Computational Materials Science 50 (2011) 2295–2310

Contents lists available at ScienceDirect

Computational Materials Science

journal homepage: www.elsevier.com/locate/commatsci

A high-throughput infrastructure for density functional theory calculations

Anubhav Jain, Geoffroy Hautier, Charles J. Moore, Shyue Ping Ong, Christopher C. Fischer, Tim Mueller, Kristin A. Persson, Gerbrand Ceder *

Massachusetts Institute of Technology, 77 Massachusetts Avenue 13-5056, Cambridge, MA 02139, United States

ARTICLE INFO

Article history:
Received 17 December 2010
Received in revised form 14 February 2011
Accepted 18 February 2011
Available online 12 April 2011

Keywords:
High-throughput computation
Density functional theory
Materials screening
GGA
Formation enthalpies

ABSTRACT

The use of high-throughput density functional theory (DFT) calculations to screen for new materials and conduct fundamental research presents an exciting opportunity for materials science and materials innovation. High-throughput DFT typically involves computations on hundreds, thousands, or tens of thousands of compounds, and such a change of scale requires new calculation and data management methodologies. In this article, we describe aspects of the necessary data infrastructure for such projects to handle data generation and data analysis in a scalable way. We discuss the problem of accurately computing properties of compounds across diverse chemical spaces with a single exchange correlation functional, and demonstrate that errors in the generalized gradient approximation are highly dependent on chemical environment.

1. Introduction

The benefits of density functional theory (DFT) [1] calculations in the design and optimization of new materials have now been demonstrated across several research fields [2–7]. The scalability of computations makes it possible – at least in principle – to make predictions on thousands of compounds, and potentially for all known inorganic materials. The objective of the "Materials Genome" project [8] described in this paper is to enable accelerated materials discovery, and ultimately to develop a database of calculated properties and structural information on all known inorganic compounds for the materials community. In this paper, we describe the calculation infrastructure used to compute some properties of approximately 80,000 compounds, encompassing the majority of unique compounds in the inorganic crystal structure database (ICSD) [9,10], as well as many newly predicted systems. The number of calculations achievable is limited only by the prevailing computing technology and resources. Subsets of calculations performed with this methodology have been applied to structure prediction [11], screening of Hg sorbents [12], band gap prediction [13], and battery design [14,15].

The potential benefits of automating and scaling computational property predictions have been demonstrated in recent years by several research groups. Curtarolo et al. investigated the effect of structure on the stability of binary alloys using about 14,000 energies calculated from first principles [16]. More recently, Curtarolo et al. presented an overview of technical issues in high-throughput band structure calculations [17]. Ortiz et al. screened about 22,000 materials to suggest new materials for radiation detectors [18]. Several smaller DFT studies, involving hundreds of materials, have been performed by various research groups for catalysis and hydrogen storage [19–24]. In addition, a small number of general-purpose online electronic structure databases are now emerging [18,25–29], including a large (~81,000 DFT calculations) alloys database by Munter et al. [30]. Given the rising interest in high-throughput DFT, we describe in this paper some of the unique challenges faced when scaling to high-throughput as well as techniques to overcome these challenges.

Fig. 1 is the data flow diagram for this work, which we expect to be typical of most high-throughput computational screening projects. Because high-throughput calculations inherently involve generating, storing, and analyzing large amounts of data, a formal data flow strategy is needed to manage data efficiently. Fig. 2 is a visual overview of the technologies and techniques we used to implement the abstract steps in the data flow diagram. In the remaining sections, we examine each of the data flow stages in more detail.

2. Data selection

In this paper, we do not focus on data selection for high-throughput; however, several algorithms exist to optimize data selection when screening materials for a particular application. Many of these algorithms, such as tiered screening and evolutionary algorithms [37–46], have been outlined previously by Bligaard et al. [47]. If the intent is to create a general-purpose database, one additional approach is to compute compounds tabulated in a

* Corresponding author. Tel.: +1 617 253 1581; fax: +1 617 258 6534.
E-mail address: gceder@mit.edu (G. Ceder).

0927-0256/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.commatsci.2011.02.023

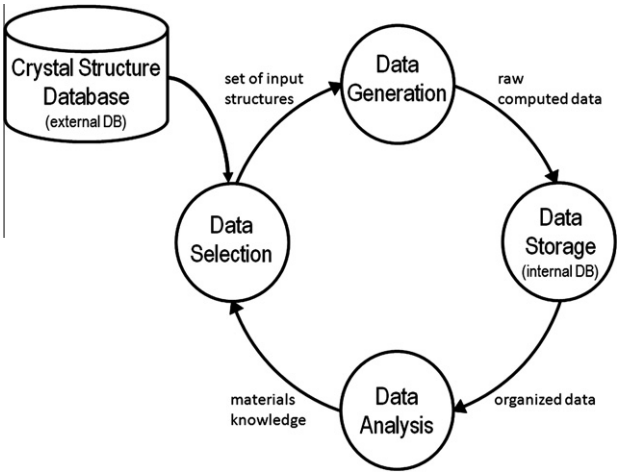

Fig. 1. Data flow diagram highlighting important steps in high-throughput computational screening. An external **crystal structure database** contains atomic positions and crystal parameters of known materials (either experimentally determined or from previous calculations), which can be used as a starting point for high-throughput screening. In the **data selection phase** interesting compounds are selected for computational testing. As an example, compounds may be taken directly from the external crystal structure database, or modified (e.g. via chemical substitution) from previously computations to generate new candidate materials, as discussed in Section 2. In the **data generation** phase, the structural data from the data selection phase are transformed into appropriate DFT input parameters and distributed over available computational resources. In the **data storage** step, the raw output files from the computations are sent to an efficient data storage/retrieval system (i.e., a database system). This internal database contains the results of all computations and may also contain experimental data. In **data analysis**, the user operates on a set of data to produce new information that may guide future data selection.

commercially-available external database such as the ICSD [9,10] or the Pauling File [48]. Because the materials in these databases are generally characterized experimentally, one can be confident that they can be synthesized. The computational space can further be expanded by using analysis of existing compounds to predict the existence of novel materials [11,31,32]. We have used a mixture of external crystal data and the prediction of novel materials to create our data set, having now computed most of the unique elements, binaries, ternaries, and quaternary compounds in the ICSD [9,10].

The search space accessible to high-throughput DFT calculations is presently limited to periodic unit cells containing up to approximately 200 atoms to ensure that accurate total energies can be obtained using reasonable computing time and memory resources. At present, the limited cell size restricts, for example, the degree of disorder that can be modeled for disordered/amorphous materials and the compositional resolution that can be probed when investigating doping or defect effects.

3. Data generation

After selecting compounds for computation, the next step in high-throughput screening (Fig. 1) is the generation of *ab initio* property data on these compounds. Although DFT calculations are well suited for high-throughput due to their relatively small number of adjustable parameters, automating DFT calculations is not yet trivial. As we will discuss, DFT calculations require making choices related to the accuracy of the computation, including the choice of the exchange–correlation functional, pseudopotentials, initial spin states, energy cutoffs, and *k*-point grid. In addition,

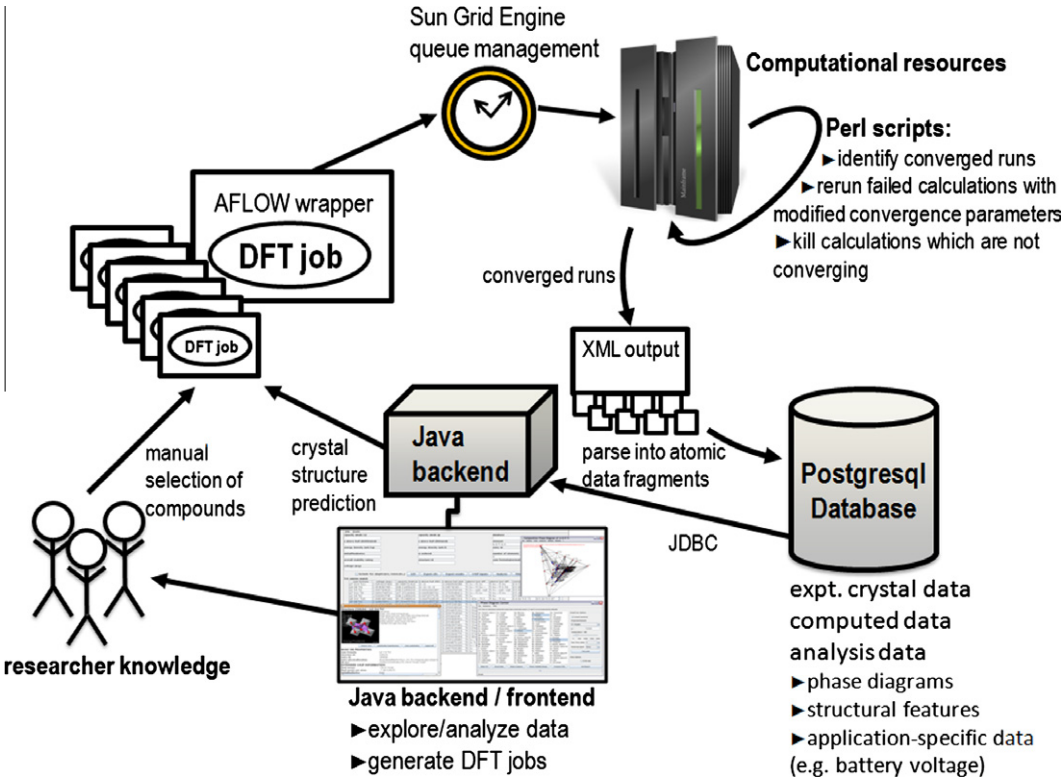

Fig. 2. Data flow implementation in our high-throughput project. Data selection is largely handled via researcher knowledge and crystal structure prediction algorithms coded in Java [11,31,32]. The Java backend is used to create batches of DFT jobs. These jobs are wrapped by the Automatic FLOW (AFLOW) [17] software, which optimizes each structure two times and handles some aspects of job convergence. The batches of jobs are submitted to a Grid Engine [33] queuing system. Active jobs are monitored and converged using Perl [34] scripts. Completed jobs are entered into a PostgreSQL [35,36] database, which interfaces with the Java backend through Java Database Connectivity (JDBC). A graphical front-end allows for data exploration and analysis, facilitating knowledge extraction.

numerical convergence must often be monitored and tuned through choice of matrix diagonalization schemes, charge mixing strategies, and *k*-space integration methods. High-throughput data generation thus involves both theoretical problems regarding the accurate treatment of diverse chemical systems and practical problems in attaining numerical convergence and monitoring DFT jobs. We describe in this section first the theoretical aspects, and then move to practical aspects of performing high-throughput DFT. Our calculations were performed using the Vienna *Ab Initio* Simulation Package (VASP) [49]; however, the vast majority of our discussion should be relevant for all DFT software.

3.1. Functional choice and +U correction

The Hohenberg–Kohn Theorem [1], which forms the foundation for DFT, states that (i) all ground state properties of a system, including the total energy, are some functional of the ground state charge density; and (ii) the correct ground state charge density minimizes the energy functional. In principle, this theorem implies that the ground state for any system can be determined by varying the charge density until the global minimum in the energy functional is found. However, the true functional relating the charge density to the energy is unknown. Much research in the DFT community is now focused on the development of approximate functionals that accurately represent the true energy functional. The different functionals, such as those based on the local density approximation (LDA) [50] and the generalized gradient approximation (GGA) [51], generally differ in the way the exchange–correlation component of the electron energy is treated. Although many other functionals exist with varying accuracies, computational expense, and target chemistries, no known functional is universally appropriate for all compounds and all materials properties. We chose the GGA functional as parameterized by Perdew et al. [52] as a good compromise between speed and accuracy when accurate ground state energies are targeted. The GGA is known to overestimate lattice constants (by an average of 0.076 Å in one test set of 40 semiconductors) [53], but is generally more accurate than LDA in computing cohesive energies and bulk moduli [54,55].

One of the challenges in high-throughput DFT is to accurately predict materials properties across wide chemical spaces that span a range of electronic arrangements (e.g., delocalized, localized). For example, both the LDA and GGA contain a spurious electron self-interaction energy (SIE) that generally over-delocalizes electrons in localized states. The SIE can result in large error in calculated reaction energies when electrons are transferred between delocalized states (as in metallic bands) and localized states (as in *d* or *f* orbitals in transition metal oxides) [56,57].

