


Material informatics for layered high- T_c superconductors

Cite as: APL Mater. **8**, 061104 (2020); <https://doi.org/10.1063/5.0004641>

Submitted: 13 February 2020 . Accepted: 04 May 2020 . Published Online: 04 June 2020

Zhong-Li Liu , Peng Kang , Yu Zhu, Lei Liu, and Hong Guo

COLLECTIONS

 This paper was selected as Featured



View Online



Export Citation



CrossMark



AIP Conference Proceedings

**The 18th International Conference
on Positron Annihilation**

ORDER PRINT EDITION



Material informatics for layered high- T_C superconductors



Cite as: APL Mater. 8, 061104 (2020); doi: 10.1063/5.0004641

Submitted: 13 February 2020 • Accepted: 4 May 2020 •

Published Online: 4 June 2020



Zhong-Li Liu,^{1,2,a)} Peng Kang,² Yu Zhu,³ Lei Liu,³ and Hong Guo^{2,4}

AFFILIATIONS

¹College of Physics and Electric Information, Luoyang Normal University, Luoyang 471934, China

²Centre for the Physics of Materials and Department of Physics, McGill University, Montreal, Quebec H3A 2T8, Canada

³Nanoacademic Technologies, Inc., Brossard, Quebec J4Z 1A7, Canada

⁴Centre for Computational Sciences, Sichuan Normal University, Chengdu 610066, China

^{a)} Author to whom correspondence should be addressed: zl.liu@163.com

ABSTRACT

Superconductors were typically discovered by trial-and-error aided by the knowledge and intuition of individual researchers. In this work, using materials informatics aided by machine learning (ML), we build an ML model of superconductors, which is based on several material descriptors with apparent physical meanings to efficiently predict critical superconducting temperature T_C . The descriptors include the average atomic mass of a compound, the average number of electrons in an unfilled shell, the average ground state atomic magnetic moments, the maximum difference of electronegativity, etc. To fully optimize the ML model, we develop a multi-step learning and multi-algorithm cross-verification approach. For known high T_C superconductors, our ML model predicts excellent T_C values with over 92% confidence. When the ML model is applied to about 2500 layered materials in the inorganic crystal structure database, 25 of them are predicted to be superconductors not known before, including 12 cuprates, 7 iron-based crystals, and 6 others, with T_C ranging from ~32 K to ~138 K. The findings shed considerable light on the mapping between the material descriptors and T_C for layered superconductors. The ML calculates that in our descriptors, the maximum difference of electronegativity is the most important one.

© 2020 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0004641>

I. INTRODUCTION

Since the discovery of superconductors by Heike Kamerlingh Onnes in 1911,¹ they have attracted scientists for more than a century. So far, the highest record of superconducting critical temperature T_C at atmospheric pressure has been close to 140 K, found in the class of cuprate superconductors² having a layered crystal structure. In 2006, Kamihara *et al.* discovered layered iron-based superconductors³ at a reasonably high temperature,⁴ which appear to contradict an established intuition that magnetism should hinder superconductivity. More recently, H₂S recorded $T_C = 203$ K at extremely high external pressures ~150 GPa,⁵ while $T_C = 260$ K was reported for lanthanum decahydride at even higher pressure in 2019.^{6,7}

Despite the tremendous knowledge of superconductivity accumulated so far, understanding the basic pairing mechanisms giving

rise to high values of T_C remains a long-lasting goal. On the other hand, the large knowledge base should allow us to discover high- T_C superconductors not known before, and it is the purpose of this work to report our investigation in this direction. Conventionally, materials discovery heavily relies on empirical knowledge or intuition of individual researchers, followed by extensive trial-and-error experimentation.⁸ Thus, for superconductors, one typically starts from some known superconducting compound and performs element substitution, alloying, doping, etc., on the compound, as has been successful for the classes of cuprates, iron-based materials, organic materials, and/or other materials. A limitation of this approach is the relatively narrow search space as it is based on the knowledge of some known materials. However, such a limitation can now be very effectively removed by the data-driven materials informatics (MI) techniques assisted by machine learning (ML), as demonstrated by many impressive works.^{8–13} Our work is based on using

MI + ML to discover possible candidates of high- T_C compounds from the Inorganic Crystal Structure Database (ICSD) that contains nearly 200 000 experimentally synthesized materials.

For superconductors, several data-driven works can be found in the literature. Analysis of the existing cuprates has found that only one single hybridized band cuts the Fermi level E_F for each CuO_2 -plane.¹⁴ Using several characteristics of the electronic structure as the descriptors, over 100 layered materials as potential superconductors were screened out from the ICSD.¹⁴ In addition, with only two geometrical parameters and a heuristic T_C formula¹⁶ as the main descriptors, Ref. 15 modeled the T_C of known superconductors, although no new candidates were proposed. Based on the superconducting ML tests, some important limitations of traditional cross-validation (CV) for evaluating the ML model performance for material discovery were identified, and the leave-one-cluster-out (LOCO) CV method was proposed and tested to be much more robust than the traditional CV.¹³

Most recently, superconducting models were built with the help of a ML algorithm and employed to predict over 30 non-cuprate and non-iron based superconductors.¹² It is noted that in Ref. 12, the screened potential superconductors were generally bulk materials, and the authors did not pay special attention to layered materials, although layered materials were known to be the main source of high- T_C superconductors. On the technical side, the potential superconductors in Ref. 12 were screened by a single machine learning algorithm; as such, the issue of the algorithm bias needs to be addressed. Therefore, in this work, we focus on the layered materials in the ICSD and screen out potential layered high- T_C superconductors by a three-algorithm cross-check strategy that significantly lowers the algorithm bias.

