Upper limit of the transition temperature of superconducting materials

Graphical abstract

In brief
The variation of transition temperature ($T_c$) in diverse superconducting materials can provide key clues for the understanding of superconductivity. However, it has long been challenging to find hidden key factors affecting $T_c$. In this work, with the help of machine learning and first-principles calculations, a parameter governing the $T_c$ upper limit of each kind of superconducting materials was discovered. It implies a close relevance between superconductivity and energy-level distribution of valence electrons.
Upper limit of the transition temperature of superconducting materials

Yang Liu,1,2,3 Haiyou Huang,1,2,7,* Jie Yuan,3 Yan Zhang,1,2 Hongyuan Feng,1,2 Ning Chen,4 Yang Li,5 Jiao Teng,4 Kui Jin,3,* Dezhen Xue,6 and Yanjing Su1,2

1Beijing Advanced Innovation Center for Materials Genome Engineering, University of Science and Technology Beijing, Beijing 100083, China
2Institute for Advanced Materials and Technology, University of Science and Technology Beijing, Beijing 100083, China
3Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China
4School of Materials Science and Engineering, University of Science and Technology Beijing, Beijing 100083, China
5Institute of Physics, Chinese Academy of Sciences, Beijing 100190, China
6Department of Engineering Science and Materials, University of Puerto Rico, Mayaguez, PR 00681-9000, USA
7State Key Laboratory for Mechanical Behavior of Materials, Xi’an Jiaotong University, Xi’an 710049, China
7Lead contact
*Correspondence: huanghy@mater.ustb.edu.cn (H.H.), kuijin@iphy.ac.cn (K.J.)

SUMMARY

Why are the transition temperatures ($T_c$) of superconducting materials so different? The answer to this question is not only of great significance in revealing the mechanism of high-$T_c$ superconductivity but also can be used as a guide for the design of new superconductors. However, so far, it is still challenging to identify the governing factors affecting the $T_c$. In this work, with the aid of machine learning and first-principles calculations, we found a close correlation between the upper limit of the $T_c$ and the energy-level distribution of valence electrons. It implies that some additional inter-orbital electron-electron interaction should be considered in the interpretation of high-$T_c$ superconductivity.

INTRODUCTION

Predicting the critical temperature ($T_c$) of superconductors has long been a great challenge.1-3 Although a number of correlations between $T_c$ and various parameters have already been discovered,4-21 it is still a tough job to find new superconductors with higher $T_c$.

Recently, the utilization of machine learning (ML) techniques inspired high hopes.25-27 ML models using different algorithms were trained to predict the existence of superconductivity and the $T_c$ of superconductors.28-54 Progress has been made in several areas, such as how $T_c$ varies with doping.28-35 the descriptors indicating superconducting mechanism,36-39 structural factors affecting $T_c$,43,44 and candidates of new high-$T_c$ superconductors.46,51 So far, ML models predicting $T_c$ have yielded good predictive scores.78,43-46 The feature variables in those models also indicated some governing factors for high $T_c$, such as “flat band,” “atomic mass,” and “number of d valence electrons,” which are in good agreement with the human experience.43,44

However, it is too early to be optimistic because the ML studies have not yet given any successful (experimentally...
verified) prediction on new high-$T_c$ superconductors, as well as any theoretical inductions beyond human expertise. Even more, the state-of-the-art models by far gave inconsistent results when proposing candidates of high-$T_c$ superconductors. In general, ML is a powerful tool for exploring patterns in the variation of $T_c$, but ML models need better generalization ability and interpretability to get trustable and useful results.

In this work, we were looking for an approach to reveal key factors affecting the $T_c$ of superconductors. ML and first-principles calculations were exploited to investigate the connection between the $T_c$ and band structure of electrons. Considering the practical need in materials science, we took the $T_c$ maximum ($T_{c\,\text{max}}$) of each kind of superconducting materials, instead of the $T_c$ of each chemical composition, as the target variable in the ML. The feature variables in the ML are all derived from the electronic orbital characteristics with explicit physical meaning, which is essential for the interpretability of ML model. Enlightened by the ML results, we found that high-$T_c$ superconductivity has a close connection with the energy-level distribution of valence electrons. Then, based on high-throughput calculations of band structure, a correlation between $T_{c\,\text{max}}$ and a band structure parameter in diverse materials was discovered. It can provide convenient guidance for the design of new superconducting materials and a clue to the pairing mechanism of high-$T_c$ superconductivity.

**RESULTS**

**Goal and target variable in machine learning**

First of all, the target variable in our ML study is the $T_{c\,\text{max}}$ of each type of superconducting materials instead of the $T_c$ of each chemical composition. Here is a domain knowledge that for each type of superconducting materials, there is a maximum of $T_c$, no matter how the $T_c$ varies with doping, pressure, or gate voltage.

In practice, the $T_{c\,\text{max}}$ is often called “$T_c$” for convenience. But, in fact, the $T_c$ and the $T_{c\,\text{max}}$ are of distinct physical meanings and practical application scenes. The key factors affecting $T_c$ and the key factors affecting $T_{c\,\text{max}}$ are also quite different (see more details in Table S1).

$T_c$ is the transition temperature of superconductivity for a specimen at a particular chemical composition. It is a physical parameter. Each value of $T_c$ is obtained through one single measurement operation. Particularly, for unconventional superconductors, the variation of $T_c$ with doping or pressure is known as the “phase diagram” of cuprate, iron-based superconductors, or heavy-Fermion superconductors. Physicists are keen to investigate the phase diagrams so as to get a deeper insight into the mechanism of high-$T_c$ superconductivity.

On the other hand, $T_{c\,\text{max}}$ is the $T_c$ maximum of each type of superconducting materials. It is a value of material performance. Each type of superconducting material has one $T_{c\,\text{max}}$ value. Particularly, in a phase diagram of an unconventional superconductor, the optimal $T_c$ is the $T_{c\,\text{max}}$. Driven by the desire to find room temperature superconductors, materials scientists are more interested in the variation of $T_{c\,\text{max}}$.

