

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

**Andrij Vasylenko**

University of Liverpool

**Dmytro Antypov**

University of Liverpool <https://orcid.org/0000-0003-1893-7785>

**Vladimir Gusev**

University of Liverpool

**Michael Gaultois**

University of Liverpool <https://orcid.org/0000-0003-2172-2507>

**Matthew Dyer**

University of Liverpool

**Matthew Rosseinsky** (✉ [m.j.rosseinsky@liverpool.ac.uk](mailto:m.j.rosseinsky@liverpool.ac.uk))

University of Liverpool <https://orcid.org/0000-0002-1910-2483>

---

## Article

### Keywords:

**Posted Date:** February 18th, 2022

**DOI:** <https://doi.org/10.21203/rs.3.rs-1334648/v1>

**License:** © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

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

A. Vasylenko<sup>1</sup>, D. Antypov<sup>1</sup>, V. Gusev<sup>1</sup>, M. W. Gaultois<sup>1</sup>, M. S. Dyer<sup>1</sup>, M. J. Rosseinsky<sup>1\*</sup>

<sup>1</sup> Department of Chemistry, University of Liverpool, Crown Street L697ZD, UK

\*corresponding author

## Abstract

At the high level, the fundamental differences between materials originate from the unique nature of the constituent chemical elements. Before specific differences emerge according to the precise ratios of elements (composition) in a given crystal structure (phase), the material can be represented by its phase field defined simply as the set of the constituent chemical elements.

Classification of the materials at the level of their phase fields can accelerate materials discovery by selecting the elemental combinations that are likely to produce desirable functional properties in synthetically accessible materials. Here, we demonstrate that classification of the materials' phase field with respect to the maximum expected value of a target functional property can be combined with the ranking of the materials' synthetic accessibility.

This end-to-end machine learning approach (PhaseSelect) first derives the atomic characteristics from the compositional environments in all computationally and experimentally explored materials, and then employs these characteristics to classify the phase field by their merit.

PhaseSelect can quantify the materials' potential at the level of the periodic table, which we demonstrate with significant accuracy for three avenues of materials' applications: high-temperature superconducting, high-temperature magnetic and targeted energy band gap materials.

## Introduction

Conceptualization of novel materials begins at the level of the periodic table with selection of chemical elements for synthetic investigation. There is a variety of possible ratios or compositions that can be formed from a set of chemical elements leading to different materials (phases); the field of these potential realizations can be defined as a material's phase field. The choice of a phase field to investigate ultimately determines the outcome of the synthetic work and the functional properties of the prospective materials.

The fundamental differences between atoms result in the variance in the materials' properties in thousands of compositions accumulated in materials databases<sup>1-3</sup>. Harvesting these 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<sup>4,5</sup>. Recently, the data-driven approach has also been used in the inverse design ML aspiring to predict the materials with the targeted properties<sup>6</sup>. Ranging from formation enthalpy<sup>7</sup> to energy band gap<sup>8</sup> to superconducting transition temperature<sup>9</sup>, 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. At the same time, many high-performance ML models based on the deep learning<sup>10</sup> or ensembles<sup>11</sup> methods lack interpretability<sup>12,13</sup>, hence they are not readily adopted by experimental teams. Improvement of interpretability of ML approaches without compromises on performance could bridge powerful ML methods with experimental workflows to form trusted ML-expert systems in material sciences.

Codification of the materials for statistical treatment involves description of the constituent chemical elements, often represented as vectors of their chemical and physical characteristics, that are combined linearly to describe a compound<sup>7</sup>. 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 the atomic descriptions in materials representations. This selection determines the quality of the model<sup>14</sup>. The composition-based models are predisposed to data leakage between training and validation datasets via compositionally close datapoints, that impedes the extrapolation of patterns in materials-properties relationships onto unexplored materials that have distinct chemistries from those in the training set<sup>15</sup>.

In this work, our goal is to assess the attractiveness of candidate functional materials at the high level of the periodic table by identifying the most promising phase fields that are likely to contain these candidates. This circumvents the combinatorial challenge of individual assessment of all possible compositions built from the chosen elements and aligns with the experimental challenge of identifying new functional materials from previously uncharted chemistry. We demonstrate that unsupervised learning of chemical elements combined with the attention technique for learning elemental contributions can be used for the accurate classification of the materials' functional performance at the level of the phase fields, while improving interpretability of the ML reasoning. This end-to-end integrated machine learning (PhaseSelect) of the materials databases can prioritise the materials with respect to both probability of a merit (maximum achievable value of the target property) and synthetic accessibility of the phase fields, while the existing vast chemical knowledge is learnt each time in the context of the specific target material function.

