Leveraging composition-based energy material descriptors for machine learning models

Materials Today Communications 36 (2023) 106579

Contents lists available at ScienceDirect

Materials Today Communications

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

Leveraging composition-based energy material descriptors for machine learning models

Giovanni Trezza, Eliodoro Chiavazzo *

Department of Energy, Politecnico di Torino, C.so Duca degli Abruzzi 24, Torino 10129, Italy

ARTICLE INFO

Dataset link: DOI:10.5281/zenodo.7725592

Keywords:
Machine learning
Material classification
Composition-based descriptors
Energy materials
Superconductors

ABSTRACT

A comparison of several classifiers is presented, with a focus on the key choice and construction of a minimal set of suitable material features. To this end, an investigation is conducted over a properly selected and high quality database reporting low temperature superconductors, featurized by composition-based descriptors. Fully general strategies to reduce the number of descriptors for material classification are proposed and discussed. The first strategy aims at testing possible invariance of the target material property (here the critical temperature) with respect to (binary) groups of composition-based features in the form $x_i^a x_j^b$, $a, b \in \mathbb{R}$. In addition, a multi-objective optimization procedure for reducing the set of composition-based material descriptors is also suggested and tested on the chosen use case. The latter procedure is then proven to be particularly convenient to be used in combination with Bayesian type classifiers. Finally, by means of the best-performing classification models, an analysis is conducted over all the $\sim 40,000$ inorganic compounds without Ni, Fe, Cu, O in Materials Project (and not in the SuperCon database, here used for model training) and the corresponding predictions are provided. Among those, 41 materials are classified to show $T_c \geq 15\,\mathrm{K}$ with a probability higher than or equal to 0.6.

1. Introduction

Construction of reliable and predictive models for material properties is becoming an aspect of general interest in a number of areas. With a special focus on materials for energy applications, the *in silico* prediction of physical properties without resorting to time consuming simulations or expensive experiments is of utmost importance. One of the reason being that low-cost, long-lasting materials with high performance is key for energy storage technologies, as it may be responsible from most of the total cost [1].

A number of technological areas ranging from the energy up to healthcare sector are being transformed by superconducting materials and may greatly benefit from the discovery of new high performance materials. Superconductors are materials characterized by zero electrical resistivity when cooled below a superconducting critical temperature T_c [2]. Due to this fundamental property, such compounds have attracted attention in a wide range of different fields. Superconducting Magnetic Energy Storage (SMES) systems allow to store energy by means of a DC current flowing through a superconducting coil; as a consequence, energy can be stored in the resulting magnetic field with almost no loss and can be released back by discharging the coil [3]. Superconducting electromagnets are employed in fusion reactors like tokamak [4], Magnetic Resonance Imaging (MRI) [5,6], Nuclear Magnetic Resonance (NMR) machines [7,8], particle accelerators [9]. Other applications include Superconducting Quantum Interference Devices (SQUIDs) [10], particle detectors [11], fast fault current limiters [12]. As such, discovery of new superconductors in the near future is highly desirable and can have a crucial impact on the energy sector (among others).

Therefore, several recent research studies have made extensive use of Machine Learning (ML)-based approaches. In particular, Stanev et al. [13] trained and validated models both for classification–prediction of the classes *superconductor/non-superconductor* - and for regression-prediction of the critical temperature, employing composition-based features together with the experimental T_cs of known superconductors. Konno et al. [14] represented each chemical formula with four tables, corresponding to the periodic table blocks s, p, d, f, with such information being the input of a convolutional Deep Neural Network (DNN) able to predict the critical temperature. Le et al. [15] trained and validated a Variational Bayesian Neural Network using superconductors composition-based features for the T_c prediction. Roter et al. used only chemical elements and stoichiometry, with no extracted features, to predict the critical temperature [16] and to cluster superconductors [17].

* Corresponding author.
E-mail address: eliodoro.chiavazzo@polito.it (E. Chiavazzo).

https://doi.org/10.1016/j.mtcomm.2023.106579
Received 17 May 2023; Received in revised form 15 June 2023; Accepted 29 June 2023
Available online 6 July 2023

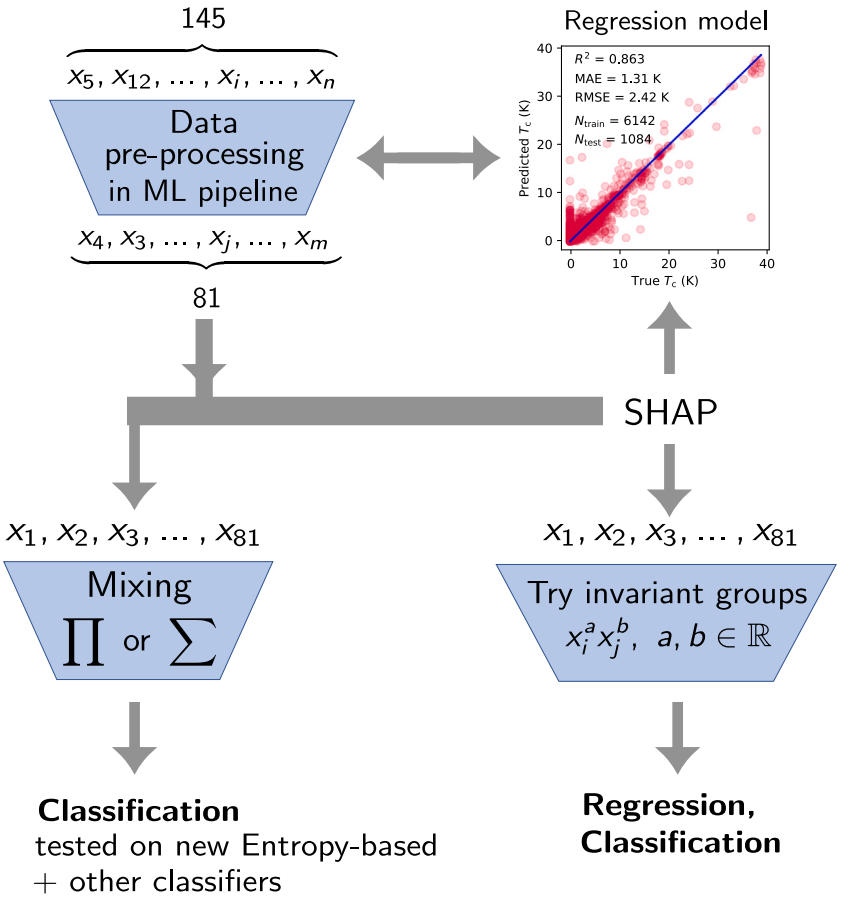

Fig. 1. Overview of the protocol used to find a reduced set of ruling descriptors for conventional superconductivity and for the construction of optimized mixed features. Over 7000 chemical compositions were featurized with 145 descriptors. A regression model has been trained and validated over this dataset, and during the pre-processing routines (i.e., feature reduction by means of linear correlation analysis, descriptors variance analysis, correlation analysis with the T_c, see Supplementary Note 4 for details), many of those features are discarded, ending up with 81 descriptors. By means of SHAP, those 81 features are ranked in terms of importance. The work aimed at finding optimized mixed features for both regression/classification in the form $x_i^a x_j^b$; and for classification, with power or linear combination of the primitive features. The latter descriptors were tested over both new entropy-based classifiers and other classifiers.

In this work, the focus is deliberately directed to classical low temperature superconductors, given the possibility to rely upon a high-quality database [18]. In light of this, during the data collection process from the database for training purposes, the exclusion of all materials containing the elements Fe, Ni, and Cu was implemented (to prevent unconventional superconductivity [19]). Furthermore, the removal of materials containing oxygen was also carried out to avoid oxides and increase the likelihood of including materials in this analysis that are more prone to exhibit ductile behavior.

After the extraction of 145 composition-based features by means of Matminer [20] for each material formula, a tree-based regression model was trained and validated for the prediction of the critical temperature, over which insights were obtained of the most important features by means of SHAP [21–23]. Based on those features, several binary classifiers are thus compared, to distinguish compounds with the critical T_c exceeding a predefined threshold value from the remaining samples.

A special focus of this study is on the identification and construction of a minimal and optimal set of key material descriptors (or features) to be adopted for classification purposes. To this end, two main strategies were pursued as briefly described below and schematically represented in Fig. 1:

- First, the aforementioned SHAP analysis was performed, establishing a descriptor ranking based on the relevance of single features $\{x_{i=1,\dots,n}\}$, where n is the maximum number of adopted features;
- In the spirit of the work by Tegmark and collaborators [24], a general approach is proposed for investigating possible symmetries of the target quantity (here T_c) with respect to groups of the originally chosen features (according to the order suggested by SHAP). Without loss of generality, the focus is on feature binary groups in the form $x_i^a x_j^b$, with $a, b \in \mathbb{R}$ being properly selected constants. To this end, a proper algorithm based on the computation of the output gradient with respect to the input features by means of a Deep Neural Network (DNN) is discussed.
- Ultimately, a general framework for drastically reducing the number of the classifier features is proposed in the form of both single and multi-objective optimization problem. Among other purposes, the latter approach proves particularly convenient to *synthetically* construct new descriptors particularly suited for Bayesian type classifiers, including a novel entropy-based classifier introduced and tested in this work.

Finally, the classifiers with the best performance was employed to rank ~40,000 compounds from Materials Project [25] and not occurring in the SuperCon.

2. Methods

2.1. Dataset creation

First, the focus is directed towards the construction of a database suitable for ML regression to predict the critical temperature. In particular, the SuperCon database [18] collects both inorganic (under the

class "Oxide and Metallic") and organic materials (under the class "Organic"). Only the entire subset of inorganic compounds was considered, consisting of ~33,000 entries, of which ~7000 have no T_c; for those latter compounds, $T_c = 0\,\mathrm{K}$ was assumed. All materials whose formulae contain symbols like '–', '+', ',', strings like 'X', 'Z', 'z' when not included in meaningful elements symbols (e.g., 'Zn'), and with $T_c > 150\,\mathrm{K}$, were dropped. Those exclusions resulted in a reduction of the number of compounds to ~26,000. Furthermore, after normalizing the formulae stoichiometry, the same approach explained by Stanev et al. [13] was used for dealing with the duplicates. In particular, when the same compound was reported with different T_c values, it was retained with the average critical temperature only if $\mathrm{std}(T_c) \leq 5\,\mathrm{K}$, otherwise all of its occurrences were dropped, ending up with ~16,000 unique compounds. Moreover, only the classical superconductivity was taken into account, by dropping materials with Ni, Fe, Cu and O (to avoid oxides). Finally, four outliers with $T_c > 50\,\mathrm{K}$ are dropped (see Supplementary Note 6 for details). These latter steps left ~7200 materials for classical superconductivity, of which ~6700 have $T_c < 15\,\mathrm{K}$. Those cleaning pre-processing were addressed by employing the Python Pandas package [26].

