Physica C: Superconductivity and its applications 575 (2020) 1353689

Contents lists available at ScienceDirect

Physica C: Superconductivity and its applications

journal homepage: www.elsevier.com/locate/physc

Predicting new superconductors and their critical temperatures using machine learning

B. Roter, S.V. Dordevic*

Department of Physics, The University of Akron, Akron, OH 44325, USA

ARTICLE INFO

Keywords:
Machine learning
High temperature superconductors

ABSTRACT

We used the superconductors in the SuperCon database to construct element vectors and then perform machine learning of their critical temperatures (T_c). Only the chemical composition of superconductors was used in this procedure. No physical predictors (neither experimental nor computational) of any kind were used. We achieved the coefficient of determination $R^2 \simeq 0.93$, which is comparable and in some cases higher than similar estimates using other artificial intelligence techniques. Based on this machine learning model, we predicted several new superconductors with high critical temperatures. We also discuss several factors that limit the learning process and suggest possible ways to overcome them.

1. Introduction

Quantum supremacy was recently achieved by Google using a superconducting microprocessor Sycamore [1]. An avalanche of similar results is now expected. The future of superconductors has never looked brighter. However, if Sycamore and other superconducting microprocessors are to find a wider circle of users, their operating temperature will have to be increased significantly. Sycamore is made of aluminum (T_c = 1.175 K) and indium (T_c = 3.41 K) and operates at temperatures below 20 mK [1]. Such low temperatures require a dilution refrigerator with ^{3}He, which is exceedingly rare and expensive. This clearly illustrates the need for new superconducting materials with higher critical temperatures. However, finding new superconductors, especially with high T_c, is a very difficult endeavor [2,3].

In recent years, there has been a surge of interest in using artificial intelligence (AI), in particular machine learning (ML) and deep learning (DL), in materials physics [4–6]. The idea is that by using the existing information in materials' databases, one can predict new materials with certain desired properties. In particular, several attempts have been made in predicting the critical temperatures of superconductors, or more generally, predicting new materials with potentially high T_c. Several prominent efforts have been by Stanev et.al [7], Hamidieh [8], Konno et.al [9], Matsumoto et.al [10] and Zeng et al. [11]. Different AI approaches were used in these papers: Stanev et al. used both classification and regression models, Hamidieh used an XGBoosted statistical model, Konno et.al used deep learning, Matsumoto et.al used a machine learning algorithm and Zeng et al. used convolutional neural networks (CNN). Materials informatics was also used to predict new hydride superconductors [3,12].

In a recent work by Zhou et.al [13], the properties of the atoms were learned from the chemical compositions of compounds from a large database, without any additional information. Inspired by this approach, we made a similar attempt in predicting new superconducting materials and their critical temperatures. The *only* predictor used is the chemical composition of compounds (both superconducting and non-superconducting), which is readily available in the existing databases and does not require any post-processing. We employed a combination of both unsupervised and supervised machine learning and achieved statistical parameters comparable, and in some instances exceeding previous attempts. Below we describe in details the procedure used, and then the results of our study. We also discuss the factors that limit the learning process, most notably the wrong entries into the database.

2. SuperCon database

SuperCon is currently the biggest and most comprehensive database of superconductors in the world [14]. It is free and open to the public, and it has been used in almost all AI studies of superconductors [7,8,10,11,9]. At the time when we downloaded it, it contained almost 34,000 entries. Fewer than 100 of them had errors in their chemical formulas and were removed. About 7000 entries did not have the values of T_c reported and were also removed. The remaining 27,000 were used for our ML calculations. However, fewer than 100 entries had T_c = 0. The importance of non-superconducting compounds for AI calculations

* Corresponding author.
E-mail address: dsasa@uakron.edu (S.V. Dordevic).

https://doi.org/10.1016/j.physc.2020.1353689
Received 21 April 2020; Received in revised form 22 May 2020; Accepted 25 May 2020
Available online 27 May 2020

was noted previously [11]. In order to improve the predictive power of ML models, we supplemented SuperCon database with about 3000 non-superconducting compounds, mostly insulators, semiconductors and some non-superconducting metals and alloys. In total, our database had about 30,000 entries. There were several thousands of multiple entries which were all kept for the reasons discussed below. The entire database was used as the training set for both classification and regression models discussed in Sections 5 and 6 below.