In our approach, the machine learns all atomic elements and their specific characteristics responsible for materials formation. This is achieved by exploring possible compositional combinations in all theoretically and experimentally studied materials<sup>16</sup>, similarly to the concept in reference<sup>17</sup>. For each atom, a machine learns a vector, that encodes atomic characteristics learnt from the co-existence of atoms within some compositional environments and the absence of such co-existence with others. Thus built atomic vectors are then combined linearly to form a phase field representation, whereas attention mechanism<sup>18</sup> is trained to derive the weights to the

atomic vectors that magnify the most prominent atomic contributions specific to a particular property. This offers a statistically-derived alternative to the expert knowledge-based manual selection of relevant chemical characteristics and their contributions, and enables the high-level ranking and classification of materials for functional applications. Furthermore, by aggregating compositions into the phase fields in the input data, this high-level approach eliminates concerns of data leakage at the compositional level as all compositions within a phase field represent a single data entry.

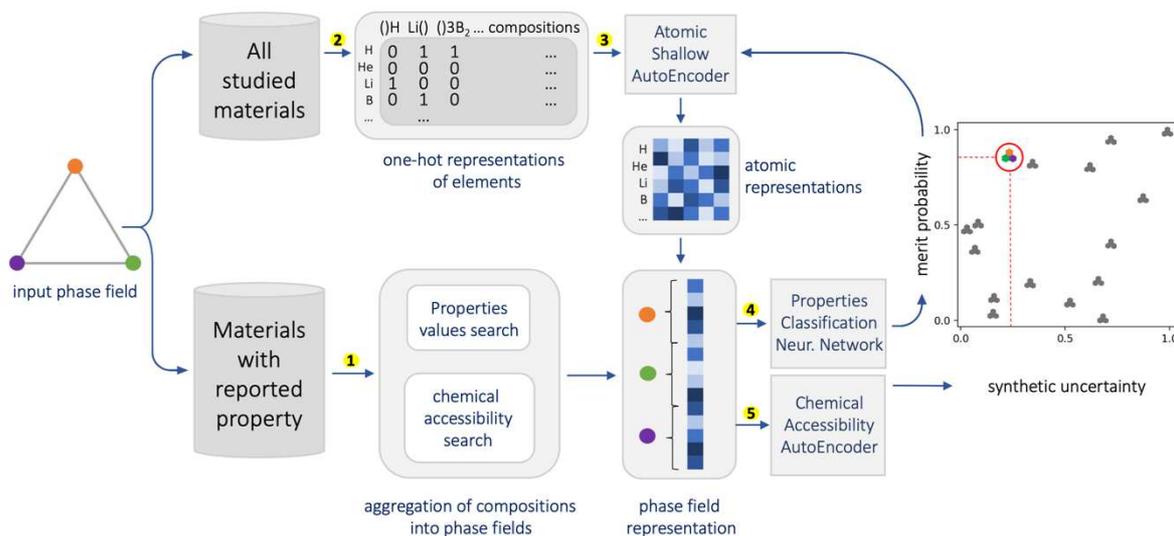
We demonstrate a significant accuracy of PhaseSelect in classification of the materials with respect to three different properties: superconducting transition temperature, Curie temperature, and energy band gap, when learning the relevant property from SuperCon<sup>3</sup> and Materials Platform for Data Science (MPDS)<sup>1</sup> databases. Within these training and test sets, each phase field is labelled according to the maximum reported value of all materials within it. This maximum value is compared to the chosen thresholds (10K, 300K, 4.5eV) that reflect practical interests in high-temperature superconducting, magnetic materials and dielectrics respectively, and a class label is allocated accordingly.

