npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning

Cameron J. Hargreaves[1], Michael W. Gaultois[1,2], Luke M. Daniels[1], Emma J. Watts[1,2], Vitaliy A. Kurlin[2,3], Michael Moran[1,2], Yun Dang[1], Rhun Morris[1], Alexandra Morscher[1], Kate Thompson[1], Matthew A. Wright[1], Beluvalli-Eshwarappa Prasad[1], Frédéric Blanc[1,2,4], Chris M. Collins[1], Catriona A. Crawford[1], Benjamin B. Duff[1,4], Jae Evans[1], Jacinthe Gamon[1], Guopeng Han[1], Bernhard T. Leube[1], Hongjun Niu[1], Arnaud J. Perez[1], Aris Robinson[1], Oliver Rogan[1,2], Paul M. Sharp[1], Elvis Shoko[1], Manel Sonni[1], William J. Thomas[1], Andrij Vasylenko[1], Lu Wang[1], Matthew J. Rosseinsky[1,2] and Matthew S. Dyer[1,2]✉

The application of machine learning models to predict material properties is determined by the availability of high-quality data. We present an expert-curated dataset of lithium ion conductors and associated lithium ion conductivities measured by a.c. impedance spectroscopy. This dataset has 820 entries collected from 214 sources; entries contain a chemical composition, an expert-assigned structural label, and ionic conductivity at a specific temperature (from 5 to 873 °C). There are 403 unique chemical compositions with an associated ionic conductivity near room temperature (15–35 °C). The materials contained in this dataset are placed in the context of compounds reported in the Inorganic Crystal Structure Database with unsupervised machine learning and the Element Movers Distance. This dataset is used to train a CrabNet-based classifier to estimate whether a chemical composition has high or low ionic conductivity. This classifier is a practical tool to aid experimentalists in prioritizing candidates for further investigation as lithium ion conductors.

npj Computational Materials (2023)9:9 ; https://doi.org/10.1038/s41524-022-00951-z

INTRODUCTION

Energy storage is a key technology to meet growing energy demand by harnessing renewable sources. Liquid electrolyte-based lithium ion batteries have been extensively deployed in the portable electronic and electric vehicle markets. Alternative batteries that utilize solid state electrolytes (SSEs) avoid the safety issues associated with organic liquid electrolytes and offer high energy density by enabling the use of a lithium metal anode[1,2]. The most significant obstacle to the adoption of SSEs is the realization of solid-state materials with the full suite of required properties, including sufficiently high ionic conductivity, stability against both lithium metal and the oxidizing cathode material (in practice this is often kinetic and associated with the formation of stable electronically insulating interfaces) together with appropriate mechanical properties[3]. As such, considerable research has been devoted to the discovery and development of SSEs that meet these requirements[4,5].

The amount of time and effort required to discover a suitable material in any domain has driven the application of machine learning methods to predict material properties[6]. Recent works have used previously published data[7,8] to train machine learning models and predict the ionic conductivity performance of materials using only their composition[9]. This approach is limited by the quality and quantity of the data available to train models. Literature reports in materials science tend to focus on subsets or particular families of materials with favourable or promising properties, leading to many reports on a limited range of materials[10,11]. While natural language processing (NLP) tasks have access to billions of training examples, in experimental materials science even large datasets typically contain fewer than 10,000 entries[12]. Due to these comparatively small training sets, it is imperative that the highest quality data are used to avoid providing inaccurate data to predictive models. As there are no large repositories of experimental ionic conductivities currently available for solid lithium ion conductors to perform a machine learning investigation, the first step must be sourcing high quality data.

Machine learning models for material's figure-of-merit performance can be built from knowledge of either the composition alone, or the structure and composition. While models built from knowledge of both structure and composition are generally superior in performance, composition-only models are important both for general reasons and for specific considerations relevant to lithium ion conductors. The experimentally measured conductivity of a material derives from its non-averaged structure which is defined by its composition. This will include structural defects that cannot be captured fully in an average crystal structure recorded in a database such as the inorganic crystal structure database (ICSD), unless the material is fully ordered without fractional site occupancy or substitutional disorder. Most structures with lithium ion conductivity that have been reported in detail (i.e., with the lithium positions) exhibit considerable disorder of this type. Even the average structure is unavailable for potential compositions that have not been experimentally studied, and in addition, many experimental reports of ionic conductivity give composition but not structural analysis of the materials investigated. Reported average crystallographic structures for lithium ion conductors frequently do not give precisely determined lithium positions because of the low X-ray scattering power and extensive structural disorder, again raising the important technical question of the connection between the potentially decisive local structure and the crystallographically-

[1]Department of Chemistry, University of Liverpool, Liverpool L69 7ZD, UK. [2]Leverhulme Research Centre for Functional Materials Design, Materials Innovation Factory, University of Liverpool, Liverpool L7 3NY, UK. [3]Department of Computer Science, University of Liverpool, Liverpool L69 3BX, UK. [4]Stephenson Institute for Renewable Energy, University of Liverpool, Liverpool L69 7ZF, UK. ✉email: msd30@liverpool.ac.uk

determined average structure. We thus build a dataset for machine learning models to predict lithium ion conductivity based on composition. There will be limitations of this approach, for example, the model will be unable to discriminate between polymorphs of a given compound. Nevertheless, crystal structure is not always known nor can it be for entirely novel compositions, thus a compositional model with low computational requirements is necessary for screening unexplored chemical space. The most direct measurement of the ionic conductivity of a material is via a.c. impedance spectroscopy (ACIS) measurement, usually on a dense ceramic[13]. All of the ionic conductivities for the materials included in this database were measured via ACIS.

