathbf{FCNN}_a^n(\mathbf{v}_i^{(t)} \oplus \mathbf{v}_j^{(t)} \oplus \mathbf{u}_{ij}^{(t)}) \\ \mathbf{a}_{ij}^n &= \frac{e^{s_{ij}^n}}{\sum_j e^{s_{ij}^n}}, \end{aligned} \tag{75}$$

where $\oplus$ is the concatenation. The vertices are then updated with a hyper-FCNN (FCNNs that output FCNN) [21]

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \mathbf{HFCNN}^t(\oplus_n \sum_j \mathbf{a}_{ij}^n \mathbf{m}_{ij}^n). \tag{76}$$

A similar attention-based pooling layer and a fully connected network with residual connections ($\mathbf{FCNN}_{\mathrm{RS}}$) give the final graph embedding [21].

$$\begin{aligned} \mathbf{s}_i^n &= \mathbf{FCNN}_a^n(\mathbf{v}_i^{(t)} \oplus \mathbf{C}) \\ \mathbf{a}_i^n &= \frac{e^{s_i^n}}{\sum_i e^{s_i^n}} \\ \mathbf{m}_i^n &= \mathbf{FCNN}_m^n(\mathbf{v}_i^{(t)}) \\ \text{Output} &= \mathbf{FCNN}_{\mathrm{RS}}(\oplus_n \sum_{i,n} \mathbf{a}_i^n \mathbf{m}_i^n). \end{aligned} \tag{77}$$

The CGAT architecture is trained for predicting the thermodynamic stability (explained later in this Chapter). However, if dynamic properties are needed, graph-based models can also be adapted to fit the DFT data into an interatomic force field.

As an example, such a neural-network force field, named M3GNET [22], was applied in Chapter 6 [WPhD4]. In M3GNET, in order to train the model to predict the forces and stresses, information of atomic coordinates $\mathbf{x}_j$ and lattice matrix $\mathbf{M}$ are added in the features. Similarly, the updating is through the convolution function. For example, the edge updating can be written as

$$\begin{aligned} \widetilde{\mathbf{u}}_{ij} &= \sum_k j_l(z_{ln}\frac{r_{ik}}{r_c})Y_{l0}(\theta_{ijk}) \odot \sigma(\mathbf{W}_v\mathbf{v}_k + \mathbf{b}_v)f_c(r_{ij})f_c(r_{ik}) \\ \mathbf{u}_{ij}^{(t+1)} &= \mathbf{u}_{ij}^{(t)} + g(\widetilde{\mathbf{W}}_2\widetilde{\mathbf{u}}_{ij} + \widetilde{\mathbf{b}}_2) \odot \sigma(\widetilde{\mathbf{W}}_1\widetilde{\mathbf{u}}_{ij} + \widetilde{\mathbf{b}}_1) \\ f_c(r) &= 1 - 6(r/r_c)^5 + 15(r/r_c)^4 - 10(r/r_c)^3, \end{aligned} \tag{78}$$

where $\mathbf{W}$ and $\mathbf{b}$ are weights from the network, j_l is the spherical Bessel function with the roots at z_{ln} and with r_c the cutoff radius, Y_{l0} is the spherical harmonics function with $m = 0$, θ is the j–i–k bond angle, $\odot$ denotes element-wise product, σ is the sigmoid activation function, and $g(x) = x\sigma(x)$ is the nonlinear activation function, f is the smooth decaying cutoff function. The output of the network is the energy E, and through auto-differentiation the forces $\mathbf{f} = -\partial E/\partial \mathbf{x_i}$ and stresses $\sigma = V^{-1}\partial E/\partial \epsilon$ are obtained.

Thermodynamic Stability

The total energy of a single system calculated through DFT (or predicted by ML) is not enough when we discuss the thermodynamic stability of the candidates. The most intuitive indicator of thermodynamic stability is the (standard) formation energy ($\Delta G_\mathrm{f}^\circ$). Unfortunately, several errors must be examined and corrected before extrapolating DFT results to ambient condition formation energies.

Firstly, the straightforward DFT calculations results are at zero temperature and pressure. The zero-temperature DFT enthalpies ($\Delta H_\mathrm{f}^\mathrm{DFT}$) can be partially corrected by considering phonon contributions, which are, however, too expensive for most of the systems. Luckily the zero-point vibrational and thermal phonon contributions to the formation energies are relatively small [132]. Moreover, with different signs, they often cancel each other, leaving the neglect of both a fair approximation when calculating formation energies.

A second error arises from the approximation of the exchange-correlation functional, as discussed in the last Chapter. This can be improved by applying more accurate while still affordable functional, e.g., SCAN functional [133], to get better DFT total energies on top of the geometries obtained via applying cheaper semi-local functionals [WPhD10]. However, fully applying expensive and accurate functionals in HT search, i.e., including geometric optimization, is still impractical, due to the poor scaling or numerical stability of high accuracy functionals.

Thirdly, stability could be underestimated for anti-ferromagnetic systems because most calculations only consider the ferromagnetic ordering. The systematic correction scheme of $\Delta H_{\mathrm{f}}^{\mathrm{DFT}}$ relies on exhaust search in the magnetic configuration space for every (magnetic) system, which could scale very rapidly with the system size. The effect of magnetism is estimated from a few to hundreds meV/atom [134], highly depending on the system under discussion.

Moreover, for systems including heavy atoms, neglecting relativistic effects may lead to erroneous total energies, but in most cases, the error is again systematic.

Although many types of error sources are systematic, error cancellation in calculating $\Delta H_{\mathrm{f}}^{\mathrm{DFT}}$ from DFT total energies is usually incomplete. The deviation of DFT from experimental data is often beyond the required chemical accuracy ($\sim$ 40 meV/atom) [135–138].

Numerous attempts have been devoted to systematically correcting DFT formation enthalpies while avoiding applying more expensive schemes. For example, to correct the error from the exchange-correlation, one can try to optimize the semi-local functional to give better formation energies [139]. A more expensive alternative is to combine non-self-consistent exact Hartree–Fock exchange with random phase approximation (RPA) correlation [140, 141].

Instead of improving the exchange-correlation functional, empirical corrections on formation energies are also explored, where usually, the correction is made by fitting the DFT formation energies of a chemical family to experimental ones and extracting corrections on the most strong-correlated system(s) in this family. Wang *et al.* [142] suggested a correction to O_2 for evaluating the formation energies of oxides. Similar approaches were then proposed to correct more diatomic gaseous systems for different functionals[137, 138, 143]. Due to the popularity and practicality of using the DFT + U scheme to correct the error of semi-local exchange-correlation functionals for

transition metals, Jain *et al.* [144] proposed an empirical correction scheme for mixing GGA and GGA + U calculations to compute $\Delta H_{\mathrm{f}}^{\mathrm{DFT}}$. Furthermore, fitting to many dissimilar chemical families instead of one family (so-called elemental-phase reference energies, FERE, method [145]) could generate a general element-specific correction. A further step can be made by considering the coordination number (number of bonds forming) and extracting correction per specific type of bond (e.g., per metal-oxygen bond) instead of per atom of a specific element. This bond-wise correction is the so-called coordination correction [136].

However, the corrected formation energies alone cannot describe the compound's stability in reality. The stability of stoichiometries $\mathrm{A}_x\mathrm{B}_y\mathrm{C}_z\cdots$ should be decided by comparing its formation energy with all other possible combinations of competing compositions and polymorphs, $\{\mathrm{A}_{x'}\mathrm{B}_{y'}\mathrm{C}_{z'}\cdots\}$, instead of comparing with the pure constituent elements **A**, **B**, **C**, $\cdots$. It is then necessary to construct the energy surface in the compositional space for all competing phases, and we can draw the lowest convex envelope of all the points. This envelope is called the convex hull (CH). A sketch for binary and ternary CHs is illustrated in Fig 2. It can be seen that any system located *above* the convex hull will have a positive reaction energy to the closet vertices

$$\Delta H_{\mathrm{r}} = \Delta H_{\mathrm{f}} - \sum_{v} \Delta H_{\mathrm{f}}^{v} > 0, \tag{79}$$

which means it will spontaneously decompose to the systems on those vertices. This reaction energy $\Delta_{\mathrm{r}}H$ is also called the distance to the convex hull, usually denoted as E_{hull}, which serves as the indicator of the thermodynamic stability of a compound.

From Eq 79, the systematic error of formation energy would further cancel because the summation is mainly confined in the vicinity of the evaluated system, where the error in DFT is systematic. In practice, empirical corrections on the DFT energies are sufficient to produce a reasonable estimation of E_{hull} even from inaccurate semi-local (GGA) functionals [146, 147]. Of course, one has to beware of the error in calculated E_{hull}, and usually, a "tolerance" is applied when using E_{hull} to estimate the thermodynamic stability of a compound. Commonly accepted tolerance is around 50 meV/atom for three-dimensional systems considering the error of DFT formation energy can be around 100 meV/atom [148, 149]. However, for two-dimensional systems, it can be as high as 250 meV/atom [150–152].

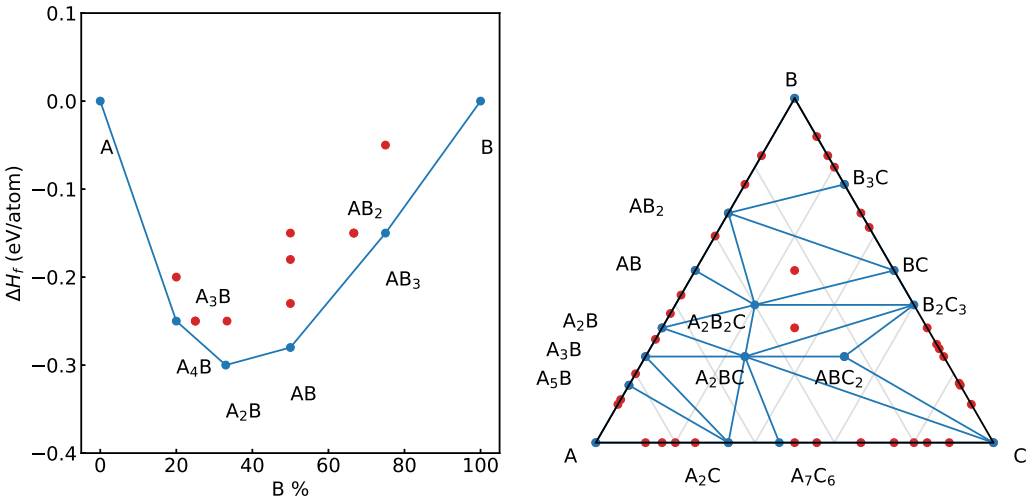


Figure 2: Sketch of a (left) binary and a (right) ternary convex hull.

Databases

The calculation of E_{hull} first needs a robust and reliable hull as reference. As the DFT calculations result heavily depends on the parameters setups (functional, pseudopotential, cutoff, k-grid densities, etc.), a standardized setup is required to allow researchers to share and co-contribute to the hull construction. This is also one of the initial motivations for building computational databases.

In Table 1 we listed several of the most relevant databases by their aliases for simplicity, their capacity, and the mainly utilized code, i.e., VASP [153, 154], QUANTUM ESPRESSO (QE) [155], or GPAW [156, 157]. Note the experimental inorganic crystal structure database (ICSD) [10], which has around 260,000 entries (with duplicated entries on the same structures).

The MP contains all experimental known solid-state materials and thousands more unknown materials. Their newest hull contains around 30,000 structures. For part of the systems, other properties, including elastic tensor, surface energy, electron properties, phonon, dielectric constant, etc., are also calculated. The MP convex hull construction has several systematic corrections on the errors of GGA functional, and the same correction scheme is also applied to all the works related to this Thesis.

The OQMD also contains all the experimental structures (with less than 34 atoms in the unit cell, $\approx$ 32,000) and exhaustive searches on several typical prototypes. A statistic calculated for the database shows that the mean absolute error between

Table 1: Some examples of large online computational databases.

Database	Alias	Entries (10^3)	Software
Materials Project [146]	MP	146	VASP
Open Quantum Materials Database [148, 149]	OQMD	1,022	VASP
Automatic-Flow for Materials Discovery [11]	AFLOW	3,528	VASP
Materials Cloud [158]	MC	38	QE
Dataset CGAT[WPhD12]	DCGAT	2,200	VASP
SCAN and PBEsol convex hull[WPhD10]	SCANHULL	250	VASP
Computational 2D Materials Database [150–152]	C2DB	5	GPAW
2D Materials Encyclopedia [159]	2DMatpedia	6.3	VASP
Computational 1D Materials Database	C1DB	< 3	GPAW
Exhaustive 2D Dataset[WPhD4]	EX2D	68	VASP

experimental formation energies and their DFT calculations is 96 meV/atom. In contrast, the mean absolute error among experimental results is 82 meV/atom. Therefore, the experimental uncertainties, besides the error from approximation in functionals, could be one primary source of the deviation of calculated formation energies from experimental results.

The AFLOW contains the phase diagrams for about 1.7k binary, 30k ternary, and 150k quaternary alloy systems. For now, it is possibly the largest database on calculated (formation) energies of inorganic crystals. It also contains nearly twice the amount of electronic structure entries (370k). One advantage of AFLOW is that it includes a few million bcc-/fcc-derived and a similar number of hcp-derived superstructures, which enables further study of chemical, spin, and defect (dis-)ordering.

The MC seems dwarfed in capacities compared with the three databases introduced above. However, this is merely because the tabulated number counts the entries in the Material Cloud 3D and 2D sets. At the same time, there are more than 11 million (although no check for duplicates has been performed) structures in the entire MC archive system contributed by researchers worldwide. For example, our group has

several large ones (in total around 2 million), including the DCGAT [WPhD12, 160] set, the recalculated hull using SCAN and PBEsol functional [WPhD10, 161], as well as the datasets obtained in Chapter 5 [WPhD3, 162] and 6 [WPhD4, 163]. The integration of different datasets on MC could be beneficial, but checking compatibility and removing duplicates across datasets can be technically complicated [21].

The above databases are mainly focused on bulk structures, and the rest are mainly about low-dimensional materials (LDM). These LDM databases are much smaller than the 3D ones, mainly because fewer experimental prototypes are proposed in experimental LDM databases. The LDM computational databases are intuitively constructed by examining the (thermodynamic) possibility of exfoliating corresponding 3D crystals. This bias inherited from experiments can be partially compensated through a systematic search on all possible species and Wycoff position combinations within each specific two-dimensional space group, which we will discuss further in Chapter 6 [WPhD4].

As shown in the table, unfortunately, there is no unique standard of the parameters among databases, not just because of the variety of computational codes favored but also the slight difference in the choice of some crucial parameters, for example, regarding the choice of U values for the GGA + U schemes. This reality sabotages the compatibility of the hull between databases. For example, different correction schemes when mixing GGA with GGA + U data are required. However, the bright side is that repeating calculations on the same set of structures, especially those that originate from experiments and thus should locate near the hull, could confirm the accuracy across all databases. More importantly, it enables exploration of how minor changes in crucial parameters affect systematic errors [149].

It is also possible to alleviate incompatibility by combining all the proposed hulls in all databases (or even rerunning them with the same parameters). For example, a recent attempt tried to provide a hull uniting several databases/sets (mainly MP, AFLOW, and datasets accumulated within our group [WPhD1–WPhD3, 101, 164]) and calculated with more accurate SCAN and PBEsol functionals [WPhD10]. Of course, an ideal but expensive approach is to scan the combinatorial space and get a complete hull exhaustively. Fortunately, by using powerful machine learning models [21, 22] to pre-exclude the majority of very unstable systems, the number of DFT validation calculations can become manageable, as shown in recent achievements on large datasets and a more complete convex hull [WPhD4, WPhD12].

Chapter 3 Searching New Double Perovskites as Transparent Semiconductors

In the following publication "**Double perovskites as p-type conducting transparent semiconductors: a high-throughput search**" [WPhD1][1], we perform a systematic search of the family of quaternary halide perovskites having general formula $A_2B^{(1)}B^{(2)}C_6$ in order to find promising candidates for transparent p-type conduction.

On the other hand, the halides could be seen as an alternative route to design TCM, for example, in CuI, as Cu^{1+} causes the valence band edge to develop hybrid character between the localized O $2p$ and dispersive Cu $3d$-orbitals[165], the m_h^* can be largely lowered (0.2–0.25 m_0) [166]. Previous works in our group [167] show that there are several low effective hole masses halide perovskites, although the gap is usually not wide enough.

Therefore we take a step further to perform a high-throughput search of double perovskites. We scanned all stoichiometries of the form $A_2B^{(1)}B^{(2)}C_6$, where A is either Rb or Cs, C is a halogen (F, Cl, Br, or I), and B is from hydrogen to bismuth, excluding the rare gases and the lanthanides (except La, which is included). The total number of candidates is 16384. We first filter the ones with a threshold of distance to the convex hull less than 25 meV/atom. This gives us a list of 1699 candidates, roughly around 10%. The reason to achieve such a rather high success rate (for stability) is that we confine the search around the compositional space. For A = Rb or Cs, numerous experimentally synthesized double halide perovskites exist [168–170]. We further filter the candidates with a gap (at the PBE level) $E_{gap}^{PBE} \geq 1.8$ eV, which leads to a list of 633 (37%). Further, only 17 double perovskites and a ternary perovskite ($CsPbF_3$) fulfill the third threshold $m_h^* < 1\ m_e$. Fortunately, 10 out of 17 do not include toxic or rare chemical elements.

Wide bandgap and low m_h^* are *necessary* for p-type TCM. However, these two conditions are *not sufficient* for good performance of the material. Another condition that has to be taken into consideration is the possibility of creating holes in the valence band, i.e., the p-type dopability. The dopability is usually checked by calculating the formation energies of the point defects. References on TM oxides show that the p-

[1] Reproduced with permission from the Royal Society of Chemistry.

type dopability is generally poor [171–173] because of the formation of energetically favoured hole killers such as the oxygen vacancy. The defect calculation for $CsPbCl_3$ shows that under the Pb-poor condition, the formation energies of shallow acceptor defects (e.g., the intrinsic V_{Cs}^{-}), extrinsic K_{Pb}^{-}, etc.) are lower than that of donor defects [174].

Note that the defect formation energies depend on not only the enthalpy differences between defected and pristine supercells but also the chemical potentials of elements. For quaternary systems, based on the synthesis condition, the investigating system is under equilibrium with stable ternary, binary, or elementary systems, and the chemical potential of each element can change drastically under different conditions. Therefore, the number of all possible equilibrium conditions can be large enough to prevent exhaustive calculations. Partially because of this reason, we did not perform the check on dopability. However, compared to ternary perovskites (e.g., $CsPbCl_3$ [174]), one could adjust the chemical potential of the two B-sites elements, which provides additional degree of freedom to tune the defect formation energies. Therefore, the formation of shallow acceptor defects in double perovskites should be easier to control, although further calculations or experiments are needed to validate this.

Journal of Materials Chemistry A

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. A, 2019, 7, 14705

Double perovskites as p-type conducting transparent semiconductors: a high-throughput search†

Hai-Chen Wang, [a] Paul Pistor, [a] Miguel A. L. Marques *[a] and Silvana Botti [b]

Received 7th February 2019
Accepted 20th May 2019

DOI: 10.1039/c9ta01456j

rsc.li/materials-a

We perform a systematic study of the family of quaternary halide perovskites in order to find good candidates for transparent p-type conduction. This is achieved by using high-throughput techniques based on density-functional theory, and by screening the materials with regard to their stability, electronic band gap, and hole effective masses. We find a total of 17 double perovskites with promising properties, 10 of which not including toxic or rare chemical elements. Furthermore, in most of these systems, doping might be achieved by adjusting the chemical potential of the two cations during the growth process. Due to chemical similarity, we expect that these materials are compatible with current photovoltaic technology based on organic halide perovskites.

1 Introduction

Transparent conducting semiconductors (TCSs) form a large family of materials that combine both high conductivity and transparency. Researchers have found many n-type TCSs, including In_2O_3, ZnO, and SnO_2, which are by now routinely used as transparent electrodes and thin films transistors in solar cell devices, infrared reflective coatings, and electrochromic displays, to name a few examples.[1,2] Unfortunately, good p-type TCSs, which are fundamental for the development of new electronic architectures such as transparent p–n junctions, are much rarer.[2] Since the first report on a p-type TCS made of NiO,[3] a class of promising p-type TCSs was identified at the end of the 90s in the family of Cu oxides with the delafossite structure. $CuAlO_2$ was first investigated by Hosono *et al.* in 1997,[4] leading to an extensive research effort of the whole family of $CuMO_2$ delafossite compounds.[5] During the past two decades, a large effort, both experimental and theoretical, has been made to explore potential p-type TCSs in other crystal families, and potential candidates have been found in the Sn–O system,[6,7] in spinel oxides[8–10] or in chalcogenides.[11–13] Unfortunately, due to the low valence band dispersion of localized O-2p orbitals in these candidates, existing p-type TCSs still do not have a performance comparable to their n-type counterparts[2,7,11] and new solutions for p-type transparent conductivity are under active research.

There are a series of conditions that are required for a good p-type TCS. First of all, the energy gap should be large in order to avoid absorption in the visible range.[2,7,11] Then one usually searches for holes with small effective masses at the top of the valence band, in order to ensure the mobility of charge carriers. Unfortunately, from $k\cdot p$ theory we know that effective masses are usually inversely proportional to the band gap, which complicates considerably the finding of materials with large gaps and small masses.[2,7,11] Finally, we must be able to generate a large density of carriers, which often means a large concentration of p-type defects. Unfortunately, several materials exhibit native n-type defects, which may compensate any p-type doping.[2,7,11] This is evident, for example, in many oxides where oxygen vacancies typically have low formation energies and n-type character.[2,7,11]

A material that was recently found to have excellent p-type conduction properties is CuI.[14] This compound exhibits a very dispersive valence band, leading to light holes with a mass of around $0.2–0.3m_e$.[15,16] Furthermore, the low energy defects are in this case copper vacancies (with a formation energy estimated to be around 0.5 eV), which are naturally p-type.[14] Finally, the gap above 3 eV makes this material transparent in the optical regime.[14] These remarkable findings suggest that perhaps the holy grail material for transparent conduction is not an oxide.

Copper(I) iodide is not the only halide material that has recently been in the spotlight for opto-electronic applications. In fact, it was recently shown that halide perovskites have extraordinary properties as absorbers for photovoltaic devices, achieving efficiencies of more than 20%.[17–19] Perovskites form a large family of materials with a wide variety of chemical compositions and material properties. Besides the applications in the field of photovoltaics, it has been shown that, within the

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

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ta01456j

perovskite family, one can find materials with a wealth of interesting physical properties, relevant for instance for high-k dielectrics, superconductivity, piezoelectricity, magnetoelectricity, *etc.*[20–23]

In view of the above, we asked ourselves if it is also possible to find perovskite systems relevant for transparent p-type conduction. Here, we will try to answer this question by performing a computational high-throughput study based on density-functional theory (DFT). Our objective is to find perovskites which are (i) thermodynamically stable, (ii) have a wide band gap, and (iii) have low effective hole mass. Of course, there have been quite a few high-throughput studies published in the recent literature trying to find either new p-type transparent conductors[13,24–27] or novel perovskites for a wealth of applications.[20,21,23,28–31] In this context, some of us already looked in ref. 21 at p-type conduction in ternary inorganic perovskites, finding a few promising systems. Here we go a step further and look at quaternary halide double perovskites. This class of materials has been recently considered[32,33] as promising absorber layers in an attempt to solve the two main problems in the field of perovskite photovoltaics: stability and the use of Pb.

Double perovskites have a number of advantages with regard to their ternary counterparts. First of all, the phase space of possible compounds is substantially larger, which increases considerably the probability of finding promising candidates with the desired properties. Furthermore, double perovskites most often crystallize in a cubic lattice, avoiding the complicated distortions that are often present in ternary perovskites.[34] Finally, if the double perovskite includes cations in different oxidation states it may be possible to dope it either n or p by simply changing the relative chemical potentials during growth.

Another advantage of double perovskites is that they are more stable to moisture in comparison with hybrid perovskites.[35,36] For this reason they were at the beginning proposed to replace hybrid perovskites as absorbers. However, their band gaps turned out to be all too large for use in single-junction solar cells.[37,38]

As an example, we can report that Slavney *et al.* found that 30 days-exposure in 55% relative humidity caused no material decomposition in double perovskite $Cs_2AgBiBr_6$.[39] Moreover, solar cells using inorganic-only transport materials are more stable against moisture.[36] We can therefore expect that inorganic double perovskites working as TCS layers will also improve the performance of hybrid solar cell devices under humid conditions.

On the negative side, it might be more complicated to synthesize quaternary perovskites, not only due to the large number of possible secondary phases, but also due to the difficulty of avoiding exchange of the two cations.

To screen the composition space searching for p-type perovskite $A_2B^{(1)}B^{(2)}C_6$ compounds, we decided to use high-throughput density-functional theory. In the first step, we study the thermodynamic stability of quaternary double perovskites in a large set of compositions. Compounds that are stable or close to thermodynamic stability are filtered out for further theoretical characterization using state-of-the-art *ab initio* methods. The remainder of this article is structured as follows: in Section 2 we describe our computational workflow and give the numerical parameters of the calculations. The results for thermodynamic stability and electronic properties are analysed in Section 3. Finally, we draw some conclusions in Section 4. More details on the results are included as ESI.†

2 Computational methods

We scanned all stoichiometries of the form $A_2B^{(1)}B^{(2)}C_6$, where A is either Rb or Cs, and C is a halide (F, Cl, Br, or I). For the B atoms we took all combinations of elements from hydrogen to bismuth, excluding the rare gases and the lanthanides (with the exception of La, which was included). This amounts to 64 chemical elements, leading to 2048 combinations for each choice of alkali metal and halide, and a total number of more than 16 000 possible compounds. In the choice of the composition space, we focused on Rb and Cs compounds, for which stable systems with interesting properties are already reported.[33,38,40,41]

A priori we can expect to find in the B positions either two divalent metals or one monovalent metal and one trivalent metal. Therefore, we could have restricted our search to only elements exhibiting those oxidation states. However, we decided not to bias our search by these considerations, and to allow for the possibility of having elements in less common oxidation states. We should nevertheless note that, for practical applications, the presence of two divalent metals is probably undesired due to the difficulty of avoiding cationic exchange and consequential disorder. Having two B cations in different oxidation states, thereby creating two different chemical environments, is therefore preferred.

We used the standard face-centered double perovskite prototype shown in Fig. 1, with 10 atoms in the primitive unit cell. We then optimized the lattice constant and calculated the total energy, which can be done very efficiently due to the high-symmetry of the cubic structure. To this end we applied *ab initio*

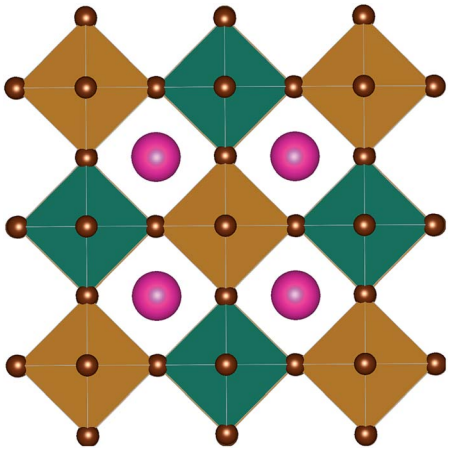


Fig. 1 The double perovskite structure ($A_2B^{(1)}B^{(2)}C_6$) used in this work. The A (C) atoms of the original perovskite structures are in pink (brown), and two different species $B^{(1)}$ and $B^{(2)}$ are positioned at the center of the green and brown octahedra, respectively.

density functional theory as implemented in the computer code VASP.[42,43] All parameters were set to guarantee compatibility with the data available in the materials project database[44] and open quantum materials database.[45] We used the PAW[46] datasets of version 5.2 with a cutoff of 520 eV and Γ-centered k-point grids, as dense as required to ensure an accuracy of 2 meV per atom in the total energy. All forces were converged to better than 0.005 eV Å^{-1}. All geometry optimizations were performed with spin-polarization using the Perdew–Burke–Ernzerhof[47] (PBE) exchange–correlation functional, with the exception of fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V, and 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 position of d-states.