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 field 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. The arising metrics of the phase fields – the merit probability (probability of achieving a high value of a property) and synthetic uncertainty (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

At the level of the phase fields, relationships between elemental combinations and their synthetic accessibility have been studied with unsupervised machine learning and validated experimentally<sup>14</sup>. Here, we employ an integrated statistical description of atomic elements and their combinations to learn what elemental combinations have high probabilities of both synthetic realization and high values of target properties. The architecture of the model 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: 1) experimentally confirmed compositions are aggregated into the phase fields; the maximum values of the properties in the phase fields are selected; 2) compositional environments (elemental co-occurrence in materials) are aggregated from all theoretically and experimentally studied materials; 3) unsupervised learning of atomic representation from data collected in 2); 4) supervised classification of phase fields by maximum achievable values of the properties; the predicted probability of entering the high-value class is used as a merit probability; 5) unsupervised ranking of the phase fields by synthetic uncertainty; metrics derived in 4) and 5) result in a map

of the phase fields' likelihood to form stable compounds with desired properties. The model is trained end-to-end so the losses of learning the atomic representation (3) and classification (4) are minimised simultaneously.

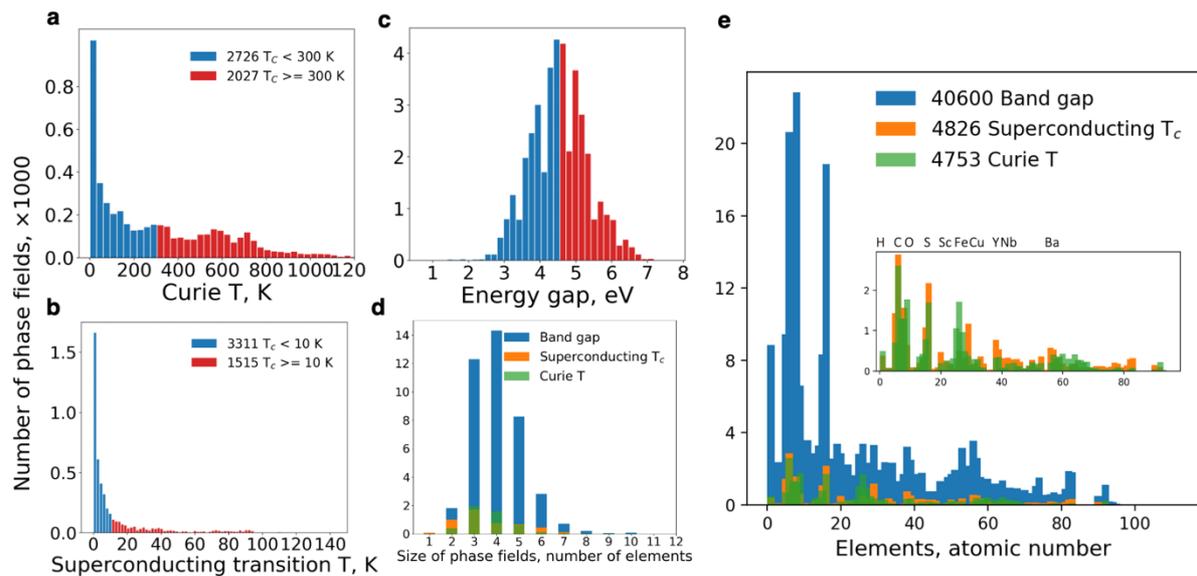
PhaseSelect consists of several connected modules (depicted as the sharp-corner rectangles in Figure 1) that pass information from the databases, while transforming the data (different data representations are depicted as the rounded-corner windows in Figure 1) and are trained simultaneously, while minimising the compound loss. We describe the data processing and the mechanisms of these modules in the following sections.

### Aggregation of compositions into phase fields

For the classification and accessibility ranking of the phase fields (See bottom stream in Figure 1) we process the materials databases, where experimentally verified values of the target property are reported for a large number of compositions<sup>1,3</sup>. Materials built from the same constituent elements are aggregated into one phase field, with the associated property value corresponding to the maximum reported property value among all reported materials within this phase field. For example, in the SuperCon database, there are many compositions reported in Y-Ba-Cu-O phase field with a high critical temperature, including  $\text{YBa}_2\text{Cu}_3\text{O}_7$  ( $T_c = 93$  K) and  $\text{Y}_3\text{Ba}_5\text{Cu}_8\text{O}_{18}$  ( $T_c = 100.1$  K) – the highest reported temperature in Y-Ba-Cu-O. Hence, Y-Ba-Cu-O enters the data for training our classification model for superconductors with 100.1 K as the corresponding maximum value. Aggregation of materials with reported superconducting transition temperature, Curie temperature and energy band gap forms three 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. Furthermore, the remaining imbalances are

taken into account by class-weighting in the corresponding classification models<sup>19</sup>. The 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)<sup>9</sup>. 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 the two classes divided by the chosen threshold value would allow fast screening for novel high-temperature superconductors. Similarly, a binary classification enables fast screening of novel materials for applications as high-temperature magnetic materials and targeted band gap materials.

