The Thermodynamic Scale of Inorganic Crystalline Metastability

Submission to Challenge.gov, Materials Science and Engineering Data Challenge.

Under submission guidelines, personally identifying information has been removed.

Supplementary Information

S.I.1 Methods

S.I.1.1: High-throughput DFT Calculation Parameters

Density functional theory calculations for the Materials Project are performed with the Vienna ab initio software package (VASP),[12] using the projector augmented-wave method with the generalized gradient approximation (GGA) within the Perdew–Burke–Ernzerhof (PBE) framework.[3] Plane-wave basis cutoff energies are set to 520 eV, which is 1.3 times the maximum cutoff as specified by the VASP pseudopotentials. We use a *k*-point grid of 500/n points, where *n* is the number of atoms in the unit cell, distributed within the Brillouin zone in a Monkhorst-Pack grid,[4] or on a Gamma-centered grid for hexagonal cells. For oxides with strongly correlated electrons – Ag, Co, Cr, Cu, Fe, Mn, Mo, Nb, Ni, Ti, Re, Ta, V, W, Y – the GGA+U method is employed, with U representing the Hubbard-parameter.[5] The values of U are set according to the methodology of Wang et al[6] and refined by Jain et al.[7] Electronic energy convergence is set to $n \times 5 \times 10^{-5}$ eV, and ionic convergence is set to $n \times 5 \times 10^{-4}$ eV. All compounds are structurally optimized twice in two consecutive runs using the FireWorks package.[8] These DFT calculation parameters are all specified in the Python Materials Genomics (PyMatGen)[9] within the MPVaspInputSet. For reactions with gas phase decompositions (O_2, N_2, etc) , the energy of the gas phase is fit using known experimental reaction energies by the method of Wang *et al.*[10] Cohesive energies are attained by calculating

$$E_{Cohesive} = \frac{1}{n} \sum_{i=atom} n_i E_i - E_{Crystal}$$

where *n* is the number of atoms of chemistry *i* in the unit cell. E_{atom} is attained from 'atom-in-a-box' calculations of a single atom in a $10\text{Å} \times 10\text{Å} \times 10\text{Å}$ unit cell.

S.I.1.2: Phase Stability Determination

Phase stability is calculated using the convex hull construction, which compares the energy of a structure to the linear combinations of the energy of other phases at the same composition. Figure S.1. shows a hypothetical convex hull in a binary A-B system, plotted as formation energy versus composition. The stable phases in this system are given by the thermodynamic convex hull, which is constructed from the convex line that joins A–α–β–B. The γ and δ phases do not fall on or below the convex hull, indicating that they are metastable phases. The γ phase is metastable with respect to α, which is a compound of the same composition, meaning that γ is a metastable *polymorph*, whereas the δ phase is metastable with respect to a linear combination of α and β, signifying that it is metastable with respect to *phase-separation.* The metastability of γ and δ is quantified by the *energy above the hull*. For the polymorph γ, this energy of metastability would be given by the energy above the stable phase, α, by $E_{polymorph} = E_{\gamma} - E_{\alpha}$. For the phase separating compound δ, the energy above the hull is calculated by $E_{phase-separating} = E_{\delta} - \sum_i x_i E_i$, where *i* is from the set of the stable compounds adjacent to δ in composition (in this case, α and β) , and *x* is the phase fraction of those compounds in δ . Further details of the phase stability calculations can be found in Reference 11.

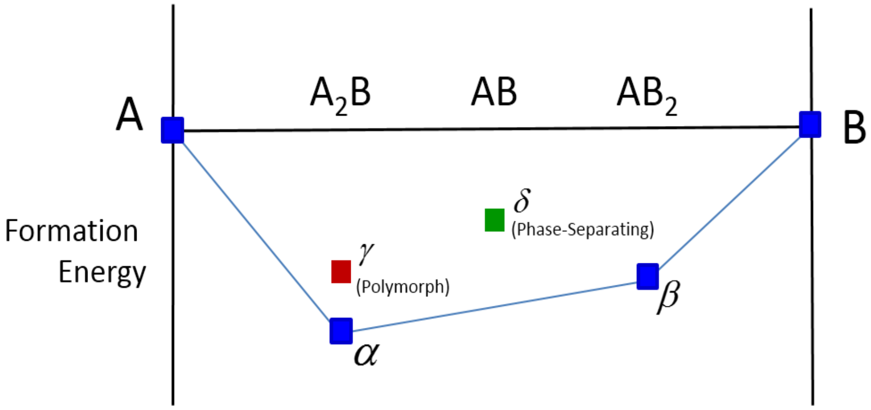

Figure S1. Convex hull in the binary A-B system. A, α, β and B are stable phases in this system, while γ is a metastable polymorph, and δ is metastable with respect to phase separation into α and β.

SI.1.3: Methods – Statistics and Visualization

Data from the MaterialsProject was collected and processed in python with Pymatgen[12] via the *mpquery* command from the MaterialsProject REST API interface.[13] Statistical analyses were completed within the numpy and scipy packages. Cumulative distribution functions from Figure 1a were smoothed using kernel distributions, using the *ksdensity* function in the MATLAB Statistics and Machine Learning Toolbox. Visualization for Figures 2a and 3 was performed using the Seaborn statistical data visualization package.[14] Bivariate kernel density estimates for Figure 2a were generated from the Seaborn *kdeplot* module. Figure 3 is generated from a modified Seaborn *violinplot* module. Figure 4a was constructed from a custom kernel distribution function combined with a normalized bar histogram from within the scipy package.

Bandwidths for Gaussian kernels were determined by Silverman's Rule of Thumb, given as the bandwidth $h \sim 1.06\hat{\sigma}n^{-1/5}$, where *n* is the number of samples and $\hat{\sigma}$ is the standard deviation of the samples. This resulted in bandwidths that ranged between 15 to 27 meV/atom.