In fact, for the issue of $T_c$, there are two different but equally important questions: (1) how to understand the phase diagram of unconventional superconductors. Or, in other words, for one type of materials, how the $T_c$ varies with doping, pressure, gate voltage, or other conditions. (2) How the $T_{c\,\text{max}}$ varies in different types of materials. During the past decades, there have been plenty of experimental and theoretical achievements on the first question\textsuperscript{55} but much less on the second one. In this work, we focus on question two, searching for unknown patterns in the data of $T_{c\,\text{max}}$. Accordingly, the $T_{c\,\text{max}}$ is taken as the target variable in our ML investigation.

**Dataset and data distribution in machine learning**

The dataset of $T_{c\,\text{max}}$ is a subset of the dataset of $T_c$, which was derived from the SuperCon database of NIMS.\textsuperscript{56} For the dataset-$T_c$, the entries with doubtful $T_c$ values, obvious typing errors in the chemical formula, and $T_c = 0$ K were all removed. A few well-known or newfound superconductors\textsuperscript{57,58} were added, but the superconductors at ultrahigh pressure (e.g., H$_2$S, LaH$_2$) are not included. Thus, dataset-$T_c$ has 12,196 entries.

To get the dataset-$T_{c\,\text{max}}$, superconductors in dataset-$T_c$ were categorized into 1,000+ groups, according to the chemical composition (please see more details experimental procedures). In each group, the superconductor having the highest $T_c$ (regarded as the $T_{c\,\text{max}}$ of a type of materials) was picked out. The dataset-$T_{c\,\text{max}}$ has 1,008 entries.

Twenty-four most-concerned superconductors (see Table S2) were chosen as the test data (unseen during model training) to assess the models’ generalization ability. The close neighbors of those 24 superconductors were taken out from the dataset-$T_{c\,\text{max}}$. Here, “close neighbor” means a small Manhattan distance ($<1.5$) in the chemical composition (elemental contents) space. At last, a subset of $T_{c\,\text{max}}$ was used to train the models. The subset-$T_{c\,\text{max}}$ has 957 entries.

It should be stressed that the data distribution is vital for the model’s performance. For the ML models predicting superconductors, their generalization ability is often in suspense because of imbalanced data distribution, including data bias, dataset shift, and clustered data distribution. Both the data bias and dataset shift are due to the fact that the explored superconductors are only a small proportion of all explored materials. Specifically, there are about 200,000 entries in ICSD,\textsuperscript{59} belonging to 70,000+ materials, while a typical dataset of $T_c$ derived from SuperCon\textsuperscript{56} usually has 10,000+ entries after data washing, belonging to about 3,000 materials. For the dataset-$T_c$, some elements (e.g., Cu, O, Fe, As, Se, H) dominate in high-$T_c$ superconductors, leading to a significant data bias. Meanwhile, when predicting new materials, the chemical composition of each candidate (data to be predicted) would be quite different from those of explored superconductors (train and test data), which is a significant dataset shift. Meanwhile, in the dataset-$T_c$, many data are of the same material with different doping contents. The chemical composition of those data are close, resulting in a clustered data distribution. Imbalanced data distribution often causes overfitting. There is a good chance to get a model having a high predictive score on the test data, but the model may be no longer good when predicting new materials. It is hard to get a reliable and trustable ML model until the data distribution is improved.

In this work, the data distribution of the dataset-$T_c$ and the subset-$T_{c\,\text{max}}$ are quite different. It can be seen in Figure 1 that the unconventional superconductors (only three materials...
families, i.e., cuprates, iron-based, and heavy-Fermion superconductors) account for nearly 52% of the total, and the high-Tc superconductors are only a small part of the total. It is an obvious data bias. It can be seen in Figure 2 that the data distribution of the dataset-Tc is highly clustered. A majority of data have at least one close neighbor (Manhattan distance <1 in chemical composition space). Whereas, in subset-Tcmax, the Manhattan distance of any neighbors is larger than 1.5, which means the data distribution is quite dispersed. Models trained with subset-Tcmax are supposed to be more trustable because of better data distribution, and the predictive score on the test data (24 most-concerned superconductors) is a good measure of the generalization ability, especially for the prediction of new materials.

**Feature design and feature selection in machine learning**

For the feature set, all features are denoted as “[orbital attribute].[shells selection].[math operator 1].[math operator 2]”, where [orbital attribute] means what attribute of the orbitals is considered, and [shell selection] means which electron shells are considered. As shown in Figure 3, three [orbital attribute], nine [shells selection], and seven [math operator] were defined by a few simple rules. Discarding the features containing empty values and the features having zero variance, we got 441 usable features, each of which has explicit physical meanings.

After feature selection, a feature subset consisting of 4 features was selected to train the final model. More details of feature design and feature selection can be found in the experimental procedures and Tables S3–S6.

**Results of machine learning model**

Figure 4 shows the result of the final model. The coefficient of determination (R²) on the test data (24 most-concerned superconductors) is 0.84. This value is not as high as those in previous ML studies (R² is often better than 0.9), but please note that the extrapolating distance of our test data (Manhattan distance >1.5) is much larger than usual. In this case, our R² can be considered a reliable indicator of the generalization ability, especially for the prediction of new materials.

According to the features in the final model, we can learn useful information. It can be seen in Figures 4B and 4C that for the feature “Nu.all.wavg.wavg” as well as “Nu.outer.min.wavg,” all cuprates have similar values. “Nu.all.wavg” means the weighted value of the unfilled number of all the valence electron shells in
each element, while “Nu.outer.min” means the minimum of the unfilled number of the outer valence electron shells (s and p shells) in each element.

As most high-\(T_c\) superconductors are cuprates, those two features are of much importance because they indicate some unique characteristics of cuprates. Unfortunately, we cannot know whether they are associated with the origin of high \(T_c\) because the cuprates have many unique characteristics, but not every of those characteristics is associated with high \(T_c\).

