npj | Computational Materials

www.nature.com/npjcompumats

ARTICLE OPEN

A high-throughput data analysis and materials discovery tool for strongly correlated materials

Hasnain Hafiz[1,7], Adnan Ibne Khair[2], Hongchul Choi[3], Abdullah Mueen[2], Arun Bansil[1], Stephan Eidenbenz[4], John Wills[3], Jian-Xin Zhu[3], Alexander V. Balatsky[5,6] and Towfiq Ahmed[3]

Modeling of *f*-electron systems is challenging due to the complex interplay of the effects of spin–orbit coupling, electron–electron interactions, and the hybridization of the localized *f*-electrons with itinerant conduction electrons. This complexity drives not only the richness of electronic properties but also makes these materials suitable for diverse technological applications. In this context, we propose and implement a data-driven approach to aid the materials discovery process. By deploying state-of-the-art algorithms and query tools, we train our learning models using a large, simulated dataset based on existing actinide and lanthanide compounds. The machine-learned models so obtained can then be used to search for new classes of stable materials with desired electronic and physical properties. We discuss the basic structure of our *f*-electron database, and our approach towards cleaning and correcting the structure data files. Illustrative examples of the applications of our database include successful prediction of stable superstructures of double perovskites and identification of a number of physically-relevant trends in strongly correlated features of *f*-electron based materials.

npj Computational Materials (2018)4:63 ; doi:10.1038/s41524-018-0120-9

INTRODUCTION

There is an ever-increasing need for developing new materials with novel electronic functionalities in our technology-driven modern society. To this end, a recent focus has been to integrate computational and experimental techniques for the purpose of obtaining robust, predictive tools for gaining insights into structure–property relationships in functional materials in order to reduce the time and cost in the materials discovery process. In this connection, it is important to expand the search space of materials to include compounds containing elements from the last rows of the periodic table, which present an especially rich playground for hosting novel functionalities that arise through the interplay of effects of spin–orbit coupling, strong electron correlations and interactions of highly localized *f*-electrons with free-electron-like conduction electrons. Although widely known for nuclear energy applications, actinides and lanthanides have recently attracted attention also in connection with superconductivity,[1] magnetism,[2] and Kondo physics[3] as well as energy-related applications.[4,5]

While a "1-to-1" approach (one system—one calculation—one experiment) has been the traditional strategy in lanthanide and actinide research, this approach is quite inefficient due to the large experimental cost of handling high-Z materials. Moreover, *f*-electron systems have proven notoriously hard to model within the first-principles density functional theory (DFT)[6,7] framework, while dynamical mean field theory (DMFT) methods[8–10] invoke semi-empirical correlation parameter *U*, which reduces their predictive power. Here, in order to address these challenges, we discuss a novel predictive strategy, which is based on adopting an integrated data and theory-driven approach. In this connection, we have built the *f*-electron structure database (fESD, http://correlatedmaterials-lanl.org) that combines experimental and DFT simulated data using state-of-the-art algorithms[11–13] to train our machine learning models. We will show how the analysis of our database reveals new insights into the intricate electronic and structural properties of *f*-electron systems.

A unique feature of the fESD is that it incorporates high-quality simulated electronic structure data which capture the effects of electron-correlations and spin–orbit coupling in an all-electron environment. Many popular databases (e.g., AFLOW,[14] Materials Project,[15] Organic Materials Database[16]) use pseudo-potentials and plane-wave-based techniques for DFT simulations. In contrast, we use full potential, all-electron, relativistic linearized-augmented-plane-wave (LAPW) results,[17,18] which is important for obtaining a predictive model for strongly correlated materials, computational cost notwithstanding.

RESULTS AND DISCUSSIONS

Database design and collection of data

Our database is primarily an electronic-structure database with query tools that allow searches of compounds with desired crystal and/or band structures. In addition, the database contains DFT simulated data with a variety of query capabilities. The database management system (DMS) is built on MySql and Java Development platforms. Additionally, a number of Python-based scripts are used for data parsing at various levels (e.g., crystallographic information files (CIF), DFT output). Work toward designing an

[1]Department of Physics, Northeastern University, Boston, MA 02119, USA; [2]Department of Computer Science, University of New Mexico, Albuquerque, NM 87131, USA; [3]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; [4]The Information Science and Technology Institute, Los Alamos National Laboratory, Los Alamos, NM 87545, USA; [5]Nordita, KTH Royal Institute of Technology and Stockholm University, Roslagstullsbacken 23, Stockholm SE-106 91, Sweden and [6]Department of Physics, University of Connecticut, Storrs, CT 06269, USA
Correspondence: Hasnain Hafiz (hafiz.h@husky.neu.edu) or Arun Bansil (ar.bansil@northeastern.edu) or Towfiq Ahmed (atowfiq@lanl.gov)
[7]Present address: Department of Mechanical Engineering Carnegie, Mellon University, Pittsburgh, PA 15213, USA

Received: 3 December 2017 Accepted: 29 October 2018
Published online: 22 November 2018

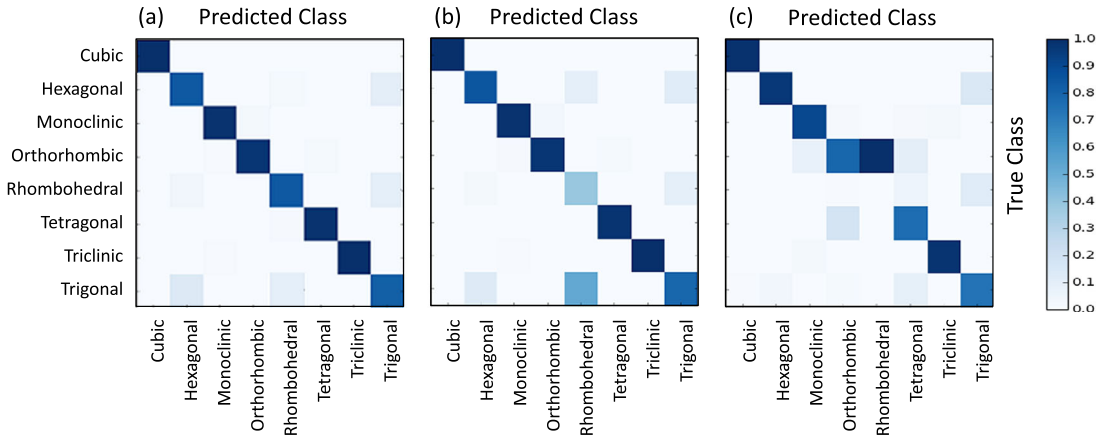

Fig. 1 Confusion matrices for different machine-learning algorithms. Confusion matrices for three methods: **a** Multilayer Perceptron (MLP), **b** K-Nearest-Neighbor (KNN), and **c** Logistic Regression (LR). MLP, which is a neural network approach, captures the correct crystal system with maximum accuracy

application-programming interface (API), which will integrate various query tools and provide a Python-based library of functions for programmed access to the database, is currently in progress.

