An open access *f*-electron database (FESD): a machine learning approach for the crystal and electronic structures of strongly correlated f-electron materials

Adnan Khair,[1] Towfiq Ahmed,[2] Abdullah Mueen,[1] and Alexander V Balatsky[3]

[1] *Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87131*
[2] *Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*
[3] *Institute for Materials Science, Los Alamos National Laboratory, Los Alamos, New Mexico 87545*

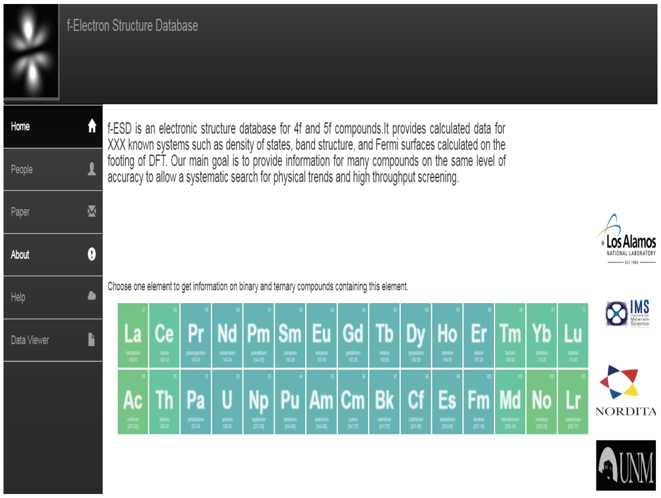
(Dated: April 11, 2017)

We are currently developing a database and data driven computational tools for theoretical understanding of structural and electronic properties in *f*-electron based materials. Due to a complex interplay among the hybridization of f-electrons to non-interacting conduction electrons, spin-orbit coupling, and strong coulomb repulsion of *f*-electrons, no model or first-principles based theory can fully explain all the structural and functional phases in such systems. Thus, motivated by the large need for the predictive modeling of *f*-electron compounds, we adopted a data-driven approach where the machine will learn from the DFT simulation of the existing crystals of actinides and lanthanides, and apply its knowledge to predict new stable systems with desired electronic properties. The newly predicted materials can then be synthesized and characterized at LANL. Towards this goal, we have started developing electronic structure database which will be aided by machine learning (ML) algorithm to extract complex electronic, magnetic and structural properties in *f*-electron system.

PACS numbers: 78.70.Dm, 71.10.Fd, 71.10.-w, 71.15.Qe

I. INTRODUCTION

A major class of strongly correlated systems includes materials with constituent f-orbital elements such as Lanthanides and Actinides. On one hand, the presence of high Z elements in some compounds gives it unique electronic, thermal, optical and magnetic properties and makes it extremely useful and interesting to study. But on the other hand, complex electronic interactions stemming from the localized *f*-orbital electrons make it very difficult to develop a predictive theoretical model, and thus leave the experimentalists to continue exploring based on serendipity. Our motivation for focusing on strongly-correlated compounds to develop our learning algorithm are two folds: (1) the complex physical properties of *f*-electron materials offer us an opportunity space to use a unique domain knowledge to develop our learning algorithm, and (2) materials predictability of new functional correlated compounds possess unprecedented promise for enhanced national security and energy related applications. To elaborate on these points, we note that the physics of Actinides is essentially a quantum many-body problem, and is notoriously hard to model, even with the state-of-the-art simulations techniques, e.g. density functional theory (DFT. For example, even the best available DFT+DMFT methods for Actinides require, as an input, a semi-empirical correlation parameter U, which significantly reduces its predictive power from first principles. This major theoretical challenge for predicting U stems from the complex and correlated mutual dependence between U and other features (e.g., lattice structure, energy, charge, density, spin, U) and can be captured using DFT. Our first innovation is to incorporate domain-specific expertise into the structure of the machine learning models. Specifically, we will use generalized kernel learning methods, and to determine similarity between compounds, we will seek a smart comparison between their respective quantum features obtained from DFT. Our second innovation is to apply the latest semi-supervised learning techniques: Our goal is to learn from the entire dataset of many compounds, even though only a small fraction contains experimentally verifiable values (e.g., U parameter) initially. To accomplish these goals we needed to construct a flexible and robust database. At the incubation stage of the process, we have founded f-electron database at IMS (NSEC). This database is running and it is unique. It is one-of-a-kind material database for the f electron materials. Our current database has reached a stage where it hosts now over 60,000 lanthanides and actinides based compounds and their crystal structure and DFT generated band-structure information. It also implements several machine learning (supervised and unsupervised) algorithms for predicting new compounds in a high throughput manner. With these preliminary works, we have so far demonstrated the enormous potential of the f-electron database which can


