Supplementary Information

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

Andrij Vasylenko[1], D. Antypov[1], V. Gusev[1], M. W. Gaultois[1], M. Dyer[1], M. J. Rosseinsky[1,*]

[1]Department of Chemistry, University of Liverpool, L697ZD Liverpool, UK

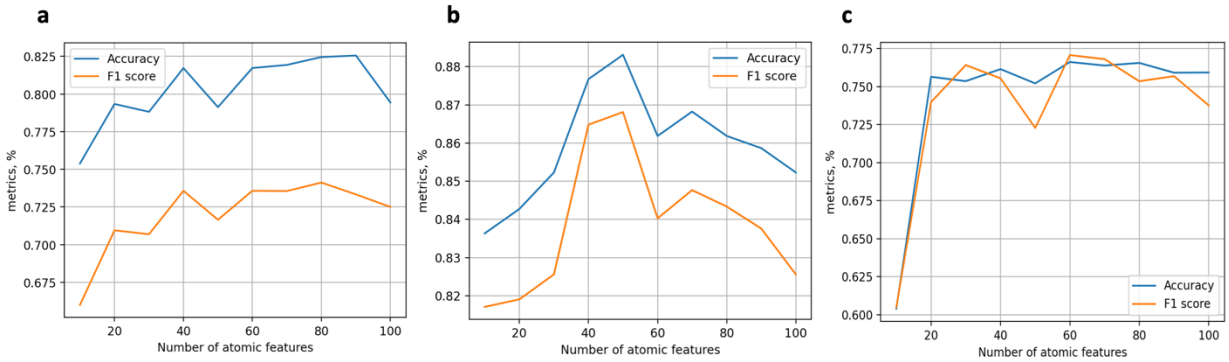
[*]corresponding author

Contents

Atomic features encoding

For unsupervised learning of atomic features from the materials database[1], we employ an approach similar to reference[2], in which we substitute single value decomposition with a shallow autoencoder. A shallow autoencoder is a 3-layer neural network, in which the input and output layers have a large number of neurons that corresponds to the size of the input vectors – sparse one-hot encoding

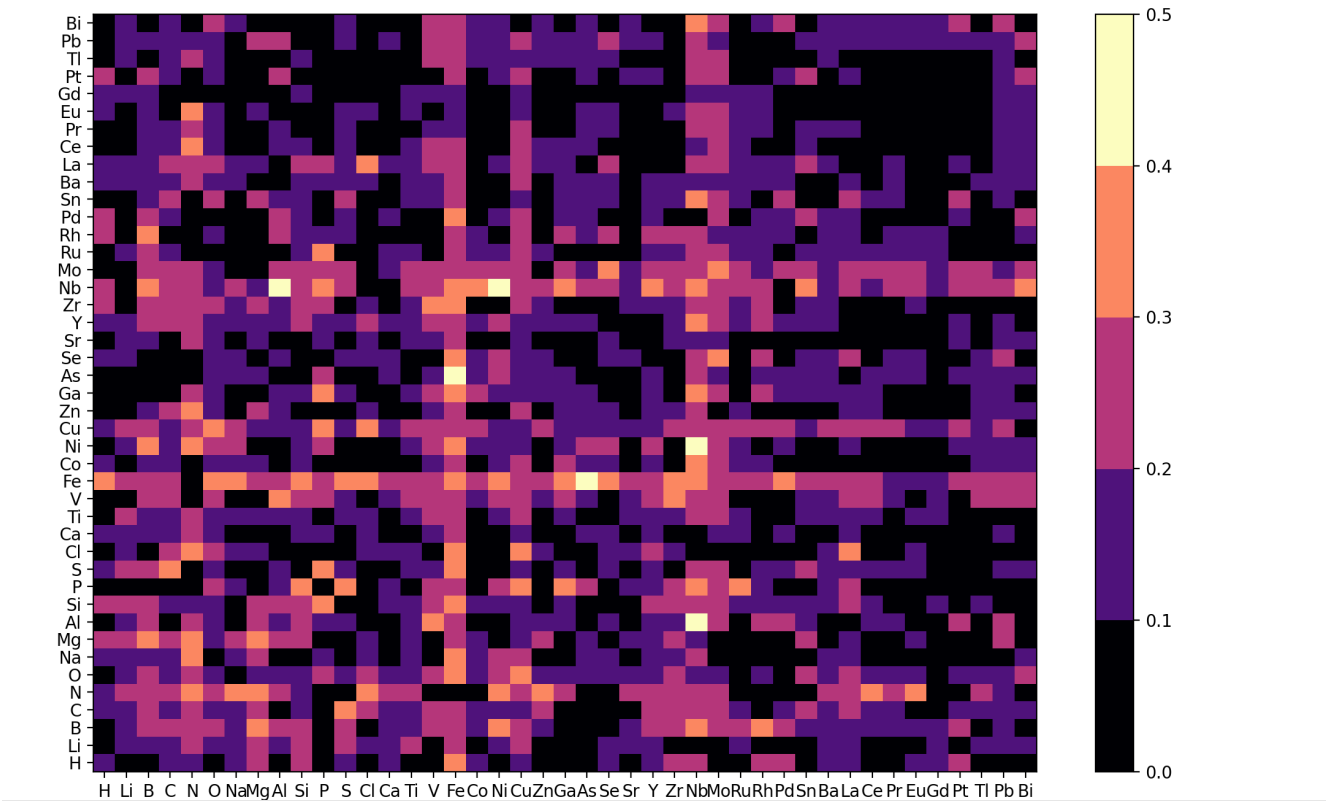
representations of atoms in the database. A single latent layer in between the input and output is a bottleneck aiming to extract the essential patterns in the data, while decreasing its dimensionality and filtering out the less representative and noisy information. One can further use thus trained representations 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 the 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 sizes of atomic vectors have the metrics depicted in Supplementary Figure 1 a, b, c respectively. Although the best performance for classification of different properties is achieved at different numbers of atomic features in each of the three cases, 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.