Inorganic Crystal Structure Database (ICSD)[19] and Crystallographic Open Database (COD)[20] are our initial sources, which currently contain 188,000 and 364,331 CIF files, respectively. There are 54,465 structure files in ICSD and 27,502 in COD with compounds containing lanthanide and actinide elements (e.g., outer shell *f*-orbital electrons). We started by downloading structure information such as crystal structures, lattice parameters, and atomic positions from these CIF files into our database. At present, the scope of our database is limited to ground state DFT-based electronic structures using available crystal structure data without sensitivity to pressure and temperature dependencies or the specifics of how the data was acquired (e.g., theoretically or experimentally). A richer labeling with more detailed attributes of the data, will be undertaken in future extensions of the database.

Data cleaning, verification, and crystal structure determination

While restructuring the data format, we developed a number of query tools with new search features which, to the best of our knowledge, are not available in other existing databases.[14–16] In particular, we provide extended search capabilities such as space-group or lattice-system-based search, and search for a super-structure of desired symmetry. As a part of this development process, we performed the "data cleaning" step by employing supervised learning tools. This resolved ambiguities in CIF files resulting from missing information, and corrected the structure files needed for generating reliable electronic structure data, which are the core content of our *f*-electron database. The details of the generation of electronic and crystal structure data are described in the Methods section.

In our SQL database, some crystallographic features (e.g., lattice parameters, space groups) are missing or ambiguous in many compounds because of incomplete original CIF files that were parsed to create the database. These compounds cause convergence problems or inaccurate results in DFT simulations. We resolve this problem by adopting a machine learning approach for verifying and cleaning the incorrect information. Currently, our database hosts crystal data for approximately 82,000 *f*-electron compounds, out of which 8711 compounds contain missing or incorrect lattice systems. In this connection, we discuss three different supervised machine-learning algorithms to predict one of the seven possible "lattice systems" as follows. (1) Logistic Regression (LR),[12] where one predicts categorical target variables using a logistic function involving linear combinations of feature values. (2) K-Nearest-Neighbor (KNN) algorithm,[11] which predicts category of an unknown instance based on K "similar" instances. We use Euclidean distance as a measure of similarity and pick the majority of the three (K = 3) most similar materials to predict the lattice parameters of an incomplete system. (3) Multilayer Perceptron (MLP)[13] in which a forward feeding network of perceptrons is used, as opposed to a single perceptron that is equivalent to a LR model. We use three layers of perceptrons with eight perceptrons in each layer.

Table 1. Benchmarking the performance of three machine-learning techniques (LR, KNN, and MLP) discussed in the text

Technique	Accuracy	Precision	Recall
LR	90.4%	76.5%	65.6%
KNN	98.1%	87.6%	86.2%
MLP	99.1%	90.8%	91.3%

The models were trained on a set of 275,926 labeled instances from the COD with accurate lattice constants (*a, b, c, α, β, γ*), volume, and the space group as training features. We ran a 10-fold cross-validation, and calculated the prediction accuracy, precision, and recall. Using the miss and hit between the predicted-class and true-class outcome, we constructed a representation widely known as the confusion (or error) matrix in the machine-learning community. As shown in Fig. 1, the confusion matrices for the three methods evaluated indicate clearly the superiority of the MLP algorithm, which essentially is a neural network approach. MLP yields an accuracy of 99.1% in predicting the crystal system, see Table 1. Although computationally more complex and expensive, we employed this algorithm for cleaning and obtaining high-fidelity structure information for our database.

However, all three models exhibit some off-diagonal non-zero values in the confusion matrix indicating errors in prediction. Interestingly, we found such instances to be restricted to three unique classes, namely, trigonal, hexagonal, and rhombohedral, see Fig. 1a. In the crystallographic convention,[21] there are two equivalent systems, i.e., the crystal system and the lattice system. Trigonal and hexagonal lattices belong to the crystal convention, while the rhombohedral and hexagonal are defined in the lattice system. We resolved such ambiguities of definition in the original CIF files by deploying our neural-network-based MLP tool (Fig. 1a), and thus obtained cleaner datasets for further *ab initio* simulations. In this way, we correctly predicted the missing crystal systems for 8711 compounds with the distribution given in Table

2. Note that our main purpose in determining crystal system information was to illustrate the viability of our machine learning tools. However, the correct coding of the crystal system and the space group for materials in the database can be of interest in searching the database for materials belonging to a given space group or a crystal system. For example, certain space groups and crystal systems have proven useful in successful searches of viable topological materials.[22] Moreover, useful physical insights can often be obtained by comparing and contrasting the evolution of properties of materials within a space group or a given crystal system as well as by comparisons across different or related space groups or crystal systems.

Perovskite heterostructure prediction

In the recent years, there has been considerable theoretical and experimental interest in creating layered compounds and superstructures by design due to their enormous potential for achieving various emergent functionalities (e.g., superconductivity,[23,24] multiferroics[25,26]). A common strategy employed in synthesis efforts is to attempt the creation of a new superstructure with desired electronic properties by overlaying two or more compounds with different chemical compositions but similar crystal symmetries. In this connection, we investigated a set of perovskite compounds in our database with the goal of identifying compounds that would be most likely to form stable superstructures. We used lattice information and chemical intuition related to "ionic radii and Goldschmidt tolerance factors"[27] as our primary screening tools for the initial narrowing-down of the search space for perovskite heterostructures. We then performed a data-mining query, and identified sets of paired superstructure combinations with the highest likelihood of forming superstructures. The next stage would then involve a more careful computational analysis of the small number of promising candidate systems so identified before it will be appropriate to encourage an experimental synthesis and validation cycle. Generally, we should expect that the prediction–synthesis–validation loop may well require more than one iteration to successfully identify a viable new material.

Table 2. Distribution of the crystal systems, which were predicted correctly by using MLP tools

Label	Count
Cubic	1355
Hexagonal	4577
Monoclinic	1428
Orthorhombic	23
Tetragonal	2
Trigonal	1326
Total	8711

Figure 2 shows the results of a compatibility test for the ordering of superstructures of a number of double perovskites with chemical formula $AA'BB'C_3C'_3$. Here, A/A′-site cations are either rare earths or actinides, B/B′ cations are transition metal elements, and C/C′ anions are mostly oxygen or halogens. Using the ionic radii[28,29] of the two existing single perovskites ABC_3 and $A'B'C'_3$ (see the inset in Fig. 2b), we assess the geometric stability of a possible double perovskite $AA'BB'C_3C'_3$ by calculating the Goldschmidt tolerance factor,[27,30] $t = \frac{r_A + r_C}{\sqrt{2}(r_B + r_C)}$, where r_A, r_B, and r_C are the average ionic radii of (A, A′), (B, B′), and (C, C′), respectively; here, we have neglected other contributing factors such as the effects of octahedral tilting. We considered both ordered and disordered compounds on B and B′ sites, and following the method described by Vasala et al.,[30] we treat all compounds with the same chemical formulae regardless of the ordering at the B-sites. Note that Goldschmidt tolerance factor is a good measure for testing the stability of $AA'BB'C_3C'_3$ type perovskite compounds, where B or B′-site cations can be ordered or disordered.[30] Considering each data point as a paired combination of a double-perovskite superstructure, we plot the associated tolerance factor t and the lattice-parameter-ratios $\frac{a1}{a2}$ and $\frac{c1}{c2}$ in Fig. 2a, where ($a1$, $c1$) and ($a2$, $c2$) are lattice parameters of ABC_3 and $A'B'C'_3$ perovskites, respectively. Clustering of the data in Fig. 2a reflects the correlation between the type of anion/cation involved and the space group of the individual perovskites with the tolerance and lattice parameters. High tolerance factors can be achieved when $\frac{a1}{a2}$ and $\frac{c1}{c2}$ are close to unity, i.e., when the difference in ionic radii of A ($r_A - r_{A'}$) or B ($r_B - r_{B'}$) site cations is small. We also see the important role of space groups here such as two cubic perovskites usually result in a stable double perovskite in contrast to the cases