Each *brute* formula was thus converted into 145 composition-based descriptors by means of Matminer [20]. Specifically, as stated by Ward et al. [27], they include stoichiometric attributes (depending on the elements' ratios), elemental property statistics (representing mean, absolute deviation, minimum and maximum of 22 atomic properties, e.g., atomic number, atomic radii), electronic structure attributes (corresponding to the average fraction of electrons in s, p, d, f valence shells over all the elements in the compound) and ionic compound attributes (including whether it is possible to form an ionic compound assuming all elements are present in a single oxidation state).

2.2. Regression models and descriptors choice

As a second step, two different regression ML models were trained and validated models for the prediction of the critical temperature. The former is a tree-based model, allowing the exact computation of the coefficients of importance in terms of the T_c by means of the Tree SHAP interpretation algorithm [21,22]. The latter is a Deep Neural Network (DNN), allowing the computation of the gradient of the critical temperature with respect to the input features, namely $\nabla T_c(x_1, \dots, x_n)$, which is necessary for the identification of the invariant groups in the form $x_i^a x_j^b$.

Specifically, the former model is an ETR-based pipeline, with hyperparameter tuning in 5-fold cross validation, trained over the 85% of the dataset and tested over the remaining 15%. The latter is a DNN trained and validated over the 85% of the database – of which the 85% was used for the training and the remaining 15% for the validation – and tested over the remaining 15%. For further details about the regression models, please refer to Supplementary Notes 3 and 4.

2.3. Invariant groups identification procedure

For the identification of the invariant binary groups in the form $x_i^a x_j^b$ the following procedure was applied.

The critical temperature is a function of more variables, namely $T_c = T_c(x_1, \dots, x_n)$. If T_c is invariant with respect to a group of features in the form $x_i^a x_j^b$, when this group is a constant $\bar{c}$ – even varying the components x_i, x_j separately – the critical temperature does not change as well. This yields

$$a \ln(x_i) + b \ln(x_j) = c \tag{1}$$

where $c = \ln(\bar{c})$. If c is constant, $\mathrm{d}c = 0$; so, Eq. (1) can be rewritten as,

$$a \frac{\mathrm{d}x_i}{x_i} + b \frac{\mathrm{d}x_j}{x_j} = 0. \tag{2}$$

An orthogonal vector $\mathbf{n}$ to the locus of points with $\mathrm{d}c = 0$ in a point $\overline{\mathbf{x}_0} = (x_{i,0}, x_{j,0})$ has components $(a/x_{i,0}, b/x_{j,0})$, which normalized becomes the following unit vector:

$$\hat{\mathbf{n}} = \left(\frac{a}{x_{i,0}\sqrt{\left(\frac{a}{x_{i,0}}\right)^2 + \left(\frac{b}{x_{j,0}}\right)^2}}, \frac{b}{x_{j,0}\sqrt{\left(\frac{a}{x_{i,0}}\right)^2 + \left(\frac{b}{x_{j,0}}\right)^2}} \right). \tag{3}$$

The condition of invariance with respect to the group $x_i^a x_j^b$ requires that the components of the gradient $\nabla T_c(x_1, \dots, x_n)$ are aligned with $\hat{\mathbf{n}}$ in $\overline{\mathbf{x}_0}$. This yields the system

$$\begin{cases} \left(\frac{\partial \widetilde{T_c}}{\partial x_i}\right)_{\overline{\mathbf{x}_0}} - \hat{n}_1 = 0 \\ \left(\frac{\partial \widetilde{T_c}}{\partial x_j}\right)_{\overline{\mathbf{x}_0}} - \hat{n}_2 = 0 \end{cases} \tag{4}$$

where

$$\begin{aligned} \left(\frac{\partial \widetilde{T_c}}{\partial x_i}\right)_{\overline{\mathbf{x}_0}} &= \left(\frac{\partial T_c}{\partial x_i}\right)_{\overline{\mathbf{x}_0}} \left(\left(\frac{\partial T_c}{\partial x_i}\right)^2_{\overline{\mathbf{x}_0}} + \left(\frac{\partial T_c}{\partial x_j}\right)^2_{\overline{\mathbf{x}_0}} \right)^{-1/2} \\ \left(\frac{\partial \widetilde{T_c}}{\partial x_j}\right)_{\overline{\mathbf{x}_0}} &= \left(\frac{\partial T_c}{\partial x_j}\right)_{\overline{\mathbf{x}_0}} \left(\left(\frac{\partial T_c}{\partial x_i}\right)^2_{\overline{\mathbf{x}_0}} + \left(\frac{\partial T_c}{\partial x_j}\right)^2_{\overline{\mathbf{x}_0}} \right)^{-1/2} \end{aligned} \tag{5}$$

and $\hat{n}_1, \hat{n}_2$ represent the two components of the unit vector $\hat{\mathbf{n}}$. If the nonlinear system in Eq. (4) is satisfied for the same exponents $(\bar{a}, \bar{b})$ over all the domains of the variables (x_i, x_j) - namely $x_{i,\min} \leq \forall x_i \leq x_{i,\max}$ and $x_{j,\min} \leq \forall x_j \leq x_{j,\max}$ - the group $x_i^{\bar{a}} x_j^{\bar{b}}$ is an intrinsic variable.

From the practical viewpoint, this has required the computation of the gradient $\nabla T_c(x_1, \dots, x_n)$, where the function $T_c(x_1, \dots, x_n)$ is represented by the DNN – built by means of Tensorflow [28] – linking the critical temperature with the input features. In particular, once the network was trained and validated, an automatic differentiation was employed to compute those partial derivatives. Specifically, for getting e.g., $\left(\frac{\partial T_c}{\partial x_j}\right)$ over all the domain of the variable x_j, all the other variables $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n)$ were fixed to their average values in the original database. Finally, for each group of two features x_i and x_j, the values of a and b were computed 100 different times respectively, comparing them for getting insight of possible invariance. The above approach was tested in the Supplementary Note 2 by means of properly designed synthetic example.

2.4. QEG-based probabilistic classifier

In addition to more classical classifiers, an attempt was made to construct a maximum Shannon entropy-based probabilistic classifier employing the notion of Quasi Equilibrium Manifold as defined in [29,30] and implemented in the discrete version of the Quasi Equilibrium Grid (QEG) as discussed in [31–34]. The main idea is described below. Given a number s of important descriptors, those features from the original dataset were first discretized by means of a $s-$dimensional binning, where each descriptor accounts for a number of bins $N_1, \dots, N_s$. The aim was thus to build a probability distribution $p(x_1, \dots, x_s)$ having the same mean vector and covariance matrix of the original binned data; among the infinite distributions respecting those bounds, this methodology focuses on the one maximizing the Shannon Entropy. Given the total number of $s-$dimensional bins $N = N_1 \times \cdots \times N_s$, the general idea consists in starting with a flattened probability distribution $\mathbf{p}^0 = (p_1, \dots, p_N)^0$ and ending up with a corrected distribution, which respects the imposed constraints of mean vector and covariance matrix. The QEG guarantees that, if $\mathbf{p}^0$ lies on the surface of maximum Shannon Entropy, also any corrected distribution will lie on the same surface. For this reason, $\mathbf{p}^0$ as the uniform distribution was always chosen, where each entry is $1/N$.

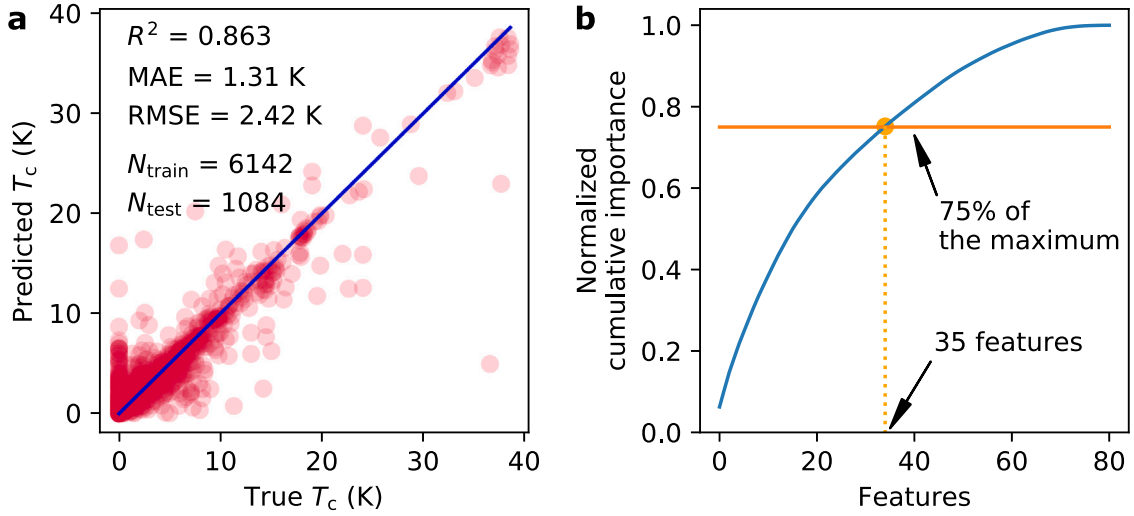

Fig. 2. Results of the ETR model. a Predictions and b corresponding normalized cumulative curve for the coefficients of importance. Model performances are shown in terms of coefficient of determination R^2, mean absolute error (MAE), and root mean squared error (RMSE), with the size of training and testing sets N_{train} and N_{test}, respectively.

Table 1
Relevant composition-based descriptors and their meaning [27].