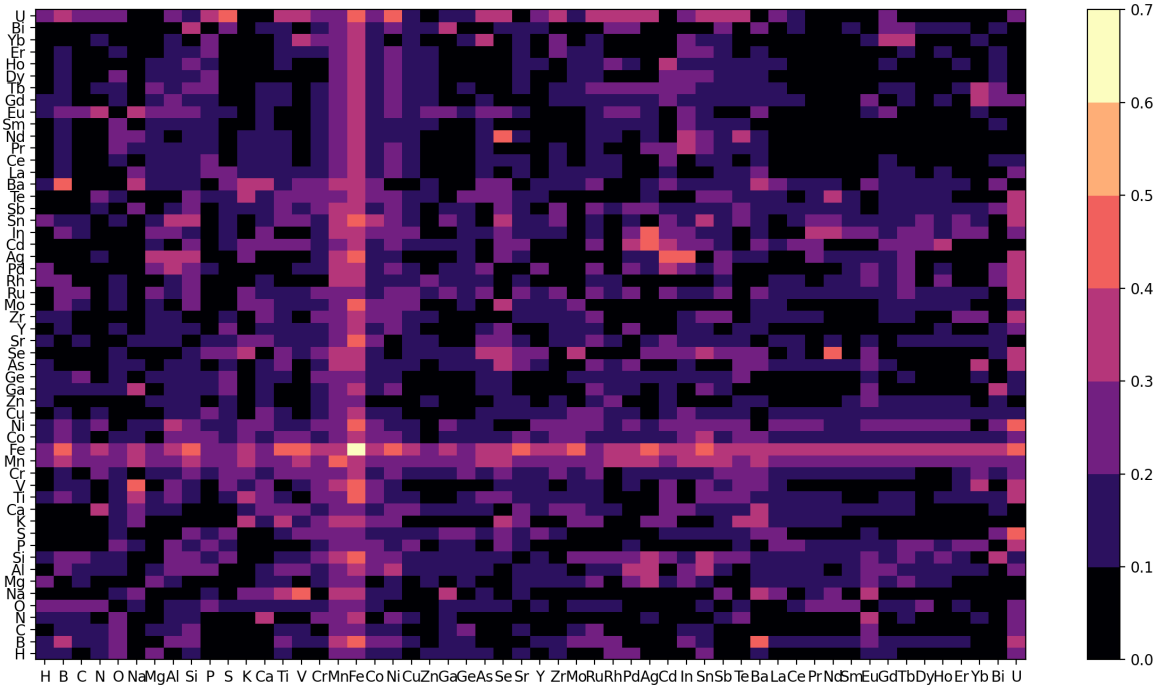
Supplementary Figure 1. Changes in classification metrics for models with different number of atomic features. **a** Accuracy and F1 score for classification of materials with respect to the maximum of superconducting transition temperature threshold 10 K: the best performing model has 80 atomic features;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

