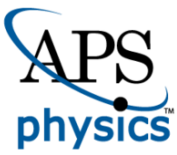

High-throughput crystal structure solution using prototypes

Sean D. Griesemer, Logan Ward, and Chris Wolverton

Phys. Rev. Materials **5**, 105003 — Published 25 October 2021

DOI: 10.1103/PhysRevMaterials.5.105003

High-Throughput Crystal Structure Solution Using Prototypes

Sean D. Griesemer[1], Logan Ward[2], Chris Wolverton[1]

[1] Department of Materials Science and Engineering, Northwestern University, Evanston, Illinois 60208, USA
[2] Data Science and Learning Division, Argonne National Laboratory, Lemont, Illinois 60439, USA

ABSTRACT

Databases of density functional theory (DFT) calculations, such as the Open Quantum Materials Database (OQMD), have paved the way for accelerated materials discovery. DFT calculations require crystal structure information as input; however, due to inherent challenges in solving a compound's structure from powder diffraction data alone, there are thousands of experimentally synthesized compounds whose structures remain unsolved. We present a rapid DFT-based structure solution method capable of resolving numerous outstanding structure solution problems at low computational cost. The method involves: 1) searching inorganic compound databases for all prototypes that match known structural characteristics of the compound, such as stoichiometry, space group, and number of atoms per cell, 2) performing DFT calculations of the target composition in each of the structural prototypes, and 3) evaluating these prototypes as candidates using a combination of DFT energy and match between calculated and experimental diffraction pattern. As this approach is straightforward and inexpensive, we employ it to solve 520 previously unsolved compounds from the Powder Diffraction File, resulting in a 1.4% expansion of the set of all experimental compounds in the OQMD. DFT calculations of these newly solved compounds could yield valuable properties.

I. INTRODUCTION

Crystal structure is a fundamental descriptor of inorganic compounds and is necessary input for first principles calculations. Indeed, the composition and crystal structure of a compound, comprising of unit cell vectors and atomic coordinates, are the only inputs required for a DFT calculation of the compound's energetic, electronic, and magnetic properties. Thanks to knowledge of crystal structures obtained by experiment, databases of high-throughput DFT calculations, such as the Open Quantum Materials Database (OQMD), Materials Project, and Automatic Flow (AFLOW), have enabled the calculation of phase diagrams, screening of materials for future applications, and prediction of novel materials. (1) (2) (3) (4) (5) (6) (7) (8) However, due to challenges in extracting crystal structure from experimental diffraction data, there are many known compounds with unknown crystal structures. For example, there are thousands of diffraction patterns in the Powder Diffraction File (PDF) without an associated crystal structure, meaning that compounds have been synthesized, diffraction patterns measured, and yet there is still no solved structure. (9) Identifying the structure of these materials would enable DFT calculations of their properties and open the door to a full exploration of their potential.

There are several reasons why a complete crystal structure is not always obtained in a diffraction experiment. For instance, while the unit cell parameters, space group, formula units per unit cell, and elemental composition can often be determined from high quality data by indexing the diffraction peaks, the determination of atomic coordinates, known as structure solution, is especially challenging because the process of measuring a diffraction intensity does not capture the phase of diffracted waves, complicating the inversion from reciprocal space to real space. (10) Another common reason is that the compound either is part of a multi-phase material, is impure or contains elements that only weakly scatter x-rays, hindering the ability to capture relevant information in the diffraction pattern. When attempting to solve structures, crystallographers

routinely use structure optimization algorithms, in which atomic coordinates are optimized to match the diffraction pattern, i.e. minimize the R-Factor. When only R-Factor is used for the objective function, structure optimization algorithms are sometimes challenged by the existence of multiple solutions with similar R-Factors. A promising workaround is to supplement R-Factor with DFT calculations in order to rule out candidates that are unphysically high in energy. (11) For example, the First-Principles-Assisted Structure Solution (FPASS) method, which uses a genetic algorithm with R-Factor and DFT energy as a combined objective function, has been used to effectively resolve several long-standing problems. (12) (13) (14) (15) Other DFT-based structure optimization algorithms that can be constrained using experimental input include USPEX, (16) CALYPSO, (17) XtalOpt, (18) PEGS, (19) and AIRSS. (20) DFT provides a highly accurate estimate of the energetic stability of candidate structures; however, DFT is computationally expensive to use in structure optimization algorithms like FPASS, USPEX and others, where as many as hundreds or thousands of structures are considered over the course of optimization.

On the other hand, a simpler and cheaper way to solve structures is to search existing databases. The OQMD contains DFT calculations of over 800,000 compounds, including experimentally observed compounds from a 2011 version of the ICSD as well as many hypothetical compounds. As we will show, structures from the ICSD portion of the OQMD can be grouped into 10,203 prototypes, distinguished by space group, stoichiometry, and Wyckoff site occupancies. This grouping allows us to identify a relatively small number of symmetrically distinct prototypes as candidates for a given unsolved structure, according to the experimentally determined stoichiometry, number of formula units and space group. Furthermore, we find that 83% of distinct compounds in the ICSD share a common prototype with at least one other ICSD

compound, giving us confidence that we can solve many (but not all) new structures using a "prototype searching" method. In this prototype searching method, we search for candidate prototypes in the OQMD, select a representative structure for each prototype, decorate the structures with the experimental composition, and evaluate them by computing R-Factors and DFT energies. A related prototype searching method has been used to predict new compounds for hydrogen storage applications; (21) (22) (23) (24) however, without experimental input to constrain the search, the prototype searching method is still computationally expensive. On the other hand, when used for structure solution with experimental input, our prototype searching method usually requires evaluating up to 3 prototypes, far fewer than what is needed for other structure solution methods, allowing us to solve structures at low cost. We note while structure optimization algorithms like FPASS, USPEX and others can leverage experimental input to constrain the search, as optimization algorithms they still typically require DFT calculations of many structures over the search space, including highly unphysical structures whose atomic coordinates are consistent with the experimental space group, stoichiometry, and Wyckoff site occupancies. On the other hand, our prototype searching method gets us straight to the answer with just a few DFT calculations for prototypes that are known to exist in the ICSD. Since the prototype searching method is inexpensive, it can be used to quickly solve numerous unsolved compounds and expand crystal databases with a limited computational budget.

In this paper, we leverage the low computational cost of the prototype searching method to solve the structures of 520 compounds from experimental diffraction patterns in the PDF. All 520 compounds were missing from the ICSD and OQMD, and thus are newly solved, and constitute a 1.4% expansion of all experimentally known compounds in the OQMD. We have provided the solved structures as VASP-formatted files in the Supplemental Material, (25) and they are

available in the latest release of the OQMD. Confident that we have identified structures that both match experimental input and are energetically stable, we open the door to analyzing the properties of these materials and considering their use in a wide range of future applications.

II. THE PROTOTYPE SEARCHING METHOD

In this section, we detail the prototype searching method to solve the structure of a compound given experimental data.

A. Searching for Candidate Structures

A completely solved structure is one where we know all descriptive details; in particular, the unit cell parameters and the coordinates of all atoms in the unit cell. For the compounds we address in this paper, we have from experimental data the unit cell parameters, elemental composition, space group, and number of formula units per unit cell, but we do not have the atomic coordinates, suggesting that the diffraction peaks were successfully indexed but the structure solution step was not completed. Our approach to solve for the atomic coordinates of the structure is to take the stoichiometry, space group, and the number of formula units per unit cell, and search the OQMD for prototypes with the same attributes. We define the prototype of a crystal structure as the set of the following attributes:

1. Stoichiometry, e.g. ABC_3
2. Space group
3. Set of Wyckoff site occupancies in the unit cell

For example, the calcite prototype ($CaCO_3$) has ABC_3 stoichiometry, $R\bar{3}c$ space group, six atoms on the ‘6a’ (0,0,1/4) Wyckoff site, six atoms on the ‘6b’ (0,0,0) Wyckoff site, and eighteen atoms

on the '18e' (x,0,1/4) Wyckoff site. Leveraging this definition allows us to classify 32 compounds within the OQMD with these attributes as having the prototype of calcite, irrespective of the elements comprising {A,B,C}, value of x, or unit cell parameters. This classification allows us to treat this group as one, symmetrically unique candidate solution to an experimental structure.

Having identified a relatively small number of prototypes as possible solutions to the experimental structure, we proceed to generate candidate structures by populating the prototypes with the experimentally determined unit cell parameters and elements from the composition. We consider all possible arrangements of elements in the structure, e.g. the two distinct ways to swap Ca and C onto the Wyckoff sites of the calcite prototype. We must also account for the fact that a single prototype can produce multiple structures that, while symmetrically identical, have different local geometries. For example, the structures C23, C25, C29, and C37 ($PbCl_2$, $HgCl_2$, SrH_2, and Co_2Si respectively) have the same stoichiometry, space group, and Wyckoff site occupancies (AB_2, Pnma, {4c,4c,4c}), but are distinct structures. In order to decide which of these structures to select as a candidate, we compute the R-Factor of all compounds with this prototype in the OQMD but with the target composition and experimental lattice parameters substituted in. Since the calculation of R-Factor is very fast, we can quickly select the structure with the lowest R-Factor as the candidate (see section IIBii for an explanation of the R-Factor calculation). By the end of our procedure, we have generated a set of candidate structures, usually no more than seven structures across three prototypes (see Results section), as possible solutions for the experimental structure.

We note that we initially assume the experimental structure does not have any partially occupied sites. In some cases, this assumption will be inevitably incorrect. We can justify the

assumption if we obtain a structure that has a satisfyingly low energy and R-Factor; otherwise, we say that none of our candidate structures are valid solutions.

B. Evaluating Energy and R-Factor of Candidate Structures

i. Calculating Formation Energy Using Density Functional Theory

We use DFT to compute the formation energy of all candidate structures. All DFT calculations are performed using the Vienna Ab-Initio Simulation Package (VASP) v5.4.4, (26) (27) using the PBE exchange correlation functional, (28) and potentials supplied by VASP with the projector augmented-wave method. (29) We use the same DFT settings as those used for hundreds of thousands of compounds in the OQMD, so that the energetic stability of our candidate structures can be directly compared to other existing compounds. To measure the energetic stability of a candidate structure, we compute the convex hull of formation energies of all OQMD compounds in the relevant phase space and compute the difference between the candidate structure's formation energy and the convex hull energy at the target composition. (30) (31) If this difference is less than zero, then the candidate structure is energetically stable and a new convex hull that includes this structure can be constructed. The reader is referred to Ref. 29 for a complete discussion of DFT settings used as well as how formation energies and the convex hull are computed for all compounds in the OQMD. (32)

ii. Calculating Match to Diffraction Pattern, or R-Factor

In addition to measuring the energetic stability of candidate structures using DFT, we also evaluate how well they match the experimental diffraction pattern. We do this by generating a hypothetical diffraction pattern associated with each candidate structure, computing the R-Factor as a metric for the distance between the hypothetical and experimental diffraction patterns, and slightly refining the atomic coordinates to minimize the R-Factors. Our methods of computing R-Factor are implemented in the Materials Interface (Mint) software (33) and are described in detail in Ward et al. (14) Here, we will provide a brief summary.