Descriptor name	Meaning
MagpieData range MeltingT	Range of melting T over the elements of a compound
0-norm	Number of different chemical species
MagpieData mode NdUnfilled	Mode of d unfilled orbitals over the elements
MagpieData mode NsUnfilled	Mode of s unfilled orbitals over the elements
MagpieData avg_dev MeltingT	Average absolute deviation of melting T over the elements

To this end, the matrix $\mathbf{m} \in \mathbb{R}^{l\times N}$, where $l = (3s+s^2)/2$ was defined. The first s rows of $\mathbf{m}$ represent the binning of those s descriptors. The remaining $l-s$ rows represent the covariance matrix entries of those s descriptors; namely, given the integers $i,j \in [1,s]$, with $i \geq j$, the generic row of $\mathbf{m}$ among the last $l-s$ rows is the result of the element-wise product $(\mathbf{m}_i - \mu_i)(\mathbf{m}_j - \mu_j)$, where μ_i, μ_j are the means of the ith and jth descriptor respectively, while $\mathbf{m}_i, \mathbf{m}_j$ are the ith and jth rows of $\mathbf{m}$ respectively. Furthermore, the matrix $\mathbf{E} = (\mathbf{m}, \mathbf{1})^\top$ was defined, where $\mathbf{1} = (1,\dots,1) \in \mathbb{R}^N$ represents the normalization condition for probability. Let denote the null space of $\mathbf{1}$ with $\boldsymbol{\rho} \in \mathbb{R}^{N\times(N-1)}$ and the null space of $\mathbf{E}$ with $\mathbf{t} \in \mathbb{R}^{N\times(N-l-1)}$. A square matrix $\mathbf{A} \in \mathbb{R}^{(N-1)\times(N-1)}$ and a vector $\mathbf{b} \in \mathbb{R}^{N-1}$ were constructed. For the first $N-l-1$ rows, the generic elements correspond to

$$\begin{aligned} A_{ij} &= \left\langle \mathbf{t}_i, \langle \mathrm{diag}(-1/\mathbf{p}), \boldsymbol{\rho}_j \rangle \right\rangle \\ b_i &= \langle (1+\ln(\mathbf{p})), \mathbf{t}_i \rangle \end{aligned} \tag{6}$$

while for the remaining l rows they are

$$\begin{aligned} A_{ij} &= \left\langle \boldsymbol{\rho}_j, \mathbf{m}_i \right\rangle \\ b_i &= 0 \end{aligned} \tag{7}$$

where $\mathbf{t}_i$, $\boldsymbol{\rho}_j$ and $\mathbf{m}_i$ are the ith column of $\mathbf{t}$, the jth column of $\boldsymbol{\rho}$ and the ith row of $\mathbf{m}$ respectively, $\mathbf{p}$ represents the flattened probability distribution at the current iteration step, $\langle\cdot,\cdot\rangle$ denotes the dot product.

The correction procedure for the ith bound is carried out as follows: (i) the starting point is computed as $\langle \mathbf{m}_i, \mathbf{p} \rangle$, (ii) the desired value is computed as $\langle \mathbf{m}_i, \widetilde{\mathbf{p}} \rangle$, where the jth entry of $\widetilde{\mathbf{p}}$ is the number of items belonging to the jth s-dimensional bin over the total number of items (namely, the frequency), (iii) the resulting residual is filled by solving the system $\mathbf{A}^k\mathbf{p}^{k+1} = \mathbf{b}^k$ iteratively, by replacing time by time $b_{N-l-1+i}$ with a correction step ε, where $\mathbf{p}^{k+1} = \mathbf{p}^k + \delta\mathbf{p}^k$ and $\delta\mathbf{p}^k$ represents the correction resulting from the kth iteration, (iv) when the correction over the ith bound is complete, the correction over the $i+1$th bound can start, by imposing $b_{N-l-1+i} = 0$ and $b_{N-l-1+i+1} = \varepsilon$.

3. Results and discussion

As mentioned above, and in line with others in the literature [13–17], a convenient source of data was adopted, namely the SuperCon database [18] which collects the values of critical temperatures T_c for superconducting materials known from literature. To the best of the available knowledge, SuperCon turns out to be the largest database of its kind, from which a list of ~16,000 materials was extracted. Beyond the T_c values, the SuperCon database provides only the chemical composition of a compound. The latter info was thus converted into meaningful features by means of Matminer [20], allowing to associate the normalized brute formula of each compound with 145 composition-based descriptors (see Methods for further details).

Armed with such features, one can compare the performance of several classifiers aiming at predicting the probability for a compound to be a superconductor candidate. In this study, known classifiers are employed. In addition, a Bayesian type classifier based on the concept of Quasi-Equilibrium Manifold [29,30,33] is also investigated, as detailed above.

3.1. Models for predicting the critical temperature value

First, an Extra Trees Regressor (ETR)-based pipeline was trained and validated, with hyperparameter tuning in 5-fold cross-validation (see Supplementary Notes 4 and 7 for details). By means of the Tree SHAP algorithm [21,22], the input features were sorted in terms of their relevance with respect to the prediction of the T_c. Model performances with the corresponding cumulative importance curves of the ruling descriptors are reported in Fig. 2. During the data preprocessing routines, the trained pipeline (i.e., feature reduction by means of linear correlation analysis, descriptors variance analysis, correlation analysis with the T_c and ML with hyperparameter tuning, see Supplementary Note 4 for details) already drops a significant number of the 145 features, thus confirming that many of the initially selected descriptors do not significantly affect the chosen target property. In particular, the final model only includes 81 descriptors.

Importantly, Fig. 3 shows the SHAP rankings of the five most meaningful descriptors for the aforementioned model. Table 1 summarizes the physicochemical meaning of the identified descriptors, based on the complete list by Ward et al. [27]. The entire list of variables, together with their cumulative importance, the trained models, and the datasets on which they have been trained are publicly available online (see Data availability and Code availability).

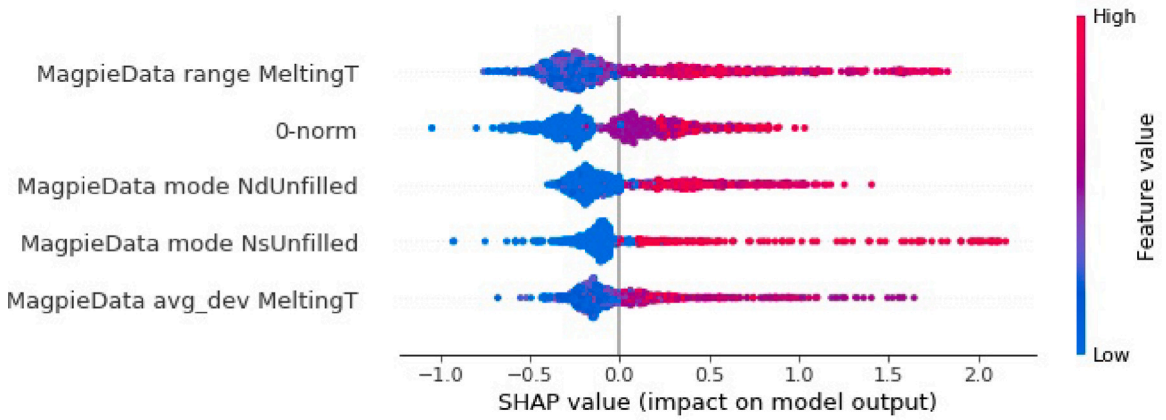

Fig. 3. The five most important features according to SHAP ranking for T_c. For each feature (i.e., each line), 1084 dots are shown, representing the entire testing sets used for computing the related SHAP values (impacts on the model output, horizontal axes); the color represents the corresponding feature value, the features are sorted according to the mean over the absolute SHAP values. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

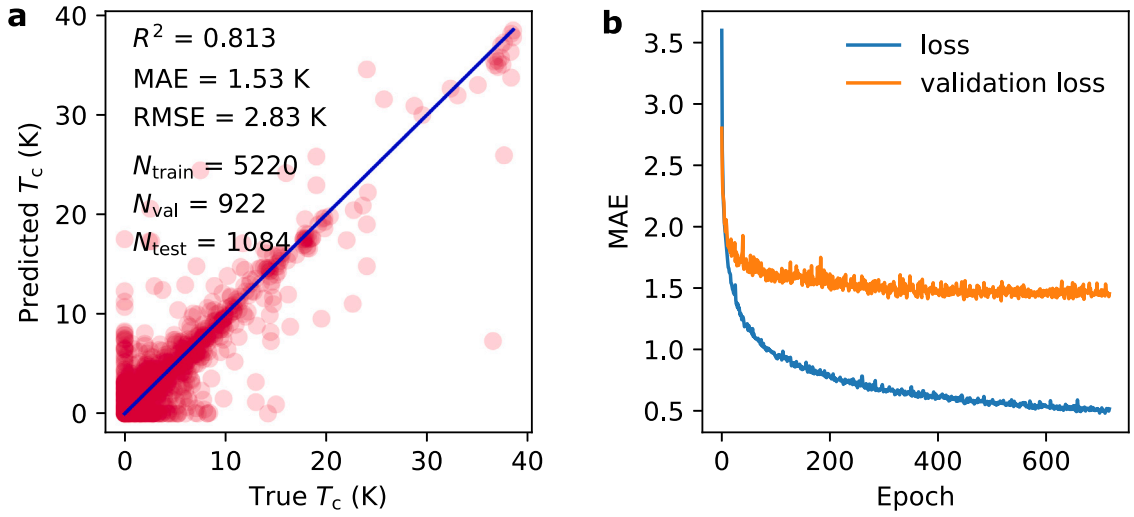

Fig. 4. Results of the DNN regression model. a Predictions over the testing set and b corresponding loss curves for the DNN regression model. Model performances are shown in terms of coefficient of determination R^2, mean absolute error (MAE), and root mean squared error (RMSE), with the sizes of the training, the validation and the testing sets, N_{train}, N_{val}, N_{test} respectively.