In the following, we build a ML model of superconductors to efficiently predict critical superconducting temperature T_C . Our philosophy is the following: (i) We avoid using high level descriptors, such as those extracted from electronic structures, which require lengthy computation. In contrast, we carry out MI to find several simple material descriptors with apparent physical meaning, which prove to be very effective indicators for high values of T_C . Our descriptors include the average atomic mass of a compound, the average number of electrons in an unfilled shell, the average ground state atomic magnetic moments, the maximum difference of electronegativity, etc., which are readily and easily obtained from chemical components of the materials in the database. (ii) We search from the experimental database, so that our predicted candidates are known to have been synthesized before. In particular, we apply our ML to 2500 layered materials, which we find by MI from the ICSD. (iii) We optimize our ML model of superconductivity by developing a multi-step learning and multi-algorithm cross-verification approach. Here, the cross-verification is done by three independent ML algorithms to increase the reliability. As a result, our ML gives high confidence of its prediction, at about 92%. This research philosophy is built into our MI + ML platform named iMAT²⁸ (more detailed features are described in the [supplementary material](#)), for intelligence based materials discovery. For known high- T_C superconductors, our ML model predicts excellent T_C values as compared to experimental measurements. Then, 25 of the layered compounds from the ICSD are predicted to be superconductors not known before, including 12 cuprates, 7 iron-based crystals, and 6 others, with T_C ranging from ~ 32 K to ~ 138 K. The findings shed

considerable light on the mapping between our material descriptors and T_C for layered superconductors.

This paper is organized as follows. In Sec. II, the material informatics analysis of superconductors is presented; Sec. III presents our machine learning technique for building superconductor models; Sec. IV is for results and discussion, and a short summary is presented in Sec. V.

II. SUPERCONDUCTOR INFORMATICS

The MI + ML for material discovery requires two types of databases as input. The first one, named *property database*, is with labeled specific property obtained by experimental measurements. For this work, we use SuperCon,¹⁷ which documents measured T_C values for over 30 000 records. The property database is used by ML algorithms to learn and build a material model, in our case, a superconductor model that statistically predicts T_C values. The second database is the one that collects a large number of material structures, named *structure database*, which is used by ML algorithms to screen materials having structures and chemical components similar to the material model of ML. Here, we build our structure database by screening layered materials from the ICSD because cuprates and iron-based high- T_C superconductors have layered structures. Finally, the ML material model is applied to the structure database to make statistical predictions for the specific property, in this work, the T_C values of the MI candidates. In the MI + ML methodology, it is critical to have material descriptors that are easy to find and full of physical meanings, and effective for prediction, in our case, a set of descriptors that somehow relates to high values of T_C . We build our descriptor set by physical arguments and MI.

A. Property and layered-structure databases

We use SuperCon¹⁷ and focus on metals, alloys, cuprates, iron-based superconductors, and other inorganic superconductors. After purifying ~ 30 000 entries of SuperCon, 15 819 non-replicate records are retained¹⁸ to form our property database for building the ML superconductor model. The resulting T_C database contains the cuprate, iron-based superconductors, and other superconductors.¹⁹ The reason why we did not train the superconducting models on any specific class of superconductors is because such ML models are not likely to be able to extrapolate to entirely different classes.^{12,13}

Next, to build the structure database, we start from the ICSD, which is a large crystal database containing ~ 200 000 experimentally identified inorganic crystal entries.¹⁸ Since cuprates and iron-based high- T_C superconductors are often found to possess a layered pattern of alternating oxides,^{19,20} our structure database is built by identifying layered materials from the ICSD, using the dimensional formula of Ref. 21. As shown in Fig. 1, by counting the *maximum number of bonded atoms* in the original $1 \times 1 \times 1$ unit cell and also in the $2 \times 2 \times 2$ supercell, we can determine the dimensionality of a crystal via²¹

$$2^d = \frac{N_{2 \times 2 \times 2}}{N_{1 \times 1 \times 1}}. \quad (1)$$

Whether two atoms are bonded or not is determined by the distance between them. To this end, we start from the smallest

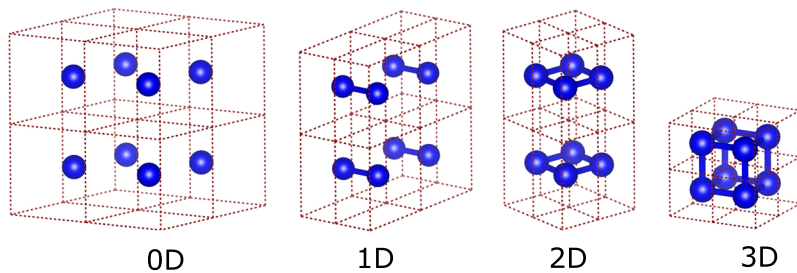


FIG. 1. A schematic diagram illustrating the algorithm for identifying the dimensionality of a crystal structure. In a $2 \times 2 \times 2$ supercell, the maximum number of bonded atoms is 2^0 , 2^1 , 2^2 , and 2^3 times that of the unit cell, corresponding to 0D, 1D, 2D, and 3D, respectively.

bonding distance R_1 among all atoms in the supercell and determine d by Eq. (1). If $d = 0$, the crystal is temporarily judged as a molecular crystal. Then, we gradually increase the distance criterion to R_2 , which makes $d = 2$, and the crystal is temporarily labeled as layered. Finally, the bonding distance is further increased to R_3 that makes $d = 3$. Then, $\Delta R = R_3 - R_2$ is the range of distance for crystals to hold layered characteristics. We took $\Delta R = 0.75$ Å in scanning bonding distance with an interval of 0.25 Å. Using this algorithm, we screened out ~ 3500 layered structures from $\sim 200\,000$ ICSD entries, which are further purified by eliminating replicate structures using a nearest-triangle algorithm,²² and by removing materials having very low symmetry because known superconductors have relatively high structural symmetries. The resulting ~ 2500 layered materials become the structural database, from which we apply the ML superconductor model to find candidates with high T_C .