SI.1.4: Hypothetical Structure Generation

Novel structures were predicted via the data-mined ionic substitution algorithm developed by Hautier *et al.*[15] as implemented in the pymatgen package. Ordered binary compounds from the 2012 version of the ICSD[16] were used as seeding structures for the ionic substitution. The probability acceptance threshold was set to 10^{-3} and the probability for unobserved substitutions (α) was set to 10^{-5}. Using this method, novel compounds were predicted for the following compositions (number of structures predicted): ZnO (34), Fe_2O_3 (35), Bi_2O_3 (33), Al_2O_3 (38), TiO_2 (92), ZrO_2 (82), HfO_2 (87), SnO_2 (88), V_2O_5 (12), Ta_2O_5 (11), WO_3 (21), MoO_3 (23). Compositions that belong to different AB formula classes have different numbers of predicted structures: structures with AB composition A_2B_3 have on average 35 structures, whereas AB composition AB_2 have on average 88 structures. This can be explained via the non-uniform distribution of unique ordered structure prototypes for given compositions in the ICSD: AB compounds have 77 unique ordered structure prototypes, while A_2B_3 compounds have 46, AB_2 compounds have 112, A_2B_5 have 13, and AB_3 have 71. The number of predicted compounds for each composition is further reduced from the number of prototypes by the substitution probability threshold specified earlier.

Lattice stability calculations were performed in the Fe_2O_3 system for the six predicted structures 100 meV/atom below the ground-state. Dynamical stability was evaluated from finite-difference phonon calculations, calculated from the *phonopy* package.[17] Of the six calculated structures, three were found to have no negative frequencies, and therefore no imaginary phonon modes, confirming dynamical lattice stability at T=*0K*. These structures are Fe_2O_3 substituted from:

- Bi_2O_3, (*Pa3* space group), mp-153042, 45 meV/atom above the hull
- Ga_2O_3, (*C2/m* space group), mp-95272, 47 meV/atom above the hull
- Sc_2S_3, (*Fddd* space group), mp-164944, 70 meV/atom above the hull

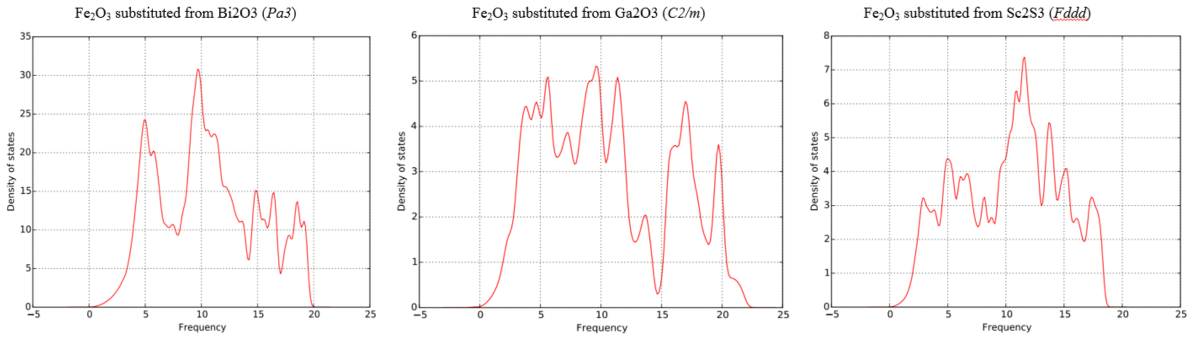

Figure S2. Phonon density of states for lattice-stable low-energy predicted Fe_2O_3 polymorphs

S.I.2: Dataset Creation, Validation and Provenance

S.I.2.1: DFT Error Discussion and Analysis

DFT-GGA offers an optimal combination of thermochemical accuracy and computational efficiency for a large-scale analysis of thermodynamic properties. A systematic study of DFT-GGA predictions of phase stability in binary metallic alloys demonstrated that DFT-GGA can successfully predict the experimental ground-state phase over 90% of the time.[18] To the best of our knowledge, there are no systematic investigations for other chemistries, but numerous studies on individual systems demonstrate that DFT-GGA can correctly reproduce the thermodynamic ground-states in oxides,[19] III-V semiconductors,[20] chalcogenides,[21] nitrides,[22] and borides[23]. Furthermore, the ability of DFT to rank phase-stability as a function of pressure is foundational to the field of high-pressure *in silico* DFT structure prediction, of which there are many successes.[24,25]

While there are notable errors in phase stability prediction, such as the failure of DFT to predict rutile as the ground-state of TiO_2,[26] the overall consistency of DFT thermochemical accuracy is sufficient to probe the energy *scale* of crystalline metastability. We have previously quantified the errors in DFT formation free energies of ternary oxides from binary oxides, showing that they are normally distributed, centered around zero with a standard deviation of 24 meV/atom.[27] We have further developed compatibility schemes between GGA and GGA+U mixing that reduces calculated formation energy errors between compounds with both localized and delocalized electronic states (strongly-correlated oxides, etc) from 464 meV/atom (mean absolute relative error of 21%) to 45 meV/atom (MARE of 2%). Furthermore, we emphasize that *absolute* formation energies are not directly comparable to *relative* energies when computing phase stability. In density functional theory, errors tend to be lower when comparing between compounds in the same chemistry (polymorphs) or between compounds adjacent in phase space (multinary phase-separating compounds).