For each candidate structure, we search for the locations of all peaks that would be expected based on crystal symmetry and lattice parameters. We then use the following equation (from Pecharsky & Zavalij, equation 8.41) to compute the intensity of a peak located at (hkl): (10)

$$I_{hkl} = K \times m_{hkl} \times LP(\theta) \times T_{hkl} \times |F_{hkl}|^2 \qquad \text{(equation 1)}$$

where K is the scaling factor, m_{hkl} is the number of lattice planes corresponding to the peak, $LP(\theta)$ is the Lorentz-polarization factor at the peak's diffraction angle, T_{hkl} is a March-Dollase function used to describe the grain orientation distribution, (34) and F_{hkl} is the structure factor, which involves atomic positions. In order to compare this diffraction pattern to the experimental pattern, we use the integrated peak intensities provided by the PDF4+ software to compute R-Factor:

$$R = \frac{\sum_{peaks}(I_{calc} - I_{obs})^2}{\sum_{peaks} I_{obs}^2} \qquad \text{(equation 2)}$$

where I_{calc} and I_{obs} are the candidate structure's calculated peak and experimentally reported peak intensities, respectively. Peaks between the two patterns are paired together if they are within 0.15 degrees of each other; if there are multiple peaks within this range, then they are added together, and if there are no peaks within this range, then a peak of zero intensity is used. While Mint also has the capability of performing Rietvield refinement to obtain R-Factor for continuous patterns, we do not utilize this feature, since many of the PDF entries we attempted to solve only provided

integrated diffraction peaks. Prior to reporting the diffraction pattern match for any candidate structure, we first refine the atomic coordinates of the structure in order to minimize R-Factor. For refinement, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm implemented in Dlib, (35) during which we optimize the free parameters of Equation 1, including atomic positions, thermal factors, and texturing.

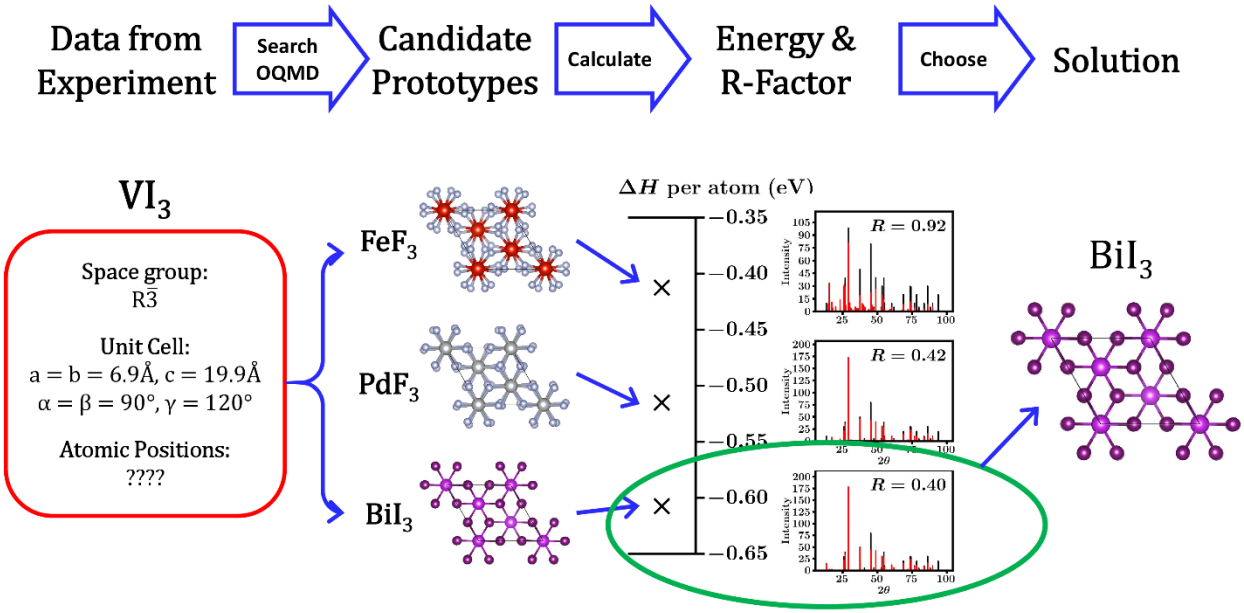

Figure 1: Flow chart of the prototype searching method to solve structures. The compound VI_3 (PDF #: 00-023-0719) is presented here as an example. Using experimentally determined structure attributes absent atomic coordinates, we search the OQMD for all prototypes with the space group ($R\bar{3}$), stoichiometry (AB_3), and formula units per cell (6). We then evaluate each of the three prototypes found (FeF_3-type, PdF_3-type, BiI_3-type) using DFT and R-Factor. We find that the BiI_3 prototype is the most plausible solution because it has the lowest formation energy and R-Factor.

iii. Choosing a Structure as the Solution

After computing the energy and R-Factor of each candidate structure, we select the best performing structure as the final solution. To do so, we take all candidate structures with an R-Factor within 0.2 above the lowest R-Factor found, and then select the lowest-energy structure among these. We then decide whether the final solution is valid, based on the values of energy and R-Factor; we provided a detailed description of the validation procedure in the Results section.

A schematic diagram of our prototype searching method is given in Figure 1 for an example PDF entry, VI_3 (PDF# 00-023-0719), that contained a diffraction pattern, space group, and unit cell parameters, but no atomic coordinates. We obtain three candidate prototypes (FeF_3-type, PdF_3-type, BiF_3-type) from the OQMD, generate one structure of VI_3 for each prototype, and evaluate their DFT formation energies and R-Factors. The BiI_3-type structure has both the lowest formation energy and the lowest R-Factor and is thus the best-performing prototype of the three. The BiI_3-type structure also has sufficiently low energy and R-Factor according to our validation criteria (see Results section), and so we declare it to be the solution of the VI_3 measurement.

III. RESULTS

A. Prevalence of Prototypes among Known Inorganic Compounds

The OQMD contains DFT calculations of experimentally observed inorganic compounds from a 2011 version of the ICSD, excluding those with partial occupancy or very large unit cells. Using the definition of a prototype outlined in Section IIA, we build a database of all prototypes that exist among 36,807 nonduplicate, stoichiometric, and inorganic compounds in the ICSD portion of the OQMD. An exhaustive database like this one can be compared to existing prototype databases such as the one built from AFLOW. (36) (37) The AFLOW prototype database distinguishes prototypes in a similar manner, i.e. by space group, stoichiometry, and Wyckoff sites, but also distinguishes prototypes with different local geometries, e.g. C23, C25, C29, and C37. A key distinction of our prototype database is that it is exhaustive and contains many more prototypes than the 1,100 prototypes in AFLOW. From 36,807 nonduplicate, stoichiometric compounds, of which 7,852 are binary, 18,482 are ternary, and 8,076 are quaternary, we identify a total of 10,203 prototypes, of which 1,617 are binary, 4,120 are ternary, and 3,062 are quaternary. Although this implies that there is an average of 3.6 compounds per prototype, some prototypes are shared by

hundreds of compounds. In Figure 2, we plot the sorted number of compounds per prototype. There are 77 prototypes with fifty or more compounds, accounting for 27% of the total number of compounds in the ICSD set; these prototypes are listed in Table 1. Such prototype-sharing reflects that compounds with similar chemistries tend to arrange in the same or similar geometries. For example, binary compounds containing a cation and an anion most commonly have NaCl, $PbCl_2$, CaF_2, and CdI_2 prototypes, while half-Heusler and related prototypes are often observed for compounds with metals and metalloids where the sum of valence electrons equals 8 or 18. (38) (39)

While most compounds share common prototypes, there are also 6,394 prototypes, of which 981 are binary, 2,428 are ternary, and 1,928 are quaternary, that are associated with just a single compound in the ICSD. These "one-hit wonders" highlight a shortcoming of the prototype searching method for solving crystal structures: some structures, ~17% of the ICSD, are unique and cannot be solved by searching for already-known prototypes. We can acquire an insight about the one-hit wonders by investigating their statistics. For instance, we note that hydrogen is disproportionately represented among the one-hit wonders: 37% of compounds containing H are one-hit wonders, much higher than the average of 17%. Other elements that are disproportionately represented are N (34%), F (31%), and Xe (56%). Nonmetals and alkali metals in general are disproportionately represented ($\gtrsim$20%), while lanthanide and actinide elements rarely occur by themselves (<10%), except for La (15%) and U (18%). Previous studies have identified which element pairs commonly appear together in compounds of the same prototype. (40) (41) In addition, many of the one-hit wonders have unique stoichiometries, such as $Fe_{107}O_{125}$. We find that 941 compounds do not share the same stoichiometry with any other compound in the ICSD. Many-component compounds tend to be unique as well: 955 of 2,156 compounds (44%) with five

or more components are one-hit wonders. One-hit wonders tend to also have larger unit cells: 29% of compounds with forty or more atoms are one-hit wonders, compared to 16% of compounds between twenty and forty atoms and just 6% of compounds with fewer than twenty atoms. Furthermore, most space groups are rarely observed. We find that 159 of 230 space groups have an above average proportion of one-hit wonders (>17%), and 11 space groups are not observed at all. On the other hand, a select few space groups account for a much larger proportion of ICSD compounds. One such space group is $Fm\bar{3}m$, which is found in 1,464 compounds, of which only 33 (2%) are one-hit wonders.

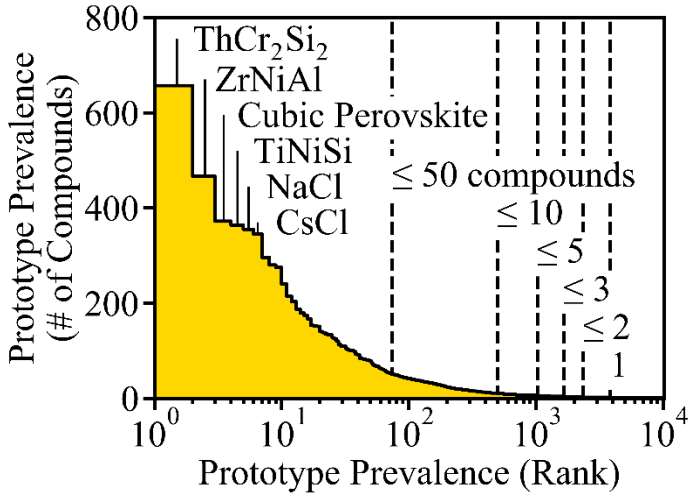

Figure 2: Sorted numbers of compounds associated with prototypes in the 2011 version of the ICSD, present in the OQMD. The total number of compounds in this ICSD set is 36,807. The most prevalent (rank 1) prototype is $ThCr_2Si_2$, with 657 compounds in the ICSD; and the second-most prevalent (rank 2) prototype is ZrNiAl, with 466 compounds. Beginning at rank 10 or so, the trend in prototype prevalence smoothly decays with a wide tail.