f-Electron Structure Database
Home
People
Paper
About
Help
Data Viewer
La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr
Los Alamos NATIONAL LABORATORY
IMS
NORDITA
UNM

f-ESD is an electronic structure database for 4f and 5f compounds.It provides calculated data for XXX known systems such as density of states, band structure, and Fermi surfaces calculated on the footing of DFT. Our main goal is to provide information for many compounds on the same level of accuracy to allow a systematic search for physical trends and high throughput screening.

Choose one element to get information on binary and ternary compounds containing this element.

FIG. 1: (Color online) Snapshot of *f* electron database. Currently hosting over 28000 compounds, their crystal structure and most of their bandstructure

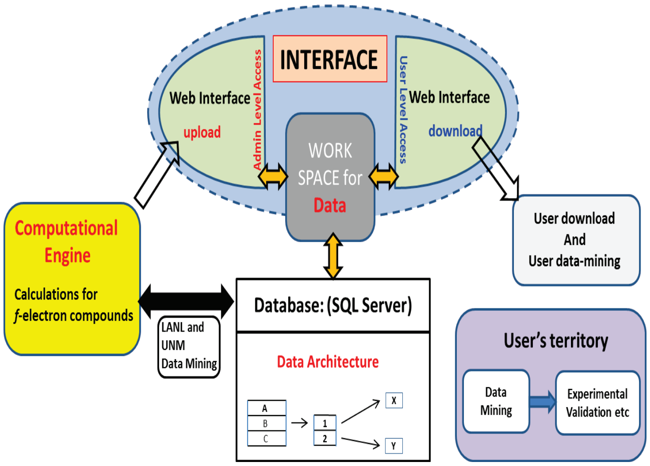


FIG. 2: (Color online) A schematic diagram showing the synergistic effort between computational modeling, Web-interface, database, and machine learning

be dedicated to serve and participate in LANLs core mission with actinide research.

During our database development, we have reached a stage where it hosts not only the 60,000 lanthanide and actinide based compounds and their crystal structure and DFT generated band-structure information, but also implements several machine learning (supervised and unsupervised) algorithms for predicting new compounds in a high throughput manner. We have organized 60,000 crystal structure with their crystal space symmetry group. For all binary actinide compounds, we have their DFT calculated band-structure and density-of-states. We currently have two manuscripts under preparation (copies available in the above web address). Several python based tools have been developed for pre- and post-processing of structure and simulated data. We have benchmarked several supervised learning methods (e.g. regression, neural networking) for cleaning our structure information. We have also implemented unsupervised clustering techniques to identify the outlier data available in the literature. With these preliminary works, we have demonstrated the immense potential of the f-electron database which can be dedicated to serve for LANLs core mission of actinide research. In Fig.2, we schematically present our database structure and proposed plan for material informatics of correlated materials. This f-electron database is designed in a highly versatile manner. Our machine learning tools can be used in a multi-purpose manner: 1) Material Informatics: We have demonstrated the correlation between properties in 200 Ce based binary compounds which provided useful insight in their Kondo temperature. 2) Improving Many-body Theory: We can use database to predict strong-correlation parameter U in f-electron systems. We attached both examples here. At this incubation period, our database is temporarily hosted at a protected server on NMC: http://199.229.237.45:8080/fesd/