One method to address the SIE is the DFT + *U* framework [58], which adds an energy correction term to (typically) the *d* or *f* orbitals. There are several formulations of DFT + *U*. In this work, we use the rotationally invariant form as proposed by Dudarev [59]. Although the magnitude of the *U* correction can be determined self-consistently using a linear response scheme [56,60], a more common method of determining this correction is by fitting to some experimentally known quantity. We chose to use Wang et al.'s method [57] of fitting the *U* parameter to experimental binary formation enthalpies, which is simple and accurately reproduces phase stabilities. We apply the *U* correction to *d* orbitals only, and currently do not determine *f* electron *U* values; the full list of *U* values used is described in Table 1. The application of *U* is considered for oxides, fluorides, and sulfides, for which electron localization is particularly a problem. The choice of systems to which we apply *U* was largely determined by our experience and by systematic benchmarking [56,61–66]. All chemistries not listed in Table 1 are calculated without *U*, i.e., with the conventional GGA approach.

One disadvantage to the GGA + *U* framework is that energies from this technique cannot be directly compared with energies calculated using GGA. Our strategy to address this issue is to break down reaction energies into component reactions into one of three categories: (i) well-represented in GGA, (ii) well-represented in GGA + *U*, or (iii) binary reactions that produce systems with localized states (e.g., oxides, fluorides) from systems with delocalized electrons (such as elemental metals). The idea is that the last reaction is the source of the incompatibility between the GGA and GGA + *U*, and we can accurately bridge GGA and GGA + *U* calculations by using experimental formation enthalpies for these reactions.

Finally, we note that functionals that reduce the self-interaction error present in GGA might avoid the issue of mixing two theoretical frameworks. Hybrid functionals [67–70] are one such option and have been demonstrated to achieve good redox energies, transition metal oxide formation energies, and band gaps without the need for an adjustable *U* parameter [53,71]. However, they are at present too computationally expensive to be used on a large scale, and recent evidence indicates that they do not resolve some issues

Table 1
U values as fit using Wang et al.'s method [57]. All *U* values are for the *d* orbitals. Values marked with an asterisk are hypothesized and were not fit explicitly. We only employed a *U* correction for chemical systems where we expect a large degree of electron localization, i.e., the oxides, fluorides, and sulfides. Other chemical systems were run using conventional GGA. Our *U* values are shown alongside the values previously determined by Wang et al. for oxides (last column) [57].

Element	Environment	*U* value (eV)	*U* value reported by Wang et al. [57] (eV) (various reactions, oxides)
Ag	Oxides/fluorides	1.5	
Co	Oxides/fluorides	3.4	3.3
Cr	Oxides/fluorides	3.5	3.5
Cu	Oxides/fluorides	4.0*	4.0*
Fe	Oxides/fluorides	4.0*	3.9, 4.1
Fe	Sulfides	1.9	
Mn	Oxides/fluorides	3.9	3.5, 3.8, 4.0*
Mn	Sulfides	2.5	
Mo	Oxides/fluorides	3.5	
Nb	Oxides/fluorides	1.5*	
Ni	Oxides/fluorides	6.0	6.4
Ti	(All chemistries)	0.0	
Re	Oxides/fluorides	2.0*	
Ta	Oxides/fluorides	2.0*	
V	Oxides/fluorides	3.1	3.0, 3.1, 3.3
W	Oxides/fluorides	4.0*	
Y	(All chemistries)	0.0	

with phase stability in oxide mixtures [72]. The development of an efficient exchange–correlation functional that can deal with metallic and localized states in a consistent manner would greatly facilitate high-throughput searches through diverse chemical spaces.

3.2. Pseudopotentials, basis sets, and k-point meshes

DFT wavefunctions are expanded in a set of basis functions; for periodic solids, this basis set typically comprises plane waves. Although the computational accuracy increases with the number of plane waves in the basis set, so does the computational expense. In particular, accurately modeling the rapid changes in the wavefunction near core electrons requires large basis sets. It is often possible to obtain results with comparable accuracy using much smaller basis sets by using a *pseudopotential* to reproduce the behavior of core electrons in each element, explicitly modeling only valence or semi-core electrons [73–75]. As with the exchange correlation functional, many types of pseudopotentials are available. We chose the projected-augmented-wave (PAW) method [76,77], which can accurately reproduce the nodes in the core region of the valence wavefunctions while retaining small basis sets. For some elements with shallow semi−core states, we chose a version of the pseudopotential that explicitly solves a greater number of electrons, treating fewer electrons as core.

Calculation accuracy also depends on the number and choice of *k*-points used for Brillouin zone integrations. The standard method for converging DFT calculations is to increase the energy cutoff (or *k*-point density) until the result of interest no longer changes significantly with further parameter increases. When the energies of multiple compounds are being compared (such as when evaluating formation enthalpies), the standard procedure is to use the highest *k*-point density and energy cutoff of all compounds to eliminate systematic errors.

In a high-throughput project, this standard method of *k-point* convergence is impractical for multiple reasons. First, rigorously converging each of thousands of compounds requires considerable computational resources and complicates the data infrastructure. In addition, it does not remove the problem of systematic errors. Because compounds are calculated *a priori*, i.e., before one knows how their properties will be combined, the computed properties may have errors relating to different energy cutoffs and *k*-point densities within the set of compounds. More importantly, convergence is typically performed with regard to a particular computational property, such as final energy, position of the *d*-band center, or band gap. For a general-purpose database, the property of interest will not be known in advance, and there is no universal way to specify a convergence criterion.

The alternative approach we have adopted is to perform convergence calculations only after identifying a material and a property of interest. Materials are thus initially run with a "default" set of parameters that are likely to converge the total energy relatively well but that are unlikely to converge many other properties (such as band structure). We set the "default" energy cutoff to 1.3 times the maximum energy cutoff specified by pseudopotential [78]. We use a "default" *k*-point grid of (500)/*n* points, where *n* represents the number of atoms in the unit cell, distributed as uniformly as possible in *k*-space. A Gamma-centered grid is used for hexagonal cells, and a Monkhorst–Pack grid [79] for all other cells. We set the "default" electronic energy difference required for convergence to $n \times 5 \times 10^{-5}$ eV, and the energy difference required for ionic convergence to 10 times the electronic energy difference ($n \times 5 \times 10^{-4}$ eV).

To test if our relatively sparse *k*-point grids and somewhat loose electronic/ionic energy cutoffs produce reasonable total energies, we calculated 182 chemically diverse compounds with two sets of parameters:

(a) Our "default" parameters, as specified above.
(b) "Accurate" parameters, where the electronic energy difference and ionic energy differences for convergence were tightened to a constant value (5×10^{-7} eV for electronic convergence, 5×10^{-5} eV for ionic convergence), and a finer *k*-point mesh equal to (5000)/number of atoms in the unit cell was used.

Full ionic and electronic convergence was achieved starting from experimental lattice constants in the ICSD [9,10] for both the "default" and "accurate" parameters.

The results of our calculations are tabulated in Table A1 in the Appendix. We summarize these results in Fig. 3. In most cases, the "default" parameter set produces energies and cell volumes that are very close to the much more expensive "accurate" parameter set. The majority of compounds (∼71%) show less than a 5 meV/atom difference in energy and less than a 2.5% difference in cell volume (Table 2) between the two parameter sets. Almost all compounds (∼96%) are within a 15 meV/atom difference in energy and a 7.5% difference in cell volume (Table 2). From these results, we expect that the "default" parameters should be reasonably sufficient for screening purposes in applications for which energetics and structural details are targeted. However, screening applications that are particularly sensitive to small differences in energy or cell volume may benefit from using the "accurate" input parameters exclusively. In addition, we suggest that promising results be re-run with accurate input parameters after initial screening.

In a small number of compounds (∼4%), we find that results from the two parameter sets differ by over 15 meV/atom in energy or by more than 7.5% in cell volume. We find that these disagreements fall roughly into two classes. The first class encompasses compounds for which the energies obtained with the two parameter sets are in fairly good agreement, but the cell volumes are

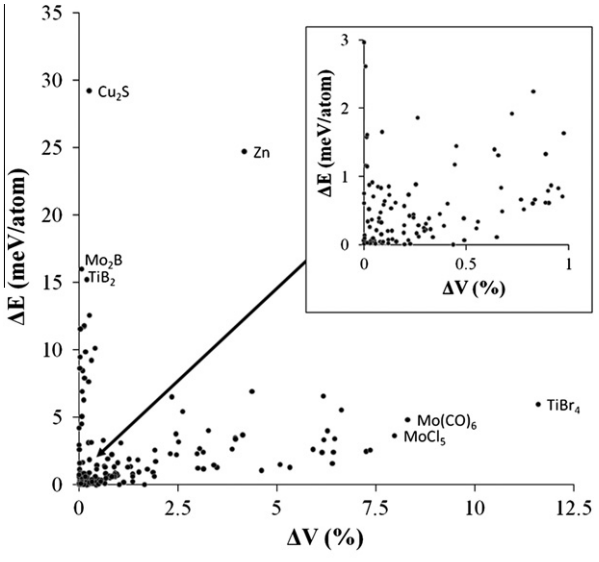

Fig. 3. Distribution of energy (ΔE) and cell volume (ΔV) differences computed using the "default" parameter set for convergence and the "increased *k*-point" parameter set (see text for details). The majority of compounds show close agreement between the two parameter sets (energies are within 5 meV/atom or less and volumes within 2.5%). We find two broad classes of disagreement: materials whose energies are in agreement but volumes are not, and vice versa. Examining the most egregious of these disagreements indicates that energy differences are likely to be related to the size of the *k*-point mesh, and volume differences are related to the convergence cutoff (see text and Table 3 for additional details).

Table 2
Summary of differences in results using "default" and "accurate" parameters. More detailed results are presented in Fig. 3 and in Table A1 in the Appendix. When comparing parameter sets, the majority of compounds (~71%) show very similar energies (<5 meV/atom difference) and very similar volume (<2.5% difference). However, almost 4% of compounds show fairly large disagreement of either over 15 meV/atom in energy or over 7.5% in cell volume. More information on these compounds can be found in Table 3.

Difference between "default" and "accurate" parameter sets	Number of compounds	Percent of compounds (%)
<5 meV/atom energy difference, <2.5% volume difference	130	71.4
<10 meV/atom energy difference, <5% volume difference	157	86.3
<15 meV/atom energy difference, <7.5% volume difference	175	96.2
Total compounds	182	100

Table 3
Compounds for which "default" and "accurate" parameters show large (>15 meV/atom or >7.5% cell volume) differences. The "tight" parameters employ the *k*-point mesh of the "default" parameters (500/atom) with the convergence criteria of the "accurate" parameters (5×10^{-7} eV for electronic convergence, 5×10^{-5} eV for ionic convergence). In cases where the "default" parameters show large disagreements in cell volume, but fairly good agreement in energy ($MoCl_5$, $Mo(CO)_6$, $TiBr_4$), the "tight" parameters reasonably reproduce the "accurate" parameters. However, in other cases, a large *k*-point mesh is needed to obtain accurate energies. In Cu_2S, the difference in energy between the "default" and "accurate" parameters is related to small atom rearrangements that can be modeled using the "tight" parameters.

Compound	Cell volume (Å^3), "default" parameters	Cell volume (Å^3), "tight" parameters	Cell volume (Å^3), "accurate" parameters	Energy (eV/atom), "default" parameters	Energy (eV/atom), "tight" parameters	Energy (eV/atom), "accurate" parameters
$MoCl_5$	1024.75	1112.03	1113.51	−4.135	−4.138	−4.138
$Mo(CO)_6$	913.27	990.16	995.99	−7.754	−7.758	−7.758
$TiBr_4$	1613.71	1815.48	1825.56	−4.068	−4.074	−4.074
TiB_2	25.75	25.71	25.70	−8.070	−8.070	−8.085
Mo_2B	73.03	72.97	73.08	−9.869	−9.869	−9.853
Zn	29.42	29.43	30.71	−1.242	−1.241	−1.267
Cu_2S	180.08	179.24	179.63	−3.956	−3.985	−3.985

largely in error. In Fig. 3, we have labeled $MoCl_5$, $Mo(CO)_6$, and $TiBr_4$ as compounds in this class. We find that in these compounds, the results can be greatly improved by using the tighter convergence cutoffs in the "accurate" parameters (5×10^{-7} eV for electronic convergence, 5×10^{-5} eV for ionic convergence), but without changing the *k*-point density (Table 3). These results suggest that if accurate cell volumes are targeted, tight convergence parameters may be more important than large *k*-point meshes. A second class of compounds shows large differences in energy between parameter sets but with similar cell volumes. In Fig. 3, we have labeled TiB_2, Mo_2B, Zn, and Cu_2S as compounds in this class. With the exception of Cu_2S, we find that simply increasing the convergence cutoffs do not significantly improve the "default" results for these materials (Table 3). We expect instead that these materials have somewhat complex Fermi surfaces for which large *k*-point meshes are necessary for accurate energy integration. For Cu_2S, we find somewhat anomalous behavior in that the energy difference is largely due to small atom rearrangements when running with higher convergence cutoffs. In this sense, Cu_2S behaves more like $MoCl_5$, $Mo(CO)_6$, and $TiBr_4$ in which improvements to the convergence cutoff are more important than the *k*-point mesh (Table 3).