Prototype	Other Name	Typical Chemistry	Strukturbericht	Space Group	Wyckoff Sites	Number of Compounds
Binary						
NaCl	Rock Salt	MT + NMT	B1	$Fm\bar{3}m$	4a, 4b	355
CsCl		MT + MT	B2	$Pm\bar{3}m$	1a, 1b	345
Cu_3Au	Bogdanovite	MT + MT	$D0_9$, $L1_2$	$Pm\bar{3}m$	1a, 3c	296
$MgCu_2$	Laves	MT + MT	C9, C15	$Fd\bar{3}m$	8b, 16c	240
Mn_5Si_3	Mavlyanovite	MT + {PTM,MTO}	$D8_8$	$P6_3/mcm$	4d, 12g	167
CrB	Herzenbergite, Westerveldite	Variable	B33, B_c, B_f	Cmcm	8c	135
GeS		Variable	B14, B16, B27, B29, B31, B_d, B_m	Pnma	8c	120
$MoSi_2$		Variable	$C11_a$, C11b	I4/mmm	2a, 4e	110
$PbCl_2$	Cotunnite	Variable	C23, C25, C28, C29, C37	Pnma	12c	109
CaF_2	Fluorite	MT + NMT	C1, C_l	$Fm\bar{3}m$	4a, 8c	104
$CaCu_5$		MT + MT	$D2_d$	P6/mmm	1a, 2c, 3g	102
AlB_2	Hexagonal ω	MT + {PTM,MTO}	C32	P6/mmm	1a, 2d	101
NiAs		{LT,AT} + {PG,CG}	$B8_1$	$P6_3/mmc$	2a, 2c	84
ZnS	Zincblende	MT + {PTM,MTO,NMT}	B3	$F\bar{4}3m$	4a,4c	83
CdI_2	Khatyrkite	Variable	C6	$P\bar{3}m1$	1a, 3d	82
Th_3P_4		{LT,AT} + {PG,CG}	$D7_3$	$I\bar{4}3d$	12a, 16c	81
$MgZn_2$	Laves	MT + MT	C14	$P6_3/mmc$	2a, 4f, 6h	80
Cu_2Sb		{LT,AT} + {PG,CG}	C38	P4/nmm	2a, 4c	69
CuAu		MT + {MT,PTM,MTO}	$L1_0$	P4/mmm	1d,1d	67
Cr_3Si		{Groups 4-6} + {Groups 8-12,PTM,MTO}	A15	$Pm\bar{3}n$	2a, 6c	67
Fe_3C	Cementite	Variable	$D0_{11}$, $D0_{20}$	Pnma	8c,8d	63
FeS_2	Pyrite	MT + {PG,CG}	C2	$Pa\bar{3}$	4a, 8c	61
Th_2Ni_{17}		{LT,AT} + {Mg,Co,Fe,Ni,Al}		$P6_3/mmc$	2b, 2d, 4f, 6g, 12j, 12k	60
$CuAl_2$	Khatyrkite	MT + {MT,PTM,MTO}	C16	I4/mcm	4a, 8h	58
Ni_2In		MT + {PTM,MTO}	$B8_2$	$P6_3/mmc$	2a, 2c, 2d	54
Ternary						
$ThCr_2Si_2$		MT + {MT,PTM} + {PTM,MTO}		I4/mmm	2a, 4d, 4e	657
ZrNiAl		MT + MT + {PTM,MTO}		$P\bar{6}2m$	1b, 2c, 3f, 3g	466
$CaTiO_3$	Cubic Perovskite	MT + {MT,PTM,MTO} + NMT	$E2_1$	$Pm\bar{3}m$	1a, 1b, 3c	372
TiNiSi		MT + MT + {PTM,MTO}		Pnma	12c	364
Cu_2MnAl	Full Heusler	MT + MT + {PTM,MTO}	$L2_1$	$Fm\bar{3}m$	4a, 4b, 8c	280
PbFCl	Matlockite	Mixed anion, MT + {MT,PTM,MTO} + {PTM,MTO,NMT}	$E0_1$	P4/nmm	2a, 4c	275
$CrNaS_2$	Caswellsilverite	MT + {MT,PTM,MTO} + CG	$F5_1$	$R\bar{3}m$	3a, 3b, 6c	214
$CaTiO_3$	Orthorhombic Perovskite	MT + {MT,PTM,MTO} + NMT		Pnma	4a, 8c, 8d	203
LiBC		MT + MT + {PTM,MTO,NMT}		$P6_3/mmc$	2a, 2c, 2d	187
Mo_2FeB_2		MT + MT + {PTM,MTO}		P4/mbm	2a, 4g, 4h	174
Ce_2SO_2		MT + {MT,PTM,MTO} + {PTM,MTO,NMT}		$P\bar{3}m1$	1a, 4d	153
MgAgAs	Half Heusler	MT + MT + {PTM,MTO}	$C1_b$	$F\bar{4}3m$	4b, 4c, 4d	152
$SrCuO_2$		MT + MT + {Si,Ge,Sn}		Cmcm	16c	140
$AlMg_2O_4$	Spinel	MT + {MT,PTM,MTO} + CG	$H1_1$	$Fd\bar{3}m$	8b, 16c, 32e	138
Cr_2AlC	MAX phase	MT + {MT,PTM,MTO} + NMT		$P6_3/mmc$	2a, 2d, 4f	133
$CaBe_2Ge_2$		MT + LT + {PTM,MTO}		P4/nmm	2a, 2b, 6c	127
$FeSb_2S_4$	Berthierite	MT + LT + CG	$E3_3$	Pnma	28c	125
$Ca_2Nb_2O_7$	Pyrochlore	MT + {MT,PTM,MTO} + O	$E8_1$	$Fd\bar{3}m$	8a, 16c, 16d, 48f	113
K_2PtCl_6		AA + {MT,PTM,MTO} + {HG,H}	$J1_1$	$Fm\bar{3}m$	4a, 8c, 24e	110

$LiGaGe$		MT + {MT,PTM} + {PTM,MTO}		$P6_3mc$	2a, 4b	101
$YCrB_4$		{LT,AT} + TM + B		Pbam	8g, 16h	100
$Mg_6Cu_{16}Si_7$		{Groups 3-4} + {Ni,Cu,Al,Ga,Zn} + {Groups 8-14}		$Fm\bar{3}m$	4b, 24d, 24e, 64f	98
$CuHf_5Sn_3$		{Bi,Groups 3-7} + {Ti,Mn,Fe,Co,Ni,Cu,Zn,B,N,P,O,} + {MT,PTM,MTO}		$P6_3/mcm$	2b, 4d, 12g	93
Co_3GdB_2		{LT,AT} + {Groups 8-13} + {Ni,Pd,Cu,B,Al,Ga,Si}		$P6/mmm$	1a, 2c, 3g	84
$MgCu_4Sn$	Friauf-Laves	{LT,AT} + {Pd,Ag,Au,Zn,Cd,In,Sn} + {Co,Ni,Cu,Pt}		$F\bar{4}3m$	4a, 4d, 16e	82
$CeMn_4Al_8$		{LT,AT} + {Cu,Cr,Mn,Ni,Fe} + {Mn,Al,Co,Fe}		$I4/mmm$	2a, 8f, 8i, 8j	80
$Zr_2Fe_{12}P_7$		{LT,AT} + {Mn,Fe,Co,Ni} + {P,As}		$P\bar{6}$	1a, 1d, 1e, 9j, 9k	79
$FeO(OH)$	Goethite, Diaspore, Chalcostibite	Metal Hydroxides, MT + {MT,PTM,MTO} + {Groups 13-16}	$F5_6$	Pnma	16c	78
$KCuO_2$		MT + MT + {MT,PTM}		Cmcm	8c, 8f	74
$CePO_4$	Monazite	MT + {MT,PTM,MTO} + {O,H,HG}		$P2_1/c$	24e	71
Sr_2TiO_4	Ruddlesden-Popper	AA + MT + {CG,HG}		$I4/mmm$	2a, 4c, 8e	69
$Yb_3Rh_4Ge_{13}$		{LT,AT} + Group 9 + {Si,Ge,Sn,Pb}		$Pm\bar{3}n$	2a, 6d, 8e, 24k	65
K_2SO_4	Arcanite	AA + {MT,PTM,MTO} + {H,CG,HG}		Pnma	20c, 8d	62
$KAsF_6$		MT + {MT,PTM,MTO} + F		$R\bar{3}$	3a, 3b, 18f	62
$ZrSiO_4$	Zircon	{Groups 3-5} + O-ate	$S1_1$	$I4_1/amd$	4a, 4b, 16h	59
$CuFeS_2$	Chalcopyrite	MT + {PTM,MTO} + {MTO,PG,CG}	$E1_1$	$I\bar{4}2d$	4a, 4b, 8d	56
$[NH_4]CdCl_3$		MT + {MT,PTM} + {CG,HG}		Pnma	20c	55
$BaSO_4$	Barite	MT + {PTM,MTO,NMT} + NMT	$H0_2$	Pnma	16c, 8d	53
K_2UF_6		MT + {MT,PTM} + {PTM,MTO}		$P\bar{6}2m$	4a, 2b, 4d, 8e	52
$CaWO_4$	Scheelite	MT + O-ate	H4	$I4_1/a$	4a, 4b, 16f	52
$U_2Co_3Si_5$		{LT} + {Groups 8-10} + {Si,Ge}		Ibam	4a, 4b, 8g, 24j	51
YNi_5Si_3		MT + {MT,PTM,MTO} + {MTO,NMT}		Pnma	36c	50
$BaNiSn_3$		{Groups 1-3} + {Groups 8-10} + {Si,Ge,Sn}		I4mm	6a, 4b	50
$Cu_4Gd_3Ge_4$		{LT,AT} + {Ni,Pd,Cu,Ag} + {Si,Ge,Sn}		Immm	2a, 4h, 4i, 4j, 8l	50
Quaternary						
K_2NaAlF_6	Elpasolite	AA + MT + {MT,PTM} + {F,Cl,O}		$Fm\bar{3}m$	4a, 4b, 8c, 24e	179
La_3CuSiS_7		LT + TM + {PTM,MTO} + {S,Se}		$P2_1$	2a, 4b, 18c	151
La_2LiSbO_6		AA + {LT,AT} + {MT,PTM} + O		$P2_1/c$	2b, 2c, 16e	133
$KAlP_2O_7$		MT + MT + pyro-O-ate		$P2_1/c$	44e	92
$KCuZrS_3$		AA + {LT,AT} + {Groups 11-12} + {S,Se,Te}		Cmcm	4b, 12c, 8f	91
$CuZrSiAs$		MT + MT + 2 Anions		$P4/nmm$	2a, 2b, 4c	67
$K_3NaFeCl_6$	Rinneite	AA + MT + {MT,PTM} + O		$R\bar{3}c$	6a, 6b, 18e, 36f	50
Quinary						
$[NH_4]Mg[SO_4]_2 \cdot [H_2O]_6$		AA + MT + H + {S,Se} + O		$P2_1/c$	2a, 60e	54

Table 1: Prototypes shared by 50 or more unique compounds sourced from ICSD in the OQMD. Notation under "Typical Chemistry": MT = metal (groups 1-12), NMT = nonmetal, AA = alkali or alkaline earth metals, TM = transition metal (groups 3-12), PTM = post-transition metal, MTO = metalloid, LT = lanthanide, AT = actinide, PG = pnictogen, CG = chalcogen, HG = halogen. "+" denotes AND and "{…}" denotes OR. The "Strukturbericht" designations are a commonly used classification of crystal structures that is similar to our classification based on prototype. Not every prototype as we've defined it has a corresponding Strukturbericht designation, and some prototypes have multiple Strukturbericht designations.

