Reflections on one million compounds in the open quantum materials database (OQMD)

Jiahong Shen¹,², Sean D Griesemer¹,², Abhijith Gopakumar¹, Bianca Baldassarri¹, James E Saal¹, Muratahan Aykol³, Vinay I Hegde² and Chris Wolverton¹,*

¹ Materials Science and Engineering, Northwestern University, Evanston, IL 60208, United States of America
² Citrine Informatics, Redwood City, CA 94063, United States of America
³ Toyota Research Institute, Los Altos, CA 94022, United States of America
⁴ J S, S G, A G and B B contributed equally.
* Author to whom any correspondence should be addressed.

E-mail: c-wolverton@northwestern.edu

Keywords: high-throughput DFT, machine learning, materials discovery, materials database

Abstract

Density functional theory (DFT) has been widely applied in modern materials discovery and many materials databases, including the open quantum materials database (OQMD), contain large collections of calculated DFT properties of experimentally known crystal structures and hypothetical predicted compounds. Since the beginning of the OQMD in late 2010, over one million compounds have now been calculated and stored in the database, which is constantly used by worldwide researchers in advancing materials studies. The growth of the OQMD depends on project-based high-throughput DFT calculations, including structure-based projects, property-based projects, and most recently, machine-learning-based projects. Another major goal of the OQMD is to ensure the openness of its materials data to the public and the OQMD developers are constantly working with other materials databases to reach a universal querying protocol in support of the FAIR data principles.

1. History and motivation

For much of the history of materials science, the generation of phase-based materials thermodynamic data has been slow and tedious, relying on delicate calorimetric measurements of carefully prepared single-phase materials samples. Decades of early effort resulted in compilations of thermodynamic quantities (e.g. JANAF [1], Kubaschewski [2], etc) containing entries for ~1000–2000 phases (including solids, liquids, and gases). Similarly, the exploration of phase equilibria in complex multi-component systems required such time, cost, and precision, resulting in similar compilations of phase diagrams (ASM [3], ACerS [4], and Pauling File [5]). Under these early conditions, the task of materials discovery (i.e. finding a novel single-phase structure to satisfy some performance metric) was difficult, relying extensively on expert intuition or gleaned correlations in the known materials space (e.g. Pettifor/Zunger maps [6, 7] or the Miedema model [8]).

Density functional theory (DFT) was a critical enabling technology for accelerating modern materials discovery. By generating ab initio predictions of total energy for arbitrary arrangements of atoms, a self-consistent dataset of thermodynamic quantities and phase stability could be produced for entirely unknown systems. From early calculations of individual molecules and simple crystalline solids, improving DFT software performance and computing performance eventually enabled many complex structures to be readily simulated in a given study.

The prospect of calculating the DFT properties of every known crystal structure became possible, and several teams took on this task, including the Materials Project [9], AFLOW [10], and our open quantum materials database (OQMD) [11]. These databases each contain predicted thermodynamic properties of tens of thousands of experimentally observed structures from the inorganic crystal structure database (ICSD) [12], dwarfing the earlier experimental compilations and becoming foundations for materials discovery efforts.
Efforts to develop the OQMD began in late 2010 amongst members of the Wolverton Research Group at Northwestern University. Builders of the University’s new Quest High Performance Computing Cluster sought volunteers to test their hardware, so significant computing resources were available. After many false starts from continuous tweaking to DFT parameter selection, DFT error checking, and automation algorithm optimization, the first data was published in late 2013 at oqmd.org. Early work using the results of the data include cathode coatings for Lithium-Ion batteries [13], electrodes for hybrid Li-ion–Li–O$_2$ cells [14], Li-ion battery anodes [15], strengthening precipitates in alloys [16], and perovskites for thermochemical water splitting applications [17].

We recognized early in the process of developing the OQMD that predictions of phase stability, highly important for materials discovery, require not just calculations of known compounds, but also unknown competing phases. Therefore, we supplemented the 40 000 or so known ICSD compounds with hundreds of thousands of hypothetical compounds based on known prototype lattices (e.g. L1$_2$ and perovskite). In this way, unexplored materials systems with potentially stable compounds and particular novel stable compounds could be identified. Every new prediction, calculated with a predefined set of DFT parameters, is consistent with all the others, resulting in an ever-growing pool of compounds that are continually being used to construct the largest phase diagram ever built. A summary of the number of all stable compounds in each prototype is summarized in figure 1.

2. Continuously expanding the OQMD

2.1. High-throughput DFT calculations

One of the major advantages of the OQMD is its flexibility to allow researchers to explore hypothetical compounds (not previously known), which enhances our scientific understanding in providing candidate compounds for future applications. We are continually performing this exploration via high-throughput DFT calculations on a wide variety of hypothetical compounds. We often use ‘structure prototypes’ as a blueprint for generating hypothetical compounds. We define the prototype of a structure as the combination of its stoichiometry, space group, and Wyckoff site occupancies. This blueprint is useful for two reasons. Reason #1 is that most known compounds share common prototypes, like NaCl-, CsCl-, and Heusler-type,
among many others. In fact, 83% of all compounds in the ICSD share a prototype with another compound, and 27% of all compounds share a prototype with \( \geq 50 \) other compounds [18]. Reason #2 is that we can use a prototype to arrive at a stable or metastable crystal structure for hypothetical compounds. This is done by taking a known compound of that prototype, substituting in elements of the hypothetical composition, and using DFT to relax unit cell parameters and atomic coordinates along the symmetry directions.

