

ARTICLE OPEN



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

Andrij Vasylenko¹, Dmytro Antypov¹, Vladimir V. Gusev^{1,2}, Michael W. Gaultois¹, Matthew S. Dyer¹ and Matthew J. Rosseinsky¹✉

The unique nature of constituent chemical elements gives rise to fundamental differences in materials. Assessing materials based on their phase fields, defined as sets of constituent elements, before specific differences emerge due to composition and structure can reduce combinatorial complexity and accelerate screening, exploiting the distinction from composition-level approaches. Discrimination and evaluation of novelty of materials classes align with the experimental challenge of identifying new areas of chemistry. To address this, we present PhaseSelect, an end-to-end machine learning model that combines representation, classification, regression and novelty ranking of phase fields. PhaseSelect leverages elemental characteristics derived from computational and experimental materials data and employs attention mechanisms to reflect the individual element contributions when evaluating functional performance of phase fields. We demonstrate this approach for high-temperature superconductivity, high-temperature magnetism, and targeted bandgap energy applications, showcasing its versatility and potential for accelerating materials exploration.

npj Computational Materials (2023)9:164; <https://doi.org/10.1038/s41524-023-01072-x>

INTRODUCTION

The conceptualisation of novel materials begins at the level of the periodic table with the 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 (e.g., {Cu, O, B}) leading to different materials (phases, e.g., Cu₂BO₄); the field of these potential realisations can be defined as a phase field (e.g., {Cu, O, B}). The choice of phase field ultimately determines the outcome of the synthetic work and the functional properties of the prospective materials. Hence, this high-level discrimination of which phase field to investigate is essential before significant resources are committed to the investigation of individual compositions within a phase field.

The fundamental differences between chemical elements result in a gamut of material properties in thousands of compositions accumulated in materials databases^{1–3}. These data have been exploited by a surge of machine learning (ML) methods aiming to predict material properties from the knowledge of their compositions and structures^{4,5}. Both structure- and composition-based approaches demonstrate powerful capabilities of ML for the acceleration of materials discovery^{6–8}. Searching for truly new materials, for which neither the composition nor the crystal structure is known beforehand, and open-ended approaches of curious formulations⁹, generative approaches¹⁰ and serendipity-based recommender systems¹¹ have been applied to navigate experimentation in the uncharted chemistry spaces. Exhaustive enumeration of unexplored compositions is impossible, making existing ML approaches based on extensive screening susceptible to missing a potential new material. For example, the synthesised compositions often differ from the computationally explored models^{12–14}; even small differences in composition can prove critical for materials properties^{15–17}. Hence tools to assess the

functional applicability of the phase field as a whole can be invaluable for navigating materials discovery. Moreover, as the quality of ML models is heavily dependent on the available data for training, composition-based models inherit the historical bias of the past research preferences towards particular materials families, such as extensive studies of cuprates as superconductors¹⁸. The imbalanced datasets are known to have detrimental effects on the model performance and the capability to extrapolate the patterns of composition–property relationships into unexplored chemistry^{18–21}.

This highlights the need for the evaluation of materials at the high governing level of the constituent elements prior to their compositional assessment. By aggregating materials into the phase fields, one retains the fundamental differences between the sets of chemical elements while eliminating the risk of missing promising compositions. Additionally, by consolidating compositions into phase fields, the presence of different material families in historical data is redistributed: data balance is improved, with materials represented uniformly in the datasets. This improves ML model accuracy and capability to extrapolate composition–property relationships into uncharted chemistry²⁰. This phase field level of approach has already shown merit, with the experimental realisation of new stable materials in phase fields prioritised using similar methods¹².

In this work, our goal is to assess the attractiveness of unexplored candidate inorganic functional materials at the level of the periodic table by identifying unexplored phase fields that are likely to contain these candidates. This circumvents the combinatorial challenge of exhaustive individual assessment of all possible compositions built from the chosen elements. Further, this workflow assists decision-making in new areas of experimental solid-state inorganic chemistry by prioritising which

¹Department of Chemistry, University of Liverpool, Crown Street, Liverpool L69 7ZD, United Kingdom. ²Department of Computer Science, University of Liverpool, Ashton Street, Liverpool L69 3BX, United Kingdom. ✉email: rossein@liverpool.ac.uk

elements to combine in a reaction; currently, this is the only ML tool addressing this challenge. The high-level prioritisation aims to provide a computationally undemanding guide for research at its earliest stage, which is only a part of the total materials discovery challenge, as there are undiscovered phases within partially explored phase fields. The proposed unexplored phase field prioritisation can be followed by more computation- and data-intensive investigations of materials. The guidance is broadly applicable to unexplored inorganic materials, where a change of constituent elements plays a determining role in stability and function; the same approach does not apply to the prioritisation of synthetic routes, nor in organic chemistry, which may be more suitably addressed by other computational guides^{22,23}.

We present an end-to-end integrated (from elemental representation to the phase fields assessment) machine learning approach, PhaseSelect, that can prioritise phase fields with respect to both functional performance (e.g., the maximum value of a target property within a phase field) and chemical similarity (*i.e.*, similarity to phase fields with stable compositions).

PhaseSelect starts with semi-supervised learning of representations for chemical elements from the elemental co-occurrence in all calculated and experimentally reported materials (inspired by the approach in reference) coupled with the supervised assessment of materials' functional performance – regression and binary classification. The coupling is achieved through the 'attention' weighting of the contributions of constituent chemical elements to the functional performance of the material. The attention mechanism originates from computer science research in natural language representation²⁴ and is implemented in our model to learn the elemental representations that align best with the resulting representation of a phase field, such that its functional performance is quantified most accurately.

We demonstrate the predictive power of PhaseSelect in quantitative assessments of phase fields with three functional properties of interest: superconducting transition temperature, Curie temperature, and bandgap energy. The models for each dataset are trained independently; composition-level data with associated properties from SuperCon³ and/or Materials Platform for Data Science (MPDS)¹ databases is first aggregated into collections of phase fields, and each phase field is labelled according to the maximum reported value of all materials within it.

In a regression task, we verify (a) the viability of the description of the material simply as elemental sets and (b) PhaseSelect's capability to learn informative phase field representations while predicting a maximum value achievable within phase fields. In binary classification, we discriminate materials with respect to performance thresholds, which we define for each property ($T_c = 10$ K, $T_c = 300$ K, $E_{\text{gap}} = 4.5$ eV) that reflect practical interests in high-temperature superconductors, magnetic materials, and dielectrics/ruling out candidates for photovoltaics. In both regression and classification, PhaseSelect demonstrates significant improvement of performance in comparison to the baseline model – default random forest²⁵ with Magpie descriptors of elements²⁶ – by 1.5% MAE \times (value range)⁻¹ and 0.1 AUC, on average across 3 datasets. We combine binary classification with regression to first assign phase fields to low- or high-performing classes of materials and then predict the maximum expected values of a property of interest. This develops reliable quantitative metrics for fast high-level discrimination and screening of materials phase fields at scale.