To determine the thermodynamic stability of a compound, we calculated the distance of its formation enthalpy from the convex hull of stability, *i.e.* we compared its formation enthalpy to that of all other known competing phases and phase mixtures, including decomposition into elemental, binary, ternary, and other quaternary phases. We chose to use the materials project database for our reference energies, and to determine the distances to the convex hull of stability with PYMATGEN.[48] The materials project database includes most of the experimentally known inorganic crystals that are present in the ICSD database[49,50] and an increasing number of theoretically predicted phases.

Our goal is to propose double perovskites that (i) are stable, (ii) yield large band gaps to ensure transparency, and (iii) possess small effective masses.

Starting from the initial set of 16 000 compounds, we first screened out all systems with a distance to the convex hull larger than 25 meV. The threshold was not set strictly to zero as we should keep in mind that there is an error inherent to our theoretical approach and to the numerical modeling (the PBE approximation, neglect of zero point motion and temperature effects, neglect of disorder, neglect of defects, *etc.*). Furthermore, thermodynamically unstable phases under ambient conditions can be sometimes stabilized by temperature, pressure, strain, *etc.* and therefore synthesizable.

After applying this first filter, we calculated band gaps and effective masses for a remaining set of compounds, always using the PBE exchange–correlation potential.[47] It is well known that a Kohn–Sham PBE band structure systematically underestimates the gap, often by about 50%. As our objective is to guarantee the transparency of our candidate materials, we set therefore a threshold of 1.8 eV for the PBE gap. This threshold value is equal to the PBE band gap of $CuAlO_2$. A successive comparison with band gaps calculated with the hybrid functional proposed by Heyd–Scuseria–Ernzerhof[51] (HSE06) confirms the validity of this choice.

The final descriptor that we use to filter our results is the hole effective mass. In this case, a semi-local approximation such as PBE already yields reliable values, but unfortunately the calculations are rather sensitive to the number of k-points. To circumvent this problem, we used a dense grid consisting of 8000 k-points per atom. The Kohn–Sham eigenvalues were then Fourier interpolated with the BoltzTrap[52] software, following the same approach as ref. 24. We calculated the average hole effective mass tensor for a carrier concentration of 10^{18} cm^{-3} and a temperature of 300 K. We then used the higher limit estimation for m_h^* (see ESI of ref. 24). We used the threshold value of $m_h^* < 1m_e$ to further filter the number of candidate structures.

Finally, accurate band structures of selected compounds were evaluated with the Heyd–Scuseria–Ernzerhof HSE06 (ref. 51) hybrid functional. The results are discussed in the following section.

3 Results and discussion

The distance to the convex hull of thermodynamic stability gives us a strong constraint on the systems that are possible to synthesize.[45,53] As such, the first step of our procedure is the calculation of the convex hull of stability for all the initial set of double perovskites.

As an example, we plot, in Fig. 2, the distance to the convex hull (in meV per atom) for the $Cs_2B^{(1)}B^{(2)}Cl_6$ system. Analogous plots for the other halides can be found in the ESI.† The chemical elements are ordered according to a modified Pettifor scale,[54] a one-dimensional representation of the periodic table where chemically similar elements are placed in neighboring positions. Furthermore, only the lower triangle is shown due to the equivalence of the $B^{(1)}$ and $B^{(2)}$ atoms. The diagonal corresponds, evidently, to the standard ternary perovskites.

To calculate the distance to the hull for each system in Fig. 2, we excluded it from the hull itself. In this way we obtain negative distances for stable materials. Of course, the magnitude of this value is the energy gained by forming the quaternary perovskite. We then used a color scale for visualization: green dots correspond to theoretically stable structures, while red denotes (very) unstable phases. As discussed above, one should not entirely disregard systems that are slightly unstable (orange points).

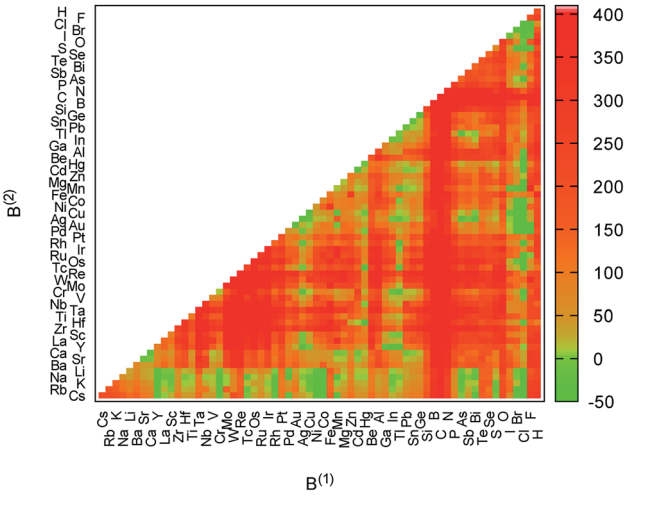


Fig. 2 Distance to the convex hull for the $Cs_2B^{(1)}B^{(2)}Cl_6$ double perovskites in meV per atom. Green corresponds to theoretically stable compounds.

Table 1 Candidate double perovskite systems compared with CuI, $CuAlO_2$, and $CuInO_2$: calculated lattice constant a, ionic radii of $B^{(1)}$ ($r_{B(1)}$) and $B^{(2)}$ ($r_{B(2)}$), effective electron (m_e^*) and hole (m_h^*) masses, band gaps calculated at the PBE level (Gap_{PBE}), with the HSE06 functional (Gap_{HSE}), and with the HSE06 functional considering spin–orbit coupling (Gap_{HSE}^{SOC})

Formula	E_h (meV)	a (Å)	$r_{B(1)}$ (Å)	$r_{B(2)}$ (Å)	Gap_{PBE} (eV)	Gap_{HSE} (eV)	Gap_{HSE}^{SOC} (eV)	m_e^* (m_e)	m_h^* (m_e)
CuI	0	—	0.96	—	1.21	2.60	2.38	0.17	0.99
$CuAlO_2$	0	—	—	—	1.83	3.36	3.35	0.77	2.59
$CuInO_2$	0	—	—	—	0.28	1.52	1.51	0.33	0.78
$Cs_2AgBiCl_6$	0	10.948	1.15	1.03	1.86	3.14	2.78	0.38	0.52
Cs_2AsTlF_6	4	9.558	0.58	1.50	2.68	3.68	3.20	0.55	0.42
$Cs_2GaLaBr_6$	0	11.962	0.62	1.03	2.89	4.12	4.08	295	0.60
Cs_2GaLaI_6	0	12.742	0.62	1.03	2.50	3.50	3.39	160	0.56
Cs_2InBiF_6	0	9.844	0.80	1.03	2.26	3.09	2.26	0.59	0.40
$Cs_2InLaBr_6$	0	12.234	0.80	1.03	2.68	3.78	3.74	312	0.59
Cs_2InLaI_6	9	13.024	0.80	1.03	2.31	3.23	3.13	193	0.52
$CsPbF_3$	0	4.897	1.19	—	2.91	3.97	3.03	0.78	0.39
Cs_2SbTlF_6	0	9.819	0.76	1.50	2.47	3.41	2.88	0.59	0.40
Cs_2SnPbF_6	26	9.693	0.69	1.19	2.22	3.14	2.47	0.53	0.34
Cs_2TlBiF_6	0	9.937	1.50	1.03	3.08	4.11	3.14	0.60	0.44
Cs_2YInI_6	1	12.764	0.90	0.80	2.47	3.31	3.21	0.74	0.60
$Rb_2AgBiCl_6$	0	10.867	1.15	1.03	1.83	3.10	2.75	0.38	0.49
Rb_2AsTlF_6	0	9.434	0.58	1.50	2.71	3.70	3.20	0.55	0.39
$Rb_2GaLaBr_6$	26	11.899	0.62	1.03	2.81	3.83	3.81	201	0.59
Rb_2SbTlF_6	16	9.714	0.76	1.50	2.54	3.48	2.92	0.68	0.39
Rb_2TlBiF_6	0	9.845	1.50	1.03	3.09	4.15	3.13	0.73	0.43
Rb_2YInBr_6	17	11.912	0.90	0.80	2.82	3.73	3.70	0.69	0.60

In general, we see an increase of stability going from iodides to fluorides. Furthermore, for each choice of A and C there is a very large number of stable or nearly stable $A_2B^{(1)}B^{(2)}C_6$ systems (see figures in the ESI†), testifying the remarkable flexibility of the perovskite crystal structure. From Fig. 2 we can clearly see that alkali metals (with oxidation state +1) have the ability to form stable quaternary perovskites by combining with a large number of +3 elements. Other cations with propensity to form stable systems are Ag, Hg, group 13 (Ga, In, Tl), and group 15 (As, Sb, Bi) elements. We also find that other halogens can also occupy the B positions, but this alters significantly the chemistry, so it is very unlikely that these systems can retain the perovskite structure.

We have imposed perfect cubic symmetry to our crystal structures to keep the number of atoms per unit cell as small as possible. However, deviations from the ideal cubic structure, caused by tilting and rotation of the BO_6 octahedra, have been reported in many double perovskite systems.[55] To check the

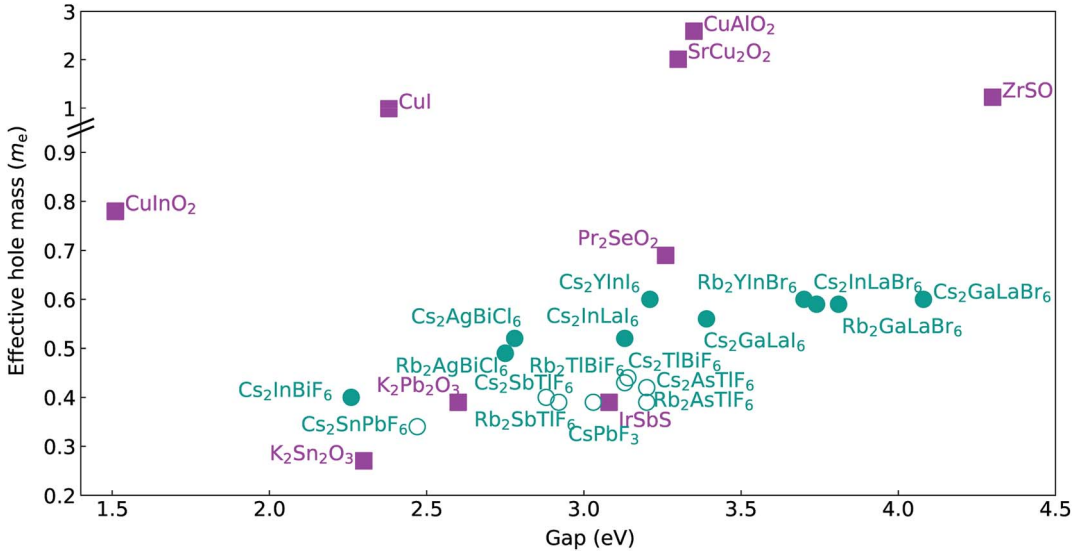


Fig. 3 The band gap *versus* effective hole mass of double perovskite p-type TCS candidates. Candidates being friendly to the environment and including toxic elements are plotted as filled and empty turquoise circles, respectively. For comparison, CuI, $CuAlO_2$, $CuInO_2$, and other p-type TCS candidates from ref. 13, 24 and 27 are marked with magenta squares. Please note the different scales in the y-axis.

effects of such distortions, we randomly rattled the structures of $Cs_2InLaBr_6$ and $Rb_2AgBiCl_6$. After re-optimization, small distortions of the octahedra are indeed observed in both systems. However, we find that the variations of the structural parameters are very small (below 3%). Moreover, these octahedra distortions have only a minor effect on the electronic structure of $Cs_2InLaBr_6$ and $Rb_2AgBiCl_6$, in agreement with the general findings that octahedral distortions do not significantly alter the electronic structure of perovskites.[56,57] We therefore consider only cubic structures in the following. We should however keep in mind that the final structures are likely to be further stabilized by small distortions of the octahedra.

At this point we filter our candidates by removing all structures that have a distance to the convex hull larger than 25 meV per atom. We obtained a list of 1699 candidates, which is around 10% of the whole set. The next step is the calculation of the electronic structure at the PBE level. At this step, the average gap value of almost 1700 systems is 2.32 eV and about 37% of them (633) have a gap above 1.8 eV. The final descriptor that we use to filter our results is the hole effective mass: $m_h^* < 1m_e$. For all 633 compounds, the average m_h^* is $6.47m_e$, implying that systems with low effective hole mass are indeed rare among them. Indeed, there are only 17 double perovskites, and a ternary perovskite ($CsPbF_3$), that remain after our final filtering.

Most of these systems involve heavy atoms like Pb, Bi, or Tl, but there are other compounds with lighter and non-toxic elements. Table 1 displays the candidate structures together with the calculated values of the distance to the convex hull of stability, PBE band gap and electron and hole masses. The maximum PBE gaps we found were 3.09 eV for Rb_2TlBiF_6 and 3.08 eV for Cs_2TlBiF_6. Furthermore, to obtain a better prediction of the band gaps, we re-calculated these by applying the more accurate hybrid HSE06 functional. Considering that heavy elements are involved, we also performed HSE06 calculations including spin–orbit coupling. In Table 1, we included for comparison CuI, as well as the delafossites $CuAlO_2$ and $CuInO_2$. We note that HSE06 still underestimates considerably the gap of CuI, and for which the value for the hole mass is here an average over its three valence bands.[58]

Few of the materials listed in Table 1 had already been proposed for photovoltaic applications. This is the case of $Cs_2BiAgCl_6$ (ref. 32 and 59) or more generally the $Cs_2\{Sb,Bi\}\{Cu,Ag,Au\}\{Cl,Br,I\}_6$ family.[33] The first system was experimentally found to have an indirect gap of 2.2 eV,[32] unfortunately too large for photovoltaics and too small for p-type transparent conduction.

There are several systems with a large band gap and low m_h^* formed by toxic elements Tl and Pb. However, we find also more interesting double perovskites such as $Cs_2InLaBr_6$, $Rb_2AgBiCl_6$, and Cs_2InBiF_6 which are more friendly to the environment. As shown in Fig. 3, our candidates have much lower effective hole masses and wider band gaps than CuI, $CuAlO_2$, and $CuInO_2$. Furthermore, compared with previous results of searching p-type TCSs,[13,24,27] most candidates screened out in this work, especially those without Pb and Tl, are located closer to the lower right corner of Fig. 3.

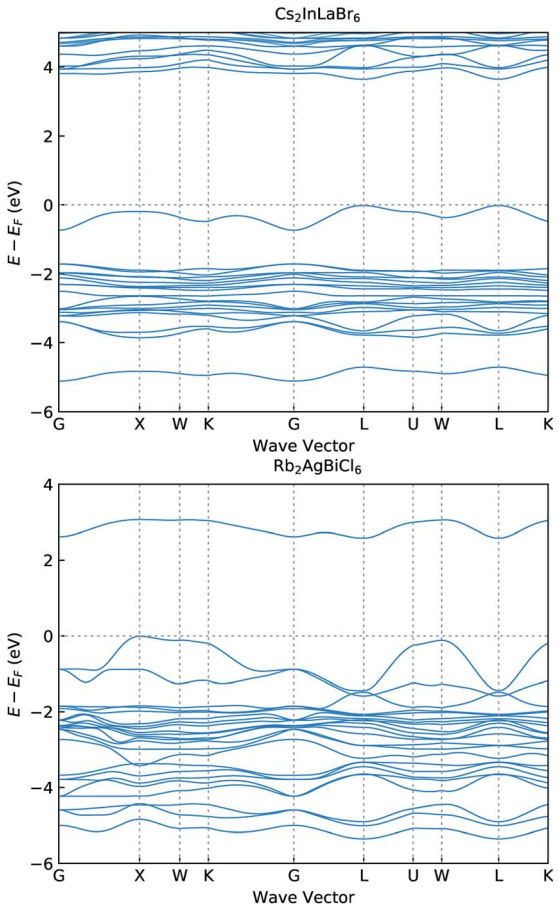


Fig. 4 The band structures of $Cs_2InLaBr_6$ (top) and $Rb_2AgBiCl_6$ (bottom) calculated within the HSE06 approximation including spin–orbit coupling.

We can also divide the candidate double perovskites into two categories based on their effective electron masses: one with both light electrons and holes, for example $Rb_2AgBiCl_6$, and the other having light holes but much heavier electrons such as $Cs_2InLaBr_6$. The band structures of $Cs_2InLaBr_6$ and $Rb_2AgBiCl_6$, calculated at the level of HSE06 including spin–orbit coupling, are shown in Fig. 4. $Cs_2InLaBr_6$ has a direct gap at the L point, while in $Rb_2AgBiCl_6$ the band gap is indirect from the top of the valence band at X to the bottom of the conduction band at the L point. The dispersive valence bands are formed from the hybridization of In-s and Br-p states, or Bi-s and Cl-p states near the Fermi level. This is consistent with the small effective hole mass calculated for these double perovskites. Furthermore, the bottom of the conduction band is more localized in $Cs_2InLaBr_6$ than in $Rb_2AgBiCl_6$, giving a much higher m_e^* in the former. Similar dispersive valence bands also exist in other candidate systems (see figures in the ESI†).

4 Conclusions

In this work we applied a high-throughput DFT calculation scheme to scan the periodic table for transparent conducting semiconductors (TCSs) with the double perovskite structure.

We studied stoichiometries of the form $A_2B^{(1)}B^{(2)}C_6$, where B ranges from hydrogen to bismuth (excluding rare gases and lanthanides except La), A is Rb or Cs, and C is a halogen element. In this phase space containing more than 16 000 possible compounds we spotted 633 structures sufficiently close to the convex hull of thermodynamic stability, and thus stable enough for an easy experimental synthesis.

These systems were then filtered with respect to their electronic band gap and hole effective mass values. We selected at the end 18 candidates, ten of them free of toxic elements such as Pb, As, and Tl. The best compounds have small effective hole masses due to a strong hybridization between the s-states of the B-site element and the p-states of the halogen near the Fermi energy.

We remind that finding a large band gap and low hole effective masses is not a sufficient condition for a good p-type TCS. In fact, another essential condition, much harder to translate into the minimization/maximization of a simple material property, is the p-type dopability of the system. However, in these quaternary systems dopability may be achieved by adjusting the chemical potential of the two cations during the growth process.

Moreover, due to the similarity between these materials and the halide perovskites used as absorbers in photovoltaics, we believe that it should be possible to easily integrate these p-type TCSs in current technology, in a step towards completely transparent photovoltaic modules.

These results are meant to provide experimentalists with essential information on the stability and electronic properties of double perovskites for application as transparent semiconductors. The crystal structures of these compounds are available for more accurate theoretical characterization, hoping that some other interesting properties that have not been screened in this first study can come on the scene and motivate experimentalists to try to synthesize some of these compounds.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

S. B. and M. A. L. M. acknowledge financial support from the DFG through projects SFB-762, MA 6787/1-1, and BO 4280/8. Computational resources were provided by the Leibniz Supercomputing Centre through the SuperMuc Project No. p1841a and pr48je.

References

1 T. Minami, *MRS Bull.*, 2000, **25**, 38–44.
2 R. A. Afre, N. Sharma, M. Sharon and M. Sharon, *Rev. Adv. Mater. Sci.*, 2018, **53**, 79–89.
3 H. Sato, T. Minami, S. Takata and T. Yamada, *Thin Solid Films*, 1993, **236**, 27–31.
4 H. Kawazoe, M. Yasukawa, H. Hyodo, M. Kurita, H. Yanagi and H. Hosono, *Nature*, 1997, **389**, 939.
5 R. Nagarajan, N. Duan, M. Jayaraj, J. Li, K. Vanaja, A. Yokochi, A. Draeseke, J. Tate and A. Sleight, *Int. J. Inorg. Mater.*, 2001, **3**, 265–270.
6 Y. Ogo, H. Hiramatsu, K. Nomura, H. Yanagi, T. Kamiya, M. Hirano and H. Hosono, *Appl. Phys. Lett.*, 2008, **93**, 032113.
7 Z. Wang, P. K. Nayak, J. A. Caraveo-Frescas and H. N. Alshareef, *Adv. Mater.*, 2016, **28**, 3831–3892.
8 H. Kawazoe and K. Ueda, *J. Am. Ceram. Soc.*, 1999, **82**, 3330–3336.
9 T. J. Coutts, D. L. Young, X. Li, W. Mulligan and X. Wu, *J. Vac. Sci. Technol., A*, 2000, **18**, 2646–2660.
10 C. F. Windisch Jr, K. F. Ferris and G. J. Exarhos, *J. Vac. Sci. Technol., A*, 2001, **19**, 1647–1651.
11 A. Banerjee and K. Chattopadhyay, *Prog. Cryst. Growth Charact.*, 2005, **50**, 52–105.
12 K. Ueda, S. Inoue, S. Hirose, H. Kawazoe and H. Hosono, *Appl. Phys. Lett.*, 2000, **77**, 2701–2703.
13 R. Kormath Madam, H. Wiebeler, T. D. Kühne, C. Felser and H. Mirhosseini, *Chem. Mater.*, 2018, **30**, 6794–6800.
14 M. Grundmann, F.-L. Schein, M. Lorenz, T. Böntgen, J. Lenzner and H. von Wenckstern, *Phys. Status Solidi A*, 2013, **210**, 1671–1703.
15 K. Edamatsu, T. Nanba and M. Ikezawa, *J. Phys. Soc. Jpn.*, 1989, **58**, 314–328.
16 J. Wang, J. Li and S.-S. Li, *J. Appl. Phys.*, 2011, **110**, 054907.
17 S. D. Stranks and H. J. Snaith, *Nat. Nanotechnol.*, 2015, **10**, 391.
18 R. F. Berger, *Chem.–Eur. J.*, 2018, **24**, 8708–8716.
19 Z. Xiao, Y. Zhou, H. Hosono, T. Kamiya and N. P. Padture, *Chem.–Eur. J.*, 2018, **24**, 2305–2316.
20 R. Sarmiento-Perez, T. F. Cerqueira, S. Körbel, S. Botti and M. A. Marques, *Chem. Mater.*, 2015, **27**, 5957–5963.
21 S. Körbel, M. A. Marques and S. Botti, *J. Mater. Chem. C*, 2016, **4**, 3157–3167.
22 J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti and M. A. Marques, *Chem. Mater.*, 2017, **29**, 5090–5103.
23 S. Körbel, M. A. Marques and S. Botti, *J. Mater. Chem. A*, 2018, **6**, 6463–6475.
24 G. Hautier, A. Miglio, G. Ceder, G.-M. Rignanese and X. Gonze, *Nat. Commun.*, 2013, **4**, 2292.
25 T. F. Cerqueira, S. Lin, M. Amsler, S. Goedecker, S. Botti and M. A. Marques, *Chem. Mater.*, 2015, **27**, 4562–4573.
26 J. Shi, T. F. Cerqueira, W. Cui, F. Nogueira, S. Botti and M. A. Marques, *Sci. Rep.*, 2017, **7**, 43179.
27 N. Sarmadian, R. Saniz, B. Partoens and D. Lamoen, *Sci. Rep.*, 2016, **6**, 20446.
28 A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde and C. Wolverton, *Chem. Mater.*, 2016, **28**, 5621–5634.
29 S. Chakraborty, W. Xie, N. Mathews, M. Sherburne, R. Ahuja, M. Asta and S. G. Mhaisalkar, *ACS Energy Lett.*, 2017, **2**, 837–845.
30 R. Armiento, B. Kozinsky, G. Hautier, M. Fornari and G. Ceder, *Phys. Rev. B*, 2014, **89**, 134103.
31 A. van Roekeghem, J. Carrete, C. Oses, S. Curtarolo and N. Mingo, *Phys. Rev. X*, 2016, **6**, 041061.
32 G. Volonakis, M. R. Filip, A. A. Haghighirad, N. Sakai, B. Wenger, H. J. Snaith and F. Giustino, *J. Phys. Chem. Lett.*, 2016, **7**, 1254–1259.

33 M. R. Filip, X. Liu, A. Miglio, G. Hautier and F. Giustino, *J. Phys. Chem. C*, 2017, **122**, 158–170.
34 J. Kangsabanik, V. Sugathan, A. Yadav, A. Yella and A. Alam, *Phys. Rev. Mater.*, 2018, **2**, 055401.
35 H.-S. Kim, J.-Y. Seo and N.-G. Park, *ChemSusChem*, 2016, **9**, 2528–2540.
36 F. Li and M. Liu, *J. Mater. Chem. A*, 2017, **5**, 15447–15459.
37 C. Zhang, L. Gao, S. Teo, Z. Guo, Z. Xu, S. Zhao and T. Ma, *Sustainable Energy Fuels*, 2018, **2**, 2419–2428.
38 E. Meyer, D. Mutukwa, N. Zingwe and R. Taziwa, *Metals*, 2018, **8**, 667.
39 A. H. Slavney, T. Hu, A. M. Lindenberg and H. I. Karunadasa, *J. Am. Chem. Soc.*, 2016, **138**, 2138–2141.
40 T. Deng, E. Song, Y. Zhou, L. Wang and Q. Zhang, *J. Mater. Chem. C*, 2017, **5**, 12422–12429.
41 P. Harikesh, H. K. Mulmudi, B. Ghosh, T. W. Goh, Y. T. Teng, K. Thirumal, M. Lockrey, K. Weber, T. M. Koh, S. Li, *et al.*, *Chem. Mater.*, 2016, **28**, 7496–7504.
42 G. Kresse and J. Furthmüller, *Comput. Mater. Sci.*, 1996, **6**, 15.
43 G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169.
44 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. a. Persson, *APL Mater.*, 2013, **1**, 011002.
45 J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *JOM*, 2013, **65**, 1501–1509.
46 P. Blöchl, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953.
47 J. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.
48 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
49 G. Bergerhoff and I. Brown, in *Crystallographic Databases*, ed. F. Allen, G. Bergerhoff and R. Sievers, International Union of Crystallography, Chester, 1987, vol. 360, pp. 77–95.
50 A. Belsky, M. Hellenbrandt, V. L. Karen and P. Luksch, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 364–369.
51 A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106.
52 G. K. Madsen and D. J. Singh, *Comput. Phys. Commun.*, 2006, **175**, 67–71.
53 W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson and G. Ceder, *Sci. Adv.*, 2016, **2**, e1600225.
54 H. Glawe, A. Sanna, E. Gross and M. A. Marques, *New J. Phys.*, 2016, **18**, 093011.
55 S. Vasala and M. Karppinen, *Prog. Solid State Chem.*, 2015, **43**, 1–36.
56 A. Kojima, K. Teshima, Y. Shirai and T. Miyasaka, *J. Am. Chem. Soc.*, 2009, **131**, 6050–6051.
57 Q. Sun, J. Wang, W.-J. Yin and Y. Yan, *Adv. Mater.*, 2018, **30**, 1705901.
58 Y. Li, J. Sun and D. J. Singh, *Phys. Rev. Mater.*, 2018, **2**, 035003.
59 E. T. McClure, M. R. Ball, W. Windl and P. M. Woodward, *Chem. Mater.*, 2016, **28**, 1348–1354.