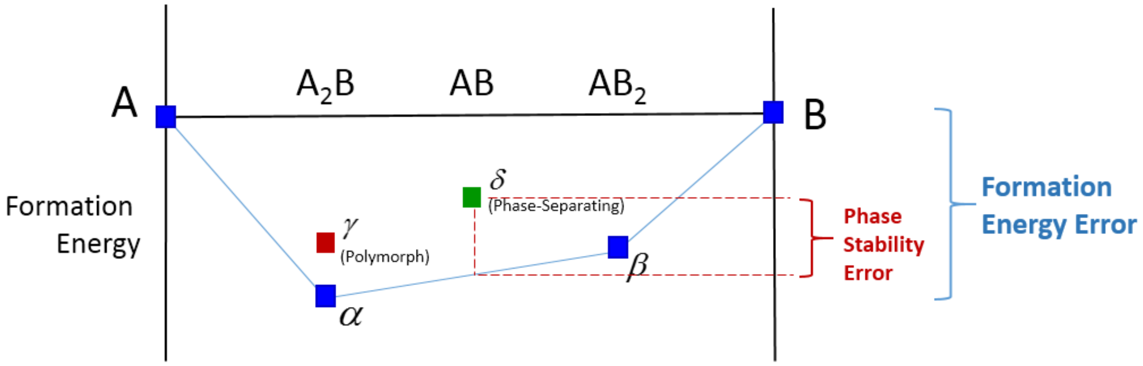

Figure S3. Phase stability should reference adjacent phases in a convex hull, rather than elemental states.

We investigate the influence of formation energy error on phase stability errors by performing Monte Carlo simulations on convex hulls, where the DFT formation energy for each entry is jittered by a zero-centered random Gaussian variable error with standard deviation of 24 meV/atom. We then re-evaluate the convex hull with jittered formation energies. We perform ten Monte Carlo samples per convex hull, on 1000 convex hulls randomly obtained from the MaterialsProject. We find that increasing the number of Monte Carlo samplings do not change the results of our findings. We emphasize that the use of random error reflects the 'worst-case scenario', where all DFT errors are independent and exhibit no correlation (consistent over- or under-prediction within a chemical space). The fact that DFT accurately predicts ground-states in elemental solids[18] and alloys suggests that DFT formation energy is *not* random and exhibits high degree of correlation, so the error bars we provide here are extremely conservative.

Figure S4 quantifies the standard deviation in *fraction* of metastable phases in a convex hull jittered by random error. We find that allotropes, and polymorphic systems in general, do not have a different fraction of metastable phases under jittered convex hulls, as there is always only one ground-state phase, and all other phases are metastable. Next we observe that the more entries in a convex hull, the more the fraction metastable tends to be fairly robust against random DFT error, where binaries have a median of 7% standard deviation in the fraction metastable, whereas quaternaries have a 2% standard deviation in the fraction metastable. The average number of entries in a convex hull is 67 entries, which we find corresponds to a 4% standard deviation in the fraction metastable.

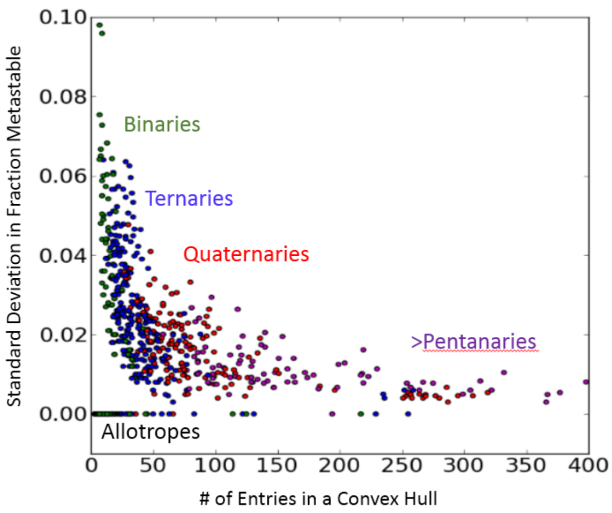

Figure S4. Fraction of metastable phases in a convex hull with formation energies jittered by 24 meV/atom random Gaussian variable.

Figure S5 shows the same analysis applied considering if the same stable phases and the same stable compositions are preserved under jittered convex hulls. Similar to Figure S4, we find that the more entries in a convex hull, the smaller the degree of variation in the both the same stable phases and the same stable compositions. On average, we find that the absolute probability of preserving the same stable phases average 67%, and is nearly always greater than 50%. The probability of attaining the same stable compositions under a jittered hull is higher, at an average 72%. Overall, these results are a product of the geometry of the convex hulls in real materials phase spaces, and the results demonstrate that a 24 meV/atom *random* error allows us to mostly preserve the qualitative features of the convex hull. Again, we have empirically validated[18] that correlation in DFT phase stability is significant, and that independent errors are representative of 'worst-case scenarios' in the phase stability error.

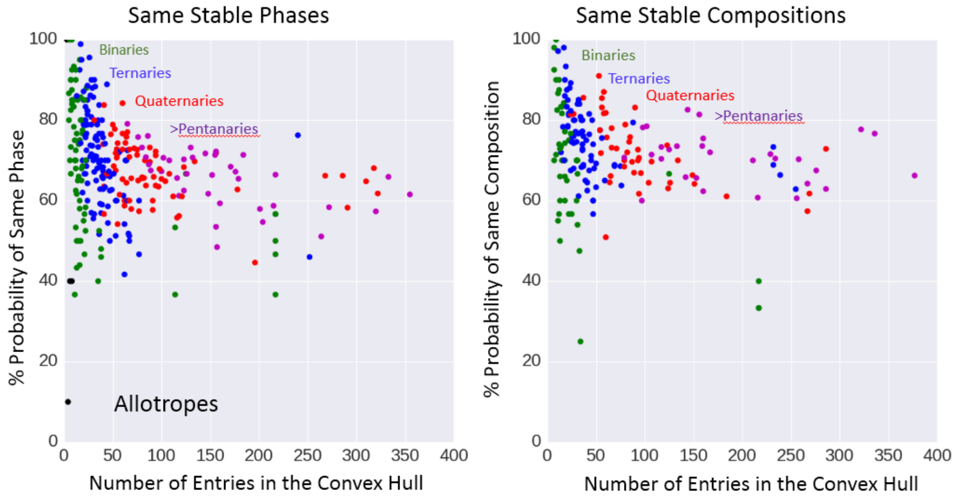

Figure S5. Probabilities of attaining same stable phases or same stable compositions under a convex hull jittered by random DFT error.