3. Element-vectors

Once the database was created, we parsed all the formulas and wrote the chemical content for each superconductor into a matrix which we call the chemical composition matrix. The corresponding values of T_c were written in a separate vector. The matrix has about 30,000 rows and 96 columns. The number of columns is determined by the elements present in the chemical formulas. The heaviest element that appears in any superconductor in the database is Curium ($Z = 96$) and consequently the chemical composition matrix has 96 columns. We also note that the matrix is extremely sparse, as more than 96% of its elements are zeros.

In Fig. 1 we show a very small portion (upper left corner) of the chemical composition matrix. By analogy with Ref. [13] we call the columns of the matrix element-vectors (or atom-vectors). They contain the information about the superconductors in the database and can be used as predictors in training the models. However, we show below that by using unsupervised machine learning the accuracy of the models can be improved.

4. Unsupervised machine learning

In this section we describe what is usually referred to as unsupervised machine learning. At the heart of this approach is the so-called Singular Value Decomposition (SVD), or equivalently Principle Component Analysis (PCA) [15]. Once the chemical composition matrix is formed as described above (Fig. 1), it is decomposed according to [15]

$$X = USV^T \tag{1}$$

where U and V are unitary matrices, and S is a diagonal matrix whose values are called singular values. The columns of matrix U will be referred as element-eigenvectors (or atom-eigenvectors) by analogy with Ref. [16]. These element-eigenvectors contain more abstract information about the chemical composition of superconductors, and were used as the *only* predictors by ML models. The rank of the chemical composition matrix described in the previous section (Fig. 1) is 83, which indicates that one can use any number of element-eigenvectors up to 83. Our calculations indicate that with as few as 10 element-eigenvectors one can achieve significant improvements over the calculations using raw element-vectors described in the previous section.

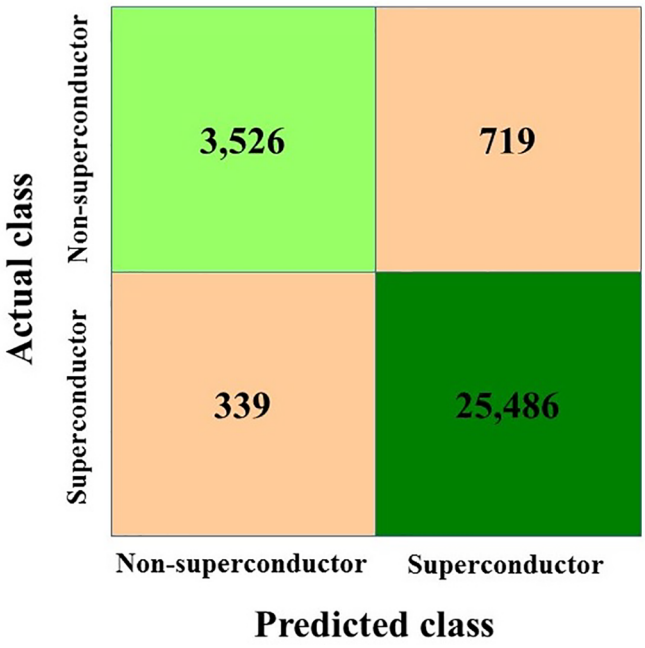

Fig. 2. (Color online). A confusion matrix for the classification model. The overall accuracy is 96.5%.

5. Classification models

To check the effectiveness of the procedure described above, we first constructed classification models using only element-eigenvectors as predictors. These classification models were designed for predicting whether a compound is a superconductor or not. A number of different training algorithms were tested, such as the Bagged Tree, Boosted Tree and Gaussian Support Vector Machines. After a number of tests, we concluded that the method called k-Nearest Neighbors (KNN) was the most accurate for this particular task.

KNN is designed to exploit the ideas of similarity and mathematical distances. Given a set of points and a distance function, KNN allows one to find the k closest points in that set to either a point or collection of points of interest, irrespective of any labeling. The latter group of points represent classes, and those classes, in our case, represent whether a compound with a given chemical composition is superconducting or not.

Applying the KNN method to our dimension-reduced chemical composition matrix U (Eq. (1)), we were able to create a classification model that was 96.5% accurate. This value exceeds those obtained by other AI techniques [7,11]. Fig. 2 shows the corresponding confusion matrix, which illustrates the need for including even more non-superconducting entries into the database.

	H	He	Li	Be	B	C	O	N	⋯
$H_{0.04}Ta_{0.96}$	0.04								
$H_{0.3}LiNbO_2$	0.3		1				2		
$Be_{13}Ru$				13					
$C_6Ir_4O_2$						6	2		
$N_{0.99}Ti$								0.99	
$BCMo_2$					1	1			
⋮									⋱

Fig. 1. (Color online). Upper left corner of the chemical composition matrix. The size of the whole matrix is (approximately) 30,000 × 96: (approximately) 30,000 entries in the database and 96 elements. The columns of the matrix represent element-vectors that we used as the *only* predictors in our calculations.