The phase field representations constructed during property classification are used further for unsupervised learning of similarity between elemental combinations in materials databases that afford stable compositions. This stage completes the end-to-end assessment of elemental combinations by producing the ranking of chemical novelty for unexplored phase fields.

The arising metrics of the phase fields – functional performance (quantified by regression and the classifier probability of belonging to a high-performance class) and chemical novelty (quantified by distance in representation space from phase fields with stable compositions) – can be coupled or used independently for any combination of elements, creating a map of potentially attractive phase fields for future research. This can provide quantitative guidance to human researchers in the consequential and costly choice of phase fields for the investigation and discovery of novel 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 experimentally¹². Here, we employ an integrated statistical description of chemical elements and their combinations to learn what elemental combinations have high probabilities of both novelty and high values of target properties. PhaseSelect architecture combines several artificial neural networks (ANN) that are trained end-to-end as an integrated model, which we describe at the high level in Eqs. (1) and (2) and each of the components in more detail in Eqs. (3)–(8):

$$S: \mathbb{R}^m \rightarrow \mathbb{R}^1, S(\mathbf{p}) = R \circ W \circ A(\mathbf{C}) \cap \mathbf{p} \quad (1)$$

$$P: \mathbb{R}^m \rightarrow \mathbb{R}^1, P(\mathbf{p}) = P(\bar{\mathbf{p}}) \quad (2)$$

where S is a supervised model for classification (regression), P is an unsupervised model for chemical similarity ranking; a phase field, \mathbf{p} , of dimensionality m (e.g., $m = 3$ for ternary) is encoded via semi-supervised model, A , which learns representations for chemical elements from the matrix of elemental co-occurrence, \mathbf{C} , in all calculated and experimentally reported materials (inspired by the approach in reference); representation learning is guided by supervised assessment (classification or regression), R , of materials' functional performance; learnt representations for phase fields, $\bar{\mathbf{p}} = W \circ A(\mathbf{C}) \cap \mathbf{p}$, minimise error of R and can be further used for unsupervised learning, P , of chemical similarity (and inversely, novelty) of phase fields in Eq. (2); the coupling between elemental and phase field representations is achieved through the 'attention' weighting, W , of the contributions of constituent chemical elements to the functional performance of the material; by symbols \circ and parenthesis in Eqs. (1) and (2) we denote connectivity of the data flow and transformation of data. The architecture of the model is illustrated in Fig. 1.

PhaseSelect consists of several connected modules (ANN depicted as the sharp-corner rectangles in Fig. 1) that pass information from the databases (dark grey cylinders in Fig. 1) while transforming the data (different data representations are depicted as the rounded-corner windows in Fig. 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 assessment of materials at the level of phase fields (see bottom stream in Fig. 1), we process the materials databases, where experimentally verified values of the target property are reported for a large number of compositions^{1,3}. Some historical bias is present in every representation of materials data. By representing materials as phase fields, the bias can be decreased via uniform representation of all phase fields, as described below. 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

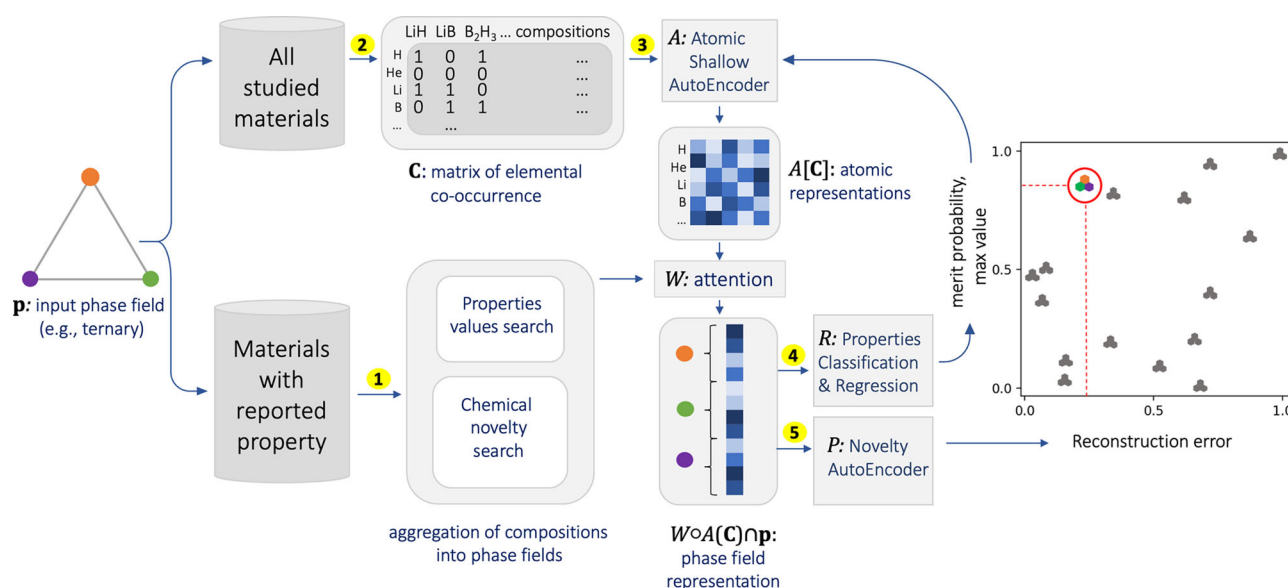


Fig. 1 PhaseSelect predicts properties and chemical accessibility of phase fields. Model architecture. Arrows show the information flow between the various components described in this paper: (1) experimentally confirmed compositions are aggregated into the phase fields; the maximum values of the properties 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 elemental 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 probability of high functional performance (merit probability); regression to maximum value achieved within phase fields (5) unsupervised ranking of the phase fields by similarity to synthetically stable materials; 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 elemental representation (3) and classification (4) contribute equally and are minimised simultaneously.

all reported materials within this phase field. Here, we focus on the maximum value because this is most likely to draw attention to a field. However, it is possible to evaluate other aggregate values, such as mean to draw out other aspects of the composition spaces. For example, in the SuperCon database, there are many compositions reported in the 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 and regression models for superconductors with 100.1 K as the corresponding maximum value. Aggregation of materials with reported superconducting transition temperature, Curie temperature and energy bandgap forms three datasets with 4826, 4753 and 40,452 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 bandgap, respectively – forms reasonably balanced data classes with 3311:1515, 2726:2027 and 20910:19690 phase fields, respectively, with data distributions illustrated in Fig. 2a–c. The data balance is further improved by class-weighting in the corresponding classification models²⁷. The rapidly decreasing number of explored phase fields with reported superconducting properties at temperatures above 10 K (See Fig. 2b) proves the development of reliable models for classification with respect to temperatures higher than 10 K particularly challenging (see Supplementary Fig. 1)²⁸. 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 a fast screening of novel materials for applications such as high-temperature magnetic materials and targeted bandgap 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 Fig. 2d). The abundance of chemical elements among the explored materials in the databases is illustrated in Fig. 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 Fig. 2e). The data distributions across different chemical elements observed in Fig. 2e reflect the biases in the input data: e.g., magnetism is associated with Fe predominantly, while superconductivity with Cu, etc.