These Monte Carlo type investigations cannot be used to estimate the influence of DFT error on the energy scale of metastability, as the addition of a Gaussian random error to a negative-exponential heavy-tailed distribution shifts the 'energy above the hull' for all phases to higher values, extending the energy scale of metastability. A more sophisticated investigation of the influence of DFT formation energy error on phase stabilities will require a quantitative analysis of correlations in DFT error. This will be performed by the authors in a future publication. However, the statistical analyses in this work all use kernel density estimates, with Gaussian bandwidths ranging from 15-27 meV/atom, further mollifying the influence of systematic DFT errors on the trends attained by statistical analysis. The characteristic energy scales of the various 'categories' in both the chemistry and composition portions of our analysis is on the order of 100 meV/atom, much larger than typical DFT error bars. Finally, due to the large number of samples in each category (~100-2000), it is unlikely that analysis on energetic differences between chemistries and compositions fail from Type 1 statistical errors.

S.I.2.2: Dataset Construction from ICSD

The MaterialsProject is constructed from DFT calculations on crystallographic entries from the 2012 Inorganic Crystal Structure Database. This dataset was retrieved November 30th, 2014, when there were 58,606 Materials Project entries. We briefly summarize how the crystal structures from the ICSD are obtained and curated for inclusion into our dataset – technical methods are described further below, and more details can be found in References 28 and 29.

The Materials Project began from the 150,042 entries in the 2012 ICSD. Of these entries, 16,626 had non-standard CIF formats and were unable to be converted into a crystal structure entry. Using the structure-matcher algorithm (detailed below), the remaining 133,416 structures were grouped into 93,307 unique materials. We next eliminated entries with disorder, incomplete unit cells, or with chemical composition not matching the description in the CIF file (these typically occurred for missing light atoms, such as AlOOH in the ICSD entry, but AlOO in the crystal structure entry). This yielded the 33,704 MaterialsProject entries with directly traceable ICSD provenance. These ICSD entries of the Materials Project can be traced from the JSON "history" field of Materials Project entries. Many of these exhibited warnings (SI.2.3), removing the entries with the listed warnings yielded in 29,902 materials entries.

Non-ICSD Materials Project entries were constructed for application-specific calculations, including hypothetical constructed structures for materials design purposes, orderings for disordered ICSD entries that we considered technologically-relevant, entries retrieved from other databases that do not have detailed provenance, and partially de-lithiated structures for lithium ion battery cathodes and anodes.

Duplicate Identification and Removal:

Duplicate entries in the ICSD are prevalent, for example, there are 122 entries for the rutile phase of TiO_2 alone. Different synthesis conditions and characterization techniques can result in the same phase being represented in separate ICSD entries, differing by minor distortions in lattice parameters, lattice angles, and atomic positions. Symmetry-breaking distortions can additionally result in the same phase being identified under different crystallographic space groups. Crystal structures are also frequently duplicated in separate entries under non-standard space group representations.

We constructed an in-house structure-matching algorithm to group isostructural entries that are robust under noisy structural distortions and non-standard lattices for the unit cell. Our algorithm compares crystal structures via solution of the linear assignment problem (LAP). This algorithm is implemented in the pymatgen,[9] under the StructureMatcher package. The algorithm proceeds in the following five steps:

1) Input structures are reduced to their Niggli primitive cells.
2) If two primitive structures have the same composition, equivalent lattice representations are enumerated.
3) From the lattice representations, determine if linear combinations of lattice vectors from the first structure can produce a lattice within angle and length tolerances of that of the second structure. Lattice vectors were given a 20% relative tolerance, and angles a 5 degree absolute tolerance.
4) The original structure is basis transformed into each of the new valid lattices.
5) With both structures on equivalent lattices, treat each structure as one half of a bipartite graph and solve the LAP by minimizing the distances between equivalent species in the two crystals.
6) If an optimal assignment solution is found, meaning the maximum distance between two sites is less than a provided absolute site tolerance, the two crystal structures are a considered equivalent. The site tolerance is defined as a fraction, in this case 50%, of cube root of the average volume per atom; allowing for site tolerance to scale with unit cell volume.

SI.2.3: MaterialsProject Error Checking

A series of automatic checks are implemented into the Materials Project to identify ICSD entries that exhibit dramatic structure changes upon calculation by DFT, deviating either in bond-lengths or unit-cell volume upon relaxation. The origin of these structure changes either result from the original structures being poorly or incorrectly characterized experimentally, or that approximations in the DFT-GGA functional are limited at describing the interatomic interactions within the crystal – such as van der Waals bonding[30] or steep gradients in the electronic surface[31]. Whether the error arises from errors in the structural characterization or from DFT, energetics attained from these entries are likely to be unphysical and should be excluded from the dataset. We discuss our error identification criteria below:

<u>Volume Change</u>

Volume changes upon relaxation from DFT were categorized for all MaterialsProject entries, calculated

$$\Delta V = \frac{V_{GGA} - V_{\exp}}{V_{\exp}}$$

It is well known that DFT-GGA tends to overestimate volume, and the average volume change in the Materials Project is found to be 3.2%, with the distribution of volume changes symmetrically distributed around this average (Figure S.3). We identify the 95th and 5th percentile volume changes to be 9.6% and -3.2%. Entries with volume changes outside of these bounds, which represent two standard deviations, are tagged as errors and excluded from the dataset.

