Element selection for functional materials discovery by integrated machine learning of atomic contributions to properties

Andrij Vasylenko[1,2]

[1]Department of Chemistry, University of Liverpool, L69ZD Liverpool, UK
[2]Institute for Condensed Matter Physics, National Academy of Science of Ukraine, 1, Sventsitskogo str., 79011 Lviv, Ukraine

Abstract

At the high level of consideration, the fundamental differences between the materials lie in the differences between the constituent chemical elements. Before the differences are detailed with the stoichiometric ratios and then atomic structures, the materials can be conceptualised at the level of their phase fields – the fields of all possible configurations of the selected chemical elements. In this work, we demonstrate that at this level, materials can be classified with respect to the likelihood to manifest a target functional property, while being synthetically accessible. In the proposed end-to-end machine learning approach (PhaseSelect), unsupervised learning of chemical environments derives atomic characteristics that are subsequently used for supervised classification of the phase fields with respect to their potential functional applications. PhaseSelect can investigate the materials' potential applicability at scale at the level of periodic table, which we demonstrate with significant accuracy on three classes of materials: for high temperature superconducting, magnetic applications and photovoltaics.

Introduction

Conceptualization of novel materials begins at the level of periodic table with selection of chemical elements. There is a variety of possible compositions that can be formed from a set of chemical elements; the field of this potential realizations can be defined as a material's phase

field. The choice of a phase field ultimately determines the outcome of the synthetic work and functional properties of the prospective compounds.

The fundamental differences between atomic elements and a variety of bonding interactions underlying the variance in the materials' properties have produced thousands of compositions accumulated in materials databases[1–3]. Harvesting this statistical data, there has been a surge of machine learning (ML) methods aiming to predict the materials' properties from the knowledge of their structures and compositions[4,5]. Ranging from enthalpy[6] to energy band gap[7] to superconducting transition temperature[8], ML predictions enable fast screening of functional inorganic materials at scale, overcoming the otherwise forbidding combinatorial challenge for precise, but significantly more resource-demanding high-throughput quantum-mechanical calculations.

Codification of the materials for statistical treatment involves description of the atomic elements, often represented as vectors of chemical and physical characteristics, that are combined linearly to describe a compound[6]. This approach relies on the expert selection of a number of exploited chemical characteristics as well as the relevance of these characteristics and the corresponding weights for atomic descriptions in materials representations. This selection determines quality of the model[9]. The composition-based models are predisposed to data leakage between training and validation datasets via compositionally close datapoints, that impedes extrapolation of patterns in materials-properties relationships onto the unexplored materials[10].

In this work, we demonstrate that unsupervised learning of chemical elements combined with the attention technique for learning elemental contributions can be used for an accurate classification of the materials' functional performance at the level of the phase fields. This end-to-end integrated machine learning (PhaseSelect) of the materials databases can render the materials with respect to the maximum achievable values of the properties in the phase fields.

In our approach, the machine learns the complex interplay of bonding interactions for each atom by exploring possible elemental co-occurrence in all studied materials[11], where a particular atom plays role in formation, similarly to the concept in [12]. Thus built atomic vectors are then combined linearly to form a phase field representation, whereas attention mechanism[13] is trained to magnify the most prominent atomic contributions specific to a particular phase field's property. This offers a statistically-derived alternative to the expert knowledge-based manual selection of relevant chemical characteristics and their contributions, and enables high-level classification of materials for functional applications, eliminating concerns of data leakage at the compositional level.

We demonstrate the significant accuracy of PhaseSelect in classification of the materials with respect to three different properties: superconducting transition temperature, Curie temperature, and energy band gap, when learning from SuperCon[3] and Materials Platform for Data Science (MPDS)[1] databases. In these applications, PhaseSelect demonstrates 80.4, 86.2, 75.6 % accuracy and 72.9, 84.2, 75.3% F1 score respectively. Furthermore, the phase fields representations derived during properties classification are exploited to recognise patterns in elemental combinations that afford stable compositions in material databases, and produce the ranking of synthetic accessibility for unexplored phase fields. Thus determined metrics of the phase fields - probability of achieving high value of a property and synthetic accessibility ranking - can be orthogonally applied to any combination of elements at scale, creating a map of potentially attractive phase fields that can provide guidance to human researchers in the consequential and costly choice of phase fields for investigations and discovery of functional materials.

Results and discussion

PhaseSelect model architecture

Our goal is to assess the attractiveness of candidate functional materials at the level of periodic table, thereby circumventing the combinatorial challenge of individual assessment of all possible compositions built from the chosen elements. At this level of the material's description, the relationship between elemental combinations and their synthetic accessibility have been studied with unsupervised machine learning and validated experimentally[9]. Here, we hypothesise that an integrated statistical description of atomic elements and their combinations
can recognise patterns of variability of atomic bonding in materials databases and that can reflect both upper bounds for the properties' values and stability for the compositions of a phase field. The proposed architecture of the model based on this hypothesis is illustrated in Figure 1.

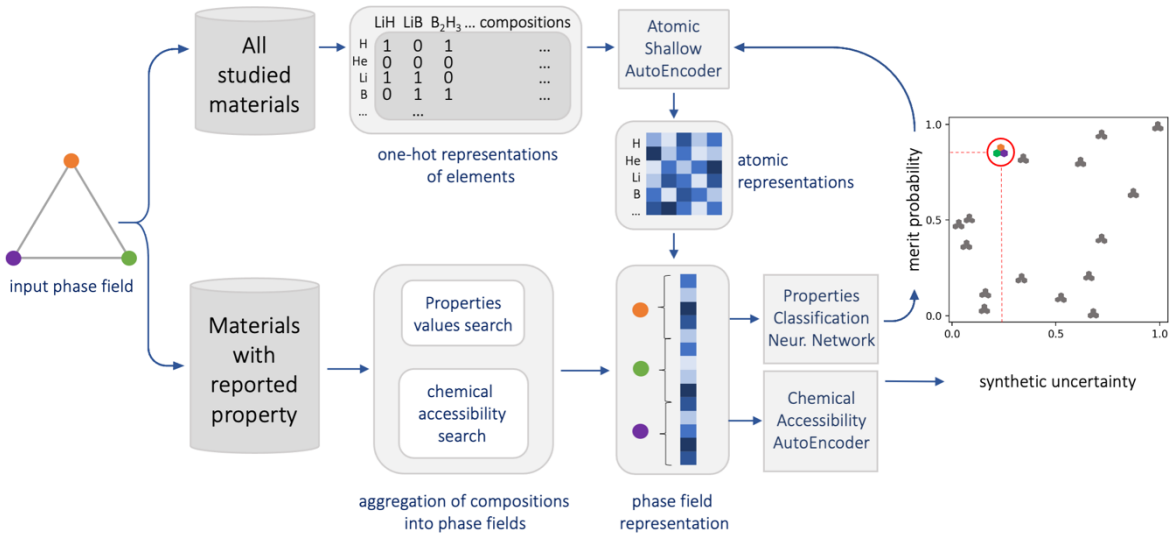


Figure 1. PhaseSelect predicts properties and chemical accessibility of phase fields. Model architecture. Arrows show the information flow between the various components described in this paper: from processing of atomic environments in all studied materials - and aggregation of data in the materials databases MPDS[1] and SuperCon[3] to learning atomic representation (top-row stream) followed by classification by properties' values and ranking of the input phase fields (bottom-row stream), resulting in a map of phase fields' likelihood to form stable compounds with desired properties.

PhaseSelect consists of several connected modules that pass information from databases to atomic representation with shallow autoencoder neural network (NN) to phase field representation with self-attention mechanism to deep NN for classification of phase fields with respect to their properties' values. The materials data is represented as collection of vectors, which are non-linearly transformed by each module via multiplication with a module's specific set of weights and biases that are derived during the training. Training of these connected modules – shallow autoencoder NN for derivation of atomic vectors, attention matrices for building phase fields representations from atomic vectors and deep NN for classification – is performed simultaneously in the end-to-end fashion, while minimising the combined loss function: the reconstruction error of atomic vectors (Euclidean distances) by autoencoder NN and binary cross-entropy error for classification by deep NN.