B. Descriptors

To train a superconductor model by ML, a set of physically motivated but easily obtained descriptors of T_C values is needed.²³ Enlightened by hydrogen-rich superconductors²⁴ and, in particular, the H_3S compound, which goes superconducting at $T_C = 203$ K (albeit under 150 GPa),⁵ the atomic mass is important according to the electron–phonon theory of superconductivity. In H_3S , the maximum mass difference ΔM of atoms is large but the average mass \bar{M} is much smaller, namely, there are heavy atoms but the majority are light atoms. We analyzed ΔM and \bar{M} for all the high- T_C materials in our property database. For a given chemical formula $A_xB_yC_z$, the average of a physical quantity F (e.g., \bar{M}) is calculated by

$$\bar{F} = \frac{x F_A + y F_B + z F_C}{x + y + z}. \quad (2)$$

Figure 2(a) plots T_C vs ΔM and \bar{M} for all compounds in the property database. We discover that materials having high values of T_C tend to have $35 < \bar{M} < 110$ amu, while ΔM is larger than ~ 120 amu. T_C values peak against these parameters, and thus ΔM and \bar{M} are used as descriptors. In the same manner, we investigated T_C vs the average number of electrons in an unfilled subshell, the atomic magnetic moment in the ground state, the average atomic electronegativity, and the maximum difference of the electronegativity. Since T_C values peak against these parameters as shown in Figs. 2(b)–2(d), they are added into the set of descriptors. Finally, the occurrence probability of the elements in superconductors is added to our descriptor set, by tabulating each element's atomic number in the feature array of our ML algorithms.

III. MACHINE LEARNING

A. Multi-algorithm cross-validation

Having prepared the property database and physics motivated descriptors, in this section, we build the ML model for superconductors. We employ three ML algorithms to train against the property database: the random forest regression (RFR),²⁵ the support vector regression (SVR),²⁶ and the artificial neural network²⁷ regression, which is often referred to as deep learning regression (DLR). For each, we optimized parameters to guarantee that neither over-fitting nor under-fitting occurs. For interested readers, the technical details are further discussed in the [supplementary material](#).

To make the three trained models more reliable, a k-fold cross-validation is applied, where the property database is split into

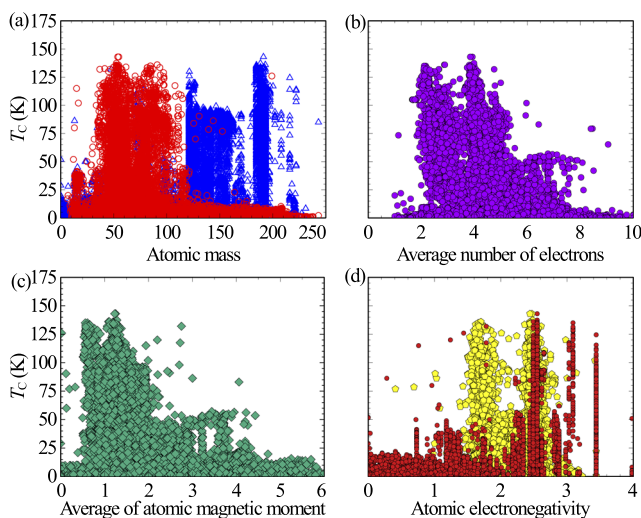


FIG. 2. T_C from the property database vs (a) the average atomic masses \bar{M} (red circles) and the maximum differences of masses (blue triangles) ΔM in amu; (b) the average number of electrons in an unfilled subshell; (c) the average atomic magnetic moment in μ_B (the theoretical value for each atomic element in the isolated state); (d) the average of atomic electronegativity (yellow pentagons) and the maximum difference of electronegativity (red solid circles) in the Pauling electronegativity scale.

k separated parts. One of them is retained as the testing database, while the remaining $k - 1$ are trained for the superconductor model. This is iterated for all k , and the mean scores are obtained. The test results are listed in Fig. S1 in the [supplementary material](#). By investigating different k values, we discover that a 10-fold cross-validation is adequate for the three ML algorithms. Finally, having validated the three individual ML algorithms with the property database, to reduce the unexpected bias of the individual ML, we combine them to perform *multi-algorithm cross-validation* of predicted T_C , which we find to help improving the reliability and accuracy of the prediction.

B. Multi-step learning

In our property database, there may exist some unknown “noisy data,” either due to errors in experimental measurements or recording, as well as other possible sources, e.g., the data not within the main stream properties of superconductors. These “noisy data” can hinder convergence toward the ML model. We deal with this problem by developing a multi-step learning approach. First, a model is learned from the original training data and used to predict T_C values of all materials in the property database. If the deviation of predicted T_C differs from the measured value by more than a pre-specified tolerance, e.g., 30 K, the data in the dataset are ignored in the next learning step, which trains a new model from the remaining dataset. This way, after each training step, some “noisy data” drop out in the next training step. In every learning, the model is evaluated by the 10-fold cross-validation discussed above. The learning is terminated until the specified mean score of R^2 of the 10-fold cross-validation is reached.