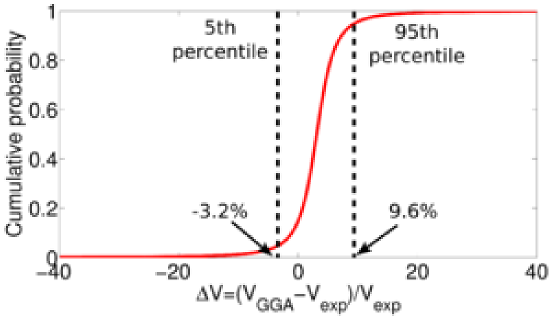

Figure S.6 Cumulative distribution of volume changes in the Materials Project

Examples:

YZn and YbZn *– materials_id: 11577 and 179 – icsd_id: 106232, 206226*
Volume differences of 370% and 350%. The paper reference mentions CsCl structures, while the entry given by the ICSD is a fcc-based alloy. The volume discrepancy comes therefore from an erroneous export of the experimental data in the paper to the ICSD.

***$TaMn_2O_3$** – materials_id: 7521, icsd_id: 15995*
-27% volume difference. The ICSD entry indicates that the experimental data is dubious, warning: "Unusual difference between calculated and measured density". The oxidation state reported for Ta: +2 is extremely rare.

***MoS_2** – materials_id: 1434, icsd_id: 644257*
18% volume difference. The MoS_2 structure is layered and van der Waals interactions between the S atoms are essential to the geometry of the structure. It is known that GGA does not model well vDW interactions, therefore this volume change is likely due to an error from DFT.

Bond Length Change

To monitor the difference in bond length between the experimental and computed data, we need to automatically determine the chemical bonds in a given crystal structure. We have taken a geometric approach to define the nearest neighbors of a given atom using the Voronoi construction.[32] After defining the nearest neighbor of each atoms in the ICSD unit cell, we monitor how the distance between each atom and his neighbors (i.e., the bond length) changes upon DFT relaxation. We have then for each atom nearest neighbor pair a change in distance that can be expressed by a change in bond length. We identified the 5th and 95th percentile bond-length changes to be -2.1% and +2.1%. Calculations with bond length changes beyond these extremes are tagged with errors and removed from the dataset.

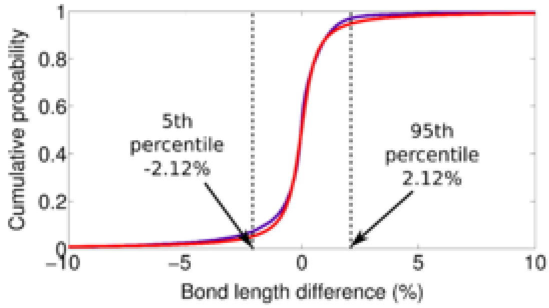

Figure S.7 Cumulative distribution of bond length changes in the Materials Project

Examples:

Ba_2CoO_4—task_id=16887, icsd_id=92321

The change in Co-O bond length is ~35%. Upon closer investigation we find that the ICSD entry shows a very long Co-O bond while the DFT relaxation brought the O much closer and form the more common regular tetrahedral local environment for Co^{4+} (see Figure S5). It is likely that the ICSD entry had an error in atomic positions.

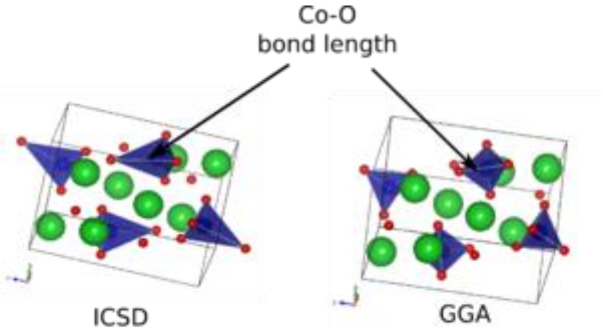

Figure S.5: Large change in bond length between the ICSD and computed entry for Ba_2CoO_4

SnF_2 – task_id: 7456 and icsd_id: 14194

Very large relaxations are observed. For instance, there is a F atom that moved dramatically after DFT relaxation. It is difficult to say if the error comes from the measurement of from DFT but this might be a DFT problem as the measurement is tagged as high quality data in the ICSD and comes from single crystal diffraction.

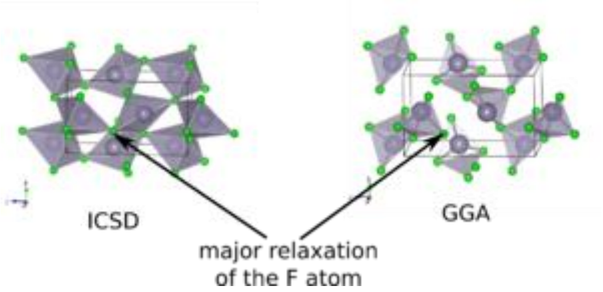

Figure S.8: Large change in bond length between the ICSD and computed entry for SnF_2

$RbInMo_2O_8$ – task_id: 7402 , icsd_id: 10186

Here, we have smaller bond length changes but still outlying the distribution. The Mo-O bond is around 1.9Å in experiments and 1.8 Å computationally. This is a quite unusual difference of 5%. Ionic radii are closer to the computed value for Mo^{6+} (1.35 Å +0.41 Å =1.76 Å).