Meanwhile, for the other two features, “E.all.range.range” and “E.s.sum.range” (Figures 4D and 4E), there is a monotonic trend in the \(T_c^{\text{max}}\)-feature plot, respectively. The \(T_c^{\text{max}}\) only can reach high values when the values of those two features are large. It is a necessary but not sufficient condition of high \(T_c^{\text{max}}\). In other words, it defines an upper ceiling of the \(T_c^{\text{max}}\). The latter two features are of importance in physics. For each element, “E.all.range” means the span of the energy level of all valence electron shells, and “E.all.range.range” for a crystal is comparable to the width of the valence band in crystal. “E.s.sum” for a crystal means the summary value of the energy level of valence s shells, and “E.s.sum.range” for a crystal is basically determined by the energy level of the deepest valence s shell in the whole crystal, which is usually at the bottom of the valence band.

As those two features are both derived from the energy-level distribution of valence electrons, and they can be traced to particular electron shells, further discussion would better be made based on the band structure of crystals.

**Correlation between \(T_c^{\text{max}}\) and a band structure parameter**

First-principles calculations were performed to see the energy-level distribution in the energy band of various superconductors (more details can be found in data and code availability). After checking tens of well-known superconductors, we found a band characteristic parameter corresponding to the selected features in the final ML model.

Figure 5 shows the orbital distribution characteristics of four representative superconductors. It can be seen that for a typical high-\(T_c\) superconductor, (Sr,Ca)CuO\(_2\), there is an orbital coupling at the bottom of the valence band, between O 2s and Sr 4p (Ca 3p). Particularly, for all cuprates, the O 2s shell has a deep energy level, leading to a large value of “E.s.sum.range.” Meanwhile, the energy level interval between O 2s and O 2p shells is large, hence the value of “E.all.range.range” is large.

It is also can be seen from Figure 5B that the energy levels of those involved orbitals can be roughly estimated according to their corresponding energy levels in isolated atoms, which are used as the input data in the feature design of ML. Thus, the results of ML and band structure simulation are mutually verified. They both indicate a materials-dependent law about the variation of \(T_c^{\text{max}}\).
involved ions also matters. For instance, as shown in Table 1, the involved electron orbitals, the atomic distance of the two superconductors, we found that in addition to the energy level of they still can participate in orbital hybridization and coupling.

As shown in Figure 6, there is a monotonous trend of $T_c^{\text{max}}$ for various superconductors. It can be expressed as high $T_c^{\text{max}}$ asks for larger $\Delta E$. This trend is universal because each data shown in Figure 6 represents not only itself but also a series of similar materials (e.g., “Y–123” is not only YBa$_2$Cu$_3$O$_7$; it also represents all LnBa$_2$Cu$_3$O$_7$, Ln = rare-earth elements). So, Figure 6 actually covers hundreds of superconducting materials, including all super-conductors ($T_c^{\text{max}} > 15$ K) as far as we know.

Like many well-known $T_c$ laws (see Table S7), the $T_c^{\text{max}}$-$\Delta E$ trend in Figure 6 is informative. Here, $\Delta E$ represents the energy level interval between the hybridized valence orbitals, usually the lowest unsaturated shell (LUS) and the highest saturated shell (HSS). $\Delta E$ is related to the magnitude of orbital hybridization, orbital coupling, and orbital interactions.

For instance, among all known superconductors at ambient pressure, the cuprates have the deepest valence band. Particularly, the superconductors $T_c^{\text{max}} > 100$ K are all Ca-containing cuprates (e.g., Hg–1223, Bi–2223), in which the calcium ions are sandwiched between Cu–O planes. The energy level of the HSS of Ca (i.e., Ca 3p; see Figure 5A) is deeper than that of most other elements. It is often overlooked that the Ca 3p and O 2s orbitals are also valence orbitals. Although their energy levels are as deep as about 15–20 eV below the Fermi level, they still can participate in orbital hybridization and coupling.

Furthermore, after checking the crystalline structures of those superconductors, we found that in addition to the energy level of the involved electron orbitals, the atomic distance of the two involved ions also matters. For instance, as shown in Table 1, a parameter $\lambda$ is defined as $\lambda = d - r_1 - r_2$ to compare the bond length in different crystals. Smaller $\lambda$ means a compacted lattice and usually stronger orbital coupling. It can be seen that the $\lambda$ of Ca–O bond in Ca-containing cuprates is much smaller than that in other Ca-containing superconductors. It also can be seen in Table 2 that for some compounds having similar crystalline structures (La–214), smaller $\lambda$ (more compacted lattice) leads to higher $T_c^{\text{max}}$.

It should be noted that the green dashed line in Figure 6 is not a fitting line but an upper ceiling of $T_c^{\text{max}}$. Actually, for the superconductors shown in Figure 6, their $\lambda$ is relatively small. It is not shown in Figure 6, but is partially in Figure S1, that there are also many superconductors, far below the ceiling line of $T_c^{\text{max}}$, each of which has a large $\lambda$.

**DISCUSSION**

**Guidance for exploring superior superconductors**

Figure 6 and Tables 1 and 2 lead to a useful criterion for screening candidates of high- $T_c$ superconductor. It suggests two necessary conditions for high $T_c^{\text{max}}$: (1) having a deep-energy-level orbital coupling and (2) small bond lengths between the ions providing those deep-energy-level orbitals.

Although it cannot tell which materials are superconductors, it does can tell which superconductors can have higher $T_c^{\text{max}}$ and which ones do not stand a chance at all. It suggests that high $T_c^{\text{max}}$ can be supported only by some particular elements.