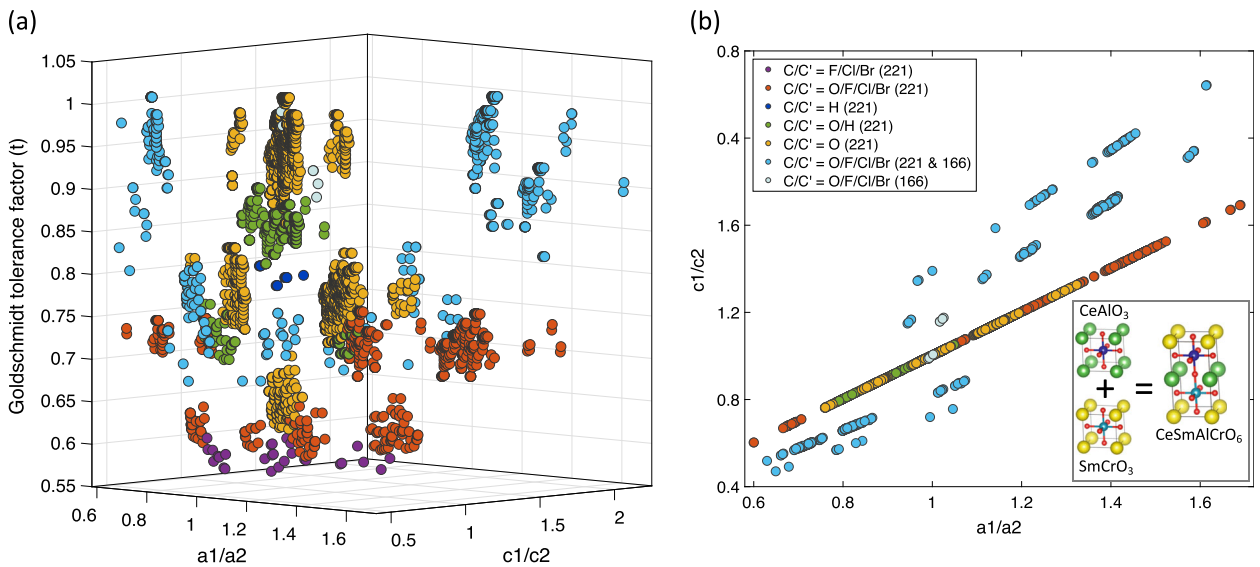

Fig. 2 Superstructure compatibility test for double-perovskites $AA'BB'C_3C'_3$. Each data point is a paired combination of two single perovskite systems as shown in the inset to (**b**). Different colors imply different combinations of C and C′ as shown in the legend in (**b**). Space group for each C/C′ type is given in the parenthesis in the legend in (**b**)

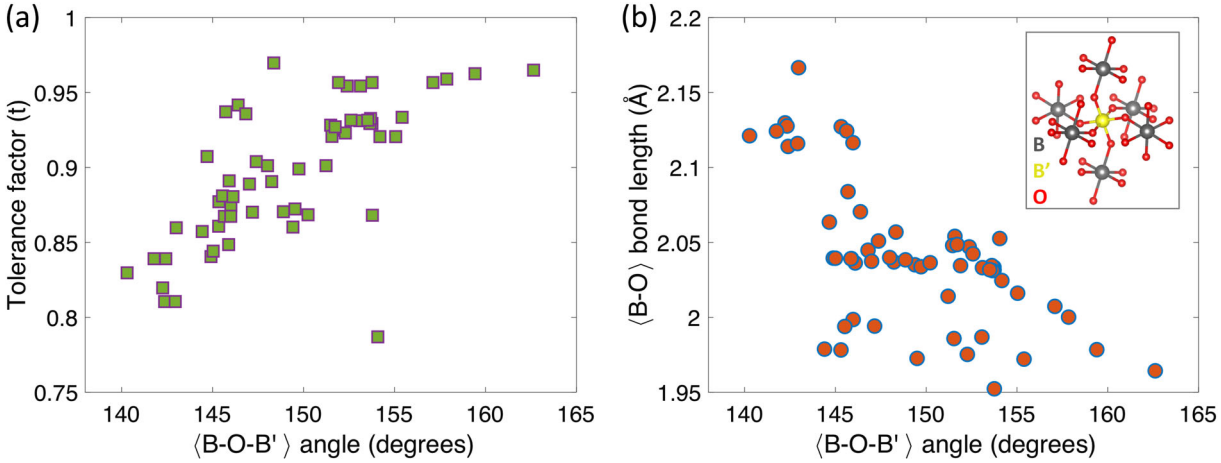

Fig. 3 Descriptors for octahedral distortion in perovskite superstructures. **a** Tolerance factor (t) is plotted against the average bond angle ⟨B-O-B'⟩. t is seen to correlate with decreased octahedral distortion as the bond angle ⟨ B-O-B'⟩ gets closer to linear. **b** A plot of the average ⟨B–O⟩ bond length against the bond angle ⟨B-O-B'⟩. The bond length is seen to correlate with decreasing distortion (increasing bond angle). For various B' cations, average bond angles and bond lengths have been obtained by considering all six neighboring B cations. The inset in (**b**) shows neighboring BO_6 and $B'O_6$ octahedra where B, B', and O ions are marked with grey, yellow, and red spheres, respectively

where the space groups are mixed. This is to be expected because similar crystal structures of the two contributing perovskites lead to lesser cation size mismatch or smaller differences in ionic radii of the cations of individual sites and result in higher tolerance factors. This is why we would expect the most abundant double-perovskite systems to be formed from two cubic systems, as demonstrated in Fig. 2b. We see a linear trend in Fig. 2b where the data points are plotted in the $\frac{a1}{a2}$ and $\frac{c1}{c2}$ plane. The central line indicates that the most abundant double-perovskite systems would be formed from ABC_3 and $A'B'C_3'$ cubic systems. Other trend lines in Fig. 2b show the cases in which mixed systems can be formed from cubic and other (non-cubic) crystal structures. For this reason, we allowed a mixture of anions to identify various possibilities in the stability trend lines. We generally see the trend of high stability (t) when both C and C′ anions are oxygen or hydrogen compared to the case when both anions are halogens (see Fig. 2a). This indicates that the most stable structures are formed when both anions are identical, which should be contrasted with the situation where anions are allowed to mix. In this way, we were able to cluster possible candidate perovskites with respect to anion mixing and distinguish them in terms of their stability.