The multi-step learning process is reminiscent of the data-cleaning step of machine learning, while its application is for the description of the main stream properties of the whole data. By tuning the pre-specified tolerance of T_C (30 K used), it is inevitable to reach models that are able to make accurate predictions in relatively small regions of chemical space, where many training data points are available. For each single algorithm, we also list the predicted potential layered high- T_C superconductors beyond 30 K in Table S-I ([supplementary material](#)), from which we clearly observe a good element divergence of the layered materials. From Table S-I, we also note that each algorithm has its bias even after the multi-step learning. First, the numbers of predicted high- T_C superconductors by the three algorithms are different. Then, the materials and predicted T_C values are also different, e.g., C_2N_3H is predicted to have T_C beyond 30 K by RFR and DLR, but, i.e., the SVR predicted value is below 30 K. Therefore, the multi-algorithm cross-validation strategy can make the predicted results more reliable. In this work, our aim is to improve the reliability of prediction with the help of the multi-algorithm cross-validation strategy, not to discover all the potential superconductors.

By testing $R^2 = 0.90, 0.92$, and 0.94 , the $R^2 = 0.92$ model is used, which avoids over-cleaning of the noisy data. All the R^2 scores of the 10-fold cross-validation for three individual ML algorithms and of the multi-step cross-validation are listed in Table S-II of the [supplementary material](#). After our multi-step learning to reach $R^2 = 0.92$ by all three ML algorithms, we carry out a further step by using 90% of the property database to train and then apply the trained models to the remaining 10% of the data. The predicted

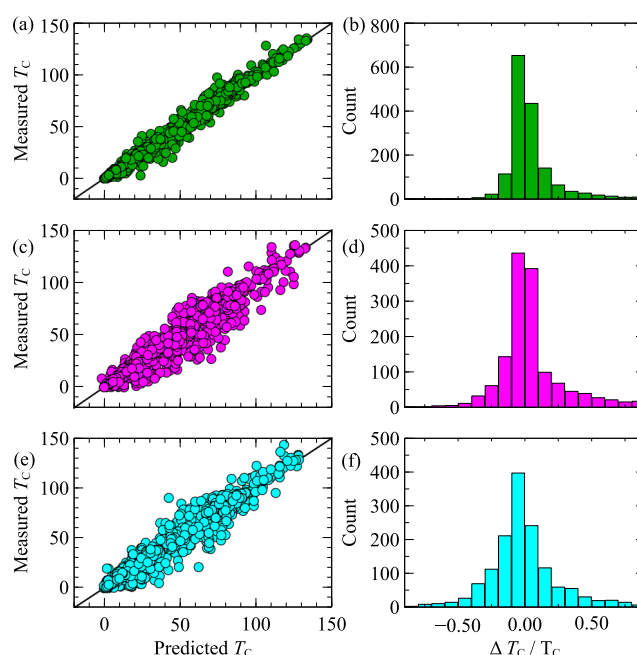


FIG. 3. Comparison of predicted and measured T_C by three ML algorithms with $R^2 = 0.92$. Left column: predicted vs measured T_C . Right column: count of relative deviation of the predicted with respect to the measured T_C . [(a) and (b)] The random forest regression. [(c) and (d)] The support vector regression. [(e) and (f)] The deep learning regression.

T_C values for the testing set by the three ML algorithms are plotted in the left column of Fig. 3 against the measured data, and the agreements are quite satisfactory. The right column of Fig. 3 plots the count of relative deviation, by which we see that the random forest regression gives the best result for this work. To numerically reflect the degrees of agreement, we calculated their coefficients, R^2 , and compared to the measured values. The R^2 values are 0.989 for the random forest regression, 0.944 for the support vector regression, and 0.957 for the deep learning regression. This quantitatively indicates that the random forest regression does the best.

IV. RESULTS AND DISCUSSION

Using the superconductor models trained by the three ML algorithms as discussed above, we now predict T_C values for all the (layered) materials in the structure database. 30 K is taken as the criterion for high- T_C materials. From the ~2500 layered materials in the structure database, ML predicts 38 high- T_C candidates as listed in Table S-III of the [supplementary material](#), which are cross-checked by the three ML algorithms only according to their chemical components. Of the 38, it turns out that 10 cuprates and 3 iron-based materials are actually in our *property database*, and their ML predicted T_C values are listed in Table S-III together with the corresponding measured values. The agreement is excellent. This is very satisfactory since our ML only relies on the easily obtained descriptors.

The most important discovery is the predicted 25 candidates not known to be superconductors before, with T_C from ~ 32 K to ~ 138 K. The average T_C predicted by the three ML algorithms is listed in Table I: there are 12 cuprates, 7 iron-based materials, and 6 others. It is very interesting that three of them have predicted $T_C > 100$ K. They are all TI-based materials with similar chemical elements to known superconductors (see Table S-III in the supplementary material) and denoted as TI-3446, TI-5669, and TI-78812 according to the notation of TI-based superconductors. Another predicted TI-based candidate is TIYBa₂Cu₂O₇, having a similar layered structure to YBa₂Cu₂O₇.