II. CLEANING CRYSTAL STRUCTURE DATA

Crystal structures play an important role in understanding physical properties and information of the materials. By evaluating crystal structures, researchers can extract the expert information which helps to contribute to their research substantially. About 100 years of this discovery of crystal structure using X-ray diffraction, researchers who learned about these structures archived their research result in scientific articles. Large amount of material crystal structures have been discovered and these structures are being archived in crystallographic databases like Inorganic Crystal Structure Database (ICSD), Crystallography Open Database (COD) etc. These crystal structures are archived as a Crystallographic Information File(CIF) in these databases. Researchers around the world use these databases to learn about material. The Crystallographic Information File contains the crystal information of the materials. Not all the CIF file archived in the databases are complete. Some of the files lack a good amount of information which describes the crystal structures. The reason these CIF files missing information could be because while generating the CIF files these information got missed by the generator or the creator didn't put it in their result. Our goal is to refine these information and generate these CIF information complete as much as possible. To do this, we have chosen a machine learning statistical approach to predict the missing value using the existing value. In Crystallography, crystal structure describes the organization of the molecular structure. The crystal structure provides the gross structure of a system which describes the three-dimensional periodic arrangement of atoms, ions, or molecules in a crystal. The crystal structure also provides bond lengths, angles and electron density distribution. When crystal information in extracted from the material all these information are organized in scientific article. Later on, different organization extract the information from this published scientific article, put it to standard data exchange format and archive it. Inorganic Crystal Structure Database (ICSD), Cambridge Structural Database (CSD), Crystallography Open Database (COD) etc different organizations are responsible for archiving of these crystal structures information to the database.

III. WHY CIF CLASSIFICATION?

ICSD now contains 184,748 inorganic crystal structures and COD has around 362,577 crystal structures of organic, inorganic, metal-organic compounds and minerals, excluding bio polymers. These databases are built overtime with existing and newly discovered crystal structures. Not all the CIF files in these databases are complete. A CIF file for material could contain more than 100 fields value. Not all CIF for material is needed to have all of these fields. But some fields are important to describe the crystallographic structure of the material. There are 8 class of symmetry systems in the crystal system. These are:

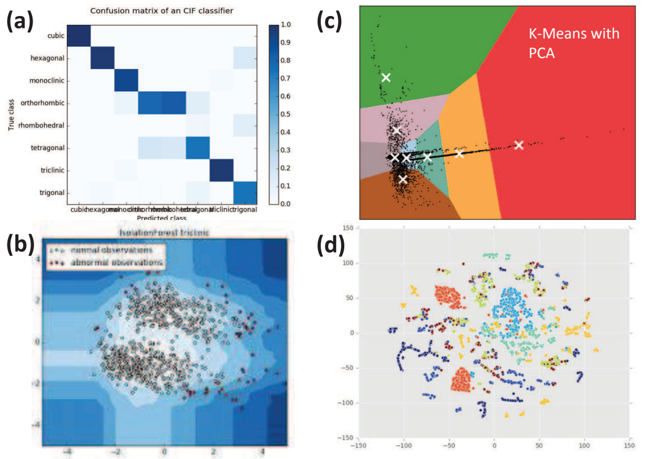


FIG. 3: (Color online) Different machine learning algorithm to correct missing crystal information. (a) Supervised KNN; (b) Outlier detection; (c) Unsupervised clustering using principle component analysis and K-Means algorithm; and (d) Clustering with TSNE algorithm; See text for details.

cubic
hexagonal
monoclinic
orthorhombic
rhombohedral
tetragonal
triclinic
trigonal

From the Crystallography Open Database (COD), out of 362,577 crystal structures 71,964 CIF file doesn't have the _symmetry_cell_setting value. These CIF files are unable to describe the symmetry system of the material. So, we use these 362,577 crystal structures of Crystallography Open Database (COD) and use machine learning approach predictive model to predict the missing 71,964 CIF files _symmetry_cell_setting that is symmetry system.

IV. DATABASE

There are 364,331 CIF files in COD database. CIF files are extensible, not all the CIF files have same set of keys or features. A single CIF file may contain 15 to more than 100 keys/features. We created a relational database 'CIFClassification' and a table 'compound' to store these CIF files data. A tool was built 'CIFDataExtractor' which extracted the set of 56 keys and values from each CIF files and inserted into the database table 'compound' to see the data availability in these keys. Not all of these keys/features have strong information of the symmetry system and a big number of these keys in CIF files have missing values. For the classification purpose, we only picked up 10 keys from the CIF files which are most available and have most crystal structure information. Followings are the chosen keys from the CIF files:

_symmetry_cell_setting
_cell_angle_alpha
_cell_angle_beta
_cell_angle_gamma
_cell_formula_units_Z
_cell_length_a
_cell_length_b
_cell_length_c
_cell_volume
_space_group_IT_number

We put these information of each CIF files in the new database table 'classficationdata'. The keys/features we picked also have null values or incorrect information in the _symmetry_cell_setting column. So we filtered out these rows in the table by removing if any column in the row has null values or symmetry value is incorrect. After filtering we got 275926 data out of 364,331.