We turn now to address the effects of octahedral distortion of the ideal cubic structure, which is very common in the perovskites.[30,31] Octahedral distortion induces changes in the B-O-B' bond angle away from 180°, and affects the B-O and B-O' bond lengths to produce elongation or contraction along the c-axis. Figure 3a shows that there is a linear trend between the tolerance factor and the average value of the B-O-B' bond angle. Systems with higher ⟨B-O-B'⟩ bond angles possess smaller octahedral distortions, yielding higher tolerance factors that favor stability. An inspection of our data indicates that ⟨B-O-B'⟩ bond angle gets closer to being linear when the B/B' cation has a small number of d electrons. However, B and B' sites with Jahn–Teller active cations such as Mn^{3+} and Cu^{2+} lead to higher distortions with the bond angle ⟨B-O-B'⟩ ≈ 142°. Majority of perovskite structures we considered showed an elongation of the c-axis, which results in shorter average B-O and B'-O bond lengths, see Fig. 3b. We thus adduce that, for a given type of A-cation, the average bond length decreases linearly as the amplitude of the distortion decreases. This linear relationship, however, may not hold when two different A-site cations are involved, which accounts for the scatter in the data of Fig. 3b.

It is also important to consider the effects of ordering of the B/B'-site cations. When the B and B' cations occupy the correct sites, we take the arrangement to be ordered. Following Vasala et al.,[30] we consider two types of cation disorder, namely, anti-site (AS) defect and anti-phase boundary (APB). Here an AS defect is defined as the interchange of B and B' cations, while APB involves the boundary of two ordered domains where occupancies of B and B' sites are reversed. Specifically, we identified 105 ordered rock-salt and 50 disordered double perovskite structures of the form $A_2BB'O_6$ in the ICSD database. Figure 4a shows the difference in B and B' cation oxidation state ($\Delta Z_B = |Z_{B'} - Z_B|$) as a function of the corresponding ionic radius difference ($\Delta r_B = |r_{B'} - r_B|$). We observe that smaller ΔZ_B values favor the disordered arrangement. This makes physical sense because B and B' cations in this case are chemically similar, and therefore, these cations will tend to occupy various sites interchangeably. We can also understand this result in terms of the competing effects of electrostatic repulsion and entropy. When B and B' cations are similar, the system will have a tendency to become disordered to increase its entropy. On the other hand, in ordered systems which tend to have higher ΔZ_B values (top red dots in Fig. 4a), highly charged B' cations always have less charged B cations as nearest neighbors, so that the Coulomb energy in the ordered state is lowered compared to the disordered state.[30,32] The difference in ionic radii, Δr_B, also provides a good descriptor for obtaining a handle on the viability of ordered vs disordered arrangements. It is known[32,33] that a larger Δr_B enhances lattice strain, which favors ordered phases.

Keeping the aforementioned trends involving ΔZ_B and Δr_B in mind, we infer that perovskites such as La_2CoFeO_6 ($\Delta Z_B = 0$ and $\Delta r_B = 0.035$ Å) are likely to be stable even in disordered arrangements.[34] In contrast, a compound like La_2NaIrO_6 ($\Delta Z_B = 4$ and $\Delta r_B = 0.45$ Å) will be expected to be stable only in ordered configurations[35] (see Table 3). In this way, ΔZ_B and Δr_B can be good descriptors for identifying ordered vs disordered superstructures. Figure 4a shows that systems with $\Delta Z_B = 0$ and $\Delta r_B \leq 0.1$ Å are always disordered, while systems with $\Delta Z_B = 4$ and $\Delta r_B \geq 0.1$ Å are always ordered. Coexistence of ordered and disordered structures for $\Delta Z_B = 2$ presumably reflects the competition between Coulombic and entropic effects. In order to gain a deeper understanding of the underlying energetics, we have computed ground state energies of various structures as shown in Fig. 4b. While the disordered configurations can achieve high tolerance factors ($t > 0.85$), they are seen to be energetically less favorable compared to the corresponding ordered systems in

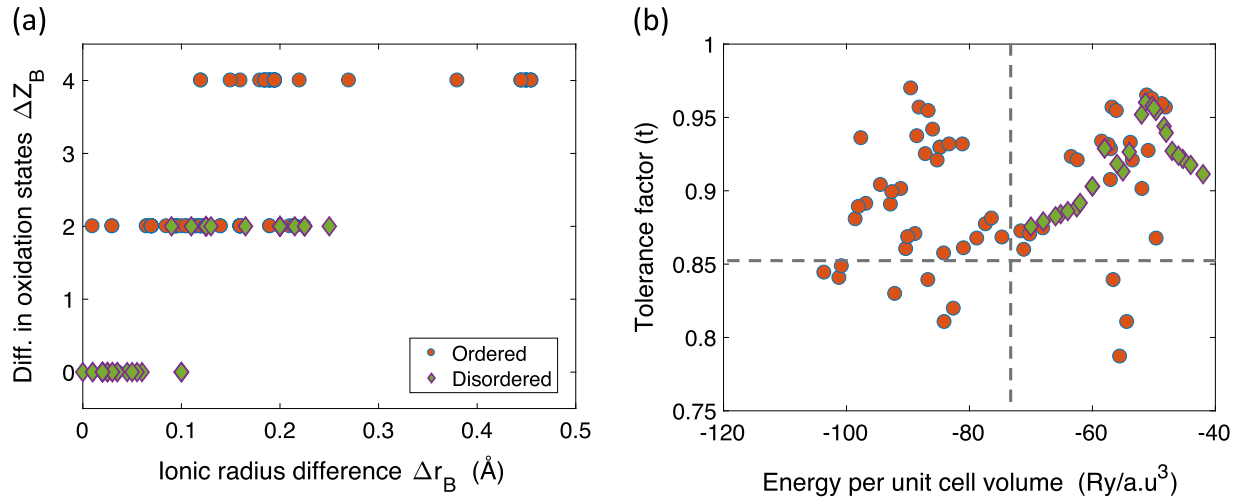

Fig. 4 Descriptors of cation ordering in double perovskites. **a** Difference in B and B' cation oxidation states, $\Delta Z_B = |Z_{B'} - Z_B|$, is plotted against the corresponding difference in ionic radii, $\Delta r_B = |r_{B'} - r_B|$. A combination of smaller values of Δr_B and ΔZ_B is seen to favor disordered arrangements, while a higher value of these descriptors favors ordered arrangements. **b** Tolerance factor (t) vs ground state energy per unit cell volume. Ground state energy is seen to be higher for disordered cases even though the associated tolerance factors can be quite large

Table 3. Predicted double perovskites obtained via our superstructure compatibility test