Chapter 4 Discovery of New Mixed Anion Perovskites

In the following publication "**A high-throughput study of oxynitride, oxyfluoride and nitrofluoride perovskites**" [WPhD2][2], we aim to search stable mix anion perovskites for photovoltaic applications.

The ternary inorganic perovskites ABC_3 are one of the most extensively studied families. As we have discussed in the preceding Chapters, in this prototype, there are an enormous amount of combinations of elements, offering an excellent opportunity to tune the properties, such as thermodynamic stability, lattice constant, band gap, etc., by varying the compositions. Several strategies can be used to go beyond ternary compositions and explore related materials for potential applications. For example, the A site could be occupied by organic molecules, leading to the hybrid perovskites. The hybrid perovskites' efficiency surpasses CdTe or CIGS (copper indium gallium selenide), and is comparable to top values achieved with single-crystalline silicon [175]. Moreover, the B sites can be occupied by two different cations to form the double perovskites, as we have shown in the last Chapter. In this Chapter, we focus on mixing two anions in the C sites. The ratio between the two anions X and Y can be controlled in experiments, but for the sake of simplicity and representation, we consider here the ABX_2Y-type of compositions.

The choice of A and B elements runs over the periodic table up to bismuth (excluding the noble gases and the lanthanides other than La). The anions X and Y are N, O, and F, thus in total six possible combinations. The total number of compositions is 6×3906 different stoichiometries. We manage to recover most of the experimentally known mixing anion perovskites with a 250 meV/atom threshold for distance to the convex hull (E_{hull}) and a lot more new compositions that have E_{hull} less than 100 meV/atom. The relatively large values of E_{hull} originate in the fact that we only consider the 5-atoms unit-cell of perovskite (the anions are fully ordered), which is usually not the case in experiments [176].

To study the effect of disorder, we construct supercells using the cluster expansion method implemented in ATAT package [177]. We restrict the size of the supercell to containing a maximum of 20 atoms. We explore all possible symmetrized configurations of X and Y atoms filling C sites (Wyckoff $3d$ position) with a 2:1 ratio. The number of configurations can be huge depending on the symmetry of the

[2]Reproduced with permission from the Royal Society of Chemistry.

optimized 5-atoms cell, so we select only $SrTaO_2N$, $LaTaN_2O$, $RbPbF_2O$, $RbBiO_2F$, $KFeF_2O$, $BaVO_2F$, and $LaMgF_2N$ as representatives. The different ordering of the anions leads to a spread of stabilization compared to fully ordered 5-atoms cell from 15 meV/atom (for $RbPbF_2O$) to 128 meV/atom (for $LaMgF_2N$). Note here we do not consider any entropy effect from disorder, which could further stabilize the mixed anion system at ambient conditions.

However, the disorder of anion is not the only effect that could affect the stability of mix-anion systems. For some compositions, it is certainly possible for them to prefer another structure. If the convex hull we used to evaluate the stability is complete, this should not become an issue. But for the quaternary with O-F, F-N, and O-N mixing anions, only limited hull information is known *a priori*. As discussed in Chapter 2, global structural prediction techniques should be able to predict the ground-state structure for a given composition but are too expensive for a high-throughput search. Therefore, we choose an alternative procedure by including as many competing structure prototypes with composition ABX_2Y. We get a list of 20 prototype systems with X and Y as non-metals from ICSD [10]. We firstly restrict the search for the seven systems discussed above on disorder effect, but we include both the cases of ABX_2Y and BAX_2Y. Other than perovskites, two new prototypes $PtCOI_2$ (ICSD#68098) and $CaBiO_2Cl$ (ICSD#84635) turn to be the ground state of $KFeF_2O$ and $LaMgF_2N$, respectively. We further use these two prototypes as well as the three most stable anion-configuration predicted above for $LaTaN_2O$, $SrTaO_2N$, $RbBiO_2F$, and $RbPbF_2O$, to re-evaluate the stability of all the (meta-)stable candidates selected based on the E_{hull} of 5-atoms cell. The results again show stabilization effects on the mix-anion systems varying from 0 to 250 meV/atom.

We have to note that we only consider five competing prototypes. Thus the E_{hull} could be changed drastically if more data on mix-anion systems are available. Unfortunately, at the time of this work, there was no systematic scan of all (ternary/quaternary) prototypes available. Recently, another work in the group managed to make significant progress on this issue, and we can revisit the stability by using a much more complete hull based on millions of DFT calculations guided by ML model pre-filtering results [WPhD12].

We also calculate the band structures for several selected candidates, and we find that the conduction bands are highly dispersive, leading to rather small electron-effective masses. However, unfortunately, the hole bands are considerably flatter. Although

those mix-anion perovskites are not expected to be suitable p-type semiconductors, mix-anion perovskites still have quite a variety on the width of band gaps. The PBE level gaps range from 0 to nearly 5 eV, confirming there is a potential ground for further band engineering via anion alloying.

Journal of Materials Chemistry A

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: J. Mater. Chem. A, 2021, 9, 8501

A high-throughput study of oxynitride, oxyfluoride and nitrofluoride perovskites†

Hai-Chen Wang, [a] Jonathan Schmidt,[a] Silvana Botti [b] and Miguel A. L. Marques *[a]

Perovskite solar devices are nowadays the fastest advancing photovoltaic technology. Their large-scale application is however restrained by instability and toxicity issues. Alloying is a promising way to stabilize perovskites, optimizing at the same time their absorption and charge-transport properties. We perform an extensive computational study of the thermodynamic stability and electronic properties of oxynitride, oxyfluoride and nitrofluoride perovskites. We consider quaternary stoichiometries of the type ABX_2Y, where A and B are any elements of the periodic table and X and Y are nitrogen, oxygen, or fluorine. As a starting point we explore the composition space using a simple five-atom perovskite unit cell. We then filter the candidate compositions according to their distance to the convex hull of thermodynamic stability. For the most stable systems, we then investigate other prototype structures, including more complex perovskite phases that allow for octahedral distortions, and a few non-perovskite geometries. Furthermore, for some paradigmatic cases, we study the effect of disorder by exhaustive enumeration of all possible disordered stoichiometric phases with up to 20 atoms in the unit cell. Our calculations are in very good agreement with data for experimentally known mixed anionic compounds, and predict a series of novel stable (perovskite and non-perovskite) oxynitride and oxyfluoride phases, including some with unexpected chemical composition, and one single nitrofluoride compound. Finally, we calculate and discuss the electronic properties of these compounds and their potential for application as photovoltaic absorbers.

Received 4th November 2020
Accepted 17th February 2021

DOI: 10.1039/d0ta10781f

rsc.li/materials-a

Introduction

Perovskites are one of the best known and more extensively studied families of compounds. They possess the general formula ABX_3, where X is a halide, a chalcogen, or even nitrogen, and A and B are two cations. Despite numerous applications of perovskites in the most diverse fields of physics and materials science,[1–4] only a restricted number of experimentally accessible ternary systems[5] exist. There are several possibilities to go beyond this limitation, and open the way to new materials with improved properties. For example, one can fill the A sites with organic molecules, leading to organic–inorganic hybrid perovskites such as $CH_3NH_3PbI_3$. Hybrid perovskites have attracted enormous interest in the past few years, in particular due to their application as absorbers in high-efficiency photovoltaic devices.[6–8] This is due to their unique properties, such as their high tolerance to defects,[9] the origin of which is still under debate among sp antibonding coupling,[8] polarons,[10] and lattice softness.[11] Alternatively, one can fill the B site with two different cations, leading to the so-called double perovskites.[12–14] These have, for instance, been proposed as absorbing layers for photovoltaics or as p-type transparent conductive oxides.[15–18] Another possibility to obtain quaternary perovskites, that we address in this article, is to mix more than one anion in the X position, leading to compositions of the type ABX_2Y.[19]

Mixed anion inorganic compounds are a versatile family of materials that contain more than one anionic species in a single phase.[20] The different radii and oxidation states of the two anions offer extra degrees of freedom with respect to the single-anion phase, enabling further control and tuning of electronic properties. In the context of perovskites, the most interesting and also the most studied systems are oxynitride and oxyfluoride compounds.

Several quaternary oxynitrides and oxyfluorides have already been synthesized and characterized in the literature. The most common methods for synthesis are solid-state reactions, low-temperature fluorination or high-pressure synthesis.[21,22] In solid-state reaction methods a mixture of metal oxides and nitrides or fluorides is simply heated in a furnace. One expects that high pressure stabilizes quaternary perovskites, as it suppresses the decomposition to oxides and nitrogen gas, and

[a]Institut für Physik, Martin-Luther-Universität Halle-Wittenberg, 06120 Halle (Saale), Germany. E-mail: miguel.marques@physik.uni-halle.de

[b]Institut für Festkörpertheorie und -Optik, Friedrich-Schiller-Universität Jena, European Theoretical Spectroscopy Facility, Max-Wien-Platz 1, 07743 Jena, Germany

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0ta10781f

Table 1 Calculated properties for ABN_2O materials. We list the composition, the Goldschmidt tolerance factor t, the τ factor, the energy distance to the convex hull of the simple five-atom cell ($E_{hull}^{(5)}$ in meV per atom), the most stable phase we found (according to the labels defined in Fig. 4), the band gap calculated with the PBE approximation (in eV), the total magnetization of the unit cell per atom (in Bohr magnetons), and the experimental bibliographic reference when available. We only show values of t and τ for materials where the oxidation states of the cations were clearly defined, and for which we had values for the ionic radii. The oxidation states were obtained with PYMATGEN.[41] The values of the band gap in parentheses are experimental results. Note that the PBE approximation underestimates the band gaps, but as we can see from comparison with the experimental numbers the error is systematic. This table includes only the most relevant materials. For more data, please refer to the ESI

Material	t	τ	$E_{hull}^{(5)}$	Str.	E_{hull}	E_{gap}	Mag.	Ref.
$BaReN_2O$	0.99	3.49	125	c	−11	0	0	
$CaReN_2O$	0.87	4.47	146	a	−5	0	0	
$KReN_2O$	1.01	3.71	248	f	7	1.94	0	
$LaNbN_2O$	0.84	3.74	127	a	−32	1.12	0	44
$LaReN_2O$	0.87	2.83	128	a	22	0	0	
$LaTaN_2O$	0.84	3.74	80	a	−45	1.29 (1.9, 2.1)	0	44–49
$LaTcN_2O$	0.86	3.09	170	a	47	0	0	
$NaReN_2O$	0.88	4.09	213	f	46	1.89	0	
$SrReN_2O$	0.93	3.78	81	c	−53	0	0	

allows the reactions to be carried out at higher temperature.[21] Experimentally synthesized compounds are presented in bold in Tables 1–5.

The interest in oxynitride perovskites comes from the numerous possible applications of these compounds. The smaller electronegativity of nitrogen with respect to oxygen leads to band gaps in the visible range, opening the way to a wealth of opto-electronic applications.[23] In fact, and in contrast to oxide perovskites that are usually colourless, these quaternary systems display bright coloring in a diverse color spectrum, enabling their use as, for example, pigments or phosphors[24–26] and photocatalysts.[22,23]

Initially, investigation into oxyfluoride perovskites was motivated by the discovery of superconductivity at 46 K in the cuprate $Sr_2CuO_2F_{2+x}$.[27–29] Oxifluoride systems can also possess interesting magnetic properties. For example, due to the interaction between the Fe^{3+} ions, $BaFeO_2F$, $SrFeO_2F$, and $PbFeO_2F$ exhibit magnetic (antiferromagnetic) ordering until a temperature of around 645 K,[30] 685 K,[31] and >500 K,[32] respectively. Iron-based oxyfluoride perovskites were also shown to exhibit multiferroic behavior.[33]

In this article, we provide a comprehensive computational study of oxynitride, oxyfluoride, and nitrofluoride perovskites. Our objective is threefold: (i) to provide a list of novel compositions that could be likely experimentally synthesized in the perovskite phase; (ii) to provide physical insight into the problem of disorder in these systems; and (iii) to study their electronic properties. We follow a systematic approach to unveil interesting materials that are not simple, evident substitutions of well-studied systems. This is particularly important for nitrofluorides, as no such perovskites are experimentally known at the moment. Our computational tool of choice is density-functional theory (DFT), a quantum approach to calculate the structural and electronic properties of materials which has demonstrated over the years unparalleled accuracy combined with reasonable computational costs.

Clearly, performing a study of the complete chemical space for oxynitride, oxyfluoride, and nitrofluoride perovskites including the effects of distortion, disorder, pressure, temperature, *etc.* is way beyond current computational possibilities. Therefore, we will follow a stepwise procedure conceived to be at the same time predictive and affordable. We start by looking

Table 2 Calculated properties for ABO_2N materials. Legend as in Table 1

Material	t	τ	$E_{hull}^{(5)}$	Str.	E_{hull}	E_{gap}	Mag.	Ref.
$BaNbO_2N$	0.95	3.54	65	a	−22	1.25 (1.8)	0	46 and 47
$BaReO_2N$	0.98	3.47	94	c	−41	0	0	
$BaTaO_2N$	0.95	3.54	33	b	−44	1.43 (1.8)	0	46 and 47
$BaTcO_2N$	0.97	3.48	82	c	−54	0	0	
$CaNbO_2N$	0.83	5.34	193	a	22	1.81 (2.1)	0	46 and 47
$CaReO_2N$	0.86	4.67	164	b	−5	0.41	0	
$CaTaO_2N$	0.83	5.34	146	a	13	1.67 (2.4)	0	46–48 and 50
$CaTcO_2N$	0.85	4.86	176	b	−7	0.44	0	
$KReO_2N$	1.01	3.64	124	b	25	0	0	
$LaHfO_2N$	0.81	5.38	177	a	20	2.53	0	
$LaTaO_2N$	0.83	4.57	220	b	140	0	0	
$LaTiO_2N$	0.86	3.13	121	a	30	1.64 (1.9)	0	44, 49 and 51
$LaZrO_2N$	0.81	5.70	260	a	45	2.53	0	51
$LiReO_2N$	0.78	5.76	223	a	40	0	0	
$NaReO_2N$	0.88	4.09	92	b	−7	0	0	
$PbReO_2N$	0.92	3.80	171	c	34	0	0	
$SrNbO_2N$	0.89	4.07	72	a	−21	1.37 (1.9)	0	46 and 47
$SrReO_2N$	0.92	3.83	87	c	−47	0	0	
$SrTaO_2N$	0.89	4.07	65	b	−14	1.68 (2.1)	0	46–48
$SrTcO_2N$	0.91	3.89	95	c	−47	0	0	

at the simple, five-atom perovskite unit cell. The most interesting systems, from the point of view of thermodynamic stability, are then selected by comparison with the experimental data. We then take into account possible distortions by using more complex prototype structures. Disorder is studied by using exhaustive enumeration methods. Finally, we calculate and discuss the physical properties of some selected compounds.

Methods

We performed DFT calculations using the VASP code,[34,35] where all parameters were set to guarantee compatibility with the data available in the materials project database.[36] We used the PAW[37] 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 *k*-points per atom (except where explicitly stated). All forces were converged to better than 0.005 eV $Å^{-1}$. All calculations were performed with spin-polarization using the Perdew–Burke–Ernzerhof[38] (PBE) exchange–correlation functional, with the exception of oxides and fluorides containing Co, Cr, Fe, Mn, Mo, Ni, V, and 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-states. Band structures were also calculated with the HSE06 functional.[39]

We prepared all possible compositions of type ABX_2Y, where A and B run over the periodic table up to bismuth (with the exception of the noble gases, including La but removing the other lanthanides), and X and Y are N, O, and F. We used in a first instance the simple five-atom unit cell shown in Fig. 1. Considering the 6 possible combinations of X and Y, this leads to 6 × 3906 different stoichiometries. We optimized the geometry of each one of these structures and calculated their formation energy and used them to build the convex hull of thermodynamic stability using PYMATGEN.[41] This robust open-source Python library for materials analysis is widely used in computational studies for a variety of tasks, including the visualization of calculations and the generation of standardized input files. In our case, we use PYMATGEN to query the materials project database and to calculate the energy distance to the convex hull. The construction of the convex hull considers all possible decomposition channels (in elementary, binary, ternary, and quaternary phases) present in the materials project database,[36] complemented with the compounds found in ref. 42. Specifically, this means that the formation energy of each ABX_2Y perovskite is compared to the formation energy of all stable crystalline phases of the A–B–X–Y phase diagram.

To study the effect of disorder, we systematically constructed supercells using the software included in ATAT.[43] We restricted the unit cells to a maximum of 20 atoms (4 formula units), and explored all possible ways to fill the Wyckoff 3d position with the X and Y atoms that respected the X_2Y stoichiometry. Equivalent unit cells that were mapped by a symmetry operation were automatically discarded by ATAT.

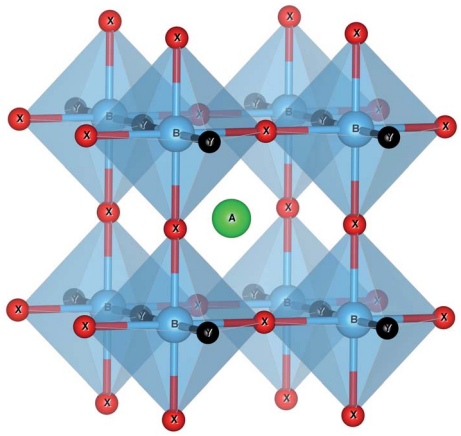


Fig. 1 The crystal structure of the ABX_2Y perovskite used for the high-throughput search. The orange ball denotes the A atom, while cyan balls are B atoms, green balls X atoms and pink balls Y atoms. The unit cell is tetragonal with space group *P*4/*m*2/*m*2/*m* (#123). Image produced with VESTA.[40]

Exploration of the chemical space

It is instructive to analyze the distance to the convex hull of stability (E_{hull}) for all chemical compositions when we use the five-atom perovskite unit cell. A histogram with this information can be found in Fig. 2. Although this plot does not give us information on specific materials, it does give us invaluable insights into the chemistry of inorganic perovskites. Interestingly, the curves for the different anion compositions exhibit different behaviors. For ABO_2N and ABN_2O the histograms are less asymmetric and are centered at around 1.5 eV. There is also a clear difference between ABO_2N and ABN_2O, with the former yielding more stable compounds than the latter. The histograms for the oxyfluorides rise very steeply until around 1 eV and then decay slowly until ∼4 eV. No noticeable difference in stability can be seen for ABO_2F and ABF_2O. The nitrofluorides display very few systems with a small distance to the hull, and show a large difference between ABN_2F and ABF_2N, with the latter yielding considerably more stable structures. The reduced stability of the ABN_2Y crystal phases can be understood by noticing that the −3 standard oxidation state of nitrogen implies that cations A and B together have to compensate for at least a −7 valence. While this is certainly possible, the number of such combinations of A and B is considerably smaller than in the case of anions with lower charge states. Furthermore, the larger difference of ionic radii between N and F (with respect to the N–O and O–F pairs) can also lead to geometrical instabilities, thereby increasing the formation energy of nitrofluoride systems.

The most important information that we can obtain from the distance of the formation energy to the convex hull is which materials are predicted by theory to be stable. It is true that experimentally one can synthesize crystal phases that are not thermodynamically stable (*i.e.*, not on the convex hull of stability), but the difficulty in realizing such phases increases considerably with their energy distance to the hull. We will therefore turn our attention to the materials contained in the left extreme of the distributions in Fig. 2. The selected

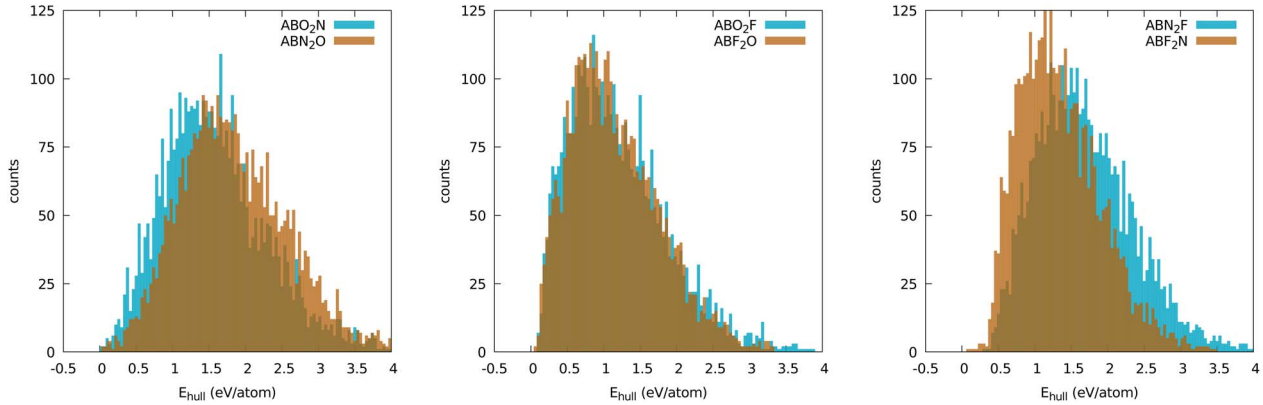


Fig. 2 Distribution of the distances to the convex hull of all oxynitride (left), oxyfluoride (center), and nitrofluoride (right) perovskites. The bins have a size of 40 meV per atom.

compounds are listed, together with information on the experimentally realized phases, in Tables 1–5 and in the ESI.†

Lowest-energy compositions

Before we start discussing our results, we have to define the criteria for filtering in low-energy compositions for further analysis. Typically, this is done by setting a reasonable threshold in the distance to the convex hull. However, several considerations have to be taken into account in our case. First, the five-atom structure we used in the high-throughput search can be further stabilized by distortion (tilting or rotation of the octahedra as common in many perovskites, for example) or by rearranging the atomic positions of the X and Y atoms. As we will see in the following, this can sometimes lead to a decrease of the formation energy of about hundred meV per atom. Then, we can expect a reasonable degree of disorder in the occupation of the X and Y sites, leading to a further decrease of the free energy due to configurational entropy. This term is typically of the order of tens of meV per atom at room temperature.[64] Note that there are still other (de)stabilization mechanisms, such as defects, temperature, pressure, *etc.*, that also lead to corrections to the free-energy. Finally, we should consider the error of the PBE approximation in the estimation of formation energies,[65–68] and hence in the determination of distances to the convex hull. This means that materials with a small positive distance to the hull might still be stable in the experiment.

In order to alleviate these issues, we decided to use a pragmatic approach. We considered our calculations for the few oxynitride and oxyfluoride phases that were synthesized experimentally and observed that all these systems have formation energies close to the convex hull of thermodynamic stability. This fact validates on the one hand our approach to find new materials, and on the other hand it provides us a valid way to set a stability threshold. The experimentally known stable composition with the highest distance to the hull in our five-atom unit cell is $LaZrO_2N$ at 260 meV per atom above the convex hull (although most other experimental compounds lie well below 150 meV per atom). This number is considerably higher than the usual criterion for metastability,[69,70] but one should keep in mind that we are filtering out compounds before having included stabilizing energy contributions coming from octahedral distortions and entropic effects. We will therefore take 260 meV per atom as the maximum distance to the convex hull of stability and pass on for further analysis only compounds that satisfy this condition.

In Tables 1–5 we summarize the most relevant results. The remainder of the data can be found in the ESI.† All structures and a summary of the calculations can be downloaded from our website.‡ At a first glance we find a variety of compositions below the 260 meV per atom stability threshold. The stability of perovskites is often discussed based on the Goldschmidt tolerance factor[71]

$$t = \frac{r_A + r_{anion}}{\sqrt{2}(r_B + r_{anion})}, \quad (1)$$

where r_A, r_B, and r_{anion} are the ionic radii of the A and B cations, and of the anion. More recently,[72] a novel data analytics approach has led to the proposition of a new factor

$$\tau = \frac{r_{anion}}{r_B} n_A \left(n_A - \frac{r_A/r_B}{\log(r_A/r_B)} \right), \quad (2)$$

where n_A is the oxidation state of A. For an experimental dataset of 576 ABX_3 materials, it was found that $0.825 < t < 1.059$ gives a classification accuracy of 74%, while $\tau < 4.18$ has an accuracy of 92%.

To use any of these formulae for mixed anions, we must decide on which value of r_{anion} to use. In line with the suggestion of ref. 72 we decided to use the arithmetic average $r_{anion} = (2r_X + r_Y)/3$. Note that, however, it has been pointed out that the Goldschmidt factor using r_{anion} can fail to capture the stability trends in mixed anionic perovskites[53,73] and pyrochlores.[74] The geometric mean has also been used to approximate the radius of a site with two ions,[75] and more complicated factors that involve, *e.g.*, octahedral factors and atomic packing fractions,

‡ https://tddft.org/bmg/data.php

have also been proposed.[75,76] For simplicity, however, we decided to build our analysis in the traditional t and on τ.

All the considered materials have values of Goldschmidt tolerance factor between 1.09 (for $CsCoF_2O$) and 0.78 (for $LiReO_2N$ and $NaInF_2O$). All in all, this range of [0.78, 1.09] is perfectly consistent with the usual range reported for perovskites.[72] Concerning τ we find that 79% of our low-energy phases have $\tau < 4.18$ and should therefore be stable perovskites according to ref. 72. This percentage of true positives is reasonable, but considerably smaller than the stated accuracy of 92%. In view of this analysis, and considering that it is nowadays possible to perform efficient high-throughput DFT calculations and apply sophisticated machine learning models,[77] it is unclear what is the benefit of using oversimplified empirical models for the prediction of novel stable materials.[78]

One last note concerning materials containing lanthanide atoms: as we removed these elements from our high-throughput search they do not appear in our results. However, one can reasonably expect that materials predicted to be low energy with La are also close to the hull after substitution of La by other lanthanides due to their chemical similarity.[79,80]

Oxynitrides

In what concerns oxynitride compounds of the ABN_2O type, we find 16 systems that satisfy our filtering conditions. Interestingly, they exhibit combinations of cations with diverse oxidation states. For example, I–VII as in $KReN_2O$, II–VI as in $CaReN_2O$, or III–V as in $LaReN_2O$. The lowest of this (at 69 meV per atom above the hull) is $LaTaN_2O$, that has already been experimentally synthesized.[44–49] At the B site, we find mostly a group V (Sr or Ba) or a group VII (Tc or Re) element. Perhaps surprisingly, the chemical element that yields more stable compounds is Re, which can combine with Na, K, Ca, Sr, Ba, La, and Pb. In view of the fact that there are, at the moment, no known Re-based oxynitrides, this finding can open the door to a completely new family of materials. We note that many of these compounds, although having low energy, require cations in less common oxidation states. Therefore, we can probably expect that, if synthesized, they will be off-stoichiometric, either due to the variation in the O–N ratio, or due to the creation of vacancies.

The other stoichiometric oxynitride family, $ABNO_2$, has considerably lower formation energies, so we find many more systems (35) within our energy threshold. In this list, we find cations in the oxidation states I–VI (as in $KReO_2N$), II–V (as in $BaTaO_2N$), and III–IV (as in $LaTiO_2N$). The most likely element that we find at the B-site is again Re, that can be combined with (in ascending order of formation energy) Sr, Na, Ba, K, Ca, Pb, Li, and Rb. Interestingly, this list includes almost all alkali metals, although Li and Rb show clearly decreased stability, but not La, as $LaReNO_2$ appears at 324 meV per atom above the hull. We find a similar situation with Tc at the B-site, yielding stable systems for A = Sr, Ba, Na, Ca, Pb, K, and La. Also many systems with Nb (A = Ba, Sr, Ca, La) and Ta (A = Sr, Ba, Ca, La, Pb, Sn) have small distances to the convex hull. When La is at the B-site, we find stable systems with the A-site being a group IV element (Ti, Zr, or Hf). Finally, we also find systems such as {Sr, Ba}{Mo, Ru}O_2N and {Sr, Ba, Na}OsO_2N.

