Quantifying uncertainty in high-throughput density functional theory: a comparison of AFLOW, Materials Project, and OQMD

Vinay I. Hegde,1,* Christopher K. H. Borg,1,* Zachary del Rosario,1,2 Yoolhee Kim,1 Maxwell Hutchinson,1 Erin Antono,1 Julia Ling,1 Paul Saxe,3 James E. Saal,1 and Bryce Meredig1,†

1 Citrine Informatics, 2629 Broadway, Redwood City, CA 94063
2 Olin College of Engineering, 1000 Olin Way, Needham, MA 02492
3 Molecular Sciences Software Institute, Virginia Tech, Blacksburg, VA 24061
(Dated: November 8, 2022)

A central challenge in high throughput density functional theory (HT-DFT) calculations is selecting a combination of input parameters and post-processing techniques that can be used across all materials classes, while also managing accuracy-cost tradeoffs. To investigate the effects of these parameter choices, we consolidate three large HT-DFT databases: Automatic-FLOW (AFLOW), the Materials Project (MP), and the Open Quantum Materials Database (OQMD), and compare reported properties across each pair of databases for materials calculated using the same initial crystal structure. We find that HT-DFT formation energies and volumes are generally more reproducible than band gaps and total magnetizations; for instance, a notable fraction of records disagree on whether a material is metallic (up to 7%) or magnetic (up to 15%). The variance between calculated properties is as high as 0.105 eV/atom (median relative absolute difference, or MRAD, of 6%) for formation energy, 0.65 Å3/atom (MRAD of 4%) for volume, 0.21 eV (MRAD of 9%) for band gap, and 0.15 μB/formula unit (MRAD of 8%) for total magnetization, comparable to the differences between DFT and experiment. We trace some of the larger discrepancies to choices involving pseudopotentials, the DFT+U formalism, and elemental reference states, and argue that further standardization of HT-DFT would be beneficial to reproducibility.

Keywords: high-throughput DFT, uncertainty quantification, reproducibility, materials databases

I. INTRODUCTION

Over the past decade, high-throughput (HT) density functional theory (DFT) has emerged as a widely-used tool for materials discovery and design [1–3]. In a standard HT-DFT workflow, software tools automate the process of calculating materials properties of interest within DFT, including submitting jobs to high-performance computing infrastructure, on-the-fly error handling, post-processing and dissemination of results, and so on, enabling researchers to evaluate typically 103–106 materials with minimal human intervention. The resulting database can then be screened for candidate materials exhibiting promising combinations of calculated properties or to search for trends amongst materials behavior to gain new chemical insights or develop surrogate models.

The increasingly widespread usage of HT-DFT in materials research can be attributed to a combination of three key factors. First, a large number of specialized codes implement fully automated calculations of specific materials properties within DFT, ranging from phonon dispersions to dielectric tensors. For example, VASP 5.1 [4, 5] introduced a feature enabling users to calculate elastic tensors by simply setting a parameter in the input file. Second, the ongoing growth of computing power has ensured that HT-DFT is now well within reach of a single university research group. Third, sophisticated, free, often open-source, software is readily available for managing large numbers of DFT calculations, post-processing output, and storing the resulting data systematically in databases. Thus, a number of HT-DFT databases with various focus areas have emerged; a list of exemplars, including any supporting workflow automation software, is given in Section S-I of the Supplemental Information (SI).

However, the entirely-automated nature of HT-DFT introduces a few key challenges. First, by definition, the volume of data from HT-DFT is too high for each individual calculation to undergo manual review or analysis [1]. How, then, are the quality and integrity of calculations monitored in high-throughput? Second, HT-DFT requires choosing, often at the outset, settings that are consistent across all calculations, encompassing all materials classes and properties being calculated. For example, it may not be known a priori whether the material being calculated is a metal or an insulator. As a result, the calculation parameters that affect, e.g., how electronic occupancies are smeared near the Fermi level must be chosen so that they are applicable to both metals and insulators. Third, practical HT-DFT calculations involve balancing accuracy and computational cost; best-practice recommendations [6] involve steps such as explicit convergence tests, which become computationally infeasible in the HT context. Of these challenges, only the first, related to monitoring the quality and integrity of calculations in high-throughput has been addressed. Software frameworks, such as Custodian [7], qmpy [8], and AiiDa [9], can store provenance information to en-
sure the integrity of calculations, and gracefully handle errors associated with catastrophic failures, e.g., those related to file read/write operations or memory issues during runtime, insufficient walltimes on high-performance computing resources, and misconfiguration of the underlying numerical libraries.

Since HT-DFT has become increasingly central to materials informatics efforts across the spectrum, from high-throughput screening to machine learning [10, 11] it is crucial to resolve the following concerns: (a) There is no one “correct” solution to some of the challenges of HT-DFT mentioned above, and different databases have tackled them slightly differently. How sensitive are the calculated materials properties to the different HT-DFT parameter choices? (b) The focus areas of many prominent HT-DFT databases in terms of the materials and properties calculated are often quite different. As a result, materials data from the various HT-DFT databases are often mixed with one another for thermochemical or other analysis. How interoperable are these various calculated materials properties across HT-DFT databases? We emphasize that such a comparison across HT-DFT databases is different from analyzing the reproducibility of DFT across software implementations and potentials, e.g., focusing on equations of state of elemental crystals: [12] the challenges of HT-DFT lie in choosing parameters that are applicable across a wide variety of materials and properties, targeting both reasonable accuracy and computational cost—very distinct from performing highly-accurate DFT calculations of a small set of materials.

Here, we analyze the reproducibility and interoperability of HT-DFT calculations. We critically compare the agreement between three databases for four properties: formation energy ($\Delta E_f$), volume ($V$), band gap ($E_g$), and total magnetization ($M$). We find certain properties (formation energies and volumes) to be more consistent across databases than others (band gap and magnetization). We then quantify the variability in each of the properties across databases and find that the typical differences between two HT-DFT databases are similar to those between DFT and experiment. Finally, we compare properties across different materials classes to identify characteristics of materials and/or properties that are harder than others to reproduce. In all cases, we identify trends, surface outliers, and investigate potential causes for an observed systematic differences between the databases.

II. METHODS

We focus on three prominent HT-DFT databases in this work: Automatic FLOW (AFLOW) [13], the Materials Project (MP) [14], and the Open Quantum Materials Database (OQMD) [3, 8]. All three databases contain calculations of a large number of mostly-experimentally reported, ordered compounds from the Inorganic Crystal Structure Database (ICSD) [15]. In addition, they contain calculations of many thousands of hypothetical compounds generated from common structural prototypes or other informatics approaches. As noted earlier, there are many other large HT-DFT databases, e.g., JARVIS-DFT [16], Materials Cloud [17], and others listed in Table S-1 of the SI. Here, we limit our focus to AFL OW, Materials Project, and OQMD as the latter (a) are among the longest-running, mature, widely-used, and general-purpose, and (b) use the VASP software package [4, 5] and projector augmented wave (PAW) potentials [18, 19] with the Perdew-Burke-Ernzerhof (PBE) parameterization [20] of a generalized-gradient approximation (GGA) to the DFT exchange-correlation functional. The variance in HT-DFT-calculated properties studied in the present work is, therefore, almost entirely due to differences in various choices involved in HT-DFT (e.g., those involving calculation parameters such as $k$-point density, the DFT+$U$ approach, post-calculation processing techniques, different versions of VASP and any associated software bugs, different versions of PBE pseudopotentials used) and not due to different implementations of DFT or approximations to the underlying exchange-correlation functional itself.

AFLOW has standardized band structure calculations [21, 22], binary alloy cluster expansions [23], finite-temperature thermodynamic properties [24], elastic and thermomechanical properties [25] calculated for many materials, and has an application programming interface (API) based on the REpresentational State Transfer (REST) standard (commonly referred to as “RESTful API”) for accessing data [26]. The Materials Project includes a variety of properties calculated for specific subsets of materials in the database, including elastic [27], thermoelectric [28], piezoelectric [29], dielectric [30], vibrational [31] properties, and X-ray adsorption spectra [32]. It also includes a collection of apps such as a Pourbaix diagram calculator [33], and the underlying data are accessible via a RESTful API [34]. Finally, the Open Quantum Materials Database (OQMD) contains calculations of a large number of hypothetical compounds based on structural prototypes, [35-37] and provides tools for the construction of DFT ground state phase diagrams at ambient and high-pressures [38-40]. The OQMD provides the entirety of the underlying database to download all at once, and a RESTful API for programmatic access [41]. License and access information for the three databases is included in Section S-II of the SI.

We query all three databases (AFLOW: queried June 2021; MP: v2019.05; OQMD: v1.2) for the calculated properties of materials whose crystal structures were sourced from the ICSD and aggregate them into a single dataset, after converting records from all sources into a unified, consistent data format, the Physical Information File (PIF) [42]. We then generate a set of comparable records for each pairwise combination of the databases—all calculations using the same initial crystal structure,
by matching their ICSD Collection Codes (hereafter referred to as "ICSD ID"). In instances where more than one calculation within a single database was labeled with the same ICSD ID, we use the lowest energy calculation for all analysis. In addition, we discard records with obviously unphysical property values (those with formation energy outside the [−5 eV/atom, +5 eV/atom] window and volumes above 150 Å³/atom), and normalize properties to the same units, where required. We then perform statistical analysis on the final curated set of comparable records across the three databases. Definitions of the metrics used in our analysis are given in Appendix A and details of the query and curation steps are provided in Section S-II of the SI.

III. RESULTS

The aggregation and processing of the data from the three HT-DFT databases results in a set of ∼70,000 total comparable DFT calculations. For each property of interest, i.e., formation energy per atom, volume per atom, band gap, total magnetization per formula unit (f.u.), the counts of records, and overlapping records for each pair of databases are shown in Table I. Approximately 15,000–25,000 comparisons can be made for each property and database pair, except for comparisons to formation energies from AFLOW, where only ∼2,200 records are reported. As mentioned earlier, overlapping records across databases were determined by using exact ICSD ID matches for the reported calculations.

A. Overall pairwise comparison statistics

Table II shows some overall statistics for comparisons of all properties across comparable records in the three databases: the median absolute difference (MAD), the interquartile range (IQR), the Pearson correlation coefficient (r), and Spearman’s rank correlation coefficient (ρ) (definitions of the metrics are in Appendix A). For band gap and total magnetization, the statistics were calculated only on subsets of overlapping records where both databases agreed that a material is non-metallic ($E_g > 0.01$ eV) and is magnetic ($M > 0.01$ $\mu_B$/atom), respectively. The latter threshold on the per-formula unit total magnetization ensures that undesired comparisons of different magnetic configurations for the same crystal structure (i.e., ferromagnetic configuration in one database being compared to antiferromagnetic configuration in another) are avoided as much as possible.

Overall, we find that: (a) The MAD in formation energy across pairs of databases can be up to 0.105 eV/atom, comparable to the ∼0.1 eV/atom difference between DFT and experimental formation energies [8]. (b) The MAD in volume across pairs of databases can be up to 0.65 Å³/atom (median absolute difference relative to mean (MRAD), of 3.8%), comparable to error between DFT and experiment [43]. (c) The MAD in band gap across pairs of databases can be up to 0.21 eV, even when comparing only records where both databases agree that a material is not metallic. For around 5%–7% of overlapping records, databases disagree whether a material is metallic. (d) The comparison of total magnetization shows high variability across database pairs. While the dispersion of differences for the MP-OQMD comparison is very small (MAD of 0.01 $\mu_B$/f.u. and IQR of 0.05 $\mu_B$/f.u.), the dispersion of differences in comparisons with AFLOW are rather large (up to MAD of 0.15 $\mu_B$/f.u. and IQR of up to 2.0 $\mu_B$/f.u.). In all cases, the correlation between calculated values is lower than for the other three properties, with both Pearson and Spearman correlation coefficients ranging from 0.6–0.8.

We further note that the latter poor correlation exists even after excluding overlapping records where the two databases disagree on whether the material is magnetic (10%–15% of the records).

B. Distribution of differences in calculated properties

We first analyze the raw differences in the calculated properties for records overlapping across pairs of databases. Figure 1 shows the distribution of the differences in calculated values for each of formation energy, volume, band gap, and total magnetization, for each pairwise combination of databases.

Formation energy: The distribution of differences in calculated formation energy across AFLOW-MP and MP-OQMD is surprisingly bimodal, with peaks around 0 and ±0.2 eV/atom. We find that the peak near 0.2 eV/atom in both pairwise comparisons corresponds mostly to oxides (see Figure S1), and is a result of different approaches in the two databases toward correcting DFT-calculated formation energies (see Section IV B). While the median difference ($\Delta x$ in Figure 1) are reasonably small across all three pairwise comparisons (up to ∼0.074 eV/atom), the difference distributions for AFLOW-MP and MP-OQMD are rather wide. The median absolute difference (MAD) and the interquartile range (IQR), both robust measures of the spread of a distribution, are up to ∼0.105 eV/atom and ∼0.173 eV/atom, respectively.

Volume: The distribution of differences in calculated volumes is skewed towards smaller volumes in the OQMD, but such a skew is absent in the AFLOW-MP comparison. Correspondingly, the median difference between AFLOW and MP volumes are ∼0.01 Å³/atom, whereas the median differences are ∼0.62 Å³/atom and ∼0.47 Å³/atom for AFLOW-OQMD and MP-OQMD, respectively. The consistently smaller volumes calculated in the OQMD can be understood to result from the choice of the plane wave energy cutoff used for DFT relaxation calculations. The OQMD chooses a plane wave cutoff that is lower than that used in AFLOW and MP ($ENMAX$ in the POTCAR file, up to 400 eV in OQMD, as opposed
to 520 eV in MP and up to 560 eV in AFLOW) for full cell relaxations. The lower plane wave cutoff results in Pulay stresses and generally smaller volumes than fully relaxed calculations. The MAD in volumes for comparisons, especially for OQMD with the other two databases, is up to ~0.65 Å³/atom. In addition, some differences in reported volumes can result from the different relaxation schemes employed in the three HT-DFT databases: AFLOW and MP perform two sequential relaxations, while the OQMD performs sequential relaxations until the volume change during a relaxation is less than 5%.

**Band gap:** The distribution of differences in the calculated band gaps is slightly skewed towards larger band gaps in the OQMD, but this skew is absent in the AFLOW-MP comparison. Correspondingly, the median difference in band gaps between AFLOW and MP is ~0.01 eV, and up to ~0.14 eV for comparisons with OQMD. The larger band gaps calculated in the OQMD might be due to smaller volumes from the choice of lower plane wave energy cutoffs. An increase in the fundamental band gap due to compressive strains (in the OQMD, due to unresolved Pulay stresses) has been observed in many semiconductor families [44–46]. In addition, the spread in the differences in calculated band gaps is quite large: with an MAD of up to ~0.21 eV and an IQR of up to ~0.36 eV for comparisons with OQMD. The spread may be, in addition to the choice of energy cutoff as discussed above, due to the different ways in which the databases calculate the band gap. For example, OQMD calculates band gap from the electronic density of states (DOS), in contrast to AFLOW and MP which calculate it from band dispersions. The energy grid used for the calculation of DOS and/or k-point meshes used for band structure calculations can also have a notable effect on the precision and accuracy of the reported band gap.