6. Regression models

One can construct the regression models using the raw element-vectors described above in Section 3 (Fig. 1). The values of statistical parameters [17] achieved using these element-vectors were: the coefficient of determination $R^2 \simeq 0.90$ and the root-mean-square error RMSE $\simeq$ 9.67 K. As good as these numbers are, they can be further improved with the help of SVD (Eq. (1)). Using element-eigenvectors as the only predictors we constructed our regression models. A number of different training methods were tested, such as Exponential Gaussian Process Elimination, Fine Tree, Boosted Tree, as well as a Gaussian Support Vector Machine (SVM). An algorithm known as the Bagged Tree was found to be the most accurate in predicting the values of T_c for our input compounds.

The Bagged Tree method is a variant of Random Forests algorithms [20]. It combines multiple decision trees in order to output better predictions - a stark contrast from just creating one decision tree. The underlying theory for this technique is that multiple weak learners should be able to combine into a much more robust form. Typically, with ML algorithms involving decision trees, altering the training data in any way can yield completely different trees, thus yielding completely different predictions. The method of bagging is designed to help greatly mitigate the high-variance nature of these decision trees.

Applying the Bagged Tree algorithm to our dimension-reduced chemical composition matrix, we were able to create a regression model that achieved $R^2 \simeq 0.93$ and RMSE $\simeq$ 8.91 K. These values of R^2 and RMSE are comparable and in some cases higher than the values achieved using other AI methods. In Fig. 3 below we display the results of our calculations. The values of predicted T_c are plotted as a function of actual T_c for all 30,000 compounds in the SuperCon database. We notice that plot does not reveal any systematic offsets, and vast majority of the data points are clustered around a line with unit slope (red line) in the $\pm$ 8.91 K range (green lines). The outliers that appear in Fig. 3 could be due to incorrect values of T_c reported in SuperCon, and are discussed in more detail in Section 8 below.

7. Predictions

Using our best model (Fig. 3) we also made predictions of new superconducting materials. For that purpose we downloaded the entire Crystallography Open Database (COD) [21], which contains about 37,000 inorganic compounds and alloys from which we made predictions. Not surprisingly, most compounds predicted to have a high T_c were oxides [7]. Some of the them are listed in Table 1. Interestingly, our calculations also predicted a number of non-oxide and non-iron based materials with T_c in the range 40–60 K. Some of the most promising examples are also listed in Table 1 [22].

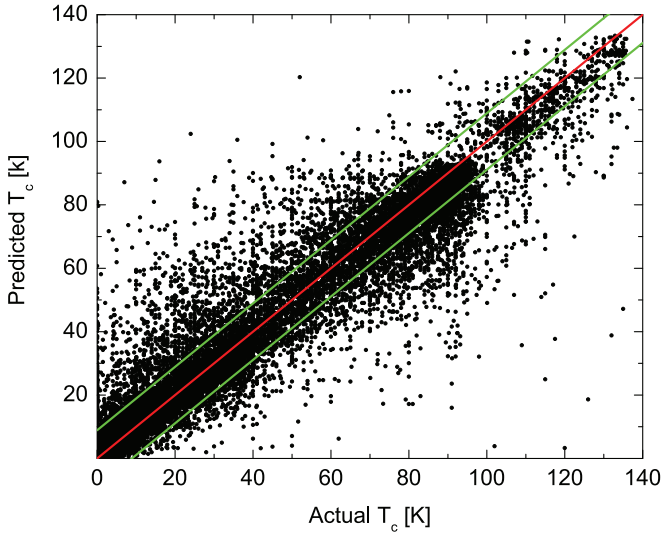

Fig. 3. (Color online). A plot of predicted T_c versus actual T_c. The achieved statistical parameters are $R^2 \simeq 0.93$ and RMSE $\simeq$ 8.91 K.

Table 1
Compounds from COD database [21] predicted to be superconducting [22]. In addition to a number of oxide materials, our models also predicted several non-oxide materials with relatively high T_c.