For example, calcium-containing superconductors should have relatively high $T_c^{\text{max}}$ because the energy level of Ca 3p is rather deep (~33.77 eV). In fact, the $T_c^{\text{max}}$ of many calcium-containing cuprates are over 100 K, higher than all other superconductors at ambient pressure. It is somehow unexpected because for years, it has been the Cu–O planes that have been regarded as the origin of high $T_c^{\text{max}}$ in cuprates, whereas the ions nearby the Cu–O planes are thought to have a minor influence. However, our results indicate that the role of Ca cations may be much more important than we used to think. Moreover, most earlier studies on cuprates have focused on the band...
structure and electron interactions around the Fermi level, which is mostly derived from Cu 3d and O 2p orbitals. Now, according to this work, the orbital coupling between Ca 3p and O 2s should also be dragged into the spotlight.

The distribution of electron orbitals could be dramatically altered at ultrahigh pressure. First-principles calculations have shown that either LaH₁₀ or H₃S has a wide valence band at ultrahigh pressure.⁶³–⁶⁵ There is a strong orbital hybridization between La 5p (or S 3p) and H 1s. The bottom of their valence band is deeper than ~20 eV, respectively. As a comparison, the bottom of the valence band of the cuprates is slightly less than ~20 eV. Besides, the lattice of LaH₁₀ and H₃S is greatly compacted at ultrahigh pressure. Although the compressed hydrides are usually thought to be BCS superconductors, which rest upon the electron-phonon interaction, our results show that their high Tₓ could have another explanation. In compressed hydrides, the elements other than H are not arbitrary. They should have deep-energy-level valence electrons.

Therefore, to design a high-Tₓ superconductor, the elements having deep-energy-level valence orbitals are preferred. Only some of the elements (alkali, alkaline, rare earth, as well as O, S, Se, As, F, Cl, ...) can provide deep-energy-level valence orbitals (see also Figures S2 and S3). However, those elements are liable to form ionic crystals, which are usually insulators. To get itinerant carriers, a third element is needed to provide covalent bonds, making the crystal metallic. In general, a high-Tₓ superconductor is supposed to consist of more than two elements, including at least one transition element and at least one element having deep-energy-level valence orbitals.

It should be emphasized that it is still unfeasible so far to make confident predictions on new high-Tₓ superconductors. To the best of our knowledge, superconductivity is affected by many factors, and high-Tₓ superconductivity has multiple necessary conditions, some of which may yet be unknown. High Tₓ would not exist unless those necessary conditions are met altogether in a superconductor. Sometimes, a material meets all the conditions of high Tₓ, but it is just not a superconductor but rather an insulator or a magnet.

ML models often mistake insulators or magnets for high-Tₓ superconductors, due to the lack of relevant information in the input data. In the absence of that essential information, the ML models were only able to give a candidate list of “materials that could be high-Tₓ superconductors, but not necessarily superconductor.”

To get a full guidance, we need to collect all the indispensable information, including but not limited to “energy level distribution
of valence electrons,” “bond lengths,” “energy dispersion near the Fermi level,” “strength of spin fluctuation,” and “strength of magnetism order.” Unfortunately, the latter three are usually not easy to obtain. That is why it is so hard to find another material family of high-$T_c$ superconductors.

Nevertheless, more superior superconductors may be found in two-dimensional (2D) materials, although their bulk materials are insulators. Gate voltage or lattice strain may produce adequate carriers in such materials.

In brief, the existence of deep-energy-level valence shells can be seen as a necessary, but not sufficient, condition of the high $T_c$ superconductors.

A clue to the pairing mechanism

What is the pairing glue in high-$T_c$ superconductors? It is a core issue in the mechanism study on superconductivity. The pairing glue (strictly speaking, the electron interaction causing the attractive potential between the two pairing electrons) is supposed to be derived from some kind of quasi-particle, elementary excitation, or fluctuation, in the lattice. For instance, in cuprates, spin fluctuation is a promising candidate for pairing glue.\(^5\) In compressed hydrogen-rich materials,\(^6\) it is the lattice vibration. However, neither spin fluctuation nor lattice vibration can explain the $T_c^{\text{max}}$-$\Delta E$ trend in this work. The $T_c^{\text{max}}$-$\Delta E$ trend shows that the excitation and fluctuation of the deep-energy-level valence electrons can also serve as a pairing glue in high-$T_c$ superconductivity. Please note that we did not rule out spin fluctuation or lattice vibration. Our results imply a physical picture of the pairing mechanism:

The unconventional high-$T_c$ superconductors may need two pairing glues ($q_k$ and $q_s$). One ($q_k$) gives rise to the angular momentum conservation of the Cooper pair, and the other ($q_s$) causes the momentum conservation of the Cooper pair. In cuprates and iron-based superconductors, the spin fluctuation can function as $q_s$, while the fluctuation of deep-energy-level valence electrons can function as $q_k$.

$q_k$ is always indispensible to any superconductor, whereas the necessity of $q_s$ depends on the coherent length. When the coherent length is long, it is easy for an electron to find its mate (the electron having opposite spin and momentum) within the range of coherent length (usually dozens of nanometers). The condition of angular momentum conservation (having opposite spin) is naturally met. In that case, only $q_k$ is necessary, and it can be contributed from the electron-phonon interaction. When the coherent length is short (the case of most high-$T_c$ superconductors), the pairing electrons need to find their mates, respectively, within several nanometers. However, in the range of several nanometers, there are only a few electrons available. The condition of angular momentum conservation is no longer naturally met. In that case, $q_s$ is indispensible, too.

Thus, a fresh understanding of the pairing mechanism of high-$T_c$ superconductivity is obtained. The pairing mechanism of all known superconductors can be put into a unified framework. Briefly, the pairing of two electrons needs two necessary conditions: momentum conservation and angular momentum conservation. The momentum conservation always needs a pairing glue of charge fluctuation, and the angular momentum conservation needs a pairing glue of spin fluctuation if the coherent length is too small.