**Total magnetization:** The median differences in AFLOW-MP and MP-OQMD are nearly zero, with reasonably small MAD values as well. However, the differences between the magnetization reported in AFLOW and the other two databases skew towards larger values in AFLOW, with long tails and correspondingly large dispersions. The difference between AFLOW and OQMD, in particular, shows an MAD of ~0.15 µB/atom and an IQR of ~2.0 µB/atom. Further, as noted earlier, a significant fraction of 10–15% overlapping records across databases disagree on whether the material has non-zero total magnetization. This disagreement may in part be due to different pseudopotential choices for various elements (and correspondingly different number of valence electrons), and sampling of different magnetic configurations, the choice of unit cell in such magnetic configuration sampling, etc. For instance, AFLOW and MP calculate ferromagnetic configurations for all materials, and ferrimagnetic and antiferromagnetic configurations for a subset of materials [48, 49], while the OQMD only calculates ferromagnetic configurations [8]. For a given material, since we only compare the lowest-energy configurations across databases with one another, it is possible that a material is predicted to be non-magnetic in one database and antiferromagnetic in another database. Alternatively, a ferrimagnetic configuration in one database could be compared to a ferromagnetic calculation in another, if both converged to finite magnetic moments.

| Property           | AFLOW | MP   | OQMD  | AFLOW-MP | AFLOW-OQMD | MP-OQMD |
|--------------------|-------|------|-------|----------|------------|---------|
| Formation Energy   | 2196  | 34907| 22248 | 2070     | 1717       | 19082   |
| Volume             | 21929 | 34907| 22248 | 19258    | 15857      | 19082   |
| Band Gap           | 21921 | 34907| 22169 | 19253    | 15790      | 19007   |
| Total Magnetization| 21929 | 34907| 22248 | 19258    | 15857      | 19082   |

**TABLE I.** The number of records after establishing ICSD ID equivalency for each property of interest in the AFLOW, Materials Project (MP), and OQMD HT-DFT databases, as well as for pairwise comparisons of the three databases.

| Property          | MAD (eV/atom) | IQR  | r   | ρ   | MAD (eV/atom) | IQR  | r   | ρ   | MAD (eV/atom) | IQR  | r   | ρ   |
|-------------------|---------------|------|-----|-----|---------------|------|-----|-----|---------------|------|-----|-----|
| Formation Energy  | 0.105         | 0.173| 0.99| 0.99| 0.019         | 0.036| 0.99| 0.99| 0.087         | 0.168| 0.99| 0.99|
| Volume            | 0.180         | 0.389| 0.98| 0.99| 0.647         | 1.117| 0.97| 0.97| 0.512         | 0.902| 0.98| 0.98|
| Band Gap          | 0.078         | 0.203| 0.94| 0.92| 0.209         | 0.364| 0.92| 0.91| 0.178         | 0.277| 0.93| 0.92|
| Total Magnetization| 0.015        | 0.759| 0.77| 0.75| 0.149         | 2.001| 0.60| 0.56| 0.012         | 0.052| 0.80| 0.74|

**TABLE II.** Overall statistics (median absolute difference (MAD), interquartile range (IQR), Pearson’s linear correlation coefficient (r), and Spearman’s rank correlation coefficient (ρ)) for the comparison of properties across HT-DFT databases. For each property, records overlapping across a pair of databases are compared (* for band gap and magnetization, only non-zero values are compared). Generally, lower MAD, lower IQR, higher r, and higher ρ values indicate better reproducibility of calculated properties.
FIG. 1. Distribution of the differences in calculated properties across HT-DFT databases. Each panel corresponds to a property and pair of databases being compared. Solid vertical black lines correspond to the first (Q₁) and third (Q₃) quartiles of the distribution. The number of records overlapping across the two databases is shown in the top right corner of each panel; the median of distribution (\(\tilde{\Delta}x\)), the median absolute difference (MAD), and the interquartile range (IQR) are noted on the left.

C. Rank-order comparisons across properties

We next seek to make comparisons across properties. Instead of comparing the raw values of the properties directly, we compare overlapping records using the ordinal
rank of the property in each database being compared (hereafter, referred to as “percentile rank”). Comparing the percentile ranks of the properties has a few advantages: (a) It allows for a single consistent metric for comparison across all four properties regardless of the magnitude of the actual value and physical units. (b) It is not affected by many systematic differences, e.g., a constant shift of 0.1 eV in all calculated band gaps in one database. Such constant shifts in calculated properties do not affect the internal consistency of a HT-DFT database, and the percentile ranks which are similarly unaffected capture this property. (c) It is a robust, uniform, identifier of outliers in calculated properties.

Figure 2 consists of percentile rank scatterplots (closely related to the quantile-quantile or Q-Q plots) of each property of interest for each database pair. Note that for band gap (total magnetization), we only include overlapping records where the two databases being compared both report the material to be non-metallic (magnetic), to avoid having to rank near-zero or zero values against one another. A compact line along the diagonal corresponds to perfect correlation between the ranked properties, with more diffuse scattering indicating lower levels of correlation.

**Formation energy**: Of the four properties, formation energy shows the best correlation between each database pair, consistent with all $r$ and $\rho$ values close to 0.99 in Table II. Nonetheless, there is some off-diagonal scatter for the MP-OQMD comparison for larger (more positive) values of formation energy that is not found in the other database pairs. These calculations correspond to compounds with smaller (positive) formation energies, where the precision necessary to reliably rank the structure approaches the accuracy of the calculation.

**Volume**: The percentile rank comparison of volume shows higher off-diagonal scatter than that seen in comparisons of formation energy. There is a skew towards higher volumes in AFLOW and MP when compared to OQMD (scatter towards top-left of the diagonal in the AFLOW-OQMD and MP-OQMD comparisons), consistent with the discussion around plane wave energy cutoffs in the previous section.

**Band gap**: The percentile rank comparison of band gap shows even higher off-diagonal scatter than that observed in comparisons of both formation energy and volume. In particular, there is meaningful scatter along the axes, corresponding to cases where one database predicts the material to have a near-zero band gap whereas the other database predicts a (much larger) non-zero band gap.

**Total magnetization**: The percentile rank comparison of total magnetization per formula unit in all three pairwise comparisons shows a few distinct clusters along the diagonal, corresponding to nominally integer values of magnetic moment per formula unit. There is considerable off-diagonal “bowing” in the comparisons with AFLOW, consistent with the distribution of differences between AFLOW and the other two databases showing a skew towards larger magnetizations in AFLOW and long tails (lower panel in Figure 1). In addition, there is considerable off-diagonal scatter (horizontal and vertical bands in the magnetization panel of Figure 2) indicating significant disagreement between the values reported in the two databases.

Overall, a comparison of rank-ordered properties across two databases shows that formation energies and volumes are more easily reproduced than band gaps and total magnetizations, consistent with correlation coefficients decreasing from $\sim 0.99$ for formation energy to $\sim 0.6$ for total magnetization (Table II).

### D. Reproducibility across materials classes

Intuitively, we expect the level of agreement among the databases to be a strong function of materials class. Therefore, we compare specific subsets of calculations based on various materials classes to elucidate potential causes of differences. The materials classes are defined based on chemical composition, the number of elemental components, the presence of magnetism, band gap, pseudopotential choices, and space group, as summarized in Table III. For classes defined by the output of a calculation (i.e., those based on magnetization and band gap), comparisons are only made if both databases agree that the property has a non-zero value. Note that according to our definition, the “Magnetic” class of materials may potentially include both ferromagnetic and ferrimagnetic materials, and the “Non-Magnetic” class may potentially include both non-magnetic and antiferromagnetic materials.
FIG. 2. Comparison of the calculated properties (formation energy, volume, band gap, and total magnetization) over records overlapping across pairwise combinations of HT-DFT databases plotted as a percentile rank (i.e., ordinal rank of the property in each database being compared). A compact line along the diagonal corresponds to perfect correlation between the ranked properties. Overall, formation energies and volumes show better reproducibility than band gaps and magnetizations. The clusters seen in the magnetization comparisons correspond to nominally integer values of magnetic moments.
| Class                  | Definition                                                                 |
|-----------------------|----------------------------------------------------------------------------|
| Oxide                 | Contains O                                                                  |
| Nitride               | Contains N                                                                  |
| Pnictide              | Contains a group 15 element                                                 |
| Chalcogenide          | Contains a group 16 element, except O                                       |
| Halide                | Contains a group 17 element                                                 |
| Alkali Metal          | Contains a group 1 element, except H                                        |
| Alkaline Earth Metal  | Contains a group 2 element                                                  |
| Transition Metal      | Contains a d-block element                                                 |
| Metalloid             | Contains B, Si, Ge, As, Sb, or Te                                          |
| Rare-Earth            | Contains an element from the lanthanide series                             |
| Actinide              | Contains an element from the actinide series                               |
| Metal-Nonmetal        | Contains at least one metal element and at least one of C, N, O, F, P, S, Cl, Se, Br, I |
| Intermetallic         | Contains only metallic elements                                            |
| Magnetic              | Both databases report a net magnetic moment $> 10^{-2} \mu_B$/f.u.         |
| Non-magnetic          | Both databases report no net magnetic moment $> 10^{-2} \mu_B$/f.u.        |
| Disagree on Magnetic  | The two databases disagree on whether a net magnetic moment $> 10^{-2} \mu_B$/f.u. is present |
| Metallic              | Both databases predict a band gap of $< 10^{-2}$ eV                        |
| Semiconductor         | Both databases predict a band gap between $10^{-2}$ and 1.5 eV             |
| Insulator             | Both databases predict a band gap larger than 1.5 eV                       |
| Disagree on Metallic  | The two databases disagree on whether a band gap $< 10^{-2}$ eV is present |
| Pseudopotentials Agree| Both databases use the same set of pseudopotentials for all elements      |
| Pseudopotentials Disagree| The databases use different pseudopotentials for at least one element |
| Use GGA+U             | Both databases use the GGA+U approach                                      |
| Use GGA               | Both databases use plain GGA                                               |
| Disagree on GGA/GGA+U | One database uses GGA whereas the other uses GGA+U                        |
| Elements              | Contains only one element                                                  |
| Binaries              | Contains two elements                                                      |
| Ternaries             | Contains three elements                                                    |
| Quaternaries          | Contains four elements                                                     |
| Triclinic             | Space group 1–2                                                            |
| Monoclinic            | Space group 3–15                                                           |
| Orthorhombic          | Space group 16–74                                                           |
| Tetragonal            | Space group 75–142                                                          |
| Trigonal              | Space group 143–167                                                        |
| Hexagonal             | Space group 168–194                                                        |
| Cubic                 | Space group 195–230                                                        |

TABLE III. Definitions for the materials classes used in this work.
FIG. 3. Median percent absolute differences between properties (formation energy, volume, band gap, total magnetization) calculated in the three databases (AFL OW, MP, OQMD), compared two at a time, across various classes of materials as defined in Table III. The numbers in parentheses indicate the number of overlapping records belonging to the respective material class for a given pair of databases. Trivial comparisons are left blank (e.g., the difference in total magnetization for non-magnetic compounds).

|                  | AFL OW vs MP | AFL OW vs OQMD | MP vs OQMD |
|------------------|--------------|----------------|------------|
|                  | $\Delta E_1$| $V$            | $E_2$      | $M$        |
|                  | $\Delta E_1$| $V$            | $E_2$      | $M$        |
|                  | $\Delta E_1$| $V$            | $E_2$      | $M$        |
| All              | 5.8          | 3.8            | 0.5        | 0.5        |
| Oxide            | (869)        | (865)          | (863)      | (863)      |
| Nitride          | 6.3          | 3.3            | 0.1        | 0.1        |
| Pnicrite         | 6.4          | 2.7            | 0.3        | 0.3        |
| Chalcogenide     | 4.2          | 4.6            | 0.1        | 0.1        |
| Halide           | 14.1         | 2.0            | 0.8        | 0.8        |
| Alkali Metal     | 5.5          | 0.7            | 0.3        | 0.3        |
| Alkaline Earth Metal | 6.9       | 1.7            | 1.0        | 1.0        |
| Transition Metal | 4.9          | 11.1           | 10.1       | 10.1       |
| Metalloid        | 5.7          | 0.8            | 2.0        | 2.0        |
| Rare-Earth       | 3.0          | 9.1            | 3.9        | 3.9        |
| Actinide         | 9.7          | 3.2            | 36.4       | 36.4       |
| Metal-Nonmetal   | 4.7          | 4.7            | 20.4       | 20.4       |
| Intermetal       | 5.5          | 14.4           | 24.0       | 24.0       |
| Magnetic         | 5.8          | 0.8            | 2.3        | 0.8        |
| Non-Magnetic     | 5.8          | 0.8            | 2.3        | 0.8        |
| Disagree On Magnetic | 5.1       | 1.3            | 17.0       | 17.0       |
| Metallic         | 3.9          | 2.0            | 8.7        | 8.7        |
| Insulator        | 4.6          | 1.7            | 4.9        | 4.9        |
| Disagree On Metallic | 5.7       | 0.9            | 3.3        | 3.3        |
| Pseudopotentials Agree | 5.8  | 1.1            | 4.9        | 4.9        |
| Use GGA+U        | 5.8          | 1.1            | 4.9        | 4.9        |
| Use GGA          | 3.9          | 0.8            | 2.3        | 0.8        |
| Disagree on GGA/GGA+U | 5.1       | 0.5            | 3.0        | 3.0        |
| Element          | 6.1          | 1.5            | 12.4       | 12.4       |
| Binary           | 5.8          | 1.4            | 17.7       | 17.7       |
| Ternary          | 5.5          | 1.0            | 4.0        | 4.0        |
| Quaternary       | 5.8          | 0.9            | 3.0        | 3.0        |
| Triclinic        | 6.0          | 0.9            | 3.9        | 3.9        |
| Monoclinic       | 6.0          | 0.9            | 3.9        | 3.9        |
| Orthorhombic     | 5.8          | 0.9            | 3.0        | 3.0        |
| Tetragonal       | 5.8          | 1.0            | 5.0        | 5.0        |
| Trigonal         | 5.4          | 0.9            | 3.5        | 3.5        |
| Hexagonal        | 4.7          | 1.0            | 3.3        | 3.3        |
| Cubic            | 5.7          | 1.1            | 2.9        | 2.9        |
Figure 3 contains the median absolute difference relative to the mean (MRAD) values for pairwise comparisons between databases, divided into materials classes as defined in Table III. Cells are colored based on the MRAD value listed. Empty cells correspond to trivial comparisons (e.g., values of band gap where both database agree the structure is metallic). We use MRAD as the metric here to reduce the effect of outliers (as compared to calculating means) as well as to enable comparisons across properties using the same metric. Overall, HIT-DFT volumes show the best agreement (lowest MRAD values), from 1–4%. Band gaps show the worst overall agreement (highest MRAD values), 4–10% across all pairwise comparisons. Formation energy comparisons with MP show MRAD values up to 6%, but the AFLOW-OQMD MRAD is only 1.3%. MRAD values for total magnetization vary highly from 0.5% for comparisons with MP to 7.6% for AFLOW-OQMD. In all cases, certain materials classes have distinctly higher or lower MRAD when compared to the MRAD averaged over all materials classes.