Across the three property datasets, the phase fields are formed from up to 12 constituent elements, with the majority of data represented by ternary, quaternary and quinary phase fields (See Figure 2d). The 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, sulphur, with an especially pronounced match between elemental distribution in datasets with materials for superconducting and magnetic applications (See inset in Figure 2e). The data distributions across different chemical elements observed in Figure 2e, reflect the biases in the input data: e.g., magnetism is associated with Fe predominantly, while superconductivity with Cu, etc.



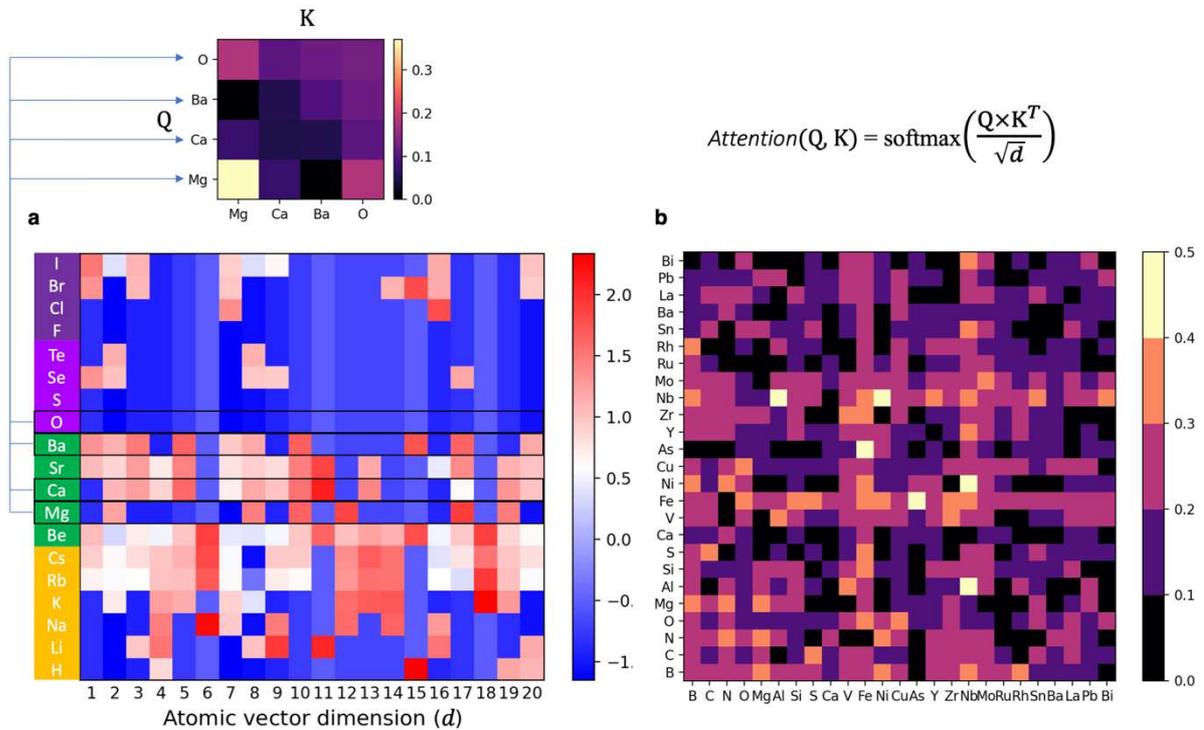
**Figure 2. Aggregation of compositions into phase fields.** **a** Distribution of phase fields of magnetic materials in MPDS<sup>1</sup> with respect to the maximum associated Curie temperature  $T_C$ . The materials' classes "low-temperature" and "high-temperature" magnets are divided at  $T_C = 300$  K as 2726:2027 phase fields. **b** Distribution of phase fields of superconducting materials (joined datasets from SuperCon<sup>3</sup> 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** Distributions of materials with respect to the number of constituent elements are similar for all datasets: the majority of the reported compositions belong to ternary, quaternary and quinary phase fields. **e** Content of individual chemical elements among the explored materials in the databases; the total numbers of phase fields in the corresponding datasets are given in the legend. All datasets have similar trends with pronounced peaks for materials containing, e.g. carbon, oxygen, silicon. The inset illustrates overlap in trends for elemental distribution in explored materials for superconducting and magnetic applications.