Attention to atomic contributions maximizing the properties

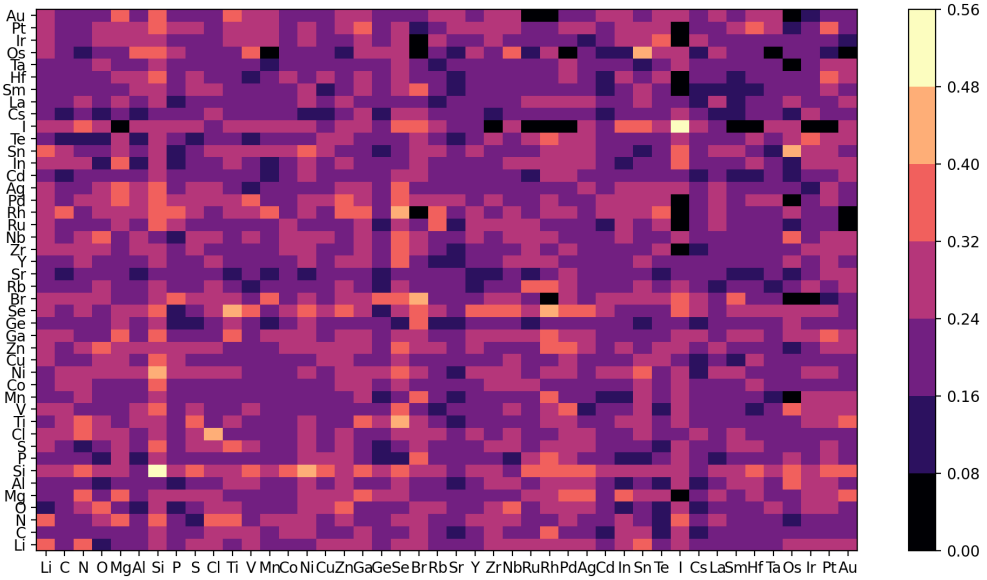
In the end-to-end classification models, we employ an attention mechanism[3] to emphasize those atomic contributions that minimize the combined loss, and hence maximize classification metrics. To incorporate information about atomic bonding interplay from all available data, the variance in size of the 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. We extract the attention scores obtained during the training of the models that illustrate atomic contributions to the properties manifested by the phase fields (Supplementary Figures 2-6). 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.



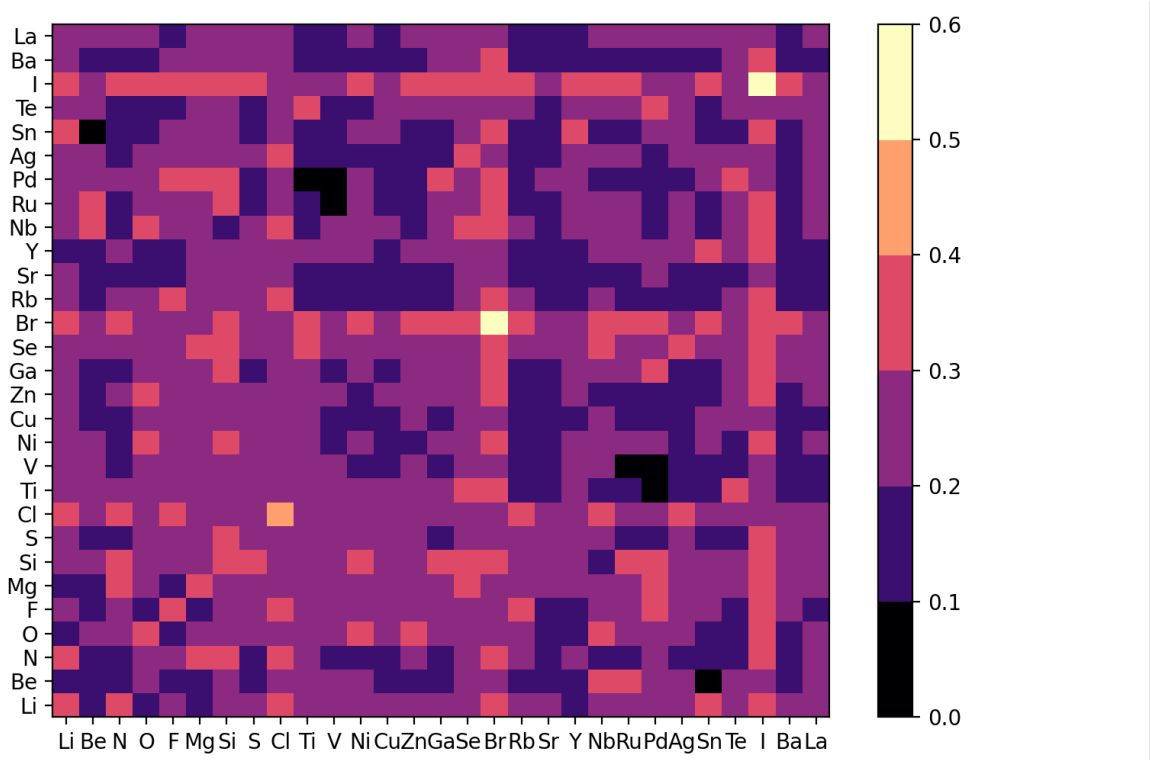
Supplementary Figure 2. Attention to atomic pairs that maximize accuracy of classification of high-temperature superconducting materials. Attention scores vary from 0 to 1. By focusing on the atomic pairs with the highest scores, when describing the phase field, accuracy of classification of these phase fields is maximized. This suggests the atomic pairs with the most prominent contributions allowing high-temperature superconductivity, e.g. Nb-Al, Nb-Ni, Cu-O and Fe-As.