Oxyfluorides

We now turn our attention to candidate oxyfluoride perovskites. At a glance these systems appear to be considerably more stable than oxynitrides, with 124 compositions of $ABFO_2$ and 134 compositions of ABF_2O below our stability threshold.

For $ABFO_2$ the most common oxidation state of the cations is I–III, although a few rare compounds with II–II (such as $BaMgF_2O$) or III–I (such as $LaLiF_2O$) do appear. The most likely elements occupying the A-site are Sr, Ba or an alkali-metal (in particular K and Rb, and to a lesser extent Na and Cs). The alkali-earths form low-energy oxyfluoride perovskites with a set of mostly first-row metals at the B-site. Particularly interesting are compounds with Mn, Fe, Co, Ni, *etc.* that should lead to materials with magnetic order. When Cs, Rb, and K are found at the A-site, one encounters at the B-site a transition metal. The ones leading to more stable compositions are the heavy elements Bi and Pb or group IV elements (Ti, Zr, or Hf). For Na, on the other hand, the B-site should contain lighter elements such as Ti, V, or Al. We find furthermore several compositions with Pb, Tl, and Ag at the A-site.

The materials that have already been synthesized (see Tables 3 and 4) are consistent with our predictions. We should note, however, that ref. 53 attempted the high pressure synthesis of

Table 3 Calculated properties for ABF_2O materials. Legend as in Table 1

Material	t	τ	$E_{hull}^{(5)}$	Str.	E_{hull}	E_{gap}	Mag.	Ref.
$AgCuF_2O$	0.93	3.75	176	f	33	0	0	
$AgFeF_2O$	0.89	3.85	152	e	64	1.18	1.00	52
$AgGaF_2O$	0.90	3.80	161	f	42	1.16	0	
$BaLiF_2O$			247	d	34	3.38	0	
$CsBiF_2O$	0.90	3.58	120	b	3	2.60	0	
$CsCaF_2O$			218	e	40	3.01	0	
$CsHgF_2O$			176	e	−20	0.55	0.20	
$CsPbF_2O$			64	d	50	1.12	0	
$CsSbF_2O$	1.01	3.23	230	b	30	3.83	0	
$CsSrF_2O$			242	d	44	2.83	0	
$KAgF_2O$	0.92	3.55	198	e	13	0.75	0.20	
$KAlF_2O$	1.02	3.57	218	f	43	4.51	0	
$KAsF_2O$	1.00	3.51	255	e	24	4.32	0	
$KBiF_2O$	0.81	5.00	181	f	−2	3.18	0	
$KSbF_2O$	0.91	3.57	245	b	21	4.19	0	
$NaAlF_2O$	0.89	3.97	190	f	26	4.84	0	
$NaFeF_2O$	0.84	4.33	165	f	44	2.31	1.00	
$NaGaF_2O$	0.85	4.21	202	f	25	3.65	0	
$NaMnF_2O$	0.84	4.33	162	e	44	0.24	0.80	
$RbAgF_2O$	0.97	3.36	189	e	1	0.46	0.20	
$RbAsF_2O$	1.05	3.45	256	a	22	3.55	0	
$RbBiF_2O$	0.85	4.09	133	f	−17	3.24	0	
$RbCuF_2O$	1.07	3.52	170	e	46	0.99	0	
$RbHgF_2O$			167	e	−31	0.89	0.20	
$RbSbF_2O$	0.96	3.36	209	b	17	3.73	0	
$RbTcF_2O$			201	e	−24	0.03	0	
$TlBiF_2O$	0.85	4.19	165	f	21	2.87	0	
$TlGaF_2O$	1.02	3.40	214	f	40	3.24	0	
$TlSbF_2O$	0.95	3.38	222	e	18	3.08	0	
$TlYF_2O$	0.90	3.63	137	f	50	2.39	0	

$NaTiO_2F$, which has essentially the same distance to the hull as $KTiO_2F$, without success (it resulted in a mixture of NaF and TiO_2). This emphasizes that the synthesis of an oxyfluoride material is a complex dynamical process whose success cannot be determined by the simple distance to the convex hull.

Concerning ABF_2O compositions, we could find information on the synthesis of $AgFeF_2O$[52] and of the ternary charge-disproportionate $Tl^{I}Tl^{III}OF_2$ compound.[81] However, our calculations indicate that this family should be at least as common as ABO_2F. For this composition we observe either the I–IV or II–III combination of cations. In particular, we find that the alkalis Cs, Rb, and K can form low-energy compounds with a variety of metals (such as Pb, Bi, Co, Ti, Sc, *etc.*). We also report several materials with Na, but only combined with lighter, first-row cations (Co, Ti, Fe, Al, *etc.*). Finally, there are a series of systems with Ta at the A-site, and with In, Sc, Y, Fe, Co, *etc.* at the B site, and with Ag at the A-site and Fe, Ga, Co, Cu, *etc.* at the B-site.

Nitrofluorides

We also looked into the possibility of obtaining nitrofluoride perovskites. To our knowledge, no such system has been synthesized experimentally. From our results, we can conclude that only one system, specifically $LaMgF_2N$, has chances of being synthesized. The five-atom unit cell is 155 meV per atom above the hull, which is a sizeable but not insurmountable energy distance. All other compositions have an energy distance of more than 300 meV per atom from the hull.

Table 4 Calculated properties for ABO_2F materials. Legend as in Table 1

Material	t	τ	$E_{hull}^{(5)}$	Str.	E_{hull}	E_{gap}	Mag.	Ref.
$AgFeO_2F$	0.81	6.20	269	e	117	0	0.82	
$AgTiO_2F$	0.90	3.81	141	f	26	2.17 (2.8)	0	53
$AgZrO_2F$	0.85	4.14	204	a	69	1.97	0	
$BaAgO_2F$	0.91	3.89	233	f	140	0	0.03	
$BaFeO_2F$	0.95	3.50	130	b	56	1.52	1.00	30 and 54
$BaGaO_2F$	0.97	3.45	248	e	110	3.52	0	
$BaInO_2F$	0.89	4.20	146	b	82	2.05	0	21
$BaMnO_2F$	0.95	3.50	92	b	64	0	0.80	
$BaScO_2F$	0.91	3.86	90	b	13	4.11	0	55
$BaTlO_2F$	0.85	4.98	222	f	75	1.98	0	
$CsTeO_2F$	0.92	3.45	254	a	40	2.41	0	
$KGaO_2F$			255	a	14	1.06	0.10	
$KHfO_2F$	0.93	3.53	141	b	36	4.40	0	
$KNbO_2F$	0.95	3.51	124	c	28	0	0	56
$KTeO_2F$	0.83	4.47	216	e	42	3.01	0	
$KTiO_2F$	0.98	3.52	148	f	11	3.61 (3.2)	0	57 and 58
$KZrO_2F$	0.93	3.54	140	b	38	3.78	0	
$NaNbO_2F$	0.82	4.59	207	a	108	0	0	56
$NaTiO_2F$	0.86	4.18	158	f	31	3.52	0	59
$NaVO_2F$	0.87	4.10	149	f	−2	2.40	0.20	
$PbFeO_2F$	0.90	4.00	195	f	81	2.04	1.00	32 and 60
$PbMnO_2F$	0.90	4.00	155	f	70	0.52	0.80	61
$PbScO_2F$	0.86	4.78	173	f	54	2.82	0.00	62
$RbBiO_2F$			83	c	50	0	0	
$RbIO_2F$			231	a	15	2.07	0	
$RbNbO_2F$	1.00	3.38	144	c	44	0	0	
$RbTeO_2F$	0.87	3.83	182	a	10	2.38	0	
$RbTiO_2F$	1.03	3.44	205	f	2	3.59	0	
$RbVO_2F$	1.05	3.48	229	f	16	2.40	0.20	
$SrCuO_2F$	0.94	3.67	178	e	42	0.84	0	
$SrFeO_2F$	0.89	4.05	171	b	90	1.57	1.00	63
$TlIO_2F$			241	a	15	2.00	0	
$TlTeO_2F$	0.87	3.90	222	a	44	2.64	0	
$TlTiO_2F$	1.03	3.45	228	f	12	3.09	0	

Effects of disorder

Having selected the compositions that possess the smallest formation energies in the perovskite structure, we now investigate how the energy depends on the specific arrangement of the anions. To that end, we pick a few interesting systems (namely $SrTaO_2N$, $LaTaN_2O$, $RbPbF_2O$, $RbBiO_2F$, $KFeF_2O$, $BaVO_2F$, and $LaMgF_2N$) and construct all possible unit-cells with up to 20 atoms by considering the different configurations that we obtain by filling the 3d Wyckoff anionic site with the two different elements. We found 285 non-equivalent structures, for which we performed a further geometry optimization (using 2000 k-points per atom for increased precision). We note that many of these supercells are consistent with the typical deformations present in perovskites, such as tilting or rotation of the octahedra. Therefore, the resulting variations in the formation energy account for contributions coming from the different anion arrangements and the structural deformation upon relaxation.

We found that the different ordering of the anions leads to a spread of energy of 100–150 meV per atom, and to a stabilization that can be as low as 15 meV per atom (for $RbPbF_2O$) to 128 meV per atom (for $LaMgF_2N$) with respect to the five-atom unit cell. We would like to note that these numbers are for the internal energy at $T = 0$ and not for the free energy. Therefore, they do not account for the term that stems from the configurational entropy that further stabilizes disordered phases.

Three examples are shown in Fig. 3, namely $LaTaN_2O$, $SrTaO_2N$, and $RbBiO_2F$. We can see three different behaviors.

In $LaTaN_2O$ the relaxation of the anions leads to a large stabilization (of 124 meV per atom) with respect to the total

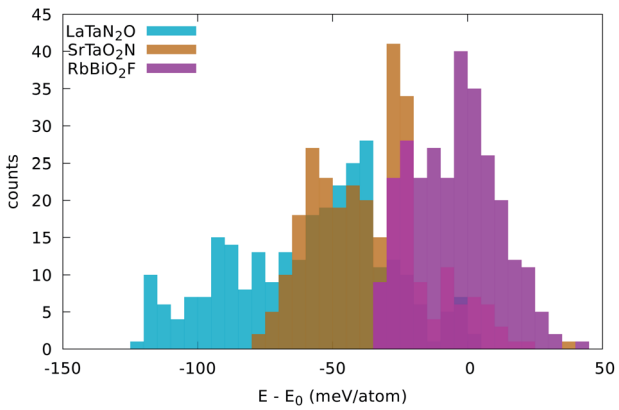


Fig. 3 Distribution of the energy of the disordered cell with respect to the energy of the five-atom unit cell. The width of the bins is 5 meV per atom.

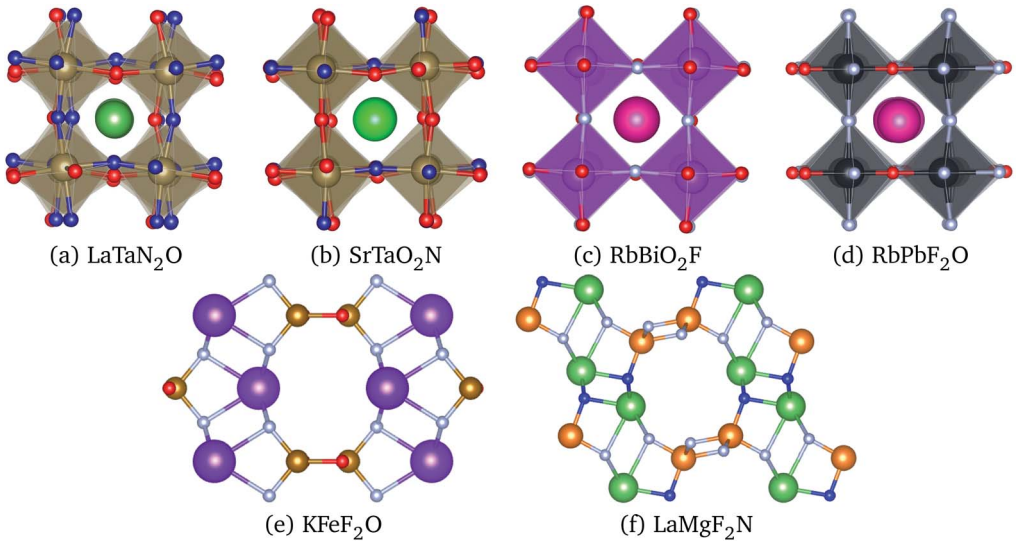


Fig. 4 Crystal structures of the lowest energy phases of (a) $LaTaN_2O$ (space group $P1$, #1), (b) $SrTaO_2N$ (space group $P3_221$, #154), (c) $RbBiO_2F$ (space group $I4/mmm$ #139), (d) $RbPbF_2O$ (space group $Pmma$, #51), (e) $KFeF_2O$ (space group $Cmcm$ #63), and (f) $LaMgF_2N$ (space group $P2_1/m$, #11). K and Bi atoms are in purple, Rb in dark pink, La and Sr in green, Ta in brown, Fe in yellow, Mg in orange, Pb in black, O in red, and F in gray and N in blue. Images produced with VESTA.[40]

energy of the simple five-atom unit cell. The lowest-energy structure we found (shown in Fig. 4) is a low-symmetry 20-atom cell, but we found 13 different anionic arrangements within 10 meV per atom from this phase. In this case the five-atom cell yields the highest energy structure. From the figure we can see the large alternating tilting of the octahedra and the distortions caused by the local mixed-anionic environment. The large distortion can also be assessed from the distributions of the atomic distances that read 3.39–3.69 for La–Ta (the half diagonal of the cube), 2.11–2.30 Å for Ta–O, and 1.92–2.20 Å for the Ta–N distance.

The compound $SrTaO_2N$ is an intermediate case. The relaxation of the anion positions in a larger supercell rearrangement leads to an energy decrease of 79 meV per atom, but we can also find supercells with energy higher by 25 meV per atom than the energy of the five-atom unit-cell. The minimum energy configuration that we found, shown in Fig. 4, has 15 atoms in the unit cell and belongs to the space group $P3_221$ (#154). Interestingly, also the lowest-energy structures of $BaVO_2F$ and $KFeF_2O$ show the same anionic arrangement. In the figure we can observe the tilting of the octahedra, which is however less pronounced than for $LaTaN_2O$. The smaller distortion is also evident from the inspection of the interatomic distances: the Sr–Ta distance is now in the range between 3.47 and 3.59 Å, the Ta–O distance is 2.00–2.15 Å, and the Ta–N distance is 1.98 Å.

Finally, the total energy of $RbBiO_2F$ assumes values essentially centered around the energy of the five-atom cell, with the lowest-energy structure 33 meV per atom below this energy. The most stable phase turns out to be a tetragonal cell with 20 atoms (space group $I4/mmm$ #139). From the small stabilization energy we can expect a small deformation of the lattice, as can be confirmed by visually inspecting Fig. 4. The Rb–Bi bond length ranges between 3.80 and 3.98 Å, the Bi–O distance is in the range 2.14–2.16 Å, and the Bi–F distance is 2.45 Å. In fact, in this structure the F atoms form perfect square motifs.

We observe that in all considered systems except $RbPbF_2O$ the minority anion is never present in opposite vertices of the octahedra, *i.e.* preferring adjacent positions.

Effects of lattice distortion: prototype search

In the previous section we considered nonequivalent configurations due to different occupation of sublattice sites and distortion of the ideal perovskite structure. However, it is certainly possible that some of the considered systems choose to crystallize in other crystallographic arrangements. Ideally, one could use global structural prediction techniques[82] that are capable of predicting the ground-state structure based solely on the chemical composition of the unit cell. Such techniques have already been used, for example, to investigate Cu, Ag, and Au ternary oxides[83] or half-Heusler compounds,[84] or nitride perovskites.[85] However, the large number of systems and the large size of the unit cells required make this approach unaffordable. Therefore, we decided for an alternative procedure that consists in trying out experimental prototype ABX_2Y crystal structures.

To this effect, we searched for stoichiometric compounds (without partial occupancy of the Wyckoff positions) within the Inorganic Crystal Structure Database (ICSD)[86] with compatible chemical compositions. We restricted the search to entries where X and Y are non-metals, but we imposed no further rules to match oxidation states, ionic radii, *etc.* The few systems found are listed in Table 6, and for many of them both X and Y are chalcogens. This again confirms that not so much is experimentally known about mixed anionic systems.

We performed geometry optimization runs for each one of these 20 prototypes, including the two cases ABX_2Y and BAX_2Y, considering again the 7 compositions included earlier in Sec. 4. We observed that several of these prototypes relaxed towards structures already studied in Sec. 4, while others resulted in very high energy phases. A couple of structures, however, turned out to be the ground state for some compositions, namely the crystal structures with ICSD references #68098 and #84635. The structure optimization of the former for $KFeF_2O$ leads to a very different geometry, depicted in Fig. 4, with space group *Cmcm* #63. This structure does not exhibit the traditional octahedral coordination of perovskites, and is 44 meV per atom lower in energy than the $P3_221$ perovskite structure. For $LaMgF_2N$ the ICSD structure #84635 led to the geometry represented in Fig. 4. This phase has the space group $P12_1/m1$ (#11) and is also not a perovskite-like structure. It is only 21 meV per atom more

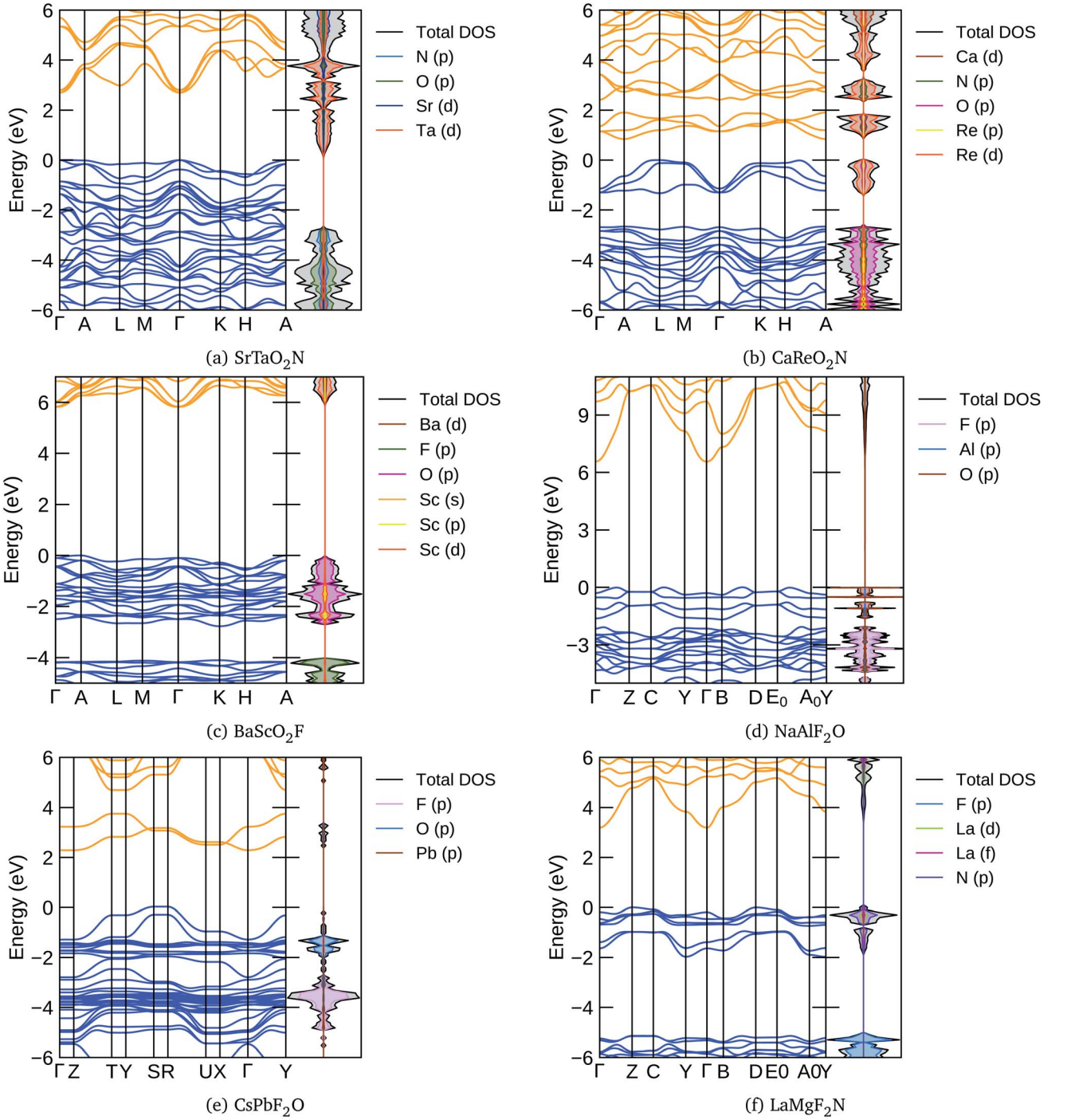


Fig. 5 The band structures and density of electronic states (DOS) of the lowest energy perovskite structures were calculated with the HSE06 functional for (a) $SrTaO_2N$, (b) $CaReO_2N$, (c) $BaScO_2F$, (d) $NaAlF_2O$, (e) $CsPbF_2O$, and (f) $LaMgF_2N$.

stable than the lowest-energy perovskite phase. Interestingly, this structure has clear similarities to the geometry of $KFeF_2O$, as can be seen in the figure.

Having identified the five structures of Fig. 4 as relevant for the low-energy phases of ABX_2Y compounds, we used them for all other compositions within the imposed energy threshold. A plot of the stabilization energy, that we define as the energy of the most stable prototype minus the energy of the five-atom cell, is depicted in Fig. 1. As for the cases discussed above, this energy can vary considerably from essentially zero to 250 meV per atom. We expected to see a correlation between the deviation of the Goldschmidt tolerance factor t from 1 and the stabilization energy. In fact, for values of $t < 0.9$ we expected the perovskite to distort from its cubic form, leading to a considerable decrease of the energy. However, we could only find a very weak correlation in the data, *i.e.* only compounds with very small values of t lowered considerably their energy by distortion. In any case, we should recall that (i) we only used 5 prototypes that clearly do not allow for all possible orthorhombic and rhombohedral distortions that may be favorable for some systems, and (ii) our energies contain two distinct contributions that unfortunately are difficult to disentangle, namely geometrical distortions and effects of disorder.

The most stable structures found for each composition can be found in Tables 1–5. We also present the distance to the convex hull of the lowest energy phase, and its electronic band gap and magnetization (per atom). We should recall that the PBE approximation tends to underestimate the band gaps by nearly a factor of two,[87,88] so real samples will have a gap larger than the one indicated in the table. Comparing with the experimental band gaps indicated in parentheses in the tables, we can see that the error is quite systematic.

Oxynitrides

We want to analyze more in detail the data in Tables 1–5. Most of the experimentally known compounds are on the convex hull of thermodynamic stability, or very close to it. This in our opinion validates the use of the distance to the convex hull as a direct measure of the probability that a certain mixed anionic perovskite can be experimentally realized.

We can also see that most systems seem to indeed crystallize in a perovskite structure, with the exception of $NaReN_2O$, $KReN_2O$, $SrTaN_2O$, $KTcO_2N$, and $SnTaO_2N$. From these non-perovskite systems, the most likely one to be realized in experiments is $KReN_2O$ which appears merely 7 meV per atom above the hull. This is a non-magnetic semiconductor, with a PBE gap of almost 2 eV.

There are a number of systems that appear listed with an ABO_2N composition, but not with an ABN_2O composition (or *vice versa*). These are, for example, the cases of $CaNbO_2N$, $CaTaO_2N$, $CaTcO_2N$, $LaHfO_2N$, $LaReN_2O$, *etc.* Sometimes both compositions appear in Tables 1–5, but one of the variants has a considerably larger distance to the hull compared to the other. Examples are $BaNbO_2N$, $BaTaO_2N$, $LaNbN_2O$, $LaTaN_2O$, *etc.* We see these results as indicating that such systems can be synthesized in the specified stoichiometries, but there is only a limited flexibility for the anion composition range. On the other hand, systems that appear with both compositions should be stable with respect to larger variations of the O/N ratio. These are particularly interesting, as they present the largest potential for the engineering of electronic (or other) properties by adjusting the anionic ratio. The best examples are $CaReO_2N$, $SrReO_2N$, $BaReO_2N$, and $KReO_2N$, although some systems with Tc in the B position or $SrTaO_2N$ are also promising. Note that this last compound is experimentally known[46–48] but not the others in this list.

Finally we notice in the list a few materials with a finite magnetic moment ($BaMoO_2N$, $KTcO_2N$, $SrMoO_2N$, $BaNbN_2O$, $BaTaN_2O$, and $SrNbN_2O$). However, none of these latter is particularly close to the hull, so we will not discuss them in more detail.

In Fig. 5 we present, as an example, the electronic band structure and density of states for $SrTaO_2N$ and $CaReO_2N$. The first crystallizes in the perovskite structure shown in Fig. 4b and presents a rather isotropic band structure with a direct band gap at Γ of 2.70 eV in the HSE06 functional. The lowest conduction bands are highly dispersive, with an effective mass of $m_e^* \sim 0.8\ m_0$. The upper valence bands are much less dispersive, which is reflected in the heavier hole mass of $m_h^* \sim 5\ m_0$. This significant difference between electron and hole masses is present in many of our systems, and has already been discussed in ref. 89. The valence states are mostly composed by p states of N and O with a small contribution coming from Ta d states, while the conduction bands have mainly Ta d character with a small O p character. The reduced hybridization between the anionic p-states and the B-metal states is probably the cause of the heavy holes. Finally, we see very few states associated with Sr in the $[-6, 6]$ eV energy windows, which is compatible with the interpretation that the A atom is fully ionized in the perovskite structure (Fig. 6).

The band structure of $CaReO_2N$ is rather different from that of $SrTaO_2N$, even if they share the same crystal structure. The band gap of 0.87 eV is indirect, with the bottom of the conduction band at A and the top of the valence band along the line connecting H and A. Both electron and hole bands are

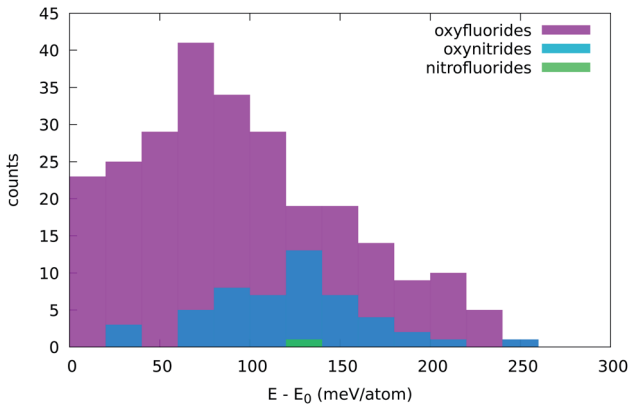


Fig. 6 Distribution of the stabilization energy, defined as the lowest energy of all prototypes used minus the energy of the five-atom cell depicted in Fig. 1. The width of the bins is 20 meV per atom.

mainly composed of Re d states hybridized with p states of O (valence band) or O and N (conduction band). The valence states are split into two manifolds, with the lowest group (starting at around 2.5 eV below the Fermi surface) mostly composed of the anionic p states with a small contribution from Re. The Ca states are found only in the conduction band above 5 eV, again indicating that this atom is ionized in this structure.