**Formation Energy:** In the comparisons with AFLOW, two materials classes, “Halides” and “Disagree on Metallic”, show the highest MRAD values of up to 14% and 40%, respectively. The high MRAD in halide formation energies can be understood to result from post hoc corrections to the effective elemental reference energies performed in MP and OQMD, but not in AFLOW, for the halide group of elements (see discussion in Section IVB). The high MRAD of the “Disagree on Metallic” class is likely an artifact of the small formation energies of the few records (~30–50) in the comparison. As noted earlier, since AFLOW reports notably fewer formation energy values than the other databases, the comparisons are made with a much smaller set of records (~2,000). Therefore, we ignore here some of the MRAD outliers in cases where the number of records being compared is very small (e.g., the material class “Magnetic” shows an MRAD of 13% between AFLOW and MP but there are only 5 records in the comparison). Further, the formation energies dataset has very few transition metal, rare-earth, and actinide element-containing compounds (Figures S3 and S7). New, different insights are likely to result from a larger dataset. In the MP-OQMD comparison, with a much larger comparable dataset (~19,000), the “Nitride”, “Pnictide”, and “Chalcogenide” material classes show the highest MRAD values, 14%, 8%, and 11% respectively. This is partly due to differences in fitted elemental chemical potentials for pnictogen and chalcogen elements in MP and OQMD (Section IVB).

**Volume:** The best agreement is observed in the AFLOW-MP comparisons, with only the “Actinide” material class showing an MRAD greater than 2%. For comparisons with OQMD, the MRAD in volume is generally higher—due to the choice of lower plane wave energy cutoff used for cell relaxation, as discussed earlier (Section III B). The highest MRAD values in the comparisons with OQMD volumes are for the “Nitride” and “Halide” classes (~7–9%). The default plane wave energy cutoffs in the VASP PAW potentials (ENMAX parameter) for N and F are among the highest (400 eV) of all elements. Thus, the lower energy cutoff used by OQMD for relaxation impacts the calculated volumes of nitrdes and fluorides the most (Figures S8 and S12). Another material class, “Triclinic”, shows similarly high MRAD values of ~8% in comparisons with OQMD. Upon examination, we find that most triclinic materials in the comparisons are oxides, nitrdes, and halides, and thus the high MRAD values are due to the chemical composition of these compounds rather than their crystal symmetry.

**Band gap:** While band gap comparisons show the highest MRAD values across properties, some materials classes in particular show MRAD values much greater than ~10%. Of these, in the “Intermetallic” and “Semiconductor” material classes, the MRAD values are expectedly high due to small average band gaps relative to which differences are reported, even though the absolute differences themselves are not conspicuously large (Figure S2). In other cases, the high MRAD values are a result of (a) different pseudopotential choices for elements (e.g., Cu/Cu_pv, Ce/Ce_3, Eu/Eu_2 choices in the “Disagree on Magnetic” class for the MP-OQMD comparison with an MRAD of ~53%; see Figure S13), (b) disagreement on whether to use the GGA or GGA+U approach to calculate properties (e.g., the “Actinide” material class with MRAD of up to 43% in comparisons with MP, the “Disagree on GGA/GGA+U” class in all three comparisons with MRAD of 12–25%), or a combination of both factors (e.g., for the “Magnetic” material class with an MRAD of up to 27% in comparisons with AFLOW), (c) non-overlapping sampling of magnetic configurations across databases. For instance, the “Magnetic” (MRAD of 13–27% across comparisons) and “Disagree on Magnetic” (MRAD of 17–53% across comparisons) classes may respectively include comparing ferromagnetic vs ferrimagnetic and non-magnetic vs antiferromagnetic ground states across two databases (note, however, that both the “Magnetic” and “Disagree On Magnetic” comparisons also include effects from other HT-DFT choices, such as choice of pseudopotential used). Note also that the errors in band gaps for the “Use GGA+U” materials class are larger than those for the “Use GGA” materials class across all three pairwise comparisons, the choice of slightly different effective U values used in the three databases being a likely contributor. Further discussions of some of the above parameter choices are in Section IV.

**Total magnetization:** While MRAD values in the MP-OQMD comparison are generally small (<5%), some material classes show much higher MRAD values, especially in comparisons with AFLOW. As in the case of band gap values, we find these comparisons to be influenced by pseudopotential choice (of rare-earth elements in particular, e.g., Nd, Nd_3, Nd_3 in AFLOW, MP, and OQMD, respectively; see Figures S10 and S14), choice of using GGA or GGA+U (e.g., MRAD of up to ~40%
in AFLOW-OQMD comparisons for the “Disagree on GGA/GGA+U” class), or both (e.g., the “Metalloid” and “Rare-Earth” material classes in the AFLOW-OQMD comparisons, “Intermetallic” and “Metallurgical” classes in the AFLOW-MP and AFLOW-OQMD comparisons). We note that some other material classes show high MRAD values, e.g., “Element”, “Binary”, “Ternary”, “Tetragonal”, “Hexagonal”, and “Cubic” (up to MRAD values up to ~50%) due to, upon further examination, the parameter choices discussed above rather than due to number of components in the compound or crystal symmetry.

Finally, we note that while our scheme of constructing a set of comparable records across pairs of databases (by matching ICSD IDs exactly) ensures comparisons between the same initial crystal structures, it excludes a number of experimentally well-studied materials with multiple ICSD entries associated with them. We investigated whether this “bias away from well-studied materials” affects our results by using a larger comparison set constructed by linking very similar ICSD entries using the crystal structure matching algorithm employed by the Materials Project (see Section S-II in the SI). While some of the quantitative metrics we report varied by a few percent in the expanded comparison, the overall conclusions remain unchanged (see Tables S-XI, S-XII, and Figures S15–S18 in the SI), consistent with recent findings [50].

IV. DISCUSSION

We discuss some of the most important factors affecting the differences across HT-DFT calculations of properties below. Some of the other factors that either have a minor effect (e.g., post hoc calculation of band gap from band dispersions or density of states) or are specific to a database/property (e.g., plane wave cutoff energy for full cell relaxations in OQMD) have been discussed in the earlier sections.

A. Effects of pseudopotential choice

For nearly all elements, VASP provides multiple PAW potentials to choose from, with different numbers of electrons in the valence. The choice of pseudopotential varies across the HT-DFT databases due to factors such as changes in VASP recommendations and issues of calculation convergence or reproduction of experimental thermochemical data [51, 52]. Interestingly, the choice of pseudopotential has minimal effect on the calculated formation energies and volumes (up to a difference of 1% in cases where pseudopotentials do or do not match; see rows “Pseudopotentials Agree” and “Pseudopotentials Disagree” in Figure 3). On the other hand, the number of valence electrons and consequently the choice of pseudopotential affects the calculated band gaps and magnetization values severely. Especially egregious differences across those properties in material classes such as “Rare-Earth” and “Magnetic” (Figure 3) can be directly traced to different pseudopotential choices. For rare-earth and actinide elements in particular, with f-electrons that are poorly described by DFT [53], using pseudopotentials that treat f-electrons in core or valence can have a significant impact on the calculated band gap (e.g., “Intermetallic” and “Magnetic” classes in Figure 3) and magnetization (e.g., “Rare-Earth” and “Intermetallic” classes in Figure 3) values.

B. Elemental references and energy corrections

The largest disagreements in HT-DFT formation energies can be understood to result from different elemental reference states and/or post-calculation energy corrections performed in the databases. To our knowledge, the formation energies reported in AFLOW use DFT total energies of the bulk elements as the reference states [54]. MP and OQMD both correct DFT-calculated energies to closely reproduce experimental formation enthalpy data. While MP adds corrections to the compound formation energies [51, 52], OQMD fits the elemental reference energies using a FERE-like approach [8, 55]. Such correction schemes involve some more HT-DFT choices: (a) Should all elemental reference energies and/or compound formation energies be effectively fit to experimental data or only a subset? For instance, MP corrects the compound formation energies of nitrides, fluorides, chlorides, hydrides, sulfides of alkali, alkaline earth, and aluminum-containing compounds [56]. The OQMD fits the reference energies of only elements whose DFT ground states are poor representation of the experimental reference states (i.e., elements that are gases or that have a solid-solid phase transition below room temperature) [8]. (b) What experimental thermochemical data should be used such correction schemes, given a lack of a single, widely-accepted set of standard experimental dataset for solids? For instance, MP and OQMD use experimental formation energies from different sources to fit elemental reference energies: MP uses data from Materials Thermochemistry [57], while OQMD uses data from SGTE SUBstance Database (SSUB) [58] in addition to others (see Refs. 8 and 52 for details of the fitting data used in the two databases). Some other standard reference databases are also widely used, such as the NIST-JANAF Thermochemical Tables [59]. Since a given material may have experimental data in one or more such reference databases of experimental properties, the choice of the source of experimental data affects the fitted formation energies in HT-DFT databases, even in cases where other parameters such as pseudopotentials used are held constant. This effect of fitted elemental reference states is shown in the calculated formation energies averaged over compounds containing each element in Figures S3, S7, and S11.
C. GGA vs. GGA+U approach

One of the ways to treat the issue of over-delocalization in DFT is to use the DFT+U approach [60, 61] (or “GGA+U” when used with GGA). Similar to the case of fitting elemental references, using the GGA+U approach requires additional HT-DFT choices. (a) Whether or not to use GGA+U for calculating properties of a given material. All three HT-DFT databases have slightly different sets of compounds for which the GGA+U approach is applied. The OQMD uses GGA+U only for oxides of certain 3d transition metals (the V–Cu series) and actinide metals [8]. MP uses GGA+U for oxides, fluorides, and sulfides of a larger set of transition metals, but not actinides [52]. AFLOW applies it to an even larger set of compounds, nearly all those containing d- or f-block elements [62]. (b) What effective U value should be used for each element? The three HT-DFT databases all use different effective U values for each element, obtained either from previous work (OQMD) or in-house parameterization by fitting to experimental data (AFLOW and MP) [21, 63]. Such choices around when to use the GGA+U approach to calculate a compound and what effective U value to use can impact some properties more than others, e.g., discrepancies in total magnetization values in the AFLOW-OQMD comparisons, particularly for “Rare-Earth”, “Intermetallic”, and “Metallic” classes. For some properties, such as formation energies, post hoc corrections are required to maintain consistency between those calculated using the GGA and GGA+U approaches, especially while constructing phase diagrams involving compounds calculated using the two different approaches. Such corrections are obtained by fitting to experimental reaction energies, and can be different between HT-DFT databases based on the source of such reaction energies.

V. CONCLUSION

Recent years have seen a dramatic increase in the application of informatics methods for materials development, using high-throughput DFT data. Several prominent HT-DFT databases exist and each uses different input parameters and post-processing techniques to calculate materials properties. Quantifying the uncertainty in calculated properties due to such parameter choices is therefore crucial to understanding the reproducibility and interoperability of such data. In this work, we centralize data from three of the largest HT-DFT databases, AFLOW, Materials Project, and OQMD, into a common data repository, allowing records to be accurately compared. We then compare four properties—formation energy, volume, band gap, and total magnetization—of materials calculated in each of the HT-DFT databases using the same initial crystal structure.

Our comparisons show that formation energy and volume are more easily reproduced than band gap and total magnetization. Interestingly, we find that the average difference in calculated properties across two HT-DFT databases is comparable to that between DFT and experiment: up to 0.105 eV/atom for formation energy, 4% for volume, 0.21 eV for band gap, and 0.15 µB/formula unit for total magnetization. Further, certain input parameter choices disproportionately affect HT-DFT properties of particular classes of materials, e.g., choice of planewave cutoff on formation energies and volumes of oxides and halides, and the choice of pseudopotential on the band gaps and magnetization of rare-earth compounds. Our results inform users of the variability to account for in reported materials properties, especially when using data from multiple HT-DFT databases in their own analyses. In addition, our quantitative uncertainty estimates can directly aid materials informatics efforts, e.g., for separation of model uncertainty and inherent noise in data.

As HT-DFT databases continue to mature, systematic comparisons, interoperability, and standardization of calculations become increasingly crucial. Efforts to improve the interoperability of materials databases, e.g., by the development of a common data schema by the OP-TiMaDe consortium [64], are already ongoing. Toward improving the standardization of calculations, HT-DFT choices and reproducibility in particular, we list a few recommendations for next-generation and new iterations of current HT-DFT databases:

(a) In-depth, versioned documentation of the various parameter choices made in a high-throughput project, including the data-driven rationale for the choices, if any.

(b) Visibility for possible uncertainty in reported properties (in both the web and programmatic interfaces used to interact with HT-DFT data) for which HT-DFT choices are expected to have a significant impact. Further, we recommend providing estimated uncertainties in calculated properties, either determined from literature references (e.g., this work), or from in-house investigations (e.g., by performing a set of HT-DFT calculations with different input parameters as part of a sensitivity analysis).

(c) Community-led initiative to reach a consensus on which HT-DFT choices ought to be standardized (e.g., energy cutoffs, fitting sets for empirical corrections, post-processing steps to determine properties such as band gap) and which HT-DFT choices could be a source of greater scientific insight if they were more diverse (e.g., DFT codes, pseudopotentials, DFT exchange-correlation functionals).

CONFLICTS OF INTEREST

ZdR was previously employed by Citrine Informatics. PS has worked as a subcontractor to Citrine Informatics.
ACKNOWLEDGEMENTS

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Small Business Technology Transfer Program under Award Number DE-SC0015106. The authors would like to thank Cormac Toher for advice on using the AFLUX RESTful API, and Anubhav Jain, Shyue Ping Ong, Matthew Horton, Eric B. Isaacs, and Chris Wolverton for their comments on an earlier version of this manuscript.

AUTHOR CONTRIBUTIONS

Conceptualization: C.K.H.B., V.H., P.S., M.H., J.E.S., B.M.; Methodology: C.K.H.B., V.H., E.A., Y.K., M.H., P.S., J.E.S.; Software: C.K.H.B., V.H., E.A., Y.K.; Validation: C.K.H.B., V.H., B.M.; Formal analysis: C.K.H.B., V.H., Z.d.R., E.A., Y.K.; Investigation: C.K.H.B., V.H.; Data Curation: C.K.H.B., V.H.; Writing – Original Draft: C.K.H.B., V.I.H., M.H., P.S., J.E.S.; Writing – Review & Editing: all authors; Visualization: E.A., Y.K., C.K.H.B., V.H.; Supervision: J.E.S., B.M., J.L.