Description of materials as sets of constituent chemical elements should help mitigate the biases in the data accumulated over time due to the focused studies of particular families of materials: both understudied (e.g., single composition) and established phase fields (e.g., multiple compositions, solid solutions, etc.). However, phase field representation of materials cannot completely eliminate the bias related to factors including historical interest, availability of particular chemical elements and similarity of the chemistry studied to that of minerals (Fig. 2e). A historical trend in materials research may also be suggested by weak correlations of property variance with maximum values and with the number of compositions reported in a phase field (see Supplementary Fig. 16). At this level of description, we ignore the property variance and focus on the maximum values achievable by materials built from selected chemical elements.

Elemental representation and phase field representation

To learn elemental characteristics from the compositional environments – explored chemical compositions, where the chemical elements are found to form a variety of stable and metastable materials – we build a module for elemental representation based on a large materials database that includes both experimental and theoretical materials^{29,30}. For each chemical element, one can build a binary vector indicating its

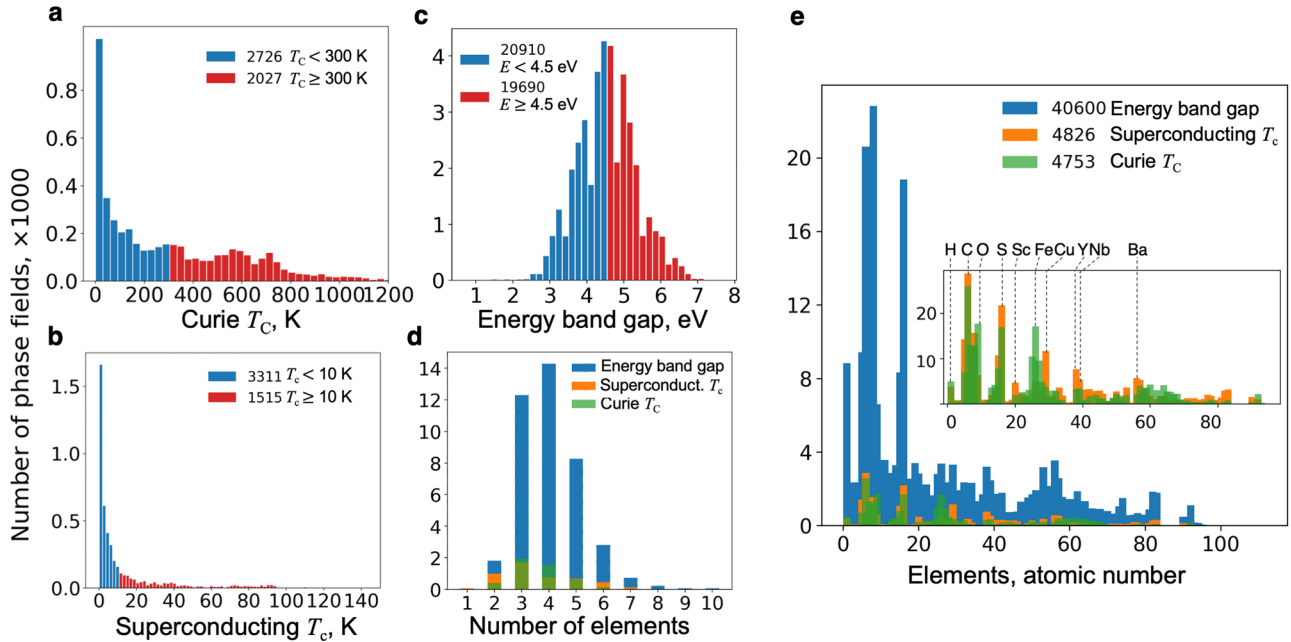


Fig. 2 Aggregation of compositions into phase fields. **a** Distribution of phase fields of magnetic materials in MPDS¹ with respect to the maximum associated Curie temperature T_C . The materials' classes "low-temperature" and "high-temperature" magnets are divided at $T_C = 300$ K as 2726:2027 phase fields. **b** Distribution of phase fields of superconducting materials (joined datasets from SuperCon³ 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 a reported value of energy gap in MPDS with respect to the maximum associated bandgap. 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 and silicon. The inset illustrates an overlap in trends for elemental distribution in explored materials for superconducting and magnetic applications, where the peaks of the prevalent constituent elements are highlighted.

presence in chemical formulae in the database. The database is expanded into a table similar to the approach proposed in reference (depicted as a matrix of coexisting elements and compositional environments in the materials in Figs. 1 and 2)). The rows of the table correspond to the chemical elements, the columns are the remains of the compositional formulae of the reported compounds, which we define here as compositional environments. For example, from the stability of Li_3PO_4 we can learn about its constituent elements, Li, P, O and their compositional environments, "() 3PO_4 ", "() Li_3O_4 " and "() 4LiP ", 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 "() 3PO_4 ", while trivalent elements do not, as they form the one-"element" phosphates with "() PO_4 " instead. In the proposed matrix representation²⁹, the intersections of the rows for elements with the columns for compositional environments are filled with ones if the resulting composition is reported in reference and with zeros otherwise. The resulting sparse matrix, \mathbf{C} , represents the coexistence of the n chemical elements and d compositional environments in the materials. We then employ a shallow autoencoder neural network – an unsupervised ML technique – to reduce the dimensionality of the matrix \mathbf{C} :

$$A: \mathbb{R}^d \rightarrow \mathbb{R}^k, A(\mathbf{C}) = \sigma_{\text{ReLU}} \left(\sum_i \omega_i c_i + b_i \right), \quad (3)$$

which forms an encoder part of the autoencoder, where ω_i and b_i are weights and biases in ANN, c_i are rows of the matrix \mathbf{C} and σ_{ReLU} is ReLU activation function³¹. An autoencoder is employed to condense the information into the rich latent space of

dimensionality k , in which similar elemental vectors (of length k) are grouped close to each other; during the autoencoder training, the encoder A is tuned in conjunction with the mirrored-size decoder, D , to minimise the loss $\mathcal{L}(D, A)$:

$$D: \mathbb{R}^k \rightarrow \mathbb{R}^d, \mathcal{L}(D, A) = \|\mathbf{C} - D \circ A(\mathbf{C})\|^2, \quad (4)$$

which is the Euclidian distance between original and reconstructed elemental vectors in matrix \mathbf{C} .