Aggregation of compositions into phase fields

For classification and accessibility ranking of phase fields (See bottom stream in Figure 1) we process the materials databases, where experimentally verified values of properties are reported for large number of compositions[1,3]. Compositions built from the same constituent elements are aggregated into the phase fields, with associated properties values corresponding to the maximum reported property value among the compositions within the phase fields. Aggregation of materials with reported superconducting transition temperature, Curie temperature and energy band gap has produce datasets with 4826, 4753 and 40452 phase fields respectively. Division of the datasets into two classes by the threshold values for the corresponding properties – 10 K, 300 K and 4.5 eV for superconducting transition temperature, Curie temperature and energy band gap, respectively – forms reasonably balanced data classes with 3311:1515, 2726:2027 and 20910:19690 phase fields, respectively, with data distributions illustrated in Figure 2a-c. Rapidly decreasing number of explored phase fields with reported superconducting properties at

temperatures above 10 K (See Figure 2b) proves development of reliable models for classification with respect to temperatures higher than 10 K challenging (See Supplementary Fig. 1)[8]. Nevertheless, despite the broad aggregation of high-temperature superconducting materials into a single class (with $T_c > 10$ K), accurate classification of unexplored materials into such classes would allow fast screening for novel high-temperature superconductors[8]. The chosen threshold values for Curie temperature and energy gap reflect practical interests in high-temperature magnetic materials and photovoltaics[14], and divide the datasets into classes of comparable size, convenient for high-precision classification. Furthermore, the remaining imbalances are taken into account by class-weighting in the corresponding classification NN models[15].

Across the datasets, the phase fields are formed from up to 12 constituent elements, with the majority data represented by ternary, quaternary and quinary (See Figure 2d) phase fields. To incorporate information about atomic bonding interplay from all available data, the variance in size of the phase fields is alleviated by zero-padding in the phase fields representation module that further allows extrapolation of the patterns derived from the explored materials onto the candidate phase fields of arbitrary number of elements.

Abundance of chemical elements among the explored materials in the databases is illustrated in Figure 2e. All datasets have similar trends with peaks for materials containing, e.g., carbon, oxygen, silicon, with especially pronounced match between elemental distribution in datasets with materials for superconducting and magnetic applications (See inset in Figure 2e). To learn atomic characteristics from the atomic environments – explored chemical compositions, where the atoms are found to form the variety of stable and metastable materials, we build a module for atomic representation based on a large materials database that includes both experimental and theoretical materials[11,12].

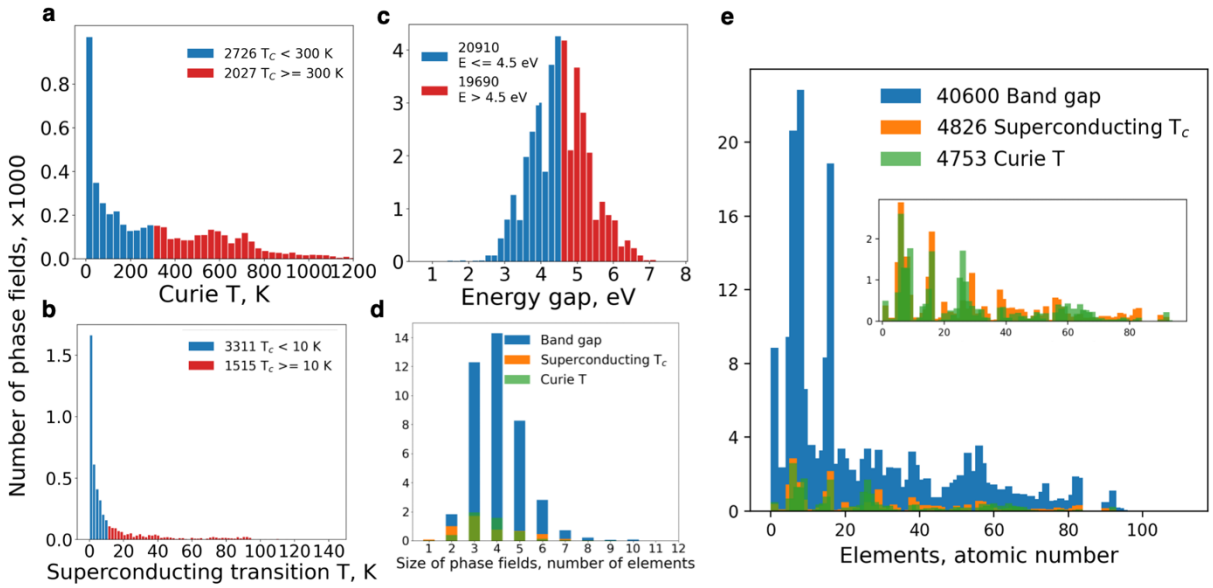


Figure 2. Aggregation of compositions into phase fields. a Distribution of phase fields of magnetic materials in MPDS[1] with respect to the maximum associated Curie temperature T_C. The materials' classes "low-temperature" and "high-temperature" magnets are divided around T_C = 300 K as 2726:2027 phase fields. **b** Distribution of phase fields of superconducting materials (joined datasets from SuperCon[3] and MPDS) with respect to the maximum associated superconducting transition temperature T_c. The materials' classes "low-temperature" and "high-temperature" superconductors are divided around T_c = 10 K as 3311:1515 phase fields. **c** Distribution of phase fields of materials with reported value of energy gap in MPDS with respect to the maximum associated band gap. The materials' classes "small-gap" and "large-gap" are divided around E = 4.5 eV as 20910:19690 phase fields. **d** Distribution of materials with respect to the number of constituent elements is similar for all datasets: the majority of the reported compositions belong to ternary, quaternary and quinary phase fields. **e** Abundance of chemical elements among the explored materials in the databases; total numbers of phase fields in the corresponding datasets are highlighted in the legend. All datasets have similar trends with pronounced peaks for materials containing, e.g. carbon, oxygen, silicon. Inset illustrates overlap in trends for elemental distribution in explored materials for superconducting and magnetic applications.

Atomic representation and phase field representation

For each chemical element one can build an one-hot encoding vector from its instances in the database. The database is expanded into a table similarly to the approach proposed in [12] (Depicted as a matrix of coexisting elements in the materials in Fig. 1). The rows of the table correspond to the chemical element, the columns are the remainders of the compositional formulas of the reported compounds – atomic environments, e.g., a composition Li_2S is represented in two rows (Li, S) and two columns: "()2S" and "Li", with ones placed at the intersections of these rows and columns. The cells of the table where chemical elements and the atomic environments do not coexist to form a material from the database are filled with zeros. The resulting sparse matrix represents population and coexistence of the chemical elements and atomic environments in the materials database. We then employ a shallow autoencoder neural network – an unsupervised ML technique – to reduce the dimensionality of this matrix, and to condense the information into the rich latent space of reduced dimensionality, in which similar vectors are grouped close to each other. We use these vectors of the latent space as atomic representations that constitute the basis for further phase fields descriptions (see Figure 3a). A comparison of the thus-derived atomic vectors with the chemical characteristics-based vectors[6] in the tasks of the supervised phase fields classification and unsupervised learning of synthetic accessibility demonstrates an increased accuracy of classification and improved trainability of the unsupervised models with the statistical representation-based approach to atomic descriptions. (See Supplementary Fig. 2).

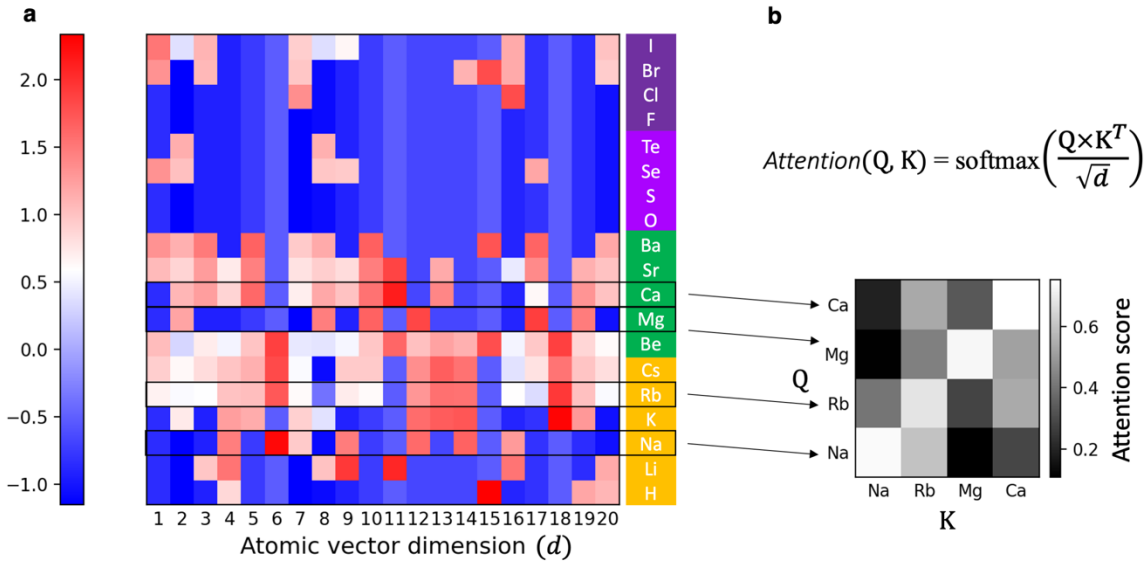


Figure 3. Atomic representations and their contributions to the phase fields' properties. a Atomic representation vectors for the 1st, 2nd, 16th and 17th atomic groups of the periodic table, illustrating differences and correlations between atomic features (vectors' components) in the neighbouring atoms and groups. The full stack of atomic vectors for the whole periodic table is extracted by PhaseSelect's atomic autoencoder shallow neural network, from the sparse matrix of chemical elements and atomic environments built for the Materials Project database[11,12]; **b** Prior to the weighted combination of atomic vectors into the phase field representation, the corresponding weights for the constituent element are calculated as the attention scores [ref], trained during the fitting of the model for phase fields classification by properties.