3.3. Spin state and magnetic ordering

Spin state and magnetic ordering pose further challenges for high-throughput DFT. The true ground state is the global minimum of the energy functional, but current DFT codes typically only find local minima within this landscape. Hence, DFT calculations often do not converge to the correct spin state or the correct magnetic ordering unless they are initialized near that state. To obtain a magnetic ground state, one must in practice compute many initializations of magnetic ordering and treat the lowest-energy result as the ground state.

Although it is possible that automated magnetic ground state searches could be scaled to high-throughput, we have thus far found it to be computationally prohibitive to search for the correct magnetic state for all compounds. Instead, we initialize all ions that can be magnetic (Ag, Au, Cd, Ce, Co, Cr, Cu, Dy, Er, Eu, Fe, Gd, Hf, Hg, Ho, Ir, La, Lu, Mn, Mo, Nb, Nd, Ni, Os, Pa, Pd, Pm, Pr, Pt, Re, Rh, Ru, Sc, Sm, Ta, Tb, Tc, Th, Ti, Tm, U, V, W, Y, Yb, Zn, Zr) in our compounds ferromagnetically with high-spin, relying on the minimization algorithm to converge the magnetic ground state. Unfortunately, this strategy does not always find low-spin states correctly; for example, $LiCoO_2$ is known to contain low-spin Co^{3+} [80], but when initialized ferromagnetically, our calculations maintain Co^{3+} as high-spin. Initializing the $LiCoO_2$ calculation with a low-spin configuration correctly reproduces Co^{3+} low-spin as the lower-energy state.

To compromise between speed and accuracy, we initialize both high- and low-spin calculations for several ions that are known to often display low-spin configurations. These ions currently include Co-containing oxides (Co^{3+} in octahedral environments, for example, is well-known to be low-spin in several compounds due to a d^6 electron configuration [81]) and Mn, Fe, Cr, and Co-containing sulfides. As this list is not exhaustive, additional ions may be added in the future. While this strategy doubles the number of calculations performed for these compounds, it increases the possibility of finding the correct spin ground state. In general, we do not find antiferromagnetic states with this strategy. The energy penalty in incorrectly specifying the magnetic state depends heavily on the chemical system. As an example, the difference between antiferromagnetic and ferromagnetic energies in the lithium metal phosphates was found to be about 10–60 meV per transition metal by Zhou et al. [63]. However, other chemical systems may be more sensitive to magnetic ordering.

3.4. Convergence to the electronic ground state and error handling

In practice, the electronic ground state is solved via the Kohn–Sham formulation of DFT [50]. This formulation maps the physical system of interacting electrons to a new system of *non-interacting* electrons under an external potential (the Kohn–Sham Hamiltonian) that yields a solution charge density identical to the original interacting system. We follow the conventional iterative approach, which first chooses a trial charge density and initial wave function, and then iteratively improves these quantities until

a self-consistent solution is reached. An excellent and detailed discussion on these topics has been compiled by Kresse and Furthmuller [49]. The challenge in high-throughput DFT is to choose an algorithm that efficiently, correctly, and reliably carries out this procedure for many thousands of compounds without user intervention.

The numerical convergence to the ground state is affected, for example, by the choice of matrix diagonalization scheme used to solve for the wavefunctions, the charge mixing strategy, and the method of performing numerical integration of the energy over a finite k-point grid. While simple systems often converge rapidly and effectively using preset minimization algorithms and parameters found in modern electronic structure codes, difficult systems often require such parameters to be tweaked by the researcher to converge in reasonable time frames and with available memory. In high-throughput DFT, where hundreds or thousands of jobs are executed simultaneously, manual intervention in job convergence is rarely possible. We have therefore developed scripts coded in the Perl language [34] to monitor errors and adjust convergence parameters accordingly.

We describe our default convergence settings and some of the automated changes that are triggered when the electronic structure code fails to converge or produces an error. As the diagnosis and solution to some errors are specific to the electronic structure code used (and even specific to different versions of the same software), we will not discuss all errors, but instead focus on errors common across several software implementations of DFT.

For our default Hamiltonian matrix diagonalization algorithm, we use a built-in routine in VASP that uses the blocked Davidson approach [82] for the first few iterations, and then switches to the residual minimization method-direct inversion in the iterative subspace (RMM-DIIS) method [83,84]. Such a scheme is attractive because RMM-DIIS is known to be quite fast, but it requires good initial wavefunctions to converge to the correct ground state [49,78]; the initial iterations of the blocked Davidson method ideally generate a reliable wave function to pass into the RMM-DIIS method. However, this mixed scheme can still sometimes converge extremely slowly. Our algorithm therefore switches to a pure blocked Davidson approach, which is generally more reliable, if the electronic ground state is not found after 100 iterations.

As suggested in VASP 5.2 [78], we use as our default charge mixing strategy a Pulay mixer [84] that combines the input charge densities from all previous iterations in a manner that minimizes their residual vectors while conserving the number of electrons in the system. To further aid in convergence, the charge densities are preconditioned to dampen charge density changes at small wavevectors according to a scheme proposed by Kerker [85]. Although the magnitude of this preconditioning can be varied, results from Kresse and Furthmuller [49] indicate that the convergence behavior is quite good over a range of values. However, for slabs, magnetic systems, and molecules, the mixing parameters may need to be tuned [49,86]. When convergence problems are encountered, our algorithm sets the Kerker cutoff wave vector to 0.001 (in effect leading to a "linear mixing scheme") and multiplies the mixing parameter by the mean eigenvalue [78] to obtain more reliable mixing.

Finally, convergence is affected by the method used for numerical integration of the band structure energy over k space with a finite k-point mesh. By default, we employ the tetrahedron method, which linearly interpolates energies between calculated k-points, along with Blöchl's corrections [87] that both simplify the implementation of this technique and correct quadratic errors. Although the tetrahedron method converges quickly with respect to the number of k-points [49,77], it can fail for a small number of k-points and may require large meshes for metals and small-gap insulators where the Fermi surface separating occupied and unoccupied orbitals is complex and discontinuous. Therefore, when the VASP software reports that it cannot determine the Fermi level accurately by the tetrahedron method, we switched to a Gaussian smearing method that smoothens the discontinuity in the Fermi function at zero temperature.

Our software also handles many additional convergence issues and job errors that are specific to the electronic structure software employed. We do not discuss these errors here, and instead plan in the future to release open source versions of our job control scripts.

4. Data storage and retrieval

Data storage and retrieval require dedicated attention when scaling up to a high-throughput project. A well-designed architecture for data storage allows researchers to explore large amounts of data intuitively and naturally, greatly enhancing the possibility of finding new and interesting compounds for an application or discovering scientific trends in the data.

Many relational database systems for data storage are now available to researchers. These include, for example, MySQL [88] and PostgreSQL [35,36], which are free of charge, and commercial database systems such as Sybase [89] and Oracle [90]. Relational databases have the capability to store data compactly and efficiently query data, making them a good option for data storage in high-throughput projects. Such databases can also interface with many popular programming languages. In our high-throughput project, we chose the free relational database system PostgreSQL and interfaced it with a Java codebase using Java Database Connectivity (JDBC).

Considerable effort is involved in the setup and maintenance of a database system for storing and managing *ab initio* calculations. To design a relational database, the relevant data for storage needs to be identified, split into atomic pieces, and the relations between these individual pieces must be determined. Next, a database blueprint, or schema, must be designed to minimize redundant information while optimizing query and database insertion speeds. Several standards and techniques, such as the Atomicity, Consistency, Isolation, Durability (ACID) guidelines [91], database normalization [92], and $E-R$ diagrams [93] can aid in database design.

While we do not discuss database architecture extensively in this paper, we present in Fig. 4 a portion of our database schema for storing periodic crystal structures. The aim of this schema is to minimize redundant information by compartmentalizing information into multiple tables. For example, the *Elements* table contains basic properties of the elements (symbol, mass, atomic number, common valences); each element is defined only once in this table. Because the Elements table links to sites of our crystal structures stored in the *Site* table ('joins' to the *Site* table, in this case indirectly through the *Species* table), the basic element information is automatically stored within each crystal structure. This relational structure makes it possible, for example, to search for all crystal structures containing an element with an atomic mass greater than a specified value without individually storing the atomic masses associated with each crystal structure. Compartmentalizing data into several tables thereby minimizes redundancy, leading to more compact data storage and less potential for error during data updates.

However, because table joins can slow query performance, a relational database may need to balance query speed and data compactness. In Fig. 4, the *Element* table and *Entry* table are separated by four joins (through *Structure*, *Site*, *Species*, and finally *Elements*), which leads to somewhat slow performance when searching for entries containing particular elements (even after indexing). Because this is a very common query, we have created

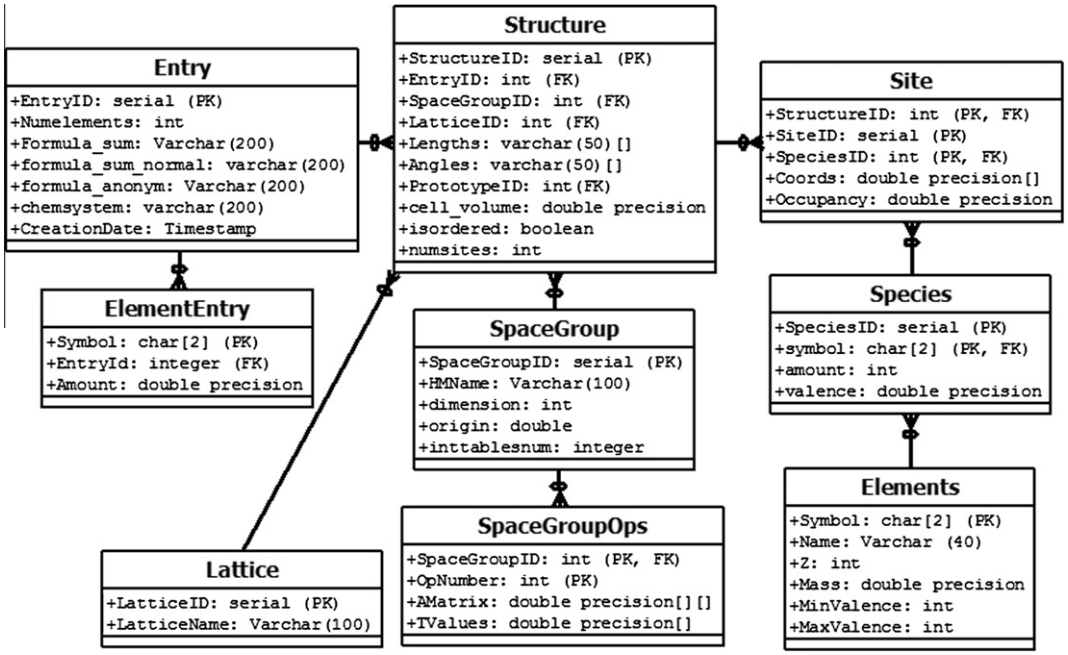

Fig. 4. Basic database schema for storing periodic crystal structures. Black connections indicate table joins, and a forked end indicates a *one-to-many* relation. Primary keys are indicated by (PK), and foreign keys are indicated by (FK). The *Entry* table contains basic chemical information about the entry, such as the chemical formula, the number of unique elements in the entry, and the chemical system (e.g., "Li–O–P"). An entry is not required to have a crystal structure; for example, in some experimental data the XRD pattern is known but the structure has not yet been refined. For entries where the crystal structure is known, the *Structure* table stores the lattice vectors and cell angles, along with somewhat redundant information, such as the cell volume and number of sites, for fast querying. The coordinates of the various sites are stored in the *Site* table, and the content of the site is stored in the *Species* table. The *Species* may be atomic (e.g., "Fe"), or it may be molecular (e.g., "H_2O"). The *Species* are composed of component elements, of which the properties are stored in the *Element* table. The *Lattice* table categorizes the lattice type of the structure (e.g., "fcc"). The *spacegroup* table contains basic information such as the spacegroup number and Hermann–Mauguin symbol, while the *SpacegroupOps* table lists the symmetry operations. As discussed in the text, the *ElementEntry* table allows for quicker searching over element symbols and contains redundant information with the *Element* table.

an additional table *ElementEntry*, which bypasses these joins and allows for quick searching over element symbols. Although the *ElementEntry* table introduces redundant information (the element symbols contained in a particular entry are stored both in *Elements* and *ElementEntry*), it has improved query speeds over element symbols.