We study the effects of the size of dimensionality k of thus derived elemental vectors on the mean absolute error of regression network and classification accuracy to select the most efficient elemental description (Supplementary Fig. 1). We use the vectors of the most efficient latent space as elemental representations to build up the phase fields descriptions as matrices $\tilde{\mathbf{p}} = A(\mathbf{C}) \cap \mathbf{p}$ of size (m, k) , where by the intersection sign \cap we signify selection of the rows in $A(\mathbf{C})$ that correspond to the chemical elements in \mathbf{p} , m is a number of constituent elements in a phase field, and rows are the corresponding elemental vectors (Fig. 3a).

To emphasise the differences in the contributions of individual chemical elements to each of the studied phase field's properties, we employ the multi-head local attention, a particularly suitable technique for weighing the relevance of elements within a set when generating representations²⁴:

$$W: \mathbb{R}^m \rightarrow \mathbb{R}^m, W_h(\tilde{\mathbf{p}}) = \sum_i a_{\mathbf{q}\mathbf{k}_i}^h \mathbf{v}_{\mathbf{k}_i}^h, \quad (5)$$

where for each head, h , attention scores $a_{\mathbf{q}\mathbf{k}_i}^h = \frac{\exp(\mathbf{q}^h \cdot \mathbf{k}_i^h)}{\sum_j \exp(\mathbf{q}^h \cdot \mathbf{k}_j^h)}$, and $\mathbf{q}^h, \mathbf{k}^h, \mathbf{v}^h$ are different aspects representations of the phase fields: $\mathbf{Q}^h \tilde{\mathbf{p}}, \mathbf{K}^h \tilde{\mathbf{p}}, \mathbf{V}^h \tilde{\mathbf{p}}$, respectively. The aspects representations matrices

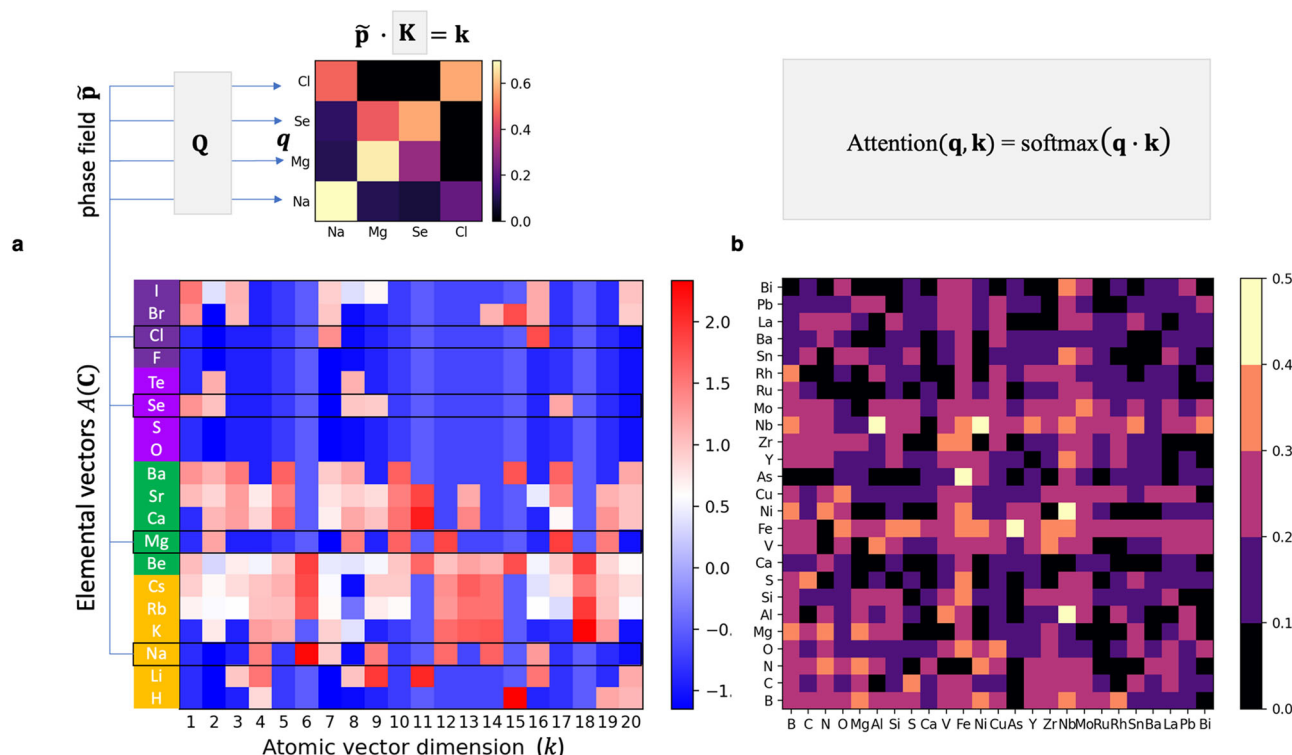


Fig. 3 Elemental representations and their contributions to the phase fields' properties. **a** Elemental representation vectors learnt via autoencoder $A(C)$ (Eq. (3)) in $k = 20$ dimensions for the 1st, 2nd, 16th and 17th atomic groups of the periodic table. The values (corresponding colour) illustrate differences and correlations between constructed elemental features (vectors' components) in the neighbouring chemical elements and groups. The full stack of elemental vectors for the whole periodic table is extracted by PhaseSelect's elemental autoencoder shallow neural network from the sparse matrix of chemical elements and compositional environments built for the Materials Project database^{29,30}; for an example, unexplored quaternary phase field, Na–Mg–Se–Cl, the corresponding contributions of the chemical elements to the likelihood of high-temperature superconductivity of this combination are calculated as the attention scores²⁴ (Supplementary Figs. 2–6). **b** Attention scores – a scaled matrix multiplication of the aspect representation matrices Q and K of a phase field \tilde{p} – are trained during the fitting of the model for phase fields classification by the target property (Eq. (5)). Here, attention to the elemental contributions to superconducting behaviour is visualised: elemental combinations that include, e.g., Fe, Nb, Cu, Ni, Mo receive high attention in the prediction of high-temperature superconductivity.