Perovskite	Tolerance (t)	$\Delta r_B = \|r_{B'} - r_B\|$ (Å)	$\Delta Z_B = \|Z_{B'} - Z_B\|$	Ref.
La_2NaIrO_6	0.89	0.45	4	35
La_2LiMoO_6	0.94	0.15	4	40
Nd_2NaRuO_6	0.87	0.45	4	41
La_2CoMnO_6	0.96	0.215	2	36,37
La_2CrCoO_6	0.97	0.005	0	38
La_2CoFeO_6	0.98	0.035	0	34,39
La_2FeMnO_6	0.95	0	0	42
Nd_2MnRhO_6	0.92	0.02	0	43
Ce_2AlCrO_6	0.98	0.08	0	This work
Ce_2GaCrO_6	0.96	0.005	0	
$CeSmAlCrO_6$	0.96	0.08	0	
$LaCeAlMnO_6$	0.98	0.11	0	

The four compounds in the top four rows of the table show large values of both Δr_B and ΔZ_B, and these are identified as ordered perovskites. Single crystals of these four compounds have been synthesized using the conventional solid-state methods.[35,36,40,41] The four compounds in the middle portion of the table have $\Delta Z_B = 0$ with very small values of Δr_B. We predict these four compounds to harbor disordered B cation arrangements. These four compounds have indeed been synthesized as solid solutions;[34,38,39,42,43] notably, as expected, these compounds achieve very high tolerance factors. The bottom four compounds in the table are predicted as possible superstructures, which would be interesting to explore experimentally

most cases, consistent with our analysis above based on the descriptors ΔZ_B and Δr_B. Notably, however, some ordered systems exhibit higher energy (>−75 Ry/a.u.^{3}) and stability ($t > 0.85$), which is possible when Coulombic and entropic effects become comparable.

Our analysis discussed above based on using ionic radii, bond angles, oxidation states, lattice parameter ratios, ground state energies, and Goldschmidt tolerance factors as descriptors correctly predicts several existing double perovskites, see Table 3 (top eight rows). These compounds have high tolerance factors and have indeed been synthesized at ambient pressure.[30,34–43] We also predict four new double perovskites with high tolerance factors (bottom four rows in Table 3), which will be interesting to explore experimentally.

Electronic structure calculations and orbital-resolved electronic structure search

Interpretation of specific features in band structures can be complicated in the presence of various competing effects, particularly in the strongly-correlated *f*-electron systems of interest to this study. In this connection, contributions of various atomic sites and different orbital angular momentum channels to the Bloch wave functions associated with the band structure near the Fermi energy can provide key information for understanding the electronic behavior of the material. Keeping this in mind, we have implemented a unique search capability based on identifying different atomic orbitals contributing to various bands at *k*-points along the high-symmetry directions. As an example, Fig. 5 shows an advanced band structure query on Uranium Nitride (UN) that has dominant *f*-orbital contributions near the Fermi energy. Such information is helpful for gaining insight into the complex electronic, magnetic, and optical properties of *f*-electron materials, and for developing machine-learning models and predictive tools. Note that local dynamical correlation effects are missing in the simulated DFT results in our database. We address this aspect in the following section.

A search for strongly correlated actinides

We have computed frequency-dependent hybridization functions for all the compounds in our database to supplement our ground state DFT calculations. Our earlier work[44] has shown that hybridization is a good descriptor for detecting localization trends, although an accurate description of localization phenomenon will require self-consistent DMFT calculations.[9,10] However, our findings[44] indicate that a high-throughput, local Green's function based hybridization can at least qualitatively capture physically-relevant trends in the localization of the *f*-states. Our analysis with the present *f*-electron database using ≈350 Ce-based binary compounds shows that the maximum hybridization value near the Fermi energy decreases with increasing lattice spacing.

We comment briefly on technical aspects of our frequency-dependent hybridization function computations. In the Anderson impurity model,[45] when an impurity electron is immersed in a sea of itinerant electrons, it hybridizes with the Bloch states of the

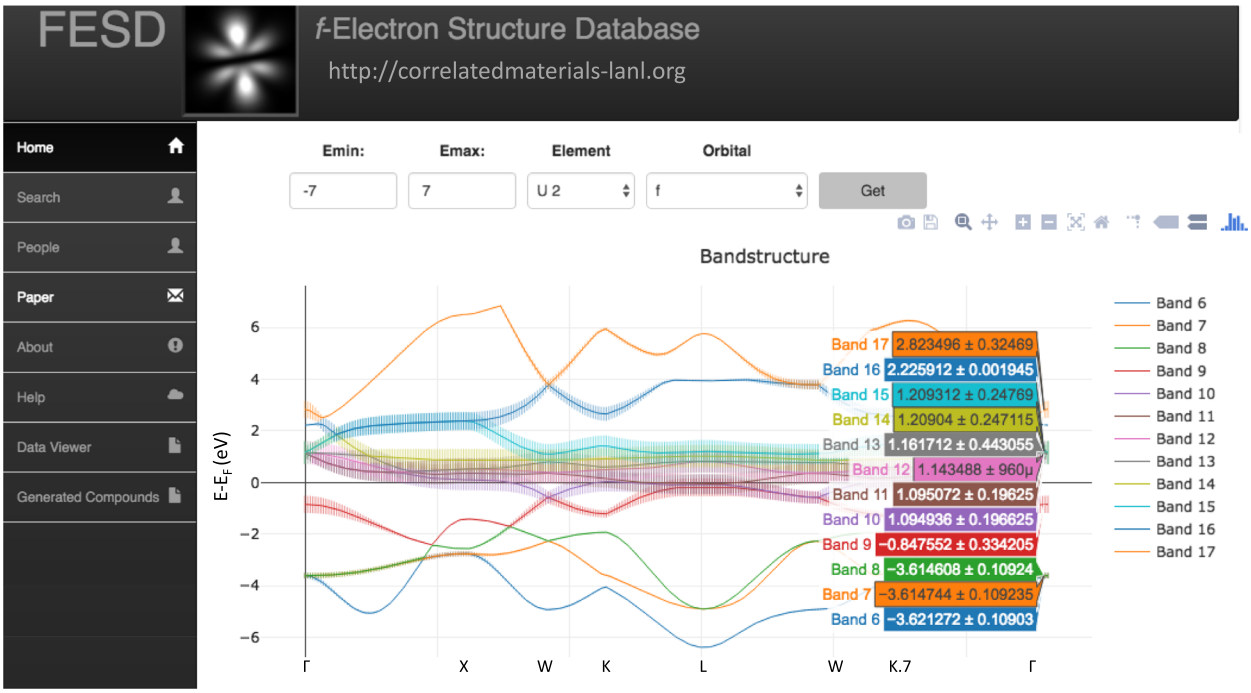

Fig. 5 Web interface for band structures. As an example, we show the band structure of UN displaying *f*-orbital characters near the Fermi level. The different drop-down boxes at the top show that an energy window of −7 to 7 eV is chosen for uranium and that the *f* orbital characters are being considered

surrounding electron sea.[46] The interaction of the impurity cluster, which in our case are the *f*-electrons, with its surrounding bath can be monitored via the energy-dependent hybridization function Δ(*E*). Using Dyson equation, we can formulate an expression for Δ(*E*) as[9,44,47]

$$G_0^{-1} = \left(E - E^{OI}\right) - \Delta(E) = (E - H) - \Delta(E) \quad (1)$$

where G_0 is the site projected Green's function obtained from DFT and *H* is the hybridization-free impurity Hamiltonian corresponding to the quantum impurity energy E^{OI}. A bigger overlap of *f*-states with the surrounding itinerant electrons will produce a larger hybridization function, which can thus serve as a descriptor for the localization/delocalization of *f*-electrons.