DATA AVAILABILITY

All data and Python scripts required to perform the analysis presented in this work are made available via the GitHub repository at https://github.com/CitrineInformatics-ERD-public/htdft-uq.
Appendix A: Definitions of statistical quantities

The definitions of statistical quantities and their symbols used in this work throughout are as follows ($x_i$ and $y_i$ refer to the two sets of data being compared, e.g. from two different databases):

1. Median difference ($\bar{\Delta}x$):

$$\bar{\Delta}x = \text{median}(x_i - y_i) \quad (A1)$$

2. Median absolute difference (MAD):

$$\text{MAD} = \text{median}(|x_i - y_i|) \quad (A2)$$

3. Interquartile range (IQR):

$$\text{IQR} = Q_3 - Q_1 \quad (A3)$$

where $Q_1$ and $Q_3$ are the first and third quartiles (25th and 75th percentiles), respectively.

4. Median relative absolute difference (MRAD):

$$\text{MRAD} = \text{median}\left(\frac{|x_i - y_i|}{|x_i + y_i|/2}\times 100\right) \quad (A4)$$

5. Pearson correlation coefficient ($r$):

$$r(x, y) = \frac{\sum^n_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum^n_i (x_i - \bar{x})^2} \sqrt{\sum^n_i (y_i - \bar{y})^2}} \quad (A5)$$

where $\bar{x} = \frac{1}{n} \sum^n_i x_i$ is the sample mean, and $n$ is the sample size.

6. Spearman’s rank correlation coefficient ($\rho$) is defined as the Pearson correlation coefficient between rank variables $x^R_i$ and $y^R_i$ corresponding to raw data values $x_i$ and $y_i$, respectively:

$$\rho(x, y) = r(x^R, y^R) \quad (A6)$$

---

[1] S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, and O. Levy, The high-throughput highway to computational materials design, Nat. Mater. 12, 191 (2013).

[2] A. Jain, Y. Shin, and K. A. Persson, Computational predictions of energy materials using density functional theory, Nat. Rev. Mater. 1, 15004 (2016).

[3] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65, 1501 (2013).

[4] G. Kresse and J. Furthmüller, Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set, Comput. Mater. Sci. 6, 15 (1996).

[5] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab-initio total energy calculations using a plane-wave basis set, Phys. Rev. B 54, 11169 (1996).

[6] A. E. Mattsson, P. A. Schultz, M. P. Desjarlais, T. R. Mattsson, and K. Leung, Designing meaningful density functional theory calculations in materials science—a primer, Modell. Simul. Mater. Sci. Eng. 13, R1 (2004).

[7] S. P. Ong, S. M. Blau, X. Qu, W. Richards, S. Dwarknath, S. Dacek, J. Montoya, R. Kingsbury, A. Jain, JSX, M. Horton, D. Waroquiers, R. Tran, H. Tang, P. Huck, G. Hautier, G. Petretto, sivonxay, C. Zheng, KeLiu, and A. Rutt, materialsproject/custodian: v2020.4.27 (2020).

[8] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, npj Comput. Mater. 1, 15010 (2015).

[9] S. P. Huber, S. Zoupanos, M. Uhrin, L. Talirz, L. Kahle, R. Häuselmann, D. Gresh, T. Müller, A. V. Yakutovich, C. W. Andersen, et al., AiiDA 1.0, a scalable computational infrastructure for automated reproducible workflows and data provenance, Scientific data 7, 1 (2020).

[10] Y. Zhuo, A. M. Tehrani, A. O. Oliynyk, A. C. Duke, and J. Brug, Identifying an efficient, thermally robust inorganic phosphor host via machine learning, Nat. Commun. 9, 1 (2018).

[11] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, Combinatorial screening for new materials in unconstrained composition space with machine learning, Phys. Rev. B 89, 094104 (2014).

[12] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, et al., Reproducibility in density functional theory calculations of solids, Science 351, 13000 (2016).

[13] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, et al., AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. 58, 227 (2012).

[14] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1, 011002 (2013).

[15] A. Belsky, M. Hellenbrandt, V. L. Karen, and P. Luksch, New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design, Acta Crystallogr. B 58, 364 (2002).

[16] K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. H. Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, et al., The joint
automated repository for various integrated simulations (JARVIS) for data-driven materials design, npj Comput. Mater. 6, 1 (2020).

[17] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, et al., Materials Cloud, a platform for open computational science, Sci. Data 7, 1 (2020).

[18] P. E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50, 17953 (1996).

[19] G. Kresse and D. Joubert, From ultrasoft pseudopotentials to the projector augmented-wave method, Phys. Rev. B 59, 1758 (1999).

[20] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77, 3865 (1996).

[21] W. Setyawan and S. Curtarolo, High-throughput electronic band structure calculations: Challenges and tools, Comput. Mater. Sci. 49, 299 (2010).

[22] W. Setyawan, R. M. Gaume, L. Lam, R. S. Feigelson, and S. Curtarolo, High-throughput combinatorial database of electronic band structures for inorganic scintillator materials, ACS Combi. Sci. 13, 382 (2011).

[23] O. Levy, G. L. Hart, and S. Curtarolo, Uncovering compounds by synergy of cluster expansion and high-throughput methods, J. Am. Chem. Soc. 132, 4830 (2010).

[24] C. Toher, J. J. Plata, O. Levy, M. de Jong, M. Asta, M. B. Nordelli, and S. Curtarolo, High-throughput computational screening of thermal conductivity, diebye temperature, and grüneisen parameter using a quasiharmonic debye model, Phys. Rev. B 90, 174107 (2014).

[25] C. Toher, C. Oses, J. J. Plata, D. Hicks, F. Rose, O. Levy, M. de Jong, M. Asta, M. Fernari, M. B. Nordelli, et al., Combining the AFLOW GIBBS and elastic libraries to efficiently and robustly screen thermomechanical properties of solids, Phys. Rev. Mater. 1, 015401 (2017).

[26] R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. B. Nordelli, and S. Curtarolo, A RESTful API for exchanging materials data in the AFLOWLIB.org consortium, Comput. Mater. Sci. 93, 178 (2014).

[27] M. De Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Shuiter, C. K. Ande, S. Van Der Zwaag, J. J. Plata, et al., Charting the complete elastic properties of inorganic crystalline compounds, Sci. Data 2, 150009 (2015).

[28] W. Chen, J.-H. Pöhl, G. Hautier, D. Broberg, S. Bajaj, U. Aydemir, Z. M. Gibbs, H. Zhu, M. Asta, G. J. Snyder, et al., Understanding thermoelectric properties from high-throughput calculations: trends, insights, and comparisons with experiment, J. Mater. Chem. C 4, 4414 (2016).

[29] M. De Jong, W. Chen, H. Geerlings, M. Asta, and K. A. Persson, A database to enable discovery and design of piezoelectric materials, Sci. Data 2, 150053 (2015).

[30] I. Petousis, D. Mrdjenvovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T. D. Schladt, K. A. Persson, and F. B. Prinz, High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials, Sci. Data 4, 160134 (2017).

[31] G. Petretto, S. Dwaraknath, H. P. Miranda, D. Winston, M. Giantomassi, M. J. Van Setten, X. Gonze, K. A. Persson, G. Hautier, and G.-M. Rignanese, High-throughput density-functional perturbation theory phonons for inorganic materials, Sci. Data 5, 1 (2018).

[32] C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. Piper, et al., Automated generation and ensemble-learned matching of X-ray absorption spectra, npj Comput. Mater. 4, 1 (2018).

[33] K. A. Persson, B. Waldwick, P. Lazic, and G. Ceder, Prediction of solid-aqueous equilibria: scheme to combine first-principles calculations of solids with experimental aqueous states, Phys. Rev. B 85, 235438 (2012).

[34] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, Comput. Mater. Sci. 97, 209 (2015).

[35] S. Kirklin, J. E. Saal, V. I. Hegde, and C. Wolverton, High-throughput computational search for strengthening precipitates in alloys, Acta Mater. 102, 125 (2016).

[36] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton, High-throughput computational screening of perovskites for thermochemical water splitting applications, Chem. Mater. 28, 5621 (2016).

[37] D. Wang, M. Amsler, V. I. Hegde, J. E. Saal, A. Issa, B.-C. Zhou, X. Zeng, and C. Wolverton, Crystal structure, energetics, and phase stability of strengthening precipitates in Mg alloys: A first-principles study, Acta Mater. 158, 65 (2018).

[38] A. R. Akbarzadeh, V. Ozoliņš, and C. Wolverton, First-principles determination of multicomponent hydride phase diagrams: Application to the Li-Mg-N-H system, Adv. Mater. 19, 3243 (2007).

[39] V. I. Hegde, M. Aykol, S. Kirklin, and C. Wolverton, The phase stability network of all inorganic materials, Sci. Adv. 6, eaay5606 (2020).

[40] M. Amsler, V. I. Hegde, S. D. Jacobsen, and C. Wolverton, Exploring the high-pressure materials genome, Phys. Rev. X 8, 041021 (2018).

[41] QO MD RESTful API, http://oqmd.org/static/docs/restful.html (2019), accessed: May 2020.

[42] K. Michel and B. Meredig, Beyond bulk single crystals: a data format for all materials structure–property–processing relationships, MRS Bull. 41, 617 (2016).

[43] P. Haas, F. Tran, and P. Blaha, Calculation of the lattice constant of solids with semiclassical functionals, Phys. Rev. B 79, 085104 (2009).

[44] G. Olsen, C. Nuese, and R. Smith, The effect of elastic strain on energy band gap and lattice parameter in III-V compounds, J. Appl. Phys. 49, 5523 (1978).

[45] C. Kuo, S. Vong, R. Cohen, and G. Stringfellow, Effect of mismatch strain on band gap in III-V semiconductors, J. Appl. Phys. 57, 5428 (1985).

[46] S.-H. Wei and A. Zunger, Predicted band-gap pressure coefficients of all diamond and zinc-blende semiconductors: Chemical trends, Phys. Rev. B 60, 5404 (1999).

[47] J. M. Munro, K. Latimer, M. K. Horton, S. Dwaraknath, and K. A. Persson, An improved symmetry-based approach to reciprocal space path selection in band structure calculations, npj Comput. Mater. 6, 1 (2020).

[48] S. Sanvito, C. Oses, J. Xue, A. Tiwari, M. Zic, T. Archer, P. Tozman, M. Venkatesan, M. Coey, and S. Curtarolo, Accelerated discovery of new magnets in the Heusler alloy family, Sci. Adv. 3, e1602241 (2017).

[49] M. K. Horton, J. H. Montoya, M. Liu, and K. A. Persson, High-throughput prediction of the ground-state collinear
magnetic order of inorganic materials using density functional theory, npj Comput. Mater. 5, 1 (2019).

[50] J. Marquez Chavez and B. Kiefer, Matcor, a program for the cross-validation of material properties between databases, Comput. Mater. Sci. 187, 110103 (2021).

[51] Materials Project: Calculations Guide, https://materialsproject.org/docs/calculations (accessed: December 2019).

[52] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, and G. Ceder, A high-throughput infrastructure for density functional theory calculations, Comput. Mater. Sci. 50, 2295 (2011).

[53] L. Eyring, K. A. Gschneidner, and G. H. Lander, Handbook on the physics and chemistry of rare earths, Vol. 32 (Elsevier, 2002).

[54] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, et al., Aflow: An automatic framework for high-throughput materials discovery, Comput. Mater. Sci. 58, 218 (2012).

[55] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies, Phys. Rev. B 85, 115104 (2012).

[56] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68, 314 (2013).

[57] O. Kubaschewski, C. B. Alcock, and P. Spencer, Materials Thermochemistry (Pergamon Press, 1993).

[58] SGTE, Thermodynamic Properties of Inorganic Materials, Vol. 19 (Springer-Verlag, 1999).

[59] M. W. Chase, NIST-JANAF Thermochemical Tables 4th Ed., J. Phys. Chem. Ref. Data Monograph No. 9, 1529 (1998).

[60] V. I. Anisimov, J. Zaanen, and O. K. Andersen, Band theory and mott insulators: Hubbard U instead of Stoner I, Phys. Rev. B 44, 943 (1991).

[61] H. J. Kulik, Perspective: Treating electron over-delocalization with the DFT+U method, J. Chem. Phys. 142, 240901 (2015).

[62] C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. B. Nardelli, and S. Curtarolo, The AFLOW standard for high-throughput materials science calculations, Comput. Mater. Sci. 108, 233 (2015).

[63] A. Jain, G. Hautier, S. P. Ong, C. J. Moore, C. C. Fischer, K. A. Persson, and G. Ceder, Formation enthalpies by mixing GGA and GGA+U calculations, Phys. Rev. B 84, 045115 (2011).