To emphasise differences in the contributions of individual atoms to the phase field's properties, we employ the multi-head local attention[13] that calculates the attention scores – weights for the constituent atomic vectors contributing to the accuracy of classification of a corresponding phase field's property – that are derived during the training (See Figure 3b). When building a phase field representation for downstream tasks of property classification and synthetic accessibility ranking, the phase field's atomic vectors are multiplied by their attention scores and then concatenated to form a $(n \times 20)$-dimensional vector, where n is a number of constituent elements in a phase field. The atomic contributions weights are also used for building a model for an

arbitrary number of elements in a phase field. For this, we create all phase fields representations vectors of an equal size *l*, corresponding to the largest phase field in a database, and pad the smaller phase field vectors, of size *s*, with *l* - *s* zeros, that will have zero attention weights, but will further enable formation of a neural network layer for processing of all input data in a single model. The described construction of a phase field representation with local attention weights also makes the model insensitive to the order in which atomic elements are listed in a phase field, without the need to take into account all possible permutation of the elements.

Classification by properties' values and ranking by synthetic accessibility

Classification in PhaseSelect is a deep neural network that assigns the phase fields representation vectors to the corresponding classes of the properties values. The phase fields in each dataset are divided into two classes (See Figure 2a-c) that are labelled with '1' for the phase fields with associated property values above the chosen thresholds, and with '0' for the remaining phase fields. Three different deep neural networks classifications are then trained on the collected data for superconducting materials, magnetic materials, and materials with a reported value of energy gap, respectively. Convergence of the training processes is illustrated in Supplementary Fig. 2a-c. Validation of the trained models is performed in 5-fold cross-validation, where 5 models are trained on different 80% portions of available data, with remaining 20% used for testing. The average accuracy across the validation sets is 80.4, 86.2, 75.6 % for classification with respect to Curie temperature, superconducting transition temperature, and energy gap respectively. The validation datasets are used to tune the parameters of the NN models, such as weights, biases, stochastic dropping, learning rate, activisations of neurons and early stopping. For the predictive models, we adopt all available data in the three datasets for training. Noting the stochastic nature of the machine learning NN, we employ averaging of the predicted probabilities over the ensemble of 300 models, this minimises the differences in training processes and derived

models' parameters (See Supplementary Fig. 3a). The ensemble with the minimised variance in predictions enables assessment of the materials' properties not only by the assigned binary classes, that are threshold-dependent (See Figure 4d, Supplementary Fig. 4, Supplementary Table 1), but also by the continuous values of probabilities as a measure of likelihood of achieving a desired property value.

In parallel to the classification module, a deep AutoEncoder neural network learns patterns of chemical accessibility by mapping phase fields representations to themselves, while reorganising them through the latent space, where similar phase fields are grouped. This reorganisation via the AutoEncoder enables ranking of the phase fields by their reconstruction errors, that reflect differences of individual entries from general patterns in data, therefore elemental combinations that are unlikely to manifest conventional bonding chemistry nor to form synthetically accessible compositions exhibit high reconstruction errors[9]. We also find that predicted reconstruction errors converge to their average values when an ensemble of models is trained (See Supplementary Fig. 3b).

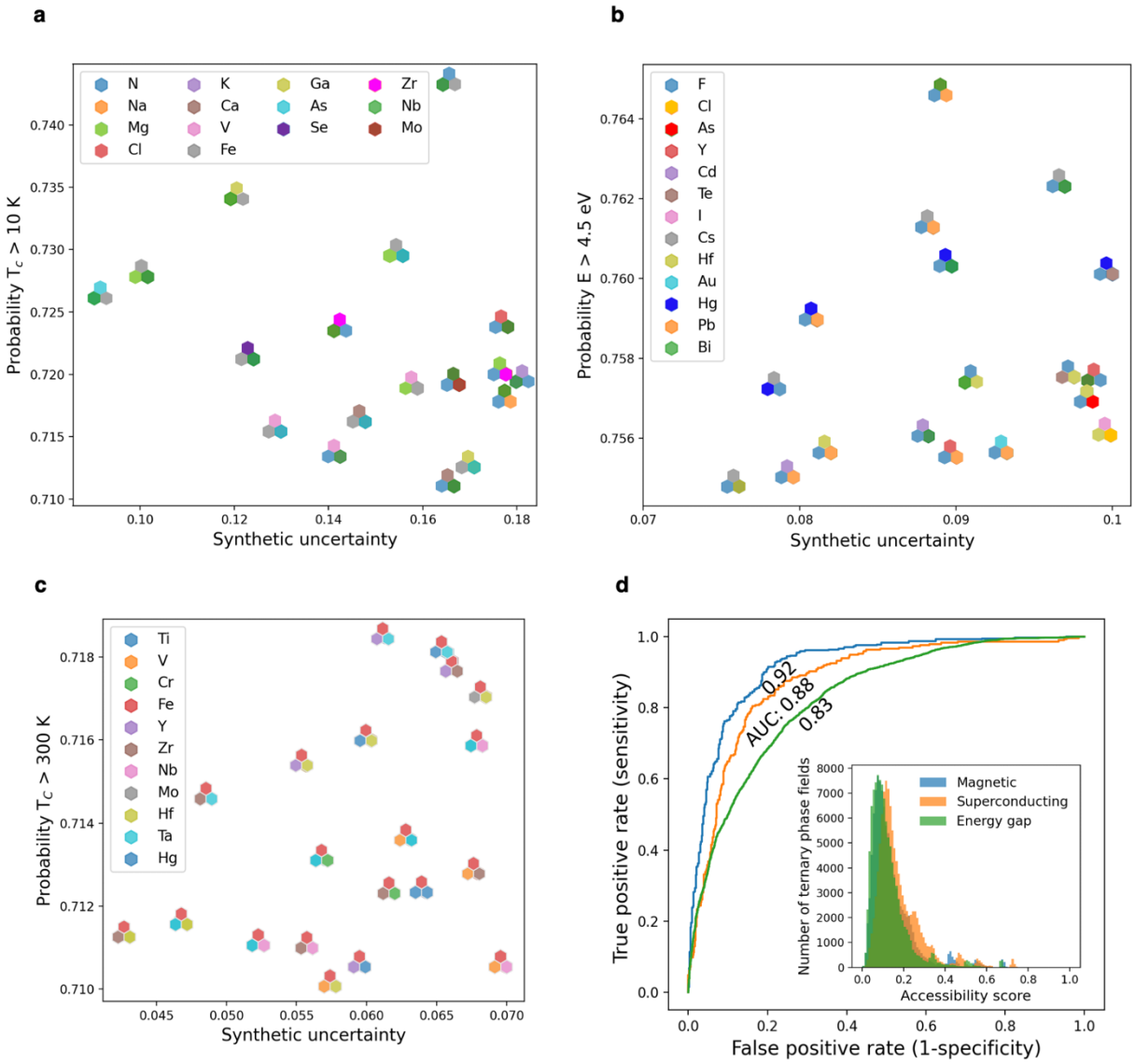


Figure 4. Probability of high-values properties and synthetic accessibility for unexplored ternary phase fields. a Unexplored ternary phase fields that are classified to exhibit superconductivity at T > 10 K with more than 70% probability and that have high likelihood of forming stable compounds (synthetic accessibility ranking < 0.2) demonstrate trends in constituent elements: most of the top 50 phase fields are predicted to contain Mg, Fe, Nb and N. **b** Unexplored ternary phase fields that are classified to exhibit energy band gap > 4.5 eV with more than 75% probability and that have high likelihood of forming stable compounds (with synthetic accessibility score < 0.1) demonstrate trends in distribution by constituent elements: different combinations of Hg-, F-, Bi-, Hf- and Pb-based phase fields have the highest probabilities. **c** Unexplored ternary phase fields that are classified to exhibit magnetic properties at Curie T > 300 K with more than 71%

probability and that have high likelihood of forming stable compounds (with synthetic accessibility score < 0.1) demonstrate trends in constituent elements: all top-ranked phase fields are Fe-based, with many phase fields containing Co and Y. **d** Receiver operating characteristics (ROC) of the classification models demonstrate high sensitivity and specificity of classifications for the range of thresholds of probabilities. The corresponding areas under the curves (AUC) demonstrate overall excellent performance of the model for magnetic materials, and good performance for both superconducting transition temperature and energy gap classifications. The inset illustrates close match of the distributions of 105995 unexplored candidate ternary phase fields with respect to their synthetic accessibility scores for all three datasets.