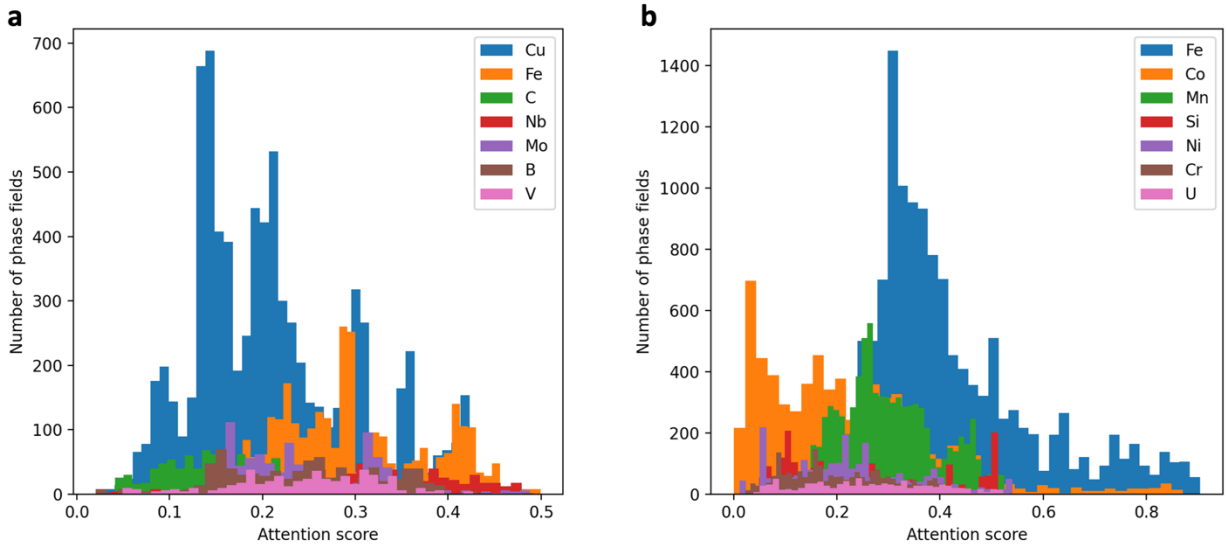
Supplementary Figure 3. Attention to atomic pairs that maximize the accuracy of classification of high-temperature magnetic materials. Attention scores vary from 0 to 1. By focusing on the atomic pairs with the highest scores, when describing the phase field, accuracy of classification of these phase 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 4. 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 of low energy gaps.



Supplementary Figure 5. 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 6. Distribution of attention scores for the most contributing atoms to the functional materials a High-temperature superconducting materials; **b** high-temperature magnetic materials.