Description of atoms by means of their compositional environments, which are shared by chemically similar elements, should mitigate the biases in the data accumulated over time due to the focused studies of particular families of materials.

## Atomic representation and phase field representation

To learn atomic characteristics from the compositional 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<sup>16,17</sup>. For each chemical element one can build a one-hot encoding vector from its instances in the database. The database is expanded into a table similarly to the approach proposed in reference<sup>17</sup> (depicted as a matrix of coexisting elements and compositional environments in the materials in Figure 1, 2)). The rows of the table correspond to the chemical elements, the columns are the remainders of the compositional formulas of the reported compounds, which we define here as compositional environments. For example, from stability of  $\text{Li}_3\text{PO}_4$  we can learn about its constituent elements, Li, P, O and their compositional environments, “()3PO4”, “()Li3O4” and “()4Li3P” respectively. In this notation, empty parentheses denote an element that by combining with the compositional environment forms a composition. Similarly, all alkali metals form the tri-“element” phosphates with “()3PO4”, while trivalent elements do not, as they form the one-“element” phosphates with “()PO4” instead. In the proposed matrix representation<sup>17</sup>, the intersections of the rows for elements with the columns for compositional environments are filled with ones if the resulting composition is reported in<sup>16</sup> and with zeros otherwise. The resulting sparse matrix represents coexistence of the chemical elements and compositional environments in the materials. 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 dimensionality  $k$ , in which similar atomic vectors (of length  $k$ ) are grouped close to each other. We study the effects of the size of dimensionality  $k$  of thus derived atomic vectors on the classification accuracy to select the most efficient atomic description (Supplementary Fig. 1).

We use the vectors of the most efficient latent space as atomic representations to build up the phase fields descriptions (Figure 3a).



**Figure 3. Atomic representations and their contributions to the phase fields' properties.** **a** Atomic representation vectors in  $k = 20$  dimensions for the 1<sup>st</sup>, 2<sup>nd</sup>, 16<sup>th</sup> and 17<sup>th</sup> atomic groups of the periodic table. The values (corresponding colour) illustrate differences and correlations between derived 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 compositional environments built for the Materials Project database<sup>16,17</sup>; for an example unexplored quaternary phase field, O-Ba-Ca-Mg, the corresponding contributions of the atomic elements to the likelihood of high-temperature superconductivity of this combination are calculated as the attention scores<sup>18</sup> (Supplementary Fig. 2-6). **b** Attention scores are trained during the fitting of the model for phase fields classification by the target property. Here, attention to the atomic contributions to superconducting behaviour is visualised: combinations with e.g., Fe, Nb, Cu, Ni, Mo receive high attention in prediction of high-temperature superconductivity.

To emphasise the differences in the contributions of individual atoms to the phase field's properties, we employ the multi-head local attention<sup>18</sup> that calculates the attention scores – weights for the constituent atomic vectors contributing to the accuracy of the phase field classification for the target property. The attention scores are derived during the training and highlight the intermediate and interpretable results of the ML reasoning process well-aligned with the human understanding of chemistry of materials (See Figure 3b, Supplementary Fig. 2-6). When building a phase field representation for the 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  $(n \times k)$ -dimensional vector, where  $n$  is a number of constituent elements in a phase field,  $k$  is a chosen length of the atomic vector.