SI.2.4: Manual Data Provenance Investigation

The ICSD serves as a general utility crystallographic reference, containing structural entries for bulk phases, hypothetical predicted phases, defect unit cells, etc. Many of these structural entries can pass the screening algorithms we described above, but are not within the scope of our desired dataset, which we limit to *observed, bulk crystalline phases*. We consider the ICSD entries that do not fall within this category ‘spurious’ – that is, able to produce an energy via DFT, but not relevant to our analysis on the energy scale of crystalline metastability. To identify the origin and frequency of these entries, we performed a manual investigation of data provenance. We consulted the original journal publications where the structures are reported, and then identified and categorized the types of ICSD entries that do not correspond to observed, bulk crystalline phases. Because it would be infeasible to manually investigate the provenance of all 33,000+ entries in our dataset, we devised a strategy that would enable a *statistical description* of the energy scale of metastability that discounts a majority of the entries in the spurious categories detailed below.

Our strategy centers on the assumption that the ICSD curates entries similarly across chemistry and composition space, meaning that the ratio of bulk crystalline phases to spurious entries is similar over different subsets of the overall dataset. This assumption is a reflection of the nature of the ICSD dataset, rather than of the investigative biases of materials researchers. We pick a smaller representative set, specifically the binary oxides, to characterize the origin and frequency of spurious entries in the larger ICSD dataset. We choose the binary oxides because they are a large and well-studied class of materials, meaning that the categories of spurious structures can be sampled completely. We determine that the majority of the spurious entries, notably those corresponding to unit cell descriptions of defect structures (point defects, superstructure frameworks, thin-film structures, etc) and hypothetical predicted structures, form the highest-energy structures, and compose approximately 20% of the dataset (Figure S7). There is not enough information in the ICSD entries to remove these structures with great precision, and so we limit our statistical investigations to the 80% lower-energy structures for a given dataset, with the claim that this will approximate the statistical median and 90th percentile of the true distribution excluding spurious entries. The descriptions of the categories of spurious entries are detailed below.

Experimentally-Unobserved Hypothetical Structures from First-Principles Calculations

First-principles crystal structure prediction involves the generation of candidate crystal structures by simulated annealing, genetic algorithms, Monte Carlo algorithms, and database-sampling techniques, whose energies are then evaluated by a first-principles Hamiltonian, oftentimes density functional theory. In the process of finding the ground-state structure, many *unphysical* candidate structures are typically created, many very high in energy. Despite these *in silico* structures being unobserved, they are occasionally still inserted into the ICSD, as a reference. Entries that fall under this category should not be considered in our dataset.

Examples:
- ICSD 161062: Ono, Shigeaki, John P. Brodholt, and G. David Price. "First-principles simulation of high-pressure polymorphs in MgAl2O4." *Physics and Chemistry of Minerals*35.7 (2008): 381-386.

- ICSD 170553: Foster, Martin D., et al. "Chemical evaluation of hypothetical uninodal zeolites."*Journal of the American Chemical Society* 126.31 (2004): 9769-9775.

Thin-Film Structures

A large number of ICSD entries are thin-film structures that form on the interfaces of bulk solids. Occasionally, it is explicitly stated that these thin-film structures do not possess bulk counterparts. Because density functional theory utilizes periodic-boundary conditions, such thin-film unit cells would be tessellated to form an infinite bulk, which is not the intention of the ICSD entry.

Examples:
- ICSD 150543: Kumar, Jitendra, and Rakesh Saxena. "Formation of NaCl-and Cu_2O-type oxides of platinum and palladium on carbon and alumina support films." *Journal of the Less Common Metals* 147.1 (1989): 59-71.
- ICSD 61543: Kubaschewski, O., and B. E. Hopkins. "Oxidation mechanisms of niobium, tantalum, molybdenum and tungsten." *Journal of the Less Common Metals* 2.2 (1960): 172-180.

Defect-Structures

Many ICSD entries provide the single unit-cell of a defect structure, which under periodic-boundary conditions, would represent a bulk with a 100% defect concentration. Such phases are unphysical and can be up to 10 eV above the ground-state phase.

Examples:

- ICSD 166273: Johnsen, Rune E., and Poul Norby. "A Structural Study of Stacking Disorder in the Decomposition Oxide of MgAl Layered Double Hydroxide: A DIFFaX+ Analysis."*The Journal of Physical Chemistry C* 113.44 (2009): 19061-19066.
- ICSD 181039: Boonchun, Adisak, and Walter RL Lambrecht. "Critical evaluation of the LDA+ U approach for band gap corrections in point defect calculations: The oxygen vacancy in ZnO case study." *physica status solidi (b)* 248.5 (2011): 1043-1051.

Crystalline framework of a solid solution

These ICSD entries are the crystalline framework or structural matrix of a solid solution, either describing a sublattice of atoms where intercalants or diffusants pass through, or are of solid solutions that can be predominantly the composition of the reported entry, but the structure may not be stable without the inclusion of solution alloying elements. These ICSD entries do not correspond to the experimentally described composition, so these entries are excluded from our analysis.

Examples:

- ICSD 95729: Lidin, Sven, Franziska Rohrer, and Ann-Kristin Larsson. "The structure of $Nb_5O_{12}F$." *Solid state sciences* 4.6 (2002): 767-772.
- ICSD 1503: Ghedira, M., et al. "Structural aspects of the metal-insulator transitions in $V_{0.985}Al_{0.015}O_2$." *Journal of Solid State Chemistry* 22.4 (1977): 423-438.

Superseded Crystal Structures

There are ICSD entries that have high energies, but for which there are more recent entries that share the same structural topology but have lower bulk lattice energies. In these cases, the high-energy structure are considered superseded and is not considered in the dataset. The origin of these high-energy compounds can be that the initial refinement is poorly conducted, or that the calculation exhibited a convergence error.

Examples:

- ICSD 15568: Loopstra, B.O.; Cordfunke, E.H.P. “On the structure of alpha UO3” *Recueil des Travaux Chimiques des Pays-Bas et de la Belgique*, (1966), 85. **Revised:** Greaves, C. T., and B. E. F. Fender. "The structure of α-UO3 by neutron and electron diffraction." Acta Crystallographica Section B: Structural Crystallography and Crystal Chemistry 28.12 (1972): 3609-3614.