3.2. Invariant groups

In a first attempt of reducing the number of input features within the above models, the possible existence of symmetries of the obtained regression models was investigated. In particular, the focus has primarily been directed towards investigating the potential invariance of the target property (here the critical temperature) with respect to binary groups of the form: $x_i^a x_j^b$. This study is restricted to binary groups, although there is confidence about the generalization of the approach to groups concurrently involving a larger number of features. To this end, as discussed in the Methods section below, it is necessary to get access to the gradient of the critical temperature with respect to the input features, namely $\nabla T_c(x_1, \ldots, x_n)$. The function $T_c(x_1, \ldots, x_n)$ was thus approximated with a Deep Neural Network (DNN), which is a convenient model allowing to compute that gradient by means of automatic differentiation.

As input features of the DNN, the same 81 relevant descriptors of the above ETR-based pipeline were employed. The dataset was thus split into three parts: (i) a training set, (ii) a validation set to get insight of possible overfitting, (iii) a testing set to effectively evaluate the model performances. Fig. 4 shows the predictions over the testing set, together with the model performances and the corresponding loss with respect to the number of epochs. Specifically, no overfitting is found. More details about the DNN structure are shown in the Supplementary Note 3. Existence of possible invariant groups in the form $x_i^a x_j^b$ was looked for among the 45 different combinations of the most relevant 10 features according to the SHAP-based ranking above. On the basis of such investigations, it can be concluded that the critical temperature of the examined materials presents no invariance with respect to the tested binary groups.

3.3. Entropy-based binary classifiers

In this section, a special Bayesian type classifier is introduced and tested, as detailed below. The first two features of the SHAP ranking were considered for constructing a Shannon Entropy-based probabilistic classifier according to the Quasi Equilibrium Grid (QEG)-based procedure reported above in the Methods. In particular, those two features were binned separately for superconductors with both $T_c < 15\,\mathrm{K}$ (class 0) and $T_c \geq 15\,\mathrm{K}$ (class 1) among the 85% of the materials – namely, the training set – thus obtaining the pair of 2 dimensional binnings in Fig. 5a. For each of those binnings, the five needed constraints were computed, namely the means of those two features and their three variance terms (see Methods). A surface of maximum Shannon entropy was thus constructed for each of the two classes by means of the QEG algorithm, as depicted in Fig. 5b.

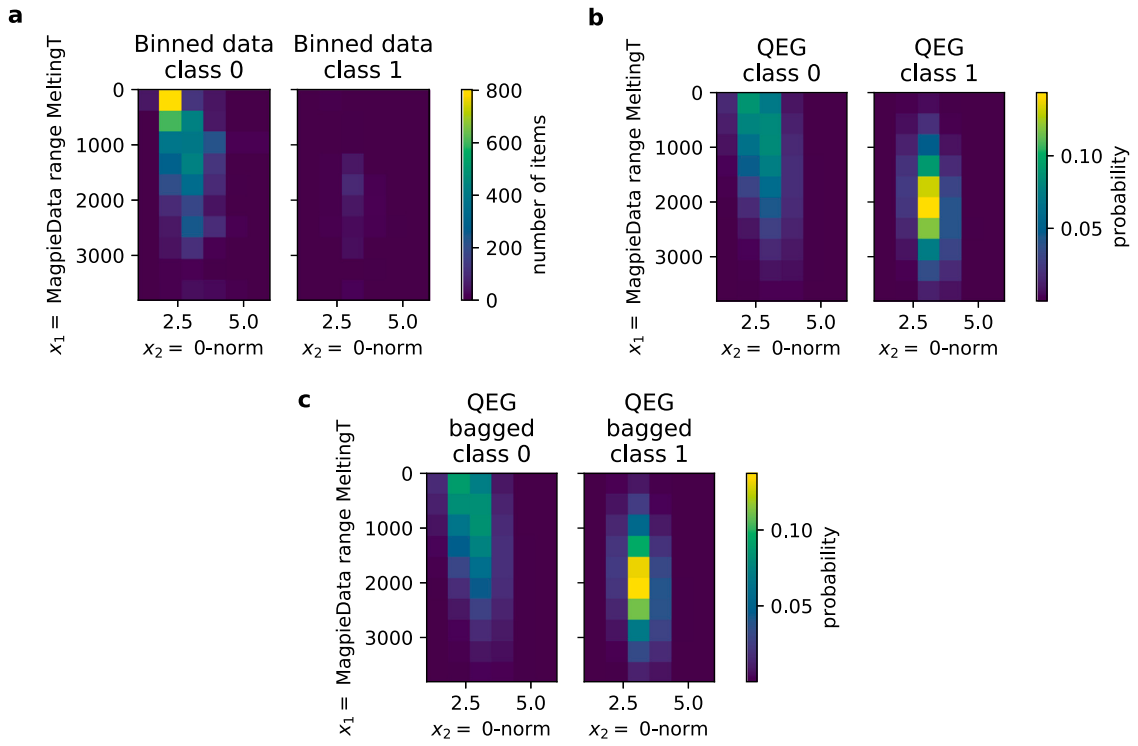

Fig. 5. Probabilistic classifier. a 2-dimensional binning, with 10 bins for the first variable and 5 bins for the second, of the two most relevant features x_1, x_2 according to the SHAP ranking for superconductors showing $T_c < 15\,\mathrm{K}$ and $T_c \geq 15\,\mathrm{K}$ respectively among the training set (namely, 85% of materials); b QEG solution of corresponding maximum Shannon entropy probability distribution; c QEG solution of corresponding maximum Shannon entropy probability distribution, bagged case.

Finally, probability distribution was computed by subtracting the QEG solution for class 0 from the QEG solution for class 1 – both multiplied by the cardinality of the corresponding class in the training set – and up-shifting the result by the minimum, in such a way to have probabilities ≥ 0. The latter distribution represents the 2 dimensional QEG probabilistic classifier. Moreover, having in mind the idea of Random Forests, which employ *bagging* (creation of more decision trees and aggregation of the results by taking the mean) [35], 100 QEG 2D distributions per class were produced, each based on a different random subset of training set. A mean distribution per class was taken; Fig. 5c shows that the bagged results are in accordance with the *non-bagged* case of Fig. 5b.

The same procedure was repeated taking into account the first three features according to the SHAP-based ranking above — namely, the range of the melting temperature, the number of different chemical species, the mode of d unfilled orbitals. A 3-dimensional binning of dimensions $10 \times 6 \times 10$ can be represented as an ensemble of ten 2-dimensional binnings, each of dimensions 6×10. Fig. 6a shows such discretization of the data, where ten matrices of axes x_2, x_3 act for the ten bins of feature x_1, from bin $x_1^{(1)}$ to $x_1^{(10)}$ for each of the two classes. Fig. 6b shows the corresponding the QEG solutions for classes 0 and 1 separately.

3.4. Other standard binary classifiers

Furthermore, two Extra Trees Classifier (ETC) models were trained and validated with default hyperparameters over the same training set accounting for the 85% of materials, including only the first two and the first three features by the aforementioned SHAP ranking respectively. Moreover, the entire dataset was used – with all the features – to train and validate two further ETC-based pipelines, both with pre-processing and hyperparameter tuning in stratified 5-fold cross validation (see Supplementary Note 4 for details). Specifically, since the cardinalities of two classes are unbalanced, the Synthetic Minority Over-Sampling TEchnique (SMOTE) algorithm was employed in one of the two pipelines, which, through interpolation, produces samples in the underrepresented class [36].

In particular, the Scikit-learn Python package [37] offers the possibility of predicting not only the class, but also the class probabilities; the predicted class is automatically chosen to be the one accounting for the highest probability. Hence, by considering only the probabilities of the class 1, i.e., the material is predicted to be superconductive, the discriminating threshold was moved from 0 (all the materials are predicted in class 1) to 1 (all the materials are predicted in class 0). For each threshold, a different confusion matrix, with different number of true positives (TP), false negatives (FN), false positives (FP), true negatives (TN), was constructed. For each confusion matrix, the true positive rate (TPR) and the false positive rate (FPR) were computed, where TPR = TP/(TP+FN) and FPR = FP/(FP+TN). The same procedure was repeated for the QEG based probabilistic classifiers, where the order of magnitude of the thresholds is lower, since the probability does not sum up to 1 over two classes but over 60 (QEG 2D) or 600 bins (QEG 3D).

Supplementary Note 8 reports a comprehensive comparison of the Receiver Operating Characteristic (ROC) curves for the all the classifiers.

The performance of a classifier can be measured by means of the Area Under Curve (AUC) of the ROC: the larger the AUC, the better the classifier. Furthermore, given a ROC curve, its best discriminating threshold ξ - above which a sample is classified as 1 and below which is classified as 0 - can be identified by means of the Youden's statistics, maximizing the quantity $J = \mathrm{TPR} - \mathrm{FPR}$ [38]. Another metric for choosing the best threshold is the maximization of the F_1 score, by definition $F_1 = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$ [39–41].