Compound	Predicted T_c (K)	Comments
$AlBaCaF_7$	46	
As_4BaCu_8	50	
$BaCu_4S_3$	31	
$CrCuSe_2$	26	
LiRbS	21	
AlB_4Cr_3	50	
$AlBa_3P_3$	38	
$BaCuTe_2O_7$	54	see Ref. [11]
$BaCu_3Br_2O_4$	60	see Ref. [11]
$Ba_5Br_2Ru_2O_9$	57	see Ref. [7]
$CaCu_2Eu_2O_6$	65	
Cr_2CuO_4	50	
$Cu_3Na_7O_8$	67	
$Cl_2Sr_2CuO_2$	27	

We also included in this analysis the materials previously predicted to be superconducting by Stanev et al. [7] (without T_c) and Zeng et al. [11]. We find some of them to be superconducting with almost the same T_c. For example, Zeng et al. [11] predicted $BaCu_3Br_2O_4$ to be superconducting with T_c= 60 K, which is the same T_c that we predicted. On the other hand, Zeng et al. found $Na_3(TiS_2)_{10}$ to be superconducting with T_c = 40.71 K, which we found to be non-superconducting. This illustrates intrinsic fragileness of AI methods.

8. Limiting factors

As the most important limiting factor of ML we identified the wrong entries into the database. To illustrate that point in Fig. 4 we display the doping (x) dependence of T_c for all $La_{2-x}Sr_xCuO_4$ (LSCO) superconductors in the SuperCon database. There are a total of 550 entries, which are represented by red points. The green curve represents the

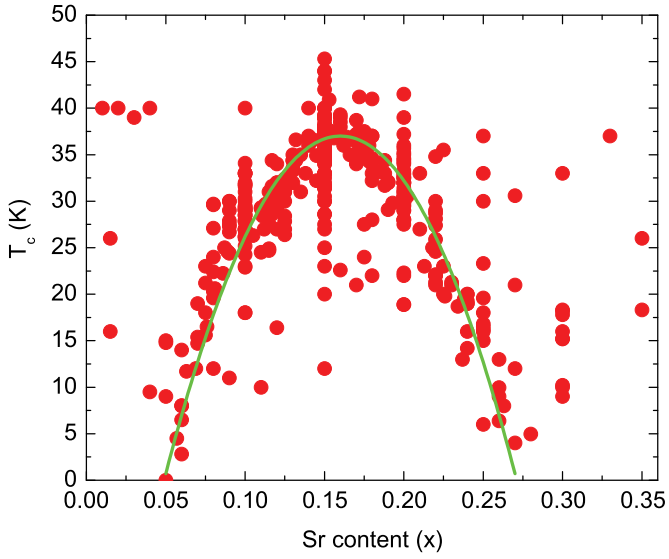

Fig. 4. (Color online). A plot of T_c versus strontium doping (x) for all 550 $La_{2-x}Sr_xCuO_4$ superconductors from the SuperCon database. Green line is the expected $T_c(x)$ behavior for this family of cuprates [23].(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

expected doping dependence of T_c for this family of cuprate superconductors [23]. As can be seen from the plot, there are a number of outliers. In addition, for many doping levels, i.e. for the same chemical composition, there are a large number of points located either below or above the expected value [24]. Similar wrong entries were found in other cuprate families, as well as in iron-based superconductors. We estimate that there could be as many as 20% of wrong entries in the entire database. Our calculations reveal that when they are removed from the database, the values of the statistical parameters can be further improved.

9. Summary

In summary, we developed and implemented a simple, yet powerful machine learning method to make predictions of new compounds that are possibly superconducting and their critical temperatures T_c. Using *only* the chemical composition of superconductors from the SuperCon database we created a number of models and achieved $R^2 \simeq 0.93$ and RMSE $\simeq$ 8.91 K. These statistical parameters are comparable and in some cases exceed those obtained with other AI studies that used a number of physical predictors (both experimental and computational).

Our results indicate that one does not need predictors such as the number of valence electrons, electronegativity, covalent radius, electron affinity or the number of unfilled orbits to achieve significant predictive power. We argue that those predictors are not directly relevant for superconductivity and that is the reason they did not lead to any significant improvements of statistical parameters (R^2 and RMSE). We suggest that physical predictors more closely related to superconductivity should be used [25]. Those include, for example, normal state resistivity (or conductivity), superfluid density (or penetration depth), band structure features, Fermi energy, specific heat, etc. They have been shown to be closely related to T_c [26,27,29,28] and, in our opinion, would lead to improved models. Unfortunately there is currently no comprehensive database that contains the values of these parameters for a large number of superconductors. SuperCon does have some of them, such as the penetration depth, critical fields, energy gap, etc. However, the number of entries with these values reported is too small for any meaningful AI calculations. For example, the values of the penetration depth or the energy gap are reported for fewer than 1200 superconductors in the SuperCon.