For the 25 ML predicted candidates not known to be superconductors before, the ultimate verification shall be experimental. Here, we can discuss their likelihood from several points of view. For known cuprate superconductors, a hybridized band cuts the Fermi level for each CuO₂-plane,¹⁴ and one or more flat-bands below the energy of highest occupied electronic state lead to a large peak in the electronic density of states (DOSs) to help enhancing T_C .¹² DOS peaks close to E_F due to the Van Hove singularity may also enhance T_C .^{29,30} This DOS phenomenology can be checked by DOSs of known superconductors from the OQMD,^{31,32} and in Fig. S2 of the supplementary material, the total DOSs of four different high- T_C superconductors belonging to different material classes are plotted. From Fig. S2, it appears that the four known high- T_C superconductors do have DOS peaks near E_F , although in Fig. S2(d), the

peak is actually not large. Following this scenario, we have calculated electronic DOSs for several ML predicted candidates in Table I by density functional theory (DFT). The computational details are similar to that of the OQMD and described in the supplementary material. The calculated DOSs are presented in Fig. 4. Indeed, there are high DOS peaks close to E_F for TI₃Ba₄Ca₄Cu₆O₁₉, Ba₂InBrO₃, and Ba₃In₂O₆, while for Sr₂Cu₂O₃, the DOS peak is less pronounced. Finally, DOSs of candidates TIYBa₂Cu₂O₇ and others were calculated in the OQMD^{31,32} and are shown in Fig. S3 of the supplementary material: most of them have peaks, although some are not very close to E_F .

Another support for the ML predicted candidates in Table I to be superconductors may come from a heuristic formula.¹⁶ It states that the upper limit of T_C of layered materials having two types of interacting layers corresponding to two types of carrier reservoirs I (electrons) and II (holes) is well-fit by the following expression:

$$T_{C0} = k_B^{-1} e^2 \Lambda / l \zeta = \beta / l \zeta k_B, \quad (3)$$

where $\beta = 0.1075 \pm 0.0003$ eV Å², l is the spacing distance between the interacting charges within the layer, and ζ is the distance between the interacting layers. The parameter l can be calculated from $l = (\eta \sigma_I / A)^{-1/2}$, where η is the total number of planes, σ_I is the fractional charge per type I layer, and A is the area of the basal plane. $\eta \sigma_I / A$ is defined as the superconducting interaction density.¹⁶

TABLE I. The predicted T_C s of 25 candidates from RFR, DTR, and DLR ML algorithms and their average, as well as from the heuristic formula estimate. The cuprates are in italics, iron-based materials are in bold, and others are in bold italics. The third column named SGN is the space group number.

Chemical formula	ICSD entry	SGN	RFR (K)	SVR (K)	DLR (K)	Average (K)	Heuristic (K)
Tl ₃ Ba ₄ Ca ₄ Cu ₆ O ₁₉	09 965	139	113.6	114.3	108.0	112.0	132.8
Tl ₅ Ba ₆ Ca ₆ Cu ₉ O ₂₉	094 966	123	113.3	113.8	106.9	111.3	137.2
Tl ₇ Ba ₈ Ca ₈ Cu ₁₂ O ₃₉	094 967	139	113.5	113.6	106.3	111.1	138.3
TIYBa ₂ Cu ₂ O ₇	074 163	123	48.4	65.6	61.7	58.6	88.8
Ba ₂ CuF ₆	021 055	69	49.5	47.9	49.5	49.0	77.2
Ba ₂ Cl ₂ Cu ₃ O ₄	081 196	139	70.3	60.8	37.1	56.1	80.6
Ba ₂ Cu ₃ Br ₂ O ₄	036 128	139	55.3	67.7	45.4	56.1	55.6
BaCuFSe	075 585	129	64.1	46.2	35.0	48.4	
TeBaCuF	245 624	129	54.5	30.4	33.6	39.5	
Sr ₂ ZnS ₂ Cu ₂ O ₂	084 735	139	56.8	35.4	34.8	42.3	
GdSeCuO	080 358	129	37.3	33.4	34.2	35.0	46.9
Sr ₂ Cu ₂ O ₃	150 912	64	60.1	37.1	84.0	60.4	87.8
Sr ₃ Cu ₂ O ₅ Fe ₂ Se ₂	154 203	139	32.1	50.0	30.5	37.5	
Sr ₂ F ₂ OFe ₂ Se ₂	249 690	139	34.8	46.3	34.5	38.5	35.5
Ca ₂ CuO ₃ FeSe	169 993	129	53.7	48.7	42.1	48.2	
NdOFeAs	164 676	129	37.2	48.1	49.3	44.9	42.9
NdOFeAs	236 650	67	37.2	48.1	49.3	44.9	34.1
GdOFeAs	422 003	129	39.5	48.0	52.2	46.6	44.6
GdOFeAs	425 015	67	39.5	48.0	52.2	46.6	31.7
Ba ₃ In ₂ O ₆	065 258	139	36.9	42.4	58.3	45.9	78.0
Ba ₃ In ₂ Cl ₂ O ₅	069 636	139	39.0	38.2	39.0	38.7	83.7
Ba ₂ InClO ₃	081 877	129	39.7	38.2	39.4	39.1	81.8
Ba ₃ In ₂ Br ₂ O ₅	071 603	139	39.4	40.3	54.7	44.8	82.6
Ba ₂ InBrO ₃	081 878	129	40.3	40.3	55.0	45.2	82.0
Ba ₂ N	067 510	166	42.6	31.3	40.4	38.1	