Performances computed over the same testing set of 1084 materials are shown in Table 2. The comparison encompasses QEGs with two and three features (QEG 2D and QEG 3D respectively), ETCs with the top two and three features of the SHAP ranking (ETC 2D-high and ETC 3D-high respectively), ETCs with the two (33rd, 34th) and three (33rd, 34th, 35th) features of the SHAP ranking (ETC 2D-middle and ETC

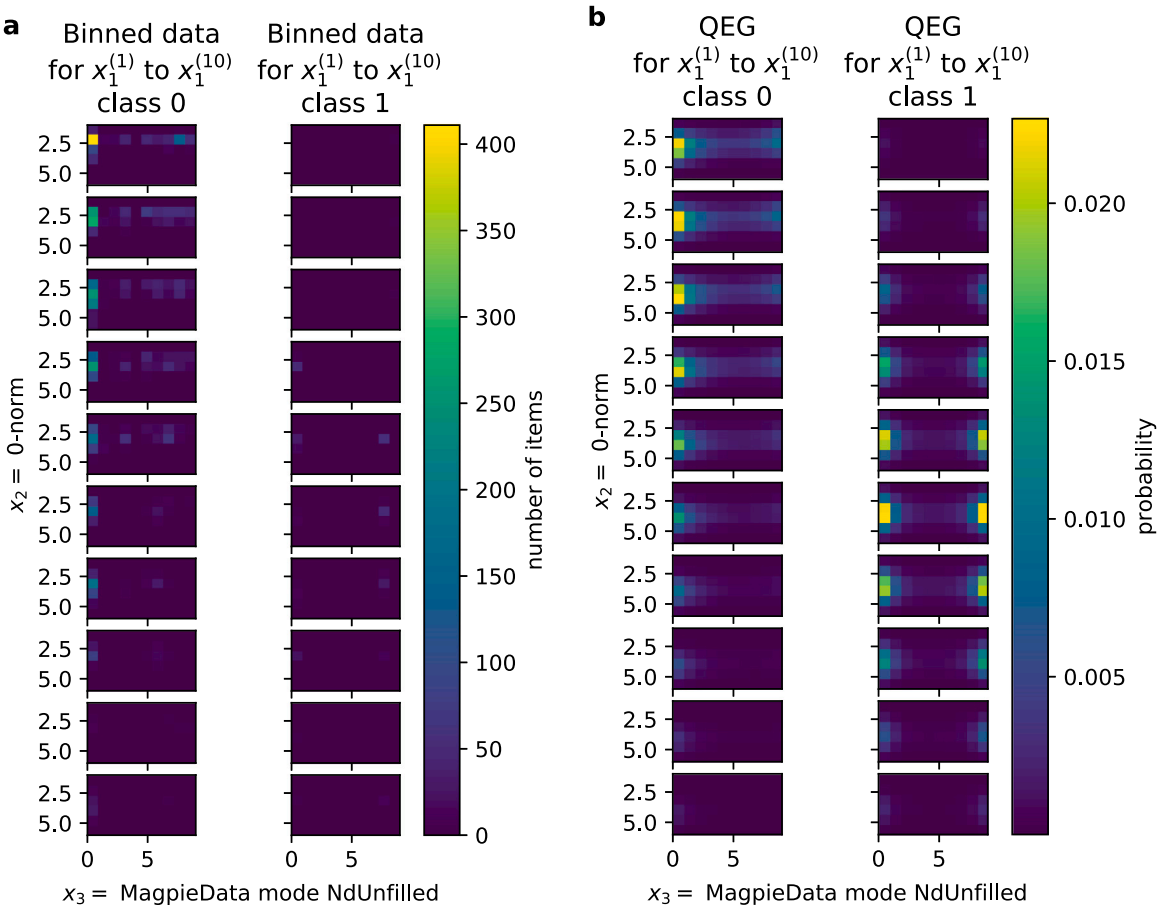

Fig. 6. Probabilistic classifier. a 3-dimensional binning, with 10 bins for the first variable, 5 for the second, 10 for the third, of the two most relevant features x_1, x_2, x_3 according to the SHAP ranking for superconductors showing $T_c < 15\,\mathrm{K}$ and $T_c \geq 15\,\mathrm{K}$ respectively; b QEG solution of corresponding maximum Shannon entropy probability distribution.

Table 2
Performances of the trained classifiers.

	AUC	$\xi_{J,\max}$	$J_{\max}$	$\xi_{F_1,\max}$	$F_{1,\max}$
No skill	0.50	–	–	–	–
QEG 2D	0.71	0.004	0.40	0.017	0.23
QEG 2D bagged	0.71	0.004	0.40	0.017	0.23
QEG 3D	0.60	0.002	0.29	0.002	0.22
ETC 2D-high	0.96	0.120	0.90	0.313	0.73
ETC 3D-high	0.96	0.125	0.89	0.333	0.73
ETC 2D-middle	0.86	0.028	0.63	0.317	0.38
ETC 3D-middle	0.82	0.167	0.62	0.167	0.52
ETC 2D-low	0.54	0.072	0.08	0.072	0.14
ETC 3D-low	0.54	0.073	0.08	0.073	0.14
ETC-vanilla	0.99	0.110	0.91	0.560	0.85
ETC-SMOTE	0.98	0.216	0.92	0.780	0.83
ETC-vanilla-81	0.98	0.040	0.91	0.630	0.84
ETC-SMOTE-81	0.99	0.140	0.91	0.732	0.84
Naive 2D	0.85	0.071	0.73	0.134	0.37

3D-middle respectively), with the least two and three features of the SHAP ranking (ETC 2D-low and ETC 3D-low respectively), ETC with all the database (ETC-vanilla), ETC with the additional SMOTE algorithm (ETC-SMOTE), ETC with all the database and all the 81 features (ETC-vanilla-81), ETC with the additional SMOTE algorithm and all the 81 features (ETC-SMOTE-81), Gaussian Naive Bayesian classifier (Naive 2D, see Supplementary Note 10 and Ref. [42] for details), together with a *No skill* classifier, in which TPR and FPR are always equal. ETC models always outperform QEG-based classifiers both in terms of $J_{\max}$ and in terms of $F_{1,\max}$; in particular, the ETC-vanilla and ETC-SMOTE turn out to be the best classifiers in terms of $F_{1,\max}$ and $J_{\max}$ respectively.

The probability of classes 0 and 1 was thus predicted with ETC-vanilla and ETC-SMOTE for all the ~40,000 materials in Materials Project without Ni, Fe, Cu, O and not in the SuperCon database. Those predictions are publicly available on GitHub repository related to this work (see Code availability).

3.5. Optimal reduction of the composition-based material descriptors

Although the above SHAP analysis can be conveniently adopted while ranking and reducing the number of material descriptors for both regressors and classifiers, the following aspects have to be stressed. On one hand, as visible on the right-hand side of Fig. 2, for achieving a sufficiently high (i.e. in the order of 70% or higher) cumulative importance over 30 features are needed. On the other hand, the larger the number of feature the higher the over-fitting possibility. Therefore, this work attempts the following possible reduction of the material descriptors. Given the original set of n features $(x_1, \ldots, x_n)$, let $(\tilde{x}_1, \ldots, \tilde{x}_n)$ be the corresponding dimensionless quantities:

$$\tilde{x}_i = \frac{x_i - x_{i,\min}}{x_{i,\max} - x_{i,\min}} + 1 \tag{8}$$

where $x_{i,\min}$ and $x_{i,\max}$ represent the minimum and maximum observed values for the ith feature over the training set, respectively. All dimensionless quantities were thus normalized by construction to a value range within the interval $[1-2]$ to avoid singularities in the expressions below.

The following new set of $m \ll n$ mixed features $(y_1, \ldots, y_m)$ is defined, as follows:

$$y_j = \prod_{i=1}^{n} \tilde{x}_i^{\alpha_{ij}} \tag{9}$$

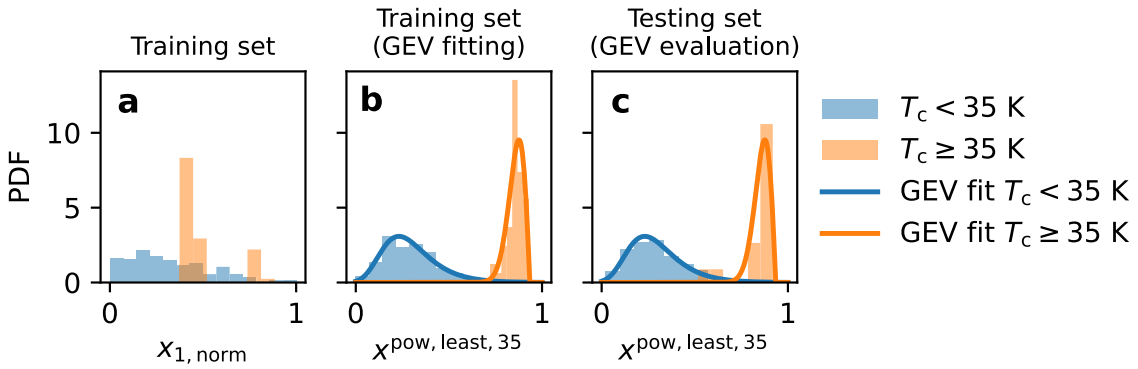

Fig. 7. One-dimensional example. a: PDFs over binned data of the training set for the two classes ($T_c < 35$ K and $T_c \geq 35$ K) reported against the normalized first most important feature according to the SHAP ranking. b: PDFs over binned data of the training set for the two classes reported against the mixed feature $x^{\mathrm{pow,least,35}}$, constructed according to Eq. (9) and choosing the point of the Pareto front with the least overlapping of the two classes according to the Battacharyya distance, together with a GEV analytical fitting of those two binnings (see text for details). c: PDFs over binned data of the testing set for the two classes reported against the same mixed feature $x^{\mathrm{pow,least,35}}$ together with the same GEV fittings of the b subfigure.

where $\{\alpha_{ij}\}$ represents an $n \times m$ matrix optimally estimated as reported below. Alternatively, the new set of m reduced mixed features can also be defined by the following linear transformation:

$$y_j = \sum_{i=1}^{n} \alpha_{ij} \tilde{x}_i \tag{10}$$

Finally, the new variables y_j can be conveniently normalized within the interval $[0 - 1]$ as follows:

$$\tilde{y}_j = \frac{y_j - y_{j,\min}}{y_{j,\max} - y_{j,\min}} \tag{11}$$

With the basic idea of Bayesian classification in mind, the following multi-objective optimization criterion can be defined. The matrix $\{\alpha_{ij}\}$ in Eq. (9) and/or Eq. (10) lies on the Pareto front while concurrently attempting: (i) maximization of a properly chosen distance between the two classes; (ii) minimization of a norm of the covariance matrix of the first class distribution; (iii) minimization of a norm of the covariance matrix of the second class distribution.

In this study, genetic algorithms were used for optimization. Moreover, for the evaluation of the distance between the two classes, a number of approaches have been tested including:

- Data in the two classes are equally binned and histograms used to evaluated the Bhattacharyya distance [43,44] between the two classes to be maximized during the above multi-objective optimization;
- Data in the two classes are equally binned and histograms used to evaluate the Earth mover distance [45] between the two classes to be maximized during the above multi-objective optimization;
- The average number of neighbors within a fixed radius of non superconducting materials to each sample of the superconducting material class in the reduced space to be minimized during the above multi-objective optimization

Finally, for the remaining two objective functions, while for one-dimensional cases a numerical estimate of the standard deviation of the binned data in the two classes is computed, in the two (or higher) dimensional cases the determinant of the covariance matrix can be adopted. More details about Pareto front calculations can be found in Supplementary Note 9.