By applying the trained ensembles of models to 105995 unexplored ternary phase fields that do not have any related compositions reported in neither MPDS, Supercon, nor ICSD[2], we classify new elemental combinations with respect to the threshold values of superconducting transition temperature, Curie temperature and energy band gap and orthogonally rank candidate phase fields by their synthetic accessibility - degree of similarity with experimentally synthesized materials that are reported to exhibit these properties. The top-performing phase fields according to both probability of exhibiting high-values of properties and synthetic accessibility rank demonstrate trends in the constituent chemical elements: Mg, Fe, Nb are predicted to constitute most of the top 50 phase fields that would yield stable compositions with superconducting transition temperatures above 10 K; similarly top 50 magnetic ternary materials are Fe-based; while different combinations of Bi, Hf, Hg, Pb and F are predicted as most likely phase fields to form stable compounds with energy gap of more than 4.5 eV (See Figure 4a-c).
While these predictions may align well with the human experts' understanding of chemistry, hence emphasizing the models' ability to infer complex atomic characteristics and phase fields-properties relationship from historical data, the models can also be used to identify unconventional and rare prospective elemental combinations as well as to rank the attractive candidate materials for experimental investigations.

Finally, we apply PhaseSelect ensembles of classification models to identify likely candidates for novel superconducting, magnetic and photovoltaic materials among the phase fields that have been reported to form stable compounds in ICSD, but were not investigated from the perspectives of aforementioned applications (See Supplementary Table 2).

Conclusions

Selection of elements is the cornerstone of the materials design. Quantitative assessment of the materials at the level of their constituent elements mitigates the high-risk of the consequential decisions in the materials science. Classification of the materials for functional applications agglomerated into phase fields is also a route to the several orders of magnitude reduction of a bewildering combinatorial space. End-to-end integrated architecture of PhaseSelect has demonstrated this capability of rendering the materials' phase fields in two orthogonal and equally challenging dimensions: merit probability and synthetic uncertainty. By employing ML PhaseSelect at the stage of conceptualization of the materials synthesis, human researchers can guide themselves towards the selection of chemical elements that are most likely to synthesize new stable compounds with high probability of superior functional properties. The attention mechanism of PhaseSelect allows extrapolation of the knowledge of materials databases to the unlimited number of unexplored phase fields. These include multi-elemental materials, with prospective performance that could not be computationally assessed with the methods developed to date.

Acknowledgements

We thank the UK Engineering and Physical Sciences Research Council (EPSRC) for funding through grant number EP/N004884.

Data availability

The raw data and the software used in this study are available at https://www.github.com/lrcfmd/PhaseSelect.

Competing Interests Statement

The authors declare there are no competing interests.

Supporting Information

Supplementary Methods (Machine Learning methodology and models, Training set for the Classification, Autoencoder Configuration, Model validation, Atomic attention scores).

References

1. Villars, P., Cenzula, K., Savysyuk, I. & Caputo, R. Materials project for data science, https://mpds.io. (2021).
2. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Cryst.* **52**, 918–925 (2019).
3. National Institute of Materials Science, Materials Information Station, SuperCon, http://supercon.nims.go.jp/index_en.html. (2011).
4. Schleder, G. R., Padilha, A. C. M., Acosta, C. M., Costa, M. & Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* **2**, 032001 (2019).
5. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

6. Jha, D. *et al.* ElemNet : Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **8**, 1–13 (2018).
7. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
8. Stanev, V. *et al.* Machine learning modeling of superconducting critical temperature. *npj Comput. Mater.* **4**, 1–14 (2018).
9. Vasylenko, A. *et al.* Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. *Nat. Commun.* **12**, 5561 (2021).
10. Meredig, B. *et al.* Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.* **3**, 819–825 (2018).
11. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
12. Zhou, Q. *et al.* Learning atoms for materials discovery. *PNAS* **115**, E6411–E6417 (2018).
13. Vaswani, A. *et al.* Attention Is All You Need. *arXiv:1706.03762 [cs]* (2017).
14. Nayak, P. K., Mahesh, S., Snaith, H. J. & Cahen, D. Photovoltaic solar cell technologies: analysing the state of the art. *Nat Rev Mater* **4**, 269–285 (2019).
15. Martín Abadi *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015).

Supplementary Information

Element selection for functional materials discovery by integrated machine learning of atomic contributions to properties

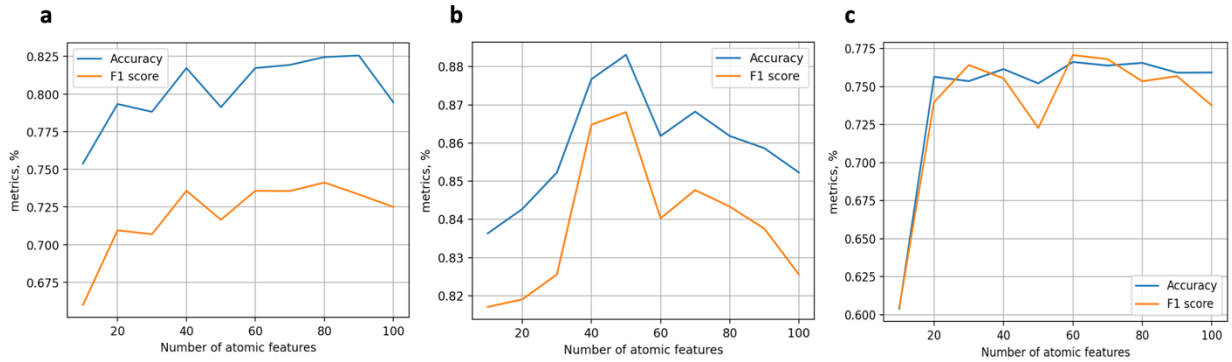
Andrij Vasylenko[1,2]

[1]Department of Chemistry, University of Liverpool, L69ZD Liverpool, UK
[2]Institute for Condensed Matter Physics, National Academy of Science of Ukraine, 1, Sventsitskogo str., 79011 Lviv, Ukraine

Atomic features encoding

For unsupervised learning of atomic features from the materials database[1], we employ an approach similar to[2], in which we substitute single value decomposition with a shallow autoencoder. Shallow autoencoder is a 3-layer neural network, in which input and output layers have a large number of neurons that corresponds to the size of the input vectors – sparse one-hot encoding representations of atoms in the database. A single latent layer in between the input and output is a bottleneck aiming to extract the essential patterns in data, while decreasing its dimensionality and filtering out the less representative and noisy information. One can further use thus trained representations as the atomic features. To maximise the quality and the descriptive power of the extracted atomic features, we study the effect of the size of the latent layer on the metrics of the downstream classifications. In this work, we train shallow autoencoder simultaneously with the classification neural network in the end-to-end fashion. When trained separately for classification of superconducting, magnetic materials, and materials with a reported band gap the end-to-end models based on the different size of atomic vectors have the metrics depicted in Supplementary Figure 1 a, b, c respectively. Despite that the best performance of classification of different properties is achieved at different number of atomic features, there is similar trend for these dependencies. This trend suggests that a small number (<

40) of features cannot fully capture the variation in data, and a large set of features (> 80) contains too much noise, hence there is an optimal number of atomic descriptors for each model.