The nine tables represented in Fig. 4 are only a small portion of our overall database, which currently contains 108 tables. The additional tables store properties of experimental crystal structures (e.g., journal information), properties of computed crystal structures (e.g., computational input parameters, user information, motivation for calculation), and post-processed information about the crystals (e.g., structure prototype, formal valence on ions).

The large number of database tables, while minimizing redundancy, can make the SQL statements for performing queries complex and unwieldy for novice users. We have therefore coded two types of mappings between the database and the Java codebase. The first mapping allows database tables to be translated into Java objects, so that users operate on high-level objects within Java rather than query individual data pieces from the database. The second mapping serves a translation layer between the user and the SQL database; the user needs only specify the constraints and the desired properties within a Java API, and the translation layer creates the necessary SQL syntax (including JOIN statements). This second mapping layer separates the content of the user query from the database architecture, thus allowing changes in the architecture to proceed without affecting user behavior.

We note that identifying the data that should be stored in the database is itself a challenging problem. It is at present impractical to store all data from a DFT calculation in the relational database, as one calculation may easily produce several hundred megabytes of uncompressed information. Our solution has been to compress and archive the largest components of the computed data, such as the charge densities and wavefunctions, to a dedicated file server. The database stores only the most heavily queried items; for the remaining data, the database stores links to the location of the original calculation files on the file server. With this setup, all data from a calculation is linked by the database to the external file server, and the most heavily accessed data can be quickly searched, retrieved, and sorted using the built-in functionality of the database.

We have included several features in the database that allow for more robust searches. As an example, calculations are automatically classified by their exchange–correlation functional type, pseudopotentials, and use of the DFT + *U* methodology. The software can thus automatically prevent combining the results of calculations that use, for example, two separate pseudopotentials for the same element. In addition, analysis codes can easily restrict themselves to a chosen theoretical framework. For example, we can create phase diagrams using GGA calculations only, GGA + *U* calculations only, or a mixture of the two. We expect that the ability to naturally accommodate calculations with heterogeneous input parameters will become increasingly important as more theoretical frameworks are tested and added to our data set.

Almost all calculated compounds in the database are structurally linked to experimental crystal data or other calculations. At present, the links are documented at the time of input file generation; when the software creates DFT input files (e.g., by direct transcription of an existing compound or by substitution of chemical species), it also notes the original compound and the structural transformations that were applied to it. The database later incorporates these links, so that it is possible to trace back the origin of any compound in the database. These links are helpful in several ways. For example, when discovering an interesting compound in the database, it is possible to trace back the original experimental research paper as a starting point for synthetic routes to that compound. The structural links can also serve to restrict analysis to

compounds known to exist experimentally, or to filter compounds that were designed via chemical substitution or structure prediction. An example of such a use is the generation of phase diagrams based only on experimental crystal structures [11,14]. In the future, we expect to complement these structural links with an algorithm that crawls the database and automatically determines relations between compounds.

We have added the ability for users to add custom notes to compounds. Users can then search within their own notes or across all notes. We are also beginning to add experimental data, such as measured binary and ternary formation enthalpies, natively to the database and link them with computed entries. Thus, a report of computed properties of a compound may also list known experimental data. We hope that by organizing materials data, new paths to materials discovery and insight will be uncovered.

5. Data analysis

The data analysis process will depend heavily on the area of application. In Section 6, we present one example of data analysis that examines the error of binary and ternary formation enthalpy calculations calculated with GGA. Here, we summarize several types of analyses that have broad applications:

(i) *Structural equivalence* – Many properties of materials are correlated to their crystal structure, but it can be difficult to determine equivalence between crystal structures. Raw crystal structure data containing atomic positions may need to be classified into crystal structure prototypes so that structural "equivalence" or "similarity" to other structures can be detected. We have implemented such a crystal structure prototyping scheme based on affine mapping [94,95] and have used it to prototype entries in the ICSD. A version of this algorithm is available online [8].

(ii) *Valence states of ions* – Similarly, to enable property correlations to the formal valence state of ions, we have implemented a valence designation scheme based on bond-valence sums [96] and Bayesian probabilities.

(iii) *Phase diagrams* – The phase diagram of a multi-component system is of interest in many materials design problems, such as the design of multi-phase materials with optimized functionality [97], the understanding of the stability of a material under various experimental and usage conditions [98,14], and reaction paths. With our comprehensive database of energies for a large number of materials, the ground state phase diagram for most multi-component systems can be calculated almost instantaneously. We have developed this capability in the form of a phase diagram module that provides the ability to generate compositional and grand canonical phase diagram constructions, perform thermal stability analyses, and elucidate decomposition paths. As an example, the thermal stability analysis of the delithiated $LiMPO_4$ (M = Fe, Mn) systems by Ong et al. [14] was performed using this high-throughput infrastructure and the above analysis tools.

(iv) *Electronic structure* – Electronic structures are of interest to many applications, e.g., solar cells, thermoelectric, and transparent conductors. We have developed tools to view the calculated total and projected DOS of materials and determine band gaps for high-throughput searching. Optical properties and band structures may be pursued in the future.

We have developed a user interface coded as a Java application capable of performing the above analyses (Fig. 5). The authors are currently working on a tool to bring these features to the larger materials community over the World Wide Web [8].

6. Cancelation of errors in GGA and formation enthalpy dependence on reference states

No single exchange correlation functional achieves consistent accuracy across a diversity of chemical environments. High-throughput data sets can be used to probe the accuracy of a particular exchange–correlational functional either within a single chemical class or across chemistries. Such a study was carried out by Curtarolo et al. for binary metals [16] and pure elements [99] to evaluate the accuracy of LDA and GGA predictions of the relative stabilities of crystal structures. More recently, Lany examined 61 binary formation enthalpies (from the elements) of semiconducting and insulating compounds under the LDA and GGA frameworks, and found large differences (an rms deviation of 0.18 eV/atom for LDA and 0.24 eV/atom for GGA) between experimental and computed values [100]. Lany subsequently fitted an energy correction for each element in the study; by fitting 14 element energies, Lany was able to reduce the rms error to 0.07 eV/atom for both LDA and GGA in the 61-compound test set [100].

In the same spirit as these earlier studies, we examined the accuracy of GGA in predicting the formation of ternary polyanion-containing compounds (silicates, borates, phosphates, carbonates, sulfates) from the elements. In particular, we investigated whether element corrections may be needed to accurately describe ternary polyanion-containing formation reactions, as was found previously by Wang et al. for binary oxides [57] and Lany for semiconducting and insulating binary compounds [100]. We also discuss the universality of such element corrections.

6.1. Methods and results

Our calculations were extracted from our database rather than performed specifically for this analysis, demonstrating the value of a general-purpose electronic structure database. The number of compounds examined was largely limited by available experimental data. In total, we examined 57 ternary compounds, 6 binary oxides, and 15 other binary compounds. All calculations were generated using the methodology and parameters described in Section 3. For Na_2O formation, we used a PAW pseudopotential that explicitly modeled 7 electrons; we found that a one-electron pseudopotential gave inaccurate results. The O_2 gas energy was calculated using a $13 \times 10 \times 11$ Å^3 supercell with a Gamma-centered $2 \times 2 \times 2$ k-point mesh. Our formation enthalpies were approximated using zero-temperature total energies that neglected pressure and zero-point effects, as detailed in a prior publication [12]. Experimental formation enthalpies were compiled from Kubaschewski et al. [101].

The ternary test set includes compounds of the form $A_iX_jO_k$, in which A represents group I and II metals plus aluminum, X includes the nonmetals {Si, B, P, C, S}, and i, j, and k represent arbitrary coefficients. We have restricted our metals A to alkali, alkaline earth, and Al to avoid errors arising from incomplete cancelation of self-interaction errors due to localized d or f orbitals. This latter problem is more appropriately addressed via the GGA + U framework, as was demonstrated by Wang et al. for binary oxides [57]. The full dataset of calculations is presented in Appendix Table A2.

A known problem in computing GGA formation reactions involving oxygen is the choice of O_2 reference energy [57,102,103]. Wang et al. earlier analyzed the oxidation energies for a large number of binary oxides [57] and identified two sources

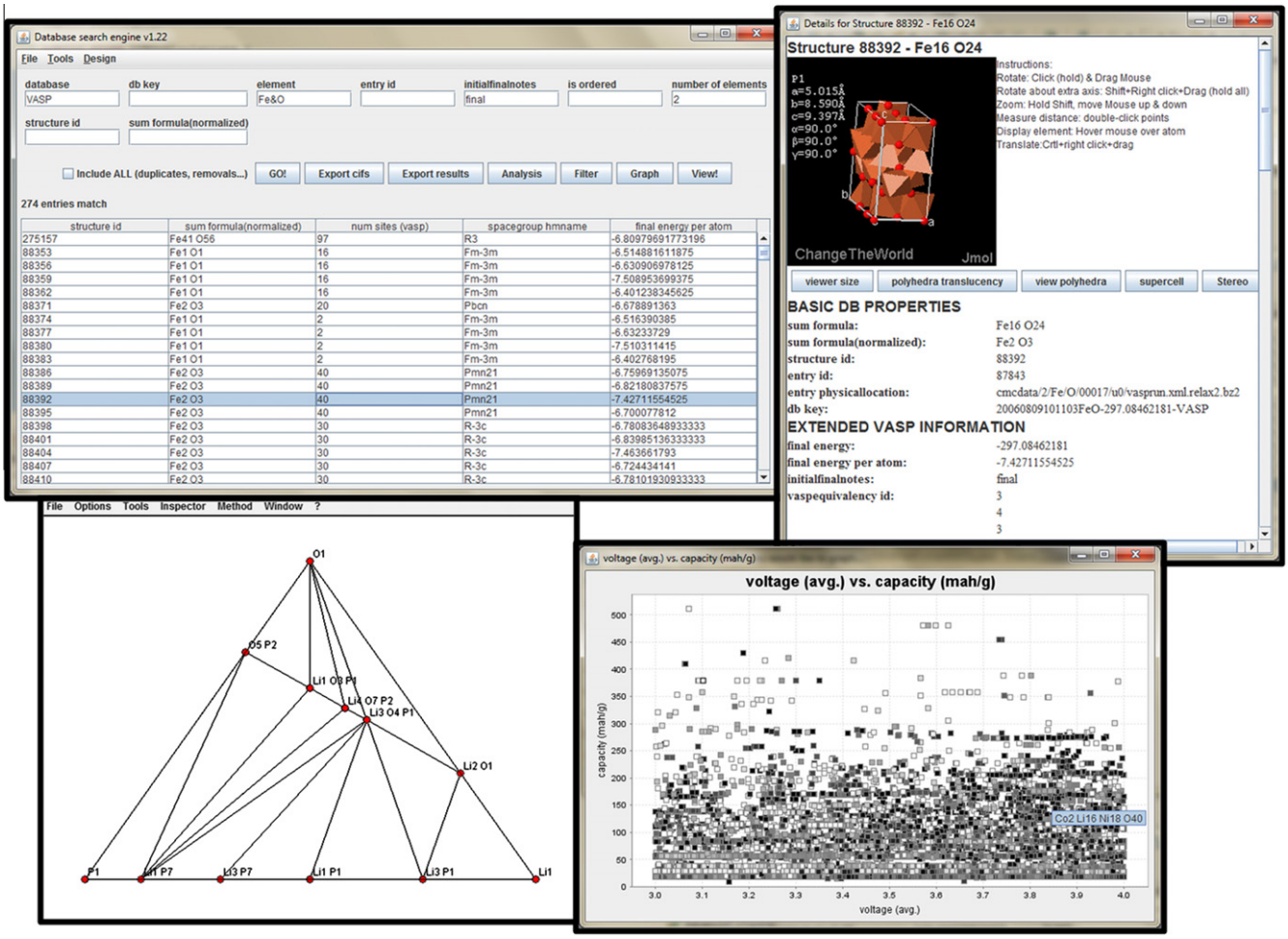

Fig. 5. Screenshot of Java application for performing exploratory data analysis. Clockwise from top-left: the main query window for finding results given user-defined constraints, a detailed view of a single compound, a user-defined interactive chart for exploring data on Li ion battery compounds, and the phase diagram tool to generate convex hull constructions. Not shown are interfaces for comparing structures, finding formal valence/coordination, calculating reaction energies, viewing density of states for an entry, and interactively defining and generating VASP input files from database entries.

of errors in the calculated formation energies: (i) an overbinding of the O_2 molecule in GGA by about 0.79 eV, and (ii) an additional error of 0.57 eV due to incomplete error cancelation when transferring electrons to the O^{2-} p orbitals in oxides. Wang et al. therefore proposed that the GGA O_2 energy be destabilized by the sum of both errors, or 1.36 eV/O_2, when calculating metal oxidation energies [57]. For the remainder of this section, we will refer frequently to two O_2 reference states based on Wang et al.'s results [57]:

1. Binding Energy Corrected Oxygen (BECO), for which the O_2 energy has been destabilized by 0.79 eV to correct for GGA overbinding the O_2 molecule.
2. Reduction + Binding Energy Corrected Oxygen (R + BECO), for which the O_2 energy has been destabilized by 1.36 eV (0.79 eV for the binding energy error, plus 0.57 eV for the error of O_2 to O^{2-} reduction and incorporation into the solid phase).