3.5.1. Application to one- and two-dimensional cases

As an example, Fig. 7 shows the Probability Density Functions (PDFs) of the two material classes $T_c < 35$ K and $T_c \geq 35$ K. Specifically, Fig. 7a reports the PDF binning of the training set data over the two classes, against the normalized most important feature according to the SHAP ranking. Fig. 7b shows the same PDFs against the mixed feature $x^{\mathrm{pow,least,35}}$, constructed according to Eq. (9) by power combination of the 30 most important features of the SHAP ranking and choosing the point of the Pareto front with the least distributions overlap according to the Bhattacharyya distance.

Interestingly, when plotted against the new mixed feature, the two classes appear well separated, whereas it is worth observing that the same two classes show a higher degree of overlapping when reported against the first SHAP feature. As a result, it appears particularly convenient to attempt an analytical bet-fitting of the two functions reported Fig. 7b, approximated by a Generalized Extreme Value (GEV) distribution, whose density has equation

$$g\left(x^{\mathrm{pow,least,35}}\right) = \frac{1}{\sigma}\left(1 + \zeta \frac{x^{\mathrm{pow,least,35}} - \gamma}{\sigma}\right)^{-\frac{\zeta+1}{\zeta}} \times \exp\left(-\left(1 + \zeta \frac{x^{\mathrm{pow,least,35}} - \gamma}{\sigma}\right)^{-1/\zeta}\right). \tag{12}$$

In this specific case, the GEV distribution for materials with $T_c < 35$ K turns out to have factors $\gamma = 0.228$, $\sigma = 0.119$, $\zeta = -0.033$. Analogously the GEV distribution for materials with $T_c \geq 35$ K turns out to have factors $\gamma = 0.847$, $\sigma = 0.046$, $\zeta = -0.539$. Such fittings were performed by means of the SciPy Python package [46]. Fig. 7c shows the PDFs over the binned data of the testing set reported against the same mixed feature, together with the GEV fittings computed on the training set. It is worth noticing that the classes are still well separated, with a good agreement between the GEV distributions and the testing set densities. The number of bins has been chosen separately for the two classes, according to the Sturges rule [47].

Furthermore, Fig. 8 shows the PDFs of the same two material classes ($T_c < 35$ K and $T_c \geq 35$ K) in a two dimensional case. Specifically, Fig. 8a reports the PDF two dimensional binning of the training set data over the two classes, against the normalized two most important features according to the SHAP ranking. Fig. 8b shows the same PDFs against the mixed features $x_1^{\mathrm{pow,uto,35}}, x_2^{\mathrm{pow,uto,35}}$, constructed according to Eq. (9) by power combination of the 52 most important features of the SHAP ranking and choosing the Utopia point of the Pareto front. As in the one dimensional case, the two classes, when plotted against the new mixed features, appear well separated. Fig. 8c shows the PDFs over the binned data of the testing set reported against the same mixed features; the two classes are still well separated. Each plot of Fig. 8 accounts for 400 two dimensional bins, on a grid 20×20. Moreover, Supplementary Note 11 shows a sharp improvement of a Naive Gaussian Bayesian classifier trained with the mixed features $x_1^{\mathrm{pow,uto,35}}$, $x_2^{\mathrm{pow,uto,35}}$ with respect to an analogous model trained with the two most relevant features according to the SHAP ranking.

All the relevant data of the Pareto fronts used for constructing those mixed features, together with the coefficients α_{ij} of each case, are publicly available on the GitHub repository related to this work (see Code availability).

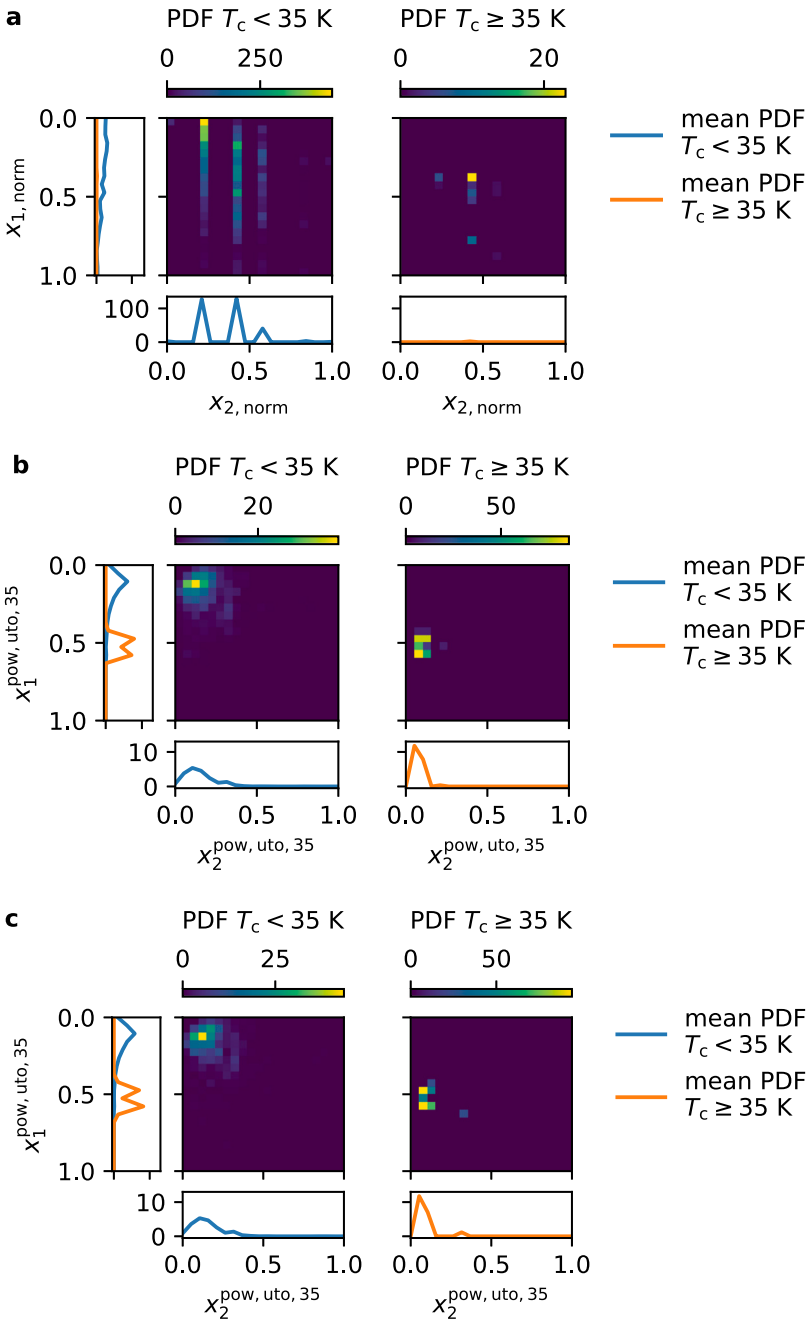

Fig. 8. Two dimensional example. a: PDFs over binned data of the training set for the two classes ($T_c < 35$ K and $T_c \geq 35$ K) reported against the normalized first most important feature according to the SHAP ranking. b: PDFs over binned data of the training set for the two classes reported against the two mixed features $x_1^{\text{pow,uto,35}}$ and $x_2^{\text{pow,uto,35}}$, constructed according to Eq. (9) from mixing the 52 most important features according the SHAP ranking and choosing the Utopia point of the Pareto front. c: PDFs over binned data of the testing set for the two classes reported against the same mixed features $x_1^{\text{pow,uto,35}}$ and $x_2^{\text{pow,uto,35}}$.

3.6. Possible generalizations

The authors are conscious that the mixed features found in this work might be still sub-optimal, as herein there is not the ambition of comprehensively exploring all possible cases. Clearly, several generalizations and variations can be studied while performing the material descriptor reduction as discussed above. Obvious generalizations might adopt different functions for reducing variables as compared to Eqs. (9) and (10), as well as different distance functions between the classes. Alternatively, other strategies for constructing optimal mixed features might also focus on distances only between classes thus neglecting minimization of variance terms, with the primary aim being the best separation between classes. In this respect, the following examples are reported:

- The training dataset is split in two classes (i.e. materials with a critical temperature above or below a certain threshold value) and a single objective optimization is performed only aiming at maximizing the distance between two classes (see Fig. 9a);
- The training dataset is split in multiple classes (i.e. > 2) and a multi-objective optimization is performed aiming at concurrently maximizing the pairwise distances between the classes (see Fig. 9b).

For further details, please refer to Supplementary Note 9.