Supplementary Figure 1. Changes in classification metrics for model with different number of atomic features. a Accuracy and F1 score for classification of materials with respect to the maximum of superconducting transition temperature 10 K: best performing model has 80 atomic features; **b** Accuracy and F1 score for classification of materials with respect to the maximum of Curie transition temperature 300 K: best performing model has 50 atomic features; **c** Accuracy and F1 score for classification of materials with respect to the maximum of energy band gap 4.5 eV: best performing model has 60 atomic features.

Models' training and validation

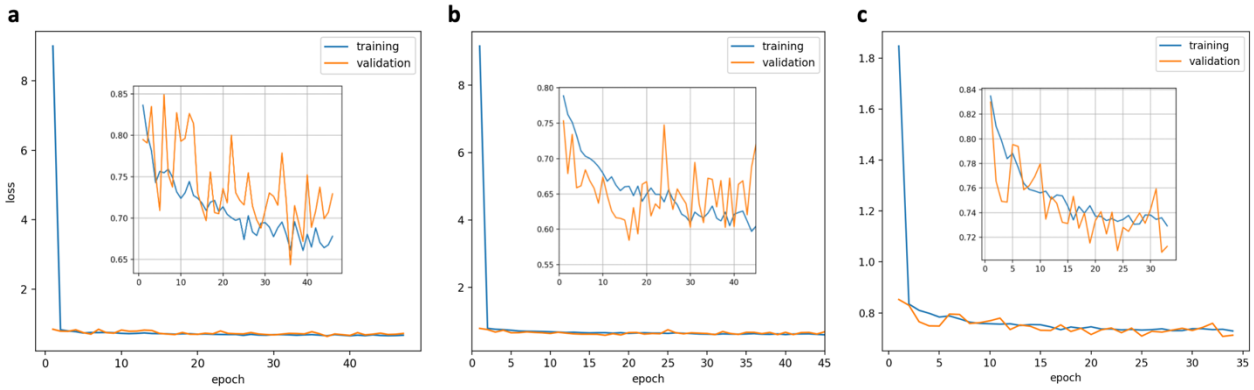
To validate the models' performance we employ 5-fold cross-validation for each dataset: phase fields with reported values of superconducting transition temperature, phase fields with reported values of Curie transition temperature, phase fields with reported values of energy gap. In 5-fold cross-validation, the data is divided into the training and test sets (80% and 20% of data respectively) in 5 different ways so 5 different models are examined with respect to the ability of the models' chosen architecture to generalise and extrapolate the information learnt from 5 different subsets of data onto the unseen areas. The accuracy and F1 scores of the classification models are presented in Supplementary Table 1.

Supplementary Table 1. Accuracy and F1 scores for classification models in 5-fold cross-validation

	Superconducting T_c=10K		**Magnetic T_c=300K**		**Energy gap 4.5 eV**	
test data subset	Accuracy,%	F1 score,%	Accuracy,%	F1 score,%	Accuracy,%	F1 score,%
0-20%	80.9	73.3	86.8	84.5	75.5	75.6
21-40%	83.6	77.1	86.7	85.7	75.2	74.8
41-60%	78.7	71.7	85.9	82.1	75.9	75.6

61-80%	79.7	71.3	85.5	84.1	76.0	75.1
81-100%	79.2	71.0	86.0	84.4	75.7	75.5
Average:	**80.4**	**72.9**	**86.2**	**84.2**	**75.6**	**75.3**

The performance metrics from the 5 models for each dataset are then averaged to describe a general ability of the models' architecture to learn from the available data. During the training of the end-to-end classification models, the weights and biases of the autoencoder and classifier neural networks are trained simultaneously, while the corresponding losses – reconstruction error and binary cross-entropy, respectively – are minimized as a combined loss during back propagation with Adam optimization[3]. The typical training of the classification models for the superconducting, magnetic and energy band gap datasets are converged under 50 epochs as illustrated in the Supplementary Figure 2.



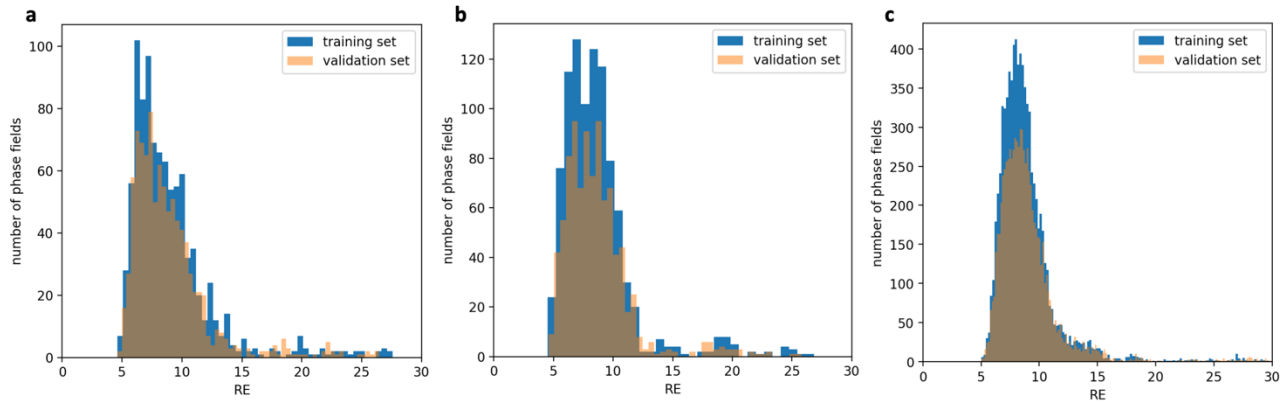
Supplementary Figure 2. Training progress of the end-to-end classification models. a Classification of the superconducting materials, training on 4826 phase fields; **b** Classification of the magnetic materials, training on 4753 phase fields; **c** Classification of the materials' energy band gap, training on 40452 phase fields.

For validation of the unsupervised models for the phase fields ranking with respect to synthetic accessibility, we employ an approach developed in [4]. We perform 5-fold cross validation, in which the validation error is defined as the percentage of entries in the test set that evaluated with normalized reconstruction errors in the 20% of the maximum Supplementary Table 2. Additionally, we compare the predicted reconstruction errors for the validation sets with the

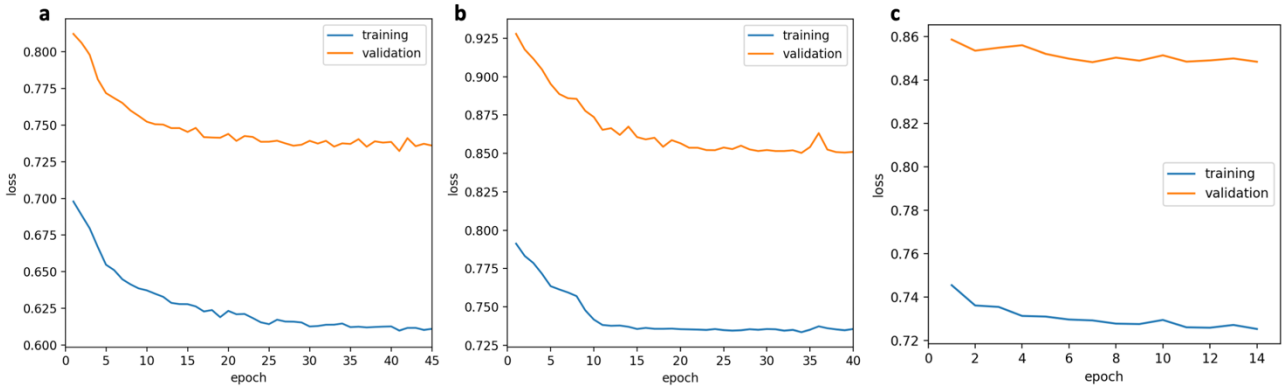
ground truth reconstruction errors obtained for the same entries in unsupervised training, when the entries are included in the training data (Supplementary Figure 3) and calculate mutual information score adjusted against chance[5] (Supplementary Table 2). The typical training process of the ranking autoencoder neural network for different dataset are depicted in Supplementary Figure 4.