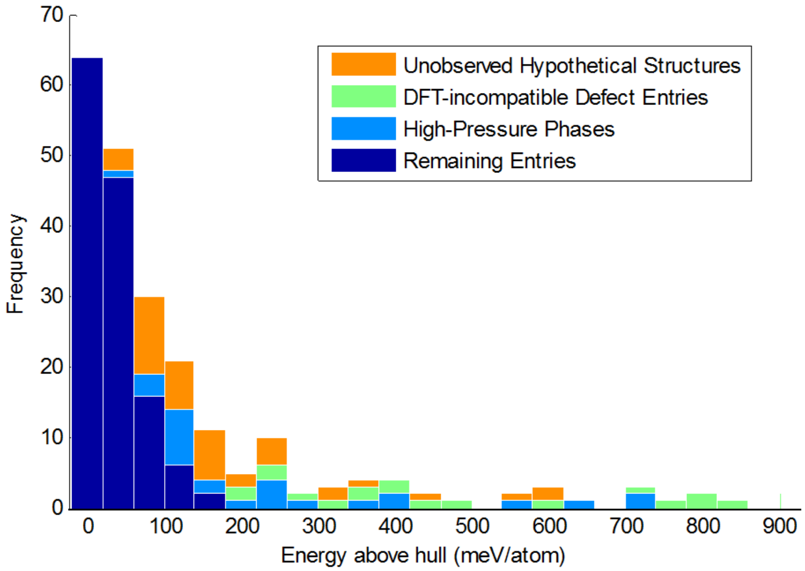

Figure S9. Energy distribution of metastable binary oxide polymorphs in the MaterialsProject, sorted by provenance. Unobserved, hypothetical phases and DFT-incompatible defect entries compose of approximately 20% of the dataset, and are distributed on the upper-end of the energy spectrum.

SI. 3. Supplemental Data Tables and Figures

SI.3.1: Data Table for Figure 1: Metastability by Chemistry

Figure 1:

Group	Chemistry	Num. Entries	Pauling Electro-negativity	Median Cohesive Energy (eV/atom)	Median E_abv_hull (meV/atom)	90th Percentile E_abv_hull (meV/atom)
VII	F	819	3.98	-5.12	12.9 [11.1,15.8]	104 [91.4,117]
	Cl	371	3.16	-4.18	13.1 [11.7,16.5]	58.0 [48.1,63.1]
	Br	137	2.96	-3.31	10.2 [8.1,12.0]	29.6 [24.2,38.6]
	I	149	2.66	-2.70	5.2 [4.2,7.9]	23.6 [20.4,30.2
VI	O	5522	3.44	-6.26	15.4 [14.8,15.9]	62.3 [59.6,66.1]
	S	574	2.58	-4.52	9.7 [8.5,11.0]	45.3 [38.6,52.2]
	Se	219	2.55	-4.24	12.2 [9.5,13.6]	47.9 [39.8,57.4]
	Te	102	2.1	-3.66	9.2 [7.3,11.6]	24.2 [21.1,39.9]
V	N	501	3.05	-6.38	62.8 [55.7,74.5]	195 [177,212]
	P	102	2.19	-5.59	8.2 [6.4,12.6]	34.6 [26.7,50.3]
	As	93	2.18	-4.87	18.5 [14.5,24.1]	60.6 [48.0,79.8]
Others	Intermetallics	943	-	-4.88	15.4 [13.8,16.1]	60.4 [49.9,67.4]

Chemistries are grouped by the most electronegative element in the compound. We do this also for complex anions, (for example, a phosphate groups (PO_4^{3-}), counts as an oxide, not a phosphide. Intermetallic compounds have all elements within the Alkalis, Alkali Earths, Transition Metals, Rare-Earths, Al, Ga, In, Sn, Pb, or Bi. Cohesive energies are referenced with respect to vapor phase atoms as calculated from DFT by a single atom in a 10Å x 10Å x 10Å box. The brackets correspond to 95% confidence intervals on the medians and 90th percentiles, as given from order statistics.

SI.3.2: Data Table for Figure 2: Metastability by Composition

Decomposition	Polymorphs			Phase-Separating			All		
Composition	Num. Entries	Median Ehull (meV/atom)	90th Percentile Ehull (meV/atom)	Num. Entries	Median Ehull (meV/atom)	90th Percentile Ehull (meV/atom)	Num. Entries	Median Ehull (meV/atom)	90th Percentile Ehull (meV/atom)
Allotropes	106	45 [29.9,68.8]	145 [141,191]	0	-	-	106	45 [29.9,68.8]	145 [141,191]
Binaries	924	9.8 [8.7,11.1]	52 [45.7,61.3]	1051	20 [18.8,21.7]	76.8 [71.1,82.0]	1975	14.7 [13.6,15.7]	67.7 [62.1,73.4]
Ternaries	1086	6.8 [6.3,7.8]	27.1 [24.8,29.9]	3146	17.4 [16.8,18.3]	62 [59.3,66.6]	4233	13.8 [13.4,14.6]	52.7 [50.2,55.8]
Quaternaries	276	3.4 [2.7,4.1]	14.3 [11.6,17.2]	2422	14.8 [13.9,15.8]	62.4 [57.8,68.8]	2698	12.6 [12.2, 13.6]	56.3 [53.0,59.6]
Pentanaries+	19	1.4 [0.8,5.7]	13.9 [5.7,27.2]	1431	36 [32.4,39.5]	150 [136,160]	1450	34.3 [31.1,38.1]	147 [134,158]

Brackets correspond to 95% confidence intervals on the medians and 90th percentiles, as given from order statistics.