It is natural to start with the most common prototypes as the blueprint for generating hypothetical compounds. In the OQMD, we used several common prototypes to conduct ‘exhaustive’ high-throughput DFT of hypothetical compounds, i.e. calculate nearly all possible compounds by substitution of elements from the periodic table. The prototypes completed include binaries \( B_1 ^\text{(NaCl)}, B_2 ^\text{(CsCl)}, B_3 ^\text{(zincblende)}, B_\text{h} ^\text{(WC)}, C_{15} ^\text{(MgZn_2)}, D_0 ^\text{3} ^\text{(BiF_3)}, D_{02} ^\text{2} ^\text{(Al}_3 ^\text{Ti)}, L_1 ^0 ^\text{(CuAu)}, L_1 ^1 ^\text{(CuPt)}, L_2 ^2 ^\text{(Cu}_3 ^\text{Au)}; \) and ternaries \( C_{1b} ^\text{(Half Heusler)}, \) and \( L_2 ^1 ^\text{(Full Heusler)}. \) This has amounted to 393 879 DFT calculations (54 030 binary and 339 849 ternary), and has been quite fruitful: 1973 of these hypothetical compounds (396 binary and 1577 ternary) are on the convex hull of stability and do not have an ICSD polymorph that is close in energy. On the other hand, most of the hypothetical compounds are above the convex hull and therefore much less likely to be synthesizable (see figure 2). DFT calculations of unstable compounds are still useful to have for boosting training sets of machine learning (ML) models as well as general understanding. However, while we can continue these exhaustive high-throughput DFT calculations using other common prototypes, we cannot do this for all 10 203 prototypes we have from ICSD. If we wanted to do exhaustive DFT for all prototypes up to five components using 76 elements in the periodic table (\( Z \leq 83 \) excluding noble gases, \( \text{Tc}, \) and \( \text{Pm} \)), then we would have to do trillions of DFT calculations.

We are interested in computationally more efficient methods to cheaply target hypothetical compounds that are on or near the convex hull. For example, we can take advantage of several already-developed methods that efficiently recommend compounds based on chemical similarity to other, known compounds. One such method, executed by Marques et al, recommends compounds starting from known compounds based on how often the relevant elemental substitution, e.g. \( \text{Fe} \rightarrow \text{Co} \), occurs in the ICSD between compounds of the same prototype. They used this approach to discover a staggering 18 479 new compounds on the convex hull, a major expansion which we have incorporated into the OQMD [19, 20]. Other methods based on chemical trends include one developed by Hautier et al, which performs common ion substitutions to generate new ionic compounds [21], as well as a method developed by Fischer et al which exploits correlations between prototypes that occur across phase diagrams [22].

2.2. Machine learning accelerated materials discovery

Machine learning (ML) algorithms are designed to automatically extract new knowledge out of data. We can use ML to learn more about materials and to create models that can accelerate the materials discovery. Among the first efforts in applying ML in materials informatics, Meredig et al [23] used a data-driven approach to learn the rules of chemistry from DFT calculations and make predictions for new compounds without knowing the structure information. Their ML model predicted around 4500 unknown ternary compounds by screening 1.6 million compositions, which largely increased the computational efficiency. To enable the fast development of ML-based models in materials discovery, Ward et al created a general approach and the associated software the Materials Agnostic Platform for Informatics and Exploration (Maggie) [24] to compute attributes for new materials, which includes both composition-dependent
attributes of elemental properties and crystal structure-dependent attributes derived from Voro

tessellation of crystal structures [25]. Magpie has been widely used to create materials ML models, which
include the discovery of new compounds for photovoltaic applications [24], the design of bulk metallic
glasses (BMG) [26], and the prediction of ternary mixed-anion semiconductors with tunable band gaps [27].
Recent advances in deep learning have also increased the prediction power in materials informatics. The
crystal graph convolutional neural network (CGCNN) framework developed by Xie and Grossman [28]
involves a material representation in the format of a crystal graph that encodes both atomic information and
bonding interactions between atoms, and can be trained to make predictions of various materials properties
through a convolutional neural network. An improved variant of the CGCNN model called iCGCNN was
developed by Park and Wolverton [29], which augmented the original model by incorporating the Voro
tessellation and explicit three-body correlations of neighboring constituent atoms in the crystal graphs. The
iCGCNN was demonstrated to have a mean absolute error of just 30.5 meV atom

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

by

b
Figure 3. Statistical summary of current OQMD. (a) Stability (convex hull distance) of all ICSD compounds calculated in the OQMD. (b) Stability (convex hull distance) of all ICSD and non-ICSD compounds. (c) Number of element types of all compounds in the OQMD. (d) Number of element types, (e) number of atoms, and (f) band gaps of all stable compounds in the OQMD. Space groups of (g) all stable ICSD compounds and (h) all stable non-ICSD compounds in the OQMD.

The majority of entries in the OQMD are ternary compounds, followed by the quaternary and binary ones (see figure 3(c)), which aligns with our material design decision that started from the simplest structures and gradually moves into more complex ones. The number of binary compounds is relatively less than that of ternary or quaternary ones due to its limited degrees of freedom. For the same reason, our high-throughput approach discovers 2118 new stable binary compounds, while 19,322 ternary and 14,375 quaternary hypothetical compounds are predicted to be stable (see figure 3(d)). With the improvement in computing power, we are now able to calculate more complex multi-component compounds and 2130 new stable quinary compounds have already been predicted in our ongoing calculations. Most of the hypothetical stable compounds in the OQMD have less than 20 atoms in their primitive cell, as shown in figure 3(e), but
we are continuously expanding our database with compounds having more than 20 atoms. The band gaps of stable compounds in the OQMD are also summarized in figure 3(f). Our current research mainly focuses on conducting materials and semiconductors, and we expand the phase space of materials with band gap 0, (0, 1] and (1, 2] eV by ~360%, ~220% and ~170%, respectively. As can be visualized in figures 3(g) and (h), within all stable ICSD compounds, the top three occurrences of space groups are Pnma (62), P2_1/c (14) and I4/mmm (139), while the top three space groups of hypothetical compounds are P6_2 m (189), Fm3m (225), and C2/m (12). This deviation is mainly caused by the specific research focuses, and several major projects performed using the OQMD platform are summarized in the following section 3.2.