Figure 6 presents the calculated hybridization functions Δ(*E*) for 4*f* and 5*f* monopnictides LnX and AnX (Ln = Ce, An = Th, U, Pu, and X = N, P, As, Sb, Bi) with space group $Fm\bar{3}m$ and a structure similar to NaCl. We reproduce previously reported results on Δ(*E*) for CeX[44] and the established trends in similar compounds,[48,49] indicating the high fidelity of our calculations. Our analysis further identifies a number of similarities and trends in Δ(*E*) features in various 4*f* and 5*f* monopnictides as follows. (1) A distinct peak in Δ(*E*) is found between 2 and 3.5 eV below the Fermi level in both LnN and AnN. The size of this peak decreases monotonically (with small variations) as we scan from the top to the bottom of group 15 of the periodic table. The reduced width of Δ(*E*) peak in Sb and Bi compounds indicates a more localized nature of *f*-electrons, whereas a sharp corresponding peak in N compounds with two times larger area under the curve shows a greater degree of delocalization. (2) Figure 7 (top panel) shows that there is an overall increase in the value of the hybridization function at the Fermi energy in going from 4*f* to 5*f* monopnictides, reflecting the more extended nature of the 5*f* shells compared to their 4*f* counterparts. (3) Within the 4*f* and 5*f* shells, if we move from the left to the right of the periodic table, we see an increase in localization with an increase in *f*-electron count. (4) The bottom panel in Fig. 7 shows an anti-correlation between the degree of localization and lattice constant for both the LnX and AnX series. And, finally, (5) moving from the top to the bottom of group 15, we observe that a higher unit-cell volume produces a smaller $\Delta(E_F)$ value. This can be understood from the fact that a larger ligand distance will reduce the interaction of the *f*-electron atom with the neighboring atoms, and thus yield a more localized band structure. Along this line, the 4*f* monopnictides exhibit higher volumes overall and increased localization compared to the corresponding 5*f* compounds.

Although we have focused on identifying trends in terms of the hybridization functions, such an analysis could also be carried out based on an examination of features in DOS and/or band structures in conjunction with lattice symmetry information. We emphasize that any predictive search must first narrow down the search space by using domain-specific knowledge by using hybridization and/or other appropriate descriptors as a prelude to deploying more sophisticated techniques (e.g., charge self-consistent DMFT and/or structure relaxation). In this way, practical schemes can be obtained for robust theory-guided discovery of new correlated functional materials.

In conclusion, we have presented the design of our newly developed *f*-electron structure database and discussed the associated high-throughput tools for analyzing structural and electronic properties of materials, including query tools for identifying orbital characteristics of electronic states near the Fermi energy. fESD currently contains data on about 80,000 *f*-electron compounds. All structure data in fESD have been cleaned and corrected using machine-learning tools. We have included DFT-based high-quality electronic structure data using all-electron self-consistent computations on the large number of compounds in fESD as its core content. Potential for materials discovery via fESD is illustrated by considering a stability search of superstructures composed of two perovskite compounds. Using ionic radii, bond angles, oxidation states, ground state total energy, lattice parameter ratios, and Goldschmidt tolerance factors as descriptors, we show how stable double-perovskite superstructures can be predicted, and how insight into the roles of anion types and space groups involved for achieving higher structural

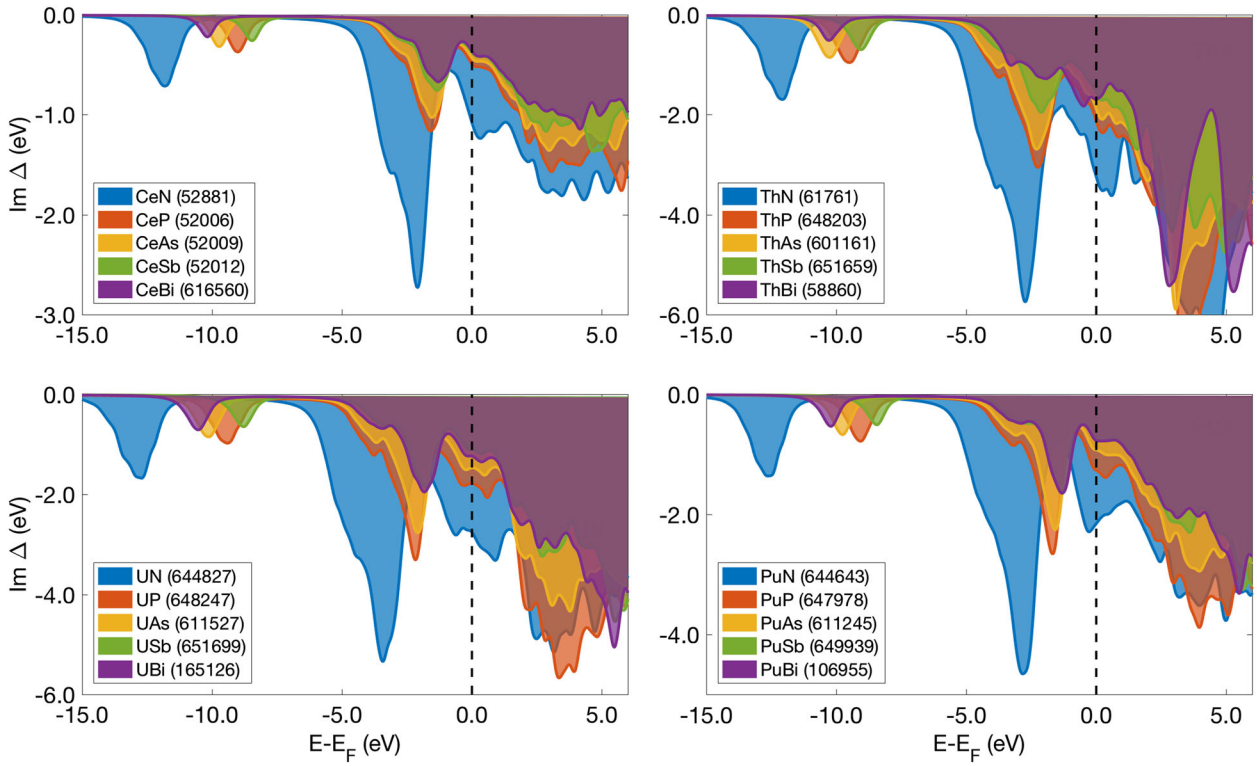

Fig. 6 Trends in *f*-orbital hybridization function. The calculated hybridization function Δ(*E*) for the series of 4*f* and 5*f* monopnictides, LnX and AnX (Ln = Ce, An = Th, U, Pu, and X = N, P, As, Sb, Bi). The ICSD identification numbers are shown in brackets