[64] C. Andersen, R. Armiento, E. Blokhin, G. Conduit, S. Dwaraknath, M. L. Evans, A. Fekeete, A. Gopakumar, S. Gražulis, V. Hegde, M. Horton, S. Kumbhar, N. Marzari, A. Merkys, F. Mohamed, A. Morris, C. Oses, G. Pizzi, T. Purell, G.-M. Rignanese, M. Scheffler, M. Scheidgen, L. Talirz, C. Toher, M. Uhrin, D. Winston, and C. Wolverton, The OPTIMADE Specification (2020).
S-I. HT-DFT DATABASES AND MANAGEMENT CODES

Given the popularity of DFT as a method generating materials data, a number of HT-DFT databases with various focus areas have emerged; a list of exemplars is given in Table S-I. Similarly, codes for the management of these databases, including the workflows to generate the data, have been developed by several groups around the world, such as the packages listed in Table S-II.

| Database     | Link                     | Materials     | Properties                          | Reference |
|--------------|--------------------------|---------------|-------------------------------------|-----------|
| Aflowlib     | materials.duke.edu/aflow.html | inorganic solids | electronic structure, thermodynamics | [1]       |
| Alloy database | alloy.phys.cmu.edu     | intermetallics  | structure, cohesive energies        | [2]       |
| CatApp       | slac.stanford.edu/~strabo/catapp | molecules on surfaces | reaction/activation energies          | [3]       |
| CCCDB        | cccbdb.nist.gov          | atoms, molecules | thermochemical properties            | [4]       |
| CMR          | cmr.fysik.dtu.dk         | perovskites, 2D materials | energetics, electronic structure    | [5]       |
| CompES-X     | compes-x.nims.go.jp      | inorganic solids | electronic structure                 |           |
| Crystalium   | crystalium.materialsvirtuallab.org | elemental solids | surface, grain boundary energetics   | [6]       |
| CEP          |                          | organic photovoltaics | HOMO-LUMO energies                  | [7]       |
| JARVIS-DFT   | jarvis.nist.gov          | 2D/solid inorganics | elastic, thermoelectric properties   | [8]       |
| Materials Cloud | www.materialscloud.org | porous, 2D materials | structural, topological properties   | [9]       |
| Materials Project | materialsproject.org     | inorganic solids | mechanical, dielectric, piezoelectric | [10]      |
| NoMaD        | nomad-coe.eu             | inorganic solids | raw DFT calculation files            |           |
| NRELMatDB    | materials.nrel.gov       | inorganic solids | quasiparticle energies               | [11]      |
| OQMD         | oqmd.org                 | inorganic solids | energetics, electronic structure    | [12]      |
| phonondb     | phonondb.mtl.kyoto-u.ac.jp | inorganic solids | phonons, thermal properties          |           |
| TE Design Lab |                          | semiconductors  | electronic, thermoelectric properties | [13]      |

* These authors contributed equally to this work
† bryce@citrine.io
### TABLE S-II. A selection of software tools for automating HT-DFT workflows and property calculations.

| Code   | Functionality                        | Link                                      | Reference |
|--------|--------------------------------------|-------------------------------------------|-----------|
|        | High-throughput workflows             |                                           |           |
| AFLOW  | calculation setup, submission        | materials.duke.edu/aflow.html             | [14]      |
| AiiDA  | calculation setup, submission, storage| aida.net                                  | [15]      |
| ASE    | calculation setup, submission, analysis| wiki.fysik.dtu.dk/ase                    | [16]      |
| MPInterfaces | surface calculation setup, analysis | github.com/henniggroup/MPInterfaces    | [17]      |
| pymatgen | calculation setup, analysis          | pymatgen.org                             | [18]      |
| qmpy   | calculation setup, management, analysis| pypi.org/project/qmpy                   | [19]      |
|        | Calculation of properties/model building |                                           |           |
| Amp    | atomistic potentials                 | amp.readthedocs.io                       | [20]      |
| ATAT   | cluster expansions                   | brown.edu/Departments/Engineering/Labs/avdw/atat | [21]      |
| atomate| electronic structure, dielectric tensors| atomate.org                             | [22]      |
| CALYPSO| crystal structure prediction          | www.calypso.cn                           | [23]      |
| MAST   | defects, diffusion                   | pythonhosted.org/MAST                    | [24]      |
| Mint   | crystal structure utilities          | github.com/materials/mint                |           |
| phonopy| phonons, high-temperature properties | atztogo.github.io/phonopy                 | [25]      |
| SeeK-path| high-symmetry paths in Brillouin zone| materialscloud.org/work/tools/seekpath  | [26]      |
| SLUSCHI| melting temperatures                  | blogs.brown.edu/qhong/?page_id=102       | [27]      |
| USPEX  | crystal structure prediction          | uspex-team.org/en/uspex/overview         | [28]      |
| Xtalopt| crystal structure prediction          | xtalopt.github.io                       | [29]      |
S-II. DATA MANAGEMENT

As described in the main text, here we focus on three high-throughput DFT databases: Automatic FLOW (AFLOW), the Materials Project (MP) and the Open Quantum Materials Database (OQMD). MP and OQMD distribute the HT-DFT data and related content via the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0). Individual AFLOW records contain the following license statement: “The data included in the AFLOW repository is free for scientific, academic, and non-commercial purposes, and any other use is prohibited.”

The primary challenge in comparing calculations from different HT-DFT databases is to collect the relevant database entries and determine equivalency between unique calculations of the same structure. This section outlines the workflow of ingesting data from AFLOW, Materials Project, and OQMD into a central repository, all post-extraction processing of the data, the method for curating the post-processed data, and the method of determining whether two records were comparable.

A. Querying Data from HT-DFT Databases

To create a complete and searchable system of records, entries from each HT-DFT database were imported into the a central repository, after standardization into the Physical Information File (PIF) format [30] using the pypif [31] package (see Section S-VII for a few example PIFs with data queried from HT-DFT databases). The data in this work represents the data found on the public-facing HT-DFT databases, aggregated in June 2021.

Each HT-DFT database was queried independently using their respective APIs:

1. AFLOW: The query was performed using the AFLUX API [1], retrieving all records in the icsd catalog. The query resulted in 60,324 records, which were stored as PIF objects.

2. Materials Project: The query was performed using the MPRester API [32], retrieving all records with non-empty icsd_ids field. The Materials Project (MP) release log (https://matsci.org/t/materials-project-database-release-log) indicates that the version corresponding to when data in this work was queried is v2019.11. The query resulted in 48,833 records, which were stored as PIF objects. In addition, some normalized properties such as per-atom volume and per-formula-unit magnetization that were not queried directly but calculated from queried data were added to the PIF objects.

3. OQMD: The query was performed using qmpy and a version 1.2 of the underlying database, filtering on FormationEnergy objects associated with converged "static" Calculation objects and Entry objects with "icsd" in the "path" field. The query resulted in 37,989 records, which were stored as PIF objects. In addition, some properties that were not directly queried but retrieved as metadata, such as labels of PAW potentials and crystal system, were added to the PIF objects.

B. Aggregation of HT-DFT Datasets

The HT-DFT database-specific tags for each of the queried properties are given in Table S-III. All three datasets are aggregated in the PIF format, using a unique key for each property irrespective of the field names queried from the HT-DFT databases. For instance, the number of atoms in the unit cell is stored as number_of_atoms for all three databases despite varying labels (e.g., nsites, natoms, calculation__output__natoms) across databases. The aggregated dataset is available in the data folder in this GitHub repository: https://github.com/CitrineInformatics-ERD-public/htdft-uq.

C. Structure Equivalency

In order to perform a fair comparison between two HT-DFT databases, it is essential to generate a set of equivalent records for each pair of databases considered. We use the ICSD Collection Code(s) (hereafter, “ICSD ID(s)”) in the metadata of each entry to generate a set of comparable records.
TABLE S-III. HT-DFT database-specific tags for properties compared in this work. In cases where they were not directly queried, normalized values such as per-atom values for “volume” and per-formula unit values for “total magnetization” were calculated from the non-normalized queried data.

| Property                        | AFLOW            | Materials Project | OQMD              |
|---------------------------------|------------------|-------------------|-------------------|
| Record ID                       | auid             | material_id       | entry_id          |
| Record URL                      | aurl             | -                 | -                 |
| Composition                     | compound         | unit_cell_formula | composition_formula |
| Number of atoms                 | natoms           | nsites            | output_natoms     |
| Total energy (/unit cell)       | energy_cell      | final_energy      | calculation_energy |
| Total energy (/atom)            | energy_atom      | final_energy_per_atom | calculation_energy_pa |
| Formation energy                | enthalpy_formation_atom | formation_energy_per_atom | formation_delta_e |
| Convex hull distance            | -                | e_above_hull      | formation_stability |
| Volume (/unit cell)             | volume_cell      | volume            | output_volume     |
| Volume (/atom)                  | volume_atom      | -                 | output_volume_pa  |
| Magnetization (/unit cell)      | spin_cell        | total_magnetization | -                 |
| Magnetization (/atom)           | spin_atom        | -                 | calculation_magmom_pa |
| Band gap                        | Egap             | band_gap          | calculation_band_gap |
| Space group ITC #               | spacegroup_relax | spacegroup        | spacegroup__number |
| Crystal system                  | lattice_system_relax | crystal_system    | -                 |
| Pseudopotentials                | species_pp       | pseudo_potential  | calculation__settings |

1. Exact ICSD ID Matching

The main text contains analysis only for records with the exact same ICSD ID across the two databases being compared at a time, ensuring that the crystal structure of the materials being compared are the same. Since the ICSD ID of each material is retrieved as part of the metadata queried from each of the three HT-DFT databases, materials can be “matched” across databases by directly comparing their ICSD IDs.

2. Aliasing for Multiple ICSD IDs Per Record

Since the former process of exact ICSD ID matching results in a smaller set of records (less than 50%) when compared to the total ICSD entries in each database, we investigated if our results hold on a larger comparison set generated by linking similar ICSD entries. We describe the process of linking similar ICSD entries below and provide tables and figures corresponding to this larger dataset in Section S-VI.

For the process of linking similar ICSD entries, we use the structure comparison and matching algorithm implemented within the Materials Project. The process is involves the following two steps:

**STEP 1. Annotation of extracted data with “ICSD UID”**

1. **Generation of a set of ICSD UIDs from Materials Project:** An ICSD UID is defined as a set of ICSD Collection Codes (ICSD IDs) belonging to the same material. Materials Project already groups ICSD IDs per material but due to the affine mapping based structure matching implemented (J. Appl. Cryst. 39, 6–16 (2006) DOI: 10.1107/S0021889805032450), the material can be part of two different “groups of equivalent structures”. For example, there are three different records for AgO in Materials Project that share the ICSD ID 60625. This task in the curation pipeline thus aggregates the ICSD IDs from all three such records to generate a super “ICSD UID”. For instance, GaN (mp-830) was associated with 10 ICSD IDs, and the corresponding ICSD UID is the set of all 10 ICSD IDs: “187047–190412–67781–156260–41546–157511–248504–191770–185155–290614”.

2. **Annotation of extracted records with an ICSD UID:** First, every record retrieved from Materials Project is matched to an ICSD UID from the set above, depending on the ICSD IDs in the record (the set of ICSD IDs in the record will be a subset of exactly one ICSD UID). Second, it is attempted to match every record retrieved from AFLOW and OQMD to ICSD UID. If no match is found, the ICSD ID in such a record constitutes a new ICSD UID.

The resulting extracted properties dictionary has the following format:

{
STEP 2. Reverse-mapping properties to ICSD UIDs

This step involves “inverting” the dictionary of extracted properties above such that the ICSD UIDs are the keys and a list of data records corresponding to each UID is the value. The inverted dictionary at this step has the following format:

```json
...
... {
  "AFLOW": {
    "icsd_id_1-icsd_id_6-icsd_id_7": [
      {
        "db_id": 1,
        "prop_name_1": prop_val_1,
        "prop_name_2": prop_val_2,
        "icsd_uid": icsd_id_1-icsd_id_6-icsd_id_7,
      },
      {
        "db_id": 3,
      
      },
      ...
    ],
    "icsd_id_2-icsd_id_3": [
      {
        "db_id": 2,
      
      },
      ...
    ],
  },
  "MP": {
    ...
  }
}
```
D. Data Curation

1. Removing composition inconsistencies

From the ICSD UID to properties dictionary above, an ICSD UID key is removed if the entries within it do not have matching compositions. This process is done first within each of the three databases (discarded entries in Table S-IV), and then for UIDs common to pair-wise combinations of the databases (discarded entries in Table S-V). Most records filtered out at this step are materials with different number of H and Li atoms (e.g., BaGaH$_4$ vs BaGaH$_5$) or small changes in composition (e.g., Y$_3$Fe$_{29}$ vs Y$_3$Fe$_{31}$).

2. Filtering for the lowest energy entry per ICSD UID

For each ICSD UID, since there may exist multiple entries (calculations) in every database, only the entry with the lowest total\_energy\_per\_atom value is retained from this step onward.

3. Removing records with unphysical properties

At this step, any records with unphysical values of certain properties are removed. This includes all boride formation energies from AFLOW, due to an error in the B chemical potential (this bug was discovered in the course of this work and confirmed by the AFLOW developers [33]). Beyond AFLOW borides, unphysical properties are defined as per-atom formation energies outside $-5$ to $+5$ eV/atom, per-atom volumes above 150 Å$^3$/atom, for all three databases (entries discarded at this step in Table S-VI).

4. Converting magnetizations into absolute values

Finally, all total\_magnetization\_per\_atom values in all three databases are converted into absolute values.

When querying the Materials Project for total magnetization via the RESTful API, the value returned is not the total magnetic moment of the unit cell, as documented [34], but rather the per formula unit value. This issue has been communicated to the Materials Project development team. The magnetization values queried from Materials Project were normalized suitably.

Lastly, we identify the largest outliers for each property in each pairwise database comparison post-curation: formation energy in Table S-VII, volume in Table S-VIII, band gap in Table S-IX, and total magnetization in Table S-X.
TABLE S-IV. Instances where composition did not match for records within a database.