3.2. Major projects

The growth of the OQMD mostly depends on the project-based high-throughput DFT studies performed by members of the Wolverton Research Group. We herein summarize six projects categorized by two major materials discovery logistics—structure-based projects (perovskites and mixed-anions) and property-based projects (thermoelectrics, solar fuels, batteries, and high-strength alloys).

3.2.1. Perovskites

Perovskite oxides have been at the center of numerous computational and experimental studies for a variety of applications, exhibiting properties suitable for technologies ranging from gas sensors [35] to solid oxide fuel cells [36, 37]. As such, a significant portion of the computational efforts through which the OQMD has grown over the years has been dedicated to populating the phase space with this structure type across compositions, considering not only several distortions of the ideal cubic phase but also mixing on both the A and B sites.

For simple ABO_3 perovskites, with over 5000 calculations, Emery et al. computed all possible permutations of metallic elements on both A and B sites in the cubic Pm-3 m structure, as well as (for cases where the energy of the cubic phase was within 500 meV atom by the convex hull) in its most common distortions (orthorhombic Pnma, rhombohedral R3m and tetragonal P4/mmm), finding over 300 stable compounds [17]. For A_2BB'O_6 and AAB'_BO_6 mixed perovskites, He and Wolverton performed a wide-ranging study comprising over 35,000 calculations [38]. To extensively cover the more commonly reported B site ordering, 10 different distortions of A_2BB'O_6 perovskites were considered (P2_1/c, C2/m, P1, P1, I4/m, I4/mmm, R3, R3, R3m, Fm3m), with permutations of Ca, Sr, Ba, and La (and, for the most common P2_1/c structure, Zn, Cd, Hg, and Pb) on the A site, and 50 metal elements on the B sites, totaling around 10,000 compositions, for over 500 of which a stable perovskite structure was identified. The more rarely observed A site ordering was then also investigated with over 3000 calculations of AA'B'_BO_6 perovskites considering 2 different distortions (Cmmm and P4/mmm) and combining rare earth, alkaline earth and transition metals on the A sites and again 50 metal elements on the B site. In order to increase the completeness of the convex hull, and thus the reliability of stability results, the studies have also included a number of competing phases by decorating the most common non-perovskite structures observed on the ICSD for generic ABO_3, A_2BB'O_6 and AAB'_BO_6 compositions.

3.2.2. Mixed anions

Materials with two or more types of anions (mixed-anion or heteroanionic materials) bring in new opportunities in materials design, which could lead to new properties that are not easily accessible in single-anion (homoanionic) materials [39]. For example, functional mixed-anion compounds have various energy-related applications including thermoelectrics [40], battery cathodes [41], and hydrogen evolution photocatalysts [42]. However, mixed-anion materials known to date only account for a small portion of the total number of combinatorial possibilities with periodic table elements, raising the question of whether is due to the scarcity of known mixed-anion materials being underexplored or simply relatively uncommon. Therefore, high-throughput DFT studies have the potential in unveiling the mystery of mixed-anion materials.

Various types of mixed-anion materials have been studied using the platform of the OQMD. Shen et al. [43] performed DFT calculations on 1188 ternary oxypnictides and found 42 of them that are experimentally unknown but predicted to be thermodynamically stable, largely expanding the number of stable oxypnictides. Amsler et al. [27] utilized an ML approach combined with a high-throughput method to investigate the ternary mixed-anion materials and predicted 21 new stable X_1Y_2Z compounds exhibiting band gaps suitable for energy applications. Quaternary oxycalogenides are a well-known material family having promising applications including thermoelectrics, transport conductors and solid-state electrolytes for Li/Na ion batteries. He et al. [44] screened over 5000 oxycalogenides compounds and found 129 hitherto unreported stable compounds and most of them are semiconductors. Given the large deficiency in the exploration of mixed-anion compounds, the OQMD is continuously working on the high-throughput
calculations of these materials, hoping to provide more insights for both experimental and theoretical research in mixed-anion materials.

3.2.3. Thermoelectrics
Thermoelectric (TE) materials have been widely studied in the past decade for their capacity to directly convert heat into electricity, which has significant promise in energy-related applications. Thermoelectric performance (conversion efficiency) is characterized by the figure of merit $ZT = S^2\sigma T / (\kappa_e + \kappa_l)$, where $S$ is the thermopower, $\sigma$ is the electrical conductivity, $T$ is the temperature and $\kappa_e$ and $\kappa_l$ are the electronic and lattice thermal conductivities, respectively. To increase the $ZT$ value of a material, we need to maximize its power factor (PF) $S^2\sigma$ and minimize its thermal conductivity ($\kappa_e$ and $\kappa_l$) simultaneously.

The OQMD provides a means to perform high-throughput studies to search for hitherto unknown materials exhibiting low lattice thermal conductivities, which may be potential thermoelectric materials. Through high-throughput screening, He et al. [45] discovered a new class of stable full Heusler compounds with ten valence electrons that have ultralow lattice thermal conductivity close to the theoretical minimum. Further investigation showed that the strong anharmonic rattling of the heavy noble metals in these materials accounts for the low lattice thermal conductivity and makes them excellent 'phonon-glass electron-crystal' materials. In another example, Pal et al. [46] predicted an enormous number of 628 thermodynamically stable quaternary chalcogenides AMM’Q3, and showed many of these compounds to be potential TE materials by validating the presence of low lattice thermal conductivity using the Peierls-Boltzmann transport equation. High-throughput DFT approach can also be used to search for new multicomponent bulk-nanostructured TE materials. As stated by Doak et al. [47], the screening-and-sorting procedure sorted out 130 candidates for high-performance thermoelectrics from a search space of 29 700 five-element systems, largely increasing the success rate in exploring new TE materials. Using a similar approach, Kocevski and Wolverton [48] screened for nanostructured two-phase Heusler TEIs in the OQMD and predicted 29 pairs that have not been considered previously for TE applications.