Classification by properties' values and ranking by synthetic accessibility

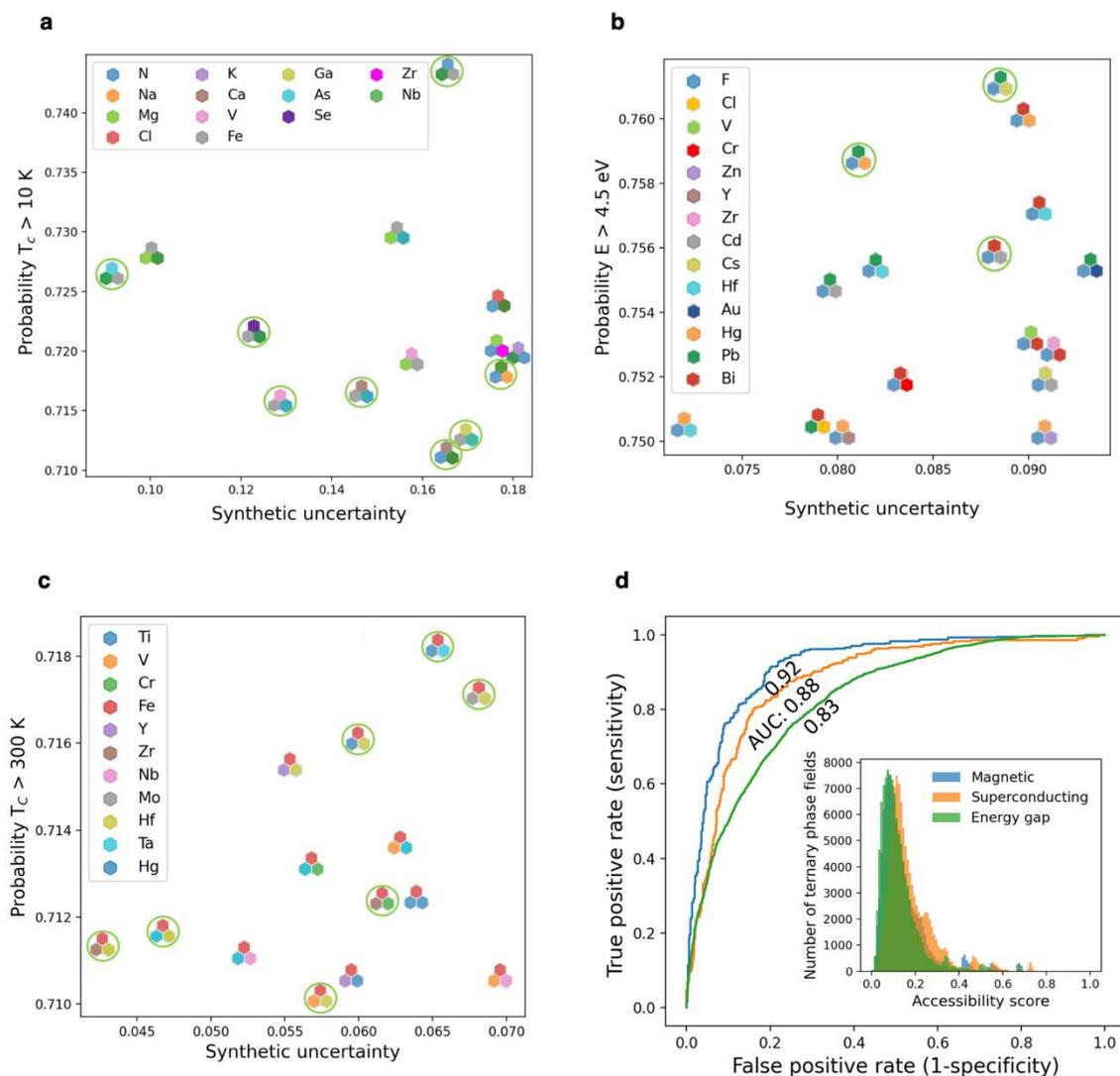
Classification in PhaseSelect is performed by a deep neural network (NN) that assigns the phase fields representation vectors to the corresponding classes of the properties' values. The phase fields in each dataset are divided into two classes (Figure 2a-c) that are labelled with '1' for the phase fields with associated property values above the chosen thresholds, and with '0' for the remaining phase fields. Three different classification models, one for each dataset - for superconducting materials and magnetic materials, and materials with a reported value of energy gap - are trained end-to-end with the architecture described in Fig. 1. Because the atomic characteristic and their relation to the materials properties are learnt from the reported chemistry, where the reports of the negatives (materials not possessing certain properties) are absent, the classification models are not trained to predict manifestation of target properties or their absence. Instead, for the phase fields that may contain compositions with target properties, the classification models predict the probability of reaching high values of these properties within the phase fields. For example, in the training set for the materials with reported values of energy gap, none were reported with zero value (Fig. 2c). To verify the predictive power of the model

trained on such data for the energy band gap classification, we have tested all 9816 intermetallic ternaries that do not have energy band gap values reported in MPDS (Supplementary discussion). 99.96% of the intermetallic ternary phase fields were classified as low energy gap materials (<4.5 eV) demonstrating the model's ability to extrapolate chemical patterns of atomic combinations – properties relationships, in absence of the zero-gap examples. On the other hand, this demonstrates vast generalisation of a model for the data regions where information is lacking.