Given the uncertainty in the proper oxygen reference state, we present our ternary formation energy results so that they may be evaluated independently of the O_2 energy. Fig. 6 plots the difference between computed and experimental formation enthalpies (per O_2) for our $A_iX_jO_k$ test set against the X/O_2 ratio when using an uncorrected O_2 gas energy. Also plotted in Figure are two horizontal dashed lines representing the BECO and R + BECO adjustments to the O_2 energy. The distance of points from these two lines represent the error if the O_2 energy is corrected based on the respective scheme. From Fig. 6, we may make the following observations:

1. The binary A_yO_z oxides, corresponding to an X/O_2 ratio of zero, are positioned along the y-axis in Fig. 6 and are represented by a brown 'x'. In agreement with the previous results of Wang et al. [57], we find that using the GGA O_2 gas energy greatly underestimates the magnitude of binary formation enthalpies by an average of about 130.4 kJ/mol-O_2. The R + BECO adjustment of 131.22 kJ/mol-O_2 reduces the mean absolute error (MAE) over our oxides data set to about 8.5 kJ/mol-O_2, which is more than an order of magnitude improvement.
2. The error/O_2 of ternary polyanion-containing compound data (all points for which X/O_2 > 0) in Fig. 6 are in general lower than the R + BECO dashed line, demonstrating that this adjustment overstabilizes ternary compounds. For example, the MAE for carbonates and sulfates (orange ' + ' symbols and green squares) using the R + BECO adjustment is 54.3 and 55.4 kJ/mol-O_2, respectively. These errors are comparable to the reduction component of the R + BECO adjustment. If this portion of the correction is removed and only a binding energy correction is employed (BECO adjustment), the MAE is reduced by over an order of magnitude to 6.6 and 6.5 kJ/mol-O_2 for carbonates

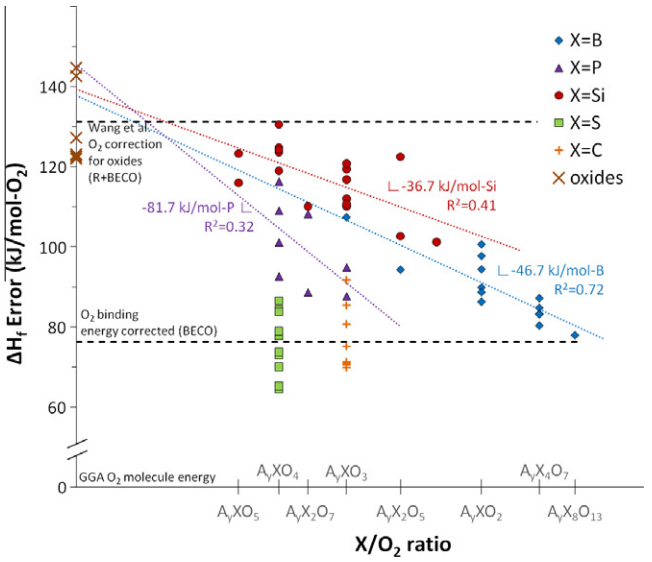

Fig. 6. Difference in formation energies of experimental [101] and GGA formation energies (ΔH_f error) per O_2 for $A_xX_yO_k$ compounds as a function of X/O_2 ratio when using an uncorrected GGA O_2 energy (positive errors correspond to the compound being under-stabilized in GGA). Each point represents one compound and is classified by the type of nonmetal X bonded with oxygen in the polyanion group. Errors when using two other O_2 reference energies can be measured by distance from the dashed lines labeled 'O_2 binding energy corrected (BECO)' and 'Wang et al. O_2 correction for oxides (R + BECO)' [57]. No oxygen reference state accurately describes all results, and errors per O_2 appear to depend linearly on X/O_2 ratio. We have fit least-squares regression lines through the errors (dotted lines) to determine the O_2 energy (y-intercept of dotted lines) and X elemental energy (slope of dotted lines) that best describes the data for each chemical class. These regression lines suggest that we can model the errors fairly well by using the R + BECO oxygen adjustment and applying a constant shift to the X elemental energies. The full data set is presented in Appendix Table A2.

and sulfates, respectively. In all cases, the uncorrected GGA O_2 calculation severely underestimates the formation enthalpy (by about 60–130 kJ/mol-O_2).

3. For phosphates, borates, and silicates, there appears to be a linear dependence of the error per O_2 on the X/O_2 ratio. We performed a least-squares linear regression for each polyanion series (dotted lines), where the y-intercept represents an O_2 correction corresponding to that polyanion series, and the slope represents a correction related to the X in the polyanion. We note that the R^2 coefficients of our regression analysis are not particularly high (R^2 of 0.32, 0.41, and 0.72 for phosphate, silicates, and borates, respectively), which may partially be due to the limited data available. Nonetheless, we may observe that in all cases, the y-intercepts are within 15 kJ/mol of the R + BECO adjustment. This suggests that we can model the errors using constant element shifts by constraining the fitted O_2 energy to the R + BECO adjusted energy and fitting constant energy adjustments to the elemental X reference states.

Table 5
Errors per mol-atom for binary phosphides, silicides, carbides, and sulfides when using pure GGA (first data column) or the element corrections derived in Table 4 for ternary polyanion-containing systems. Errors are with respect to experimental values from the Kubaschewski table [101] as listed in Appendix Table A2. All formation energies are underestimated both with GGA and GGA with element corrections. The errors are much smaller when using pure GGA, suggesting that element corrections derived for a particular chemical class should not be broadly applied outside that class.

	ΔH_f error(kJ/mol-atom) pure GGA	ΔH_f error(kJ/mol-atom) element corrections derived for ternary polyanion-containing systems
AlP	23.19	49.62
Mg2Si	10.69	17.84
Carbides (4)	19.96	68.40
Sulfides (9)	26.34	73.42

To present our errors in a more universal way, we compile in Table 4 errors normalized per mol-atom rather than per mol-O_2 for each chemical class and three methods of calculating ternary formation enthalpies. In column (1) are the formation energy errors if the O_2 energy is corrected using the R + BECO adjustment. The data in column (2) are the formation energy errors using the BECO adjustment. In column (3), we present the formation energy errors if the O_2 energy is corrected using the R + BECO adjustment and the element X energies are adjusted to minimize the least-squares error, a strategy suggested by our linear fits in Fig. 6.

Table 4 demonstrates that neither the R + BECO nor the BECO adjustments provide good results across several chemical spaces. The R + BECO adjustment demonstrates errors of over 10 kJ/mol-atom for the borates, carbonates, and sulfates, whereas the BECO adjustment demonstrates errors of over 10 kJ/mol-atom for silicates and binary oxides. Thus, correcting only the O_2 reference state does not provide good accuracy across both oxides and ternary polyanion-containing compounds. Much better accuracy can be obtained by combining R + BECO adjustment with constant shifts to the set of element energies for X. Although this increases the number of fitted elements over the data set from one to six, it reduces the MAE over all 63 compounds in the test set to under 2 kJ/mol-atom.

6.2. Discussion

The GGA errors in the formation of of A_yXO_z ternary polyanion-containing compounds depend on both the O_2 reference employed

Table 4
Errors per mol-atom over ternary polyanion-containing compounds and binary oxides, divided by chemical class and element adjustments to GGA. Errors are with respect to experimental values from the Kubaschewski table [101] as listed in Appendix Table A2. The first two data columns represent errors when adjusting only the O_2 energy. The third data column represents errors when adjusting the O_2 energy and fitting adjustments to the X element energy energies. While this last method produces the best results, it also contains the highest number of adjustable parameters. In addition, it appears that these element adjustments do not generalize well beyond binary oxides and ternary polyanion-containing systems (discussed in Section 6.2).

System	ΔH_f error(kJ/mol-atom) R + BECO adjusted O_2	ΔH_f error(kJ/mol-atom) BECO adjusted O_2	ΔH_f error(kJ/mol-atom) R + BECO adjusted O_2 plus fitted element corrections	Fitted adjustment to elemental X energy (kJ/X)	Pauling electronegativity difference between oxygen and X
Silicates	4.05	11.48	1.58	−21.48	1.54
Borates	11.21	3.56	1.08	−39.03	1.4
Phosphates	9.33	7.08	2.30	−52.85	1.25
Carbonates	14.68	1.72	1.86	−79.77	0.89
Sulfates	17.3	2.02	2.03	−107.74	0.86
Binary oxides	2.11	12.13	2.11	N/A	N/A

Table A1
Results of computations for 182 compounds using "default" and "accurate" parameters as described in the text.