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.

Models' training and validation

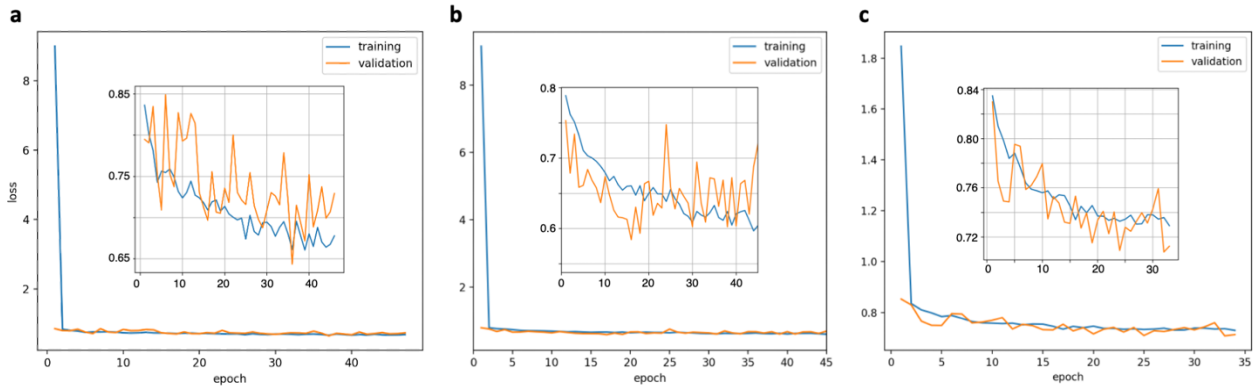
To validate the models' performance we employ 5-fold cross-validation for each dataset: phase fields with reported values of superconducting transition temperature, phase fields with reported values of Curie transition temperature, phase fields with reported values of energy gap. In 5-fold cross-validation, 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[4]. 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 7.



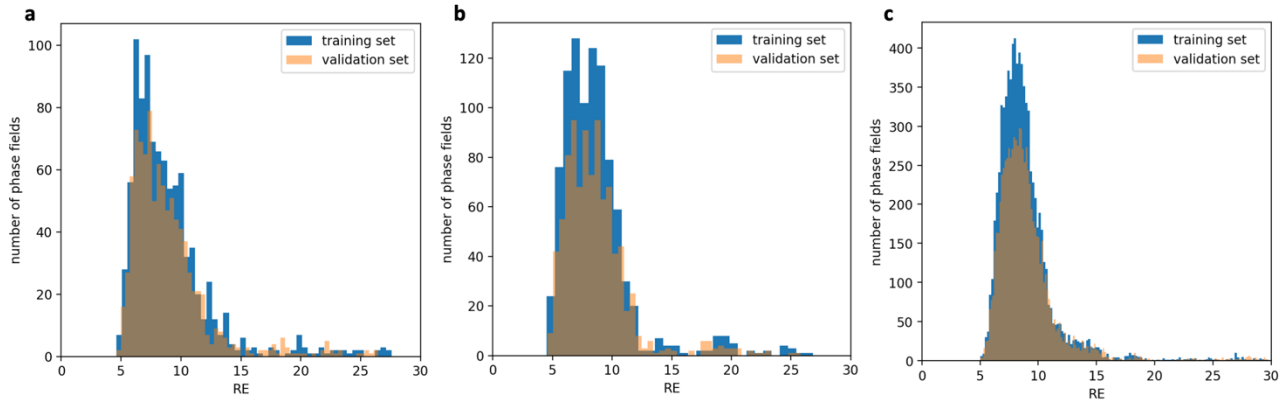
Supplementary Figure 7. Training progress of the end-to-end classification models. a Classification of the superconducting materials, training on 4826 phase fields; **b** Classification of the magnetic materials, training on 4753 phase fields; **c** Classification of the materials' energy band gap, training 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 [5]. 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 Fig. 8) and calculate the mutual information score adjusted against chance[6] (Supplementary Table 2). The typical training process of the ranking autoencoder neural network for different datasets are depicted in Supplementary Fig. 9.