3.2.4. Solar fuel materials
Solar thermochemical water splitting (STCH) offers a renewable alternative to fossil fuels by enabling the production of hydrogen through a carbon-free process that can theoretically harness the energy of the entire solar spectrum. STCH consists of a two-step cycle consisting of a reduction step, in which an oxide compound loses oxygen upon heating, and a water-splitting step, where $\text{H}_2\text{O}$ is exposed to the reduced material, which re-oxidizes producing hydrogen [49]. Numerous efforts are being directed to the search for metal oxides capable of increasing the currently limited reported efficiencies, seeking compounds capable of reduction at suitable temperatures with fast kinetics while maintaining their structural stability through temperature change and oxygen loss.

In this context, high-throughput DFT and materials databases are powerful tools to expedite materials search by allowing for screening based on properties of interest on a large scale. Over the last few years, multiple such screenings have been performed utilizing the OQMD. Focusing on structure types with reported attractive STCH properties such as the ability to withstand oxygen nonstoichiometry and ease of oxygen diffusion, the OQMD has been populated with thousands of decorations of such prototypes. For the hundreds of new stable compounds thereby identified, the enthalpy change in the reduction step was then quantified by computing the oxygen vacancy formation energy, and utilized as a criterion to further screen the compounds. The screening was based on thermodynamic analysis [50], which showed the range of reduction enthalpies that allows for both steps of the STCH reactions to be favorable. Applying this methodology to perovskite, pyrochlore and spinel structure types, hundreds of new candidates [17, 51, 52] have been identified in recent years, with two mixed perovskites being successfully experimentally synthesized and cycled and displaying favorable redox thermodynamics and high fuel production [53, 54].

3.2.5. Batteries
To meet the increasing energy storage demands that accompany the expansion of the renewable energy and electric vehicle market, extensive effort is being directed to increasing batteries performance on multiple fronts. Leveraging the OQMD platform, numerous studies have been conducted to both accelerate the discovery of new anodes, cathodes, electrolytes and coatings materials for traditional lithium-ion batteries (LIBs) and aid in the development of batteries exploiting alternative chemistries.

To identify new anode materials with superior performance to the traditionally employed graphitic carbon, Kirklin and Wolverton, as one of the earliest applications of the OQMD, computed the lithiation reactions of transition metal silicides, stannides and phosphides and assessed their potential based on gravimetric and volumetric capacity, cell potential and volume expansion [15]. Examining Li, Na and Mg based anodes, Snyder et al. addressed the degradation in cell performance brought about by their reactivity...
with the electrolyte by surveying the database for thermodynamically stable materials to serve as coatings, therefore exhibiting chemical equilibrium with each anode type, then ranking the candidates according to their electronic insulation [55]. Addressing cathode degradation instead, Aykol et al screened the OQMD for various cathode coatings tailored to different battery chemistries, identifying the top 30 candidates for physical barrier, HF-barrier and HF-scavenger coatings by selecting thermodynamically and electrochemically stable materials and applying multi-objective optimization of filters relative to each functionality [56].

In the search for new cathode materials, Bhattacharya and Wolverton predicted several new candidates by surveying spinel oxides of the 3d transition metal series, computing their thermodynamic stability, cation site preference, and delithiation voltage [57]. Focusing on the \( \text{Li}_2\text{MO}_3 \) family, Kim et al identified a number of \( \text{Li}_2\text{MO}_3 \) active inactive electrode pairs, again computing their thermodynamic stability and delithiation voltage [58]. In a subsequent study, the authors then explored the possibility of embedding lithiated cobalt spinel in \( \text{Li}_2\text{MO}_3 \) based cathodes searching the \( \text{Li}–\text{Co}–\text{Mn}–\text{Ni}–\text{O} \) space and identified \( \text{Li}\text{Co}_{0.1875}\text{Ni}_{0.8125}\text{O}_2 \) as a promising candidate [59]. In a more open-ended approach, Aykol et al explored the \( \text{Li}_2\text{X}_3\text{Y}_2\text{O}_{12} \) garnet type structure identifying several stable compositions, then classified their potential applications as either anodes, cathodes, or electrolytes based on their computed electrochemical window [60].

Going beyond traditional LIB, alternative avenues to address limitations in capacity and cost are being explored through the development of battery materials exploiting different chemistries. Examining the possibility of the development of hybrid Li-ion Li–O2 cells, Kirklin et al screened the database for \( \text{Li}_2\text{O} \)-containing compounds stable upon \( \text{Li}_2\text{O} \) removal, then further selecting for a maximum reaction voltage and band gap on both products and reactants of each reaction, identifying 22 possible materials and reactions [14]. Focusing on reversible Ca-ion batteries, Yao et al identified new candidate anode materials by exploring all Ca–M systems reported in the ICSD, selecting the ones with a reported M compound, computing the voltage, and ranking them based on their maximum energy density and capacity [61].

### 3.2.6. High-strength alloys

Strong structural materials have various technological and constructional applications in nearly all aspects of human life. Age hardening, one way to increase the yield strength and hardness of pure metals and alloys, is realized by forming precipitates from the supersaturated solid solution through quenching and aging. Many known age-hardenable alloy systems have been discovered and optimized by extensive experimental trial and error. However, it is not feasible to exhaustively search for new precipitate strengtheners using brute force experimentation due to the extremely large number of possible compositional combinations.