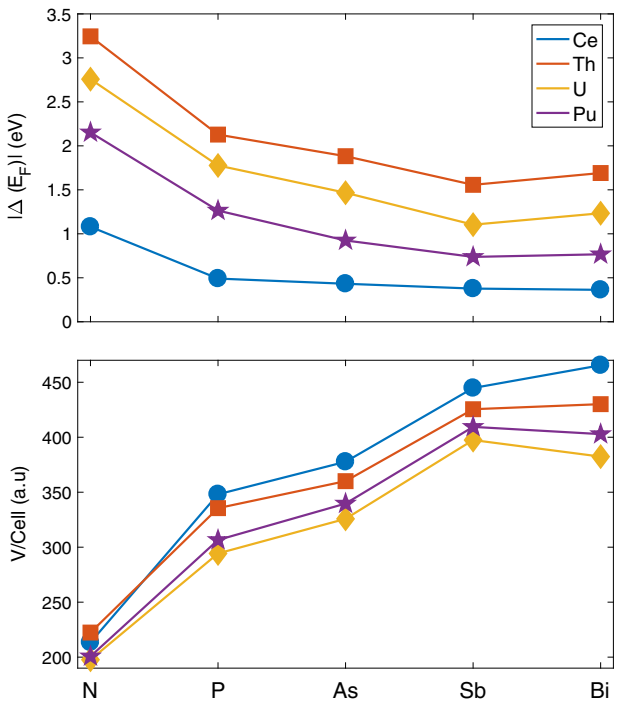

Fig. 7 Volume vs hybridization function at the Fermi level. The value of hybridization function at the Fermi level, $\Delta(E_F)$ (top), and the volume per unit cell (bottom) for the series of monopnictides considered in Fig. 6, shows an anti-correlation between the degree of localization and lattice constant

stability can be obtained. We also demonstrate that our database can be used to gain a handle on the strongly correlated aspects of *f*-electron materials. For this purpose, we have carried out calculations of hybridization functions to supplement our DFT results that provide a descriptor for the strength and nature of interactions between the localized *f*-electrons and itinerant conduction states. In this way, a number of trends in the lanthanide and actinide based series of compounds are identified, including an anti-correlation between the hybridization strength at the Fermi energy and the volume of the unit cell. fESD is thus well equipped to spur the discovery of next generation *f*-electron-based materials with novel functionalities.

METHODS

Generation of electronic and crystal structure data

Electronic ground states were obtained within the DFT framework using full-potential LAPW scheme as implemented in the WIEN2k package.[18] Exchange-correlation effects were accounted for by using the generalized gradient approximation (GGA) with Perdew–Burke–Ernzerhof (PBE) functional.[50] Calculations start with the transformation of CIF files into WIEN2k input files using the tool "cif2struct". During this preparation stage, the space group symmetry of the system is correctly captured based on all atomic positions. Computations were carried out using a large plane wave cut-off (RKmax = 9.0), and a 10 × 10 × 10 *k*-mesh in the first Brillouin zone (BZ) in order to obtain well-converged energies and electronic structures.

Band structure calculations were carried out along the standard high-symmetry lines in the BZs. Density of states (DOS) was obtained over a uniform energy mesh and resolved into the basis of real orbitals. Hybridization functions Δ(*E*) for 4*f* and 5*f* electrons were calculated by solving the local Green's function on the real axis.[9,10] Since we are only interested in total Δ(*E*), we used a real harmonic basis with spin but without any symmetry to project the local Green's function. In order to obtain an accurate estimate of Δ(*E*), the energy window was set to −20 to +10 eV around the Fermi level with a large *k*-mesh and a Lorentzian broadening of 0.25 eV. All calculations were maintained at the same level

of numerical accuracy, so that trends over large datasets can be captured correctly.

DATA AVAILABILITY

The *f*-electron structure database (fESD) is available at http://correlatedmaterials-lanl.org. The datasets generated during and/or analyzed during the current study are available from the corresponding authors on reasonable request.

ACKNOWLEDGEMENTS

We are grateful to Kipton Barros, Filip Ronning, Heike Harper, Bernardo Barbiellini, Robert S. Markiewicz, Joel Kress, and Avadh Saxena for important discussion. This work is supported by the Institute for Materials Sciences (IMS), NSEC at LANL and by the U.S.D.O.E at LANL under Project No. 20170680ER (T.A.) through the LANL LDRD program. Work at LANL was supported in part by U.S. DOE Basic Energy Sciences Core Program LANL E3B5 (J.-X.Z. and A.V.B.). The work at Northeastern University is supported by the US Department of Energy, Office of Science, Basic Energy Sciences grant number DE-FG02-07ER46352, and benefitted from Northeastern University's Advanced Scientific Computation Center (ASCC) and the NERSC supercomputing center through DOE grant number DE-AC02-05CH11231.