Q^h , K^h , V^h are trainable parameters that capture different areas (aspects) of phase field representations; they are initialised randomly. The resulting representation is obtained by concatenating the heads representations W_h . The attention scores $\sigma_{q_k}^h$ are the weights for the constituent elemental vectors contributing to the accurate prediction of the phase field values of a targeted property, these weights highlight the intermediate and interpretable results of the ML reasoning process well-aligned with the human understanding of the chemistry of materials (Fig. 3b, Supplementary Figs. 2–6).

From the calculated attention scores, one can infer elemental, pair and complex many-element contributions to the targeted functional properties. For example, one can obtain a distribution of attention scores for the constituent elements that affect a particular property the most across the phase fields (Supplementary Fig. 6) and the maps for pairs of elements which contribute the most in co-occurrence (Fig. 3b, Supplementary Figs. 3–5). From the latter, while some pairs may be familiar to human researchers, e.g., Nb–Al, Nb–Ni, Cu–O and Fe–As exhibiting high-temperature superconductivity, other less obvious pairs, e.g., N–Na, La–Cl and P–Si motivate further research into the chemistry of superconductivity. Furthermore, more complex correlations in ternary and quaternary phase fields, readily evaluated with our approach, are difficult to visualise and assess by simple statistics, which underlines the additional utility of this approach.

When building a phase field representation for the downstream tasks of property values assessments and chemical similarity

ranking, the element attention-weighted phase field matrix $W(\tilde{p})$ (Eq. (5)) is flattened to form a $(m \times k)$ -dimensional vector \bar{p} , where m is a number of constituent elements in a phase field, k is the chosen length of the elemental vector. Vector \bar{p} is then padded for length justification of phase fields with different numbers of chemical elements (see 'Methods').

Regression and classification by properties' values and ranking by chemical novelty

Supervised assessment of properties in PhaseSelect is performed by two separate ANN that are (a) a regressor that predicts targeted property values of phase fields (b) a classifier that assigns the phase fields to the corresponding classes of the properties' values:

$$R: \mathbb{R}^{m \times k} \rightarrow \mathbb{R}^1, R(\bar{p}) = \sigma_q \left(\delta_{\text{ReLU}} \sum_i [\delta_i \omega_i p_i + b_i] \right), \quad (6)$$

$$(a) \sigma_q(z) = \sum_i \omega_i z_i + b_i, \mathcal{L}(R, t) = \text{MAE}(R, t); \quad (7)$$

$$(b) \sigma_q(z) = \frac{1}{1 + \exp(-z)}, \mathcal{L}(R, t) = \text{BCE}(R, t), \quad (8)$$

where δ stands for a dropout function³², ω_i and b_i are weights and biases of ANN nodes, and p_i are elements of the phase field vector; loss functions of prediction R of the target value t , $\mathcal{L}(R, t)$, are mean absolute error (MAE) and binary cross entropy (BCE), for regression and classification respectively. The corresponding loss