Kirklin et al [16] conducted a high-throughput computational study on \( \sim 200,000 \) compounds to search for effective strengthening precipitates for three major alloy/precipitate systems, including FCC metals (with \( \text{L1}_1 \text{X}_2\text{Y}_2 \) precipitates), HCP metals (with \( \text{D0}_{19} \text{X}_2\text{Y}_2 \) precipitates), and BCC metals (with \( \text{L2}_1 \text{X}_2\text{YZ} \) precipitates). By using the OQMD, the thermodynamic stability of each compound was evaluated and the precipitate compounds were screened to be either in stable equilibrium with, or likely to form metastable precipitates in, the host matrix. Kirklin et al proved the effectiveness of their effort by recovering most of the widely known precipitate strengthening systems through their screening metrics. In addition, they also predicted 34 \( \text{L1}_1 \) precipitate compounds, 29 \( \text{D0}_{19} \) precipitate compounds and 50 \( \text{L2}_1 \) precipitate compounds as promising precipitate strengtheners in several common host matrices (e.g. Al, Fe, Mg, Ni, Co).

### 3.3. Data accessibility

The OQMD is comprised of a SQL database, an Application Programming Interface (API), and a web interface created using the open-sourced Django webframework. The function of the API in the OQMD is to populate the SQL database with new data generated via DFT and transfer relevant data from the database to the web interface upon the users’ request. The API for OQMD, called qmpy, is under active development at https://github.com/wolverton-research-group/qmpy. Users may download the entire OQMD database from OQMD.org and if desired, host it on their local machines. A locally hosted OQMD database can be accessed and modified using qmpy following the standard Django API protocols. Alternatively, users are able to interact with the database via the HTML-based web interface at OQMD.org to browse materials, create phase diagrams, and analyze the data generated from DFT calculations. The OQMD also supports the RESTful querying of the database via OQMD.org which enables data retrieval over basic HTTP connections in JSON format without requiring a web browser on the user side to parse the responses. The OQMD RESTful queries are to be made in accordance with the Open Databases Integration for Materials Design (OPTIMADE) protocols [62]. The OPTIMADE team has designed a universal API to make materials databases accessible and interoperable. The adherence to OPTIMADE is in support of creating an open and universal RESTful querying system for all material databases. Currently, the OQMD.org interface provides data including
material crystal structures, bandgap energy, formation energy, thermodynamic stability, electronic density of states, and phase diagrams, among others.

The qmpy API is constantly updated to use the stable python and Django versions in the interest of achieving network safety and data robustness. As of the writing of this article, the current version of qmpy (v1.4) uses python 3.7 and Django v2.2. Recent upgrades to qmpy are backward-incompatible with previous versions of MySQL DB files due to the drastic changes in dependent packages.

4. The openness of OQMD

4.1. Open access to the public
From the very beginning of the OQMD, a central guiding philosophy was to ensure open access to all public data. From the first release of the database in 2013, users have had the option to download the entirety of the database for use under a highly permissive license (CC-BY 4.0). We believed that there was significant value in performing science on the entirety of the database (rather than on individual entries or small samples) and that there would have great benefit to the community by allowing everyone that capability. This vision is strongly aligned with the notion of FAIR data principles, introduced one year later in 2014. The current success and proliferation of data-driven materials informatics methods can be partially attributed to the availability of these large, uniform, self-consistent HT-DFT datasets.

4.2. External use of OQMD
The openness of the OQMD has facilitated its use in advancing materials research across the world. In the years since its public release, the database contents have been leveraged by a large number of researchers across five continents for a variety of applications. Thanks to the richness of the phase space covered by the over one million compounds included, several computational studies have referred to the OQMD convex hull to select stable phases for applications such as photovoltaics [63], hydrogen production [64], and fuels for nuclear reactors [65]. Screening based on properties other than stability has also been widely employed in materials search for technologies ranging from batteries [66] to magnetic compounds [67]. Multiple phase spaces included in the database have also been utilized for studies based on structure types [68–71], and specific phase spaces [72–74].

On an even larger scale, OQMD data has found widespread employment in the training and validation of newly developed ML models. Focusing on formation energy prediction, Krajewski et al employed the compositional and structural features introduced by Ward et al [24, 25] and tested several neural network architectures, identifying the best-performing ones and developing an open sources tool with a user interface [75]. Including more generic structural information, Jain et al trained a representation learning feed-forward neural network on the 20 most common structure types on the OQMD exploiting exclusively atomic number and crystallographic symmetry information [76], and Jørgensen et al developed a message passing neural network based only on local information about bonding and symmetry [77]. Approaching the problem in a structure agnostic fashion, Jha et al developed a deep neural network model relying only on the compositional attributes [78], and, in a later work, employed residual learning [79] to tackle the decrease in accuracy observed with increasing depth of the network architecture. Targeting the bandgap as well as the formation energy, Goodall et al once again took a structure agnostic approach taking only stoichiometric data as input and reformulating it as a weighted graph between elements [80]. Min et al on the other hand, employed a gradient boost regressor model and an active learning process to identify new ferroelectric materials [81], and Stanley et al utilized a kernel ridge regressor to identify new mixed halide perovskites for photovoltaic applications [82]. Focusing on structural properties, Williams et al [83] and Zheng et al [84] both employed convolutional neural networks to predict the lattice parameters of, respectively, cubic perovskite oxides and full heusler compounds. In very recent works, Dan et al [85] and Zhao et al [86] both pursued the goal of designing new chemically valid hypothetical materials by employing generative adversarial networks, and Hu et al developed a global optimization-based algorithm to reconstruct crystal structures from atomic contact maps [87].