| Database | Compositions                                  | ICSD UID |
|----------|-----------------------------------------------|----------|
| AFLOW    | Ba$_2$CeCl$_7$, AgSbTe$_2$, YIO, Bi, Ba$_2$LaCl$_7$, Ba$_2$YCl$_7$, YBrO | 0        |
| AFLOW    | Cr$_2$Te$_4$O$_{11}$, LiB                    | 1        |
| AFLOW    | SrGaH$_5$, SrGaH$_4$                         | 240697   |
| AFLOW    | BaGaH$_4$, BaGaH$_5$                         | 240693   |
| OQMD     | ZnCoCuAg, LiMgSnPd                           | 16478    |
| Databases       | Compositions                                                                 | ICSD UID |
|-----------------|------------------------------------------------------------------------------|----------|
| AFLOW-MP        | LaHO₂, LaO₂                                                                  | 60675    |
| AFLOW-MP        | XeSb₂F₁₀, Xe₂Sb₂F₁₉                                                          | 157664   |
| AFLOW-MP        | Ba₃P₂HO₁₃, Ba₃P₂O₁₃                                                          | 62283    |
| AFLOW-MP        | Li₂ScP₂HO₈, Li₂Sc(PO₄)₂                                                        | 409955   |
| AFLOW-MP        | KAl₂P₂H₅O₁₁, KAl₂P₂H₄O₁₁                                                       | 407355   |
| AFLOW-MP        | RbH₅O, Rb₂O₂                                                                  | 61048    |
| AFLOW-MP        | H₂O₂BiH₈, H₂O₂BiO₈                                                            | 413928   |
| AFLOW-MP        | Rb₃H(SO₄)₂, Rb₃(SO₄)₂                                                         | 60050    |
| AFLOW-MP        | Ta₃Al₁HO₁₄, Ta₃Al₁O₁₄                                                         | 67673    |
| AFLOW-MP        | Gd₂CBr₂, GdBr                                                                | 47226    |
| AFLOW-MP        | HO₅F, O₃I                                                                    | 26621    |
| AFLOW-MP        | LiZnBi, ZnBi                                                                 | 100115   |
| AFLOW-MP        | Y₃Fe₃I, Y₃Fe₂₃                                                             | 107259   |
| AFLOW-MP        | RbO₉H₂O₉, RbO₉O₉                                                             | 20611    |
| AFLOW-MP        | Ba₅Re₂O₁₇, Ba₅Re₂O₁₆                                                          | 100777   |
| AFLOW-MP        | Na₅VP₂O₉₂, Na₅VP₂O₉₂                                                         | 50760    |
| AFLOW-MP        | Na₅TiHF₄, Na₅TiF₈                                                            | 14131    |
| AFLOW-MP        | C₂₂NH₂₂S₂(O₂F₃)₂, C₂₂NS₂(O₂F₃)₂                                               | 50524    |
| AFLOW-MP        | BaHF₃, BaF₃                                                                   | 35409    |
| AFLOW-MP        | HgH₂O₄Cl, HgO₄Cl                                                             | 29038    |
| AFLOW-MP        | LiTa₃(Bi₂O₇)₂, Ta₃(Bi₂O₇)₂                                                     | 415141   |
| AFLOW-MP        | YH₂O₂, YO₂                                                                   | 28442    |
| AFLOW-MP        | Dy₂B₂H₈O₈, Dy₂B₂O₈                                                            | 413927   |
| AFLOW-MP        | CaHOCI₃, CaOCI                                                                | 24403    |
| AFLOW-MP        | NdAl₂, Nd₂Al                                                                   | 608745   |
| AFLOW-MP        | Ga(Bi₄O₇)₃, Ga(Bi₃O₅)₄                                                        | 68648    |
| AFLOW-MP        | HoH₂O₂, HoO₂                                                                  | 2944     |
| AFLOW-MP        | Y₁₂(ReC₃)₅, Y₁₂Re₅C₆                                                         | 658805   |
| AFLOW-MP        | SnPHO₃, SnPO₃                                                                  | 25034    |
| AFLOW-MP        | LiCH₃O₃, LiCH₂O₃                                                               | 109604   |
| AFLOW-MP        | CaCu₃AsH₂O₅, CaCu₃AsO₅                                                       | 64694    |
| AFLOW-MP        | K₂Cr₂AsH₂O₁₀, K₂Cr₂AsO₁₀                                                       | 30533    |
| AFLOW-MP        | SrHF₃, SrF₃                                                                   | 35408    |
| AFLOW-MP        | NdMoH₂O₅I₄, I₄                                                               | 281173   |
| AFLOW-MP        | BaAl₃H₂O₉, BaAl₅O₉                                                             | 33282    |
| AFLOW-MP        | SrNH, SrN                                                                     | 410656   |
| AFLOW-MP        | Si₃N₃H, Si₃N₃                                                                  | 202970   |
| AFLOW-MP        | La₆TaH₂(O₂C)₃, La₆Ta(O₂Cl)₃                                                   | 62189    |
| AFLOW-MP        | NaCHO₂, NaCO₂                                                                  | 109643   |
| AFLOW-MP        | LaNHO₄, LaNO₄                                                                  | 413563   |
| AFLOW-MP        | Ba₅Cr₃H₂O₁₃, Ba₅Cr₃O₁₃                                                        | 21034    |
| AFLOW-MP        | H₂O₇, H₂O₂                                                                   | 27844    |
| AFLOW-MP        | CdHOCI₃, CdOCI                                                                 | 26752    |
| AFLOW-MP        | Ti₃H₂S₂O₉, Ti₃S₂O₉                                                             | 35358    |
| AFLOW-MP        | CdNH₄O₄, CdNO₄                                                                  | 35355    |
| AFLOW-MP        | Mn₃P₂H₂O₇, Mn₃P₂O₇                                                             | 415152   |
| AFLOW-OQMD      | LiB, MnCoSiO₄                                                                  | 2        |
| AFLOW-OQMD      | Y₃Fe₃I, Y₃Fe₂₃                                                               | 107259   |
| AFLOW-OQMD      | Ga(Bi₄O₇)₃, Ga(Bi₃O₅)₄                                                       | 68648    |
| AFLOW-OQMD      | H₂O₂, H₂O₂                                                                   | 27844    |
| MP-OQMD         | Bi₂(PO₅)₂, SmPd                                                                | 107679   |
| MP-OQMD         | SrGaH₅, SrGaH₄                                                                  | 240697   |
| MP-OQMD         | BaGaH₅, BaGaH₄                                                                  | 240693   |
| Database          | Composition | Property value | ICSD UID |
|-------------------|-------------|----------------|----------|
| AFLOW             | SiO₂        | −11.950        | 170547   |
| MP                | Ta          | 5.113          | 54207    |
| OQMD              | CuO₂        | 1126.321       | 51426    |
| OQMD              | SiO₂        | 84.972         | 155252   |
| OQMD              | Sc₅Ga₃      | 8.575          | 165189   |
| Volume (Å³/atom)  |             |                |          |
| MP                | HoVO₄       | 364.707        | 152694   |
| MP                | HgH₂IO₆     | 188.228        | 409499   |
| MP                | Rb          | 523.351        | 109016   |
| MP                | CaC₂        | 896.388        | 252718   |
| MP                | EuAg        | 962.227        | 58257    |
| MP                | SnSe        | 182.422        | 52425    |
| MP                | Na          | 349.842        | 70067    |
| MP                | TeB         | 733.428        | 168896   |
| MP                | K           | 433.990        | 157565   |
| MP                | TiCl₂       | 386.012        | 20762    |
| MP                | Ta          | 309.658        | 54207    |
| MP                | Fe₃C        | 217.282        | 76827    |
| MP                | Cu          | 603.475        | 150682   |

TABLE S-VI. Records with physically unreasonable values of formation energy (outside the [−5, +5] eV/atom window) and/or volume (>150 Å³/atom).

| Databases       | Composition | ΔE₁^f (eV/atom) | ΔE₂^f (eV/atom) | Δ¹⁻² ICSD UID |
|-----------------|-------------|-----------------|-----------------|---------------|
| AFLOW-MP        | AlPO₄       | −2.780          | −0.307          | −2.474        | 162670        |
| AFLOW-MP        | SiO₂        | −1.041          | −2.870          | 1.829         | 25632         |
| AFLOW-MP        | O           | 0.004           | 1.669           | −1.665        | 92775         |
| AFLOW-MP        | AlClO       | −1.741          | −2.789          | 1.048         | 27812         |
| AFLOW-MP        | Bi          | 0.135           | 1.090           | −0.954        | 51675         |
| AFLOW-MP        | YAl₃        | 0.515           | −0.437          | 0.952         | 58220         |
| AFLOW-MP        | PCl₅        | −0.192          | −1.061          | 0.869         | 76731         |
| AFLOW-MP        | K₂BeO₂      | −1.259          | −2.124          | 0.865         | 23633         |
| AFLOW-MP        | Li₄P₂O₇     | −1.899          | −2.762          | 0.863         | 39814         |
| AFLOW-MP        | RbC₈        | 0.751           | −0.030          | 0.781         | 200563        |

| AFLOW-OQMD      | AlPO₄       | −2.780          | −0.237          | −2.543        | 162670        |
| AFLOW-OQMD      | O           | 0.004           | 1.395           | −1.391        | 92775         |
| AFLOW-OQMD      | MgO         | −2.877          | −1.879          | −0.998        | 181459        |
| AFLOW-OQMD      | AlClO       | −1.741          | −2.608          | 0.867         | 27812         |
| AFLOW-OQMD      | RbC₈        | 0.751           | −0.040          | 0.791         | 200563        |
| AFLOW-OQMD      | K₂BeO₂      | −1.259          | −2.002          | 0.743         | 23633         |
| AFLOW-OQMD      | SrP₂S₃      | −0.277          | −1.016          | 0.739         | 412766        |
| AFLOW-OQMD      | Li₄P₂O₇     | −1.899          | −2.634          | 0.735         | 39814         |
| AFLOW-OQMD      | PCl₅        | −0.192          | −0.908          | 0.716         | 76731         |
| AFLOW-OQMD      | Mg₅P₂O₈     | −2.346          | −2.930          | 0.584         | 9849         |

| MP-OQMD         | MnNiAs      | 3.547           | −0.249          | 3.796         | 161716        |
| MP-OQMD         | Nd₃PbN      | −0.945          | 1.319           | −2.264        | 76397         |
| MP-OQMD         | IrC₄        | 0.981           | 3.161           | −2.180        | 181498        |
| MP-OQMD         | Nd₃SnN      | −1.029          | 1.112           | −2.141        | 76398         |
| MP-OQMD         | Sm₃AlN      | −0.894          | 1.109           | −2.003        | 52640         |
| MP-OQMD         | LiNO₃       | 0.690           | −1.252          | 1.942         | 33661         |
| MP-OQMD         | Pr₃AlN      | −0.788          | 1.101           | −1.888        | 52639         |
| MP-OQMD         | Cr₃O        | 1.971           | 0.141           | 1.830         | 15904         |
| MP-OQMD         | Eu(Cu₂Sn)₂  | 1.296           | −0.322          | 1.617         | 416796        |
| MP-OQMD         | GdMg₂Ag     | 1.502           | −0.102          | 1.605         | 107733        |

TABLE S-VII. Top ten outliers in calculated formation energy across pairwise comparisons of databases.
TABLE S-VIII. Top ten outliers in calculated volume across pairwise comparisons of databases.

| Databases     | Composition | $V^1$ | $V^2$ | $\Delta V^{1-2}$ | ICSD UID |
|---------------|-------------|-------|-------|-------------------|-----------|
| AFLOW-MP      | Bi          | 31.51 | 103.72| $-72.21$         | 51675     |
| AFLOW-MP      | MnNiAs      | 15.70 | 84.29 | $-68.59$         | 161716    |
| AFLOW-MP      | Hg          | 28.86 | 95.21 | $-66.35$         | 79804     |
| AFLOW-MP      | HoS         | 39.85 | 103.31| $-63.46$         | 66357     |
| AFLOW-MP      | SrPSe₃      | 89.79 | 27.94 | $61.85$           | 412766    |
| AFLOW-MP      | NbTeBr₃     | 75.01 | 34.02 | $40.99$           | 35376     |
| AFLOW-MP      | CeBr₃       | 67.72 | 31.52 | $36.20$           | 31582     |
| AFLOW-MP      | Se          | 63.04 | 31.90 | $31.14$           | 150731    |
| AFLOW-MP      | FeSeBr₇     | 66.74 | 37.16 | $29.58$           | 39528     |
| AFLOW-MP      | SnSe        | 27.09 | 56.42 | $-29.34$          | 71338     |
| AFLOW-OQMD    | SrPSe₃      | 89.79 | 27.65 | $62.14$           | 412766    |
| AFLOW-OQMD    | NbTeBr₃     | 75.01 | 30.71 | $44.30$           | 35376     |
| AFLOW-OQMD    | CeBr₃       | 67.72 | 31.01 | $36.71$           | 31582     |
| AFLOW-OQMD    | H            | 37.20 | 2.73  | $34.47$           | 28465     |
| AFLOW-OQMD    | CsTl        | 14.22 | 47.71 | $-33.49$          | 165344    |
| AFLOW-OQMD    | Se          | 63.04 | 31.39 | $31.65$           | 150731    |
| AFLOW-OQMD    | FeSeBr₇     | 66.74 | 37.22 | $29.52$           | 39528     |
| AFLOW-OQMD    | Ca₃PSe₃     | 53.48 | 25.50 | $27.98$           | 412765    |
| AFLOW-OQMD    | SnSe        | 27.09 | 52.60 | $-25.52$          | 71338     |
| AFLOW-OQMD    | TiO₂        | 35.12 | 11.39 | $23.73$           | 97008     |
| MP-OQMD       | Bi          | 103.72| 31.53 | $72.20$           | 51675     |
| MP-OQMD       | MnNiAs      | 84.29 | 14.74 | $69.55$           | 161716    |
| MP-OQMD       | Hg          | 95.21 | 27.10 | $68.11$           | 79804     |
| MP-OQMD       | CoO₂        | 57.55 | 11.71 | $45.84$           | 89837     |
| MP-OQMD       | H₂          | 43.46 | 5.38  | $38.07$           | 28344     |
| MP-OQMD       | Cd₆Sb₅      | 62.94 | 31.30 | $31.65$           | 52832     |
| MP-OQMD       | Cl₂         | 70.98 | 41.47 | $29.50$           | 22406     |
| MP-OQMD       | LiBH₄       | 37.31 | 9.87  | $27.44$           | 168803    |
| MP-OQMD       | HfPd₅       | 41.53 | 15.69 | $25.84$           | 168289    |
| MP-OQMD       | Xe          | 83.51 | 58.86 | $24.65$           | 9786      |
| Databases       | Composition                  | $E_1^g$ | $E_2^g$ | $\Delta E^{1-2}$ | ICSD UID |
|-----------------|------------------------------|---------|---------|-----------------|----------|
| AFLOW-MP        | CeF$_3$                     | 5.87    | 0.00    | 5.87            | 42470    |
| AFLOW-MP        | LiAlPHO$_5$                 | 0.02    | 5.72    | -5.70           | 68921    |
| AFLOW-MP        | Na$_2$PHO$_3$               | 0.00    | 5.58    | -5.58           | 155976   |
| AFLOW-MP        | KCeF$_4$                    | 5.44    | 0.00    | 5.44            | 23229    |
| AFLOW-MP        | RbYbF$_3$                   | 1.09    | 6.52    | -5.43           | 40590    |
| AFLOW-MP        | K$_2$NaCe$_2$S$_6$O$_{24}$  | 5.49    | 0.08    | 5.41            | 281576   |
| AFLOW-MP        | CsYbF$_3$                   | 1.78    | 7.05    | -5.26           | 49579    |
| AFLOW-MP        | BaTm$_2$F$_8$               | 1.97    | 7.24    | -5.26           | 20103    |
| AFLOW-MP        | CePO$_4$                    | 5.23    | 0.00    | 5.23            | 184550   |
| AFLOW-MP        | Mg$_3$P$_2$O$_8$            | 0.00    | 5.18    | -5.18           | 9849     |
| AFLOW-OQMD      | BeF$_2$                     | 8.04    | 0.00    | 8.04            | 173557   |
| AFLOW-OQMD      | KYb$_3$F$_{10}$             | 0.97    | 8.44    | -7.47           | 28258    |
| AFLOW-OQMD      | NaCaAlF$_6$                 | 7.12    | 0.00    | 7.12            | 80542    |
| AFLOW-OQMD      | H$_2$OsC$_8$N$_2$F$_6$      | 0.00    | 6.68    | -6.68           | 151185   |
| AFLOW-OQMD      | YbCl$_2$O$_{12}$            | 0.00    | 6.38    | -6.38           | 85762    |
| AFLOW-OQMD      | LiAlPHO$_5$                 | 0.02    | 6.20    | -6.18           | 68921    |
| AFLOW-OQMD      | LiEuP$_4$O$_{12}$           | 0.33    | 6.41    | -6.08           | 416878   |
| AFLOW-OQMD      | Na$_2$PHO$_3$               | 0.00    | 6.07    | -6.07           | 155976   |
| AFLOW-OQMD      | CsSiH$_2$O$_6$F$_6$         | 0.00    | 5.96    | -5.96           | 2900     |
| AFLOW-OQMD      | CsYbF$_3$                   | 1.78    | 7.73    | -5.94           | 49579    |
| MP-OQMD         | KYb$_3$F$_{10}$             | 0.00    | 8.44    | -8.44           | 28258    |
| MP-OQMD         | BeF$_2$                     | 7.96    | 0.00    | 7.96            | 173557   |
| MP-OQMD         | KCeF$_4$                    | 0.00    | 7.75    | -7.75           | 23229    |
| MP-OQMD         | H$_2$                      | 0.00    | 7.22    | -7.22           | 28539    |
| MP-OQMD         | NaCaAlF$_6$                 | 7.11    | 0.00    | 7.11            | 80542    |
| MP-OQMD         | EuMgF$_4$                   | 0.29    | 7.22    | -6.94           | 86246    |
| MP-OQMD         | CsEuF$_3$                   | 0.00    | 6.93    | -6.93           | 49577    |
| MP-OQMD         | H$_2$OsC$_8$(NF$_3$)$_2$    | 0.17    | 6.68    | -6.51           | 151185   |
| MP-OQMD         | Yb(CIO$_4$)$_3$             | 0.00    | 6.46    | -6.46           | 85763    |
| MP-OQMD         | LiEu(PO$_4$)$_4$            | 0.00    | 6.41    | -6.41           | 416878   |
| Databases          | Composition | $M^1$ | $M^2$ | $\Delta^{1-2}$ | ICSD UID |
|--------------------|-------------|-------|-------|----------------|----------|
| AFLOW-MP           | YbMn$_{28}$| 124.89| 7.16  | 117.72         | 643923   |
| AFLOW-MP           | Pr$_6$Mn$_{23}$| 121.85| 32.42 | 89.43          | 643337   |
| AFLOW-MP           | BaMn$_{28}$ | 87.51 | 8.61  | 78.90          | 615966   |
| AFLOW-MP           | Yb$_6$Mn$_{23}$| 105.23| 29.18 | 76.05          | 643920   |
| AFLOW-MP           | Ba$_{6}$Co$_{25}$S$_{27}$| 66.99 | 0.00  | 66.99          | 71939    |
| AFLOW-MP           | Mn$_{20}$W$_3$C$_6$| 74.10 | 16.64 | 57.46          | 618270   |
| AFLOW-MP           | Pr$_6$Mn$_{23}$| 73.37 | 16.53 | 56.84          | 618260   |
| AFLOW-MP           | Gd$_7$Rh$_5$C$_{12}$| 56.95 | 0.15  | 56.80          | 617956   |
| AFLOW-MP           | Nd$_{12}$Co$_6$Sn | 56.33 | 0.00  | 56.33          | 240094   |
| AFLOW-MP           | Yb$_6$Co$_{30}$P$_{19}$| 56.51 | 1.12  | 55.39          | 67950    |
| AFLOW-OQMD         | EuMn$_{28}$ | 125.94| 4.13  | 121.82         | 631390   |
| AFLOW-OQMD         | Gd$_{13}$Ge$_{6}$O$_{31}$F| 90.97 | 0.01  | 90.96          | 62329    |
| AFLOW-OQMD         | Pr$_6$Mn$_{23}$| 121.85| 33.03 | 88.82          | 643337   |
| AFLOW-OQMD         | BaMn$_{28}$ | 87.51 | 5.58  | 81.93          | 615966   |
| AFLOW-OQMD         | Yb$_6$Mn$_{23}$| 105.23| 30.78 | 74.45          | 643920   |
| AFLOW-OQMD         | Ba$_{6}$Co$_{25}$S$_{27}$| 66.99 | 2.23  | 64.76          | 71939    |
| AFLOW-OQMD         | Nd$_{12}$Co$_6$Sn | 56.33 | 0.01  | 56.32          | 240094   |
| AFLOW-OQMD         | Gd$_7$Pd$_3$ | 54.76 | 0.89  | 53.86          | 104112   |
| AFLOW-OQMD         | Gd$_7$CoI$_{12}$| 51.02 | 0.00  | 51.01          | 245279   |
| AFLOW-OQMD         | ThMn$_{12}$ | 52.10 | 2.25  | 49.86          | 104986   |
| MP-OQMD            | Gd$_{13}$Ge$_{6}$O$_{31}$F| 91.00 | 0.01  | 90.99          | 62329    |
| MP-OQMD            | ZnFe$_{16}$Ni$_{17}$O$_{32}$| 38.00 | 94.00 | −56.00         | 182238   |
| MP-OQMD            | Eu$_7$Au$_3$ | 50.72 | 0.01  | 50.71          | 611842   |
| MP-OQMD            | Gd$_6$Zn$_{23}$| 42.88 | 0.01  | 42.87          | 636504   |
| MP-OQMD            | Gd$_6$C$_3$Cl$_5$| 42.69 | 0.00  | 42.68          | 202547   |
| MP-OQMD            | Gd$_{10}$S$_{19}$| 42.00 | 0.00  | 42.00          | 416804   |
| MP-OQMD            | Ba$_8$Eu$_7$Cl$_{34}$| 45.00 | 4.02  | 40.98          | 408479   |
| MP-OQMD            | Mn$_{8}$Au$_{31}$| 0.28 | 39.05 | −38.77         | 58552    |
| MP-OQMD            | Eu$_5$Pd$_2$ | 36.02 | 0.00  | 36.02          | 631525   |
| MP-OQMD            | Eu$_5$Pt$_2$ | 35.94 | 0.00  | 35.94          | 631557   |
FIG. S1. Distribution of compounds in the MP-OQMD comparison set, near the peak at $\Delta E_f = 0.2 \text{ eV/atom}$ in the histogram of formation energy differences. Only compounds containing the top few most-frequently-occurring elements are shown.
**FIG. S2.** Median absolute differences between properties (formation energy, volume, band gap, total magnetization are in units of eV/atom, Å\(^3\)/atom, eV, and µB/formula unit, respectively) calculated in the three databases (AFLOW, MP, OQMD), compared pairwise, across various classes of materials as defined in Table III of the main text. The numbers in parentheses indicate the number of overlapping records belonging to the respective material class for a given pair of databases. Trivial comparisons are left blank (e.g., the difference in total magnetization for non-magnetic compounds).
To study the source of differences between the various HT-DFT databases, we collect statistics for the four properties being compared—formation energy, volume, band gap, total magnetization—averaged over all records containing a certain element in the periodic table. For each element, we also present the pseudopotential (psp) used in the two databases being compared, and the number of comparable records of compounds containing the element over which statistical quantities are computed. Note that any missing elemental block in the periodic tables in Figures S3–S14 implies that there were no materials in that comparison set with that element.