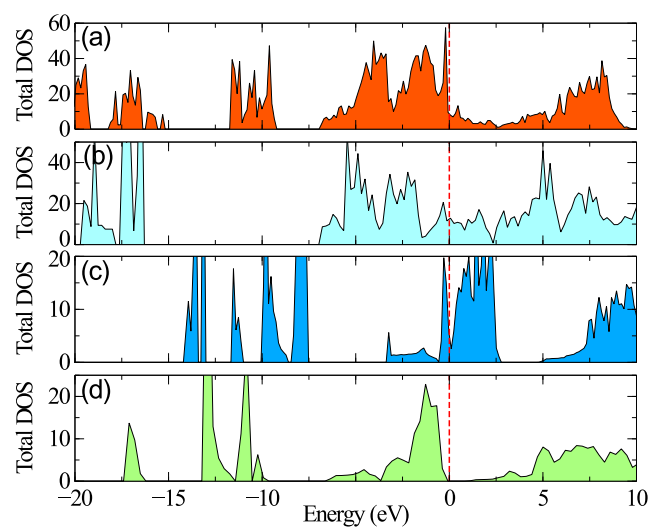


FIG. 4. The calculated total DOSs of four superconductor candidates. (a) $\text{Ti}_3\text{Ba}_4\text{Ca}_4\text{Cu}_6\text{O}_{19}$; (b) $\text{Sr}_2\text{Cu}_2\text{O}_3$; (c) $\text{Ba}_2\text{InBrO}_3$; (d) $\text{Ba}_3\text{In}_2\text{O}_6$. Significant DOS peaks are seen near E_F (red-dashed line).

Reference 16 shows that the above heuristic formula well fits many known high- T_C superconductors. Using Eq. (3), we estimated T_C of the candidates as shown in Table I and Table S-III of the supplementary material. For materials with apparent layered characteristics, the values of heuristic T_C are in reasonable agreement with experiments (see Table S-III). For the ML predicted candidates, the heuristic estimations largely support the average prediction (see Table I). We note that the heuristic formula cannot be used to predict new superconductors, and it is only used to “estimate” a T_C value for the layered candidates found by the MI + ML algorithms. Note that this heuristic formula only fits to layered crystals, which have two types of interacting layers corresponding to two types of carrier reservoirs, electrons and holes.¹⁶ Taking $\text{YBa}_2\text{Cu}_3\text{O}_{6.92}$ as an example, the type I reservoir is $\text{BaO}-\text{CuO}-\text{BaO}$, and the type II reservoir is $\text{CuO}_2-\text{Y}-\text{CuO}_2$. In Tables I and S-III, there are several materials, which do not possess similar reservoirs, and their T_C s cannot be estimated using this heuristic formula.

It is very interesting to learn how the ML algorithms picked the layered superconductor candidates listed in Table I, by investigating the physical descriptors of them. In Fig. S4 of the supplementary material, we plot the predicted T_C against the descriptors calculated by ML for the candidates, together with those from our property database shown in Fig. 2. For the predicted layered candidates, we note that the average atomic mass \bar{M} is in a very narrow region, from ~ 50 to 75 amu. Such a narrow concentration is also observed on the average number of electrons in an unfilled subshell—ranging from 3.9 to 4.7 ; and on the average of the electronegativity—in the range from 2.2 to 2.6 . We also note that these averaged values become more concentrated for those candidates with higher T_C values. For the maximum atomic mass differences ΔM , not only it distributes in very narrow regions [see Fig. S4(a) in the supplementary material], it takes very large values beyond ~ 120 amu. The ML indeed picked candidates, which have \bar{M} much smaller than the largest mass in the material. In other words, the known high- T_C superconductors

in the property database are dominated by materials, whose chemical formula has light elements as the majority but has at least a much heavier atom in the unit cell. As for the ground state atomic magnetic moments, the ML picked candidates have ~ 1 to $2 \mu_B$ from our structure database, clearly learned from the iron-based superconductors³ in the property database.

These “learned” superconductor information may be helpful for the ultimate theory of the pairing mechanism of layered high- T_C superconductors. Another very interesting outcome is the relative importance of different descriptors calculated by the ML algorithms. For RFR, this can be easily obtained, and we present it using the percentage of importance in Table II. Two sets of percentages are calculated by RFR: one by taking into account the occurrence possibility of atomic elements in the superconductors and the other not. Surprisingly, the maximum difference of electronegativity $\delta\eta$ appears to be the most important descriptor for T_C with 30.7% and 56.1% for the two sets. This is followed by average atomic magnetic moment $\bar{\mu}$ and average atomic masses \bar{M} . The importance of $\delta\eta$ indicates that high- T_C compounds tend to have at least two types of atoms, one with strong abilities to attract electrons and the other much weaker ability. This tends to result in ionic bonds and from the distribution of average atomic electronegativity (>2.2) in Fig. S4(d) of the supplementary material, majority of the atoms indeed have non-metal properties.

It is interesting to observe in what way the informatics (e.g., Figs. 1 and S4) discovered from the property database can be related to the current understanding of pairing mechanisms of high- T_C superconductors. For instance, the mass descriptors, \bar{M} and ΔM [Figs. 1(a) and S4(a)], in our informatics appear to be in-line with the inter-layer coupling model, in which the phonon plays an important role,³⁴ while the prominence of the magnetic descriptor, average ground state atomic magnetic moments [Figs. 1(c) and S4(c)], appears to be in-line with the antiferromagnetic spin fluctuation model.³³ The electronegativity descriptor is very interesting by its sharp double-peak structure in Figs. 1(d) and S4(d) because the border between metal and non-metal of this quantity is ~ 2.0 . Therefore, the first high peaks in Figs. 1(d) and S4(d) are from metallic compounds, while the second peaks are from non-metallic materials. The ML predicted high- T_C compounds of this work, somehow, are mostly from the non-metal side, whose average electronegativity is beyond ~ 2.2 [see Fig. S4(d)], suggesting that there is still significant room for new high- T_C materials to be discovered in non-metallic compounds. While beyond the scope of this work, it will be quite interesting to use accurate high-level electronic structure calculations to directly substantiate the ML predicted compounds in Table I.