Oxyfluorides

From Tables 3 and 4 we can see that there are considerably more oxyfluoride systems, within our formation-energy cutoff, than oxynitrates. However, the lowest energy phases of many of these systems turn out to be a non-perovskite, *i.e.* one of the structures depicted in Fig. 4e and f. Furthermore, some of the experimentally known systems have slightly large distances to the convex hull (>50 meV per atom). Unfortunately, with the available data it is difficult to discern if this is due to the difficulty of describing fluorinated materials with the PBE approximation, or if disorder, defects or alloying have a greater stabilization role in these systems. Furthermore, we note that several of the experimentally known materials are magnetic. It is well known that the vast majority of magnetic semiconductors are indeed anti-ferromagnetic; however, all our calculations are performed for the ferromagnetic phase. This can also lead to an overestimation of the theoretical formation energy of typically a few tens of meV per atom.

With Ag in the A position, we find stable compositions for B = Cu, Fe, Ga, Ti, and Zr. The most stable seems to be $AgTiO_2F$, while $AgZrO_2F$ is the only one that the PBE predicts to crystallize in a perovskite structure. Filling the A site with Ba leads to the stabilization of a series of ABO_2F compositions. Most of these systems have considerable band gaps, with the largest gap of 4.11 eV found for $BaScO_2F$. One band structure of this kind is depicted in Fig. 5c. We can see that the HSE band gap of 5.82 eV between A and Γ is indirect, with the curvature of the conduction bands considerably larger than that of the valence bands. The top of the valence is mainly composed of O p-states with a smaller contribution of Sc states, while the bottom conduction has Sc d-character with a smaller component coming from Ba d and O p levels. The valence bands are split, with the top composed mainly of O p states and the bottom bands exhibiting mainly F p-character. Between the two manifolds there is a gap. These characteristics are shared by many of the oxynitride and nitrofluoride systems, ultimately due to the larger electronegativity of F and of the stronger ionic character of the cation–F bond when compared to the cation–O bond.

Some of the Ba containing materials are magnetic when the B site is occupied by a 3d metal such as Cr, Mn, Fe, or Ni. On the other hand, the only system with a perovskite structure that appears in our list is $BaLiF_2O$, with a rather large PBE band gap of 3.38 eV. As such, we do not expect that the F/O ratio can be considerably increased for Ba-based systems.

Most of the systems in Tables 3 and 4 have an alkali metal in the A position (Na, K, Rb, or Cs). At the B site we find a rather diverse set of metals, ranging from the alkali earths Ca and Sr, passing through the magnetic metals Mn and Fe, a series of transition metals, and even the semi-metal atoms As, Te, and I. Many of these materials retain the perovskite structure, but some relax into the structures of Fig. 4e and f. They are mostly insulators with considerably large band gaps that can reach 4.51 eV for $KAlF_2O$ and 4.84 eV for $NaAlF_2O$. In Fig. 5d we plot the band structure calculated with the HSE06 functional of this latter compound. We can see that the band gap is indirect, and the band structure is anisotropic, as can be expected from the crystal structure depicted in Fig. 4f. We again see a splitting of the valence bands with states with considerable F character in the lower valence band. The effective mass of the electrons is also clearly smaller than that of the holes. Note that the only AB combinations of cations that we can find in both short lists are CsSb and RbBi, so we expect that these systems are particularly resistant to variations of the F/O ratio.

Finally, interestingly, we find in the list the system $CsPbF_2O$ with a perovskite structure at 50 meV per atom above the hull. The counterpart composition, $CsPbO_2F$, appears at 103 meV per atom above the hull, but with the structure of Fig. 4f. The interest in this structure comes from the fact that the perovskite $CsPbI_3$ is the parent inorganic compound[90] for the famous halide perovskites used for photovoltaic applications.[6] Furthermore, the PBE electronic band gap of these systems is quite stable in what regards the composition, going from 1.12 eV for $CsPbF_2O$ to 1.18 eV for $CsPbO_2F$, a value perfectly suitable for absorbers in photovoltaic modules.[91] We should in fact consider that the PBE calculation underestimates the band gap, but we are also neglecting spin–orbit corrections that are sizable for heavy-element compounds and reduce the size of the gap. The HSE band structure of $CsPbF_2O$ is depicted in Fig. 5e. The band gap is indirect, with the top of the valence and bottom of the conduction bands composed of hybridized F p, O p, and Pb states. In this case the separation of the valence into two manifolds is not complete, leading to some overlap between the two sets of bands. The bottom of the conduction band, on the other hand, is separated by more than 1 eV from the rest of the conduction band.

Nitrofluorides

The calculated properties for $LaMgF_2N$ are shown in Table 5. We can see that this system is considerably stabilized by relaxing into the structure of Fig. 4f, which lies just 19 meV per atom above the hull. This is a large band gap semiconductor, with a PBE band gap of 2.26 eV. The HSE band structure of this material is presented in Fig. 5f. The highly dispersive lowest conduction band is mainly constructed from La d states, while the top valence has mostly N p character with smaller La d character. In the case of this nitrofluoride system we also see a clear splitting of the valence bands. However, due to the

Table 5 Calculated properties for ABF_2N materials. Legend as in Table 1

Material	t	τ	$E_{hull}^{(5)}$	Str.	E_{hull}	E_{gap}	Mag.	Ref.
$LaMgF_2N$	0.81	5.64	155	f	19	2.26	0	

Table 6 Entries with composition $ABXY_2$ and without partial occupancy of the Wyckoff positions found in the ICSD where X and Y are non-metals. We present the ICSD number, the chemical composition, the space group, and the number of atoms in the primitive unit cell

ICSD #	Composition	Spg.	N_{atoms}
80	$CeBiOS_2$	*P*4/*nmm* (#129)	10
2238	$LaGaOS_2$	*Pmca* (#57)	20
14 191	$CeCrOS_2$	*B*112/*m* (#12)	10
14 192	$LaCrOS_2$	*Pbnm* (#62)	20
38 636	$NaNbO_2F$	*Pbnm* (# 62)	20
48 024	$GaLaOSe_2$	$P2_1ab$ (#29)	20
54 075	$CeCrOSe_2$	*B*112/*m* (#12)	10
66 246	$AsPbO_2Cl$	$P2_12_12_1$ (#19)	40
68 098	$PtCOI_2$	*C*12/*c*1 (#15)	20
69 869	$CaFeO_2Cl$	*A*1*m*1 (#8)	10
84 635	$CaBiO_2Cl$	$P12_1/m1$ (#11)	10
86 229	$PbSbO_2Cl$	*Cmcm* (#63)	10
171 722	$CdSbS_2Cl$	*Pnma* (#62)	20
171 723	$CdSbS_2Br$	*C*12/*m*1 (#12)	10
244 028	$LaVSe_2O$	*C*12/*m*1 (#12)	10
411 137	$TaSrNO_2$	*I*4/*mcm* (#140)	10
411 138	$TaLaN_2O$	*C*12/*m*1 (#12)	10
413 289	$CuBiSCl_2$	*Cmcm* (#63)	10
419 916	$LaFCN_2$	*Cmcm* (#63)	10
424 505	$NaBOF_2$	*C*12/*c*1 (#15)	30

strongest electronegativity difference of N and F, the gap between the two manifolds is considerably larger than for the oxyfluoride systems.

Although $LaMgF_2N$ is the only nitrofluoride system that we predict to have possibilities to be realized experimentally, we recall that La is often easily substituted by other lanthanides (or actinides),[80] giving us hope that such nitrofluoride systems can be discovered in the future.

Conclusions

From our extensive computational study of quaternary oxynitride, oxyfluoride, and nitrofluoride perovskites it is clear that there are many more compounds that are experimentally accessible than the few systems that have been discovered to date. Some of them can be obtained by simple chemical substitutions of known compounds, such as $LaHfO_2N$ or $AgTiO_2F$ that can be obtained by substituting Zr in $LaZrO_2N$ or in $AgZrO_2F$. However, many of the predicted materials do not have an experimentally known counterpart, such as the many Re-based oxynitride systems that we found. We note that although many of the low-energy compositions seem to crystallize in a perovskite phase, some systems prefer other atomic arrangements that do not exhibit the famous octahedra. Furthermore, we show that changing the arrangement of the anions and allowing for geometrical distortions can stabilize the structure of the perovskites by up to 150 meV per atom, although this number is highly dependent on the composition. This stabilization is in many cases fundamental, as it makes the composition thermodynamically stable, even without the need to resort to entropic arguments.

Many of our systems turn out to be semiconducting or insulating, with electronic band gaps going from a fraction of an eV to several eV. Often the electron bands are highly dispersive, leading to rather small electron effective masses, while the hole bands are considerably flatter. This can be understood from the reduced hybridization between the cationic states and the anionic p levels. Furthermore, for the fluorine compounds we find that the valence states are often split, with the top of the valence characterized by a strong p-character from O or N atoms, and the lower manifold of states stemming from the F states. The splitting is considerably larger for the nitrofluoride systems than for the oxifluorides indicating that this is due to the larger electronegativity of F leading to a cation–F bond with a stronger ionic character. However, and in spite of these common features, we find a multitude of behaviors that illustrate the diversity and importance of these mixed ionic systems.

Our results confirm that alloying on the anion sublattice is a promising strategy to improve the stability and engineer the band gap of perovskite absorbers.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

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

Notes and references

1 M. Liu, M. B. Johnston and H. J. Snaith, *Nature*, 2013, **501**, 395–398.
2 S. D. Stranks and H. J. Snaith, *Nat. Nanotechnol.*, 2015, **10**, 391–402.
3 G. Schileo and G. Grancini, *JPhys Energy*, 2020, **2**, 021005.
4 D. W. deQuilettes, K. Frohna, D. Emin, T. Kirchartz, V. Bulovic, D. S. Ginger and S. D. Stranks, *Chem. Rev.*, 2019, **119**, 11007–11019.
5 J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti and M. A. L. Marques, *Chem. Mater.*, 2017, **29**, 5090–5103.
6 P. K. Nayak, S. Mahesh, H. J. Snaith and D. Cahen, *Nat. Rev. Mater.*, 2019, **4**, 269–285.
7 A. K. Jena, A. Kulkarni and T. Miyasaka, *Chem. Rev.*, 2019, **119**, 3036–3103.
8 W.-J. Yin, J.-H. Yang, J. Kang, Y. Yan and S.-H. Wei, *J. Mater. Chem. A*, 2015, **3**, 8926–8942.
9 W.-J. Yin, T. Shi and Y. Yan, *Appl. Phys. Lett.*, 2014, **104**, 063903.
10 K. Miyata, T. L. Atallah and X.-Y. Zhu, *Sci. Adv.*, 2017, **3**, e1701469.
11 W. Chu, Q. Zheng, O. V. Prezhdo, J. Zhao and W. A. Saidi, *Sci. Adv.*, 2020, **6**, eaaw7453.
12 M. Anderson, K. Greenwood, G. Taylor and k. Poepellmeier, *Prog. Solid State Chem.*, 1993, **22**, 197–233.

13 A. Hossain, P. Bandyopadhyay and S. Roy, *J. Alloys Compd.*, 2018, **740**, 414–427.
14 E. Meyer, D. Mutukwa, N. Zingwe and R. Taziwa, *Metals*, 2018, **8**, 667.
15 E. Greul, M. L. Petrus, A. Binek, P. Docampo and T. Bein, *J. Mater. Chem. A*, 2017, **5**, 19972–19981.
16 J. Kangsabanik, V. Sugathan, A. Yadav, A. Yella and A. Alam, *Phys. Rev. Mater.*, 2018, **2**, 055401.
17 E. T. McClure, M. R. Ball, W. Windl and P. M. Woodward, *Chem. Mater.*, 2016, **28**, 1348–1354.
18 H.-C. Wang, P. Pistor, M. A. Marques and S. Botti, *J. Mater. Chem. A*, 2019, **7**, 14705–14711.
19 Y. Kobayashi, Y. Tsujimoto and H. Kageyama, *Annu. Rev. Mater. Res.*, 2018, **48**, 303–326.
20 H. Kageyama, K. Hayashi, K. Maeda, J. P. Attfield, Z. Hiroi, J. M. Rondinelli and K. R. Poeppelmeier, *Nat. Commun.*, 2018, **9**, 772.
21 T. Katsumata, R. Suzuki, N. Satoh, S. Suzuki, M. Nakashima, Y. Inaguma, D. Mori, A. Aimi and Y. Yoneda, *J. Solid State Chem.*, 2019, **279**, 120919.
22 M. Ahmed and G. Xinxin, *Inorg. Chem. Front.*, 2016, **3**, 578–590.
23 M. Sakar, R. M. Prakash, K. Shinde and G. R. Balakrishna, *Inorg. Chem. Front.*, 2020, **45**, 7691–7705.
24 R.-J. Xie and H. T. B. Hintzen, *J. Am. Ceram. Soc.*, 2013, **96**, 665–687.
25 R. Aguiar, D. Logvinovich, A. Weidenkaff, A. Rachel, A. Reller and S. G. Ebbinghaus, *Dyes Pigm.*, 2008, **76**, 70–75.
26 R.-J. Xie and N. Hirosaki, *J. Adv. Mater.*, 2007, **8**, 588–600.
27 M. G. Francesconi and C. Greaves, *Supercond. Sci. Technol.*, 1997, **10**, A29–A37.
28 C. Greaves and M. G. Francesconi, *Curr. Opin. Solid State Mater. Sci.*, 1998, **3**, 132–136.
29 E. E. McCabe and C. Greaves, *J. Fluorine Chem.*, 2007, **128**, 448–458.
30 R. Heap, P. R. Slater, F. J. Berry, O. Helgason and A. J. Wright, *Solid State Commun.*, 2007, **141**, 467–470.
31 F. J. Berry, R. Heap, O. Helgason, E. A. Moore, S. Shim, P. R. Slater and M. F. Thomas, *J. Condens. Matter Phys.*, 2008, **20**, 215207.
32 Y. Inaguma, J.-M. Greneche, M.-P. Crosnier-Lopez, T. Katsumata, Y. Calage and J.-L. Fourquet, *Chem. Mater.*, 2005, **17**, 1386–1390.
33 S. T. Hartman, S. B. Cho and R. Mishra, *Inorg. Chem.*, 2018, **57**, 10616–10624.
34 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
35 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
36 A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder and K. a. Persson, *APL Mater.*, 2013, **1**, 011002.
37 P. E. Blöchl, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1994, **50**, 17953–17979.
38 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
39 A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106.
40 K. Momma and F. Izumi, *J. Appl. Crystallogr.*, 2011, **44**, 1272–1276.
41 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
42 H.-C. Wang, S. Botti and M. Marques, *npj Comput. Mater.*, 2020.
43 A. van de Walle, *Calphad*, 2009, **33**, 266–278.
44 R. Marchand, F. Pors and Y. Laurent, *Ann. Chim.*, 1991, 553–560.
45 M. Liu, W. You, Z. Lei, T. Takata, K. Domen and C. Li, *Chin. J. Catal.*, 2006, **27**, 556–558.
46 Y.-I. Kim, P. M. Woodward, K. Z. Baba-Kishi and C. W. Tai, *Chem. Mater.*, 2004, **16**, 1267–1276.
47 F. Oehler and S. G. Ebbinghaus, *Solid State Sci.*, 2016, **54**, 43–48.
48 E. Günther, R. Hagenmayer and M. Jansen, *Z. Anorg. Allg. Chem.*, 2000, **626**, 1519–1525.
49 Y.-I. Kim, PhD thesis, The Ohio State University, 2005.
50 R. Marchand, F. Pors and Y. Laurant, *Rev. Int. Hautes Temp. Refract.*, 1986, **23**, 11–15.
51 S. J. Clarke, B. P. Guinot, C. W. Michie, M. J. C. Calmont and M. J. Rosseinsky, *Chem. Mater.*, 2002, **14**, 288–294.
52 F. Takeiri, T. Yamamoto, N. Hayashi, S. Hosokawa, K. Arai, J. Kikkawa, K. Ikeda, T. Honda, T. Otomo, C. Tassel, K. Kimoto and H. Kageyama, *Inorg. Chem.*, 2018, **57**, 6686–6691.
53 Y. Inaguma, K. Sugimoto and K. Ueda, *Dalton Trans.*, 2020, **49**, 6957–6963.
54 F. J. Berry, F. C. Coomer, C. Hancock, Ö. Helgason, E. A. Moore, P. R. Slater, A. J. Wright and M. F. Thomas, *J. Solid State Chem.*, 2011, **184**, 1361–1366.
55 R. Needs and M. Weller, *J. Solid State Chem.*, 1998, **139**, 422–423.
56 W. Rüdorff and D. Krug, *Z. Anorg. Allg. Chem.*, 1964, **329**, 211–217.
57 B. Chamberland, *Mater. Res. Bull.*, 1971, **6**, 311–315.
58 T. Katsumata, H. Umemoto, Y. Inaguma, D. Fu and M. Itoh, *J. Appl. Phys.*, 2008, **104**, 044101.
59 V. Zaitsev and V. Senin, *Vestn. Otd. nauk Zemle*, 2012, **4**, 8.
60 T. Katsumata, A. Takase, M. Yoshida, Y. Inaguma, J. E. Greedan, J. Barbier, L. M. D. Cranswick and M. Bieringer, *MRS Online Proc. Libr.*, 2006, **988**, 0988-QQ06-03.
61 T. Katsumata, M. Nakashima, Y. Inaguma and T. Tsurui, *Bull. Chem. Soc. Jpn.*, 2012, **85**, 397–399.
62 T. Katsumata, M. Nakashima, H. Umemoto and Y. Inaguma, *J. Solid State Chem.*, 2008, **181**, 2737–2740.
63 F. J. Berry, X. Ren, R. Heap, P. Slater and M. F. Thomas, *Solid State Commun.*, 2005, **134**, 621–624.
64 F. Brivio, C. Caetano and A. Walsh, *J. Phys. Chem. Lett.*, 2016, **7**, 1083–1087.
65 R. Sarmiento-Pérez, S. Botti and M. A. L. Marques, *J. Chem.*, 2015, **11**, 3844–3850.

66 F. Tran, J. Stelzl and P. Blaha, *J. Chem. Phys.*, 2016, **144**, 204120.
67 V. Stevanović, S. Lany, X. Zhang and A. Zunger, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2012, **85**, 115104.
68 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 and J. Sun, *npj Comput. Mater.*, 2018, **4**, 9.
69 W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson and G. Ceder, *Sci. Adv.*, 2016, **2**, e1600225.
70 A. A. Emery and C. Wolverton, *Sci. Data*, 2017, **4**, 170153.
71 V. M. Goldschmidt, *Sci. Nat.*, 1926, **14**, 477–485.
72 C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli and M. Scheffler, *Sci. Adv.*, 2019, **5**, eaav0693.
73 G. Pilania, A. Ghosh, S. T. Hartman, R. Mishra, C. R. Stanek and B. P. Uberuaga, *npj Comput. Mater.*, 2020, **6**, 71.
74 R. Mouta, R. X. Silva and C. W. A. Paschoal, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2013, **69**, 439–445.
75 W. Li, E. Ionescu, R. Riedel and A. Gurlo, *J. Mater. Chem. A*, 2013, **1**, 12239.
76 Q. Sun and W.-J. Yin, *J. Am. Chem. Soc.*, 2017, **139**, 14905–14908.
77 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.
78 J. George, D. Waroquiers, D. Di Stefano, G. Petretto, G.-M. Rignanese and G. Hautier, *Angew. Chem., Int. Ed.*, 2020, **59**, 7569–7575.
79 H. Glawe, A. Sanna, E. K. U. Gross and M. A. L. Marques, *New J. Phys.*, 2016, **18**, 093011.
80 J. A. Flores-Livas, R. Sarmiento-Pérez, S. Botti, S. Goedecker and M. A. Marques, *J. Phys. Mater.*, 2019, **2**, 025003.
81 G. Demazeau, I. Grannec, A. Marbeuf, J. Portier and P. Hagenmuller, *C. R. Acad. Sci.*, 1969, **269**, 987–988.
82 A. R. Oganov, *Modern Methods of Crystal Structure Prediction*, John Wiley & Sons, 2011.
83 T. F. T. Cerqueira, S. Lin, M. Amsler, S. Goedecker, S. Botti and M. A. L. Marques, *Chem. Mater.*, 2015, **27**, 4562–4573.
84 F. Yan, X. Zhang, Y. G. Yu, L. Yu, A. Nagaraja, T. O. Mason and A. Zunger, *Nat. Commun.*, 2015, **6**, 7308.
85 R. Sarmiento-Perez, T. F. Cerqueira, S. Körbel, S. Botti and M. A. Marques, *Chem. Mater.*, 2015, **27**, 5957–5963.
86 *Crystallographic Databases*, ed. G. G. F. H. Allen and R. Sievers, International Union of Crystallography, Chester, 1987.
87 P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. L. Marques and S. Botti, *J. Chem.*, 2019, **15**, 5069–5079.
88 P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques and S. Botti, *npj Comput. Mater.*, 2020, **6**, 96.
89 I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo and K. W. Jacobsen, *Energy Environ. Sci.*, 2012, **5**, 9034–9043.
90 W. Xiang and W. Tress, *J. Adv. Mater.*, 2019, **31**, 1902851.
91 W. Shockley and H. J. Queisser, *J. Appl. Phys.*, 1961, **32**, 510–519.

Chapter 5 High-Throughput Exploration of Prototype Space

In the following publication "**Predicting stable crystalline compounds using chemical similarity**" [WPhD3], we systematically scan the chemical space around all the experimental known prototypes.

A systematic scan of all compositions for one prototype is manageable, as proven by the last two Chapters, but it is too expensive for all prototypes. Therefore, in this Chapter, we borrow intuition from experiments and perform "hypothetical" feasible substitution synthesis. The feasibility is approximated from the similarity between elements studied in Ref [18] and discussed in Chapter 2.

However, choosing the threshold of similarity to consider a substitution feasible is somehow arbitrary. In Table 2 and Fig. 3 below, we show the number and the percentage of stable compounds as a function of the threshold of the similarity. Note that the table is generated after the publication submission, and the convex hull in the Materials Project database [146] has changed so that the number could be slightly different to those in the paper. Furthermore, we do not have data for a threshold below 5%, as we did not perform those runs. The results for replaceability larger than 20% are statistically less significant. In our substitution workflow the first iteration we start from stable compounds in the materials project database, the next iteration is based on the stable substituted structures in the previous iteration.

We can deduce from the table that the optimum threshold is around 20%, as it maximizes the probability of finding a stable compound. However, if we had chosen 20% as the threshold value, we would have missed a considerable amount of stable compounds. On the other hand, a threshold value of 5% gives at the end of the iterations nearly three times the number of stable structures we would have obtained with a threshold of 20%. Therefore, we choose to set the threshold value to around 5% to obtain better balance between computational costs and success rate.

We start from 713 different prototypes of 9524 compounds from the Materials Project database. The first generation of substitution led to 73,375 candidates, with 59,853 not included in online databases. The most common one is the cubic full-Heusler compound, comprising 10,653 systems. Heusler compounds remain the most common prototype in the second generation with 4238 systems. In the third generation, the

Table 2: The total number of substitution pairs (A replaced by B) and percentage of stable structures at the second and third iteration of substitution at different replaceability thresholds.

threshold	iter 1 → iter 2		iter 2 → iter 3	
(%)	stable	percentage (%)	stable	percentage (%)
5.0	7423	14.4	4608	8.0
7.5	6363	18.5	3997	10.5
10.0	5206	21.0	3339	11.9
12.5	4357	21.7	2909	12.6
15.0	3540	24.0	2431	13.6
20.0	2333	25.0	1628	14.3
30.0	414	14.7	452	10.7
40.0	181	12.3	264	15.2
50.0	90	15.5	141	19.6
60.0	9	5.6	24	15.5
70.0	1	0.8	8	8.0

most common prototype becomes the hexagonal ZrNiAl–Fe_2P structure, with 5009 compounds. The total number of candidates not included in any database is 189,981. Among them, 18,479 are on the convex hull ($E_{\text{hull}} = 0$). Note that the recent amount of stable systems in the Materials Project was around 35,000, at the time of this publication, our results increased the number of theoretically synthesizable materials known by mankind by more than 50%. Moreover, the success rate is 9.72% which is one to two orders of magnitude higher than systematic searches without pre-filtering[164].

In Fig 4, we further show that the distribution of stable structures and the success rate vary across the periodic table. It can be seen that the distribution of new compounds follows, to some extent, the distribution of the initial structures. For example, there are many oxides in both sets, and the number of new stable systems (N_{new}) is approximately proportional to the number of initial systems (N_{ini}) for most $3d$ transition metals. However, there are multiple exceptions, e.g., for Al, Si, K, Ga, As, Rb, Cs, lanthanides, and actinides. In some of these cases, there are many more new compounds than expected. In contrast, for Mo and W, there are much fewer. 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,

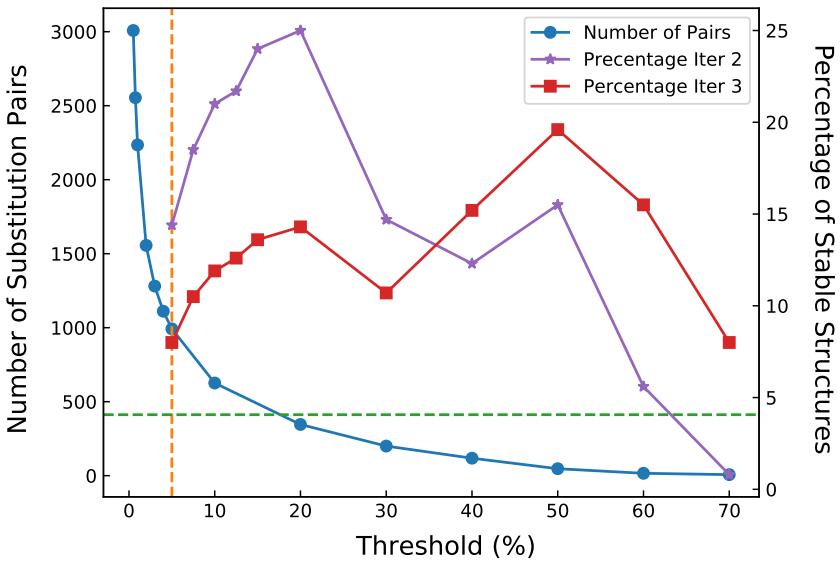


Figure 3: The total number of substitution pairs (A replaced by B) considered and the percentage of stable structures as functions of the threshold of similarity applied. The vertical orange dash shows the 5% threshold used in the present work. The green horizontal dash shows the number of pairs considering only substitution among elements in the same group of the periodic table.

but most elements have a success rate of around 10%. However, there are indeed some exceptions with very high success rates, especially for lanthanides or actinides like Pm or Pa, probably due to that most high-throughput searches overlooked them.