4.3. FAIR data in OQMD
The acronym FAIR in data stands for four aspects of data: Findable, Accessible, Interoperable, andReusable. The OQMD is committed to adhering to the FAIR data principles. Three noteworthy efforts toward FAIR-principles implementation are implemented, namely—Persistent Identifiers (PID), OPTIMADE RESTful API data transfer, and structured data for web indexing. The Handle persistent identifier registry has assigned the prefix https://hdl.handle.net/20.500.12856/ to the OQMD. Currently, every material entry in the OQMD has a PID URL associated with it in support of the Availability and Reusability of the data. The PID of a material is a permanent reference at the handle.net server, redirecting to the OQMD page of that
Figure 4. Percentage of all compounds on the OQMD convex hull that are sourced from the ICSD, plotted against the date of calculation. The OQMD started with ICSD compounds (100%), but ICSD compounds now make up just 32% of the OQMD convex hull. For this plot, all ICSD compounds are assumed to have been calculated prior to 2014, although many were calculated later than that.

material regardless of the URL changes at the oqmd.org domain. Perseverance of the PID also enables a unified reference to the original OQMD page when the OQMD database is hosted locally and modified according to external researchers’ preferences. The RESTful querying of the OQMD based on OPTIMADE API specifications contributes to the efforts of creating a unified material query system. A universal RESTful query system for all material databases is in support of the Findability, Availability, and Interoperability of FAIR principles. A structured version of materials data, in accordance with the vocabularies provided by schema.org, is included in the OQMD web pages for better indexing of relevant data by search engines and subsequently aiding in the Findability and Availability of the data. Both Material Entry pages and Material Composition pages in OQMD contain structured data, and are indexed by most of the popular web search engines.

5. Challenges for the future of HT-DFT and OQMD

The OQMD started out as a DFT database of primarily experimental compounds from the ICSD. However, recent high-throughput DFT projects, including exhaustive compound searches with common prototypes, application-driven searches, and efficient ML-based searches, have transformed the OQMD into a database of both experimental and hypothetical compounds. Since we first began with only ICSD compounds, the number of compounds that are stable, or on the convex hull, has more than doubled in size. In figure 4, we show how the OQMD convex hull has evolved over the years. Experimental compounds now make up just 32% of all compounds on the OQMD convex hull. This tells us that experimental databases like ICSD may only be scratching the surface of what’s possible for materials synthesis; there may be tens of thousands (or more) compounds that have yet to be synthesized. In addition, thousands of ICSD compounds have not been incorporated in the OQMD (e.g. solid solutions or disordered structures), which could impact the number of new predicted compounds. Going forward, we aim to explore ways to efficiently invest our computational resources to rapidly expand the OQMD to new materials. The reproducibility of high-throughput DFT data is also important to our exploration. A recent paper [88] presented a thorough study and compared the reported properties among three major HT-DFT databases including OQMD. According to their conclusion, the reproducibility of the DFT data is accepted among three databases compared to the differences between DFT and experiment.

However, a fair question to ask is how valuable the calculations of new materials will be as the database keeps growing beyond millions of compounds, especially with ML helping prioritize candidates that are more likely to be stable [89–91]. As the chemical spaces become more populated with stable compounds, finding new ones becomes more difficult due to the energetic competition with the existing convex-hull. In addition, higher-order compounds face more competition as they compete against lower-order ones as well (e.g. a new ternary compound in A–B–C chemistry competes not only with the other ternaries, but also with the binary compounds in A–B, B–C and A–C chemistries). As an outcome, the formation energies of higher-order stable compounds are heavily skewed towards lower energies [89]. Overall, this observation is in conflict with the notion that many new materials are awaiting discovery in complex, high-order chemistries [89]. Nevertheless, we should look beyond the narrow definition of the ‘material’ we have for structures in DFT databases, and for example, acknowledge that in higher-order systems, configurational
entropy effects will become a dominant factor in stability assessment, missing from the purely enthalpic convex-hull construction in the current database.

In effect, this is an indication that the next phase of high-throughput DFT databases, including the OQMD, should include calculations that require larger, complex structures such as for modeling of entropy effects in disordered compounds (e.g. via workflows for special quasirandom structures [92]) or such as modeling of point or line defects, surfaces, interfaces, thermal or ionic transport or incorporation of computationally-expensive functionals, such as HSE. There are, in fact, numerous early examples of such efforts [93–96]. In other words, we anticipate the utility of the OQMD and other HT DFT databases in the near future to be enhanced by going beyond perfect crystal structures, and starting to model more complex or ‘realistic’ aspects of materials in an HT fashion.

6. Conclusion

To celebrate the one million compounds milestone in the OQMD, we wrote this perspective to document the motivation for developing the OQMD, summarize the current status of the database, and raise some open-ended questions for the future of both high-throughput DFT calculations and materials databases. As an ever-growing materials database, the OQMD has been expanded for years through project-based HT DFT studies and is now taking the advantage of ML approaches to accelerate materials discovery. From the birth of the OQMD, the openness of its materials data to the public has been ensured as the main objective of the database and we are constantly working with other materials databases to reach a universal querying system in support of the FAIR principles. Researchers from all over the world have been intensively utilizing the OQMD to predict new materials for a variety of applications and to train some state-or-the-art ML models in materials discoveries. The OQMD has been and will continuously be serving as a pioneering fully open platform to advance the materials research for the whole community.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://OQMD.org.

Acknowledgments

The authors acknowledge support from the Award 70NANB19H0005 from the U.S. Department of Commerce, National Institute of Standards and Technology as part of the Center for Hierarchical Materials Design (CHiMaD), from the Materials Research Science and Engineering Centers (MRSEC) program (NSF Grant No. DMR-1720319) at the Materials Research Center of Northwestern University and from the U.S. Department of Energy, through the Office of Energy Efficiency and Renewable Energy (EERE) Contract DE-EE0008089. M A is currently employed at Rivian Automotive in Palo Alto, California.