TABLE II. Relative importance of each descriptor in the ML model of superconductors calculated by random forest regression. The labels are: OPES stands for the occurrence possibility of elements in ML, \bar{M} stands for the average atomic masses, \bar{N} the average number of electrons in an unfilled subshell, $\bar{\mu}$ the average atomic magnetic moment, $\bar{\eta}$ the average atomic electronegativity, ΔM the maximum difference of atomic masses, and $\delta\eta$ the maximum difference of electronegativity.

OPES	\bar{M} (%)	\bar{N} (%)	$\bar{\mu}$ (%)	$\bar{\eta}$ (%)	ΔM (%)	$\delta\eta$ (%)
Yes	6.5	3.5	12.5	2.0	3.6	30.7
No	11.3	6.0	16.6	3.2	6.8	56.1

V. SUMMARY

We have carried out a materials informatics investigation to screen new layered high- T_C superconductor candidates from layer materials in the ICSD: 25 are predicted, which are not known before to be superconductors, including 12 cuprates, 7 iron-based crystals, and 6 others, whose T_C values are predicted to range from 32 K to the impressive 138 K at atmospheric pressure.

Our MI + ML approach emphasizes three important ingredients:²⁸ (i) a set of simple but effective material descriptors, (ii) a multi-step learning and multi-algorithm cross-verification machine learning strategy to build superconductor models, and (iii) application of the models to the experimental structure database, so that the predictions are realistic materials (as opposed to hypothetical materials). Regarding (i), the simple but effective descriptor set is summarized in Table II. Our ML algorithms picked candidates from the structure database using narrow ranges of the descriptors, learned from our property database. Regarding (ii), we find that the ML models are better optimized using the multi-step learning and multi-ML algorithm cross-verification. For known high T_C superconductors, such an optimized ML model predicts excellent T_C values with over 92% confidence. Regarding (iii), we applied the ML model to about 2500 layered materials in the ICSD and discovered 25 new layered superconductor candidates, whose critical temperatures have not yet been experimentally measured. We substantiate their likelihood using DOSs calculated by DFT and find them consistent with known DOS phenomenology.¹² The new layered candidates are further substantiated by a heuristic formula,¹⁶ which leads T_C estimates to be largely consistent with the ML predictions.

In principle, to predict the superconducting phase transition temperature T_C of a material, one should already know that the ground state of that material is a superconducting state. This is, however, extremely difficult to determine theoretically. In this work, we build a machine learning model of T_C by the experimentally measured T_C in the SuperCon database, and after the model is accurately tested using experimentally known materials, we apply it to the ICSD to search for possible new candidates. Indeed, the purpose of the data-driven approach, such as that presented in this work, is to screen material candidates, which may have certain physical properties, thereby narrowing down the phase space for further experimental and/or theoretical investigations.

ACKNOWLEDGMENTS

H.G. acknowledges the financial support from the Natural Science and Engineering Research Council of Canada (NSERC) and Fonds de recherche du Québec Nature et technologies (FRQNT). We thank Compute Canada for computation facilities. Z.L.L. acknowledges support from the National Natural Science Foundation of China (Grant No. 41574076) and the Key Research Scheme of Henan Universities (Grant No. 18A140024).

DATA AVAILABILITY

The data that support the findings of this study are openly available at https://github.com/zhongliliiu/imat_data.