Formula	Spacegroup	Energy (eV/atom) "Default" parameters	Energy (eV/atom) "Accurate" parameters	Cell volume (Å^3) "Default" parameters	Cell volume (Å^3) "Accurate" parameters	ΔE (eV/atom)	ΔV (%)
Elements							
Al1	F m −3 m	−3.752	−3.741	16.37	16.43	−0.011	−0.37
Au1	F m −3 m	−3.275	−3.274	18.09	18.05	−0.001	0.22
B1	R −3 m R	−6.652	−6.652	817.32	817.64	0	−0.04
Bi1	R −3 m H	−3.871	−3.872	73.36	73.69	0.001	−0.45
Ca1	F m −3 m	−2.001	−2.004	41.45	41.89	0.003	−1.05
Cu1	F m −3 m	−3.735	−3.725	11.98	12	−0.01	−0.17
Fe1	I m −3 m	−8.317	−8.317	11.36	11.36	0	0.00
I1	C m c a	−1.515	−1.517	187.2	190.89	0.002	−1.93
K1	I m −3 m	−1.096	−1.095	71.71	71.87	−0.001	−0.22
La1	P 63/m m c	−4.928	−4.92	146.38	146.73	−0.008	−0.24
Li1	I m −3 m	−1.899	−1.896	20.17	20.3	−0.003	−0.64
Mg1	P 63/m m c	−1.534	−1.541	45.73	45.79	0.007	−0.13
Mo1	I m −3 m	−10.933	−10.945	15.65	15.63	0.012	0.13
Na1	I m −3 m	−1.305	−1.305	36.73	36.7	0	0.08
P1	P −1	−5.267	−5.269	759.68	800.35	0.002	−5.08
Si1	F d −3 m S	−5.411	−5.422	40.81	40.76	0.011	0.12
Sm1	R −3 m H	−4.715	−4.715	101.74	101.62	0	0.12
Ti1	P 63/m m c	−7.767	−7.766	34.19	34.2	−0.001	−0.03
V1	I m −3 m	−9.086	−9.083	13.5	13.47	−0.003	0.22
W1	I m −3 m	−12.941	−12.948	16.19	16.17	0.007	0.12
Y1	P 63/m m c	−6.455	−6.456	65.58	65.61	0.001	−0.05
Zn1	P 63/m m c	−1.242	−1.267	29.42	30.71	0.025	−4.20
Binaries							
Al1 Br3	P 1 21/c 1	−3.278	−3.28	647.57	692.06	0.002	−6.43
Al1 Cl3	C 1 2/m 1	−3.827	−3.828	198.81	206	0.001	−3.49
Al1 F3	C m c m	−5.891	−5.891	307.5	307.42	0	0.03
Al1 H3	R −3 c H	−3.462	−3.46	66.11	64.18	−0.002	3.01
Al1 I3	P 1 21/c 1	−2.747	−2.748	817.93	857.46	0.001	−4.61
Al1 N1	P 63 m c	−7.446	−7.446	42.49	42.49	0	0.00
Al1 P1	F −4 3 m	−5.184	−5.188	41.58	41.58	0.004	0.00
Al2 La1	F d −3 m S	−4.646	−4.637	135.28	134.86	−0.009	0.31
Al2 O3	R −3 c H	−7.481	−7.481	262.59	262.58	0	0.00
Al2 S3	P 61	−5.028	−5.028	662.8	671.92	0	−1.36
Al4 C3	R −3 m H	−6.19	−6.189	81.57	81.57	−0.001	0.00
B1 Fe1	P b n m	−7.868	−7.869	63.27	63.58	0.001	−0.49
B1 Fe2	I −4 2 m	−8.086	−8.077	53.95	53.9	−0.009	0.09
B1 Mo1	I 41/a m d S	−9.325	−9.328	83.01	83.01	0.003	0.00
B1 Mo2	I 4/m c m	−9.869	−9.853	73.03	73.08	−0.016	−0.07
B1 N1	P 63 m c	−8.784	−8.787	39.72	42.34	0.003	−6.19
B1 P1	F −4 3 m	−6.445	−6.454	23.53	23.53	0.009	0.00
B1 Ti1	P n m a	−8.04	−8.04	84.96	85.03	0	−0.08
B2 O3	P 31	−8.023	−8.024	146.83	147.99	0.001	−0.78
B2 S3	P 1 21/c 1	−5.577	−5.579	938.84	1012.32	0.002	−7.26
B2 Ti1	P 6/m m m	−8.07	−8.085	25.75	25.7	0.015	0.19
Bi1 Cl3	P n 21 a	−3.284	−3.29	476.3	510.11	0.006	−6.63
Bi1 F3	F m −3 m	−4.567	−4.567	47.83	47.83	0	0.00
Bi1 I3	R −3 H	−2.595	−2.597	394.27	399.47	0.002	−1.30
Bi2 O3	P 1 21/c 1	−5.753	−5.753	341.23	341.13	0	0.03
Bi2 S3	P b n m	−4.378	−4.38	524.55	529.25	0.002	−0.89
Br1 Cu1	F −4 3 m	−2.937	−2.938	46.29	46.72	0.001	−0.92
Br1 Li1	F m −3 m	−3.315	−3.315	41.86	41.82	0	0.10
Br2 Fe1	P −3 m 1	−4.309	−4.309	86.07	84.92	0	1.35
Br2 Ti1	P −3 m 1	−4.737	−4.737	84.72	85.27	0	−0.65
Br3 La1	P 63/m	−4.482	−4.482	258.44	257.32	0	0.44
Br4 Ti1	P a −3	−4.068	−4.074	1613.71	1825.56	0.006	−11.60
C1 Li1	I m m m	−5.555	−5.555	47.75	47.22	0	1.12
C1 Mo2	P b c n	−10.484	−10.484	150.43	150.33	0	0.07
C1 Si1	P 63 m c	−7.525	−7.527	42	42	0.002	0.00
C1 Ti1	F m −3 m	−9.251	−9.263	20.42	20.41	0.012	0.05
Cl1 Cu1	F −4 3 m	−3.169	−3.169	39.98	39.83	0	0.38
Cl1 Li1	F m −3 m	−3.687	−3.688	34.24	34.24	0.001	0.00
Cl2 Cu1	C 1 2/m 1	−2.961	−2.964	77.53	79.53	0.003	−2.51
Cl2 Fe1	R −3 m H	−4.722	−4.722	69.13	70.29	0	−1.65
Cl3 Fe1	R −3 H	−4.163	−4.166	206.02	218.97	0.003	−5.91
Cl3 Ti1	P −3 1 m	−4.889	−4.89	208	215.33	0.001	−3.40
Cl5 Mo1	C 1 2/m 1	−4.135	−4.138	1024.75	1113.51	0.003	−7.97
Cu1 F2	P 1 21/n 1	−3.699	−3.7	70.72	71.2	0.001	−0.67
Cu1 I1	F −4 3 m	−2.779	−2.781	56.64	56.09	0.002	0.98
Cu1 O1	C 1 2/c 1	−4.267	−4.267	44.31	44.28	0	0.07
Cu2 O1	P n −3 m Z	−3.702	−3.701	79.97	78.9	−0.001	1.36
Cu2 S1	P 43 21 2	−3.956	−3.985	180.08	179.63	0.029	0.25
Cu3 P1	P 63 c m	−4.17	−4.172	305.7	304.9	0.002	0.26

(*continued on next page*)

Table A1 (*continued*)

Formula	Spacegroup	Energy (eV/atom) "Default" parameters	Energy (eV/atom) "Accurate" parameters	Cell volume (Å^3) "Default" parameters	Cell volume (Å^3) "Accurate" parameters	ΔE (eV/atom)	ΔV (%)
F1 Li1	F m −3 m	−4.853	−4.853	16.8	16.76	0	0.24
F2 Fe1	P 42/m n m	−5.697	−5.697	76.26	76.08	0	0.24
F3 La1	P 63 c m	−6.811	−6.811	335.29	334.96	0	0.10
F4 Ti1	P n m a	−6.315	−6.317	882.81	905.08	0.002	−2.46
Fe1 Si1	P 21 3	−7.381	−7.382	88.48	88.09	0.001	0.44
Fe1 Ti1	P m −3 m	−8.46	−8.458	25.8	25.61	−0.002	0.74
Fe2 O3	R −3 c R	−6.781	−6.781	105.19	105.51	0	−0.30
Fe2 P1	P −6 2 m	−7.809	−7.81	100.03	100.03	0.001	0.00
Fe2 Ti1	P 63/m m c	−8.408	−8.409	152.45	152.59	0.001	−0.09
Fe3 O4	P 1 2/m 1	−6.823	−6.823	313.61	312.85	0	0.24
Fe3 P1	I −4	−7.949	−7.95	179.33	178.88	0.001	0.25
Fe4 N1	P m −3 m	−8.352	−8.354	54.65	54.64	0.002	0.02
H1 Li1	F m −3 m	−3.048	−3.05	15.58	15.68	0.002	−0.64
H2 La1	F m −3 m	−4.525	−4.524	44.65	44.64	−0.001	0.02
I1 Li1	F m −3 m	−2.911	−2.911	54.2	54.03	0	0.31
I3 La1	C c m m	−3.921	−3.923	338.98	349.39	0.002	−2.98
I4 Si1	P a −3	−2.677	−2.679	2152.29	2203.12	0.002	−2.31
I4 Ti1	C 1 2/c 1	−3.486	−3.489	415.04	443.7	0.003	−6.46
La1 S1	F m −3 m	−6.703	−6.696	50.78	50.74	−0.007	0.08
La2 O3	I a −3	−8.404	−8.405	741.12	740.74	0.001	0.05
La2 S3	P n m a	−6.648	−6.648	504.65	503.52	0	0.22
Li1 O1	P 63/m m c	−4.851	−4.851	66.56	66.45	0	0.17
Li2 O1	F m −3 m	−4.771	−4.771	24.83	24.91	0	−0.32
Li2 S1	F m −3 m	−3.99	−3.99	46.66	46.64	0	0.04
Li3 N1	P 6/m m m	−3.898	−3.898	44.45	44.39	0	0.14
Mo1 O2	P 1 21/c 1	−7.508	−7.508	140.55	140.28	0	0.19
Mo1 O3	P b n m	−7.181	−7.183	222.41	225.26	0.002	−1.27
Mo1 S2	P 63/m m c	−7.265	−7.266	117.79	118.87	0.001	−0.91
Mo1 Si2	I 4/m m m	−7.766	−7.767	40.57	40.52	0.001	0.12
Mo2 S3	P 1 21/m 1	−7.545	−7.542	166.07	166.57	−0.003	−0.30
Mo3 Si1	P m −3 n	−9.922	−9.91	116.36	116.67	−0.012	−0.27
Mo5 Si3	I 4/m c m	−9.281	−9.28	229.36	229.06	−0.001	0.13
N1 Ti1	F m −3 m	−9.73	−9.725	19.26	19.27	−0.005	−0.05
N4 Si3	P 63	−8.182	−8.183	148.6	148.62	0.001	−0.01
O1 Ti1	A 1 1 2/m	−8.856	−8.857	108.59	108.73	0.001	−0.13
O2 Si1	P 32 2 1	−7.905	−7.905	122.29	122.33	0	−0.03
O2 Ti1	I 41/a m d S	−8.831	−8.831	70.94	70.89	0	0.07
O2 Ti1	P 42/m n m	−8.804	−8.804	64.55	64.53	0	0.03
O3 Ti2	R −3 c R	−8.919	−8.919	106.02	105.94	0	0.08
O5 P2	R 3 c H	−7.011	−7.015	428.33	457.02	0.004	−6.28
O5 Ti3	C 1 2/m 1	−8.892	−8.893	178.23	177.76	0.001	0.26
P1 Si1	C m c 21	−5.568	−5.569	517.08	552.46	0.001	−6.40
P2 S3	P 1 21/c 1	−4.802	−4.804	1139.62	1230.15	0.002	−7.36
P4 S3	P m n b	−5.005	−5.007	1707.66	1803.75	0.002	−5.33
S1 Ti1	P 63/m m c	−7.358	−7.357	59.43	59.41	−0.001	0.03
S2 Si1	I c m a	−5.235	−5.237	172.7	184	0.002	−6.14
S2 Ti1	C 1 2/m 1	−6.59	−6.589	86.06	85.98	−0.001	0.09
Si1 Ti1	P n m a	−7.349	−7.35	118.94	118.96	0.001	−0.02
Si2 Ti1	F d d d S	−6.748	−6.752	84.77	84.83	0.004	−0.07
Ternaries							
Al1 Cl1 O1	P m n m S	−5.864	−5.865	95.28	98.37	0.001	−3.14
Al1 F6 Li3	P n a 21	−5.299	−5.299	392.96	394.12	0	−0.29
Al1 H3 O3	P 1 21/n 1	−5.856	−5.857	439.45	432.91	0.001	1.51
Al1 La1 O3	P m −3 m	−8.008	−8.008	55.3	55.35	0	−0.09
Al1 Li1 O2	R −3 m H	−6.614	−6.614	33.06	33.06	0	0.00
Al1 O4 P1	R −3 H	−7.476	−7.477	640.65	652.86	0.001	−1.87
Al2 Cu1 O4	F d −3 m S	−6.518	−6.518	134.43	134.59	0	−0.12
Al2 Fe1 O4	F d −3 m S	−7.338	−7.339	139.75	139.71	0.001	0.03
Al2 O12 S3	R −3 H	−6.509	−6.51	410.15	414.18	0.001	−0.97
Al2 O5 Si1	P −1	−7.633	−7.633	303.08	303.25	0	−0.06
Al2 O5 Ti1	C m c m	−7.976	−7.976	167.49	167.52	0	−0.02
Al4 B2 O9	P b a m	−7.508	−7.509	162.99	163.42	0.001	−0.26
Al6 O13 Si2	P b a m	−7.126	−7.126	226.28	226.59	0	−0.14
B1 H3 O3	P −1	−6.085	−6.088	276	287.1	0.003	−3.87
B1 Li1 O2	P 1 21/c 1	−7.018	−7.02	151.99	156.92	0.002	−3.14
B3 Li1 O5	P n a 21	−7.616	−7.616	328.65	331.2	0	−0.77
B4 Li2 O7	I 41 c d	−7.446	−7.45	475.64	495.17	0.004	−3.94
Bi1 Cl1 O1	P 4/n m m Z	−4.821	−4.824	116.32	119.98	0.003	−3.05
Bi2 O12 S3	R −3 H	−6.052	−6.056	579.73	594.28	0.004	−2.45
Br1 H4 N1	P −4 3 m	−4.46	−4.459	68.42	67.96	−0.001	0.68
C1 Cu1 O3	C 1 m 1	−6.231	−6.233	58.11	59.24	0.002	−1.91
C1 Fe1 O3	R −3 c H	−7.31	−7.311	100.41	100.22	0.001	0.19
C1 Li2 O3	C 1 2/c 1	−6.56	−6.56	120.54	121.21	0	−0.55
C6 Mo1 O6	P n m a	−7.754	−7.758	913.27	995.99	0.004	−8.31

Table A1 (*continued*)