The validation of the trained models is performed in the 5-fold cross-validation, where 5 models are trained on different 80% portions of the available data, with the remaining 20% used for testing. The average accuracy across the validation sets is 80.4, 86.2, 75.6 % for classification with respect to superconducting transition temperature, Curie temperature, and energy gap respectively. The validation datasets are used to tune the parameters of the NN models, such as dropout<sup>20</sup>, learning rate, activation<sup>21</sup>, early stopping<sup>19</sup> and stochastic optimisation algorithm<sup>22</sup>. 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 (Supplementary Fig. 10). 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 (Figure 4d, Supplementary Fig. 9, Supplementary Table 1), but also by the continuous values of probabilities as a measure of likelihood of achieving a desired property value. The latter helps to prioritise the materials for synthesis and further investigation.

In parallel to the classification module, a deep AutoEncoder neural network learns patterns of chemical accessibility from the experimentally verified materials data. Similarly to the procedure in <sup>14</sup>, an unsupervised de-noising AutoEncoder learns the patterns of similarity in data while

reducing dimensionality of the phase fields representations. The training consists of two parts: encoding into a reduced dimensionality latent space, where phase fields representations are reorganised, so the similar phase fields are aligned, and decoding from the latent representation into the reconstructed images of original vectors. 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. Hence, elemental combinations that are unlikely to manifest conventional bonding chemistry nor to form synthetically accessible compositions exhibit high reconstruction errors<sup>14</sup>. We also find that predicted reconstruction errors converge to their average values when an ensemble of models is trained (See Supplementary Fig. 10b).



**Figure 4. Probability of high-values properties and synthetic accessibility for unexplored ternary phase fields. Materials reported in ICSD<sup>2</sup>, for which property values are not in SuperCon-v2018<sup>3,9</sup> or MPDS<sup>1</sup> are circled. 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 with synthetic uncertainty (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 an 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 ternary phase fields with respect to their synthetic accessibility scores for all three datasets.

By applying the trained ensembles of models to 105995 ternary phase fields (Supplementary discussion) and focusing on the unexplored materials that do not have any related compositions with reported properties in MPDS or SuperCon-v2018, 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. We also highlight the phase fields, where compositions were synthesized and reported in ICSD, but for which there are no information about the properties discussed here in Supercon or MPDS, hence these phases fields did not enter the data for training. The large number of such phase fields among the top-performing candidates with respect to the measure of synthetic accessibility provides verification of the developed models and demonstrates that highly ranked candidates are likely to produce thermodynamically stable materials observed experimentally (See Figure 4a-c). We report the full list of likely candidates for novel superconducting materials among the phase fields that have been reported to form stable compounds in ICSD, but were not investigated from the perspectives of superconducting applications in <sup>23</sup> and its excerpt in Supplementary Table 7.

The top-performing phase fields according to both probability of exhibiting high values of properties and synthetic accessibility rank demonstrate trends produced by 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 the 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 contain stable compounds with energy gap of more than 4.5 eV, what can be expected from the simple bonding considerations as the majority of the latter are fluorides.

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.

## **Conclusions**

Selection of elements is the cornerstone of the materials design. Quantitative assessment of the potential properties of the prospective materials at the level of their constituent elements mitigates the high risk of the consequential decisions in elaborate research of materials discovery. Classification of the materials for functional applications agglomerated into phase fields is also a route to the several orders of magnitude reduction of the combinatorial space. The 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 PhaseSelect at the stage of conceptualization of the materials synthesis, human researchers can make use of numerical guidance in the selection of chemical elements that are most likely to produce new stable

compounds with high probability of superior functional properties, combining this statistically derived quantitative information with the expert knowledge and understanding. The attention mechanism of PhaseSelect presents a route to interpretation of the machine learning for materials science and allows extrapolation of the knowledge of materials databases to the large number of unexplored phase fields. These include multi-elemental materials, with prospective performance that could not be computationally assessed at scale with the methods developed to date.