FIG. S3. Median values of formation energy for compounds containing a certain element in the periodic table, for a comparison of AFLOW and MP. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).
FIG. S4. Median values of per-atom volume for compounds containing a certain element in the periodic table, for a comparison of AFLOW and MP. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).

FIG. S5. Median values of band gap for compounds containing a certain element in the periodic table, for a comparison of AFLOW and MP. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).
FIG. S6. Median values of total magnetization (per formula unit) for compounds containing a certain element in the periodic table, for a comparison of AFLOW and MP. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).

### AFLOW-MP | Total magnetization

| DB-1 | DB-2 |
|------|------|
| psp-1 | x-1 | psp-2 | x-2 |
| # records |

| Element | Total magnetization (Bohr magneton/f.u.) |
|---------|----------------------------------------|
| Li      | 3.00 |
| Al      | 3.00 |

FIG. S7. Median values of formation energy for compounds containing a certain element in the periodic table, for a comparison of AFLOW and OQMD. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).

### AFLOW-OQMD | Formation energy

| psp-1 | psp-2 |
|-------|-------|
| # records |

| Element | Formation energy (eV/atom) |
|---------|---------------------------|
| Li      | -2.32 |
| Al      | -2.16 |
FIG. S8. Median values of per-atom volume for compounds containing a certain element in the periodic table, for a comparison of AFLOW and OQMD. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).

FIG. S9. Median values of band gap for compounds containing a certain element in the periodic table, for a comparison of AFLOW and OQMD. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).
FIG. S10. Median values of total magnetization (per formula unit) for compounds containing a certain element in the periodic table, for a comparison of AFLOW and OQMD. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).

FIG. S11. Median values of formation energy for compounds containing a certain element in the periodic table, for a comparison of MP and OQMD. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).
FIG. S12. Median values of per-atom volume for compounds containing a certain element in the periodic table, for a comparison of MP and QM. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).

FIG. S13. Median values of band gap for compounds containing a certain element in the periodic table, for a comparison of MP and QM. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).
FIG. S14. Median values of total magnetization (per formula unit) for compounds containing a certain element in the periodic table, for a comparison of MP and OQMD. The VASP PAW potential used for each element and the number of records in each comparison are indicated (* indicates more than one pseudopotential used in the database overall for that element).
S-VI. RESULTS FROM THE LARGER, MULTIPLE ICSD ID COMPARISON DATASET

TABLE S-XI. The number of records after establishing ICSD ID equivalency for each property of interest in the AFLOW, Materials Project (MP), and OQMD HT-DFT databases, as well as for pairwise comparisons of the three databases.

| Property            | AFLOW | MP      | OQMD   | AFLOW-MP | AFLOW-OQMD | MP-OQMD |
|---------------------|-------|---------|--------|----------|------------|---------|
| Formation Energy    | 3411  | 48515   | 33281  | 3285     | 2827       | 30115   |
| Volume              | 32738 | 48515   | 33281  | 30067    | 25593      | 30115   |
| Band Gap            | 32727 | 48515   | 33141  | 30059    | 25466      | 29979   |
| Total Magnetization | 32738 | 48515   | 33281  | 30067    | 25593      | 30115   |

TABLE S-XII. Overall statistics (median absolute difference (MAD), interquartile range (IQR), Pearson’s linear correlation coefficient \(r\), and Spearman’s rank correlation coefficient \(\rho\)) for the comparison of properties across HT-DFT databases. For each property, records overlapping across a pair of databases are compared (* for band gap and magnetization, only non-zero values are compared). Generally, lower MAD, lower IQR, higher \(r\), and higher \(\rho\) values indicate better reproducibility of calculated properties.

| Property            | AFLOW-MP | AFLOW-OQMD | MP-OQMD |
|---------------------|----------|------------|---------|
|                     | MAD  | IQR | \(r\) | \(\rho\) | MAD  | IQR | \(r\) | \(\rho\) | MAD  | IQR | \(r\) | \(\rho\) |
| Formation Energy (eV/atom) | 0.090 | 0.171 | 0.99 | 0.99 | 0.017 | 0.028 | 0.99 | 0.99 | 0.076 | 0.164 | 0.99 | 0.99 |
| Volume (Å³/atom)    | 0.187 | 0.417 | 0.98 | 0.99 | 0.606 | 1.093 | 0.97 | 0.97 | 0.441 | 0.794 | 0.98 | 0.98 |
| Band Gap (eV)       | 0.077 | 0.201 | 0.94 | 0.92 | 0.211 | 0.367 | 0.93 | 0.92 | 0.180 | 0.277 | 0.94 | 0.93 |
| Total Magnetization (µB/f.u.)* | 0.056 | 1.024 | 0.76 | 0.74 | 0.419 | 2.267 | 0.67 | 0.56 | 0.019 | 0.061 | 0.83 | 0.74 |
FIG. S15. Distribution of the differences in calculated properties across HT-DFT databases. Each panel corresponds to a property and pair of databases being compared. Solid vertical black lines correspond to the first (Q₁) and third (Q₃) quartiles of the distribution. The number of records overlapping across the two databases is shown in the top right corner of each panel; the median of distribution (\(\tilde{x}\)), the median absolute difference (MAD), and the interquartile range (IQR) are noted on the left. Note that this figure represents data from the larger comparison datasets obtained via the structure matching algorithm described in Section S-II.C2.
FIG. S16. Comparison of the calculated properties (formation energy, volume, band gap, and total magnetization) over records overlapping across pairwise combinations of HT-DFT databases plotted as a percentile rank. Overall, formation energies and volumes show better reproducibility than band gaps and magnetizations. The clusters seen in the magnetization comparisons correspond to nominally integer values of magnetic moments. Note that this figure represents data from the larger comparison datasets obtained via the structure matching algorithm described in Section S-II C2.
FIG. S17. Median percent absolute differences between properties (formation energy, volume, band gap, total magnetization) calculated in the three databases (AFLOW, MP, OQMD), compared two at a time, across various classes of materials as defined in Table III of the main text. The numbers in parentheses indicate the number of overlapping records belonging to the respective material class for a given pair of databases. Trivial comparisons are left blank (e.g., the difference in total magnetization for non-magnetic compounds). Note that this figure represents data from the larger comparison datasets obtained via the structure matching algorithm described in Section S-II C2.

| Property          | All | Oxide | Nitride | Pnicotide | Chalcogenide | Halide | Alkali Metal | Alkaline Earth Metal | Transition Metal | Metalloid | Rare-Earth | Actinide | Metal-Nonmetal | Intermetallic | Magnetic | Non-Magnetic | Disagree On Magnetic | Metallic | Semiconductor | Insulator | Disagree On Metallic | Pseudopotentials Agree | Pseudopotentials Disagree | Use GGA+U | Use GGA | Disagree on GGA/GGA+U | Element | Binary | Ternary | Quaternary | Triclinic | Monoclinic | Orthorhombic | Tetragonal | Trigonal | Hexagonal | Cubic |
|-------------------|-----|-------|---------|-----------|------------|--------|-------------|----------------------|------------------|-----------|------------|---------|--------------|-------------|----------|-------------|------------------------|---------|-------------|---------|-------------|-------------|---------------------|---------|---------|----------------|--------|--------|--------|-----------|---------|---------|-----------|----------|---------|----------|------|
| $\Delta E_f$      | 5.5 | 6.2   | 6.6     | 5.6       | 4.8        | 3.5    | 2.6         | 1.9                  | 4.6              | 5.3       | 2.5        | 9.7     | 8.7         | 5.8        | 5.8      | 5.5         | 5.1                  | 1.9     | 5.2        | 5.9      | 15.8       | 5.4       | 5.8      | 0.5         | 8.1     | 4.8      | 5.6     | 6.1       | 5.5      | 4.6     | 4.9        |
| $V$               | 1.0 | 3.6   | 1.7     | 0.5       | 1.6        | 1.3    | 2.0         | 0.5                  | 1.4              | 0.9       | 3.6        | 9.7     | 2.6         | 4.3        | 0.1      | 0.2         | 0.1                  | 1.4     | 0.7        | 0.9      | 0.9         | 0.4       | 0.5      | 0.3         | 1.8     | 1.9      | 0.9     | 3.5       | 0.9      | 1.5     | 1.3        |
| $E_g$             | 4.0 | 0.6   | 1.7     | 0.0       | 1.1        | 0.2    | 2.0         | 0.0                  | 1.4              | 0.9       | 3.6        | 9.7     | 2.6         | 4.3        | 0.1      | 0.2         | 0.1                  | 1.4     | 0.7        | 0.9      | 0.9         | 0.4       | 0.5      | 0.3         | 1.8     | 1.9      | 0.9     | 3.5       | 0.9      | 1.5     | 1.3        |
| $M$               | 1.9 | 1.0   | 1.0     | 0.0       | 1.0        | 0.0    | 1.0         | 0.0                  | 1.0              | 1.0       | 1.0        | 1.0     | 1.0         | 1.0        | 1.0      | 1.0         | 1.0                  | 1.0     | 1.0        | 1.0      | 1.0         | 1.0       | 1.0      | 1.0         | 1.0     | 1.0      | 1.0     | 1.0       | 1.0      | 1.0     | 1.0        |

FIG. S17. Median percent absolute differences between properties (formation energy, volume, band gap, total magnetization) calculated in the three databases (AFLOW, MP, OQMD), compared two at a time, across various classes of materials as defined in Table III of the main text. The numbers in parentheses indicate the number of overlapping records belonging to the respective material class for a given pair of databases. Trivial comparisons are left blank (e.g., the difference in total magnetization for non-magnetic compounds). Note that this figure represents data from the larger comparison datasets obtained via the structure matching algorithm described in Section S-II C2.
FIG. S18. Median absolute differences between properties (formation energy, volume, band gap, total magnetization are in units of eV/atom, Å³/atom, eV, and µB/formula unit, respectively) calculated in the three databases (AFLOW, MP, OQMD), compared pairwise, across various classes of materials as defined in Table III of the main text. The numbers in parentheses indicate the number of overlapping records belonging to the respective material class for a given pair of databases. Trivial comparisons are left blank (e.g., the difference in total magnetization for non-magnetic compounds). Note that this figure represents data from the larger comparison datasets obtained via the structure matching algorithm described in Section S-IIC2.
S-VII. EXAMPLE PHYSICAL INFORMATION FILES (PIFS)

Example (serialized) PIFs with data queried from each of the three HT-DFT databases are shown below.