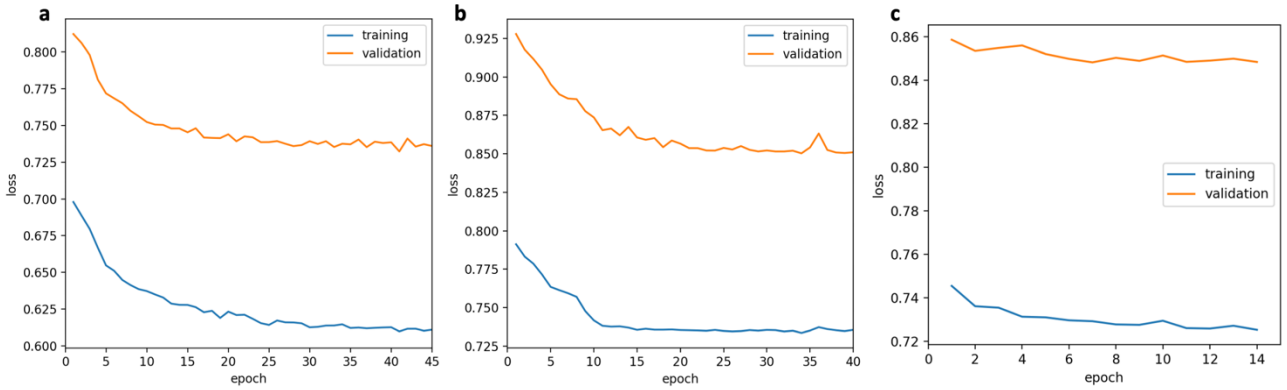
Supplementary Table 2. Accuracy and Adjusted Mutual Information Score (AMIS) for ranking autoencoder models in 5-fold cross-validation

	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**



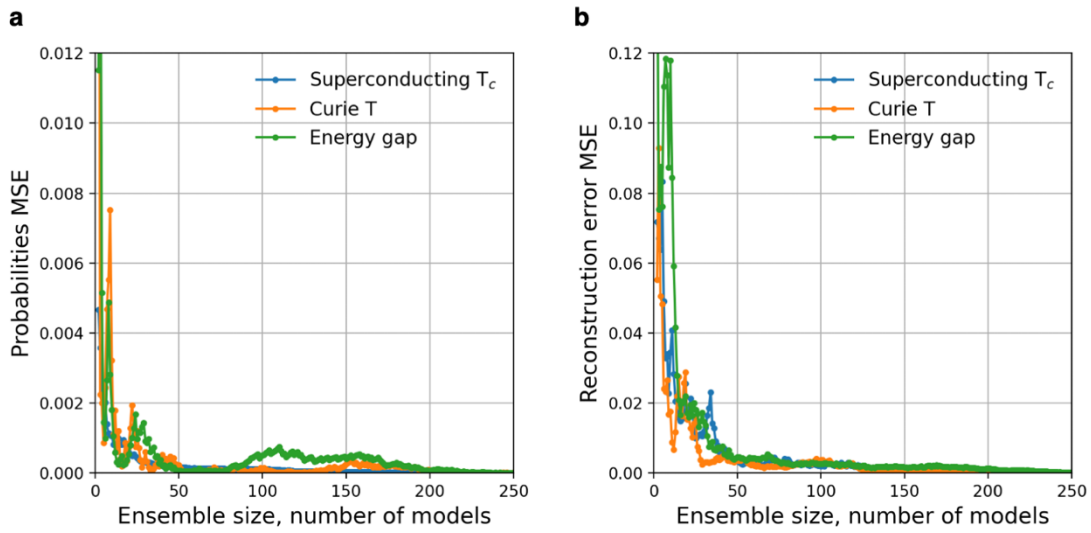
Supplementary Figure 8. Distribution of reconstructions errors (RE) for the phase fields. RE for the same phase fields are calculated in two approaches: 1) in unsupervised learning, as a part of a training set – used as ground truth RE for AMIS calculation in Supplementary Table 2; 2) predicted by the 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 9. 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.