Formula	Spacegroup	Energy (eV/atom) "Default" parameters	Energy (eV/atom) "Accurate" parameters	Cell volume (Å^3) "Default" parameters	Cell volume (Å^3) "Accurate" parameters	ΔE (eV/atom)	ΔV (%)
Cl1 Fe1 O1	P m n m S	−5.469	−5.469	106.83	108.9	0	−1.90
Cl1 H3 N1	P a −3	−4.369	−4.375	617.63	632.47	0.006	−2.35
Cl1 La1 O1	P 4/n m m S	−7.082	−7.082	119.71	119.12	0	0.50
Cl1 Li1 O4	P n m a	−4.469	−4.47	295.17	300.37	0.001	−1.73
Cu1 Fe1 O2	R −3 m H	−5.691	−5.691	140.07	140.16	0	−0.06
Cu1 Fe1 S2	I −4 2 d	−5.047	−5.046	149.5	149.8	−0.001	−0.20
Cu1 Fe1 S2	I −4 2 d	−5.046	−5.046	148.85	149.67	0	−0.55
Cu1 Fe2 O4	I 41/a m d S	−6.06	−6.06	152.7	152.71	0	−0.01
Cu1 H2 O2	C m c 21	−4.634	−4.641	84.52	88.38	0.007	−4.37
Cu1 O4 S1	P n m a	−5.485	−5.485	279.52	281.87	0	−0.83
Cu5 Fe1 S4	F 2 3	−4.353	−4.354	161.68	163.24	0.001	−0.96
Cu5 Fe1 S4	F 2 3	−4.353	−4.354	161.68	163.15	0.001	−0.90
F1 H4 N1	P 63 m c	−4.898	−4.898	123.05	122.93	0	0.10
Fe1 H1 O2	C m c m	−6.094	−6.094	72.25	72.1	0	0.21
Fe1 Mo1 O4	C 1 2/m 1	−7.194	−7.195	345.1	344.68	0.001	0.12
Fe1 O3 Ti1	R −3 H	−8.072	−8.072	108.29	108.17	0	0.11
Fe1 O4 S1	C m c m	−6.458	−6.458	143.77	144.19	0	−0.29
Fe1 O4 W1	P 1 2/c 1	−7.725	−7.725	140.53	140.53	0	0.00
Fe2 O12 S3	R −3 H	−6.321	−6.322	449.73	453.76	0.001	−0.89
Fe2 O4 Si1	P b n m	−7.358	−7.358	318.21	318.06	0	0.05
Fe2 O4 Ti1	F d −3 m Z	−7.734	−7.734	322.64	322.14	0	0.16
H1 Li1 O1	P 4/n m m S	−4.975	−4.975	55.54	55.76	0	−0.39
H4 N2 O3	P c c n	−5.53	−5.534	625.17	650.91	0.004	−3.95
La1 O4 P1	P 1 21/n 1	−8.066	−8.066	314.1	315.39	0	−0.41
Li1 N1 O3	R −3 c R	−5.986	−5.986	101.06	100.04	0	1.02
Li1 O3 P1	P 1 21/n 1	−6.725	−6.727	742.6	747.51	0.002	−0.66
Li2 O3 Si1	C m c 21	−6.563	−6.563	121.1	121.49	0	−0.32
Li2 O3 Ti1	F m −3 m	−6.995	−6.995	158.9	159.31	0	−0.26
Li2 O4 S1	P 1 21/c 1	−5.887	−5.887	338.67	339.78	0	−0.33
Quaternaries							
Al1 Cl3 H12 O6	R −3 c H	−4.826	−4.827	488.22	492.29	0.001	−0.83
Cl1 H4 N1 O4	P n m a	−4.685	−4.689	398.31	415.49	0.004	−4.13
Cl2 Fe1 H4 O2	C 1 2/m 1	−4.749	−4.753	115.07	118.96	0.004	−3.27
Cl2 Fe1 H8 O4	P 1 21/c 1	−4.828	−4.83	342.72	347.88	0.002	−1.48
Cu1 H10 O9 S1	P −1	−5.109	−5.112	369.16	372.24	0.003	−0.83
Cu1 H2 O5 S1	P −1	−5.325	−5.33	181.55	186.43	0.005	−2.62
Cu1 H6 O7 S1	C 1 c 1	−5.169	−5.178	275.26	275.19	0.009	0.03
Fe1 H4 O6 P1	P b c a	−6.056	−6.057	890.14	901.24	0.001	−1.23
H8 N2 O4 S1	P n a 21	−5.365	−5.372	507.28	540.68	0.007	−6.18

and the X/O_2 ratio. We found that although Wang et al.'s O_2 reference (R + BECO) [57] gave very good performance for binary oxides, it failed to give results accurate to within 10 kJ/mol-atom for several polyanion-containing systems (Table 4). Similarly, the O_2 reference corrected for the GGA binding energy (BECO) reproduced several polyanion-containing systems quite well but failed to adequately describe oxides or silicates (Table 4). Finally, we found that using the R + BECO adjustment along with shifting the X energies could give accurate results across binary oxide and ternary polyanion-containing systems (Table 4).

Our results do not clearly indicate whether the majority of the error in GGA ternary polyanion-containing compound formation reactions comes from an inadequacy of describing the reduction of molecular O_2 to solid oxygen in a lattice with a constant O_2 adjustment, or from an error associated with oxidizing X from the elemental state (e.g., moving from elemental P to P^{5+} in a phosphate). We note that the difference in electronegativity between O and X correlates at least qualitatively with the magnitude of the fitted correction to X (Table 4). This may suggest that the error depends on the covalency of X–O bonds, and may explain why the R + BECO adjustment gives better results for more ionic bonding (oxides, silicates) where oxygen is strongly reduced, whereas removing the oxygen reduction component of the adjustment (BECO) gives better results for more covalently-bonded compounds (carbonates, sulfates). However, because we obtain fairly good accuracy by assuming the R + BECO adjustment for all systems and correcting the X elemental energy states (Table 4), we may also interpret the data as suggesting that the R + BECO adjustment is fairly universal for all gas to solid transitions, and the majority of the error lies not in O_2 but in oxidizing X from the elemental state. Given the uncertainty, perhaps the most cautious interpretation is to state that the pure GGA results tends to understabilize the simultaneous reduction of O_2 and oxidation of X, and that the degree of this understabilization depends on X and can be tuned by adjusting the O_2 energy and X elemental energy.

While we find that using the R + BECO adjustment and fitting constant shifts to the energies of elements X improves the accuracy of ternary polyanion-containing systems (Table 4), this additional accuracy comes at the expense of having several adjustable parameters in our model. These adjustable parameters have been fit only for the ternary polyanion-containing data set, and should not be interpreted to be universal for all reactions involving X. In particular, we find that while the elemental energy adjustments from Table 4 increase the accuracy for the polyanion materials in which X is oxidized, this correction reduces the accuracy of GGA energies for A_iX_j binary systems such as phosphides, silicides, carbides, and sulfides (Table 5). For example, computing carbide formation using the elemental C correction derived for carbonates increases the error by greater than a factor of 3.

Our data suggests that oxidation and reduction of an elemental species may incur opposite errors in GGA. When artificially destabilizing the O_2 molecule by its reduction correction without adjusting elemental X states, we tend to overstabilize the formation energy of ternary polyanion systems (Fig. 6). To compensate for

Table A2
Experimental [101] and calculated (GGA) results for 57 ternary compounds, 6 binary oxides, and 15 other binary compounds (silicides, phosphides, carbides, and sulfides).

System	Enthalpy/f.u (Kubaschewski et al.)	Enthalpy/f.u. (pure GGA)
Silicates		
Al2SiO5	2589.5	2281.2415
Ba2SiO4	2272.3	2011.0852
BaSiO3	1618	1436.8389
Ca2SiO4	2328.4	2090.3832
Ca3Si2O7	3942.6	3557.2451
Ca3SiO5	2928.8	2638.7545
CaSiO3	1635.1	1469.8907
K2Si2O5	2508.7	2202.4745
K2Si4O9	4315.8	3860.2417
Li2SiO3	1648.5	1480.4849
Mg2SiO4	2176.9	1927.1785
MgSiO3	1548.5	1369.3769
Na2Si2O5	2473.6	2216.9425
Na2SiO3	1563.1	1387.8741
Na4SiO4	2101.2	1853.8622
Sr2SiO4	2302.9	2054.2892
SrSiO3	1633.4	1467.3359
Borates		
Ca2B2O5	2722.9	2487.1965
Ca3B2O6	3424.6	3102.5118
CaB2O4	2027.1	1854.5212
CaB4O7	3340.9	3049.4231
CsBO2	976.8	879.0976
K2B4O7	3326.3	3029.5551
KBO2	995	894.3966
Li2B4O7	3374.4	3069.2111
LiBO2	1019.2	929.3546
Na2B4O7	3284.9	3003.6231
Na2B8O13	5902.8	5395.9829
NaBO2	975.7	887.0226
RbBO2	974.9	880.5256
SrB4O7	3332.6	3041.5307
Phosphates		
AlPO4	1733	1547.8092
Ca2P2O7	3336.7	2957.8971
Ca3(PO4)2	4117.1	3651.7584
LiPO3	1254.8	1112.5679
Mg3(PO4)2	3780.7	3344.5344
Na3PO4	1916.9	1714.7262
Na4P2O7	3166.5	2856.5327
NaPO3	1220.1	1088.7029
Carbonates		
CaCO3	1206.9	1101.1919
Cs2CO3	1136.4	998.9219
K2CO3	1153.1	1032.0919
Li2CO3	1215.5	1110.7819
MgCO3	1095.8	989.4819
Na2CO3	1129.7	1022.7019
Rb2CO3	1133	1004.9419
SrCO3	1220.1	1107.3619
Sulfates		
Al2(SO4)3	3441.3	3053.9646
BaSO4	1481.1	1308.1272
CaSO4	1434.1	1286.7482
Cs2SO4	1444.3	1276.8002
K2SO4	1438.5	1282.8582
Li2SO4	1437.2	1306.7682
MgSO4	1284.9	1138.8792
Na2SO4	1389.5	1249.4172
Rb2SO4	1437.2	1268.5762
SrSO4	1453.1	1295.3002
Binary oxides		
Al2 O3	1675.7	1458.5
Ca O	634.9	573.61
K2 O	363.2	299.61
Li2 O	597.9	536.33
Mg O	601.6	530.23
Na2 O	415.1	354.06
Binary compounds		
Mg2Si	79.1	47.04
AlP	164.4	118.01
Al4C3	209.2	60.06
BaC2	74.1	27.30
CaC2	59.4	6.70
SrC2	84.5	8.34
Al2S3	723.4	508.86
BaS	463.6	401.92
CaS	473.2	416.39
K2S	376.6	309.72
Li2S	446.9	388.07
MgS	345.6	277.62
Na2S	366.1	318.32
Rb2S	361.1	287.88
SrS	452.7	415.40

this overstabilization, we need to account for the oxidation of X, which requires energy adjustments to X in the opposite direction from O_2. The magnitude and sign of GGA errors in formation enthalpies from element reference states may thus depend on the oxidation state and electronegativity of the elements in the compound. With this interpretation, it is not surprising that our X adjustments fitted for ternary polyanion systems (where X is oxidized) penalize the accuracy of A_iX_j formation reactions (where X is reduced). The A_iX_j formation reactions are understabilized in pure GGA, and an artificial destabilization to X is needed to accurately model these reactions rather than the artificial stabilization of X needed for $A_iX_jO_k$ compounds.

Given that fitted element corrections can vary greatly by chemical system in GGA, it is unlikely that such corrections will apply broadly across several chemical spaces. Instead, we believe our results underscore the need to better understand the particular GGA errors in different chemical systems, especially when computing reaction energies from the elements. For analyses from our database, for example, we generally employ the R + BECO adjustment but keep in mind that polyanion-containing formation reactions from an O_2 reference state may need an additional correction to X reference states to accurately model polyanion-containing compounds. In addition, our results suggest a great need for improved exchange–correlation functionals for high-throughput studies.

7. Conclusion

High-throughput density functional theory presents new opportunities for materials design and rapid computational screening, but also poses unique technical challenges regarding implementation and accuracy across wide chemical systems. In this paper, we demonstrated how data flow was managed in our high-throughput project and described the computational tools we used to scale DFT calculations to a large scale. In addition, we presented data regarding convergence for a large number of compounds, demonstrating that fairly good energy and cell volume convergence for compounds could be obtained with relatively small k-point meshes and relatively loose convergence parameters. A major challenge for high-throughput computations is that GGA errors are very dependent on the nature of the chemistry; the formation enthalpy errors of compounds depend very much on the extent to which elements undergo oxidation or reduction making a single element energy correction of limited value. This was clar-

ified by demonstrating that formation energies of ternary polyanion-containing compounds depend on the particular O_2 reference employed, and that no O_2 reference gives satisfactory results over all chemical spaces. While satisfactory formation enthalpies for ternary polyanion-containing compounds can be obtained by using Wang et al.'s O_2 reference [57] along with several element energy adjustments, it appears that such adjustments are not universal. The development and testing of more broadly applicable functionals would be a significant advance for the high-throughput application of *ab initio* methods to materials design.

The authors are currently developing a web interface to make available to the materials community the database of electronic structure calculations performed using the techniques described in this paper [8]. We hope this effort will contribute to the growth of high-throughput density functional theory for the design and understanding of new materials.