Supplementary Table 2. Accuracy and Adjusted Mutual Information Score (AMIS) for ranking autoencoder models in 5-fold cross-validation

	Superconducting materials		**Magnetic materials**		**Energy gap materials**	
test data subset	Accuracy,%	AMIS	Accuracy,%	AMIS	Accuracy,%	AMIS
0-20%	96.1	0.69	94.7	0.64	97.2	0.77
21-40%	97.4	0.76	95.3	0.66	98.6	0.78
41-60%	97.7	0.68	93.5	0.66	97.7	0.75
61-80%	95.1	0.79	94.9	0.64	98.7	0.75
81-100%	96.6	0.72	93.9	0.68	97.8	0.81
Average:	**96.6**	**0.73**	**94.5**	**0.66**	**98.0**	**0.77**

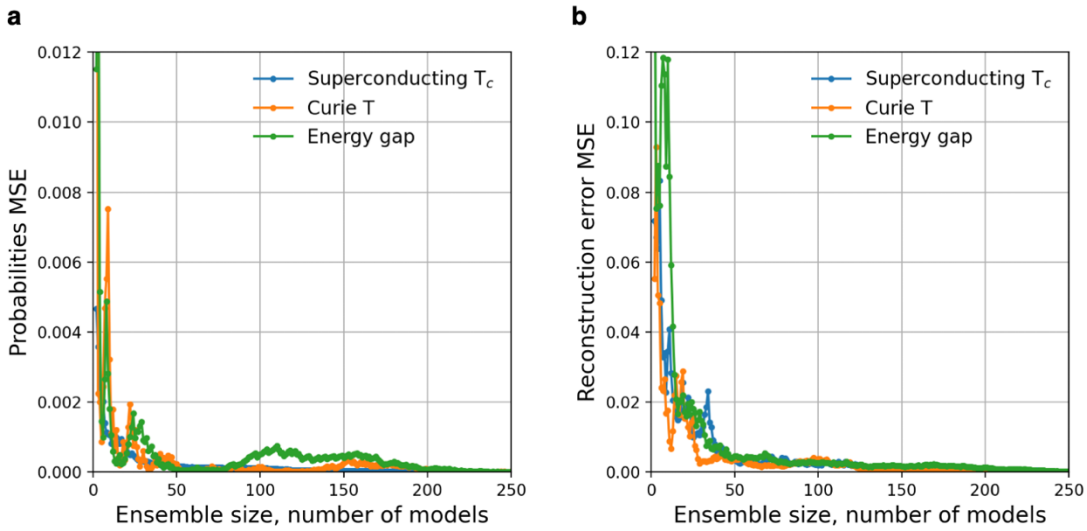


Supplementary Figure 3. Distribution of reconstructions errors (RE) for the phase fields. RE for the same phase fields are calculated in two approaches: 1) in unsupervised learning, as a part of a training set – used as ground truth RE for AMIS calculation in Supplementary Table 2; 2) predicted by the model trained on 80% of the remaining data – as a validation set. **a** Superconducting materials; **b** magnetic materials; **c** materials with reported energy gap.



Supplementary Figure 4. Training progress of the ranking autoencoder models. a ranking of the superconducting materials, training on 4826 phase fields; **b** ranking of the magnetic materials, training on 4753 phase fields; **c** ranking of the materials with the reported energy band gap, training on 40452 phase fields.

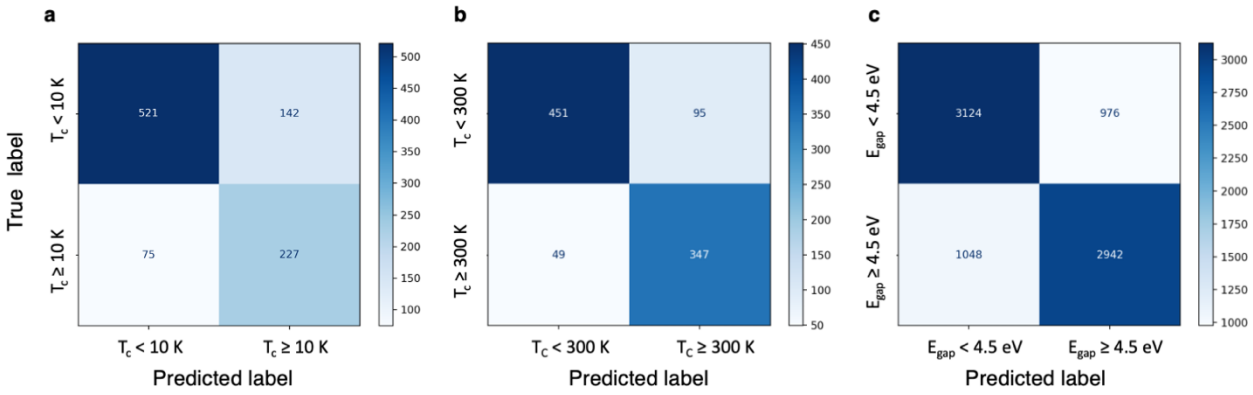
To take into account statistical variance in both supervised and unsupervised results from the neural networks trained at different instances, we average the results across the ensemble of 250 neural networks. Convergence of deviations of results in terms of the mean square errors from the running average values is illustrated in Supplementary Figure 5. For all datasets, for both supervised classifying neural network and ranking autoencoders, the average values converge when more than 200 models are considered.



Supplementary Figure 5. Convergence of the mean square errors (MSE) of the average predicted scores with a number of models in ensemble. a Probabilities of phase fields to belong to a binary class are averaged over ensemble of models. MSE of the average scores decrease below 0.001 for ensembles larger than 200 models

for all datasets. **b** MSE of the average reconstruction errors, used as synthetic accessibility scores of phase fields decrease below 0.005 for ensembles larger than 200 models.

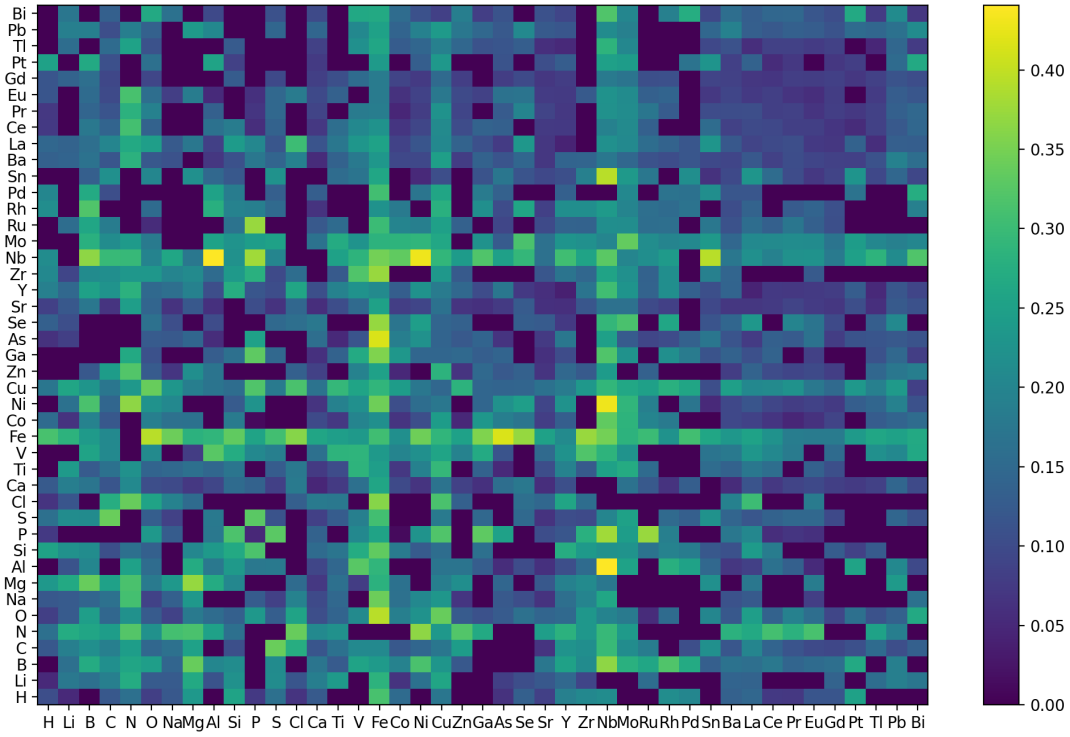
The ensembles of the trained models for each dataset are then used to classify the phase fields with respect to the corresponding properties. For randomly selected 20% of the phase fields from each dataset, the classification predictions are illustrated with the confusion matrices in Supplementary Figure 6.