Supplemental Information References

[1] Kresse, Georg, and Jürgen Furthmüller. "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set." *Physical Review B*54.16 (1996): 11169.

[2] Kresse, Georg, and Jürgen Furthmüller. "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set."*Computational Materials Science* 6.1 (1996): 15-50.

[3] Perdew, John P., Kieron Burke, and Matthias Ernzerhof. "Generalized gradient approximation made simple." *Physical review letters* 77.18 (1996): 3865.

[4] Monkhorst, Hendrik J., and James D. Pack. "Special points for Brillouin-zone integrations." *Physical Review B* 13.12 (1976): 5188.

[5] Anisimov, Vladimir I., Jan Zaanen, and Ole K. Andersen. "Band theory and Mott insulators: Hubbard U instead of Stoner I." *Physical Review B* 44.3 (1991): 943.

[6] Wang, Lei, Thomas Maxisch, and Gerbrand Ceder. "Oxidation energies of transition metal oxides within the GGA+ U framework." *Physical Review B* 73.19 (2006): 195107.

[7] Jain, Anubhav, et al. "Formation enthalpies by mixing GGA and GGA+ U calculations." *Physical Review B* 84.4 (2011): 045115.

[8] Jain, Anubhav, et al. "FireWorks: a dynamic workflow system designed for high-throughput applications." *Concurrency and Computation: Practice and Experience* (2015).

[9] Ong, Shyue Ping, et al. "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis." *Computational Materials Science* 68 (2013): 314-319.

[10] Wang, L., T. Maxisch, and G. Ceder. "A first-principles approach to studying the thermal stability of oxide cathode materials." *Chemistry of materials* 19.3 (2007): 543-552.

[11] Ping Ong, Shyue, et al. "Li– Fe– P– O2 phase diagram from first principles calculations." *Chemistry of Materials* 20.5 (2008): 1798-1807.

[12] Ong, Shyue Ping, et al. "Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis." *Computational Materials Science* 68 (2013): 314-319.

[13] Ong, Shyue Ping, et al. "The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles." *Computational Materials Science* 97 (2015): 209-215.

[14] Michael Waskom et al, *seaborn*, http://dx.doi.org/10.5281/zenodo.19108

[15] Hautier, Geoffroy, et al. "Data mined ionic substitutions for the discovery of new compounds." *Inorganic Chemistry* 50.2 (2010): 656-663.

[16] Bergerhoff, G., and I. D. Brown. "Crystallographic databases." *International Union of Crystallography, Chester* (1987): 77-95.

[17] Togo, Atsushi, and Isao Tanaka. "First principles phonon calculations in materials science." *Scripta Materialia* 108 (2015): 1-5.

[18] Curtarolo, Stefano, Dane Morgan, and Gerbrand Ceder. "Accuracy of ab initio methods in predicting the crystal structures of metals: A review of 80 binary alloys." *Calphad* 29.3 (2005): 163-211.

[19] Zhu, Y. Z., et al. "Electronic structure and phase stability of MgO, ZnO, CdO, and related ternary alloys." *Physical Review B* 77.24 (2008): 245209.

Karazhanov, S. Zh, et al. "Phase stability, electronic structure, and optical properties of indium oxide polytypes." *Physical Review B* 76.7 (2007): 075129.

Gerosa, Matteo, et al. "Electronic structure and phase stability of oxide semiconductors: Performance of dielectric-dependent hybrid functional DFT, benchmarked against G W band structure calculations and experiments." *Physical Review B* 91.15 (2015): 155201.

[20] Wang, S. Q., and H. Q. Ye. "First-principles study on elastic properties and phase stability of III–V compounds." *physica status solidi (b)* 240.1 (2003): 45-54.

[21] Gökoğlu, G., M. Durandurdu, and O. Gülseren. "First principles study of structural phase stability of wide-gap semiconductors MgTe, MgS and MgSe." *Computational Materials Science* 47.2 (2009): 593-598.

[22] Takeuchi, Noboru. "First-principles calculations of the ground-state properties and stability of ScN." *Physical Review B* 65.4 (2002): 045204.

Mancera, Luis, Jairo A. Rodríguez, and Noboru Takeuchi. "First principles calculations of the ground state properties and structural phase transformation in YN." *Journal of Physics: Condensed Matter* 15.17 (2003): 2625.

[23] Wang, Bing, et al. "Phase stability and physical properties of manganese borides: a first-principles study." *The Journal of Physical Chemistry C* 115.43 (2011): 21429-21435.

[24] Ma, Yanming, et al. "Transparent dense sodium." Nature 458.7235 (2009): 182-185.

[25] Oganov, Artem R., and Colin W. Glass. "Crystal structure prediction using ab initio evolutionary techniques: Principles and applications." *The Journal of chemical physics* 124.24 (2006): 244704.

[26] Muscat, Joseph, Varghese Swamy, and Nicholas M. Harrison. "First-principles calculations of the phase stability of TiO 2." *Physical Review B* 65.22 (2002): 224112.

[27] Hautier, Geoffroy, et al. "Accuracy of density functional theory in predicting formation energies of ternary oxides from binary oxides and its implication on phase stability." *Physical Review B* 85.15 (2012): 155208.

[28] Jain, Anubhav, et al. "A high-throughput infrastructure for density functional theory calculations." *Computational Materials Science* 50.8 (2011): 2295-2310.

[29] The Materials Project Wiki, https://materialsproject.org/wiki/index.php/Main_Page

[30] Grimme, Stefan. "Density functional theory with London dispersion corrections." *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011): 211-228.

[31] Armiento, Rickard, and Ann E. Mattsson. "Functional designed to include surface effects in self-consistent density functional theory." *Physical Review B*72.8 (2005): 085108.

[32] M. O'Keefe, Acta Crystallographica Section A 35, 772-775 (1979