We also showed that the limiting factor for achieving higher predicting power is the quality of entries in the SuperCon database. If the number of wrong entries can be reduced in the future, for example by human curation, then the predictive power of ML models will inevitably be improved.

Using our best models, we also made predictions of new superconductors and their T_c's. Running the models on inorganic compounds from the COD database we have identified a number of materials that are potentially superconducting, with relatively high T_c. Future transport and/or thermodynamic measurements will determine how accurate any of these AI predictions [7,8,11] are.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] F. Arute, et al., Nature 574 (2019) 505.
[2] S. Uchida, High Temperature Superconductivity: The Road to Higher Critical Temperature, Springer, Japan, 2015.
[3] A.P. Drozdov, et al., Nature 569 (2019) 528. Superconductivity with record high critical temperatures was recently discovered in hydrates. Unfortunately they only superconduct when exposed to extreme pressures
[4] O. Isayev, D. Fourches, E.N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, Chem. Mater. 27 (2015) 735.
[5] J. Hill, G. Mulholland, K. Persson, R. Seshadri, C. Wolverton, B. Meredig, MRS Bull. 41 (2016) 399.
[6] J. Yuan, V. Stanev, C. Gao, I. Takeuchi, K. Jin, Supercond. Sci. Technol. 32 (2019) 123001.
[7] V. Stanev, C. Oses, A.G. Kusne, E. Rodriguez, J. Paglione, S. Curtarolo, I. Takeuchi, NPJ Comput. Mater. 4 (2018) 29.
[8] K. Hamidieh, Comput. Mater. Sci 154 (2018) 346.
[9] T. Konno, H. Kurokawa, F. Nabeshima, Y. Sakishita, R. Ogawa, I. Hosako, A. Maeda, arXiv:1812.01995.
[10] K. Matsumoto, T. Horide, Appl. Phys Express 12 (2019) 073003.
[11] S. Zeng, Y. Zhao, G. Li, R. Wang, X. Wang, J. Ni, NPJ Comput. Mater. 5 (2019) 84.
[12] T. Ishikawa, T. Miyake, K. Shimizu, Phys. Rev. B 100 (2019) 174506.
[13] Q. Zhou, P. Tang, S. Liu, J. Pan, Q. Yan, S.C. Zhang, Proc. Natl. Acad. Sci. 115 (2018) E6411.
[14] SuperCon database: https://supercon.nims.go.jp.
[15] M.E. Wall, A. Rechtsteiner, L.M. Rocha, D.P. Berrar, W. Dubitzky, M. Granzow, A Practical Approach to Microarray Data Analysis, Kluwer, Norwell, MA, 2003, pp. 91–109.
[16] O. Alter, P.O. Brown, D. Botstein, Proc. Natl. Acad. Sci. 97 (2000) 10101.
[17] R^2 and RMSE are conventionally defined [18,19] as $R^2 = 1$ - SSE/SST (SSE is the sum of squared error and SST is the sum of squared total) and RMSE $= \sqrt{\sum_{i=1}^{N} (T_{c,predicted} - T_{c,actual})^2/N}$.
[18] C.M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, New York, 2006.
[19] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer, New York, 2001.
[20] L. Breiman, Mach. Learn. 45 (2001) 5.
[21] Crystallography Open Database: https://www.crystallography.net.
[22] Full list of predictions is available upon request.
[23] J.L. Tallon, J.W. Loram, Phys. C 349 (2001) 53.
[24] S. Yomo, et al., Jpn. J. Appl. Phys. 26 (1987) L603. Some of these points which deviate from the expected behavior are for samples whose T_c was altered by application of pressure, such as in However, this information is not systematically reported in the SuperCon database
[25] J.E. Hirsch, Phys. Rev. B 55 (1997) 9007.
[26] C.C. Homes, S.V. Dordevic, M. Strongin, D.A. Bonn, R. Liang, W.N. Hardy, S. Komiya, Y. Ando, G. Yu, N. Kaneko, X. Zhao, M. Greven, D.N. Basov, T. Timusk, Nature 430 (2004) 539.
[27] S.V. Dordevic, D.N. Basov, C.C. Homes, Sci. Rep. 3 (2013) 1713.
[28] Y. Liu, N. Chen, Y. Li, arXiv:1811.12171.
[29] Y.J. Uemura, Phys. Rev. Mater. 3 (2019) 104801.