B. High-Throughput Structure Solution by Prototype Searching

i. Description of Target Compounds from the Powder Diffraction File

As the prototype searching method is cheap, often costing only a few DFT calculations, we leverage this approach to perform "high-throughput" structure solution for numerous entries from the International Centre for Diffraction Data (ICDD) database within the PDF for which the atomic coordinates are missing but other structure details are known. We start with 80,624 entries missing atomic coordinates in the 2018 version of the PDF4+ software. We screen for entries that satisfy the following criteria:

- Entry is "primary" status as identified by the PDF4+ software, i.e. is not an alternative to another similar entry.
- Diffraction experiment was performed under ambient conditions. We note that the enthalpy of high-pressure compounds can be accounted for in DFT by supplying external pressure to the stress tensor. Furthermore, the enthalpies of high-pressure compounds can be compared to those of other compounds in the OQMD. (42)
- Compound is binary, ternary, or quaternary.
- Compound is inorganic and does not contain noble gases, actinides, or radioactive elements.

- Diffraction data quality is listed as “star,” “good,” or “indexed,” indicating that the diffraction pattern represents a single-phase crystal with minimal impurities. While structures with poorer quality diffraction patterns can still be solved, their R-Factors may be less useful.
- Space group and number of formula units per unit cell are already known.
- Reduced cell volumes are less than 3000 $Å^3$ and unit cells contain few enough atoms to be cheaply assessed by high-throughput DFT:
    - Cubic, hexagonal, trigonal, and tetragonal cells contain 80 or fewer atoms.
    - Orthorhombic cells contain 40 or fewer atoms.
    - Monoclinic and triclinic cells contain 20 or fewer atoms.
- The structure does not evidently contain partially occupied sites, i.e., the listed composition contains only natural numbers and it is possible to generate a structure with a set of fully occupied Wyckoff sites given the listed space group and number of formula units per unit cell. We note that a PDF entry satisfying these conditions may not necessarily represent a fully occupied structure. We can justify the validity of our prototype structures based on DFT energy and R-Factor.
- There is no existing OQMD nor ICSD compound with the same composition and space group.
- There is at least one prototype in the OQMD matching the known stoichiometry, space group, and number of formula units per unit cell.

We find 603 PDF entries that satisfy the above constraints. We additionally find hundreds of entries that satisfy all the above constraints except for the last one, i.e., there is no prototype in the OQMD that matches the provided stoichiometry, space group, and number of atoms per unit cell. However, it is possible that the listed space group is incorrect, and hence we attempt to solve these

by searching within the crystal system, e.g., tetragonal space groups, instead of the listed space group.

ii. Summary of Structures Obtained by Prototype Searching

For 603 PDF entries with diffraction data but no structure, we find at least one prototype in the OQMD that matches the provided space group, stoichiometry, and number of formula units per unit cell. In Figure 3a, we plot a distribution of the number of prototypes found per PDF entry. The highest number of prototypes is only ten. In most cases (386, or 64%), only one prototype is found. Although the number of candidate prototypes is almost always very few, each prototype can produce multiple structures representing the possible ways to arrange elements onto the prototype's Wyckoff sites. Despite this, there are rarely more than a dozen structures to evaluate (see Figure 3b).

After computing the DFT energy and R-Factor of all candidate structures, we select the structure with the lowest DFT energy among all candidates within 0.2 of the lowest R-Factor. We are thus left with 603 structure candidates, each one outperforming other candidates for every attempted PDF entry. For 10 of the 603 entries, we find a candidate with a different space group within the same crystal system that outperforms all candidates with the reported space group. In these 10 cases, the structure with the same space group fails our validation checks of energy and R-Factor (described in the following section), while the structure with a different space group passes these checks; we thus opt to present the 10 structures with a different space group. In addition, we find that for 21 of the PDF entries, while there is no prototype in the OQMD that matches the reported space group, stoichiometry, and number of formula units per cell, there is a candidate with a different space group within the same crystal system that passes our validation

checks. In total, we present 624 structures (603 + 21) in the following analysis. Of these, 520 pass our validation checks of energy and R-Factor, and we thus declare them to be solved.

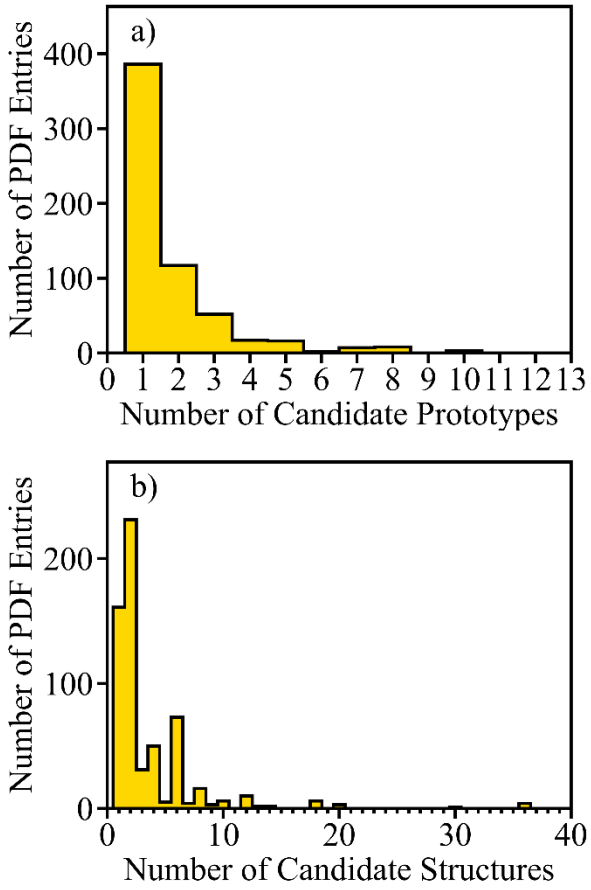

Figure 3: Distribution of the number of a) OQMD prototypes and b) candidate structures matching stoichiometry, space group, and number of formula units per unit cell of 603 PDF entries with missing atomic coordinates.

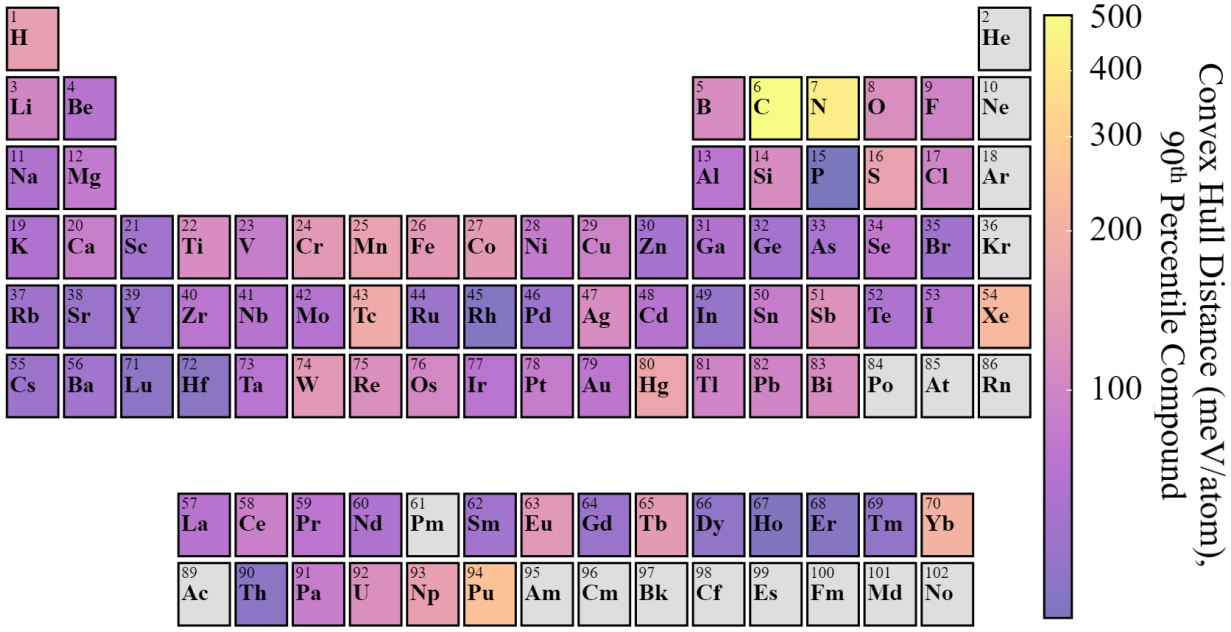

Figure 4: DFT-computed 90th percentile convex hull distance (meV/atom) of ICSD compounds containing each element on the periodic table. The metastability of materials is chemistry dependent, with carbides and nitrides standing out as being particularly high in energy. Gray shaded elements are excluded from this analysis.

iii. Analysis of Structures Obtained by Prototype Searching

After selecting the 624 best-performing candidate structures, one for each PDF entry, we proceed to assess their validity by examining the values of energy and R-Factor. As for energy, we are interested in the difference between the structure's formation energy and the OQMD convex hull at the relevant composition with this structure included. If this value is 0 meV/atom, then our candidate structure is stable and thus highly plausible. However, metastable compounds with nonzero convex hull distances are also common in nature. Although not all hypothetical structures with nonzero convex hull distances can be synthesized, they should be considered as potentially valid solutions in our structure search. Analyses of experimentally known metastable compounds calculated by DFT have revealed that, while most metastable compounds are within 100 meV/atom of the convex hull, the values of convex hull distance are highly dependent on chemistry. (43) (44) In Figure 4, for each element, we plot the 90th percentile convex hull distance for ICSD compounds

containing that element. There is a stark contrast in the convex hull distances as a function of element; carbides and nitrides are especially metastable. (45) (46) (47) We thus opt to use these values of convex hull distance as cutoff values in determining whether the structures we obtain from the prototype searching method are valid based on energy. Specifically, for a compound of interest, e.g. Ba_2CeSnO_6 (PDF #: 00-056-0332) solved in this work, we use the highest of the four 90^{th} percentile convex hull distance values as the cutoff: 116 meV/atom for oxygen. Since our best-performing candidate structure for Ba_2CeSnO_6 is 102 meV/atom above the convex hull, we deem this structure valid based on energy.