A. Example PIF with queried AFLOW data in the JSON format

```json
{
  "category": "system.chemical",
  "tags": [
    "AFLOW",
    "ICSD"
  ],
  "ids": [
    {
      "name": "AUID",
      "value": "aflow:beda768ce32fec75"
    },
    {
      "name": "ICSD",
      "value": 409808
    }
  ],
  "references": [
    {
      "url": "http://aflow.org/material.php?id=aflow:beda768ce32fec75"
    },
    {
      "url": "http://aflowlib.duke.edu/AFLOWDATA/ICSD_WEB/FCC/C16Cs2Mo1_ICSD_409808"
    },
    {
      "doi": "http://dx.doi.org/10.1016/j.commatsci.2012.02.002"
    }
  ],
  "chemicalFormula": "Cl6 Cs2 Mo1",
  "properties": [
    {
      "scalars": "aflow:beda768ce32fec75",
      "name": "auid"
    },
    {
      "scalars": "aflowlib.duke.edu:AFLOWDATA/ICSD_WEB/FCC/C16Cs2Mo1_ICSD_409808",
      "name": "aurl"
    },
    {
      "scalars": "Cl6Cs2Mo1",
      "name": "compound"
    },
    {
      "scalars": "6,2,1",
      "name": "composition"
    },
    {
      "scalars": "9",
      "name": "natoms"
    },
    {
      "scalars": "PAW_PBE",
      "name": "basis_set"
    }
  ]
}
```
"name": "dft_type"
},
{
"scalars": "-34.4871",
"units": "eV",
"name": "energy_cell"
},
{
"scalars": "-3.83189",
"units": "eV/atom",
"name": "energy_atom"
},
{
"scalars": "299.219",
"units": "$\AA^3$",
"name": "volume_cell"
},
{
"scalars": "33.2465",
"units": "$\AA^3$/atom",
"name": "volume_atom"
},
{
"scalars": "0",
"units": "eV",
"name": "Egap"
},
{
"scalars": "10,10,10;11,11,11;\Gamma-X,X-W,W-K,K-\Gamma,\Gamma-L,L-U,U-W,W-L,L-K,U-X;20",
"name": "kpoints"
},
{
"scalars": "1.99872",
"units": "$\mu_B$",
"name": "spin_cell"
},
{
"scalars": "0.222081",
"units": "$\mu_B$/atom",
"name": "spin_atom"
},
{
"scalars": "Fm-3m #225,Fm-3m #225,Fm-3m #225",
"name": "sg"
},
{
"scalars": "Fm-3m #225,Fm-3m #225,Fm-3m #225",
"name": "sg2"
},
{
"scalars": "225",
"name": "spacegroup_orig"
},
{
"scalars": "225",
"name": "spacegroup_relax"
},
B. Example PIF with queried MP data in the JSON format

```json
{
  "tags": [
    "Materials Project",
    "ICSD"
  ],
  "references": [
    {
      "url": "https://materialsproject.org/materials/mp-540537"
    }
  ]
}
```
{  
  "doi": "http://dx.doi.org/10.1063/1.4812323"
}
],
"ids": [
  {
    "name": "Material",
    "value": "mp-540537"
  },
  {
    "name": "ICSD",
    "value": 1
  }
],
"properties": [
  {
    "name": "last_updated",
    "scalars": "2019-11-14 16:07:39.266000"
  },
  {
    "name": "material_id",
    "scalars": "mp-540537"
  },
  {
    "name": "original_task_id",
    "scalars": "mp-540537"
  },
  {
    "name": "icsd_ids",
    "scalars": [1]
  },
  {
    "name": "pretty_formula",
    "scalars": "Cr2Te4O11"
  },
  {
    "name": "final_energy",
    "scalars": -208.54704595,
    "units": "eV"
  },
  {
    "name": "final_energy_per_atom",
    "scalars": -6.133736645588236,
    "units": "eV/atom"
  },
  {
    "name": "volume",
    "scalars": 539.6736877085604,
    "units": "$\AA^3$"
  },
  {
    "name": "nsites",
    "scalars": 34
  },
  {
    "name": "formation_energy_per_atom",  
}
"scalars": -1.7626992724264714,
"units": "eV/atom"
},
{
"name": "e_above_hull",
"scalars": 0,
"units": "eV/atom"
},
{
"name": "band_gap",
"scalars": 0.5996000000000001,
"units": "eV"
},
{
"name": "is_hubbard",
"scalars": true
},
{
"name": "hubbards",
"scalars": "{"Cr": 3.7, "Te": 0.0, "O": 0.0}"
},
{
"name": "volume_per_atom*",
"scalars": 15.87275552084001,
"units": "$\text{\AA}^3$/atom"
},
{
"name": "total_magnetization_per_formula_unit",
"scalars": 5.9998625,
"units": "$\mu_B$/formula unit"
},
{
"name": "total_magnetization_per_atom*",
"scalars": 0.3529330882352941,
"units": "$\mu_B$/atom"
},
{
"name": "unit_cell_formula*",
"scalars": "Cr4 O22 Te8"
},
{
"name": "reduced_cell_formula*",
"scalars": "Cr2 O11 Te4"
},
{
"tags": ["PBE", "paw"],
"name": "potentials*",
"scalars": ["Cr_pv",
"Te",
"O"
]}
},
{
"name": "spacegroup_number"
"scalars": 14,
{
"name": "spacegroup_symbol",
"scalars": "P2_1/c"
},
{
"name": "crystal_system",
"scalars": "monoclinic"
},
{
"name": "point_group",
"scalars": "2/m"
}
],
"category": "system.chemical",
"chemicalFormula": "Cr2 O11 Te4"
}

C. Example PIF with queried OQMD data in the JSON format

{
"tags": [
"OQMD v1.2",
"ICSD"
],
"references": [
{
"url": "http://oqmd.org/materials/entry/4189"
},
{
"url": "http://oqmd.org/analysis/calculation/1381320"
},
{"doi": "10.1007/s11837-013-0755-4"}
],
"ids": [
{
"name": "Entry",
"value": 4189
},
{
"name": "ICSD",
"value": 23036
},
{
"name": "Calculation",
"value": 1381320
},
{
"name": "FormationEnergy",
"value": 5636096
}
"scalars": -6.34258323861111,
"units": "eV/atom"
},
{
"name": "calculation__magmom_pa",
"scalars": -1.72333333333333e-05,
"units": "$\mu_B$/atom"
},
{
"name": "calculation__converged",
"scalars": true
},
{
"name": "calculation__output__id",
"scalars": 2069016
},
{
"name": "calculation__output__volume",
"scalars": 467.937,
"units": "$\AA^3$"
},
{
"name": "calculation__output__volume_pa",
"scalars": 12.9982,
"units": "$\AA^3$/atom"
},
{
"name": "calculation__output__natoms",
"scalars": 36
},
{
"name": "calculation__output__spacegroup__number",
"scalars": 62
},
{
"name": "formation__delta_e",
"scalars": -0.7234503675,
"units": "eV/atom"
},
{
"tags": [
"v5.2",
"PAW"
]
},
"name": "potentials*",
"scalars": [
"Ni",
"Si",
"Y_sv"
]
},
{
"name": "crystal_system*",
"scalars": "Orthorhombic"
},
{
"name": "is_hubbard",
"scalars": false

{
   "name": "hubbards",
   "scalars": "{"Y": null, "Si": null, "Ni": null}"
}
]
"category": "system.chemical",
"chemicalFormula": "Ni5 Si3 Y1"}
[1] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, et al., AFLLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations, Comput. Mater. Sci. 58, 227 (2012).

[2] M. Widom and M. Mihalkovic, Stability of Fe-based alloys with structure type C6Cr23, J. Mater. Res. 20, 237 (2005).

[3] J. S. Hummels, J. F. Abild-Pedersen, F. Studt, T. Bligaard, and J. K. Nørskov, CatApp: a web application for surface chemistry and heterogeneous catalysis, Angew. Chem. 124, 278 (2012).

[4] R. D. Johnson III, Computational Chemistry Comparison and Benchmark Database, Tech. Rep. (National Institute of Standards and Technology, 1999).

[5] D. D. Landis, J. S. Hummels, S. Nestorov, J. Greeley, M. Dulak, T. Bligaard, J. K. Nørskov, and K. W. Jacobsen, The computational materials repository, Comput. Sci. Eng. 14, 51 (2012).

[6] R. Tran, Z. Xu, D. W. Balachandran Radhakrishnan, W. Sun, K. A. Persson, and S. P. Ong, Surface energies of elemental crystals, Sci. Data 3 (2016).

[7] J. Hachmann, R. Olivares-Amaya, S. Atahan-Evenrek, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway, and A. Aspuru-Guzik, The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid, J. Phys. Chem. Lett. 2, 2241 (2011).

[8] K. Choudhary, K. F. Garrity, A. C. Reid, B. DeCost, A. J. Biacchi, A. R. H. Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, et al., The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, npj Comput. Mater. 6, 1 (2020).

[9] L. Talirz, S. Kumbhar, E. Passaro, A. V. Yakutovich, V. Granata, F. Gargiulo, M. Borelli, M. Uhrin, S. P. Huber, S. Zoupanos, et al., Materials Cloud, a platform for open computational science, Sci. Data 7, 1 (2020).

[10] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, et al., Commentary: Materials Project: A materials genome approach to accelerating materials innovation, APL Mater. 1, 011002 (2013).

[11] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies, Phys. Rev. B 85, 115104 (2012).

[12] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD), JOM 65, 1501 (2013).

[13] P. Gorai, D. Gao, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, Q. Lv, V. Stevanović, and E. S. Toberer, TE Design Lab: A virtual laboratory for thermoelectric material design, Comput. Mater. Sci. 112, 368 (2016).

[14] W. Setyawan and S. Curtarolo, High-throughput electronic band structure calculations: Challenges and tools, Comput. Mater. Sci. 49, 299 (2010).

[15] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, and B. Kozinsky, AiiDA: automated interactive infrastructure and database for computational science, Comput. Mater. Sci. 111, 218 (2016).

[16] A. Larsen, J. Mortensen, J. Blomqvist, I. Castelli, R. Christensen, M. Dulak, J. Friis, M. Groves, B. Hammer, C. Hargus, et al., The Atomic Simulation Environment: A Python library for working with atoms, J. Phys. Condens. Matter 29, 273002 (2017).

[17] K. Mathew, A. K. Singh, J. J. Gabriel, K. Choudhary, S. B. Sinnott, A. V. Davydov, F. Tavazza, and R. G. Hennig, MPIInterfaces: A Materials Project based Python tool for high-throughput computational screening of interfacial systems, Comput. Mater. Sci. 122, 183 (2016).

[18] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis, Comput. Mater. Sci. 68, 314 (2013).

[19] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, npj Comput. Mater. 1, 15010 (2015).

[20] A. Khorsheid and A. A. Peterson, Amp: a modular approach to machine learning in atomistic simulations, Comput. Phys. Commun. 207, 310 (2016).

[21] A. Van De Walle, M. Asta, and G. Ceder, The alloy theoretic automated toolkit: A user guide, Calphad 26, 539 (2002).

[22] K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I.-h. Chu, T. Smidt, B. Bocklund, M. Horton, et al., Atomate: A high-level interface to generate, execute, and analyze computational materials science workflows, Comput. Mater. Sci. 139, 140 (2017).

[23] Y. Wang, J. Lv, L. Zhu, and Y. Ma, CALYPSO: A method for crystal structure prediction, Comput. Phys. Commun. 183, 2063 (2012).

[24] T. Mayehibara, H. Wu, T. Angsten, A. Kaczmarowski, Z. Song, G. Jenness, W. Xie, and D. Morgan, The Materials Simulation Toolkit (MAST) for atomistic modeling of defects and diffusion, Comput. Mater. Sci. 126, 90 (2017).

[25] A. Togo and I. Tanaka, First principles phonon calculations in materials science, Scr. Mater. 108, 1 (2015).

[26] Y. Hinuma, G. Pizzi, Y. Kumagai, F. Oba, and I. Tanaka, Band structure diagram paths based on crystallography, Comput. Mater. Sci. 128, 140 (2017).

[27] Q.-J. Hong and A. van de Walle, A user guide for SLUSCHI: solid and liquid in ultra small coexistence with hovering interfaces, Calphad 52, 88 (2016).

[28] C. W. Glass, A. R. Oganov, and N. Hansen, USPEX: evolutionary crystal structure prediction, Comput. Phys. Commun. 175, 713 (2006).
[29] D. C. Lonie and E. Zurek, XtalOpt: An open-source evolutionary algorithm for crystal structure prediction, Comput. Phys. Commun. 182, 372 (2011).

[30] K. Michel and B. Meredig, Beyond bulk single crystals: a data format for all materials structure–property–processing relationships, MRS Bull. 41, 617 (2016).

[31] pypif: Python toolkit for working with PIFs, https://github.com/CitrineInformatics/pypif (2018), accessed: May 2020.

[32] S. P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, and K. A. Persson, The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles, Comput. Mater. Sci. 97, 209 (2015).

[33] C. Toher, Private communication (2018).

[34] The Materials Project: The Materials API, https://materialsproject.org/docs/api (accessed: December 2019).