ORCID iDs

Jiahong Shen  https://orcid.org/0000-0002-1951-2183
Sean D Griesemer  https://orcid.org/0000-0001-5531-0725
Vinay I Hegde  https://orcid.org/0000-0003-4786-823X

References

[1] Chase M W, Curnutt J L, Downey J R, McDonald R A, Syverud A N and Valenzuela F A 1982 Janaf thermochemical tables, 1982 supplement J. Phys. Chem. Ref. Data 11 695–940
[2] Kubaschewski O and Evans E L L 1958 Metallurgical Thermochemistry (International Series of Monographs in Metal Physics and Physical Metallurgy) 3d edn (New York: Pergamon)
[3] Committees, prepared under the direction of the ASM International Handbook 1992 ASM Handbook (Alloy Phase Diagrams) vol 3 (Materials Park, OH: ASM International)
[4] Mcmurdie H F and Hall F P 1949 Phase diagrams for ceramics: supplement no. 1 J. Am. Ceram. Soc. 32 154–64
[5] Villars P et al 2004 The pauling file European Powder Diffraction EPDIC 8 (Materials Science Forum) vol 443 (Trans Tech Publications Ltd) pp 357–60
[6] Pettifor D G 1984 A chemical scale for crystal-structure maps Solid State Commun. 51 31–34
[7] Perdew J P and Zunger A 1981 Self-interaction correction to density-functional approximations for many-electron systems Phys. Rev. B 23 5048–70
[8] Miedema A R, Boom R and De Boer F R 1975 On the heat of formation of solid alloys J. Less-Common Met. 41 283–98
[9] Jain A, Hauert G, Moore C I, Ping Ong S, Fischer C C, Mueller T, Persson K A and Ceder G 2011 A high-throughput infrastructure for density functional theory calculations Comput. Mater. Sci. 50 2295–310
[49] Roeb M, Neises M, Monnerie N, Call F, Simon H, Satller C, Schmuker M and Pitz-Paal R 2012 Materials-related aspects of thermochemical water and carbon dioxide splitting: a review Materials 5 2015–34
[50] Meredig B and Wolverton C 2009 First-principles thermodynamic framework for the evaluation of thermochemical H2O–Or CO2 -splitting materials Phys. Rev. B 80 245119
[51] Baldassarri B, He J, Qian X, Mastronardo E, Haile S and Wolverton C 2022 Accuracy of DFT computed oxygen-vacancy formation energies and high-throughput DFT search of solar thermochemical water-splitting compounds
[52] He J, Baldassarri B, Qian X, Haile S and Wolverton C 2022 Influence of exchange-correlation functionals on oxygen vacancy formation energy of transition metal oxides
[53] Qian X, Jiangang H, Mastronardo E, Baldassarri B, Wolverton C and Haile S M 2020 Favorable redox thermodynamics of SrTiO3−δ/4 in solar thermochemical water splitting Chem. Mater. 32 9335–46
[54] Qian X, Jiangang H, Mastronardo E, Baldassarri B, Yuan W, Wolverton C and Haile S M 2021 Outstanding properties and performance of CaTiO3−δ/4 for solar-driven thermochemical hydrogen production Matter 4 688–708
[55] Snyderack D H, Hegde V I and Wolverton C 2017 Electrochemically stable coating materials for Li, Na and Mg metal anodes in durable high energy batteries. J. Electrochem. Soc. 164 A3582–9
[56] Aykol M, Kirilkıncı S and Wolverton C 2014 Thermodynamic aspects of cathode coatings for lithium-ion batteries Adv. Energy Mater. 4 1400690
[57] Bhattacharya J and Wolverton C 2014 Prediction of quadratory spinel oxides as Li-battery cathodes: cation site preference, metal mixing, voltage and phase stability J. Electrochem. Soc. 161 A1440–6
[58] Kim S, Aykol M, Hegde V I, Lu Z, Kirilkıncı S, Croy J R, Thackeray M M and Wolverton C 2017 Material design of high-capacity Li-rich layered-oxide electrodes: Li2MnO3 and beyond Energy Environ. Sci. 10 2201–11
[59] Kim S, Hegde V I, Yao Z, Zhi L, Amsler M, Jiangang H, Shiqiang Hao J R Croy E L, Thackeray M M and Wolverton C 2018 First-principles study of lithium cobalt spinel oxides: correlating structure and electrochemistry ACS Appl. Mater. Interfaces 10 134739–90
[60] Aykol M, Kim S, Hegde V I, Kirilkıncı S and Wolverton C 2019 Computational evaluation of new lithium-3 garnets for lithium-ion battery applications as anodes, cathodes and solid-state electrolytes Phys. Rev. Mater. 3 1–7
[61] Yao Z, Hegde V I, Aspuru-Guzik A and Wolverton C 2019 Discovery of calcium-metal alloy anodes for reversible Ca-ion batteries Adv. Energy Mater. 9 1802994
[62] Andersen C W et al 2021 OPTIMIZE, an API for exchanging materials data Sci. Data 8 217
[63] Jain D, Chaube S, Khullar P, Goverapet Srinivasan S and Rai B 2019 Bulk and surface DFT investigations of inorganic halide perovskites screened using machine learning and materials property databases Phys. Chem. Chem. Phys. 21 19423–36
[64] Pandey M, Vojvodic A, Thygesen K S and Jacobsen K W 2015 Two-dimensional metal dicalcogenides and oxides for hydrogen evolution: a computational screening approach J. Phys. Chem. Lett. 6 1527–65
[65] Jiang C, Xie Y and Benson M T 2020 Efficient computational search for lanthanide-binding additive dopants for advanced U-Zr based fuels Materialia 10 100653
[66] Kercher A K, Kolopus J A, Carroll K J, Unocic R R, Kirilkıncı S, Wolverton C, Stooksbury S L, Boattner L A and Dudney N J 2016 Mixed polyionic glass cathodes: glass-state conversion reactions J. Electrochem. Soc. 163 A131–7
[67] Shen C, Qiang Gao N M Fortunato H K Singh I O, Gutfleisch O and Zhang H 2021 Designing of magnetic MAB phases for energy applications J. Mater. Chem. A 9 8805–13
[68] Yin M and Nash P 2015 Enthalpies of formation of selected Pd4YZ Heusler compounds Intermetallics 58 15–19
[69] Yin M, Nash P, Chen W and Chen S 2016 Standard enthalpies of formation of selected Ni3YZ Heusler compounds J. Alloys Compd. 660 238–65
[70] Yin M and Nash P 2015 Standard enthalpies of formation of selected Ru3YZ Heusler compounds J. Alloys Compd. 634 70–74
[71] Yin M and Nash P 2015 Standard enthalpies of formation of selected XXYZ half-Heusler compounds J. Chem. Thermodynamics. 91 1–7
[72] Gierlotka W, Dębski A, Terlicka S, Gasiор W, Peska M, Polański M and Tien Lin I 2021 Insight into phase stability in the Mg–Pt system. The ab initio calculations J. Phase Equilib. Diffus. 42 102–6
[73] Carone D, Mohammad Usman V V Klepov M D Smith V K, Besmann T M and Zur Loye H C 2020 New germanate and mixed cobalt germanate salt inclusion materials: [RbF(RbF)][Ge2O3]z and [(RbF)0.96(Rb1−xF)(Ge0.96O3.1)]0.5[(RbF)0.96(Rb1−xF)(Ge0.96O3.1)]0.5[(Ge0.96Ge2F2)2] CrystEngComm 22 8072–80
[74] Chrif–Aloaoui F Z, Nasik M, Kardellas S, Azza H and Selhaoui N 2018 Enthalpies of formation of the compounds of ternary phases equilibrium in the Al–Ni–Ti size system at 1073 K J. Mater. Environ. Sci. 258 109–109
[75] Krajeski A M, Siegel J W, Xu J and Liu Z K 2020 Extensible structure-informed prediction of formation energy with improved accuracy and usability employing neural networks SSRN Electron. J. 2081–36
[76] Jain A and Bilgaard T 2018 Atomic-position independent descriptor for machine learning of material properties Phys. Rev. B 98 1–7
[77] Jorgensen P B, Rio E G D, Schmidt M N and Jacobsen K W 2019 Materials property prediction using symmetry-labelled graphs as atomic position independent descriptors Phys. Rev. B 100 1–12
[78] Iha D, Ward L, Paul A, Liao W keng, Choudhary A, Wolverton C and Agrawal A 2018 ElemNet: deep learning the chemistry of materials from only elemental composition Sci. Rep. 8 1–13
[79] Iha D, Choudhary A, Tavaazza F, Liao W, Choudhary A, Campbell C and Agrawal A 2019 Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning Nat. Commun. 10 802
[80] Goodall R E A and Lee A A 2020 Predicting materials properties without crystal structure: deep representation learning from stoichiometry Nat. Commun. 11 1–9
[81] Min K and Cho E 2020 Accelerated discovery of potential ferroelectric perovskite: via active learning J. Mater. Chem. C 8 7866–72
[82] Stanley J C, Mayr F and Gagliardi A 2020 Machine Learning Stability and Bandgaps of Lead-Free Perovskites for Photovoltaics Adv. Theory Simul. 3 1–6
[83] Williams L, Mukherjee A and Rajan K 2020 Deep learning based prediction of perovskite lattice parameters from highfield surface fingerprint J. Phys. Chem. Lett. 11 7462–8
[84] Zheng X, Zheng P and Zhang R Z 2018 Machine learning material properties from the periodic table using convolutional neural networks Chem. Sci. 9 8426–52
[85] Dan Y, Zhao Y, Xiang Li, Shaobo Li, Ming H and Jianjun H 2020 Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials npj Comput. Mater. 6 1–7
[86] Zhao Y, Al-Fahidi M, Ming H, Edirisuriya Y S, Nasiri A and Jianjun H 2021 High-throughput discovery of novel cubic crystal materials using deep generative neural networks Adv. Sci. 8 2100566
[87] Hu J, Yang W, Dong R, Yuxin Li, Xiang Li, Shaobo Li and Siriwardane E M D 2021 Contact map based crystal structure prediction using global optimization CrystEngComm 23 1765–76