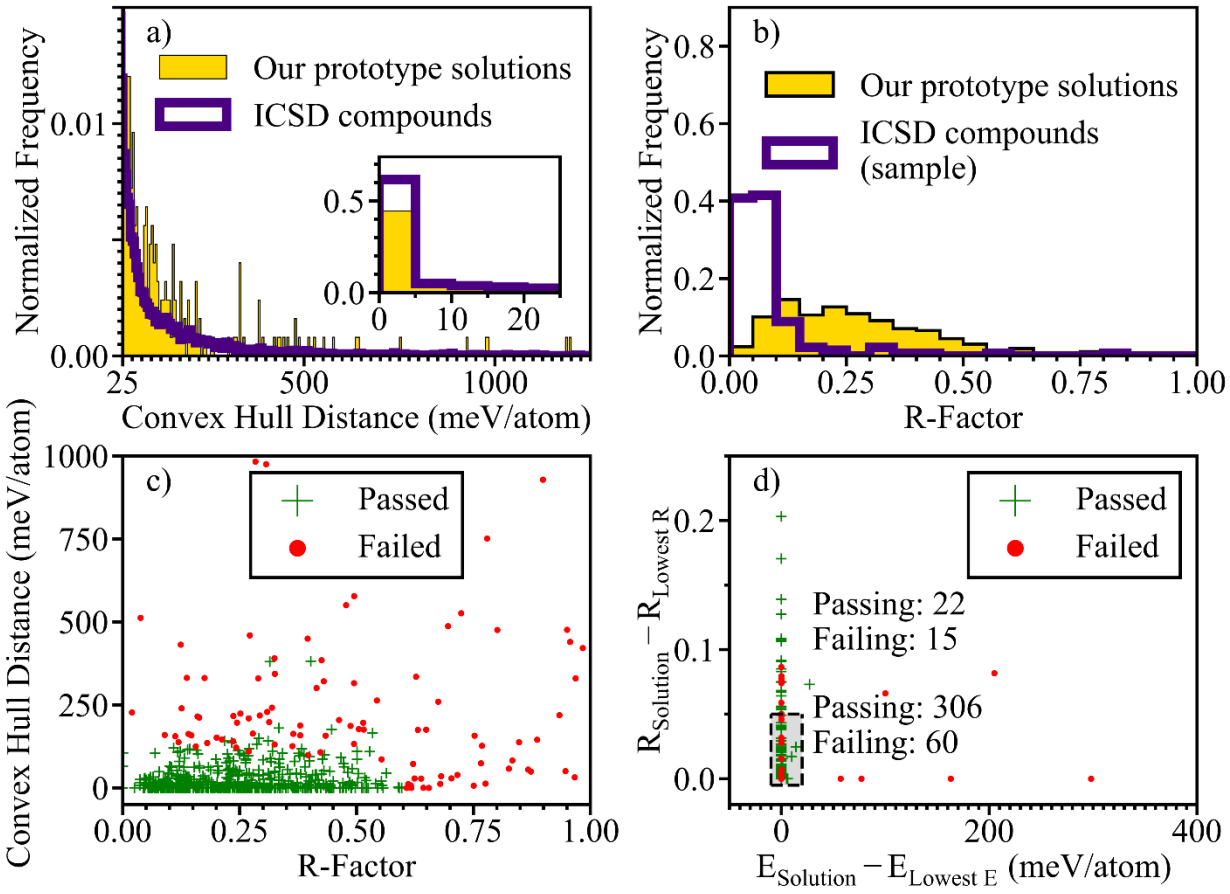

Figure 5: a) Distribution of convex hull distances of 624 compounds with structures obtained by prototype searching in this work, compared to 23,247 ICSD compounds that have been calculated in the OQMD. Inset is the same distribution between 0 and 25 meV/atom; almost half of the 624 compounds lie within 5 meV/atom of the convex hull, somewhat shy of the ICSD. b) Distribution of R-Factors of 624 compounds with structures obtained by prototype searching in this work, compared to 136 randomly chosen solved compounds from the PDF. c) Convex hull distances and R-Factors of 624 compounds with structures obtained by prototype searching. Green pluses and red dots are compounds that passed and failed validation checks, respectively. d) Discrepancies in the best-performing candidate structure energies and R-Factors with the lowest-energy and lowest-R candidate structures. Only cases with multiple candidate structures are shown (403 of 624 PDF compounds). The vast majority (91%) of cases lie within the shaded region; in other words, the best-performing candidate structure usually is close to the lowest energy and lowest R-Factor of all candidate structures. Cases with low discrepancy are also more likely to pass our validation checks (84%) than cases with high discrepancy (59%).

In Figure 5a, we plot the convex hull distances of 624 compounds with structures obtained by prototype searching in this work, along with those of ICSD compounds. If all 624 of these compounds were correctly solved, then we would expect that their convex hull distances would line up well with the ICSD distribution. Although the proportion of our compounds that lie on the convex hull is high (277 compounds within 5 meV/atom of the hull), this proportion is shy of the ICSD, where 61% of compounds are within 5 meV/atom of the convex hull. We find that 543 compounds (87%) pass our validation criterion for energy, compared to 93% of the ICSD. However, we also expect that the prototype searching method will fail to solve compounds that are "one-hit wonders," i.e. compounds that do not share a prototype with any other in the ICSD (described in Section IIIA). Since as many as 17% of known compounds are one-hit wonders, this inevitable shortcoming of our approach might explain why our convex hull distances are higher than those of ICSD compounds, on average.

R-Factors of all 624 compounds are plotted in Figure 5b. To give context to our R-Factor values, we overlay a distribution of R-Factors for 136 randomly selected already-solved ICSD compounds with diffraction patterns stored in the PDF. With a median value of 0.25, our structures have higher R-Factors overall than the already-solved structures (median of 0.06). We argue that

this discrepancy does not suggest an issue with our prototype solutions, because many of our solutions with R-Factor greater than 0.05 are clearly the right answers by inspection. For example, 44 of our compounds evidently have the elpasolite (K_2NaAlF_6) prototype, since they have A_2BCD_6 stoichiometry, space group of $Fm\bar{3}m$, and four formula units per unit cell. The only other possible prototype is typically very high in energy. Indeed, we find that 17 of the elpasolite compounds lie on the convex hull, despite R-Factors ranging from 0.05 to 0.52. The high R-Factors are not due to any issue with our refinement code either; despite elpasolite having only one degree of freedom to refine (the x coordinate of the 24e site), we still obtain high R-Factors. We argue that the high R-Factors highlight an issue with the diffraction patterns, not with our prototype searching approach. Because we cannot impose a strict R-Factor validation criterion, we look to the relationship with energy values to decide on a cutoff R-Factor value. Stable compounds tend to have low R-Factors: 51% of compounds with R-Factor below 0.1 lie on the convex hull; 54% with R-Factor between 0.1 and 0.2 lie on the convex hull; 50% between 0.2 and 0.3. Following these intervals, we have 41%, 36%, 36% between 0.5 and 0.6, 24% 0%, 0%, and 0% between 0.9 and 1.0. As the proportion of stable compounds begins dropping off at 0.6, we opt to use an R-Factor of 0.6 as the cutoff value for validation. This works out to be a generous cutoff value: 580 of our compounds (93%) have R-Factor less than 0.6.

Combining our validation checks, we declare that 520 of 624 (83%) of our compounds are "solved" based on low convex hull distance and R-Factor less than 0.6. The convex hull distances and R-Factors of all 624 compounds are plotted in Figure 5c. Although most of our compounds simultaneously pass both energy and R-Factor validation criteria, there are cases that pass only one of the criteria. Compounds with high energy and low R-Factor might have structures that happen to exhibit a close match to diffraction data while being theoretically unphysical. On the

other hand, compounds with low energy and high R-Factor could be polymorphs of the "true" structure observed in experiment. It is also possible that compounds with low energy and high R-Factor are, in fact, correctly solved; indeed, we are using an atypically high cutoff for R-Factor. Despite the high R-Factors, we argue that the R-Factors are helpful in distinguishing structures that best match experimental data. In Figure 5d, we demonstrate that even though many of our structures have high R-Factor, they are most often both the lowest-energy and lowest-R-Factor candidate out of all possible candidates. Considering 403 cases where more than one possible candidate structure exists, we find that 366 (91%) of our best-performing candidates lie within the shaded region, i.e. are within 20 meV/atom of the lowest-energy candidate and 0.05 of the lowest-R-Factor candidate. Compounds that pass our validation criteria are even more likely to lie within the shaded region (93%) than failing compounds (80%). This result demonstrates that even when all candidate structures have high R-Factor, we can still use R-Factor to distinguish the best structure from other candidates; however, DFT energy is often helpful in determining which candidates are physical.

Upon inspecting our prototypes selected for the PDF compounds, we noticed that they are quite often chemically similar to other ICSD compounds with the same prototype. For example, the solution to Ag_7SbS_6 (PDF #: 00-021-1333) is the prototype of Ag_7AsS_6, found in ICSD. We can quantify "chemical similarity" by taking advantage of a data mined Pettifor chemical scale developed by Glawe and co-workers. (41) They computed a chemical similarity metric P_{AB} for all pairs of elements A and B on the periodic table. To compute the chemical similarity between two compounds, e.g. Ag_7SbS_6 and Ag_7AsS_6, we take the product $P = \prod P_{AB}$ of chemical similarities of the closest-matching element pairs in the two compounds, setting P_{AB} to 1 when the elements are identical and 0 if the element pairs rarely or never occur in the ICSD. For all of our chosen

prototypes, we searched for the ICSD compound of the same prototype with the highest chemical similarity; the results are plotted in Figure 6. The trends in the plots demonstrate that compounds that pass our validation criteria are more likely to be chemically similar to ICSD compounds than compounds that fail. The chemical similarities we find here give us an extra layer of confidence in our solutions.

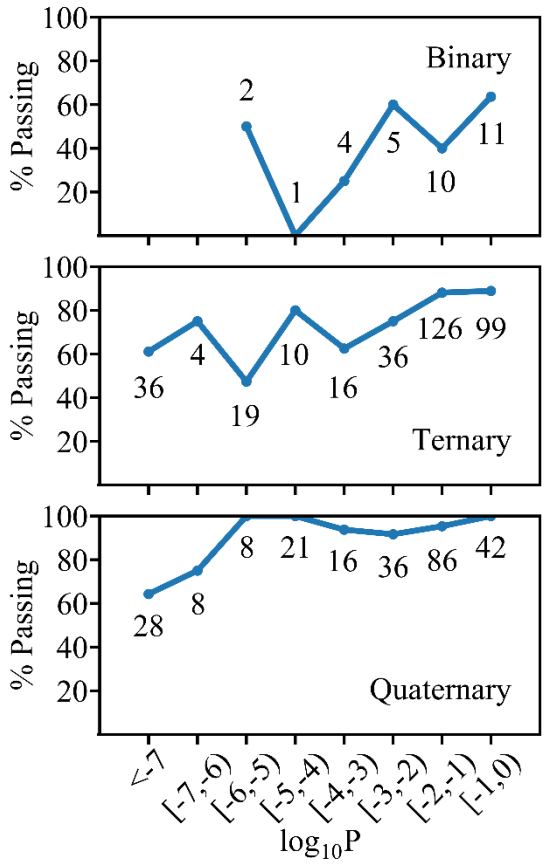

Figure 6: The percentage of the prototypes chosen for each of 624 PDF compounds that pass our validation criteria, plotted against the chemical similarities *P* of these compounds to ICSD compounds of the same prototype. The chemical similarities *P* are binned by decades on a log scale; see the text of section IIIBiii for the definition of *P* between a pair of compounds. The top, middle, and bottom plots focus on binary, ternary, and quaternary compounds, respectively. The numbers of compounds that fall within each range of chemical similarities are shown beside the data points. The trends demonstrate that compounds that pass our validation criteria are more likely to be chemically similar to ICSD compounds than compounds that fail.