Conclusion and prospects

In summary, ML models were used to predict the upper limit of $T_c$ ($T_c^{\text{max}}$) of superconducting materials. Enlightened by the ML

---

Table 1. A comparison of several Ca-containing materials

| Material | $X$ | Involved orbitals | $E_{\text{orb}}$ (eV)\(^6\) | $d_{\text{Ca},X}$ (pm)\(^5\) | $r_X$ (pm)\(^a\) | $\lambda$ (pm) | $T_c^{\text{max}}$ (K) |
|----------|-----|-------------------|-------------------|-------------------|---------|---------|------------------|
| CaO      | O   | Ca 3p – O 2s      | −29.17            | 239               | 140     | 0       | N/A              |
| Ca (ambient pressure) | Ca | Ca 3p – Ca 3p  | −33.77            | 395               | 99      | 197     | N/A              |
| CaI\(_2\) | Ca | Ca 3p – Ca 3p  | −33.77            | 383               | 99      | 185     | 13               |
| Ca (216 GPa)\(^b\) | Ca | Ca 3p – Ca 3p  | −33.77            | 231\(^c\)         | 99      | 33      | 29               |
| Bi\(_2\)Sr\(_2\)Ca\(_2\)Cu\(_3\)O\(_y\) | O | Ca 3p – O 2s  | −29.17            | 252               | 140     | 13      | 110              |
| HgBa\(_2\)Ca\(_2\)Cu\(_3\)O\(_y\) | O | Ca 3p – O 2s  | −29.17            | 236               | 140     | −3      | 134              |

---

\(^5\)“$X$” represents the neighboring atom of Ca. $E_{\text{orb}}$ is the energy level of the involved orbital in $X$. $d_{\text{Ca},X}$ is the bond length. $r_X$ and $r_M$ are ionic radii of Ca and $X$, respectively. $\lambda = d – r_{\text{Ca}} – r_X$ is a characteristic parameter for the Ca-$X$ bond.

| Material | Involved orbitals | $E_{\text{orb}}$ (eV)\(^6\) | $d_{O,M}$ (pm)\(^5\) | $r_M$ (pm)\(^b\) | $\lambda$ (pm) | $T_c^{\text{max}}$ (K) |
|----------|-------------------|-------------------|-------------------|---------|---------|------------------|
| La\(_2\)CuO\(_4\) | O 2s – La 5p | −26.36            | 255               | 115     | 0       | 35               |
| Sr\(_2\)RuO\(_4\) | O 2s – Sr 4p | −27.69            | 270               | 113     | 17      | 1.5              |
| Sr\(_2\)IrO\(_4\) | O 2s – Sr 4p | −27.69            | 277               | 113     | 23      | N/A              |

---

\(^6\)“$M$” represents the neighboring atom of O. $E_{\text{orb}}$ is the energy level of the involved orbital in $M$. $d_{O,M}$ is the bond length. $r_O$ and $r_M$ are ionic radii of O and M, respectively ($r_O = 140$ pm). $\lambda = d – r_O – r_M$ is a characteristic parameter for the O-$M$ bond.

---

\(^a\)Chu et al.\(^{62}\)

\(^b\)Pauling ionic radii.

---

\(^5\)Pauling ionic radii.

---

\(^c\)DFT simulated value.
Figure 7. Flowchart of feature selection

results, we found a correlation between $T_c^{\text{max}}$ and a band structure parameter. It suggests that the energy-level distribution of valence electrons is crucial for high-$T_c$ superconductivity. It should be noted that the $T_c^{\text{max}}$ trend is in fact based on the band structure simulation results. Here, the role of ML is just to give a hint. Thanks to the explicit meaning of the features in the ML model, theoretical induction based on our ML results is effective and efficient. It implies a physical picture of electron pairing in high-$T_c$ superconductors. It also reveals a necessary condition of the high $T_c$, which provides convenient guidance for designing superior superconductors.

Although artificial intelligence nowadays can give useful clues and guidance, they are still not good enough to predict new high-$T_c$ superconductors with confidence. Improvements may be achieved by means of graph neural network algorithms, which are good at dealing with structural information, including both the crystalline structures and the electron band structures.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact is Haiyou Huang (huanghy@mater.ustb.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The source code and input data of ML, and the results of density functional theory (DFT) calculations, are available Data S1.

Machine learning

Dataset

$T_c^{\text{max}}$ is a subset of dataset-$T_c$. According to the chemical composition, superconductors in the dataset-$T_c$ were categorized into 1,000 groups, according to the component number and the content of each element in the chemical formulas. Each group can be labeled as “$n$-$E$-$c$,” where $n$ is the component number (the compound consisting of how many elements), $E$ is the element, and $c$ is the content of the element. Before the categorization, all the values of $c$ are rounded to integers. A superconductor can appear in multiple groups. For example, $\text{YBa}_2\text{Cu}_3\text{O}_7$ appears in four groups: 4-$\text{Y}$, 4-$\text{Ba}$, 4-$\text{Cu}$, and 4-$\text{O}$; $\text{MgO}$ (0.81$\text{Al}_0.19\text{B}_2$) appears in two groups: 3-$\text{Mg}$ and 3-$\text{B}$; and $\text{FeSe}_0.5\text{Te}_0.5$ appears in three groups: 3-$\text{Fe}$, 3-$\text{Se}$, and 3-$\text{Te}$. In each group, we picked the entry having the highest $T_c$ value and considered it the $T_c^{\text{max}}$. Sometimes, several groups give the same picked entry of $T_c^{\text{max}}$. After merging the duplicate entries, dataset-$T_c^{\text{max}}$ has 1,008 entries.

Twenty-four well-known or most-studied superconductors in the dataset-$T_c$ were picked out and retained as the test set to assess the generalization ability of the final model. In pursuit of long extrapolation distances when testing models, the train set should not include those 24 superconductors, as well as their neighboring data.

In those 24 superconductors, 10 entries are in dataset-$T_c^{\text{max}}$ (1,008), and 42 entries in dataset-$T_c^{\text{max}}$ (1,008) are found to be the neighbors of those 24 superconductors. Here, the word neighbor means a small Manhattan distance in the composition space (<1.5). So, those entries (10+$42=52$) were removed from the dataset-$T_c^{\text{max}}$, resulting in a subset of dataset-$T_c^{\text{max}}$. The subset-$T_c^{\text{max}}$ has (1,008 – 52 = 957) entries.