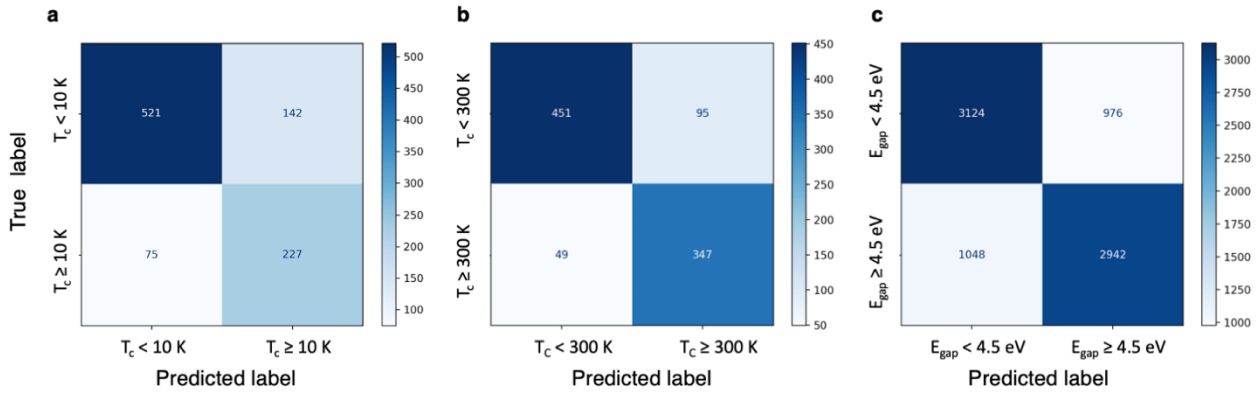
To take into account statistical variance in both supervised and unsupervised results from the neural networks trained at different instances, we average the results across the ensemble of 250 neural networks. Convergence of deviations of results in terms of the mean square errors from the running

average values is illustrated in Supplementary Figure 10. 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 10. Convergence of the mean square errors (MSE) of the average predicted scores with the number of models in the ensemble. a Probabilities of phase fields belonging to a binary class are averaged over an 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 11.



Supplementary Figure 11. Confusion matrices for binary classification models with threshold probability 0.5. a Superconducting materials classification of 20% of the collected data from MPDS[7] and SuperCon[8] 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.

The corresponding average accuracy, F1 score and the Matthews' correlation coefficients (MCC) are presented for the three models in Supplementary Table 3.

Supplementary Table 3. Average binary classifications metrics of the maximum values of exhibited properties in the phase field.

Metrics	Superconducting Tc >10 K	Magnetic Tc > 300 K	Energy gap > 4.5 eV
Accuracy, %	80.4	86.2	75.6
F1 score, %	72.9	84.2	75.3
MCC	0.608	0.711	0.523

Combination of probabilities of high-values properties (merit probability) and synthetic uncertainties

We combine the outcomes of the classifying neural network and autoencoder to rank unexplored ternary combinations of elements. For the unexplored ternary combinations we consider all possible combinations of 87 atoms, that exclude rare and toxic elements and have sufficient data in Materials Project to be reasonably well learnt with the proposed unsupervised approach described above. The total number of ternary combinations, therefore, is $87 \times 86 \times 85 / 3! = 105995$, among them 12297 have a reported value of energy band gap in MPDS (and in a peer-reviewed literature), 1953 are reported to have magnetic properties and a corresponding Curie temperature in MPDS, and 1716 are reported to have superconducting properties and a corresponding critical temperature in a combined data from SuperCon and MPDS.

The best ranking combinations, illustrated in Figure 4 in the main text are presented in the Supplementary Tables 4-6. Among the considered phase field there are entries that have been

synthesized and reported in ICSD-v2021[9], but do not have records in MPDS and SuperCon concerning the properties studied here. These entries did not enter the training datasets and are highlighted in bold in the Supplementary Tables 4-6. These entries have been predicted to have low synthetic uncertainty, that provides experimental verification of the proposed method for ML assessment of synthetic accessibility. The full list of the predicted scores for the yet experimentally unexplored ternary phase fields can be found along with the PhaseSelect software[10].

Supplementary Table 4. Predicted probabilities of the best unexplored ternary phase fields to manifest superconducting T_c > 10 K and their synthetic uncertainty scores. The phase fields, in which compounds are synthesized[9] but were not included into the training data[7,8] are highlighted in bold.

Phase fields	Probability Tc > 10 K	Synthetic uncertainty
N Fe Nb	**0.7433**	**0.1643**
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
Fe Se Nb	**0.7212**	**0.124**
N Mg Zr	0.72	0.1776
N K Nb	0.7194	0.1798
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**
Fe Ga As	**0.7126**	**0.1709**
N Ca Nb	**0.7111**	**0.1665**

Supplementary Table 5. Predicted probabilities of the best unexplored ternary phase fields to manifest Curie T_c > 300 K and their synthetic uncertainty scores. The phase fields, in which compounds are synthesized[9] but were not included into the training data[7] are highlighted in bold.