All 520 compounds solved in this work are provided in the Supplemental Material, (25) along with a complete tabular summary of all 624 attempts. In addition, all compounds can be found in the OQMD, which can be accessed via the web at oqmd.org or directly downloaded. As there are 36,807 unique ICSD compounds already in the OQMD, we have expanded the set of all experimentally observed compounds in the OQMD by 1.4%. The simplicity and efficiency of the prototype searching method presented in this paper has thus enabled us to significantly expand the set of experimentally observed compounds accessible to DFT. It will be of interest to further study the properties of these materials. For example, as shown in Figure 7, 283 of our solved compounds have nonzero bandgaps within 4 eV, making them potential candidates for semiconductor applications. In addition to the 520 newly solved compounds, we find 33 PDF "unsolved" compounds where there is either no matching prototype in the OQMD or no prototype matching the reported space group that passes our validation checks, but there is solution with a different space group within the same crystal system that not only passes our validation checks but also already exists in the ICSD. We provide these 33 solutions in a separate table in the Supplemental Material. (25)

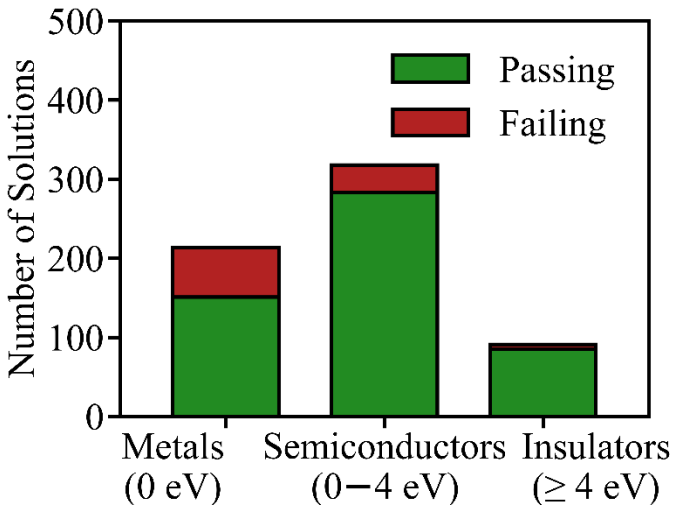

Figure 7: 151 of our solutions that pass validation criteria are metals (0 eV), 283 are semiconductors (0-4 eV), and 85 are insulators (≥4 eV); band gap was not determined for 1 solution.

iv. Examples of Solutions Obtained by Prototype Searching

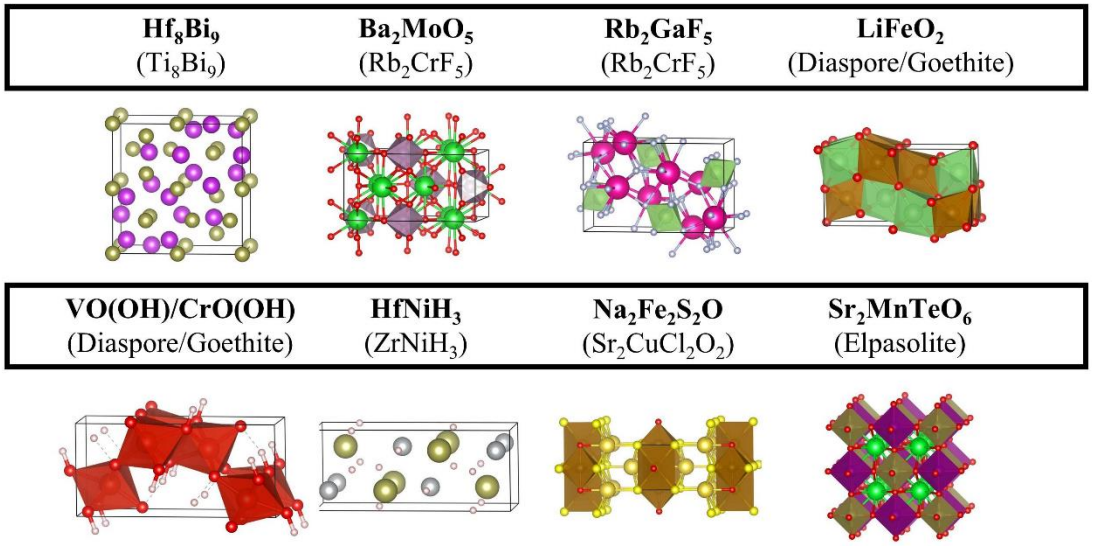

Figure 8: Crystal structures of 9 of the 520 materials solved using prototypes. The compositions of the solved materials are in bold, and the prototypes are in parentheses. Note that some of the solutions presented here have the same prototype, specifically Ba_2MoO_5 and Rb_2GaF_5 as well as $LiFeO_2$, VO(OH), and CrO(OH).

In this section, to demonstrate the prototype searching method at work, we discuss nine PDF compounds that we solved. An illustration of the solved compounds is shown in Figure 8. All of the nine compounds in this section pass our validation criteria of energy and R-Factor and are chemically similar to other compounds in the ICSD with the same prototype. For some of these compounds, the paper describing the diffraction experiment stated the name of the prototype that matches our solution but did not present atomic coordinates. Although the prototypes of these compounds were already known, our prototype searching method enabled us to obtain atomic coordinates for all structures and expand the OQMD.

1. Hf_8Bi_9

In the PDF entry for Hf_8Bi_9 (#: 00-051-0679), a diffraction pattern is supplied along with aspace group (P4/nmm), unit cell, and formula units (Z = 2), but atomic coordinates are missing. (48) Because the atomic coordinates are missing, this compound did not previously exist in the ICSD nor OQMD and has thus been excluded from DFT studies. However, in the reference for this entry, the authors presented the then-new prototype Ti_8Bi_9, complete with atomic coordinates, and stated that Hf_8Bi_9 has the same prototype as Ti_8Bi_9. As Ti_8Bi_9 is indeed already in the OQMD, we use the prototype searching method to complete the structure of Hf_8Bi_9. Specifically, our crystal structure for Hf_8Bi_9 consists of the unit cell parameters provided by the PDF entry for Hf_8Bi_9, and the DFT-relaxed atomic coordinates of Bi plus the atomic coordinates of Hf substituted for Ti in the already-solved compound Ti_8Bi_9. We find that this structure matches the diffraction pattern well (R-Factor = 0.21, see Figure 9a) and is on the convex hull (Figure 9b).

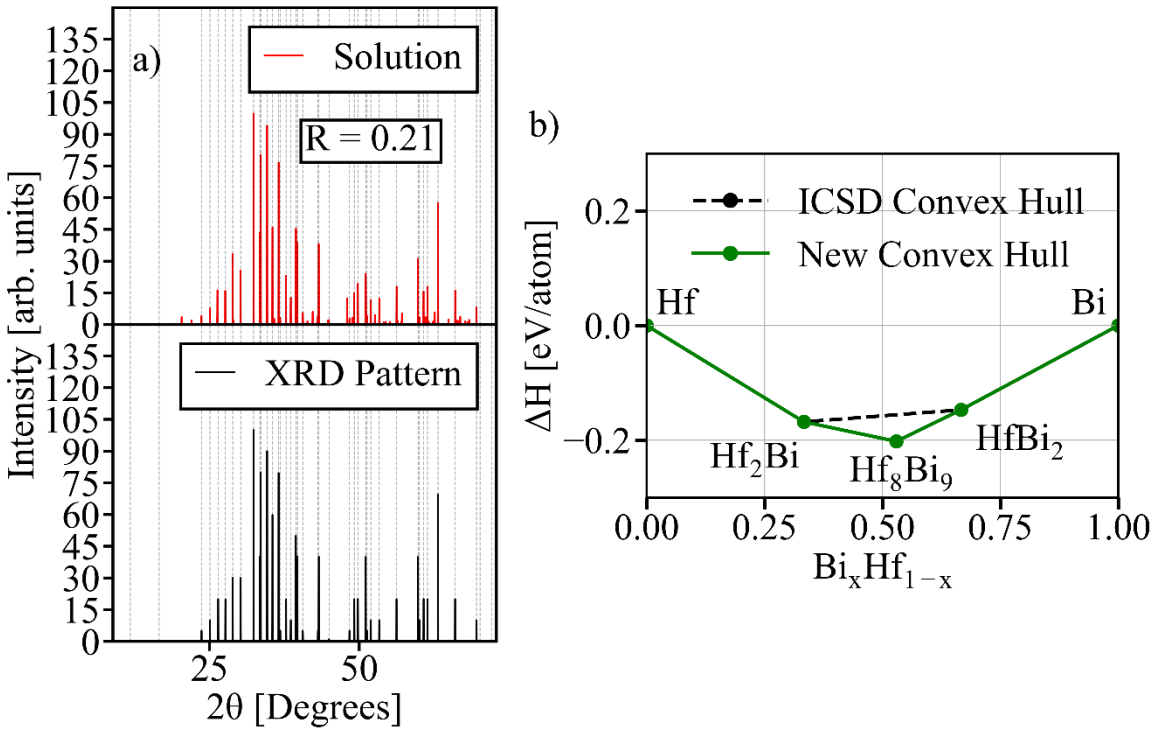

Figure 9: a) Diffraction pattern describing the prototype solution Ti_8Bi_9 for the compound Hf_8Bi_9 from the PDF (top), compared to the experimental XRD pattern reported in the PDF for Hf_8Bi_9 (bottom). The R-Factor is 0.21, highlighting a close match between the two patterns. b) "New" Hf-Bi convex hull including Hf_8Bi_9 solution generated from DFT calculations stored in the OQMD, plotted against the "old" convex hull found in the OQMD not including Hf_8Bi_9. Since our Hf_8Bi_9 solution was lower in energy by 46 meV/atom than the old convex hull at that composition, we construct a new convex hull to include our solution.

2. Ba_2MoO_5 and Rb_2GaF_5

The reference for Ba_2MoO_5 provided by the PDF (#: 00-025-0011) for this diffraction pattern describes the structure as isostructural with $K_2VO_2F_3$ with Pnma symmetry and 4 formula units per unit cell but does not provide atomic coordinates. (49) The OQMD prototype Rb_2CrF_5 is indeed isostructural with $K_2VO_2F_3$ in that it has the same space group and Wyckoff site occupancies (though with different stoichiometry), and, with the elements Ba, Mo and O substituted in, lies on the Ba-Mo-O convex hull and has an R-Factor of 0.16, indicating it is a highly plausible solution. An existing ICSD compound, Ba_2WO_5, has the same prototype and is highly chemically similar to Ba_2MoO_5 (P = 0.25). We also considered three other candidate prototypes: $BaSi_2O_5$ (convex

hull distance = +19 meV/atom, R-Factor = 0.28), KPd_2F_5 (hull distance = +102 meV/atom, R-Factor = 0.42), and CsN_2H_5 (DFT failed to converge, R-Factor = 0.19). Since the Rb_2CrF_5 prototype has both the lowest energy and lowest R-Factor out of all candidates and passes our validation criteria, we deem it to be the solution to Ba_2MoO_5.

The compound Rb_2GaF_5 from the PDF (#: 00-032-0914) has the same story. (50) The prototype Rb_2CrF_5 is the solution because it is on the convex hull and has R-Factor of 0.43, lower than other candidates. The ICSD compound Rb_2FeF_5, with the same prototype, is the most chemically similar to Rb_2GaF_5 (P = 0.04).

3. $LiFeO_2$ polymorph, VO(OH), and CrO(OH)

While several polymorphs of $LiFeO_2$ are known, to our knowledge there are no reports of the atomic coordinates of the Pnma polymorph of $LiFeO_2$ listed in the PDF (#: 00-052-0698), and consequently its structure was not previously in the OQMD. The reference listed in the PDF reported that the goethite polymorph of $LiFeO_2$ is rechargeable in lithium cells. (51) We find that goethite, or FeO(OH), is the correct prototype for this polymorph of $LiFeO_2$ when Li atoms are substituted for H, since the convex hull distance is only +36 meV/atom and the R-Factor is 0.13. We reject another candidate, YPd_2Si (convex hull distance = +272 meV/atom, R-Factor = 0.29), because it is highly unstable.