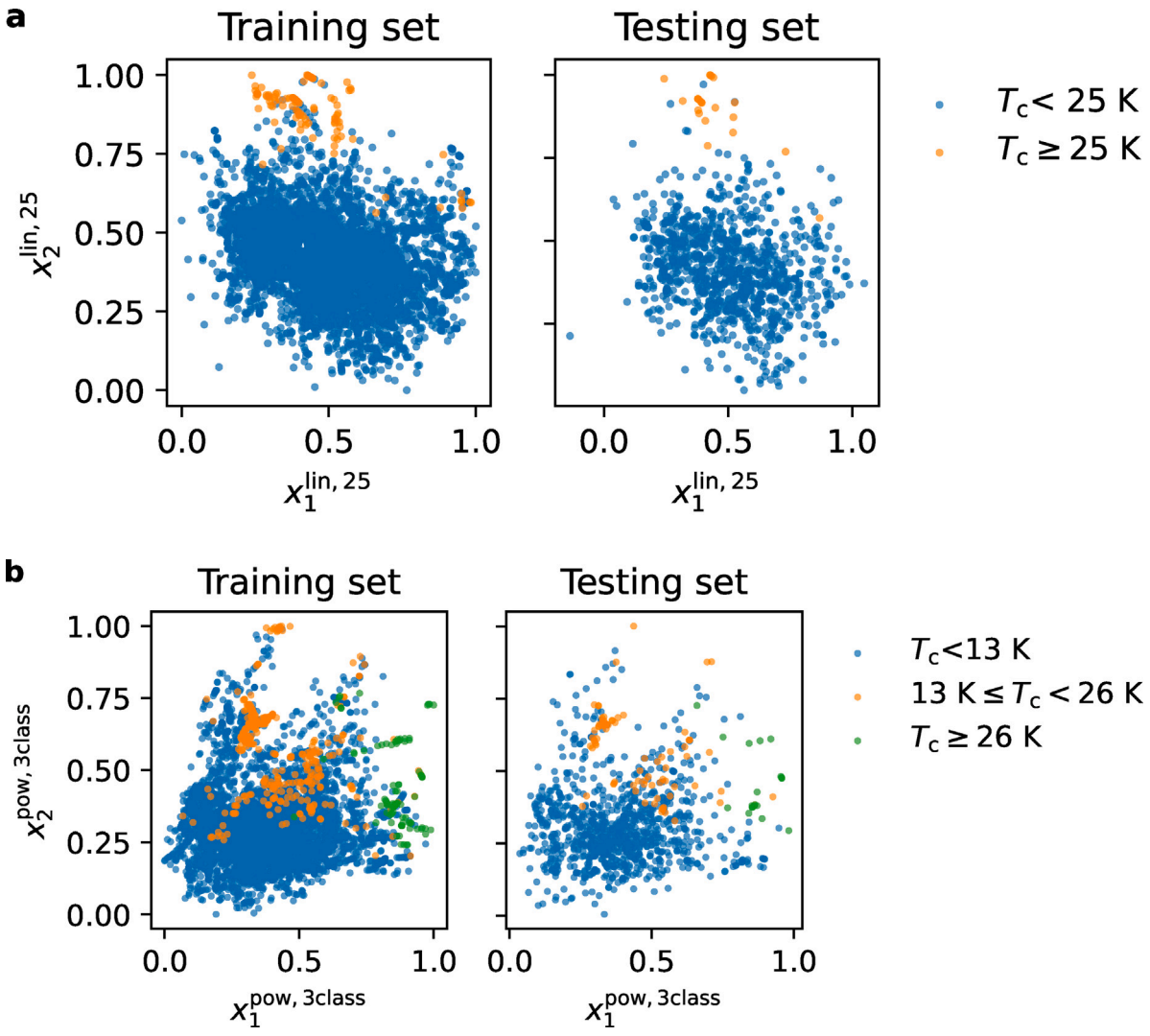

Fig. 9. Projections of training and testing sets into the reduced feature space with colors indicating the critical temperature classes. a Projection over the two mixed features $x_1^{\text{lin,25}}$, $x_2^{\text{lin,25}}$, constructed according to Eq. (10) and obtained by single objective optimization, where the Bhattacharyya distance between the two classes $T_c < 25$ K and $T_c \geq 25$ K has been maximized. b Projection over the two mixed features $x_1^{\text{pow,3class}}$, $x_2^{\text{pow,3class}}$, constructed according to Eq. (9) and obtained by multi-objective optimization where the Bhattacharyya pairwise distances between the three classes $T_c < 13$ K, $13\ \text{K} \leq T_c < 26$ K, $T_c \geq 26$ K have been concurrently maximized. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.7. Entropy-, tree-, and Bayes-based binary classifiers on the new mixed features

The results are reported for the QEG-based probabilistic classifiers and ETCs by employing the new mixed features, always constructed by aggregating the top 30 features of the SHAP ranking, in both the cases of power (Eq. (9)) and linear (Eq. (10)) transformations. In these examples, for the purposes of optimization and class separation, only the Utopia point of the Pareto front and the Bhattacharyya distance are considered respectively.

Specifically, Fig. 10a and b show the binnings of the two classes ($T_c < 15$ K and $T_c \geq 15$ K) against the two power mixed features $x_1^{\text{pow,uto,15}}$, $x_2^{\text{pow,uto,15}}$ and the corresponding QEG solution respectively. Analogously, Fig. 11a and b show the binnings of the two classes ($T_c < 15$ K and $T_c \geq 15$ K) against the two linear mixed features $x_1^{\text{lin,uto,15}}$, $x_2^{\text{lin,uto,15}}$ and the corresponding QEG solution respectively. Table 3 reports the performances of such classifiers, ending up with a consistent improvement of both the $J_{\max}$ and the $F_{1,\max}$ score with respect to the case of the QEG 2D trained with the top SHAP descriptors (see QEG 2D in Table 2); specifically, the linear transformation improves also the AUC. The same mixed features were employed to train and validate two ETCs, ending up with similar metrics – AUC, $J_{\max}$ and $F_{1,\max}$ – score with respect to the case ETC 2D-high, trained with the two most relevant features according to the SHAP ranking. The same features were finally employed to re-train also the Gaussian Bayesian Classifier, getting an improvement + for all the metrics (namely, AUC, $J_{\max}$ and $F_{1,\max}$) with respect to the same classifier trained with the top two features according to the SHAP ranking, both for the power transformation and for the linear transformation.

Table 3
Performances of the classifiers trained with mixed features.

	AUC	$\xi_{J,\max}$	$J_{\max}$	$\xi_{F_1,\max}$	$F_{1,\max}$
No skill	0.50	–	–	–	–
QEG 2D-mixed pow	0.69	0.007	0.52	0.007	0.29
QEG 2D-mixed lin	0.79	0.008	0.61	0.011	0.37
ETC 2D-mixed pow	0.95	0.104	0.82	0.480	0.72
ETC 2D-mixed lin	0.93	0.092	0.77	0.574	0.70
Naive 2D-mixed pow	0.94	0.506	0.76	0.504	0.67
Naive 2D-mixed lin	0.90	0.502	0.78	0.505	0.50

The corresponding ROC curves are reported in Supplementary Note 8.

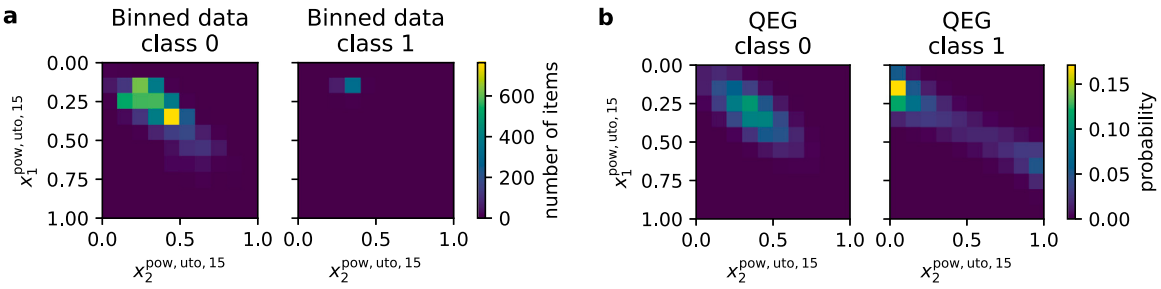

Fig. 10. Probabilistic classifier. a 2-dimensional binning, with 10 bins for each variable, of the two mixed features $x_1^{\text{pow,uto,15}}$, $x_2^{\text{pow,uto,15}}$ constructed according to Eq. (9), by selecting the utopia point of the Pareto front, for superconductors showing $T_c < 15\,\text{K}$, and $T_c \geq 15\,\text{K}$ respectively; b QEG solution of corresponding maximum Shannon entropy probability distribution.

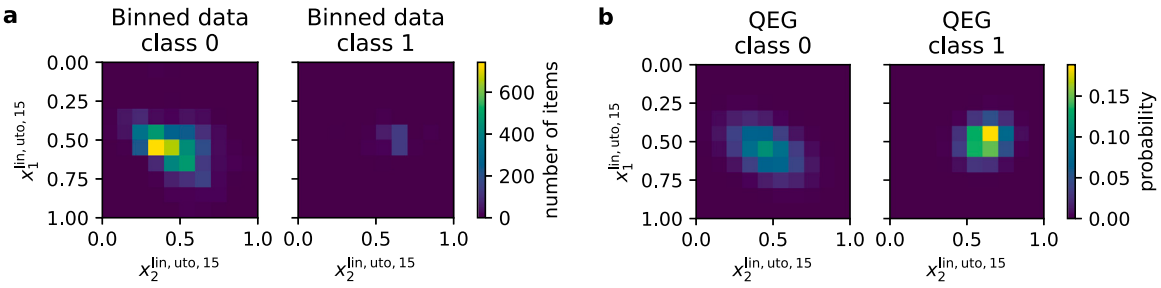

Fig. 11. Probabilistic classifier. a 2-dimensional binning, with 10 bins for each variable, of the two mixed features $x_1^{\text{lin,uto,15}}$, $x_2^{\text{lin,uto,15}}$ constructed according to Eq. (10), by selecting the utopia point of the Pareto front, for superconductors showing $T_c < 15\,\text{K}$ and $T_c \geq 15\,\text{K}$ respectively; b QEG solution of corresponding maximum Shannon entropy probability distribution.

4. Conclusions

In this work several ML tools have been developed for studying the critical temperature of superconductors. From the SuperCon database, only the inorganic compounds without Fe, Ni, Cu, O were considered, thus excluding oxides that belong to low temperature classic superconductors. By means of Matminer and on the basis of the SuperCon database, 145 composition-based features were generated for each compound. A tree-based regression model was trained and validated for the prediction of the T_c, allowing us to identify the most relevant descriptors by means of the Tree SHAP routine. Then, several different classifiers were produced, based on different sets of features and considering materials with $T_c \geq 15\,\text{K}$ in class 1 and materials with $T_c < 15\,\text{K}$ in class 0. In particular, with the idea of Bayesian classifiers in mind, a new Entropy-based classifier (here referred to as QEG) was tested, approximating the multidimensional binning of the data over the chosen set of descriptors with the surface of maximum Shannon Entropy. Other employed models include tree-based classifiers (namely ETCs) and Naive Bayesian classifiers. In particular, by comparing ETCs using only two or three of the original extracted features, the SHAP ranking – identified for regression – turns out to be consistently used for classification. Since ETCs with few features performed better than both QEGs and Naive Bayesian classifiers, two more comprehensive models - ETC-vanilla, ETC-SMOTE - were trained, both based on a number of features selected during the pre-processing routines of the respective ML pipelines. The latter uses also the SMOTE algorithm to sample, through interpolation, materials in the under-represented class of superconductors. Additionally, two further models were trained - ETC-vanilla-81 and ETC-SMOTE-81, with the same ensemble of 81 features effectively used by the regression model ETR. The best-performing models, namely ETC-vanilla and ETC-SMOTE, were employed to rank ~40,000 compounds in MaterialsProject and not occurring in the Supercon, in terms of the probability of showing $T_c \geq 15\,\text{K}$. For instance, ETC-vanilla predicts 41 of those formulae to show $T_c \geq 15\,\text{K}$ with probability not lower than 0.6. Furthermore, by means of multi-objective optimization, optimized mixed features have been found, proving to be particularly suitable for class separation. To this end, by means of power or linear combination, the top 30 features of the SHAP ranking were mixed. With such new features, the performances of both QEGs and Naive Bayesian classifiers improve, while the ETCs performances are in line with the corresponding models trained over the original features. Remarkably, in general there is no need to have access to the SHAP ranking for achieving such optimization, and, in principle, all the input features can be imported for mixing.