Label	Count
rhombohedral	821
hexagonal	3366
cubic	3935
trigonal	4846
tetragonal	7835
orthorhombic	45017
triclinic	70675
monoclinic	139431
Total	275926

We denote the CIF files with missing/incorrect symmetry value as a test class. There are 71,964 CIF files in with missing values/incorrect information. We will be predicting the missing/incorrect symmetry value of this 71,964 CIF files.

So, in the database 'classificaitondata' table '_symmetry_cell_setting
' is the target class and other 9 keys are features.

Though, these 9 features in the data suppose to have real values, some values in the features have a value in incorrect format like the following:

_cell_angle_alpha	95.560(5)
_cell_angle_beta	103.442(5)
_cell_angle_gamma	96.291(5)
_cell_length_a	9.899(5)
_cell_length_b	11.729(5)
_cell_length_c	12.
_cell_volume	1364.9(11)
_space_group_IT_number	?

so, a database function was written 'modify()' to change to all these incorrect values to correct format. After running this function to all the rows a real value like X.XX(X) changed to X.XX.

V. METHOD

Our objective here is to predict the missing/incorrect symmetry value of the 71,964 files in the Crystallography Open Database (COD). But right now we don't have the opportunity to validate our predicted class. For this reason, we use k-fold cross validation approach to verify the predication performance. We did 10 k-fold cross validation from the 275,926 label data and calculated the accuracy. In this process, we did 10 iterations. In each iteration, we picked 10% of the data as test from this label data and 90% a training data. That means in each iteration it picked 248,333 data a training data and 27592 data as test data. Then it trained the classification algorithm with 248,333 training data and tested with the 27592 test data. We accumulated all these 10 iterations result score and average the total score to get the accuracy. In this process, confusion matrix is also generated for all these 10 iterations. After 10 iteration, the average of the 10 confusion matrix was averaged. From this averaged confusion matrix we calculated the precision, recall and f-measure.

For a particular row in the data, the features value are like following:

_cell_angle_alpha	90
_cell_angle_beta	90
_cell_angle_gamma	120
_cell_formula_units_Z	6
_cell_length_a	14.2066
_cell_length_b	14.2066
_cell_length_c	33.41
_cell_volume	5839.7
_space_group_IT_number	167

We applied Z-score normalization method to scale the data.

VI. RESULT

A. Classification

Different classification algorithms was run with different configurations on the k-fold cross validation process. These are 1) Logistic Regression 2) KNN 3) Support Vector Machine 4) Multilayer perceptron. The results are evaluated below

1. *Logistic Regression CIF classification*

Here is the result of logistic regression:

Data size	Accuracy	Precision	Recall	F-Measure
275926	90.4%	0.764900	0.655588	0.706

Confusion matrix figure-4:

2. *KNN CIF classification*

Here is the result of KNN CIF classification:

With Nearest Neighbor = 3, Metric = Euclidean

Data size	Accuracy	Precision	Recall	F-Measure
275926	98.1%	0.875740	0.861904	0.87

Confusion matrix figure-5:

With Nearest Neighbor = 5, Metric = Euclidean

Data size	Accuracy	Precision	Recall	F-Measure
275926	98%	0.873555	0.860282	0.882

Confusion matrix figure-**??**:

FIG. 4: Confusion matrix of logistic regression CIF classification

FIG. 5: Confusion matrix of KNN (Nearest Neighbor = 3) CIF classification

VII. COMPUTATIONAL METHODOLOGY

VIII. CONCLUSION

IX. ACKNOWLEDGEMENT

We thank A. Svane, N. E. Christensen, M. van Schilfgaarde, and A. N. Chantis for useful discussions and collaboration on related work. This work was supported by U.S. DOE at LANL under Contract No. DE-AC52-06NA25396, the LANL LDRD Program (T.A., R.C.A. & J.-X.Z.), and the Europe VR Program (A.V.B.).