Single crystals of VO(OH) (PDF #: 00-011-0152), found in montroseite, were examined by x-ray crystallography in 1953 were found to be isostructural with diaspore, or AlO(OH). (52) Diaspore and goethite are the same prototype (as is chalcostibite). An incomplete structure for VO(OH) having only V and O positions can be found in the ICSD; (53) hydrogen positions are missing, presumably since they cannot be detected in the x-ray pattern, and as a result, the properties of VO(OH) have not been studied with DFT. We obtain a complete structure for

VO(OH), including H positions, by substituting V, O, and H into the sites of the diaspore structure and find it to be nearly stable (convex hull distance = +8 meV/atom, R-Factor = 0.48). We similarly apply our prototype searching method to fill in the H coordinates of the CrO(OH) structure (PDF#: 00-025-1497), which was previously found to resemble diaspore. (54) Our structure for CrO(OH) is close to the convex hull (+8 meV/atom), but has poor match to diffraction pattern (R-Factor = 0.62).

4. $HfNiH_3$

We report several stable hydrides in this work, including four lanthanide hydrogen chalcogenides. It is tricky to solve the hydrogen positions from x-ray diffraction data since hydrogen scattering is too weak to detect in an x-ray diffraction pattern. In the case of $HfNiH_3$ (PDF #: 00-047-1412), the peak indices could be matched to those of space group Cmcm. The authors inferred that the H atoms situate within the HfNi structure (space group Cmcm, 8 atoms per unit cell). (55) Separate DFT studies of $HfNiH_3$ utilized the assumption that H atoms occupy octahedral and tetrahedral interstices between Hf and Ni atoms in order to estimate the positions of H. (56) (57) We find ten unique prototypes having Cmcm space group and 20 atoms per cell, but the best performing prototype is that of $ZrNiH_3$ (convex hull distance = 0, R-Factor = 0.45). This is indeed a superstructure of HfNi, in which nine Hf-H bonds constitute edge-sharing polyhedra. Notably, the other nine prototypes with much higher energy are not hydrides. The $ZrNiH_3$ structure in the OQMD, complete with H positions, was obtained using neutron diffraction; (58) we utilize the solution from this past neutron diffraction study to complete the structure of $HfNiH_3$.

5. $Na_2Fe_2S_2O$

The diffraction pattern for the mixed anion compound $Na_2Fe_2S_2O$ was obtained through an ICDD Grant-In-Aid (PDF #: 00-065-0329) (59). The atomic positions are missing from the entry, but the space group and number of formula units were reported to be I4/mmm and Z = 2, respectively. We conclude that the $Sr_2CuCl_2O_2$ prototype is a convincing solution. Since there are 3! = 6 unique ways to arrange the elements Na, Fe and S onto the 4c, 4e and 4e Wyckoff sites of the $Sr_2CuCl_2O_2$ prototype, we check each one individually and find that the best arrangement is on the convex hull and has R-Factor of 0.21. Interestingly, this arrangement has cation Na^{1+} occupying the anion Cl^{-1} site of $Sr_2CuCl_2O_2$, and likewise has anion O^{2-} occupying the cation Cu^{2+} site. Such an arrangement could be a direct consequence of the balancing of oxidation states in $Na_2Fe_2S_2O$. Other arrangements are significantly higher in energy, so they are ruled out.

6. Novel Elpasolites

Many materials presented in this work share the same prototypes with one another. Forty-four of the materials in this work have the elpasolite structure, or K_2NaAlF_6, which is an ordered double perovskite. Elpasolite is one of two prototypes that are possible given the experimentally known $Fm\bar{3}m$ space group, ABC_2D_6 stoichiometry, and 40 atoms per unit cell. The other possibility is the same as elpasolite but with the D_6 atoms occupying the '24d' Wyckoff site rather than the '24e' Wyckoff site; this prototype is rare in the ICSD and is typically higher in energy by 1000-2000 meV/atom. Elpasolite is the most common quaternary prototype in nature, with 179 examples from the ICSD subset of the OQMD. All of the elpasolite-type compounds we present here are within +114 meV/atom of the OQMD convex hull (22 are on the hull), and have R-Factors below 0.52 (28 had R-Factors below 0.20), indicating that they were all stable or metastable and had reasonable pattern matches. For the metastable cases, the ground state is often a distortion of double perovskite; in the case of Sr_2MnTeO_6 (PDF #: 00-029-0897), the ground state is monoclinic

(P21/c) double perovskite, which is 24 meV/atom lower in energy than the elpasolite decoration. Recently there has been interest in identifying more elpasolite compounds. It is difficult to perform high-throughput DFT calculations of elpasolite structures using elemental substitution, since there are millions of permutations. Faber et al. developed a machine learning model to predict the energies of elpasolite compounds, and found 90 structures on the convex hull, after considerable model training and DFT calculations of 2,133 candidates. (60) We note that one of our 44 elpasolites is in their set of 90: Cs_2KGaF_6 (PDF #: 00-021-0849).

IV. DISCUSSION

Structure solution is a challenging roadblock to materials discovery. Thankfully, crystal structures are rarely unique, and a successful structure solution can often be obtained by searching among a relatively small number of prototypes as valid candidates. We apply this simple and inexpensive strategy to solve 520 structures taken from the PDF. Utilizing the OQMD as an exhaustive database of prototypes as well to validate the energetic stability of candidates along with R-Factor, we have identified potential solutions to these materials, and we have a high degree of confidence in our solutions.

The prototype strategy employed in this work can be improved upon in many ways. One way is to tweak the definition of a prototype to distinguish different structures more effectively. In our approach, we define the prototype of a structure as the combination of its stoichiometry, space group, and Wyckoff site occupancies. All structures from the OQMD sharing these characteristics are grouped into one prototype. However, within these constraints, there can be many degrees of freedom in atomic coordinates and lattice parameters, and it is possible for two structures with the

same prototype, as defined in this paper, to in fact have very different local geometries, a problem described at length by Trimarchi *et al*. (61) Our workaround is to choose the OQMD compound whose structure, with its elements replaced by the target elements, gives the lowest R-Factor, since the calculation of R-Factor is nearly instantaneous compared to DFT. A more reliable workaround would be to devise a stricter prototype definition capable of properly distinguishing structures with different local geometries. For instance, some definitions apply additional restrictions on unit cell axial ratios and angles. (62) One could also quantify the difference between structures using a distance metric, such as one devised from radial distribution functions (63) or atomic/molecular matching algorithms (64) (65) (66). Moreover, if a given prototype has many internal degrees of freedom, one could conceivably develop an algorithm to optimize DFT and R-Factor within the search space of that prototype.

Another way to improve the performance of the prototype searching method is to recommend the most plausible prototypes first, prior to evaluating them with DFT. There was no need to do so for this work, since constraining the search to the PDF-provided space group, composition, and number of atoms per unit cell of all solved materials reduced the number of candidate prototypes fewer than three in most cases. If, on the other hand, we could not constrain the search as much, there would have been too many candidates to evaluate. Existing techniques for recommending prototypes as candidates for an unsolved compound involve machine learning (67) as well as data-mined ion substitution (40).

Furthermore, we suggest incorporating prototypes as initial guesses to structural optimization algorithms as a way to improve their performance. If an existing prototype is indeed the correct answer, as is the case for most compounds in nature, then optimization algorithms would converge immediately without wasting computational resources.

V. CONCLUSION

In this work, we outline a novel prototype searching method and use it to solve the structures of 520 PDF diffraction patterns. For each diffraction pattern, we obtain all prototypes in the OQMD satisfying the known stoichiometry, space group, and number of atoms per unit cell that are provided by the PDF, and select a structure based on DFT energy and R-Factor. We then validate each structure by assessing its energetic stability with respect to competing phases in the OQMD as well as the R-Factor. The 520 solved compounds, along with a table of descriptive details, can be found in the Supplemental Material, (25) and the compounds are also available in the latest release of OQMD. Identifying structures for these experimentally observed materials enables us to explore their properties from first-principles and unveil their potential for a wide variety of future applications. To allow others to take advantage of the low cost of our prototype searching method, we plan to update the "fpassmgr" software package, currently available under an open-source license. (14) (68) Currently, the "fpassmgr" package can be used to perform FPASS calculations along with validation checks, including evaluating the energetic stability of candidate solutions against OQMD competing phases. Our update will automate the process of searching for and evaluating candidate prototypes from the OQMD for many unsolved compounds in parallel.

VI. ACKNOWLEDGMENTS

Funding for this work came from the U.S. Department of Commerce, National Institute of Standards and Technology (Award No. 70NANB14H012), as part of the Center for Hierarchical Materials Design (ChiMaD). This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. In addition, this work used the

Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation (NSF) grant number ACI-1548562; specifically, it used the Bridges system, which is supported by the NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

VII. REFERENCES

[1] 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, 15010 (2013).

[2] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Davek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation, APL Mater. 1, 011002 (2013).

[3] E. Perim, D. Lee, Y. Liu, C. Toher, P. Gong, Y. Li, W. N. Simmons, O. Levy, J. J. Vlassak, J. Schroers, and S. Curtarolo, Spectral Descriptors for Bulk Metallic Glasses Based on the Thermodynamics of Competing Crystalline Phases, Nat. Commun. 7, 1-9 (2016).

[4] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The High-Throughput Highway to Computational Materials Design, Nat. Mater. 12, 191-201 (2013).

[5] R. Woods-Robinson, D. Broberg, A. Faghaninia, A. Jain, S. S. Dwaraknath, and K. A. Persson, Assessing High-Throughput Descriptors for Prediction of Transparent Conductors, Chem. Mater. 30, 8375-8389 (2018).

[6] V. I. Hegde, M. Aykol, S. Kirklin, and C. Wolverton, The Phase Stability Network of All Inorganic Materials, Sci. Adv. 6, eaay5606 (2020).

[7] E. B. Isaacs and C. Wolverton, Inverse Band Structure Design via Materials Database Screening: Application to Square Planar Thermoelectrics, Chem. Mater. 30, 1540-1546 (2018).

[8] M. Amsler, L. Ward, V. I. Hegde, M. G. Goesten, Y. Xia, and C. Wolverton, Ternary Mixed-Anion Semiconductors with Tunable Band Gaps from Machine-Learning and Crystal Structure Prediction, Phys. Rev. Mater. 3, 035404 (2019).

[9] A. Gindhart, T. Blanton, J. Blanton, and S. Gates-Rector, The Power of Electron Diffraction Phase Analysis and Pattern Simulations Using the ICDD Powder Diffraction File (PDF-4+), Microsc. Microanal. 24, 1154-1155 (2018).

[10] V. Pecharsky and P. Zavalij, Fundamentals of Powder Diffraction and Structural Characterization of Materials, Boston : Springer US (2009).

[11] H. Putz, J. C. Schön, and M. Jansen, Combined Method for ab initio Structure Solution from Powder Diffraction Data, J. Appl. Crystallogr. 32, 864-870 (1999).

[12] B. Meredig and C. Wolverton, A Hybrid Computational-Experimental Approach for Automated Crystal Structure Solution, Nat. Mater. 12, 123-127 (2012).