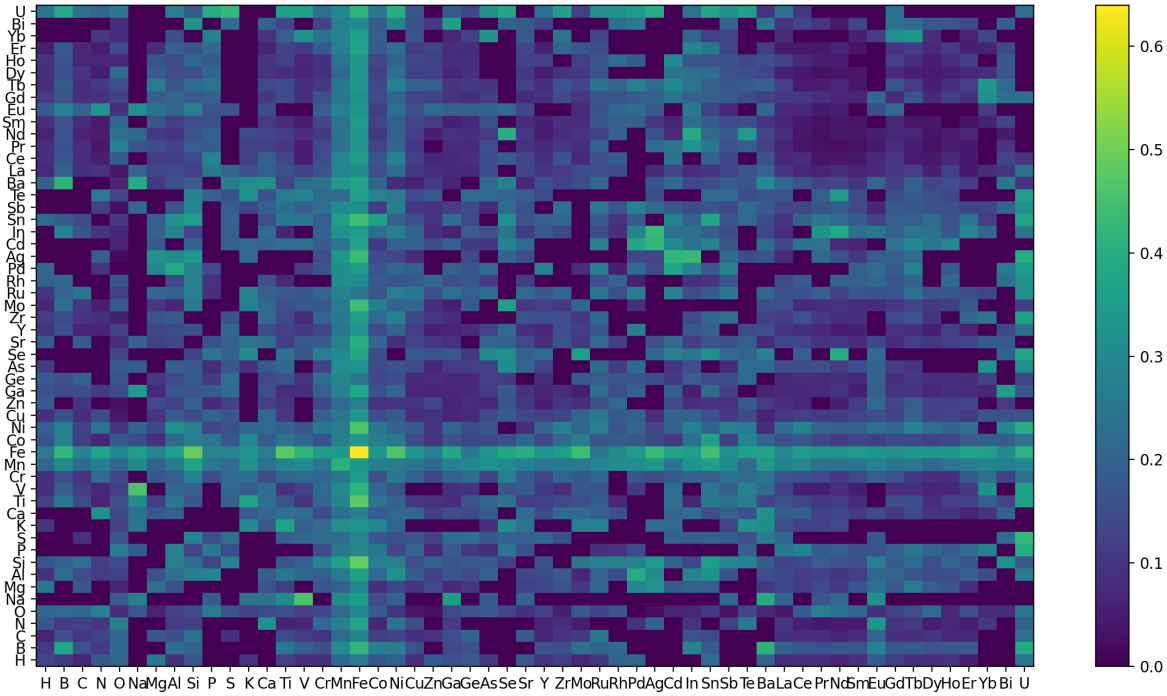
Supplementary Figure 6. Confusion matrices for binary classification models with threshold probability 0.5. a Superconducting materials classification of 20% of the collected data from MPDS[6] and SuperCon[7] with respect to transition temperature 10 K; **b** magnetic materials classification of 20% of the collected data from MPDS, with respect to Curie temperature 300 K; c classification materials with reported values of energy band gap with respect to energy gap value 4.5 eV, test set is 20% of randomly selected data collected from MPDS.

Attention to atomic contributions maximizing the properties

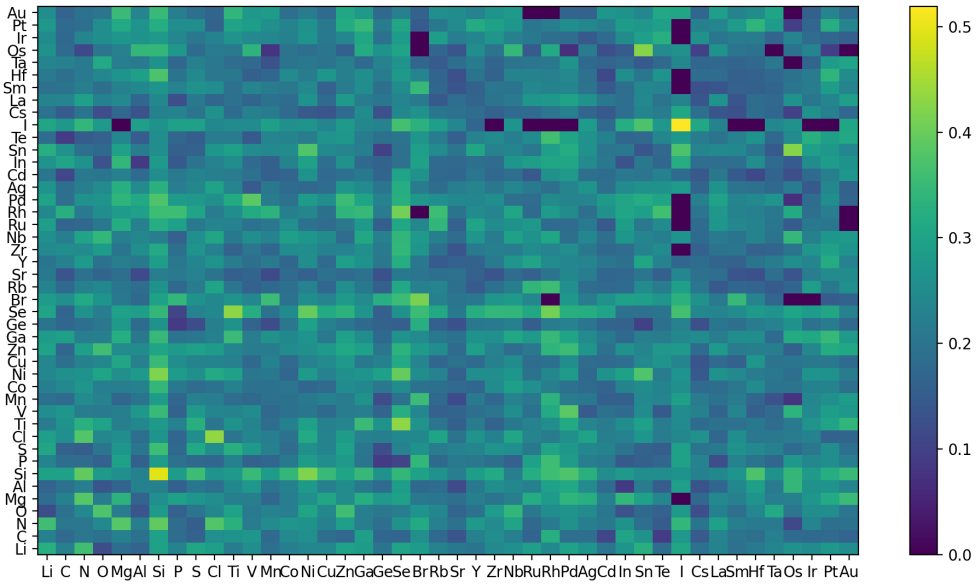
In the end-to-end classification models, we employ attention mechanism[8] to emphasize atomic contributions that minimize the combined loss, and hence maximize classification metrics. We extract the attention scores that we obtain during the training of the models that illustrate atomic contributions to the properties manifested by the phase fields (Supplementary Figures 7-11). For visualisation, the attention scores are averaged across the attention heads and across all instances of the atomic pairs in the corresponding datasets. In Supplementary Figure 11, distributions of the averaged attention scores are plotted for the atoms that contribute the most to identify phase fields that manifest particular properties.



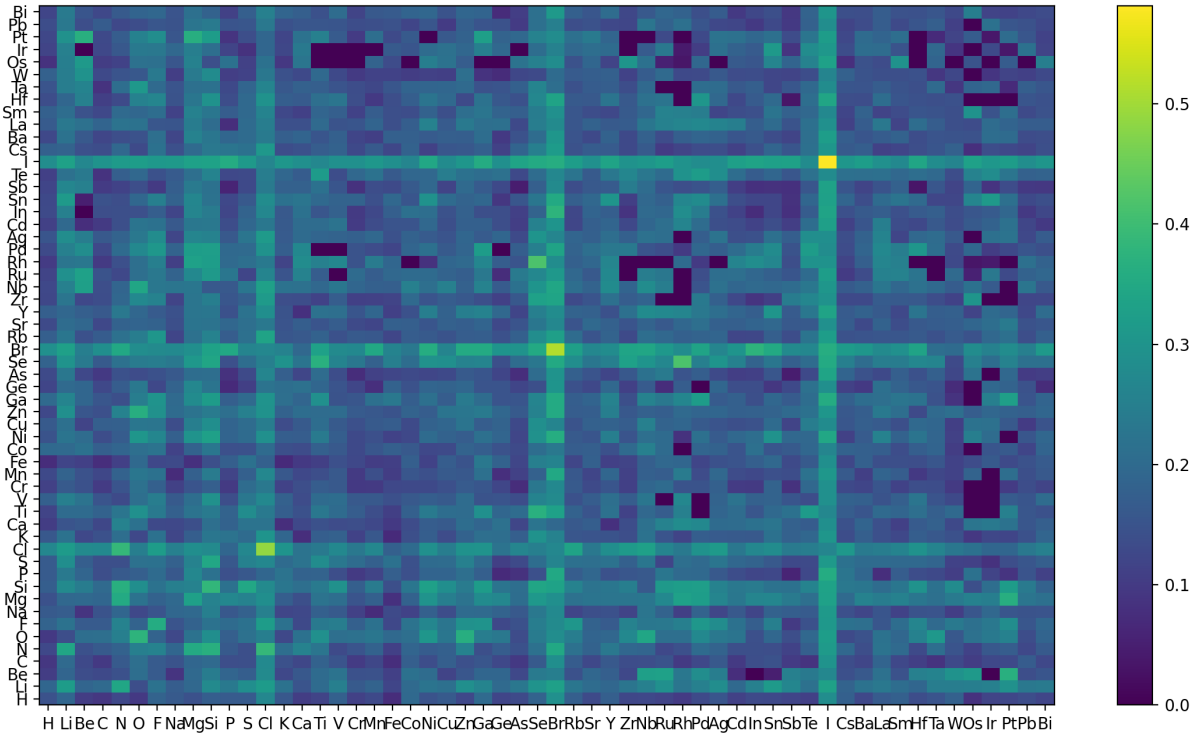
Supplementary Figure 7. Attention to atomic pairs that maximize accuracy of classification of high-temperature superconducting materials. Attention scores vary from 0 to 1. By focusing on the atomic pairs with the highest scores, when describing the phase field, accuracy of classification of these phase fields is maximized. This suggests the atomic pairs with the most prominent contributions allowing high-temperature superconductivity, e.g. Nb-Al, Nb-Ni and Fe-As.