[88] Hegde V I, Borg C K H, del Rosario Z, Kim Y, Hutchinson M, Antonino E, Ling J, Saxe P, Saal J E and Meredig B 2020 Reproducibility in high-throughput density functional theory: a comparison of allow materials project, and OQMD (arXiv:2007.01988)

[89] Hegde V I, Aykol M, Kirklin S and Wolverton C 2020 The phase stability network of all inorganic materials Sci. Adv. 6 eaa5606

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

[91] Montoya J H, Winther K T, Flores R A, Bligaard T, Hummelsboj J S and Aykol M 2020 Autonomous intelligent agents for accelerated materials discovery Chem. Sci. 11 8317–32

[92] Zunger A, Wei S-H, Ferreira L G and Bernard J E 1990 Special quasirandom structures Phys. Rev. Lett. 65 353–6

[93] Xia Y, Hegde V I, Pal K, Hua X, Gaines D, Patel S, Jiangang H, Aykol M and Wolverton C 2020 High-throughput study of lattice thermal conductivity in binary rocksalt and zinc blende compounds including higher-order anharmonicity Phys. Rev. X 10 041029

[94] Montoya J H and Persson K A 2017 A high-throughput framework for determining adsorption energies on solid surfaces npj Comput. Mater. 3 14

[95] Mathew K, Singh A K, Gabriel J J, Choudhary K, Sinnott S B, Davydov A V, Tavazza F and Hemmig R G 2016 Mpinterfaces: a materials project based python tool for high-throughput computational screening of interfacial systems Comput. Mater. Sci. 122 183–90

[96] Restuccia P, Levita G, Wolocco M, Losi G, Fatt G, Ferrario M and Righi M C 2018 Ideal adhesive and shear strengths of solid interfaces: a high throughput ab initio approach Comput. Mater. Sci. 154 517–29