We also perform an analysis of the physical properties of the stable structures. On the PBE level, there are 4840 semiconductors/insulators. Most of them are oxides and fluorides, along with other halides, chalcogenides, and pnictogenides. We also find 4187 magnetic systems with magnetizations extending up to 0.2 $\mu_B/Å^3$. We then filter out 884 structures with non-zero values for both gap and magnetic moments, i.e., potentially ferromagnetic semiconductors. However, further screening on the possible anti- or ferrimagnetic ordering is needed to validate this point. We also evaluate the hardness of our materials, and a few possible super-hard materials are filtered out, which could be interesting for further investigations. Other than these properties, we can look at the hydrostatic deformation potentials, which is the band gap variation with respect to hydrostatic pressure, with a model [WPhD12] trained on the dataset of Ref [WPhD5]. We also want to note that the large dataset in this publication is further used to train machine learning models based on neural networks, e.g., Ref [21] and Ref [120]. Especially the former [21], known as CGAT, are then applied to speed

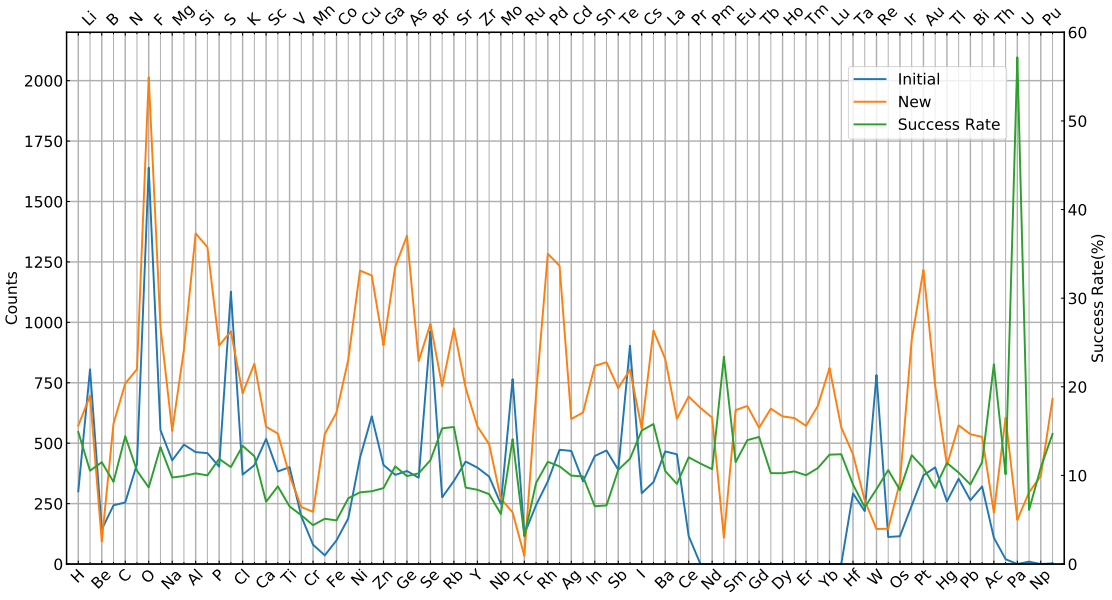


Figure 4: Distribution of the number of new and initial compounds that contain a selected chemical element, as well as the success rate for substitutions involving a selected chemical element.

up the high-throughput screening in the following chapter.

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

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj nature partner journals

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.

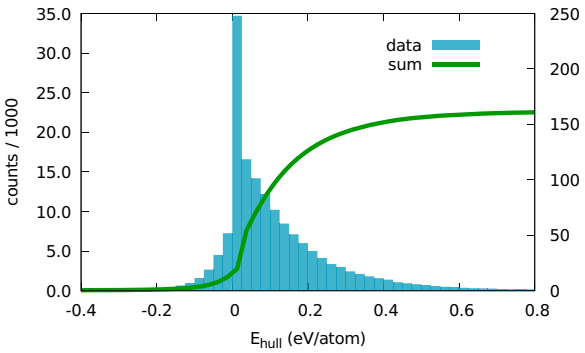


Fig. 1 Thermodynamic stability. Distribution of the distances to the convex hull of all 189,981 compounds.

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.

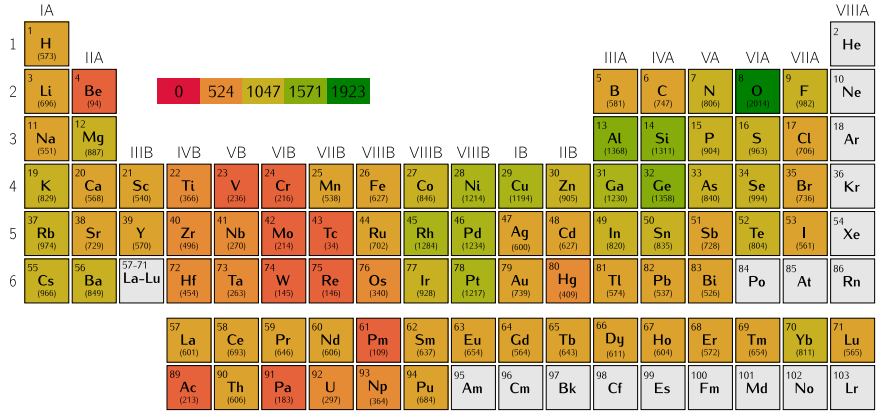


Fig. 2 Distribution of stable materials. The number of stable materials containing a given element through the periodic table.

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.

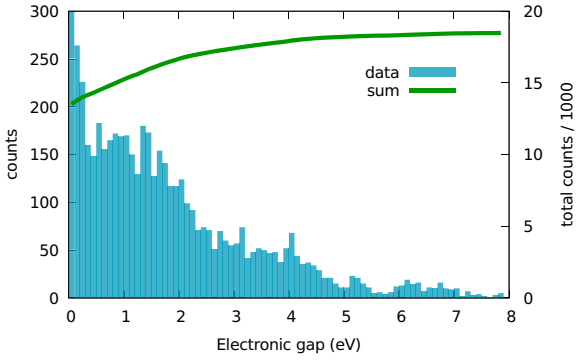


Fig. 3 Band gaps. Histogram of the electronic band-gap for all new stable compounds.

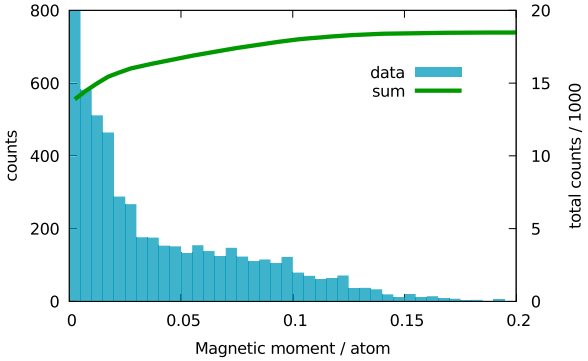


Fig. 5 Total magnetization. Histogram of the total magnetization per unit volume (in $\mu_B \cdot \text{Å}^{-3}$) for all new stable compounds.

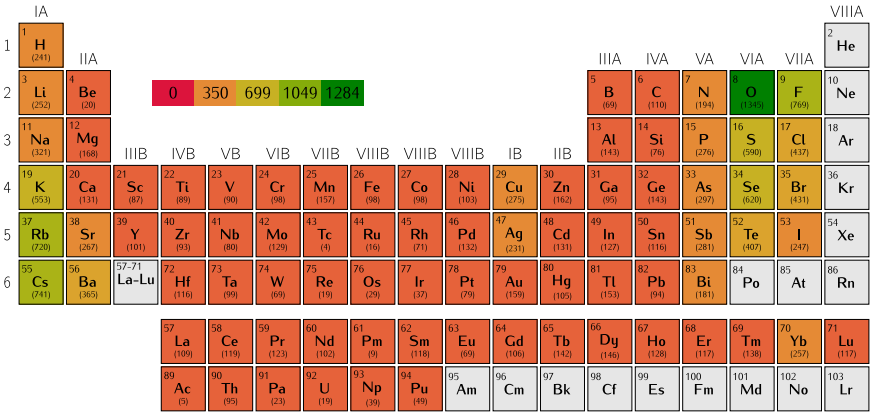


Fig. 4 Distribution of semiconductors and insulators. The number of stable semiconductors and insulators containing the given chemical element of the periodic table.

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).

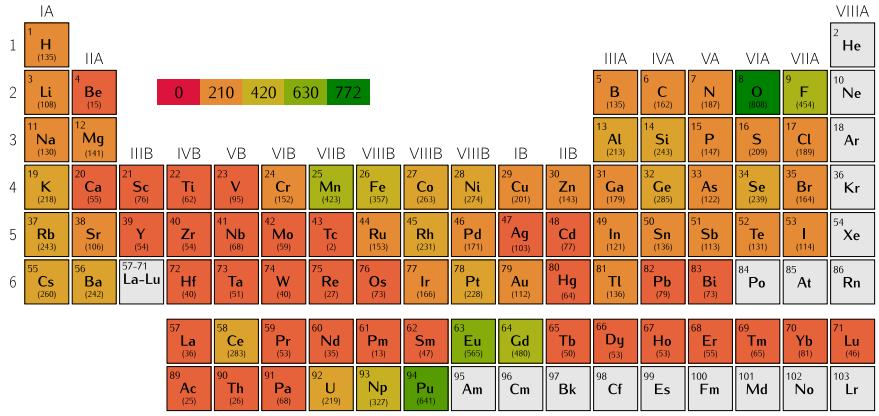


Fig. 6 Distribution of magnetic systems. The number of magnetic systems containing a given element of the periodic table.

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 $22.6\% \times 26.1\% = 5.9\%$. 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 antiferromagnets 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,

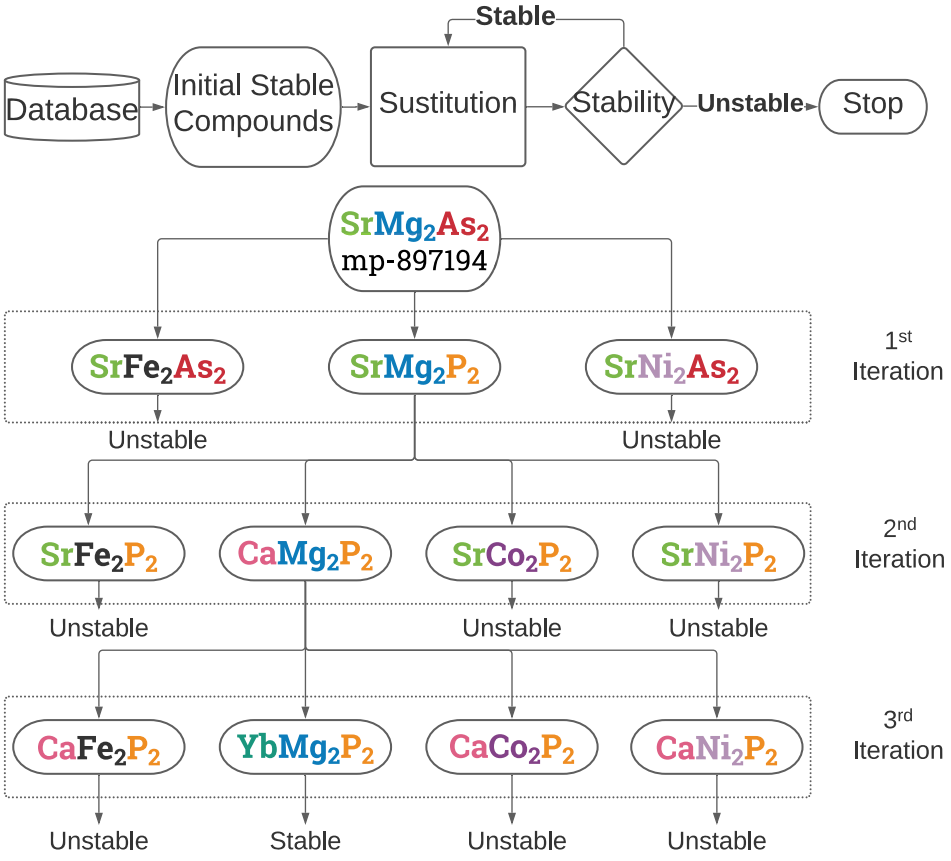


Fig. 7 Work flow. An illustration of a work flow for predicting stable materials based on substitution.

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

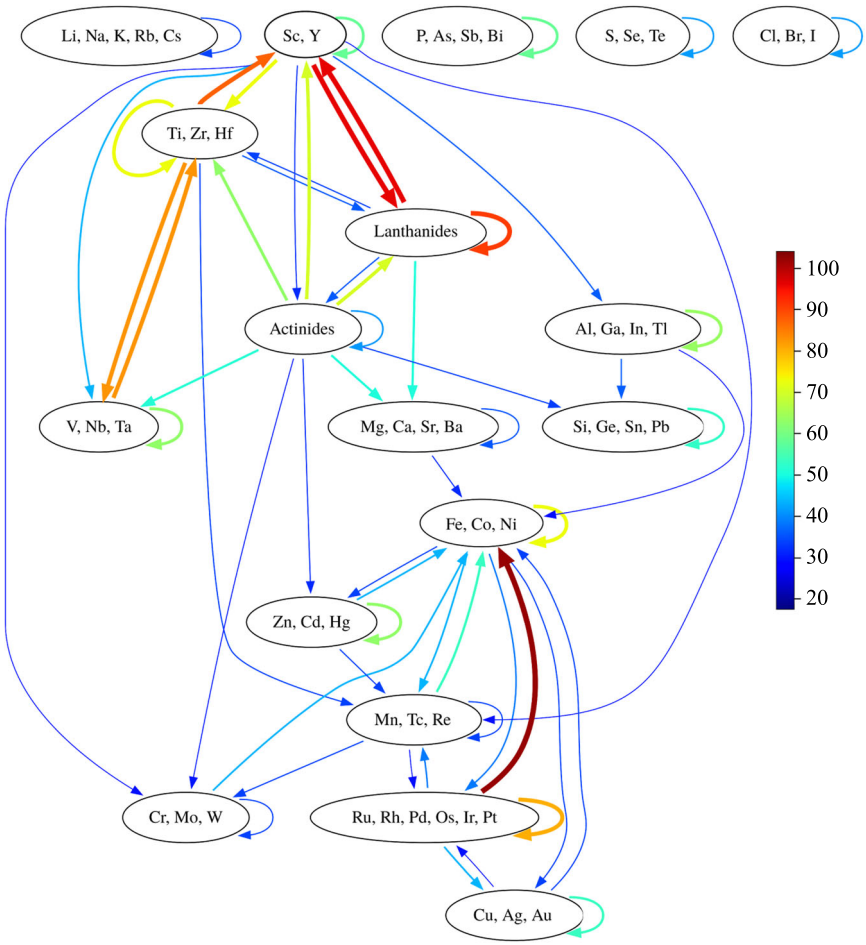


Fig. 8 Substitution schema. Replacements are shown by arrows that start from the elements being replaced. Substitutions between elements within a group are indicated by arrows starting from and pointing to the same box. The thickness of the arrow and the color scale are proportional to the number of 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]

$$S_{AB} := \frac{1}{N_A}\sum_{I,J\neq I}\delta^{IJ}_{AB} \quad (1)$$

where $\delta^{IJ}_{AB} = 1$ 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_2O_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.

Reprints and permission information is available at http://www.nature.com/reprints

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

Chapter 6 Machine Learning Aided Search of New 2D Materials

In the following publication, "**Symmetry-based computational search for novel binary and ternary 2D materials**" [WPhD4], we perform a symmetry-based exhaustive search on the structural and compositional richness of two-dimensional (2D) materials.

2D materials have attracted significant attention since the synthesis of mono-layer graphene [178], and multiple research interests, such as catalysis [179, 180], electronic transport [181, 182], optics [183, 184], and topological properties [185–187] regarding 2D materials has been widely discussed in the literature. However, as we have shown in Chapter 2, the chemical space for 2D materials is still relatively unexplored compared to bulk, three-dimensional (3D) compounds. Experiments have focused only on a few dozen exfoliatable 3D layered materials, which largely limited follow-up prototype based HT search [150–152].

In this Chapter, we adopt an exhaustive systematic strategy which does not rely on experimental prototypes. We start by generating combinations of 74 elements within the periodic table from Li until Bi, excluding radioactive (Tc, Pm, Pr) and rare gases (Ne, Ar, Kr, Xe). A total of $74 \times 73/2 = 2,701$ binary and $74 \times 73 \times 72/6 = 64,824$ ternary combinations is obtained. The list of of these combinations is denoted as X.

Then we consider all the Wycoff positions (WPs) in each two-dimensional space group [188–190]. For each space group, all permutations of the Wycoff positions are generated for afilling 2 or 3 elements. The resulting list is P_j for each space group j. Note P_j is also the list of *general* compositions (e.g., AB_2) for a given j decorated with elements **A** and **B** (as well as **C** if considering ternary). A product of list X and P is then generated. This product list $S_{i,j} = X_i \otimes P_j$ represents a list of stoichiometries for a given combination of elements and space group.

The number of stoichiometries in each list S can be too large at this stage, so we apply two restrictions: 1) The sum of most common oxidation states (Q) of elements ($\sum_{i,j} Q_{i=\text{elements}}^{j=\text{WPs}}$) can reach zero. 2) The electronegativities of cations are lower than that of anions. Note that we rule out intermetallics that often adopt uncommon oxidation states of metals. However, we can go beyond these limitations with the help of machine learning models. The stoichiometries are then input to PyxTal

package [191] to generate appropriate structure with reasonable cell parameters considering the covalent radius of the atoms.

Before performing the DFT calculation, we use a universal neural-network force field M3GNET [22] to pre-optimize the structure. Then, the optimized structures are filtered against three more criteria: 1) The layer thickness is within 7.5Å. 2) The distance to the convex hull (E_{hull}) predicted by the CGAT model [21] is below 600 meV/atom, and 3) the system is not already in the C2DB database [150–152]. The systems that pass all three criteria are then optimized with DFT, and their E_{hull} are evaluated using the hull in Ref [160].

We recover the large majority of systems already present in C2DB, showing the validity of our generating workflow. Moreover, the diversity of structural motifs is beyond that of the experimental ones (mostly square and hexagonal lattices). These interesting new motifs appear naturally in our workflow and are unlikely to be constructed by hand *a priori*. Most importantly, some tilings are unique to the 2D world, with no layered 3D counterpart known, which could lead to further interest in studying the stacking of these layers.

We also push one more step beyond the above workflow by running a systematic prototype search based on all motifs, including both new ones and the ones in C2DB. Again, we pre-filter out binary systems below 200 meV/atom and ternaries below 50 meV/atom through the CGAT model and obtained 1023 candidates. After consecutive optimizations with the M3GNET force field and DFT, we find 638 systems with E_{hull} below 250 meV/atom. The success rate is exceptionally high, at around 62%, which demonstrates how powerful and important the machine-model- based pre-filtering can be in the high-throughput search.

2D Materials

OPEN ACCESS

RECEIVED
21 November 2022

REVISED
22 March 2023

ACCEPTED FOR PUBLICATION
12 April 2023

PUBLISHED
27 April 2023

PAPER

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

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

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

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

Keywords: 2D materials, high-throughput search, machine learning, density functional theory, 2D materials database

Supplementary material for this article is available online

Abstract

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

1. Introduction

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

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

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

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

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

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

While the success of using chemical combinatorics is recognized for 3D materials, it has been a substantial handicap for predicting new 2D materials. The number of known 2D materials prototypes is unfortunately very small. Various research groups have considered different strategies to address this issue, often resorting to machine learning methods. This paper presents an entirely different approach which is not based on motifs or chemical substitutions. Instead, we create all possible combinations of chemical elements for binary (and ternary) systems for specific two-dimensional space groups. Therefore, based on symmetries and chemical criteria, we can arrive at a sizeable two-dimensional crystal structure database and a diverse set of structural prototypes. The crystal shapes show a variety of bondings and forms absent in existing databases. As our search is systematic, crystal structures contain a large number of different chemical formulae as well as almost all possible Wyckoff positions (WPs) allowed by the space groups. Here we focus on two-dimensional materials but our approach is general, and it can also be applied to three-, one-, or even zero-dimensional structures.

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

2. Strategy

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

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

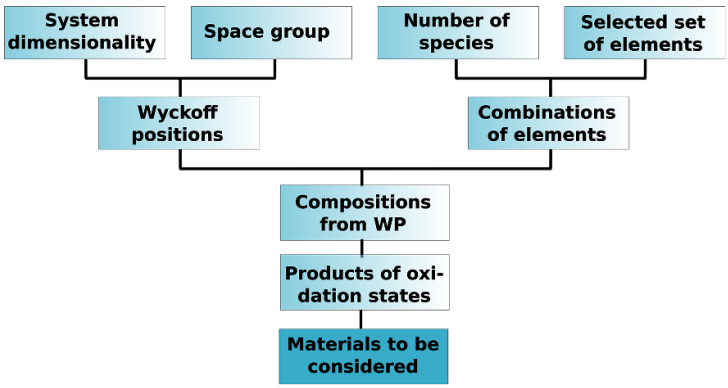


Figure 1. Summarized flowchart for choosing potential materials using group symmetries and their Wyckoff positions.

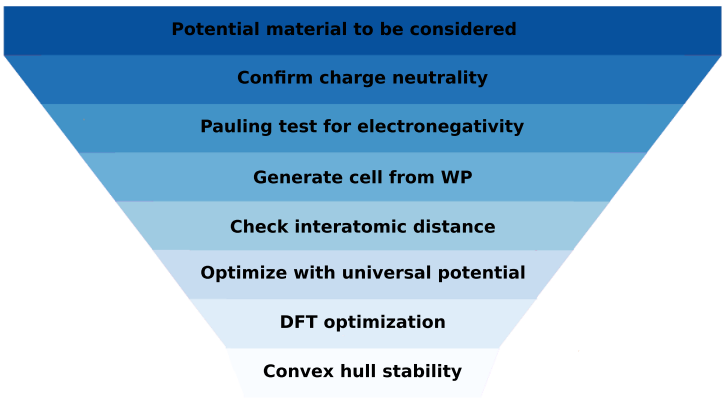


Figure 2. Pipeline for the materials screening, from materials chosen following flowchart 1.

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

The next step is the creation of all possible combinations of the WPs for each chemical element in the combined list. For example, if the number of WPs is four, we get 4! possibilities. This strategy allows different WPs for the same species and therefore broadens the number of possible stoichiometries (i.e. a chemical species can occupy more than one WP). We then create a product of this list associated with the number of selected species. This selection leads to the definition of a chemical composition based on the occupation of the different WPs for each species in the compound. As certain WPs have free parameters, our approach is not exhaustive. For example, for the $p1$ space group we only occupy the (single) position 1 a once for each atomic species. This leads to a single possibility for both the binary compound AB, and the ternary compound ABC. The number of possibilities increases, however, very rapidly with the number of different WPs available within the space group.

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

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

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

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

Space group	Wyckoff positions	$N^{(2)}_{\rm tot}$	$N^{(2)}_{<0.25}$	$N^{(3)}_{\rm tot}$	$N^{(3)}_{<0.25}$
*p*1 (01)	1a (1)	225	5		
*p*11 m (04)	2b (1), 1a (..m)	1321	339		
*p*11a (05)	2a (1)	225	47	1944	307
*p*211 (08)	2c(1), 1b(2..), 1a(2..)	6645	623		
$p2_111$ (09)	2a (1)	225	40	1944	273
*c*211 (10)	4b (1), 2a (2..)	1321	153		
*p*b11 (12)	2a (1)	225	15	1944	448
*c*m11 (13)	4b (1), 2a (m..)	1321	268		
$p2_1$/b11 (17)	4c (1), 2b (−1), 2a (−1)	3129	220		
$p2_12_12$ (21)	4c(1), 2b(..2),2a(..2)	6645	398		
$pb2_1m$ (29)	4b (1), 2a (..m)	1321	140		
*p*b2b (30)	4c (1), 2b (.2.), 2a (.2.)	6645	228		
*p*m2a (31)	4c (1), 2b (m..), 2a (.2.)	6645	478		
$pm2_1n$ (32)	4b (1), 2a (m..)	1321	148		
$pb2_1a$ (33)	4a (1)	225	19	1944	139
*p*b2n (34)	4b (1), 2a (.2.)	1321	74		
*c*m2e (36)	8c (1), 4b (m..), 4a (.2.)	6645	383		
p31m (70)	6d (1), 3c (..m), 2b (3..) 1a (3.m)	15 728	783		

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

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

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

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

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

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

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

3. 2D materials

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

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

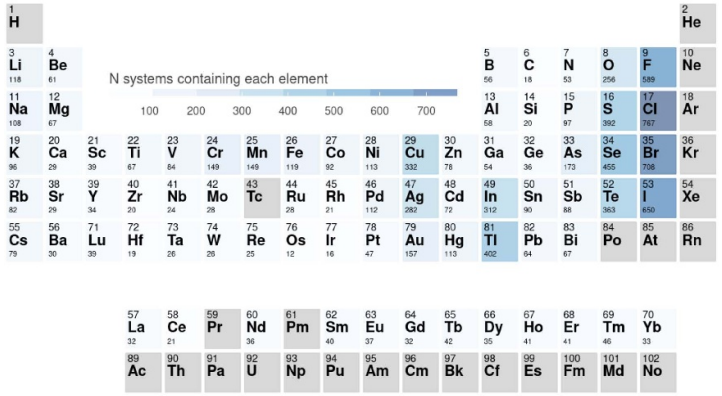
We find 2D compounds across the whole periodic table, including some with lanthanides that have been up to now excluded from previous works. Not surprisingly, the majority of compounds includes a non-metal element (due to the requirement of charge neutrality), leading to the pronounced peaks for O, S, Se, Te, F, Cl, Br, I, etc. The figure also reveals some differences in the prevalence of certain elements between our dataset and C2DB. For example, we find considerably more compounds with F than with O, while in C2DB we observe the opposite behavior. As our approach is to a large extent systematic in what concerns chemical compositions and geometries, we believe that the differences are explained by a bias already present in ICSD and other databases that were used to seed the 2D databases. For example, it is well known that oxides are over-represented in experimental works as they can be more easily synthesized and are often stable in air.

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

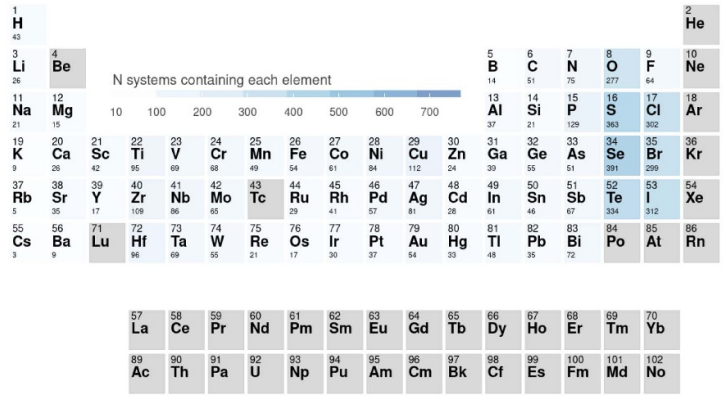
The diversity of stoichiometries is illustrated in figure 4. As our emphasis has been on binary compounds, it is not surprising that most represented stoichiometries are binary. Among these, the simple AB_2, AB_3, A_2B_3, etc dominate the low-energy structures. This fact can be easily understood by the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

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

The first two structures can be derived from a truncated square and a rhombic tiling. In the first case, Cs_2Br_2 squares are connected, forming regular empty octahedra, leading to a rather open lattice. In the second, Se_3O_3 rectangular units form bonds along the corners, leading to flattened octahedra. We then present an example of a Pythagorean tiling, a motif that is composed of two different squares



(a)



(b)

Figure 3. Periodic tables showing the frequency of the chemical elements in binary compounds below 250 meV/atom in (a) our work and (b) C2DB. We emphasize that our entries do not include, by construction, any of the compounds of C2DB.