AUTHOR CONTRIBUTIONS

T.A. and H.H. designed the research of the manuscript. T.A. conceived the concepts of the database with suggestions from A.V.B. H.H. performed the high-throughput computations with help from A.I.K. and T.A. H.H. analyzed first principles calculations and strongly-correlated results with key help from T.A, A.B., J.-X.Z., H.C., J.W., and A.V. B. A.I.K. developed the database management system, application-programming interface, and the machine learning tools with key input from A.M. and S.E. H.H. and T.A. wrote the paper with key input from A.B. and J.-X.Z.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Steglich, F. et al. Superconductivity in the presence of strong Pauli paramagnetism: $CeCu_2Si_2$. *Phys. Rev. Lett.* **43**, 1892–1896 (1979).
2. Sorace, L. & Gatteschi, D. Electronic structure and magnetic properties of lanthanide molecular complexes. in *Lanthanides and Actinides in Molecular Magnetism* (eds Layfield, R. A. & Murugesu, M.) (Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, 2015).
3. Savrasov, S. Y., Kotliar, G. & Abrahams, E. Correlated electrons in δ-plutonium within a dynamical mean-field picture. *Nature* **410**, 793–795 (2001).
4. Carlos, L. D., Ferreira, R. A. S., de Zea Bermudez, V., Julian-Lopez, B. & Escribano, P. Progress on lanthanide-based organic–inorganic hybrid phosphors. *Chem. Soc. Rev.* **40**, 536–549 (2011).
5. Yemel'yanov, V. & Yevstyukhin, A. *The Metallurgy of Nuclear Fuel: Properties and Principles of the Technology of Uranium, Thorium and Plutonium* (Pergamon Press Ltd., Oxford, 1969).
6. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
7. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
8. Georges, A., Kotliar, G., Krauth, W. & Rozenberg, M. J. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.* **68**, 13–125 (1996).
9. Kotliar, G. et al. Electronic structure calculations with dynamical mean-field theory. *Rev. Mod. Phys.* **78**, 865–951 (2006).
10. Haule, K., Yee, C.-H. & Kim, K. Dynamical mean-field theory within the full-potential methods: electronic structure of $CeIrIn_5$, $CeCoIn_5$ and $CeRhIn_5$. *Phys. Rev. B* **81**, 195107 (2010).
11. Altman, N. S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **46**, 175–185 (1992).
12. Hosmer, D. W. Jr., Lemeshow, S. & Sturdivant, R. X. *Applied Logistic Regression* (John Wiley & Sons, New York, 2000).
13. Pal, S. K. & Mitra, S. Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Netw.* **3**, 683–697 (1992).
14. Curtarolo, S. et al. Aflow: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
15. Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
16. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: an open-access online database for data mining. *PLoS One* **12**, e0171501 (2017).
17. Singh, D. J. & Nordstrom, L. *Plane Waves, Peudopotential and the LAPW Method* (Springer, US, 2008).
18. Blaha, P., Schwarz, K., Madsen, G., Kvasnicka, D. & Luitz, J. *WIEN2K. An Augmented Plane Wave + Local Orbitals Program for Calculating Crystal Properties* (Techn. Universitaet Wien, Wien, Austria, 2001).
19. Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD)—present and future. *Crystallogr. Rev.* **10**, 17–22 (2004).
20. Gražulis, S. et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* **40**, D420–D427 (2011).
21. Hahn, T. *International Tables for Crystallography Volume A: Space-group Symmetry*. (The International Union of Crystallography, Kluwer Academic Press, Dordrecht, 1996).
22. Bansil, A., Lin, H. & Das, T. Colloquium: topological band theory. *Rev. Mod. Phys.* **88**, 021004 (2016).
23. Iyo, A. et al. New-structure-type Fe-based superconductors: $CaAFe_4As_4$ (A = K,Rb, Cs) and $SrAFe_4As_4$ (A = Rb,Cs). *J. Am. Chem. Soc.* **138**, 3410–3415 (2016).
24. Wang, Z.-C. et al. Superconductivity in $KCa_2Fe_4As_4F_2$ with separate double Fe_2As_2 layers. *J. Am. Chem. Soc.* **138**, 7856–7859 (2016).
25. Ahmed, T. et al. Magnetic, electronic, and optical properties of double perovskite Bi_2FeMnO_6. *APL Mater.* **5**, 035601 (2017).
26. Ahmed, T. et al. Site-mixing effect on the xmcd spectrum in double perovskite Bi_2FeMnO_6. *Appl. Phys. Lett.* **108**, 242907 (2016).
27. Goldschmidt, V. M. Die Gesetze der Krystallochemie. *Naturwissenschaften* **14**, 477–485 (1926).
28. Shannon, R. D. & Prewitt, C. T. Effective ionic radii in oxides and fluorides. *Acta Crystallogr. B* **25**, 925–946 (1969).
29. Shannon, R. D. Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Crystallogr. B* **32**, 751–767 (1976).
30. Vasala, S. & Karppinen, M. $A_2B'B''O_6$ perovskites: a review. *Prog. Solid State Chem.* **43**, 1–36 (2015).
31. Woodward, P. M. Octahedral tilting in perovskites. I. Geometrical considerations. *Acta Crystallogr. B* **53**, 32–43 (1997).
32. Cussen, E. J. & Battle, P. D. The influence of structural disorder on the magnetic properties of $Sr_2Fe_{1-x}Ga_xTaO_6$ ($0 \le x \le 1$). *J. Mater. Chem.* **13**, 1210–1214 (2003).
33. Galasso, F. & Darby, W. Ordering of the octahedrally coordinated cation position in the perovskite structure. *J. Phys. Chem.* **66**, 131–132 (1962).
34. Karpinsky, D., Troyanchuk, I., Bärner, K., Szymczak, H. & Tovar, M. Crystal structure and magnetic ordering of the $LaCo_{1-x}Fe_xO_3$ system. *J. Phys.: Condens. Matter* **17**, 7219 (2005).
35. Davis, J. M., Mugavero, S. J., Glab, K. I., Smith, M. D. & zur Loye, H.-C. The crystal growth and characterization of the lanthanide-containing double perovskites Ln_2NaIrO_6 (Ln = La, Pr, Nd). *Solid State Sci.* **6**, 413–417 (2004).
36. Kim, M. et al. Investigation of the magnetic properties in double perovskite R_2CoMnO_6 single crystals (R = rare earth: La to Lu). *J. Phys.: Condens. Matter* **27**, 426002 (2015).
37. Asai, K. et al. Magnetic properties of $REMe_{0.5}Mn_{0.5}O_3$ (RE = rare earth element, Me = Ni,Co). *J. Phys. Soc. Jpn.* **67**, 4218–4228 (1998).
38. Gilbu, B., Fjellväg, H. & Kjekshus, A. Properties of $LaCo_{1-t}Cr_tO_3$. I. Solid solubility, thermal expansion and structural transition. *Acta Chem. Scand.* **48**, 37–45 (1994).
39. Vyshatko, N. P., Kharton, V. V., Shaula, A. L. & Marques, F. M. B. Powder x-ray study of $LaCo_{0.5}Ni_{0.5}O_{3-\delta}$ and $LaCo_{0.5}Fe_{0.5}O_{3-\delta}$. *Powder Diffr.* **18**, 159–161 (2003).
40. Aharen, T. et al. Magnetic properties of the geometrically frustrated *s* = antiferromagnets, La_2LiMoO_6 and Ba_2YMoO_6, with the B-site ordered double perovskite structure: vidence for a collective spin-singlet ground state. *Phys. Rev. B* **81**, 224409 (2010).
41. Gemmill, W. R., Smith, M. D. & zur Loye, H.-C. Synthesis and magnetic properties of the double perovskites Ln_2NaRuO_6 (Ln = La, Pr, Nd). *J. Solid State Chem.* **177**, 3560–3567 (2004).
42. de Lima, O. F., Coaquira, J. A. H., de Almeida, R. L., de Carvalho, L. B. & Malik, S. K. Magnetic phase evolution in the $LaMn_{1-x}Fe_xO_{3+y}$ system. *J. Appl. Phys.* **105**, 013907 (2009).
43. Haque, M. T. & Kamegashira, N. Synthesis, structure and properties of $LnMn_{1/2}Rh_{1/2}O_3$ (Ln = rare earth. *J. Alloys Compd.* **395**, 220–226 (2005).
44. Herper, H. C. et al. Combining electronic structure and many-body theory with large databases: a method for predicting the nature of 4*f* states in Ce compounds. *Phys. Rev. Mater.* **1**, 033802 (2017).
45. Anderson, P. W. Localized magnetic states in metals. *Phys. Rev.* **124**, 41–53 (1961).

46. Blandin, A. & Friedel, J. Propriétés magnétiques des alliages dilués. interactions magnétiques et antiferromagnétisme dans les alliages du type métal noble-métal de transition. *J. Phys. Radium* **20**, 160–168 (1959).
47. Georges, A. Strongly correlated electron materials: dynamical mean-field theory and electronic structure. *AIP Conf. Proc.* **715**, 3–74 (2004).
48. Keis, N. & Komissarov, A. Cerium in structural and stainless steels and cast iron. *Met. Sci. Heat Treat.* **5**, 439–443 (1963).
49. Litsarev, M. S., Di Marco, I., Thunström, P. & Eriksson, O. Correlated electronic structure and chemical bonding of cerium pnictides and γ-Ce. *Phys. Rev. B* **86**, 115116 (2012).
50. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).