During the feature selection, the subset-$T_c^{\text{max}}$ was used to train the ML models in genetic algorithm (GA). For each model, the subset-$T_c^{\text{max}}$ was randomly split into a train set (85%) and a test set (15%). The predictive accuracy of the models on the test set was used as the criterion to select better features.

For the final model after feature selection, all 957 entries of the subset-$T_c^{\text{max}}$ were used as the train data, and the test data are the retained 24 superconductors.

Feature design

For the feature set, all features were designed based on the orbital attributes of valence electrons. The orbital attributes include the energy level and the occupancy of valence orbitals of isolated atoms.

The feature extraction imitates the way of MAGPIE. Each feature is noted as: [orbital attribute].[shells selection].[math operator 1].[math operator 2], where orbital attribute means what property of the orbitals is considered, while shell selection means which electron shells are considered.

We defined three [orbital attribute], nine [shells selection], and seven [math operator]. That’s 3 $\times$ 9 $\times$ 7 $\times$ 7 = 1,233 features in total. Discarding the features containing empty values and the features having zero variance, we got 441 usable features.

It should be noted that valence electrons here stands for the electrons participating in the orbital hybridization and coupling. The valence electrons usually come from the outmost saturated orbital and the unsaturated orbital(s). Here, we took an energy cutoff of ~36 eV, and the electrons in the energy levels range [0, ~36] (eV) are deemed valence electrons.

More details and two examples of feature extraction can be found in Tables S3–S5.

Feature selection

As shown in Figure 7, the feature selection consists of two steps: (1) the 441 features were filtered by their Pearson correlation coefficient ($P$) and (2) feature subsets were filtered by GA, according to their performance in ML models.

In step 1, the 441 features were eliminated one by one. At first, we found the two features having the highest Pearson correlation coefficient in all features. Between those two features, we eliminated the one having the lower Pearson correlation coefficient. That’s 3 $\times$ 9 $\times$ 7 $\times$ 7 = 1,233 features in total. Discarding the features containing empty values and the features having zero variance, we got 441 usable features.

In step 2, GA was used to find the “best” subset of the 124 features. Random forest regression (RFR) models were trained to give the coefficient of determination ($R^2$) to measure the fitness of the feature subsets. For each RFR model, the dataset (subset-$T_c^{\text{max}}$, 957 entries) was randomly split into a train set (85%) and a test set (15%).

We ran the GA codes 500 times. Each GA run had 15 generations, and the population in each generation consisted of 100 feature subsets. Each feature subset had 4 features at most, which were randomly picked from the 124 features. Each GA run recommended a feature subset that had the highest $R^2$ score on the test data.
SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100609.

ACKNOWLEDGMENTS

This research was funded by Guangdong Province Key Area R&D Program (2019B010940001), National Key Research and Development Program of China (2018YFB0704300, 2020YFB0704504), and Scientific and Technological Innovation Foundation of Shunde Graduate School, USTB (BK19BE030).

AUTHOR CONTRIBUTIONS

Conceptualization, H.H., Y. Liu, N.C., J.T., Y.S., and K.J.; software and data curation, H.F., Y.Z., D.X., Y. Liu, and N.C.; writing – original draft, Y. Liu and Y.Z.; writing – review & editing, Y. Liu, H.H., Y. Li, N.C., J.Y., and K.J.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Mann, A. (2011). High-temperature superconductivity at 25: still in suspense. Nature 475, 280–282. https://doi.org/10.1038/475280a.
2. Geballe, T.H., Hammond, R.H., and Wu, P.M. (2015). What Tc tells. Physica C: Superconductivity and its Applications 514, 9–16. https://doi.org/10.1016/j.physc.2015.02.009.
3. Zhou, X., Lee, W.-S., Imada, M., Trivedi, N., Phillips, P., Kee, H.-Y., Törmä, P., and Eremer, M. (2021). High-temperature superconductivity. Nat. Rev. Phys. 3, 462–465. https://doi.org/10.1038/s42255-021-00324-3.
4. Uemura, Y.J., Le, L.P., Luke, G.M., Sternlieb, B.J., Wu, W.D., Brewer, J.H., Riseman, T.M., Seaman, C.L., Maple, M.B., Ishikawa, M., et al. (1991). Basic similarities among cuprate, bismuthate, organic, Chevrel-phase, and heavy-fermion superconductors shown by penetration-depth measurements. Phys. Rev. Lett. 66, 2665–2668. https://doi.org/10.1103/PhysRevLett.66.2665.
5. Uemura, Y.J., Luke, G.M., Sternlieb, B.J., Brewer, J.H., Carolan, J.F., Hardy, W.N., Kadono, R., Kempton, J.R., Kiefel, R.F., Kreitzman, S.R., et al. (1989). Universal correlations between Tc and n/M (carrier density over effective mass) in high-Tc cuprate superconductors. Phys. Rev. Lett. 62, 2317–2320. https://doi.org/10.1103/PhysRevLett.62.2317.
6. Homes, C.C., Dordevic, S.V., Strongin, M., Bonn, D.A., Liang, R., Hardy, W.N., Komiya, S., Ando, Y., Yu, G., Kaneko, N., et al. (2004). A universal scaling relation in high-temperature superconductors. Nature 430, 539–541. https://doi.org/10.1038/nature02673.
7. Bózóangi, I., He, X., Wu, J., and Bollinger, A.T. (2016). Dependence of the critical temperature in overdoped copper oxides on superfluid density. Nature 536, 309–311. https://doi.org/10.1038/nature19061.
8. Pines, D. (2013). Finding new superconductors: the spin-fluctuation gateway to high Tc and possible room temperature superconductivity. J. Phys. Chem. B 117, 13145–13153. https://doi.org/10.1021/jp403088e.
9. Jin, K., Butch, N.P., Kirshenbaum, K., Pklgione, J., and Greene, R.L. (2011). Link between spin fluctuations and electron pairing in copper oxide superconductors. Nature 476, 73–75. https://doi.org/10.1038/nature10308.
10. Yuan, J., Chen, Q., Jiang, K., Feng, Z., Lin, Z., Yu, H., He, G., Zhang, J., Jiang, X., Zhang, X., et al. (2022). Scaling of the strange-metal scattering in unconventional superconductors. Nature 602, 431–436. https://doi.org/10.1038/s41586-021-04305-5.
11. Maxwell, E. (1952). The isotope effect in superconductivity. Phys. Today 5, 14–18. https://doi.org/10.1063/1.3067418.
12. Chmaissem, O., Jorgensen, J.D., Short, S., Knizhnik, A., Eckstein, Y., and Shaked, H. (1999). Scaling of transition temperature and CuO planes buckling in a high-temperature superconductor. Nature 397, 45–48. https://doi.org/10.1038/16209.
13. Lee, C.H., Kihou, K., Iyo, A., Kito, H., Shirage, P.M., and Eisaki, H. (2012). Relationship between crystal structure and superconductivity in iron-based superconductors. Solid State Commun. 152, 644–648. https://doi.org/10.1016/j.ssc.2011.12.012.
14. Mizuguchi, Y., Hayashi, Y., Deguchi, K., Tsuda, S., Yamaguchi, T., Takeda, K., Kotegeawa, H., Tou, H., and Takano, Y. (2010). Anion height dependence of...
Tc for the Fe-based superconductor. Supercond. Sci. Technol. 23, 054013. https://doi.org/10.1088/0953-2048/23/5/054013.