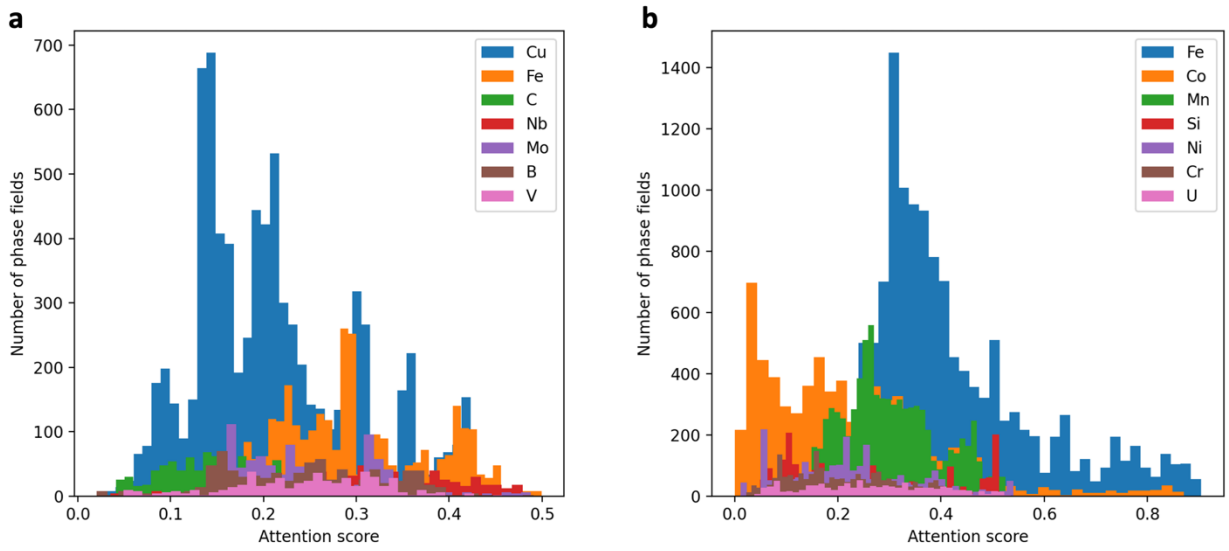
Supplementary Figure 8. Attention to atomic pairs that maximize accuracy of classification of high-temperature magnetic materials. Attention scores vary from 0 to 1. By focusing on the atomic pairs with the highest scores, when describing the phase field, accuracy of classification of these phase fields is maximized. This suggests the atomic pairs with the most prominent contributions allowing high-temperature magnetic behaviour, with Mn, Fe and Co included in the majority of such pairs.



Supplementary Figure 9. Attention to atomic pairs that maximize accuracy of classification of materials with energy gap < 4.5 eV. Attention scores vary from 0 to 1. By focusing on the atomic pairs with the highest scores, when describing the phase field, accuracy of classification of these phase fields is maximized. The majority of the atoms in the phase fields have 0.3-0.5 attention score, and contribute equally to identification low energy gap.



Supplementary Figure 10. Attention to atomic pairs that maximize accuracy of classification of materials with energy gap > 4.5 eV. Attention scores vary from 0 to 1. By focusing on the atomic pairs with the highest scores, when describing the phase field, accuracy of classification of these phase fields is maximized. This suggests atoms and atomic pairs with the most prominent contributions to the materials with energy gap > 4.5 eV, e.g. I, Br, Se, Cl, Si.



Supplementary Figure 11. Distribution of attention scores for the most contributing atoms to the functional materials a High-temperature superconducting materials; **b** high-temperature magnetic materials.

Combination of probabilities of high-values properties and minimized synthetic uncertainties

We combine the outcomes of the classifying neural network and autoencoder to rank unexplored ternary combinations of elements. The best ranking combinations, illustrated in Figure 4 in the main text are presented in the Supplementary Tables 3-5. The full list of the predicted scores for the yet experimentally unexplored ternary phase fields as well as the predictions for the phase fields, in which the experimentally synthesized compounds are reported in Inorganic Crystal Structure Database[9], but which, however, were not studied with respect to their potential as high-temperature superconductors or magnetic materials can be found along with the PhaseSelect software[10].

Supplementary Table 3. Predicted probabilities of the best unexplored ternary phase fields to manifest superconducting T_c > 10 K and their synthetic uncertainty scores.

Phase fields	Probability Tc > 10 K	Synthetic uncertainty

N Fe Nb	0.7433	0.1643
Fe Ga Nb	0.7341	0.1192
Mg Fe As	0.7295	0.1557
Mg Fe Nb	0.7278	0.1016
Fe As Nb	0.7261	0.0903
N Cl Nb	0.7238	0.1781
N Zr Nb	0.7235	0.1411
Fe Se Nb	0.7212	0.124
N Mg Zr	0.72	0.1776
N K Nb	0.7194	0.1798
N Nb Mo	0.7192	0.1677
Mg V Fe	0.7189	0.1589
N Na Nb	0.7187	0.1774
Ca Fe As	0.7162	0.1477
V Fe As	0.7154	0.1299
N V Nb	0.7134	0.1424
Fe Ga As	0.7126	0.1709
N Ca Nb	0.7111	0.1665
N Fe Nb	0.7433	0.1643

Supplementary Table 4. Predicted probabilities of the best unexplored ternary phase fields to manifest Curie T_c > 300 K and their synthetic uncertainty scores.

Phase fields	Probability T_c > 300 K	Synthetic uncertainty
Fe Y Ta	0.7185	0.0613
Ti Fe Ta	0.7181	0.0658
Fe Y Zr	0.7179	0.0661
Fe Mo Hf	0.717	0.0685
Ti Fe Hf	0.716	0.0603

Fe Y Nb	0.7159	0.0681
Fe Y Hf	0.7154	0.0557
Fe Zr Ta	0.7147	0.0485
V Fe Ta	0.7136	0.0631
Cr Fe Ta	0.7131	0.057
V Fe Zr	0.7128	0.0679
Ti Fe Hg	0.7123	0.0642
Cr Fe Zr	0.7123	0.0619
Fe Hf Ta	0.7117	0.0467
Fe Zr Hf	0.7113	0.0426
Fe Nb Ta	0.7111	0.0522
Fe Zr Nb	0.7111	0.0557
Fe Y Hg	0.7106	0.0598
V Fe Nb	0.7106	0.0699
V Fe Hf	0.7101	0.0575

Supplementary Table 5. Predicted probabilities of the best unexplored ternary phase fields to manifest energy band gap > 4.5 eV and their synthetic uncertainty scores.

Phase fields	Probability T_c > 300 K	Synthetic uncertainty
F Pb Bi	0.7649	0.089
F Cs Bi	0.7623	0.0969
F Cs Pb	0.7613	0.0885
F Hg Bi	0.7603	0.0897
F Hg Pb	0.759	0.0811
F Te Hf	0.7575	0.0975
F Y Bi	0.7575	0.0984
F Hf Bi	0.7574	0.0906
F Cs Hg	0.7572	0.0787

F As Hf	0.7572	0.0984
Cl I Hf	0.7561	0.0999
F Cd Bi	0.7561	0.0882
F Au Pb	0.7556	0.0932
F Hf Pb	0.7556	0.082
F Y Pb	0.7555	0.09
F Cd Pb	0.755	0.0796
F Cs Hf	0.7548	0.0761
F V Bi	0.7531	0.0904

Supplementary References

1. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
2. Zhou, Q. *et al.* Learning atoms for materials discovery. *PNAS* **115**, E6411–E6417 (2018).
3. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017).
4. Vasylenko, A. *et al.* Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. *Nat. Commun.* **12**, 5561 (2021).
5. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? in *Proceedings of the 26th Annual International Conference on Machine Learning* 1073–1080 (Association for Computing Machinery, 2009). doi:10.1145/1553374.1553511.
6. Villars, P., Cenzula, K., Savysyuk, I. & Caputo, R. Materials project for data science, https://mpds.io. (2021).

7. National Institute of Materials Science, Materials Information Station, SuperCon, http://supercon.nims.go.jp/index_en.html. (2011).
8. Vaswani, A. *et al.* Attention Is All You Need. *arXiv:1706.03762 [cs]* (2017).
9. Zagorac, D., Müller, H., Ruehl, S., Zagorac, J. & Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Cryst.* **52**, 918–925 (2019).
10. Vasylenko, A. PhaseSelect: Element selection for functional materials discovery by integrated machine learning of atomic contributions to properties, https://github.com/lrcfmd/PhaseSelect. (2021).