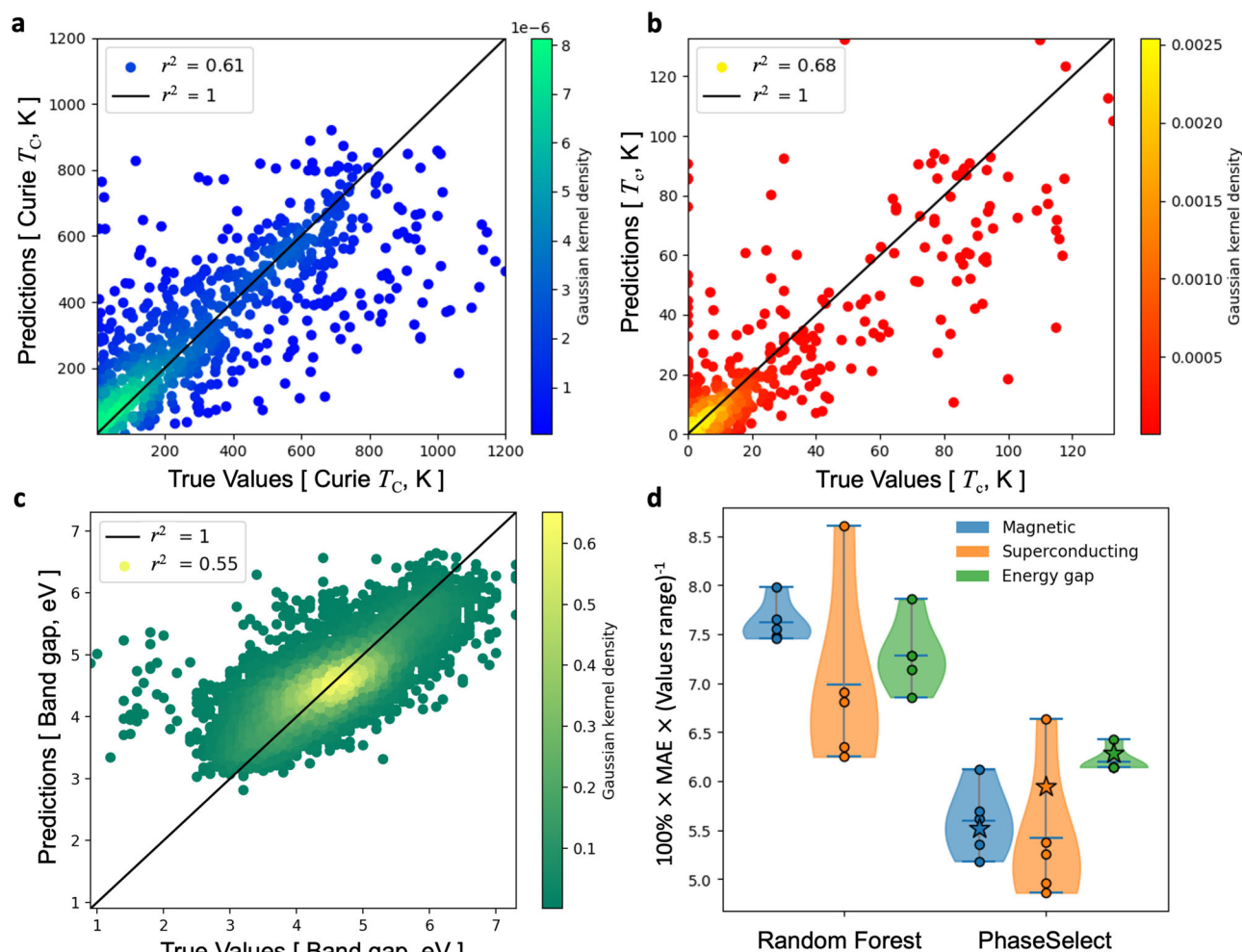


Fig. 4 PhaseSelect regressions of phase fields to targeted properties. The models are trained and tested on a random 80–20% train-test split of the datasets: MAEs and r^2 scores for the held-out 20% data (tests) are consistent with the average results in 5-fold cross-validations performed for the remaining 80% data. **a** Predictions vs true values for maximum Curie temperatures reported and peer-reviewed within phase fields in MPDS¹; **b** predictions vs true values for maximum superconducting transition temperatures experimentally reported within phase fields in Supercon³; **c** predictions vs true values for maximum energy bandgap reported and peer-reviewed within phase fields in MPDS¹; **d** 5-folds' MAEs scaled with a range of values in the corresponding database for default Random Forest²⁵ with Magpie²⁶ descriptors (see Supplementary Fig. 10) and PhaseSelect models; the error bars are the standard deviations of MAE achieved in 5 non-overlapping data folds in cross-validation; the star markers correspond to the MAEs of the tests (**a–c**).

functions are summed up with the representation learning loss $\mathcal{L}(A, D)$ (Eq. (4)) for simultaneous end-to-end training of representations and predictions: $\mathcal{L}_{\text{compound}} = \mathcal{L}(A, D) + \mathcal{L}(R, t)$.

For each dataset, we train an individual regressor with the architecture described above in Eqs. (1)–(7) and Fig. 1, in which elemental representations, phase field representations with attention to elemental contributions and predictions of a maximum achievable value for a particular property are trained end-to-end. For each dataset, we split the data so 20% is withheld for testing and perform fivefold cross-validation and then model training on the remaining 80% (Supplementary Fig. 7). In line with the best practices³³, we compare the performance of PhaseSelect with the baseline models. For the latter, we employ random forest²⁵ regression and classification models, whereas phase fields are described as $(m \times k)$ -dimensional vectors, with k Magpie elemental features²⁶ (see Supplementary Fig. 11). In Fig. 4, we illustrate the match between PhaseSelect predictions with true values and improved performance of PhaseSelect in comparison to the baseline models for all datasets studied.

The illustrated in Fig. 4a–c distributions of values, the corresponding MAE and r^2 scores are calculated for the held-out

20% of data for each property dataset, while PhaseSelect is trained on the remaining 80%; the metrics are characteristic for the results observed in k-fold cross-validations: MAE is within the standard deviations, including scaled MAE highlighted in Fig. 4d. Description of phase fields as a concatenation of elemental vectors and regression to the maximum values reported within the phase fields is a demonstrated a viable approach by Random Forest; its performance could be further enhanced via hyperparameter optimisation (see the comparison to dummy model regressions in Supplementary Fig. 11). The PhaseSelect approach further improves regression metrics for all studied properties (Fig. 4d), while capturing major trends in elemental combinations to property relationships.

For the guidance of synthetic combinatorial chemistry and acceleration of screening at scale, it is practically useful to first accurately assign candidates to the major clusters of performance before prioritisation of elements within the high-performing group. To ensure high accuracy of assignment, we employ binary classification of phase fields to the 'poor' and 'high' performing groups: the phase fields in each dataset are divided into two classes (Fig. 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 independent classifiers, 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 Eqs. (1)–(8) and Fig. 1. Because the elemental characteristics and their relation to the material's 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 bandgap classification, we have tested all 9816 intermetallic ternaries that do not have energy bandgap 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 elemental combinations – properties relationships in the absence of the zero-gap examples.

Similarly to the regressors, we study the performance of classification models in the fivefold cross-validations (Supplementary Fig. 7) and report the averaged results over the folds in accordance with the common benchmarking practice³⁴. The average accuracy across the validation sets is 80.4, 86.2 and 75.6% for classification with respect to superconducting transition temperature, Curie temperature and energy gap, respectively. For the predictive models, we adopt all available data in the three datasets for training. Noting the stochastic nature of the machine learning ANN, 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. 14). 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 (Fig. 4d, Supplementary Fig. 13), 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.

A deep AutoEncoder neural network in Eq. (2) learns patterns of chemical similarity with the experimentally verified materials data. Similar to our original approach reported in¹², an unsupervised denoising AutoEncoder learns the patterns of similarity in data while reducing the dimensionality of the phase field representations. The training consists of two parts, in principle equivalent to Auto-encoder in Eqs. (3) and (4): encoding into a reduced dimensionality latent space, where phase field 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 (i.e., combinations nonconforming with synthetically accessible compositions in training data) exhibit high reconstruction errors¹². By combining the ranking AutoEncoder with the classification of properties, we can transfer some advantages of supervised learning to unsupervised assessment of similarity between the phase fields. This is achieved by encoding the phase fields with the elemental representations that are learnt in the supervised setting of the end-to-end training.

By applying PhaseSelect to 105,995 ternary phase fields and focusing on the 90,029 unexplored phase fields (Supplementary Discussion) 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 the superconducting transition temperature, Curie temperature and energy bandgap, predict maximum values for these properties expected within the phase fields, and rank candidate phase fields by their reconstruction errors – degree of similarity with experimentally synthesised materials that are reported to exhibit these properties. We also highlight the phase fields, where compositions were synthesised and reported in ICSD, but for which there is no information about the properties discussed herein SuperCon or MPDS (encircled markers in Fig. 5a–c), hence these phase fields did not enter the data for training. The large number of such phase fields among the top-performing candidates provides verification of the developed models and demonstrates that highly ranked candidates are likely to produce thermodynamically stable materials observed experimentally. 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 (See the full list of candidates in³⁵ and its excerpt in Supplementary Table 7).

The top-performing phase fields, according to both the probability of exhibiting high values of properties and conformity with synthetically accessible materials, demonstrate trends produced by the constituent chemical elements: Mg, Fe and 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, which can be expected from 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 emphasising the models' capability to infer complex elemental characteristics in relation to properties 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. Such less expected examples include combinations of elements that do not exhibit ambient pressure or pressure-induced superconductivity as elemental solids³⁶, exclude Fe, Cu and rare-earth metals, known for forming families of superconducting materials, but are classified as high-temperature (>10 K) superconductors, when combined: C–Mg–Rb, Cr–K–N and As–C–Na among other 125 ternaries³⁵. For magnetic applications, all unexplored phase fields that were classified to exhibit magnetism at $T_c > 300$ K contain known magnetic elements Fe, Co, Ni, or Mn. However, we can highlight the combinations that are interesting also from the perspective of similarity to the synthetically accessible materials, including 4001 Fe-free ternaries³⁵, such as Co–Ti–Zr, Mo–Co–B and Hf–Mn–Ti. Among 13,070 large bandgap dielectric phases³⁵ not involving oxides nor fluorides, we can highlight Te–S–I and Ga–S–Cl.