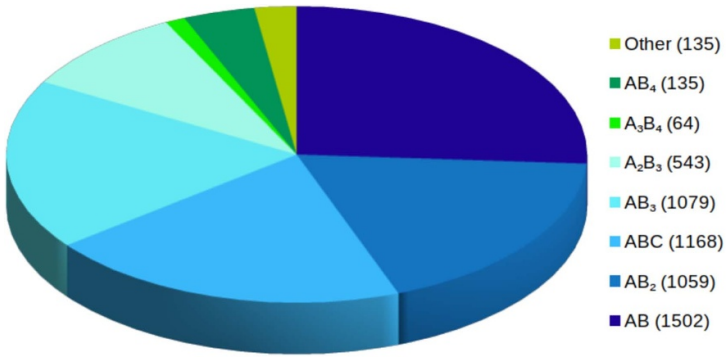


Figure 4. Pie chart of the frequency of different general chemical formula for compounds with distances to the convex hull smaller than 250 meV/atom in our database. The category *other* includes 29 general formulas.

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

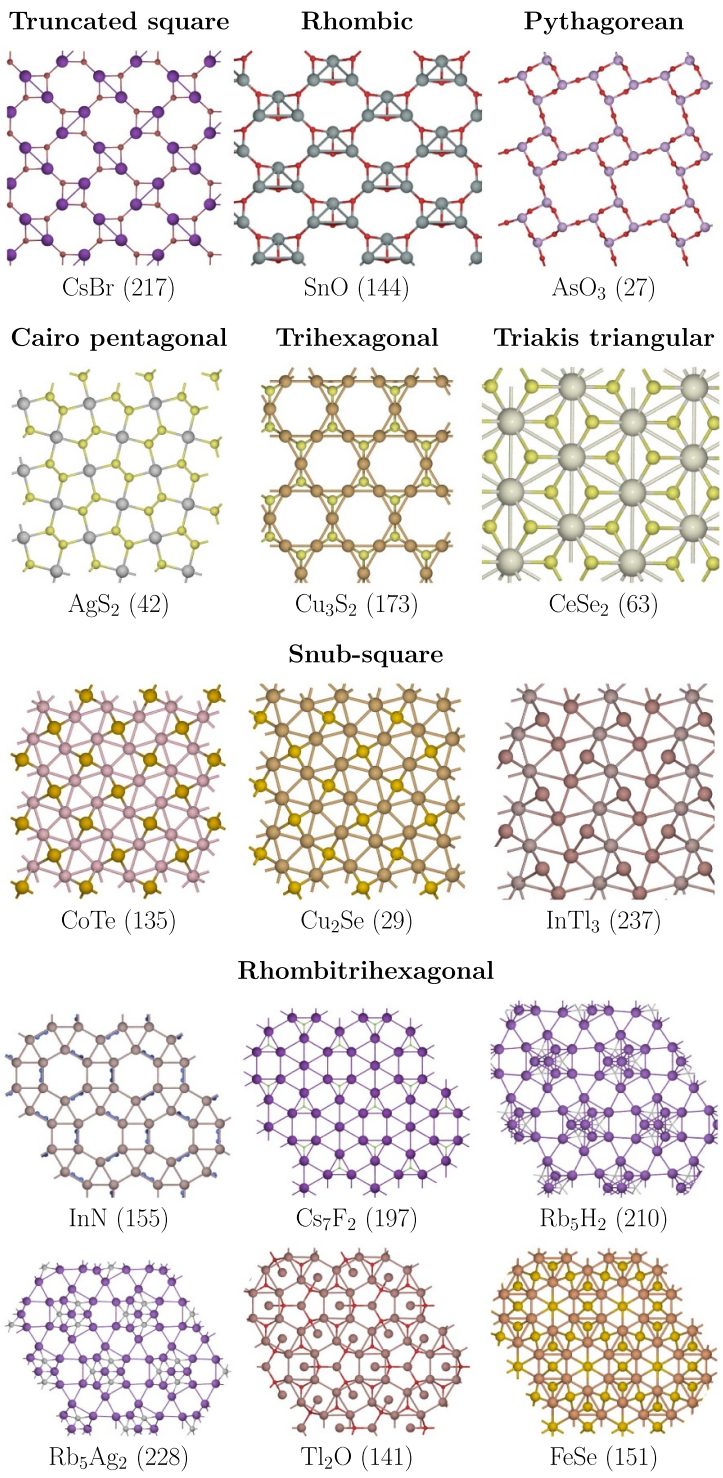


Figure 5. Diversity of structures obtained by our procedure. We label in bold the convex tiling from which the structure can be derived. Next to the chemical formula, in parenthesis, is the distance to the convex hull (in meV/atom) of the specific compound. Note that the structure depicted is not necessarily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

The following three examples are derived from snub-square lattices. In the first two, the metal forms this interesting square-triangle lattice and the non-metal decorates the squares. In the case of CoTe, Te-atoms can be found above and below the plane of the Co atoms, while in Cu_2Se, Se-atoms alternate above and below the plane of the Cu-atoms. We note that this specific lattice was recently proposed for some noble metal chalcogenide monolayers [43] and for certain Ba and Ti oxides [44]. The third example is more complex, as both In and Tl form a distorted version of the snub-square lattice, with further Tl-atoms alternating above and below the plane. Note, however, that this curious structure is almost at the

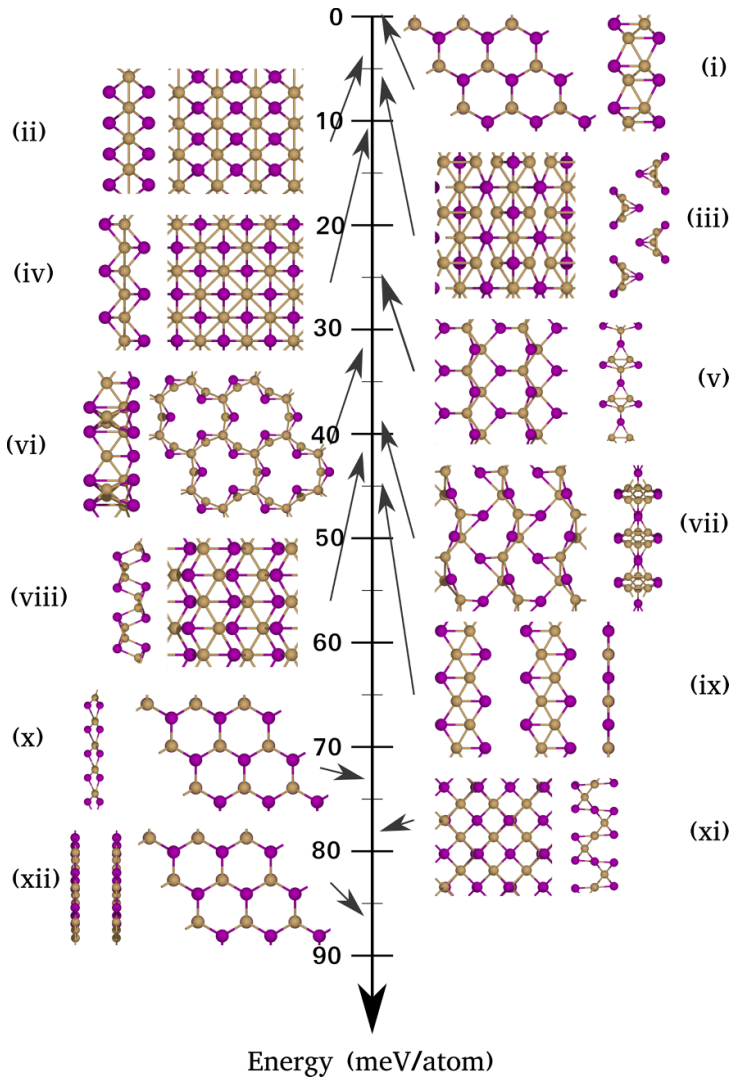


Figure 6. Polymorphism in CuI. The vertical axis denotes increasing energy distance to the convex hull of stability in meV/atom. Copper atoms are depicted in gold, while iodine is in violet. For each structure we show a top and a side view.

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

We have given several examples of the different stoichiometries in our dataset and of the structural variety stemming from them. Now, we look at the issue of polymorphism, i.e. the different phases possible for a specific chemical composition. Not surprisingly, polymorphism depends strongly on the chemical elements present in the compound. For example, for BN we found a single structure below 250 meV/atom, the well-known honeycomb lattice, while for other compounds we have an extraordinary variety in the same energy range.

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

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

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

4. Prototype search

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

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

Note that, in contrast to the systematic generation of structures based on the space groups, in the machine-learning assisted prototype search, we do not impose any constraint on the possible oxidation states. As such, the machine can, and does, propose 2D systems including chemical elements in other, less common oxidation states.

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

5. Conclusions

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

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

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

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

Data availability statement

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

Acknowledgment

We acknowledge the computational resources awarded by XSEDE, a project supported by National Science Foundation Grant Number ACI-1053575. The authors also acknowledge the support from the Texas Advances Computer Center (with the Stampede2 and Bridges supercomputers). We also acknowledge the Super Computing System (Thorny Flat) at WVU, which is funded in part by the National Science Foundation (NSF) Major Research Instrumentation Program (MRI) Award #1726534, and West Virginia University. A H R and L W were funded in part, by the Luxembourg National Research Fund (FNR), Inter Mobility 2DOPMA, Grant Reference 15627293. A H R also recognizes the support of West Virginia Research under the call research challenge grand program. J S and M A L M gratefully acknowledge the Gauss Centre for Supercomputing e.V. (www.gauss-centre.eu) for funding this project by providing computing time on the GCS Supercomputer SUPERMUC-NG at Leibniz Supercomputing Centre (https://www.lrz.de) under the Project pn25co. J S and M A L M gratefully acknowledge the computing time provided to them on the high-performance computers Noctua 2 at the NHR Center PC2. These are funded by the Federal Ministry of Education and Research and the state governments participating on the basis of the resolutions of the GWK for the national high-performance computing at universities (www.nhr-verein.de/unsere-partner).

We also thank Kristian Thygesen for kindly providing us with full access to the C2DB database.

Author contributions

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

Conflict of interest

The authors declare that they have no competing interests.

Appendix. Methods

DFT calculations

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

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

M3GNET

We employed the universal neural-network force-field M3GNET [23] that was developed to reproduce the energies and the forces of bulk structures with remarkable results. As a starting point we used the pretrained network distributed with M3GNET. We tested this model on 1300 of our systems by measuring the difference between the energy calculated with M3GNET (at the M3GNET relaxed structure) and the energy calculated with DFT (at the DFT relaxed structure). We arrived at a MAE of 320 meV/atom and a median absolute error of 223 meV/atom. These numbers are already rather small, especially when we consider that the training set of M3GNET did not include 2D systems that can exhibit very different bonding patterns compared to bulk structures. As soon as enough data was available from our own simulations, we used transfer learning techniques to retrain

M3GNET for 2D materials (see appendix). Specifically, we build a dataset comprising energies, forces, and stresses from structures calculated during the geometry optimization steps. Structures with extremely high forces above 50 eVÅ^{-1} were removed from the data as were structures with no neighboring atoms inside the cutoff radius to avoid errors during training. To balance the training set, for systems with more than 4 recorded geometry optimization steps only the first, last and N_{steps}/3 step were used. The final training set for M3GNET contained 11 612 geometry relaxations corresponding to 34 944 energies and structures. The resulting network had a validation MAE of 61 meV/atom for direct energy predictions after training. The test errors for geometry optimizations on the same dataset as the pretrained model were 198 meV/atom for the MAE and 96 meV/atom for the median absolute error proving the efficiency of our transfer learning strategy. Of course, we expect these errors to decrease further simply by adding more data to the training set. The models were trained with the base hyperparameters from M3GNET and by setting the loss function of the stress in the non-periodic direction to zero.

Crystal-graph attention networks

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

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

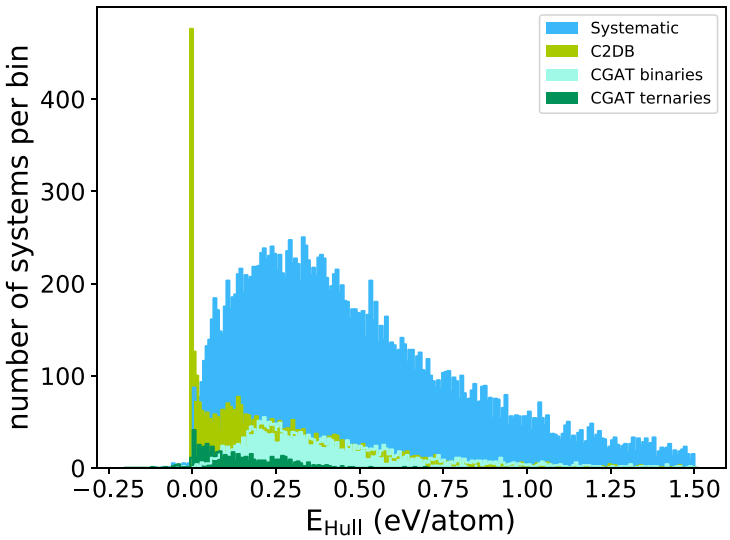


Figure 7. Distribution of the distances to the convex hull (calculated with DFT) for the compounds obtained through our systematic approach and stemming from the machine-learning assisted prototype search compared to the entries of C2DB.

ORCID iDs

Hai-Chen Wang https://orcid.org/0000-0002-2892-5879
Jonathan Schmidt https://orcid.org/0000-0001-5685-6404
Miguel A L Marques https://orcid.org/0000-0003-0170-8222
Ludger Wirtz https://orcid.org/0000-0001-5618-3465
Aldo H Romero https://orcid.org/0000-0001-5968-0571

References

[1] Novoselov K S, Geim A K, Morozov S V, Jiang D, Zhang Y, Dubonos S V, Grigorieva I V and Firsov A A 2004 *Science* **306** 666
[2] Novoselov K S, Jiang D, Schedin F, Booth T J, Khotkevich V V, Morozov S V and Geim A K 2005 *Proc. Natl Acad. Sci. USA* **102** 10451
[3] Mounet N *et al* 2018 *Nat. Nanotechnol.* **13** 246
[4] Sorkun M C, Astruc S, Koelman J and Er S 2020 *npj Comput. Mater.* **6** 1
[5] Zhou J *et al* 2019 *Sci. Data* **6** 1
[6] Haastrup S *et al* 2018 *2D Mater.* **5** 042002
[7] Gjerding M N *et al* 2021 *2D Mater.* **8** 044002
[8] Lyngby P and Thygesen K S 2022 *npj Comput. Mater.* **8** 232
[9] Song Y, Siriwardane E M D, Zhao Y and Hu J 2021 *ACS Appl. Mater. Interfaces* **13** 53303
[10] Jain A *et al* 2013 *APL Mater.* **1** 011002
[11] Gražulis S, Chateigner D, Downs R T, Yokochi A, Quirós M, Lutterotti L, Manakova E, Butkus J, Moeck P and Le Bail A 2009 *J. Appl. Crystallogr.* **42** 726
[12] Allen F H 2002 *Acta Crystallogr.* B **58** 380
[13] Van Hove M A, Hermann K and Watson P 2002 *Acta Crystallogr.* B **58** 338
[14] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 *JOM* **65** 1501
[15] Curtarolo S *et al* 2012 *Comput. Mater. Sci.* **58** 218
[16] Talirz L *et al* 2020 *Sci. Data* **7** 1
[17] Bergerhoff G and Brown I *et al* 1987 *Crystallographic Databases* vol 360, ed F Allen (Chester: International Union of Crystallography) pp 77–95
[18] Zagorac D, Müller H, Ruehl S, Zagorac J and Rehme S 2019 *J. Appl. Crystallogr.* **52** 918
[19] Belsky A, Hellenbrandt M, Karen V L and Luksch P 2002 *Acta Crystallogr.* B **58** 364
[20] Schmidt J, Hoffmann N, Wang H-C, Borlido P, Carriço P J M A, Cerqueira T F T, Botti S and Marques M A L 2022 Large-scale machine-learning-assisted exploration of the whole materials space Materials Cloud (https://doi.org/10.24435/materialscloud:m7-50)
[21] Park C W and Wolverton C 2020 *Phys. Rev. Mater.* **4** 063801
[22] Schmidt J, Hoffmann N, Wang H-C, Borlido P, Carriço P J M A, Cerqueira T F T, Botti S and Marques M A L 2023 Machine-learning-assisted determination of the global zero-temperature phase diagram of materials *Adv. Mater.* (https://doi.org/10.1002/adma.202210788)
[23] Chen C and Ong S P 2022 *Nat. Comput. Sci.* **2** 718
[24] Goodall R E, Parackal A S, Faber F A, Armiento R and Lee A A 2022 Rapid discovery of novel materials by coordinate-free coarse graining *Sci. Adv.* **8** eabn4117
[25] Zhao Y, Al-Fahdi M, Hu M, Siriwardane E M, Song Y, Nasiri A and Hu J 2021 *Adv. Sci.* **8** 2100566
[26] Long T, Fortunato N M, Opahle I, Zhang Y, Samathrakis I, Shen C, Gutfleisch O and Zhang H 2021 *npj Comput. Mater.* **7** 66
[27] Noh J, Kim J, Stein H S, Sanchez-Lengeling B, Gregoire J M, Aspuru-Guzik A and Jung Y 2019 *Matter* **1** 1370
[28] Ren Z *et al* 2022 *Matter* **5** 314
[29] Xie T, Fu X, Ganea O-E, Barzilay R and Jaakkola T 2021 arXiv:2110.06197
[30] Hohenberg P and Kohn W 1964 *Phys. Rev.* **136** B864
[31] Kohn W and Sham L J 1965 *Phys. Rev.* **140** A1133
[32] Kopsky V and Litvin D eds 2002 *Volume E: Subperiodic Groups* International Tables for Crystallography (Dordrecht: Springer)
[33] Aroyo M I, Perez-Mato J M, Orobengoa D, Tasci E, de la Flor G and Kirov A 2011 *Bulg. Chem. Commun.* **43** 183
[34] Aroyo M I, Kirov A, Capillas C, Perez-Mato J and Wondratschek H 2006 *Acta Crystallogr.* A **62** 115
[35] Fredericks S, Parrish K, Sayre D and Zhu Q 2021 *Comput. Phys. Commun.* **261** 107810
[36] Davies D W, Butler K T, Jackson A J, Morris A, Frost J M, Skelton J M and Walsh A 2016 *Chem* **1** 617

Conclusions and Outlooks

In the four publications in this Thesis, we used systematic high-throughput techniques to search for promising functional materials. The first publication focused on finding promising transparent p-type semiconductors which could broaden the choice for experiments in building the p-n junctions in photovoltaic applications. We adopted a systematic search on the double perovskite prototype, only restricting the A-site elements to be Rb or Cs based on previous experimental evidence of stable Rb/Cs-related systems. Our success rate in finding (meta-)stable systems is nearly 10%, which is higher than usually non-restricted exhaustive search, showing how pre-selection could be helpful. Then, we applied criteria on the band gap and hole effective masses that shrank the promising list drastically, resulting in 18 qualified candidates. Ten of them can be considered friendly to ecosystems.

The second publication focused on mix-anion perovskites. We again conducted a systematic search. At first, we only considered the fully ordered five-atoms unit-cell of the formula ABX_2Y. We successfully recovered the experimentally known systems. Moreover, we predicted several novel (meta-)stable oxyfluorides and oxynitrides. For nitrofluoride, we found only one system, $LaMgF_2N$, having a distance to the convex hull less than 200 meV/atom. Further considering the disorder of X and Y anions and lattice distortions, we found that both factors have varying stabilization effects from tens to a few hundreds of meV/atom, depending on the stoichiometry. We also identified some mix-anion structures that are favoured energetically over non-perovskite structures. The electronic structures of the candidates confirm the feasibility of adjusting the gap via anion-alloying.

In the third publication, we tried to explore the labyrinth of chemical space systematically. We used experimental intuition, and the concept of similarity of elements, i.e., considerable similarity between two elements potentially allows the substitution of one element by another while keeping the crystal structure stable. We systematically performed the substitutions by choosing an optimal threshold of similarity. Starting from less than 10,000 stable structures within the Materials Project (MP) database, we generated and calculated nearly 190,000 structures. We found more than 18,000 new stable ones, which increased the information about the convex hull at the time of this publication by more than 50%. Moreover, the success rate of the whole process is nearly 10%, which is an order of magnitude larger than usual systematic

high-throughput searches. We found several super-hard materials and more than 800 potential magnetic semiconductors among these new systems.

The fourth publication focused on two-dimensional systems. For 2D, there are only a few known and widely discussed prototypes. To enrich the knowledge of possible 2D motifs, we applied a systematic workflow based on enumerating all permutations of Wycoff position in the 2D space groups to generate all possible stoichiometries. Similarly, the combinatorial exploding number of candidate structures makes brute force DFT validation impractical. Instead, we applied machine models of interatomic force-field [22] and thermodynamic stability [21] to pre-select potential stable structures. We recovered the 2D structures in the C2DB database and found many new motifs (prototypes) that can not be constructed by hand or from the exfoliation of known 3D structures. We performed a machine-learning supported prototype-based high-throughput search on several compositions using all the prototypes, both those known in C2DB and the ones constructed by us. As a result, we found 638 new meta-stable structures out of 1023 DFT calculations. This exceptional success rate (>60%) shows a dramatic efficiency increase when combining well-performing machine models with the high-throughput scan.

We want to note that searching for new functional materials can be more complicated than running brute force DFT calculations. The number of possible structures combinatorially explodes with the increasing number of elements in a composition. Brute force search is more like finding a path through the labyrinth of chemical space via trial and error. Ideally, we want *a priori* a list of the highly promising candidates, i.e., a map of the labyrinth. Commonly this map is empirically drawn before actual exploration takes place. Therefore, it is inevitably human-biased. Despite this bias, the high-throughput search can be successful, as shown in Chapters 3 and 4. In Chapter 5, we tried a more sophisticated pre-selection, we minimized the human-intervened, and used the data-mining-based similarities of elements, to significantly improved success rate. Our results show that the bias could impede the efficiency of high-throughput search. In Chapter 6, we went beyond search in merely the chemical space and included configurations of Wycoff positions to explore the structural labyrinth. We show that this task is only possible by using machine models to pre-draw the map. Not surprisingly, but impressively, the pre-drawn map is quite accurate, enabling the search to extend beyond the border of the tiny known corner of the labyrinth.

Further work based on Chapter 3 could investigate the p-type dopability of the promising candidates. The workflow can be easily adapted to layered perovskites, hybrid perovskites, or even beyond. The data of effective masses can also be used to train an explanatory model to extract and better understand the underlying physics deciding the electronic structures in perovskites and reverse-design [174] potential compositions. It has been shown [WPhD5] that the gap could be changed drastically through strain caused by the lattice mismatch during the fabrication of the device. Furthermore, strain can also affect the band curvature and, thus, the effective masses. In Chapter 3, the transparent candidates were filtered out based only on their PBE band gap values. We did not consider the excitonic effect that can also affect optical absorption. Unfortunately, the proper consideration of excitons involves too expensive DFT calculations. Therefore, machine learning techniques will again be critical in enabling more accurate filtering criteria.

When performing the study in Chapter 4, we did not use any machine model. Otherwise, we could make the investigation using supercells and consider the disorder effect at the beginning. The disorder was discussed considering the space group's symmetry, and fully disordered structures can be generated and pre-optimized easily with a machine learning force-field, for example, M3GNET [22]. With a better and more complete hull [160], we would have more confidence in predicting whether the perovskite structure is the ground state of the mix-anion composition. It would also be important to predict how the gap and effective mass vary as a function of the ratio of two or more mixed anions. This task, again, requires large supercells, so it will be more efficient using machine learning techniques.

The chemically and structurally diverse dataset obtained in Chapter 5 has been used to train machine learning models to predict the stability of inorganic materials [21, 120]. Such a relatively large amount of data increased the capacity of the training sets and, consequently, helped to achieve good performance of the models. Since the total energy is not the only result of our calculations, the other properties, for example, the band structure and density of state, can also be used to train models to predict (PBE) gaps or fit tight-binding parameters. However, we did not consider anti- or ferrimagnetic configurations in the workflow. Further work in this direction can filter magnetic semiconductors or insulators, fit the Heisenberg model to extract spin exchange parameters, or train explanatory models to help unveil more physics behind them. Combined with the recent achievement of machine-aided high-throughput

results [WPhD12], we can also revisit the similarity scale between elements.

Following the work in Chapter 6, we are calculating structures generated for the rest of the space groups and more ternary compositions. We plan to run DFT calculations on about 40,000 systems pre-selected by the CGAT machine model [21]. Our dataset also broadens the horizon for potential 2D functional materials, including but not limited to topological structures, mechanoelectrics, heterostructures, twistronics, etc. Noticing that many motifs lack a 3D layered counterpart, we can stack the 2D motifs and construct new interesting 3D layered prototypes, which can further push the exploration of the labyrinth of chemical space beyond the vicinity of current experimental prototypes.

References

[1] M. N. Baibich, J. M. Broto, A. Fert, F. N. V. Dau, F. Petroff, P. Etienne, G. Creuzet, A. Friederich, and J. Chazelas, Physical Review Letters **61**, 2472–2475 (1988).

[2] G. Binasch, P. Grünberg, F. Saurenbach, and W. Zinn, Physical Review B **39**, 4828–4830 (1989).

[3] H. Amano, M. Kito, K. Hiramatsu, and I. Akasaki, Japanese Journal of Applied Physics **28**, L2112–L2114 (1989).

[4] S. Nakamura, T. Mukai, and M. Senoh, Japanese Journal of Applied Physics **30**, L1998–L2001 (1991).

[5] M. S. Whittingham, Science **192**, 1126–1127 (1976).

[6] K Mizushima, P Jones, P Wisman, and J Goodenough, Solid State Ionics **3-4**, 171–174 (1981).

[7] A. Yoshino, in *Lithium-ion batteries* (Elsevier, 2014), pages 1–20.

[8] Z. I. Alferov, Reviews of Modern Physics **73**, 767–782 (2001).

[9] C. K. Kao, Reviews of Modern Physics **82**, 2299–2303 (2010).

[10] F. H. Allen, G. Gergerhoff, and R. Sievers, editors (International Union of Crystallography, Chester, 1987).

[11] 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, and D. Morgan, Comput. Mater. Sci. **58**, 218–226 (2012).

[12] R. Sarmiento-Pérez, T. F. T. Cerqueira, S. Körbel, S. Botti, and M. A. L. Marques, Chemistry of Materials **27**, 5957–5963 (2015).

[13] B. C. Yeo, H. Nam, H. Nam, M.-C. Kim, H. W. Lee, S.-C. Kim, S. O. Won, D. Kim, K.-Y. Lee, S. Y. Lee, and S. S. Han, npj Computational Materials **7**, 137 (2021).

[14] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, Nature Materials **12**, 191–201 (2013).

[15] A. Li, R. Bueno-Perez, D. Madden, and D. Fairen-Jimenez, Chemical Science **13**, 7990–8002 (2022).

[16] K. Choudhary, K. F. Garrity, N. J. Ghimire, N. Anand, and F. Tavazza, Physical Review B **103**, 155131 (2021).

[17] A. Vishina, O. Y. Vekilova, T. Björkman, A. Bergman, H. C. Herper, and O. Eriksson, Physical Review B **101**, 094407 (2020).

[18] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, New Journal of Physics **18**, 093011 (2016).