Additionally, further examples have been produced, differing with the previous ones in terms of threshold T_c and/or optimization routines. Among those, an optimal single feature was found to separate classes $T_c < 35\,\text{K}$ and $T_c \geq 35\,\text{K}$. Interestingly, in this case it was possible to give the equation of an analytical classifier fitted on the materials binned over new mixed feature. Both the best QEG model — QEG 2D-mixed lin classifier (for $T_c \geq 15\,\text{K}$), and the analytical classifier (for $T_c \geq 35\,\text{K}$) were employed to rank the same ~40,000 materials of MaterialsProject not occurring in the SuperCon database. Such predictions are publicly available on the GitHub repository related to this work (see Code availability).

Another aim of this work was to test the possible invariance of the critical temperature with respect to binary groups of features in the form of $x_i^a x_j^b$. To this end, a second regression model – i.e., a DNN – was trained and validated for the prediction of the T_c, allowing to compute the gradient of the critical temperature with respect to the input features, namely $\nabla T_c(x_1, \ldots, x_n)$. Finally, it is important to stress that the suggested methods in this papers, namely the search for invariant groups of regression models, the optimization of mixed composition based feature and the maximum entropy based classifiers are general and not restricted to the selected case study. Hence, potential future applications of these methods can be envisioned for other energy materials such as thermal energy storage [48] and electrochemical energy storage [49] applications.

CRediT authorship contribution statement

Giovanni Trezza: Data curation, Writing – original draft, Visualization, Investigation. **Eliodoro Chiavazzo:** Conceptualization, Methodology, Software, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eliodoro Chiavazzo reports financial support was provided by European Commission.

Data availability

Processed datasets and trained models of this study are publicly available in Zenodo at DOI:10.5281/zenodo.7725592 [50].

Acknowledgments

The authors are grateful to Nicola Marzari (École Polytechnique Fédérale de Lausanne), Samuel Poncé (Université catholique de Louvain) and Marnik Bercx (École Polytechnique Fédérale de Lausanne) for the valuable discussions about the material selection in the case study reported in this work. E.C. acknowledges partial financial funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 957189.

Code availability

The codes used to obtain the results of this study are publicly available in github at https://github.com/giotre/superconductors.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.mtcomm.2023.106579.

References

[1] R. Schmuch, R. Wagner, G. Hörpel, T. Placke, M. Winter, Performance and cost of materials for lithium-based rechargeable automotive batteries, Nat. Energy 3 (2018) 267–278.
[2] J. Hirsch, M. Maple, F. Marsiglio, Superconducting materials: conventional, unconventional and undetermined, Physica C 514 (2015) 1–444.
[3] S.C. Johnson, et al., Selecting favorable energy storage technologies for nuclear power, in: Storage and Hybridization of Nuclear Energy, Elsevier, 2019, pp. 119–175.
[4] W. Yuanxi, L. Jiangang, W. Peide, et al., First engineering commissioning of east tokamak, Plasma Sci. Technol. 8 (253) (2006).
[5] R. Aarnink, J. Overweg, Magnetic resonance imaging, a success story for superconductivity, Europhys. News 43 (2012) 26–29.
[6] Hall A., et al., Use of high temperature superconductor in a receiver coil for magnetic resonance imaging, Magn. Reson. Med. 20 (1991) 340–343.
[7] K. Asayama, Y. Kitaoka, G.-q. Zheng, K. Ishida, Nmr studies of high tc superconductors, Prog. Nucl. Magn. Reson. Spectrosc. 28 (1996) 221–253.
[8] A. Rigamonti, F. Borsa, P. Carretta, Basic aspects and main results of nmr-nqr spectroscopies in high-temperature superconductors, Rep. Progr. Phys. 61 (1367) (1998).
[9] L. Rossi, L. Bottura, Superconducting magnets for particle accelerators, Rev. Accel. Sci. Technol. 5 (2012) 51–89.
[10] J. Clarke, A.I. Braginski, The SQUID Handbook, Vol. 1, Wiley Online Library, 2004.
[11] R. Cristiano, M. Ejrnaes, A. Casaburi, N. Zen, M. Ohkubo, Superconducting nano-strip particle detectors, Supercond. Sci. Technol. 28 (2015) 124004.
[12] M. Noe, M. Steurer, High-temperature superconductor fault current limiters: concepts, applications, and development status, Supercond. Sci. Technol. 20 (2007) R15.
[13] Stanev V., et al., Machine learning modeling of superconducting critical temperature, Npj Comput. Mater. 4 (2018) 1–14.
[14] Konno T., et al., Deep learning model for finding new superconductors, Phys. Rev. B 103 (2021) 014509.
[15] T.D. Le, et al., Critical temperature prediction for a superconductor: A variational bayesian neural network approach, IEEE Trans. Appl. Supercond. 30 (2020) 1–5.
[16] B. Roter, S. Dordevic, Predicting new superconductors and their critical temperatures using machine learning, Physica C 575 (2020) 1353689.
[17] B. Roter, N. Ninkovic, S. Dordevic, Clustering superconductors using unsupervised machine learning, Physica C (2022) 1354078.
[18] M. I. S. SuperCon, National Institute of Materials Science, 2011, http://supercon.nims.go.jp/index_en.html.
[19] G. Stewart, Unconventional superconductivity, Adv. Phys. 66 (2017) 75–196.
[20] Ward L., et al., Matminer: An open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69.
[21] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Process. Syst. 30 (2017).
[22] Lundberg S. M., et al., From local explanations to global understanding with explainable ai for trees, Nat. Mach. Intell. 2 (2020) 56–67.
[23] G. Trezza, L. Bergamasco, M. Fasano, E. Chiavazzo, Minimal crystallographic descriptors of sorption properties in hypothetical mofs and role in sequential learning optimization, Npj Comput. Mater. 8 (2022) 1–14.
[24] S.-M. Udrescu, M. Tegmark, Ai feynman: A physics-inspired method for symbolic regression, Sci. Adv. 6 (2020) 2631.
[25] Jain A., et al., Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Mater. 1 (2013) 011002.
[26] W. McKinney, Python for Data Analysis: Data Wrangling with PandAs, NumPy, and IPython, O'Reilly Media, Inc, 2012.
[27] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, Npj Comput. Mater. 2 (2016) 1–7.
[28] Abadi M., et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015, URL https://www.tensorflow.org/, Software available from tensorflow.org.
[29] A.N. Gorban, Model reduction in chemical dynamics: slow invariant manifolds, singular perturbations, thermodynamic estimates, and analysis of reaction graph, Curr. Opin. Chem. Eng. 21 (2018) 48–59.
[30] A.N. Gorban, I.V. Karlin, Invariant Manifolds for Physical and Chemical Kinetics, Vol. 660, Springer, 2005.
[31] E. Chiavazzo, Invariant Manifolds and Lattice Boltzmann Method for Combustion, (Ph.D. thesis), ETH Zurich, 2009.
[32] E. Chiavazzo, I.V. Karlin, Quasi-equilibrium grid algorithm: Geometric construction for model reduction, J. Comput. Phys. 227 (2008) 5535–5560.
[33] E. Chiavazzo, I. Karlin, Adaptive simplification of complex multiscale systems, Phys. Rev. E 83 (2011) 036706.
[34] E. Chiavazzo, Approximation of slow and fast dynamics in multiscale dynamical systems by the linearized relaxation redistribution method, J. Comput. Phys. 231 (2012) 1751–1765.
[35] Hastie T., et al., Random forests, in: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2009, pp. 587–604.
[36] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artificial Intelligence Res. 16 (2002) 321–357.
[37] Pedregosa F., et al., Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[38] W.J. Youden, Index for rating diagnostic tests, Cancer 3 (1950) 32–35.
[39] N. Chinchor, Proceedings of the 4th Conference on Message UnderstandIng, Muc4'92, Association for Computational Linguistics Stroudsburg, PA, 1992.
[40] C.J. Van Rijsbergen, A theoretical basis for the use of co-occurrence data in information retrieval, J. Doc. (1977).
[41] A.A. Taha, A. Hanbury, Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool, BMC Med. Imaging 15 (2015) 1–28.
[42] H. Zhang, The optimality of naive bayes, Aa 1 (2004) 3.
[43] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, Bull. Calcutta Math. Soc. 35 (1943) 99–109.
[44] A. Bhattacharyya, On a measure of divergence between two multinomial populations, SankhyĀ (1946) 401–406.
[45] C. Villani, Optimal Transport: Old and New, Vol. 338, Springer, 2009.
[46] P. Virtanen, et al., SciPy, 1.0: Fundamental algorithms for scientific computing in python, Nature Methods 17 (2020) 261–272.
[47] H.A. Sturges, The choice of a class interval, J. Amer. Statist. Assoc. 21 (1926) 65–66.
[48] L. Aghemo, L. Lavagna, E. Chiavazzo, M. Pavese, Comparison of key performance indicators of sorbent materials for thermal energy storage with an economic focus, Energy Storage Mater. (2022).
[49] Wang Z., et al., Deep learning for ultra-fast and high precision screening of energy materials, Energy Storage Mater. 39 (2021) 45–53.
[50] G. Trezza, E. Chiavazzo, Models and datasets for optimal composition-based material descriptors: A case study on superconductor classification, 2023, http://dx.doi.org/10.5281/zenodo.7725592.