Acknowledgements

This research was supported by the US Department of Energy through grants Nos. #DE-FG02-96ER4557 and DE-FG02-97ER25308. Additional funding was provided by Umicore and Bosch. The authors would like to acknowledge discussions and contributions to the high-throughput infrastructure and methodology from Dr. Fei Zhou. In addition, we thank Dr. Shirley Meng for discussions related to high-throughput Li ion battery design, Dr. Michael Kocher for conversations on methods of interaction between the codebase and database, and Dr. Maria Chan and Dr. Denis Kramer for assistance in calibrating *U* parameters.

Appendix A. Calculated data sets

See Tables A1 and A2.

References

[1] P. Hohenberg, Physical Review 136 (1964) B864–B871.
[2] J. Hafner, C. Wolverton, G. Ceder, MRS Bulletin 31 (2006) 659–668.
[3] K. Kang, Y.S. Meng, J. Bréger, C.P. Grey, G. Ceder, Science 311 (2006) 977–980.
[4] J. Wang, J.B. Neaton, H. Zheng, V. Nagarajan, S.B. Ogale, B. Liu, et al., Science 299 (2003) 1719–1722.
[5] M.L. Cohen, Solid State Communications 107 (1998) 589–596.
[6] A. Kolmogorov, M. Calandra, S. Curtarolo, Physical Review B 78 (2008) 094520.
[7] G.K.H. Madsen, Journal of the American Chemical Society 128 (2006) 12140–12146.
[8] Materials Genome <www.materialsgenome.org>.
[9] G. Bergerhoff, R. Hundt, R. Sievers, I. Brown, Journal of Chemical Information and Computer Sciences 23 (1983) 66–69.
[10] F. Karlsruhe, Inorganic Crystal Structure Database, <http://icsd.fiz-karlsruhe.de/icsd/>.
[11] G. Hautier, C.C. Fischer, A. Jain, T. Mueller, G. Ceder, Chemistry of Materials 22 (2010) 3762–3767.
[12] A. Jain, S.-A. Seyed-Reihani, C.C. Fischer, D.J. Couling, G. Ceder, W.H. Green, Chemical Engineering Science 65 (2010) 3025–3033.
[13] M. Chan, G. Ceder, Physical Review Letters 105 (2010) 196403.
[14] S.P. Ong, A. Jain, G. Hautier, B. Kang, G. Ceder, Electrochemistry Communications 12 (2010) 427–430.
[15] J.C. Kim, C.J. Moore, B. Kang, G. Hautier, A. Jain, G. Ceder, Journal of the Electrochemical Society 158 (2011) A309.
[16] S. Curtarolo, D. Morgan, G. Ceder, Accuracy of ab initio methods in predicting the crystal structures of metals: review of 80 binary alloys, (2008).
[17] W. Setyawan, S. Curtarolo, Computational Materials Science 49 (2010) 299–312.
[18] C. Ortiz, O. Eriksson, M. Klintenberg, Computational Materials Science 44 (2009) 1042–1049.
[19] J. Greeley, M. Mavrikakis, Nature Materials 3 (2004) 810–815.
[20] J. Greeley, T.F. Jaramillo, J. Bonde, I.B. Chorkendorff, J.K. Nørskov, Nature Materials 5 (2006) 909–913.
[21] J. Greeley, Surface Science 601 (2007) 1590–1598.
[22] M. Andersson, T. Bligaard, A. Kustov, K. Larsen, J. Greeley, T. Johannessen, et al., Journal of Catalysis 239 (2006) 501–506.
[23] J.S. Hummelshøj, D.D. Landis, J. Voss, T. Jiang, A. Tekin, N. Bork, et al., The Journal of Chemical Physics 131 (2009) 014101.
[24] J. Greeley, J.K. Nørskov, The Journal of Physical Chemistry C 113 (2009) 4932–4939.
[25] Japan Science and Technology Agency, Computational Electronic Structure Database (CompES), <http://caldb.nims.go.jp/>.
[26] N. Tatara, Y. Chen, Progress of Theoretical Physics Supplement 138 (2000) 755–756.
[27] H.L. Skriver, The hls Alloy Database, <http://databases.fysik.dtu.dk/hlsDB/hlsDB.php>.
[28] M. Klintenberg, Electronic Structure Project, <http://gurka.fysik.uu.se/ESP/>.
[29] S. Curtarolo, AFLOW-lib databases, <http://aflowlib.org>.
[30] T.R. Munter, D.D. Landis, F. Abild-Pedersen, G. Jones, S. Wang, T. Bligaard, Computational Science & Discovery 2 (2009) 015006.
[31] G. Ceder, D. Morgan, C. Fischer, K. Tibbetts, S. Curtarolo, MRS Bulletin 31 (2006) 981–985.
[32] C.C. Fischer, K.J. Tibbetts, D. Morgan, G. Ceder, Nature Materials 5 (2006) 641–646.
[33] Sun Grid Engine, <http://gridengine.sunsource.net>.
[34] The Perl Programming Language, <http://www.perl.org>.
[35] M. Stonebraker, L.A. Rowe, ACM SIGMOD Record 15 (1986) 340–355.
[36] M. Stonebraker, G. Kemnitz, Communications of the ACM 34 (1991) 78–92.
[37] K. Mitra, International Materials Reviews 53 (2008) 275–297.
[38] D. Scott, S. Manos, P. Coveney, Journal of Chemical Information and Modeling 48 (2008) 262–273.
[39] G.A. Gazonas, D.S. Weile, R. Wildman, A. Mohan, International Journal of Solids and Structures 43 (2006) 5851–5866.
[40] R. Giro, M. Cyrillo, D. Galvão, Chemical Physics Letters 366 (2002) 170–175.
[41] G. Johannesson, T. Bligaard, A. Ruban, H. Skriver, J.K. Norskov, Physical Review Letters 1 (2002) 255506.
[42] S. Woodley, Applications of Evolutionary Computation in Chemistry 110 (2004) 95–132.
[43] D. Deaven, K. Ho, Physical Review Letters 75 (1995) 288–291.
[44] A.R. Oganov, C.W. Glass, The Journal of Chemical Physics 124 (2006) 244704.
[45] N. Chakraborti, International Materials Reviews 49 (2004).
[46] A. Kolmogorov, S. Shah, E. Margine, A. Bialon, T. Hammerschmidt, R. Drautz, Physical Review Letters 105 (2010) 217003.
[47] T. Bligaard, M. Andersson, K. Jacobsen, H. Skriver, C.H. Christensen, J.K. Norskov, MRS Bulletin 31 (2006) 986–990.
[48] P. Villars, Journal of Alloys and Compounds 279 (1998) 1–7.
[49] G. Kresse, J. Furthmüller, Computational Materials Science 6 (1996) 15–50.
[50] W. Kohn, L.J. Sham, Physical Review 140 (1965) 1133–1138.
[51] D. Langreth, J. Perdew, Physical Review B 21 (1980) 5469–5493.
[52] J.P. Perdew, K. Burke, M. Ernzerhof, Physical Review Letters (1996) 3865–3868.
[53] J. Heyd, J.E. Peralta, G.E. Scuseria, R.L. Martin, The Journal of Chemical Physics 123 (2005) 174101.
[54] D. Rappoport, N.R.M. Crawford, F. Furche, K. Burke, C, Which functional should I choose?, in: E.I. Solomon, R.A. Scott, R.B. King (Eds.), Computational Inorganic and Bioinorganic Chemistry, Wiley-Blackwell, 2009.
[55] G. Csonka, J. Perdew, A. Ruzsinszky, P. Philipsen, S. Lebègue, J. Paier, et al., Physical Review B 79 (2009) 155107.
[56] F. Zhou, M. Cococcioni, C.A. Marianetti, D. Morgan, G. Ceder, Physical Review B 70 (2004) 235121.
[57] L. Wang, T. Maxisch, G. Ceder, Physical Review B 73 (2006) 195107.
[58] V.I. Anisimov, J. Zannen, O.K. Andersen, Physical Review B 44 (1991) 943–954.
[59] S.L. Dudarev, S.Y. Savrasov, C.J. Humphreys, A.P. Sutton, Physical Review B 57 (1998) 1505–1509.
[60] H. Kulik, M. Cococcioni, D. Scherlis, N. Marzari, Physical Review Letters 97 (2006) 103001.
[61] K. Persson, A. Bengtson, G. Ceder, D. Morgan, Geophysical Research Letters 33 (2006).
[62] K. Persson, G. Ceder, D. Morgan, Physical Review B 73 (2006) 115201.
[63] F. Zhou, M. Cococcioni, K. Kang, G. Ceder, Electrochemistry Communications 6 (2004) 1144–1148.
[64] F. Zhou, K. Kang, T. Maxisch, G. Ceder, D. Morgan, Solid State Communications 132 (2004) 181–186.
[65] R.E. Doe, K.A. Persson, G. Hautier, G. Ceder, Electrochemical and Solid-State Letters 12 (2009) A125.
[66] R.E. Doe, K.A. Persson, Y.S. Meng, G. Ceder, Chemistry of Materials 20 (2008) 5274–5283.
[67] A.D. Becke, Journal of Chemical Physics 98 (1993) 1372.
[68] J. Perdew, M. Ernzerhof, K. Burke, The Journal of Chemical Physics 105 (1996) 9982.
[69] J. Heyd, G.E. Scuseria, M. Ernzerhof, The Journal of Chemical Physics 118 (2003) 8207.
[70] O.A. Vydrov, J. Heyd, A.V. Krukau, G.E. Scuseria, The Journal of Chemical Physics 125 (2006) 074106.
[71] V.L. Chevrier, S.P. Ong, R. Armiento, M.K.Y. Chan, G. Ceder, Physical Review B 82 (2010) 075122.
[72] S.P. Ong, V. Chevrier, G. Ceder, Physical Review B 83 (2011) 075112.
[73] T. Starkloff, J. Joannopoulos, Physical Review B 16 (1977) 5212–5215.
[74] M.L. Cohen, V. Heine, Solid State Physics 24 (1970) 37–248.
[75] J.C. Phillips, Physical Review 112 (1958) 685–695.
[76] G. Kresse, D. Joubert, Physical Review B 59 (1999) 1758–1775.
[77] P.E. Blochl, Physical Review B 50 (1994) 953–979.
[78] G. Kresse, M. Marsman, J. Furthmüller, VASP the GUIDE (2010).
[79] H.J. Monkhorst, J.D. Pack, Physical Review B 13 (1976) 5188–5192.

[80] L.A. Montoro, M. Abbate, E.C. Almeida, J.M. Rosolen, Chemical Physics Letters 309 (1999) 14–18.
[81] L.E. Orgel, An Introduction to Transition-Metal Chemistry: Ligand-Field Theory, Methuen, London, 1960.
[82] E.R. Davidson, Methods in Computational Molecular Physics, Plenum, New York, 1983.
[83] D. Wood, A. Zunger, Journal of Physics A: Mathematical and General 18 (1985) 1343–1359.
[84] P. Pulay, Chemical Physics Letters 73 (1980) 393–398.
[85] G. Kerker, Physical Review B 23 (1981) 3082–3084.
[86] G. Kresse, J. Furthmüller, Physical Review B 54 (1996) 11169–11186.
[87] P.E. Blöchl, O. Jepsen, O. Andersen, Physical Review B 49 (1994) 16223–16233.
[88] MySQL, <www.mysql.com>.
[89] Sybase, <www.sybase.com>.
[90] Oracle, <www.oracle.com>.
[91] T. Haerder, A. Reuter, ACM Computing Surveys 15 (1983) 287–317.
[92] E.F. Codd, Communications of the ACM 13 (1970) 377–387.
[93] P.-S.C. Peter, ACM Transactions on Database Systems 1 (1976) 9–36.
[94] H. Burzlaff, Y. Malinovsky, Acta Crystallographica Section A: Foundations of Crystallography 53 (1997) 217–224.
[95] R. Hundt, J.C. Schön, M. Jansen, Journal of Applied Crystallography 39 (2006) 6–16.
[96] N. Brese, M. O'keeffe, Acta Crystallographica Section B: Structural Science 47 (1991) 192–197.
[97] B. Kang, G. Ceder, Nature 458 (2009) 190–193.
[98] S. Ong, L. Wang, B. Kang, G. Ceder, Chemistry of Materials 20 (2008) 1798–1807.
[99] Y. Wang, S. Curtarolo, C. Jiang, R. Arroyave, T. Wang, G. Ceder, et al., Calphad 28 (2004) 79–90.
[100] S. Lany, Physical Review B 78 (2008) 245207.
[101] O. Kubaschewski, C. Alcock, P. Spencer, Materials Thermochemistry, sixth ed., Pergamom Press, Oxford, 1993.
[102] C. Franchini, R. Podloucky, J. Paier, M. Marsman, G. Kresse, Physical Review B 75 (2007) 195128.
[103] B. Hammer, L. Hansen, J. Nørskov, Physical Review B 59 (1999) 7413–7421.