Phase fields	Probability T_c > 300 K	Synthetic uncertainty
Ti Fe Ta	**0.7181**	**0.0658**
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
V Fe Ta	0.7136	0.0631
Cr Fe Ta	0.7131	0.057
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 Y Hg	0.7106	0.0598
V Fe Nb	0.7106	0.0699
V Fe Hf	**0.7101**	**0.0575**

Supplementary Table 6. Predicted probabilities of the best unexplored ternary phase fields to manifest energy band gap > 4.5 eV and their synthetic uncertainty scores. The phase fields, in which compounds are synthesized[9] but were not included into the training data[7] are highlighted in bold.

Phase fields	Probability E_g > 4.5 eV	Synthetic uncertainty
Cs F Pb	**0.7613**	**0.8852**
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 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 Cd Pb	0.755	0.0796
F V Bi	0.7531	0.0904

Prediction of superconducting behaviour for reported phase fields in ICSD-v2021

We apply PhaseSelect ensembles of classification models to identify likely candidates for novel superconducting materials among the phase fields that have been reported to form stable compounds in ICSD-v2021, but were not investigated from the perspectives of superconducting applications and reported in MPDS and SuperCon (hence were not included into the training dataset). The excerpt of these predictions is presented in Supplementary Table 7; classification of all binary, ternary and quaternary phase field in ICSD with respect to the maximum accessible value of superconducting critical temperature is uploaded in[10].

Supplementary Table 7. Predicted probabilities of superconducting behaviour at T_c > 10 K for the best ternary phase fields reported to form stable structures in ICSD. (Excerpt for p > 0.7. The full list is in[10]).

Phase fields	Probability T_c > 10 K	Phase fields	Probability T_c > 10 K
Fe N Nb	0.7466	Mo N Nb	0.7142
Fe Li N	0.7391	C Li N	0.7131
Fe Ga N	0.7383	As Fe Nb	0.7113
C Fe N	0.7352	Al N Nb	0.7111
Fe Mo N	0.7343	C Ga N	0.7102
Ba Fe N	0.733	Ga N V	0.7101
Fe N Se	0.7324	N Nb V	0.7096
Ca Fe N	0.7322	Ca Fe O	0.709
C Mg N	0.7305	C N V	0.709
Fe Mg O	0.7302	Ba Fe O	0.7087
Li Mg N	0.7291	B Mg N	0.7077
Ga N Nb	0.7262	Fe Nb Se	0.7075
Ga Mg N	0.7261	C K N	0.7075
C N Nb	0.726	Ca Mg N	0.7065

Fe N Zr	0.725	As Ca Fe	0.7058
Mg Mo N	0.7229	C Cl N	0.7056
Fe Ga Nb	0.7227	N Na Nb	0.7056
Fe N Sr	0.7226	As Fe V	0.7055
Cl Mg N	0.7218	N Nb Zr	0.7054
Cu Fe O	0.7197	Ba N Nb	0.7037
Fe N Sn	0.7194	As Fe Ga	0.7034
Fe Mn N	0.7192	Fe N Pt	0.7023
Fe N O	0.7186	C N Na	0.702
As Fe O	0.7175	Ca N Nb	0.7019
Fe N Y	0.7173	As Ba Fe	0.7017
C Mo N	0.717	C Ca N	0.7014
Fe Ga V	0.7157	As Fe K	0.7007
Li N Nb	0.7143		

Tools and Libraries

PhaseSelect[10] has been built using Python 3.7.4, Tensorflow 2.4.1, Scikit-learn 0.24.0, Numpy 0.19.2, Pandas 1.1.4. The figures in the main text and Supplementary figures are created using Matplotlib 3.3.4.

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. Vaswani, A. *et al.* Attention Is All You Need. *arXiv:1706.03762 [cs]* (2017).
4. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017).
5. 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).

6. 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.
7. Villars, P., Cenzula, K., Savysyuk, I. & Caputo, R. Materials project for data science, https://mpds.io. (2021).
8. National Institute of Materials Science, Materials Information Station, SuperCon, http://supercon.nims.go.jp/index_en.html. (2011).
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).