REFERENCES

- H. K. Onnes, "The resistance of pure mercury at helium temperatures," *Commun. Phys. Lab. Univ. Leiden*, **12**, 120 (1911).
- A. Schilling, M. Cantoni, J. D. Guo, and H. R. Ott, "Superconductivity above 130 K in the Hg–Ba–Ca–Cu–O system," *Nature* **363**, 56 (1993).
- Y. Kamihara, H. Hiramatsu, M. Hirano, R. Kawamura, H. Yanagi, T. Kamiya, and H. Hosono, "Iron-based layered superconductor: LaOFeP," *J. Am. Chem. Soc.* **128**, 10012 (2006).
- G. Wu, Y. L. Xie, H. Chen, M. Zhong, R. H. Liu, B. C. Shi, Q. J. Li, X. F. Wang, T. Wu, Y. J. Yan, J. J. Ying, and X. H. Chen, "Superconductivity at 56 K in samarium-doped SrFeAsF," *J. Phys.: Condens. Matter* **21**, 142203 (2009).
- A. P. Drozdov, M. I. Erements, I. A. Troyan, V. Ksenofontov, and S. I. Shylin, "Conventional superconductivity at 203 kelvin at high pressures in the sulfur hydride system," *Nature* **525**, 73 (2015).
- M. Somayazulu, M. Ahart, A. K. Mishra, Z. M. Geballe, M. Baldini, Y. Meng, V. V. Struzhkin, and R. J. Hemley, "Evidence for superconductivity above 260 K in lanthanum superhydride at megabar pressures," *Phys. Rev. Lett.* **122**, 027001 (2019).
- A. P. Drozdov, P. P. Kong, V. S. Minkov, S. P. Besedin, M. A. Kuzovnikov, S. Mozaffari, L. Balicas, F. F. Balakirev, D. E. Graf, V. B. Prakapenka, E. Greenberg, D. A. Knyazev, M. Tkacz, and M. I. Erements, *Nature* **569**, 528 (2019).
- L. Ward, S. C. O'Keeffe, J. Stevick, G. R. Jelbert, M. Aykol, and C. Wolverton, "A machine learning approach for engineering bulk metallic glass alloys," *Acta Mater.* **159**, 102 (2018).
- L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," *npj Comput. Mater.* **2**, 16028 (2016).
- A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, "High-throughput machine-learning-driven synthesis of full-Heusler compounds," *Chem. Mater.* **28**, 7324 (2016).
- A. Mansouri Tehrani, A. O. Oliynyk, M. Parry, Z. Rizvi, S. Couper, F. Lin, L. Miyagi, T. D. Sparks, and J. Brgoch, "Machine learning directed search for ultraincompressible, superhard materials," *J. Am. Chem. Soc.* **140**(31), 9844 (2018).
- V. Stanev, C. Oses, A. G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, and I. Takeuchi, "Machine learning modeling of superconducting critical temperature," *npj Comput. Mater.* **4**, 29 (2018).
- B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, and L. Ward, "Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery," *Mol. Syst. Des. Eng.* **3**, 819 (2018).
- M. Klintonberg and O. Eriksson, "Possible high-temperature superconductors predicted from electronic structure and data-filtering algorithms," *Comput. Mater. Sci.* **67**, 282 (2013).
- H.-R. Zhang, Y. Zhang, D.-B. Dai, M. Cao, and W.-F. Shen, "Modelling and optimization of the superconducting transition temperature," *Mater. Des.* **92**, 371 (2016).
- D. R. Harshman, A. T. Fiory, and J. D. Dow, "Theory of high- T_C superconductivity: Transition temperature," *J. Phys.: Condens. Matter* **23**, 295701 (2011).
- See <https://supercon.nims.go.jp/index.html> for superconducting material database (SuperCon).
- In the purifying process of the property database, we eliminated "suspicious" records in SuperCon such as the ones having $T_C > 210$ K at atmosphere pressure. For replicated compounds having exactly the same chemical components, we retain the one with the highest T_C . We also kept compounds with dilute impurities since the descriptors do not change significantly and the data reflect how T_C varies with components. In purifying the structure database, we eliminate records having fractional occupation numbers of atomic sites, because such disorder appears to enhance electrical resistance and hinder superconductivity; Y. Dubi, Y. Meir, and Y. Avishai, "Nature of the superconductor to insulator transition in disordered superconductors," *Nature* **449**, 876 (2007).

- ¹⁹A. Kuzemsky and I. Kuzemskaya, "Structural sensitivity of superconducting properties of layered systems," *Physica C* **383**, 140 (2002).
- ²⁰T. A. Zaleski and T. K. Kopeć, "Dependence of the superconducting critical temperature on the number of layers in a homologous series of high- T_C cuprates," *Phys. Rev. B* **71**, 014519 (2005).
- ²¹Y. Zhu, X. Kong, T. D. Rhone, and H. Guo, "Systematic search for two-dimensional ferromagnetic materials," *Phys. Rev. Mater.* **2**, 081001 (2018).
- ²²Z.-L. Liu, "Muse: Multi-algorithm collaborative crystal structure prediction," *Comput. Phys. Commun.* **185**, 1893 (2014).
- ²³L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Phys. Rev. Lett.* **114**, 105503 (2015).
- ²⁴N. W. Ashcroft, "Hydrogen dominant metallic alloys: High temperature superconductors?," *Phys. Rev. Lett.* **92**, 187002 (2004).
- ²⁵L. Breiman, "Random forests," *Mach. Learn.* **45**, 5 (2001).
- ²⁶V. Vapnik, *The Nature of Statistical Learning Theory* (Springer-Verlag, New York, 1995).
- ²⁷H. T. Siegelmann and E. D. Sontag, "Analog computation via neural networks," *Theor. Comput. Sci.* **131**, 331 (2011).
- ²⁸The MI + ML algorithms, descriptors, purifying methods, and learning procedures used in this work are packaged in our intelligence based materials discovery platform *iMAT*. *iMAT* has many other features including automatic tests of parameters for different ML algorithms and automatic generation and optimization for combinations of descriptors. Z.-L. Liu, P. Kang, and H. Guo, "*iMAT*: A package for intelligent-design of materials" (unpublished) (2018).
- ²⁹J. Labbé, S. Barišić, and J. Friedel, "Strong-coupling superconductivity in V_3X type of compounds," *Phys. Rev. Lett.* **19**, 1039 (1967).
- ³⁰J. E. Hirsch and D. J. Scalapino, "Enhanced superconductivity in quasi two-dimensional systems," *Phys. Rev. Lett.* **56**, 2732 (1986).
- ³¹J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD)," *JOM* **65**, 1501 (2013).
- ³²S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Ruhl, and C. Wolverton, "The open quantum materials database (OQMD): Assessing the accuracy of dft formation energies," *npj Comput. Mater.* **1**, 15010 (2013).
- ³³P. Monthoux, A. Balatsky, and D. Pines, "Weak-coupling theory of high-temperature superconductivity in the antiferromagnetically correlated copper oxides," *Phys. Rev. B* **46**, 14803 (1992).
- ³⁴S. Chakravarty, A. Sudbo, P. W. Anderson, and S. Strong, "Interlayer tunneling and gap anisotropy in high-temperature superconductors," *Science* **261**, 337 (1993).