[13] L. Ward, K. Michel, and C. Wolverton, Three New Crystal Structures in the Na-Pb System: Solving Structures Without Additional Experimental Input, Acta Crystallogr. A 71, 542-548 (2015).

[14] L. Ward, K. Michel, and C. Wolverton, Automated Crystal Structure Solution from Powder Diffraction Data: Validation of the First-Principles Assisted Structure Solution Method, Phys. Rev. Mater. 1, 063802 (2017).

[15] G. L. W. Hart, L. J. Nelson, R. R. Vanfleet, B. J. Campbell, M. H. F. Sluiter, J. H. Neethling, E. J. Olivier, S. Allies, C. I. Lang, B. Meredig, and C. Wolverton, Revisiting the Revised Ag-Pt Phase Diagram, Acta Mater. 124, 325-332 (2017).

[16] A. R. Oganov and C. W. Glass, Crystal Structure Prediction Using ab initio Evolutionary Techniques: Principles and Applications, J. Chem. Phys. 124, 224704 (2006).

[17] Y. Wang, J. Lv, L. Zhu, and Y. Ma, Crystal Structure Prediction via Particle-Swarm Optimization, Phys. Rev. B 82, 094116 (2010).

[18] D. C. Lonie and E. Zurek, XtalOpt: An Open-Source Evolutionary Algorithm for Crystal Structure Prediction, Comput. Phys. Commun. 182, 372-387 (2011).

[19] E. H. Majzoub and V. Ozoliņš, Prototype Electrostatic Ground State Approach to Predicting Crystal Structures of Ionic Compounds: Application to Hydrogen Storage Materials, Phys. Rev. B 77, 104115 (2008).

[20] C. J. Pickard and R. J. Needs, Ab initio Random Structure Searching, J. Phys. Condens. Matter 23, 053201 (2011).

[21] Y. Zhong, C. Wolverton, A. Y. Chang, and Z. Liu, A Combined CALPHAD/First-Principles Remodeling of the Thermodynamics of Al-Sr: Unsuspected Ground State Energies by "Rounding Up the (Un)usual Suspects, Acta Mater. 52, 2739-2754 (2004).

[22] O. M. Løvvik, O. Swang, Structure and Stability of Possible New Alanates, Europhys. Lett. 67, 607-613 (2004).

[23] C. Wolverton and V. Ozoliņš, Hydrogen Storage in Calcium Alanate: First-Principles Thermodynamics and Crystal Structures, Phys. Rev. B 75, 064101 (2007).

[24] C. Wolverton, D. J. Siegel, A. R. Akbarzadeh, V. Ozoliņš, Discovery of Novel Hydrogen Storage Materials: An Atomic Scale Computational Approach, J. Phys. Condens. Matter 20, 064228 (2008).

[25] See Supplemental Material at [URL] for (1) 90th percentile convex hull distances of compounds containing each element on periodic table, (2) details of prototype solutions that match

an existing ICSD compound, and (3) details of 624 compounds we attempted to solve using prototypes, including 520 successfully solved compounds, and (4) VASP-formatted files of all solved compounds after DFT relaxation.

[26] G. Kresse and J. Furthmüller, Efficient Iterative Schemes for ab initio Total-Energy Calculations Using a Plane-Wave Basis Set, Phys. Rev. B 54, 11169-11186 (1996).

[27] 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, 15-50 (1996).

[28] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized Gradient Approximation Made Simple, Phys. Rev. Lett. 77, 3865-3868 (1996).

[29] P. E. Blöchl, Projector Augmented-Wave Method, Phys. Rev. B 50, 17953-17979 (1994).

[30] 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, 3233-3239 (2007).

[31] S. Kirklin, B. Meredig, and C. Wolverton, High-Throughput Computational Screening of New Li-Ion Battery Anode Materials, Adv. Energy Mater. 3, 252-262 (2013).

[32] 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).

[33] [Online] https://github.com/materials/mint.

[34] W. A. Dollase, Correction of Intensities for Preferred Orientation in Powder Diffractometry: Application of the March Model, J. Appl. Crystallogr. 19, 267-272 (1986).

[35] D. E. King, Dlib-ml: A Machine Learning Toolkit, J. Mach. Learn. Res. 10, 1755-1758 (2009).

[36] M. J. Mehl, D. Hicks, C. Toher, O. Levy, R. M. Hanson, G. Hart, and S. Curtarolo, The AFLOW Library of Crystallographic Prototypes: Part 1, Comput. Mater. Sci. 136, S1-S828 (2017).

[37] D. Hicks, M. J. Mehl, E. Gossett, C. Toher, O. Levy, R. M. Hanson, G. Hart, and S. Curtarolo, The AFLOW Library of Crystallographic Prototypes: Part 2, Comput. Mater. Sci. 161, S1-S1011 (2019).

[38] H. C. Kandpal, C. Felser, and R. Seshadri, Covalent Bonding and the Nature of Band Gaps in Some Half-Heusler Compounds, J. Phys. D: Appl. Phys. 39, 776-785 (2006).

[39] J. He, S. S. Naghavi, V. I. Hegde, M. Amsler, and C. Wolverton, Designing and Discovering a New Family of Semiconducting Quaternary Heusler Compounds Based on the 18-Electron Rule, Chem. Mater. 30, 4978-4985 (2018).

[40] G. Hautier, C. C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, Inorg. Chem. 50, 656-663 (2011).

[41] H. Glawe, A. Sanna, E. K. U. Gross, and M. A. L. Marques, The Optimal One Dimensional Periodic Table: A Modified Pettifor Chemical Scale from Data Mining, New J. Phys. 18, 093011 (2016).

[42] M. Amsler, V. I. Hegde, S. D. Jacobsen, and C. Wolverton, Exploring the High-Pressure Materials Genome, Phys. Rev. X 8, 041021 (2018).

[43] W. Sun, S. T. Dacek, S. P. Ong, G. Hautier, A. Jain, W. D. Richards, A. C. Gamst, K. A. Persson, and G. Ceder, The Thermodynamic Scale of Inorganic Crystalline Metastability, Sci. Adv. 2, e1600225 (2016).

[44] M. Aykol, S. S. Dwaraknath, W. Sun, and K. A. Persson, Thermodynamic Limit for Synthesis of Metastable Inorganic Materials, Sci. Adv. 4, eaaq0148 (2018).

[45] W. Sun, A. Holder, B. Orvañanos, E. Arca, A. Zakutayev, S. Lany, and G. Ceder, Thermodynamic Routes to Novel Metastable Nitrogen-Rich Nitrides, Chem. Mater. 29, 6936-6946 (2017).

[46] J. Odahara, W. Sun, A. Miura, N. C. Roser-Navarro, M. Nagao, I. Tanaka, G. Ceder, and K. Tadanaga, Self-Combustion Synthesis of Novel Metastable Ternary Molybdenum Nitrides, ACS Mater. Lett. 1, 64-70 (2019).

[47] W. Sun *et al.*, A Map of the Inorganic Ternary Metal Nitrides, Nat. Mater. 18, 732-739 (2019).

[48] C. G. Richter and W. Jeitschko, Preparation and Crystal Structure of the Titanium and Hafnium Bismuthides Ti_8Bi_9 and Hf_8Bi_9, J. Solid State Chem. 134, 26-30 (1997).

[49] Standard X-ray Diffraction Powder Patterns: Section 12, Data for 57 Substances, Natl. Bur. Stand. monogr. 25 (1975).

[50] C. W. F. T. Pistorius, A_2BF_5 Phases in the Systems $AF\text{-}BF_3$ (A=K, Rb, Cs, Tl; B=Al, Fe, Cr, Ga, V, Tl, Ln), MRS Bull. 10, 1079-1084 (1975).

[51] Y. Sakurai, H. Arai, S. Okada, and J. Yamaki, Low Temperature Synthesis and Electrochemical Characteristics of $LiFeO_2$ Cathodes, J. Power Sources 68, 711-715 (1997).

[52] A. D. Weeks, E. A. Cisney, and A. M. Sherwood, Montroseite, a New Vanadium Oxide from the Colorado Plateaus, Washington D. C. : U.S. Geological Survey, 1235-1241 (1953).

[53] H. T. Howard Jr. and S. Block, The Crystal Structure of Montroseite, a Vanadium Member of the Diaspore Group, Washington D. C. : U.S. Geological Survey, 1242-1250 (1953).

[54] C. Milton, D. E. Appleman, M. H. Appleman, E. C. Chao, F. Cuttitta, J. I. Dinnin, E. J. Dwornik, B. L. Ingram, and H. J. Rose, Merumite: A Complex Assemblage of Chromium Minerals from Guyana, Washington: United States Government Printing Office (1976).

[55] I. E. Nemirovskaya, A. N. Grechenko, A. M. Alekseev, and V. V. Lunin, Phase Transformations in Hydrogen Sorption-Desorption by Hydrides of Intermetallic Compounds of the CrB Structural Type, J. Struct. Chem. 32, 680-686 (1991).

[56] R. M. van Essen and K. H. J. Buschow, Hydrogen-Absorption in Various Zirconium- and Hafnium-Based Intermetallic Compounds, J. Less-Common Met. 64, 277-284 (1979).

[57] K. D. C´iric, V. J. Koteski, D. L. J. Stojic, J. S. Radakovic, and V. N. Ivanovski, HfNi and Its Hydrides – First Principles Calculations, Int. J. Hydrog. Energy 35, 3572-3577 (2010).

[58] S. W. Peterson, W. I. Korst, and V. N. Sadana, Neutron Diffraction Study of Nickel Zirconium Hydride, J. Phys. France 25, 451-453 (1964).

[59] X. Chen and S. Jin, Institute of Physics, Chinese Academy of Sciences, Beijing : ICDD Grant-in-Aid (2013).

[60] F. A. Faber, A. Lindmaa, O. Anatole von Lilienfeld, and R. Armiento, Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals, Phys. Rev. Lett. 117, 135502 (2016).

[61] G. Trimarchi, X. Zhang, V. M. J. DeVries Vermeer, J. Cantwell, K. R. Poeppelmeier, and A. Zunger, Emergence of a Few Distinct Structures from a Single Formal Structure Type During High-Throughput Screening for Stable Compounds: The Case of RbCuS and RbCuSe, Phys. Rev. B 92, 165103 (2015).

[62] E. Parthé, Elements of Inorganic Structural Chemistry, Leipzig : Pöge Druck (1990).

[63] A. R. Oganov and M. Valle, How to Quantify Energy Landscapes of Solids, J. Chem. Phys. 130, 104504 (2009).

[64] H. Burzlaff and Y. Malinovsky, A Procedure for the Classification of Non-Organic Crystal Structures I: Theoretical Background, Acta Cryst. A 53, 217-224 (1997).

[65] J. A. Chisholm and S. Motherwell, COMPACK: A Program for Identifying Crystal Structure Similarity Using Distances, J. Appl. Crystallogr. 38, 228-231 (2005).

[66] L. Zhu *et al*., A Fingerprint Based Metric for Measuring Similarities of Crystalline Structures, J. Chem. Phys. 144, 034203 (2016).

[67] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, Predicting Crystal Structure by Merging Data Mining with Quantum Mechanics, Nat. Mater. 5, 641-646 (2006).

[68] [Online] https://bitbucket.org/wolverton/fpass-manager.