## Methods

In this work, we adopt the same architecture and hyperparameters for all three problems investigated.

For unsupervised learning of atomic representations the shallow autoencoder has a single hidden layer with ReLU activation<sup>21</sup>, and sigmoid activation for the decoder layer. Effects of different numbers of nodes on the model performance are studied (Supplementary Fig. 1). The binary cross-entropy loss is added to the total loss of the model.

We employ 8-head attention<sup>18</sup> for learning the weights for atomic contributions in the phase fields representation and a padding mask for length justification of phase fields with different number of atoms. Multi-head approach ensures stabilisation of the training and improvement of the performance. For classification neural network we use 4 hidden layers with 80, 80, 80, 20 nodes respectively, with ReLU activation, L1 = 0.03, L2 = 1e-4 regularizations and 0.5 dropout<sup>20</sup>.

The ranking AutoEncoder is built with 4 hidden layers for encoder with decreasing number of nodes, 1/2, 1/4, 1/8 and 1/16 respectively of the initial length of a phase field vector, 4-dimensional latent representation and 4 hidden layers in decoder with increasing number of nodes, 1/16, 1/8, 1/4 and 1/2 of the initial length of a phase field vector. Each AutoEncoder hidden layer is followed by 0.1 dropout and activated with ReLU.

For the training, we employ Adam optimisation<sup>22</sup> with constant 1e-5 learning rate. During 100 epochs with batches of 16, we monitor the accuracy and ensure early stopping with 7 epoch patience.

## **Acknowledgements**

We thank the UK Engineering and Physical Sciences Research Council (EPSRC) for funding through grants number EP/N004884 and EP/V026887. D. A, V. G. and M.W.G. thank the Leverhulme Trust for funding via the Leverhulme Research Centre for Functional Materials Design.

## **Data availability**

The raw data used in this study is available at <https://www.github.com/lrcfmd/PhaseSelect>. The distribution of the phase fields' rankings, computed phase field's probability data generated in this study are available via University of Liverpool data repository at <https://doi.org/10.17638/datacat.liverpool.ac.uk/1613>

## **Software availability**

The software developed for this study is available at <https://www.github.com/lrcfmd/PhaseSelect>.

## **Competing Interests Statement**

The authors declare there are no competing interests.

## Supporting Information

Machine Learning methodology and models, Training set for the Classification, Model validation, Atomic attention scores, Tools and libraries.

## References

1. Villars, P., Cenzula, K., Savvysyuk, I. & Caputo, R. Materials project for data science, <https://mpds.io>. (2021).
2. Zagorac, D., Müller, H., Rühl, 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](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. Ren, Z. *et al.* An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **5**, 314–335 (2022).
7. Jha, D. *et al.* ElemNet : Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.* **8**, 1–13 (2018).
8. 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).
9. Stanev, V. *et al.* Machine learning modeling of superconducting critical temperature. *NPJ Comput. Mater.* **4**, 1–14 (2018).

10. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
11. Polikar, R. Ensemble Learning. in *Ensemble Machine Learning: Methods and Applications* (eds. Zhang, C. & Ma, Y.) 1–34 (Springer US, 2012). doi:10.1007/978-1-4419-9326-7\_1.
12. Oviedo, F., Ferrer, J. L., Buonassisi, T. & Butler, K. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *arXiv:2111.01037 [cond-mat]* (2021).
13. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **23**, 18 (2021).
14. 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).
15. 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).
16. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
17. Zhou, Q. *et al.* Learning atoms for materials discovery. *PNAS* **115**, E6411–E6417 (2018).
18. Vaswani, A. *et al.* Attention Is All You Need. *arXiv:1706.03762 [cs]* (2017).
19. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. (2015).
20. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
21. Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). *arXiv:1803.08375 [cs, stat]* (2019).

22. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]* (2017).
23. Vasylenko, A. PhaseSelect: Element selection for functional materials discovery by integrated machine learning of atomic contributions to properties, <https://github.com/lrcfmd/PhaseSelect>. (2021).

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [PhaseSelectSupplementaryInformation.pdf](#)