The selection of elements as material components is the cornerstone of materials design, as their choice delimits all future outcomes in subsequent synthetic work. Quantitative assessment of the potential properties, including novelty ranking of the prospective materials at the level of their constituent elements, supports decisions regarding where to focus experimental synthetic effort. Classification of the materials for functional applications agglomerated into phase fields avoids the challenges of the common composition-based approaches, such as the exhaustive assessment of all possible combinations and in-cluster extrapolation without data leakage. Working at the level of phase fields is also a route to reducing the combinatorial space by several orders of magnitude.

The end-to-end integrated architecture of PhaseSelect represents and quantifies phase fields in two unrelated dimensions:

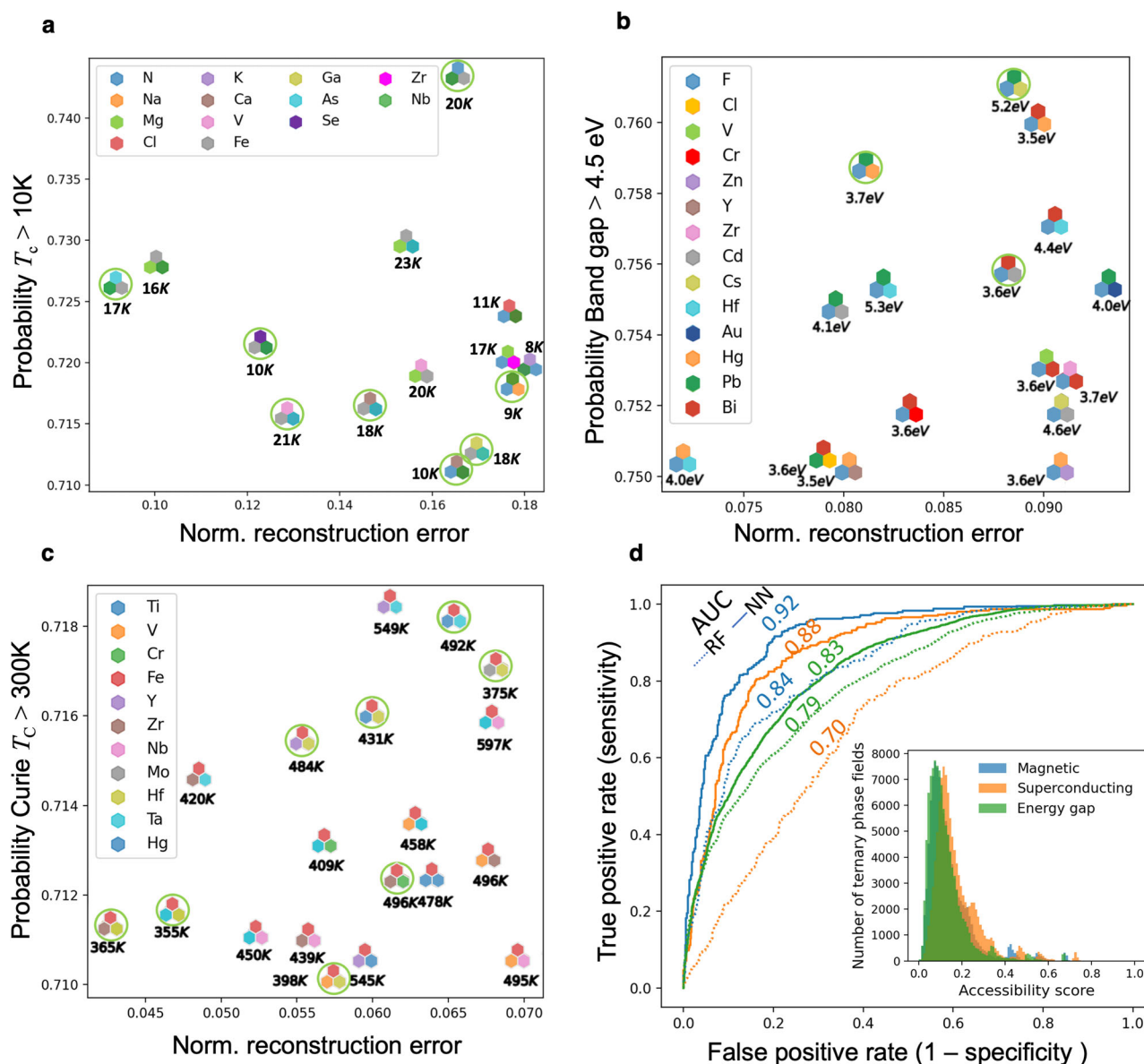


Fig. 5 Prediction of maximum property values and similarity with synthetically accessible materials. Circled materials are reported in ICSD², and are not found in SuperCon-v2018^{3,28} or MPDS¹. **a** Unexplored ternary phase fields with $>70\%$ probability of superconductivity at transition temperature $T_c > 10\text{ K}$ and normalised reconstruction error < 0.2 . **b** Unexplored ternary phase fields with $>75\%$ probability of energy band gap $> 4.5\text{ eV}$, and reconstruction error < 0.1 . **c** Unexplored ternary phase fields with $>71\%$ probability of Curie temperature $T_c > 300\text{ K}$ and reconstruction error < 0.1 . **d** Receiver operating characteristics (ROC) of the classification models trained and tested on the random 80-20% train-test splits demonstrate high sensitivity and specificity of classifications for the range of thresholds of probabilities. The corresponding areas under the curves (AUC) indicate excellent performance for magnetic materials and good performance for superconducting transition temperature and energy gap classifications. PhaseSelect considerably increases AUC for all datasets in comparison to default Random Forest classifiers trained on the same data. Inset: distributions of 105995 ternary phase fields with respect to reconstruction errors for all three datasets.

functional performance (via classification and regression) and similarity to synthetically accessible materials, which can be coupled or used independently for any combination of elements at scale to prioritise experimental targets. By employing PhaseSelect at the conceptualisation stage of material discovery and synthesis, human researchers can make use of numerical guidance in the selection of elements that are most likely to produce new stable compounds with a high probability of superior functional properties. This enables the combination of statistically derived quantitative information with expert knowledge and understanding to prioritise promising phase fields and de-risk material discovery. Finally, the attention mechanism of PhaseSelect

presents a route to the interpretation of machine learning for materials science and allows extrapolation of the knowledge of materials databases to a large number of unexplored phase fields. These include multi-element materials with prospective performance that could not otherwise be computationally assessed at scale with current methods.