[19] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, npj Computational Materials **5**, 83 (2019).

[20] C. W. Park and C. Wolverton, Phys. Rev. Mater. **4**, 063801 (2020).

[21] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, and M. A. L. Marques, Science Advances **7**, eabi7948 (2021).

[22] C. Chen and S. P. Ong, arXiv, 10.48550/ARXIV.2202.02450 (2022).

[23] L. H. Thomas, Mathematical Proceedings of the Cambridge Philosophical Society **23**, 542–548 (1927).

[24] E. Fermi, Rend. Accad. Naz. Lincei **6**, 32 (1927).

[25] R. M. Dreizler and E. K. U. Gross (Springer Berlin Heidelberg, 1990).

[26] P. Hohenberg and W. Kohn, Phys. Rev. **136**, B864–B871 (1964).

[27] T. L. Gilbert, Physical Review B **12**, 2111–2120 (1975).

[28] J. E. Harriman, Physical Review A **24**, 680–682 (1981).

[29] J. T. Chayes, L. Chayes, and M. B. Ruskai, Journal of Statistical Physics **38**, 497–518 (1985).

[30] C. A. Ullrich and W. Kohn, Physical Review Letters **87**, 093001 (2001).

[31] P. E. Lammert, The Journal of Chemical Physics **125**, 074114 (2006).

[32] M. Levy, Proc. Natl. Acad. Sci. U.S.A **76**, 6062–6065 (1979).

[33] M. Levy, Phys. Rev. A **26**, 1200–1208 (1982).

[34] E. H. Lieb, Int. J. Quantum Chem. **24**, 243–277 (1983).

[35] D. M. Ceperley and B. J. Alder, Physical Review Letters **45**, 566–569 (1980).

[36] J. P. Perdew, M. Ernzerhof, and K. Burke, The Journal of Chemical Physics **105**, 9982–9985 (1996).

[37] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865–3868 (1996).

[38] E. H. Lieb and S. Oxford, International Journal of Quantum Chemistry **19**, 427–439 (1981).

[39] Y. Wang and J. P. Perdew, Physical Review B **43**, 8911–8916 (1991).

[40] M. Levy, International Journal of Quantum Chemistry **36**, 617–619 (2009).

[41] 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, and J. Sun, npj Computational Materials **4**, 10.1038/s41524-018-0065-z (2018).

[42] J. P. Perdew and M. Levy, Physical Review Letters **51**, 1884–1887 (1983).

[43] P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. L. Marques, and S. Botti, Journal of Chemical Theory and Computation **15**, 5069–5079 (2019).

[44] P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques, and S. Botti, npj Computational Materials **6**, 96 (2020).

[45] A. V. Krukau, O. A. Vydrov, A. F. Izmaylov, and G. E. Scuseria, The Journal of Chemical Physics **125**, 224106 (2006).

[46] A. D. Becke and E. R. Johnson, The Journal of Chemical Physics **124**, 221101 (2006).

[47] F. Tran and P. Blaha, Physical Review Letters **102**, 226401 (2009).

[48] A. D. Becke and M. R. Roussel, Physical Review A **39**, 3761–3767 (1989).

[49] H. Jansen and P. Ros, Chemical Physics Letters **3**, 140–143 (1969).

[50] D. Vanderbilt, Physical Review B **41**, 7892–7895 (1990).

[51] P. E. Blöchl, Phys. Rev. B **50**, 17953–17979 (1994).

[52] G. Kresse and D. Joubert, Phys. Rev. B **59**, 1758–1775 (1999).

[53] S. Dugheri, G. Marrubini, N. Mucci, G. Cappelli, A. Bonari, I. Pompilio, L. Trevisani, and G. Arcangeli, Acta Chromatographica **33**, 99–111 (2021).

[54] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, ACS Comb. Sci. **13**, 579–633 (2011).

[55] S. K. Suram, J. A. Haber, J. Jin, and J. M. Gregoire, ACS Comb. Sci. **17**, 224–233 (2015).

[56] M. Shevlin, ACS Medicinal Chemistry Letters **8**, 601–607 (2017).

[57] K. Kim, L. Ward, J. He, A. Krishna, A. Agrawal, and C. Wolverton, Phys. Rev. Mater. **2**, 123801 (2018).

[58] D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton, and A. Walsh, Chem **1**, 617–627 (2016).

[59] A. Zakutayev, J. Perkins, M. Schwarting, R. White, K. Munch, W. Tumas, N. Wunder, and C. Phillips, 2017.

[60] S. Goedecker, The Journal of Chemical Physics **120**, 9911–9917 (2004).

[61] V. Swamy, J. Gale, and L. Dubrovinsky, Journal of Physics and Chemistry of Solids **62**, 887–895 (2001).

[62] T. F. T. Cerqueira, R. Sarmiento-Pérez, M. Amsler, F. Nogueira, S. Botti, and M. A. L. Marques, Journal of Chemical Theory and Computation **11**, 3955–3960 (2015).

[63] Y. Liu, T. Zhao, W. Ju, and S. Shi, Journal of Materiomics **3**, 159–177 (2017).

[64] S. D. Griesemer, L. Ward, and C. Wolverton, Physical Review Materials **5**, 105003 (2021).

[65] V. M. Goldschmidt, Die Naturwissenschaften **14**, 477–485 (1926).

[66] Q. Sun and W.-J. Yin, Journal of the American Chemical Society **139**, 14905–14908 (2017).

[67] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, and M. Scheffler, Science Advances **5**, eaav0693 (2019).

[68] L. Pauling, Journal of the American Chemical Society **51**, 1010–1026 (1929).

[69] R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. L. Sunde, D. Chon, K. R. Poeppelmeier, and A. Zunger, Nature Chemistry **7**, 308–316 (2015).

[70] U. Mizutani, MRS Bulletin **37**, 169–169 (2012).

[71] E. Mooser and W. B. Pearson, Acta Crystallographica **12**, 1015–1022 (1959).

[72] J. C. Phillips and J. A. V. Vechten, Physical Review Letters **22**, 705–708 (1969).

[73] A. Zunger, Physical Review B **22**, 5839–5872 (1980).

[74] P Villars, Journal of the Less Common Metals **92**, 215–238 (1983).

[75] D. Pettifor, Solid State Communications **51**, 31–34 (1984).

[76] D. G. Pettifor, Journal of Physics C: Solid State Physics **19**, 285–313 (1986).

[77] D. G. Pettifor, Journal of the Chemical Society, Faraday Transactions **86**, 1209 (1990).

[78] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, Chemistry of Materials **22**, 3762–3767 (2010).

[79] W. S. McCulloch and W. Pitts, The Bulletin of Mathematical Biophysics **5**, 115–133 (1943).

[80] S. Ray, in 2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon) (Feb. 2019).

[81] 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, and L. M. Ghiringhelli, Electronic Structure **4**, 023004 (2022).

[82] J. Carrasquilla and R. G. Melko, Nature Physics **13**, 431–434 (2017).

[83] D. Kim and D.-H. Kim, Physical Review E **98**, 022138 (2018).

[84] D. Carvalho, N. A. García-Martínez, J. L. Lado, and J. Fernández-Rossier, Physical Review B **97**, 115453 (2018).

[85] M. S. Scheurer and R.-J. Slager, arXiv, 10.48550/ARXIV.2001.01711 (2020).

[86] J. Schmidt, C. L. Benavides-Riveros, and M. A. L. Marques, The Journal of Physical Chemistry Letters **10**, 6425–6431 (2019).

[87] S. Dick and M. Fernandez-Serra, Nature Communications **11**, 3509 (2020).

[88] H. Sidky and J. K. Whitmer, The Journal of Chemical Physics **148**, 104111 (2018).

[89] M. M. Sultan and V. S. Pande, The Journal of Chemical Physics **149**, 094106 (2018).

[90] J. Behler and M. Parrinello, Physical Review Letters **98**, 146401 (2007).

[91] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Physical Review Letters **104**, 136403 (2010).

[92] M. R. G. Marques, J. Wolff, C. Steigemann, and M. A. L. Marques, Physical Chemistry Chemical Physics **21**, 6506–6516 (2019).

[93] C. Chen and S. P. Ong, Nature Computational Science **2**, 718–728 (2022).

[94] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, Chemistry of Materials **31**, 3564–3572 (2019).

[95] L. Breiman, Machine Learning **45**, 5–32 (2001).

[96] J. H. Friedman, The Annals of Statistics **29**, 1189–1232 (2001).

[97] P. Geurts, D. Ernst, and L. Wehenkel, Machine Learning **63**, 3–42 (2006).

[98] D. Jha, L. Ward, A. Paul, W. keng Liao, A. Choudhary, C. Wolverton, and A. Agrawal, Scientific Reports **8**, 17593 (2018).

[99] X. Zheng, P. Zheng, and R.-Z. Zhang, Chem. Sci. **9**, 8426–8432 (2018).

[100] X. Zheng, P. Zheng, L. Zheng, Y. Zhang, and R.-Z. Zhang, Comput. Mater. Sci. **173**, 109436 (2020).

[101] J. Schmidt, L. Chen, S. Botti, and M. A. L. Marques, The Journal of Chemical Physics **148**, 241728 (2018).

[102] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, Physical Review Letters **117**, 135502 (2016).

[103] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Physical Review Letters **108**, 058301 (2012).

[104] J. S. Hub, B. L. de Groot, H. Grubmüller, and G. Groenhof, Journal of Chemical Theory and Computation **10**, 381–390 (2014).

[105] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, International Journal of Quantum Chemistry **115**, 1094–1101 (2015).

[106] L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, Computer Physics Communications **247**, 106949 (2020).

[107] H. Huo and M. Rupp, Machine Learning: Science and Technology **3**, 045017 (2022).

[108] J. Behler and M. Parrinello, Phys. Rev. Lett. **98**, 146401 (2007).

[109] A. P. Bartók, R. Kondor, and G. Csányi, Physical Review B **87**, 184115 (2013).

[110] M. J. Willatt, F. Musil, and M. Ceriotti, The Journal of Chemical Physics **150**, 154110 (2019).

[111] R. Drautz, Physical Review B **99**, 014104 (2019).

[112] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, and R. Ramprasad, npj Computational Materials **3**, 37 (2017).

[113] G. Imbalzano, A. Anelli, D. Giofré, S. Klees, J. Behler, and M. Ceriotti, The Journal of Chemical Physics **148**, 241730 (2018).

[114] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, and P. Marquetand, The Journal of Chemical Physics **148**, 241709 (2018).

[115] H. Gao, J. Wang, and J. Sun, The Journal of Chemical Physics **150**, 244110 (2019).

[116] S. N. Pozdnyakov, M. J. Willatt, A. P. Bartók, C. Ortner, G. Csányi, and M. Ceriotti, Physical Review Letters **125**, 166001 (2020).

[117] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, Nature Communications **8**, 15679 (2017).

[118] T. Xie and J. C. Grossman, Phys. Rev. Lett. **120**, 145301 (2018).

[119] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, The Journal of Chemical Physics **148**, 241722 (2018).

[120] R. E. A. Goodall, A. S. Parackal, F. A. Faber, R. Armiento, and A. A. Lee, Science Advances **8**, eabn4117 (2022).

[121] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436–444 (2015).

[122] S. N. Pozdnyakov and M. Ceriotti, Machine Learning: Science and Technology **3**, 045020 (2022).

[123] Q. Chen, M. Chen, L. Zhu, N. Miao, J. Zhou, G. J. Ackland, and Z. Sun, ACS Applied Materials & Interfaces **12**, 45184–45191 (2020).

[124] S. Li, Y. Liu, D. Chen, Y. Jiang, Z. Nie, and F. Pan, WIREs Computational Molecular Science **12**, e1558 (2021).

[125] X. Wang, S. Ye, W. Hu, E. Sharman, R. Liu, Y. Liu, Y. Luo, and J. Jiang, Journal of the American Chemical Society **142**, 7737–7743 (2020).

[126] M. Geiger and T. Smidt, arXiv, 10.48550/ARXIV.2207.09453 (2022).

[127] C. W. Park, M. Kornbluth, J. Vandermause, C. Wolverton, B. Kozinsky, and J. P. Mailoa, npj Computational Materials **7**, 73 (2021).

[128] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley, arXiv, 10.48550/ARXIV.1802.08219 (2018).

[129] M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. Cohen, arXiv, 10.48550/ARXIV.1807.02547 (2018).

[130] R. Kondor, Z. Lin, and S. Trivedi, arXiv, 10.48550/ARXIV.1806.09231 (2018).

[131] R. Kondor, arXiv, 10.48550/ARXIV.1803.01588 (2018).

[132] S. Lany, Physical Review B **78**, 10.1103/physrevb.78.245207 (2008).

[133] J. Sun, A. Ruzsinszky, and J. P. Perdew, Phys. Rev. Lett. **115**, 036402 (2015).

[134] M. K. Horton, J. H. Montoya, M. Liu, and K. A. Persson, npj Computational Materials **5**, 10.1038/s41524-019-0199-7 (2019).

[135] G. Hautier, S. P. Ong, A. Jain, C. J. Moore, and G. Ceder, Physical Review B **85**, 10.1103/physrevb.85.155208 (2012).

[136] R. Friedrich, D. Usanmaz, C. Oses, A. Supka, M. Fornari, M. B. Nardelli, C. Toher, and S. Curtarolo, npj Computational Materials **5**, 10.1038/s41524-019-0192-1 (2019).

[137] A. Wang, R. Kingsbury, M. McDermott, M. Horton, A. Jain, S. P. Ong, S. Dwaraknath, and K. A. Persson, Scientific Reports **11**, 10.1038/s41598-021-94550-5 (2021).

[138] R. Urrego-Ortiz, S. Builes, and F. Calle-Vallejo, ChemCatChem **13**, 2508–2516 (2021).

[139] R. Sarmiento-Pérez, S. Botti, and M. A. L. Marques, Journal of Chemical Theory and Computation **11**, 3844–3850 (2015).

[140] M. Kaltak, and G. Kresse, Phys. Rev. B **90**, 054115 (2014).

[141] T. S. Jauho, T. Olsen, T. Bligaard, and K. S. Thygesen, Physical Review B **92**, 115140 (2015).

[142] L. Wang, T. Maxisch, and G. Ceder, Physical Review B **73**, 054115 (2006).

[143] S. Grindy, B. Meredig, S. Kirklin, J. E. Saal, and C. Wolverton, Physical Review B **87**, 054115 (2013).

[144] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, Phys. Rev. B **84**, 045115 (2011).

[145] , S. Lany, X. Zhang, and A. Zunger, Phys. Rev. B **85**, 115104 (2012).

[146] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, APL Mater. **1**, 011002 (2013).

[147] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, npj Computational Materials **6**, 115104 (2020).

[148] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, JOM **65**, 1501–1509 (2013).

[149] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, npj Computational Materials **1**, 115104 (2015).

[150] S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen, and K. S. Thygesen, 2D Mater. **5**, 042002 (2018).

[151] 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, and K. S. Thygesen, 2D Mater. **8**, 044002 (2021).

[152] P. Lyngby and K. S. Thygesen, Npj Comput. Mater. **8**, 115104 (2022).

[153] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 –50 (1996).

[154] G. Kresse and J. Furthmüller, Phys. Rev. B **54**, 11169–11186 (1996).

[155] P Giannozzi, O Andreussi, T Brumme, O Bunau, M. B. Nardelli, M Calandra, R Car, C Cavazzoni, D Ceresoli, M Cococcioni, N Colonna, I Carnimeo, A. D. Corso, S de Gironcoli, P Delugas, R. A. D. Jr, A Ferretti, A Floris, G Fratesi, G Fugallo, R Gebauer, U Gerstmann, F Giustino, T Gorni, J Jia, M Kawamura, H.-Y. Ko, A Kokalj, E Küçükbenli, M Lazzeri, M Marsili, N Marzari, F Mauri, N. L. Nguyen, H.-V. Nguyen, A. O. de-la Roza, L Paulatto, S Poncé, D Rocca, R Sabatini, B Santra, M Schlipf, A. P. Seitsonen, A Smogunov, I Timrov, T Thonhauser, P Umari, N Vast, X Wu, and S Baroni, J. Condens. Matter Phys. **29**, 465901 (2017).

[156] J. J. Mortensen, L. B. Hansen, and K. W. Jacobsen, Physical Review B **71**, 115104 (2005).

[157] J Enkovaara, C Rostgaard, J. J. Mortensen, J Chen, M Dułak, L Ferrighi, J Gavnholt, C Glinsvad, V Haikola, H. A. Hansen, H. H. Kristoffersen, M Kuisma, A. H. Larsen, L Lehtovaara, M Ljungberg, O Lopez-Acevedo, P. G. Moses, J Ojanen, T Olsen, V Petzold, N. A. Romero, J Stausholm-Møller, M Strange, G. A. Tritsaris, M Vanin, M Walter, B Hammer, H Häkkinen, G. K. H. Madsen, R. M. Nieminen, J. K. Nørskov, M Puska, T. T. Rantala, J Schiøtz, K. S. Thygesen, and K. W. Jacobsen, Journal of Physics: Condensed Matter **22**, 253202 (2010).

[158] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, C. S. Adorf, C. W. Andersen, O. Schütt, C. A. Pignedoli, D. Passerone, J. VandeVondele, T. C. Schulthess, B. Smit, G. Pizzi, and N. Marzari, Scientific Data **7**, 115104 (2020).

[159] J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang, and Y. P. Feng, Sci. Data **6**, 1–10 (2019).

[160] Schmidt, Jonathan, Hoffmann, Noah, Wang, Hai-Chen, Borlido, Pedro, M.A. Carriço, Pedro J., F. T. Cerqueira, Tiago, Botti, Silvana, and L. Marques, Miguel A., Materials Cloud **85**, 115104 (2022).

[161] Schmidt, Jonathan, Wang, Hai-Chen, F. T. Cerqueira, Tiago, Botti, Silvana, and L. Marques, Miguel A., Materials Cloud **85**, 115104 (2021).

[162] Wang, Hai-Chen, Botti, Silvana, and L. Marques, Miguel A., Materials Cloud **85**, 115104 (2021).

[163] Wang, Hai-Chen, Schmidt, Jonathan, L. Marques, Miguel A., Wirtz, Ludger, and Romero, Aldo H., Materials Cloud **85**, 115104 (2022).

[164] J. Schmidt, J. Shi, P. Borlido, L. Chen, S. Botti, and M. A. L. Marques, Chemistry of Materials **29**, 5090–5103 (2017).

[165] H. Kawazoe and K. Ueda, Journal of the American Ceramic Society **82**, 3330–3336 (2004).

[166] J. Wang, J. Li, and S.-S. Li, Journal of Applied Physics **110**, 054907 (2011).

[167] S. Körbel, M. A. L. Marques, and S. Botti, Journal of Materials Chemistry C **4**, 3157–3167 (2016).

[168] P. C. Harikesh, H. K. Mulmudi, B. Ghosh, T. W. Goh, Y. T. Teng, K. Thirumal, M. Lockrey, K. Weber, T. M. Koh, S. Li, S. Mhaisalkar, and N. Mathews, Chemistry of Materials **28**, 7496–7504 (2016).

[169] T. T. Deng, E. H. Song, Y. Y. Zhou, L. Y. Wang, and Q. Y. Zhang, Journal of Materials Chemistry C **5**, 12422–12429 (2017).

[170] E. Meyer, D. Mutukwa, N. Zingwe, and R. Taziwa, Metals **8**, 667 (2018).

[171] G. Hautier, A. Miglio, G. Ceder, G.-M. Rignanese, and X. Gonze, Nat. Commun. **4**, 2292 (2013).

[172] A. Zunger, Applied Physics Letters **83**, 57–59 (2003).

[173] J. Robertson and S. J. Clark, Physical Review B **83**, 075205 (2011).

[174] P. Zhang, S. Yu, X. Zhang, and S.-H. Wei, Physical Review Materials **3**, 055201 (2019).

[175] A. K. Jena, A. Kulkarni, and T. Miyasaka, Chemical Reviews **119**, 3036–3103 (2019).

[176] H. Kageyama, K. Hayashi, K. Maeda, J. P. Attfield, Z. Hiroi, J. M. Rondinelli, and K. R. Poeppelmeier, Nature Communications **9**, 115104 (2018).

[177] A. van de Walle, Calphad **33**, 266–278 (2009).

[178] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov, Science **306**, 666–669 (2004).

[179] X. Chia and M. Pumera, Nature Catalysis **1**, 909–921 (2018).

[180] X. Zhang, A. Chen, L. Chen, and Z. Zhou, Advanced Energy Materials **12**, 2003841 (2021).

[181] S. D. Sarma, S. Adam, E. H. Hwang, and E. Rossi, Reviews of Modern Physics **83**, 407–470 (2011).

[182] H. Schmidt, F. Giustiniano, and G. Eda, Chemical Society Reviews **44**, 7715–7736 (2015).

[183] G. G. Naumis, S. Barraza-Lopez, M. Oliva-Leyva, and H. Terrones, Reports on Progress in Physics **80**, 096501 (2017).

[184] Q. Ma, G. Ren, K. Xu, and J. Z. Ou, Advanced Optical Materials **9**, 2001313 (2020).

[185] L. A. Walsh and C. L. Hinkle, Applied Materials Today **9**, 504–515 (2017).

[186] X. Liu and M. C. Hersam, Nature Reviews Materials **4**, 669–684 (2019).

[187] D. Culcer, A. C. Keser, Y. Li, and G. Tkachov, 2D Materials **7**, 022007 (2020).

[188] , S. Lany, X. Zhang, and A. Zunger, edited by V. Kopský and D. B. Litvin, Vol. 85 (International Union of Crystallography, Jan. 2010), page 115104.

[189] M. I. Aroyo, J. M. Perez-Mato, D. Orobengoa, E. Tasci, G. de la Flor, and A. Kirov, Bulg. Chem. Commun **43**, edited by V. Kopský and D. B. Litvin, 183–197 (2011).

[190] M. I. Aroyo, A. Kirov, C. Capillas, J. Perez-Mato, and H. Wondratschek, Acta Crystallogr., Sect. A: Found. Crystallogr. **62**, edited by V. Kopský and D. B. Litvin, 115–128 (2006).

[191] S. Fredericks, K. Parrish, D. Sayre, and Q. Zhu, Comput. Phys. Commun. **261**, edited by V. Kopský and D. B. Litvin, 107810 (2021).

List of Publications During PhD

[WPhD1] **Hai-Chen Wang**, P. Pistor, M. A. L. Marques, and S. Botti, Journal of Materials Chemistry A **7**, 14705–14711 (2019).

[WPhD2] **Hai-Chen Wang**, J. Schmidt, S. Botti, and M. A. L. Marques, Journal of Materials Chemistry A **9**, 8501–8513 (2021).

[WPhD3] **Hai-Chen Wang**, S. Botti, and M. A. L. Marques, npj Computational Materials **7**, 12 (2021).

[WPhD4] **Hai-Chen Wang**, J. Schmidt, M. A. L. Marques, L. Wirtz, and A. H. Romero, 2D Materials **10**, 035007 (2023).

[WPhD5] P. Borlido, J. Schmidt, **Hai-Chen Wang**, S. Botti, and M. A. L. Marques, npj Computational Materials **8**, 10.1038/s41524-022-00811-w (2022).

[WPhD6] P. Pistor, M. Meyns, M. Guc, **Hai-Chen Wang**, M. A. Marques, X. Alcobé, A. Cabot, and V. Izquierdo-Roca, Scripta Materialia **184**, 24–29 (2020).

[WPhD7] J. Jacobs, M. A. L. Marques, **Hai-Chen Wang**, E. Dieterich, and S. G. Ebbinghaus, Inorganic Chemistry **60**, 13646–13657 (2021).

[WPhD8] A. W. Huran, **Hai-Chen Wang**, and M. A. L. Marques, 2D Materials **8**, 045002 (2021).

[WPhD9] A. W. Huran, **Hai-Chen Wang**, A. San-Miguel, and M. A. L. Marques, The Journal of Physical Chemistry Letters **12**, 4972–4979 (2021).

[WPhD10] J. Schmidt, **Hai-Chen Wang**, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, Scientific Data **9**, 10.1038/s41597-022-01177-w (2022).

[WPhD11] J. Schmidt, **Hai-Chen Wang**, G. Schmidt, and M. A. L. Marques, npj Computational Materials **9**, 10.1038/s41524-023-01009-4 (2023).

[WPhD12] J. Schmidt, N. Hoffmann, **Hai-Chen Wang**, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti, and M. A. L. Marques, Advanced Materials **35**, 10.1002/adma.202210788 (2023).

Copyright

Acknowledgments

It would only be possible to finally put together this thesis with the support from all the people around me over the past years.
First, I want to express my genuine gratitude to my supervisor, Prof. Miguel Marques, for providing this opportunity. I would still be stuck in China if it were not for him. I had fun working with him and learned a lot, not just physics but also teaching skills, and I am always inspired by his dedication to science and education.
I want to thank all the group members, Mario, Jonathan, Ahmad, Thomas, Carlos, and Conrad. They provided significant help in both work and life. Their kindness and patience are crucial, and I appreciate the welcoming and open-minded group environment. All these meant a lot to me. Also, I want to thank the students, Martin and Mattheo, whom I had the pleasure to work with and learned a lot from.
I am also thankful to the neighboring groups' members, Viktoriia, Dominik, Benjamin, Dimitrios, Marius, Mikheil, Micheal, and Sajad, for holding many group activities and fun topics for discussions.
I also want to thank Prof. Georg Schmidt, Prof. Kathrin Dörr, Prof. Jörg Schilling, and Katja for their deciding help in many bureaucratic procedures.
Finally, I would like to thank my parents, and my friends all around the world ;), especially Vika and Mario, for being essential sources of motivation, strength, fantastic recipes, and laughter(HoHoHoHo).

Eidesstattliche Erklärung / Statutory statement

Hiermit erkläre ich, Hai-Chen Wang, gemäß §5 der Promotionsordnung der Naturwissenschaftlichen Fakultäten I, II und III der Martin-Luther-Universität Halle-Wittenberg vom 13.07.2016, dass ich die vorliegende Arbeit "High-throughput Discovery and Characterization of Inorganic Materials Using ab-intio Methods" selbstständig und ohne fremde Hilfe verfasst, andere als die von mir angegebenen Quellen und Hilfsmittel nicht benutzt und die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Ich erkläre die Angaben wahrheitsgemäß gemacht, keine vergeblichen Promotionsversuche unternommen und keine Dissertation an einer anderen wissenschaftlichen Einrichtung zur Erlangung eines akademischen Grades eingereicht zu haben. Ich bin weder vorbestraft noch sind gegen mich Ermittlungsverfahren anhängig.

Halle (Saale), den 19.01.2023

Curriculum Vitae

Personal data

Name:	Hai-Chen Wang
Date of birth:	25.04.1988
Place of birth:	Kunming, China

Education

Aug. 2018 – Jul. 2023	Wissenschaftlicher Mitarbeiter Martin-Luther University Halle-Wittenberg, Germany
Sep. 2012 – Jun. 2017	PhD in Material Chemistry Thesis "Theoretical Investigation on Hydrogen Storage Properties of Cu-doped MgH_2, Ti/VLi Decorated $LiBH_4$, and Pure $LiCa(AlH_4)_3$." School of Chemistry and Chemical Engineering, Guangxi University, China
Sep. 2006 – Jun. 2011	B.Sc. in Chemistry Department of Chemistry and Molecular Science, Wuhan University, China
Sep. 2003 – Jun. 2006	High School No.8 Middle School of Kunming City, China