15. Peng, Y.Y., Delea, G., Minola, M., Conni, M., Amorese, A., Di Castro, D., De Luca, G.M., Kummer, K., Salluzzo, M., Sun, X., et al. (2017). Influence of apical oxygen on the extent of in-plane exchange interaction in cuprate superconductors. Nat. Phys. 13, 1201–1206. https://doi.org/10.1038/nphys4248.

16. Bud’ko, S.L., Ni, N., and Canfield, P.C. (2009). Jump in specific heat at the superconducting transition temperature in BaFe2(Co0.75Fe0.25)2As2 and BaFe2(Ni0.25Fe0.75)2As2 single crystals. Phys. Rev. B 79, 220516. https://doi.org/10.1103/PhysRevB.79.220516.

17. Stewart, G.R. (2011). Superconductivity in iron compounds. Rev. Mod. Phys. 83, 1589–1652. https://doi.org/10.1103/RevModPhys.83.1589.

18. Wilson, S.D., Dai, P., Li, S., Chi, S., Kang, H.J., and Lynn, J.W. (2006). Resonance in the electron-doped high-transition-temperature superconductor Pr0.65La0.35CuO4. Nature 442, 59–62. https://doi.org/10.1038/nature04857.

19. Wu, Y.P., Zhao, D., Wang, A.F., Wang, N.Z., Xiang, Z.J., Luo, X.G., Wu, T., and Chen, X.H. (2016). Emergent kondo lattice behavior in iron-based superconductors AFe2As2 (A = K, Rb, Cs). Phys. Rev. Lett. 116, 147001. https://doi.org/10.1103/PhysRevLett.116.147001.

20. Nakamura, S., Moriya, T., and Ueda, K. (1996). Spin fluctuation-induced superconductivity in two and three dimensional systems. J. Phys. Soc. Jpn. 65, 4026–4033. https://doi.org/10.1143/JPSJ.65.4026.

21. Cao, Y., Fatemi, V., Fang, S., Watanabe, K., Taniguchi, T., Kaxiras, E., and Jarillo-Herrero, P. (2018). Unconventional superconductivity in magic-angle graphene superlattices. Nature 556, 43–50. https://doi.org/10.1038/s41586-018-02160.

22. Hashimoto, K., Cho, K., Shibauchi, T., Kasahara, S., Mizukami, Y., Katsumata, R., Tsuruhara, Y., Terashima, T., Ikeda, H., Tanatar, M.A., et al. (2012). A sharp peak of the zero-tempereature penetration depth at optimal composition in BaFe2(As0.7P0.3)2. Science 336, 1554–1557. https://doi.org/10.1126/science.1219821.

23. Moriya, T., and Ueda, K. (2003). Antiferromagnetic spin fluctuation and superconductivity. Rep. Prog. Phys. 66, 1299–1341. https://doi.org/10.1088/0034-4885/66/8/202.

24. Lee, P.A. (2008). From high temperature superconductivity to quantum critical temperature. npj Comput. Mater. 2, 180510. https://doi.org/10.1038/s41524-018-00666-7.

25. Isayev, O., Fourches, D., Muratov, E.N., Oses, C., Rasch, K., Tropsha, A., and Curtarolo, S. (2015). Materials cartography: representing and mining materials space using structural and electronic fingerprints. Chem. Mater. 27, 735–743. https://doi.org/10.1021/cm503507h.

26. Zhang, H.-r., Zhang, Y., Dai, D.-b., Cao, M., and Shen, W.-f. (2016). Modelling and optimization of the superconducting transition temperature. Mater. Des. 92, 371–377. https://doi.org/10.1016/j.matdes.2015.12.081.

27. Stanev, V., Oses, C., Kunse, A.G., Rodriguez, E., Paglione, J., Curtarolo, S., and Takeuchi, I. (2018). Machine learning modeling of superconducting critical temperature. npj Comput. Mater. 4, 29. https://doi.org/10.1038/s41524-018-0095-8.

28. Liu, Z.-L., Kang, P., Zhu, Y., Liu, L., and Guo, H. (2020). Material informatics for layered high-Tc superconductors. Apl. Mater. 8, 061104. https://doi.org/10.1063/5.0004641.

29. Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. Comput. Mater. Sci. 154, 346–354. https://doi.org/10.1016/j.commatsci.2018.07.052.

30. Konno, T., Kurokawa, H., Nabeshima, F., Sakishita, Y., Ogawa, R., Hosako, I., and Maeda, A. (2021). Deep learning model for finding new superconductors. Phys. Rev. B 103, 014509. https://doi.org/10.1103/PhysRevB.103.014509.

31. Liu, Y., Zhang, H., Xu, Y., Li, S., Dai, D., Li, C., Ding, G., Shen, W., and Qian, Q. (2018). Prediction of superconducting transition temperature using a machine-learning method. Mater. Tech. 52, 639–643. https://doi.org/10.17222/mit.2018.043.
conventional superconductors. J. Phys. Chem. Lett. 9, 1920–1926. https://doi.org/10.1021/acs.jpcllett.8b00615.
50. Dan, Y., Dong, R., Cao, Z., Li, X., Niu, C., Li, S., and Hu, J. (2020). Computational prediction of critical temperatures of superconductors based on convolutional gradient boosting decision trees. IEEE Access 8, 57686–57878. https://doi.org/10.1109/ACCESS.2020.2981874.
51. Hutcheon, M.J., Shipley, A.M., and Needs, R.J. (2020). Predicting novel superconducting hydrides using machine learning approaches. Phys. Rev. B 101, 144505. https://doi.org/10.1103/PhysRevB.101.144505.
52. Roter, B., and Dordevic, S.V. (2020). Predicting new superconductors and their critical temperatures using machine learning. Physica C: Superconductivity and its Applications 575, 1353689. https://doi.org/10.1016/j.physc.2020.1353689.
53. Le, T.D., Noumeir, R., Quach, H.L., Kim, J.H., Kim, J.H., and Kim, H.M. (2020). Critical temperature prediction for a superconductor: a variational bayesian neural network approach. IEEE Trans. Appl. Supercond. 30, 1–5. https://doi.org/10.1109/TASC.2020.2971456.
54. Regnault, N., Xu, Y., Li, M.-R., Ma, D.-S., Jovanovic, M., Yazdani, A., Parkin, S.S.P., Feiner, C., Schoop, L.M., Ong, N.P., et al. (2022). Catalogue of flat-band stoichiometric materials. Nature 603, 824–828. https://doi.org/10.1038/s41586-022-04519-1.
55. Stewart, G.R. (2017). Unconventional superconductivity. Adv. Phys. X. 66, 75–196. https://doi.org/10.1080/00018732.2017.1331615.
56. Superconducting Material Database. http://supercon.nims.go.jp/supercon/material_menu.
57. Hirsch, J.E., Maple, M.B., and Marsiglio, F. (2015). Superconducting materials classes: introduction and overview. Physica C: Superconductivity and its Applications 514, 1–8. https://doi.org/10.1016/j.physc.2015.03.002.
58. Li, D., Lee, K., Wang, B.Y., Osada, M., Crossley, S., Lee, H.R., Cui, Y., Hikita, Y., and Hwang, H.Y. (2019). Superconductivity in an infinite-layer nickelate. Nature 572, 624–627. https://doi.org/10.1038/s41586-019-1496-5.
59. Bergerhoff, G., Hundt, R., Sievers, R., and Brown, I.D. (1983). The inorganic crystal structure data base. J. Chem. Inf. Comput. Sci. 23, 66–69. https://doi.org/10.1021/ci00038a003.
60. Herman, F., Sherwood Skillman, S., and Arents, J. (1964). Atomic Structure Calculations, 111 (Prentice-Hall).
61. Hamlin, J.J. (2015). Superconductivity in the metallic elements at high pressures. Physica C Supercond 514, 59–76. https://doi.org/10.1016/j.physc.2015.02.032.
62. Chu, C.W., Deng, L.Z., and Lv, B. (2015). Hole-doped cuprate high temperature superconductors. Physica C Supercond 514, 290–313. https://doi.org/10.1016/j.physc.2015.02.047.
63. Bernstein, N., Hellberg, C.S., Johannes, M.D., Mazin, I.I., and Mehl, M.J. (2015). What superconducts in sulfur hydrides under pressure and why. Phys. Rev. B 97, 060511. https://doi.org/10.1103/PhysRevB.97.060511.
64. Ortenzi, L., Cappelluti, E., and Pietronero, L. (2016). Band structure and electron-phonon coupling in H2S: a tight-binding model. Phys. Rev. B 94, 064507. https://doi.org/10.1103/PhysRevB.94.064507.
65. Liu, H., Naumov, I.I., Hoffmann, R., Ashcroft, N.W., and Hemley, R.J. (2017). Potential high-Tc superconducting lanthanum and yttrium hydrides at high pressure. Proc. Natl. Acad. Sci. USA 114, 6990–6995. https://doi.org/10.1073/pnas.1704505114.
66. Struzhkin, V.V. (2015). Superconductivity in compressed hydrogen-rich materials: pressing on hydrogen. Physica C: Superconductivity and its Applications 514, 77–85. https://doi.org/10.1016/j.physc.2015.02.017.
67. Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Comput. Mater. 2, 16028. https://doi.org/10.1038/npjcompmat.2016.28.
68. Lin, L.I.K. (1989). A concordance correlation coefficient to evaluate reproducibility. Biometrics 45, 255–268. https://doi.org/10.2307/2532051.
69. Sapra, R.L. (2014). Using R2 with caution. Current Medicine Research and Practice 4, 130–134. https://doi.org/10.1016/j.cmrpr.2014.06.002.
70. Clark, S.J., Segall, M.D., Pickard, C.J., Hasnip, P.J., Probert, M.I.J., Refson, K., and Payne, M.C. (2005). First principles methods using CASTEP. Z. für Kristallogr. - Cryst. Mater. 220, 567–570. https://doi.org/10.1524/zkri.220.5.567.65075.