METHODS

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

For unsupervised learning of elemental representations, the shallow autoencoder has a single hidden layer with ReLU activation³¹ and sigmoid activation for the decoder layer. The effects of different numbers of nodes on the model performance are studied (Supplementary Fig. 1).

We employ 8-head attention²⁴ for learning the weights for elemental contributions in the phase fields representation and a padding mask μ for length justification of phase fields with different numbers of chemical elements by a maximum-size phase field: $\mu: \mathbb{R}^{k \cdot m} \rightarrow \mathbb{R}^{k \cdot \max(m)}$, $\mu(\mathbf{p}_{k_1 \dots k_m}) = \mathbf{p}_{k_1 \dots k_m, z_{m+1} \dots z_{m+\max(m)}}$, $z_j = 0$. The multi-head approach ensures stabilisation of the training and improvement of the performance. For the classification neural network, we use 2 hidden layers with 80 and 20 nodes, respectively, with ReLU activation, $L1 = 0.03$, $L2 = 1e-4$ regularisations and 0.5 dropout³².

The ranking AutoEncoder is built with 4 hidden layers for the 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 the decoder with an 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 dropouts and activated with ReLU.

For the training, we employ Adam optimisation³⁷ with a starting learning rate of $1e-3$ and a scheduled decrease after every 100 epochs. During training, we monitor the accuracy (mean absolute error for regression) of the training data and randomly select 20% validation data and ensure early stopping with 7 (40) epoch patience, respectively.

Tools used for the implementation of these methods are listed in Supplementary Methods.

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 and computed phase field's probability data generated in this study are available via the University of Liverpool data repository at <https://doi.org/10.17638/datacat.liverpool.ac.uk/1613>.

CODE AVAILABILITY

The software developed for this study is available at <https://www.github.com/lrcfmd/PhaseSelect> and <https://doi.org/10.5281/zenodo.7464312>.

Received: 21 July 2022; Accepted: 25 June 2023;

Published online: 04 September 2023

REFERENCES

- Villars, P., Cenzula, K., Savysyuk, I. & Caputo, R. Materials project for data science. <https://mpds.io> (2021).
- 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).
- National Institute of Materials Science, Materials Information Station, SuperCon, http://supercon.nims.go.jp/index_en.html (2011).
- 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).
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- De Breuck, P.-P., Hautier, G. & Rignanese, G.-M. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *NPJ Comput. Mater.* **7**, 1–8 (2021).
- Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **11**, 6280 (2020).
- Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *NPJ Comput. Mater.* **7**, 1–10 (2021).

- Grizou, J., Points, L. J., Sharma, A. & Cronin, L. A curious formulation robot enables the discovery of a novel protocell behavior. *Sci. Adv.* **6**, 4237 (2020).
- Fuhr, A. S. & Sumpter, B. G. Deep generative models for materials discovery and machine learning-accelerated innovation. *Front. Mater.* **9**, 865270 (2022).
- Shekar, V. et al. Serendipity based recommender system for perovskites material discovery: balancing exploration and exploitation across multiple models. Preprint at *chemRxiv* <https://doi.org/10.26434/chemrxiv-2022-11wvf-v2> (2022).
- 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).
- Gamon, J. et al. Computationally guided discovery of the sulfide Li₃AlS₃ in the Li–Al–S phase field: structure and lithium conductivity. *Chem. Mater.* **31**, 9699–9714 (2019).
- Collins, C. et al. Accelerated discovery of two crystal structure types in a complex inorganic phase field. *Nature* **546**, 280–284 (2017).
- Telford, E. J. et al. Doping-induced superconductivity in the van der Waals superatomic crystal Re₆Se₈Cl₂. *Nano Lett.* **20**, 1718–1724 (2020).
- Budrikis, Z. Magnetism: doping rehabilitates failed materials. *Nat. Rev. Mater.* **3**, 1–1 (2018).
- Suo, Z., Dai, J., Gao, S. & Gao, H. Effect of transition metals (Sc, Ti, V, Cr and Mn) doping on electronic structure and optical properties of CdS. *Results Phys.* **17**, 103058 (2020).
- 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).
- Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 16028 (2016).
- Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data* **6**, 27 (2019).
- Korkmaz, S. Deep learning-based imbalanced data classification for drug discovery. *J. Chem. Inf. Model.* **60**, 4180–4190 (2020).
- Cova, T. F. G. & Pais, A. A. C. C. Deep learning for deep chemistry: optimizing the prediction of chemical patterns. *Front. Chem.* **7**, 809 (2019).
- Jorner, K., Tomberg, A., Bauer, C., Sköld, C. & Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nat. Rev. Chem.* **5**, 240–255 (2021).
- Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **31**, 5998–6008 (2017).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Jha, D. et al. ElemNet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 17593 (2018).
- Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. *Oper. Syst. Des. Implement.* **12**, 265–283 (2016).
- Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *NPJ Comput. Mater.* **4**, 1–14 (2018).
- Zhou, Q. et al. Learning atoms for materials discovery. *Proc. Natl Acad. Sci. USA* **115**, 6411–6417 (2018).
- Jain, A. et al. Commentary: The Materials Project: a materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).
- Agarap, A. F. Deep Learning using Rectified Linear Units (ReLU). Preprint at *arXiv* <https://arxiv.org/abs/1803.08375> (2019).
- 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).
- Artrith, N. et al. Best practices in machine learning for chemistry. *Nat. Chem.* **13**, 505–508 (2021).
- Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *NPJ Comput. Mater.* **6**, 1–10 (2020).
- Vasylenko, A. PhaseSelect: Element selection for functional materials discovery by integrated machine learning of atomic contributions to properties. <https://github.com/lrcfmd/PhaseSelect> (2021).
- Flores-Livas, J. A. et al. A perspective on conventional high-temperature superconductors at high pressure: Methods and materials. *Phys. Rep.* **856**, 1–78 (2020).
- Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at *arXiv* <https://arxiv.org/abs/1412.6980> (2017).

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.V.G. and M.W.G. thank the Leverhulme Trust for funding via the Leverhulme Research Centre for Functional Materials Design.

AUTHOR CONTRIBUTIONS

A.V. identified, developed and implemented the PhaseSelect model in discussion with D.A., V.V.G., M.W.G. and M.S.D. A.V., D.A. and M.J.R. wrote the first draft, and all authors contributed to the completion of the paper. M.J.R. directed the project.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-023-01072-x>.

Correspondence and requests for materials should be addressed to Matthew J. Rosseinsky.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023