For a specialist domain topic like solid electrolyte chemistry, the task of digesting the presented information requires significant expertize. Throughout the literature, there are inconsistencies in how data are presented, which introduces difficulties when comparing different reports. A broad knowledge of the background literature is essential for recognizing potentially problematic experimental procedures affecting both composition and conductivity, uncovering discrepancies in reported data, and identifying materials and properties that have in fact been computationally derived rather than experimentally measured (which problematically and unfortunately may not be clearly stated in the body of the text in some cases). All of these challenges increase the difficulty and time required to construct a high-quality database of experimentally reported data.

Leading NLP approaches have demonstrated their capability to extract chemical data from the extensive corpus of past scientific literature[14], a process referred to as automated scraping. Text mining has been demonstrated to be a powerful tool in creating materials datasets. For example, Court and Cole[15] created a dataset of materials and their associated magnetic ordering temperatures. This is possible as a magnetic ordering temperature is reported as a single number usually in the text. Unfortunately for ionic conductors, the task of finding and pairing compositions, temperature of measurement, and conductivities is too complex even for state of the art NLP techniques to be effective. There are the standard issues of tokenizing chemical formulae consistently, and parsing correct values in text and tables. In particular, for ionic conductors with a non-crystalline component, the composition is reported as a mixture of reactants rather than a stoichiometric chemical formula. Furthermore, as the vast majority of reported data is presented in figures with no standardized units for conductivity and extreme heterogeneity between entries, extracting relevant data is a combined challenge in both the fields of NLP and computer vision. Accordingly, the creation of a reliable database is unattainable with present automated capabilities, and thus a manual approach is employed here.

Previous investigations have predicted the ionic conductivity of solid-state materials using statistical methods. Due to the aforementioned difficulties in gathering initial datasets of sufficient size and quality these approaches build models that are based on relatively small experimentally-derived datasets (of the order of 40–82 entries)[8,9,16]. In this study, we have reviewed the literature to gather a dataset of experimentally reported solid-state lithium ion conductivities which with 403 unique compositions is an order of magnitude larger than previously available. A statistical overview of the dataset is presented, with the range of conductivities examined for each structural prototype. Unsupervised embedding and clustering techniques are used to partition this dataset into nine families by compositional similarity, thus assessing the diversity of the dataset. We develop supervised regression and classification models to predict the lithium ion conductivity and assess whether a material will possess an ionic conductivity $\log_{10}(\sigma) \geq 4$ at room temperature, where the conductivity is reported in units of S cm^{-1}. The best regression models achieve a mean absolute error for $\log_{10}(\sigma)$ of 0.85, and the best classification models have a Matthews Correlation Coefficient (MCC) of 0.63, assessed under *k*-folds cross-validation in both cases.

RESULTS AND DISCUSSION

Database construction

A large collection of solid-state lithium electrolyte literature was gathered, and the ionic conductivities were extracted for the materials reported in each study. The experimental procedures in a given source were critically assessed to understand how each sample was synthesized, characterized, and processed into a ceramic. We ensure that in each of the studies, samples had clearly defined compositions and reported direct measurements of the conductivity taken via ACIS. The values of ionic conductivity in the database are a mixture of bulk and total values, as the two are not always distinguished, with only a small number of studies providing sufficient detail in labelling the reported values as such. Where exact stoichiometry may be unclear from the given reagents, any studies that lacked supporting characterization (such as ICP analysis) to confirm the presence of lithium, were discarded. The ionic conductivity and material composition are both of equal importance in the database, as the predictive models are constructed with these two variables. By ensuring that data is exclusively gathered from experimental studies of high calibre, we gain confidence in the quality of the results of subsequent machine learning analysis. Typically, this requires extracting the values from an Arrhenius plot and converting each value from the plotted units (commonly plotted as either σ in S cm^{-1} or S m^{-1}, $\log_{10}(\sigma)$, $\log_{10}(\sigma T)$, or $\ln(\sigma T)$) to conductivity in S cm^{-1} at a specific temperature. In some reports these values may also be provided in tables, or stated in the main body of text along with supporting discussion, allowing for cross-checking of the reported value.

The first stage of the initial literature review was carried out by an undergraduate student to collate source papers of reported conductivities from keyword searches using search engines, and reviews of the field[16–20]. This survey focussed on tabulating the physical properties reported in each paper: composition, ionic conductivity, temperature at which the conductivity was measured, activation energy, and structural prototype. Following this initial tabulation, the activation energy was excluded from the final database as it is not reported frequently enough to warrant inclusion.

Owing to the complexities described above, further expert validation of the data was required. The ionic conductivity of a material is typically determined using ACIS, although it can also be calculated through molecular dynamics simulations[21], or examined by NMR diffusion experiments[22], ion migration studies[23], or entirely different measurements not directly related to ion transport (e.g., maximum entropy method analysis of diffraction data[24]). Even experimental papers which report a measured conductivity for a material through ACIS may themselves involve a variety of measurements and sample preparations, creating uncertainty around reported values. Postgraduate and postdoctoral researchers with more than two years direct experience of battery research with a broad knowledge of background literature assessed experimental procedures, consistency in sample preparation, quality, and other aspects of the reported data based on the details provided. Each researcher handled a selection of entries and was tasked with validating the database entry against the source report.

Dealing with such a large table of data in spreadsheet form adds significant challenges. Specifically, working with an online spreadsheet directly with twenty researchers leads to issues with version conflicts, edit histories, issues with concurrent user access, merging changes from multiple users, as well as assigning and tracking tasks. These issues were avoided by reducing the individual tasks to their core components through a bespoke interface developed with the streamlit prototyping library, shown in Supplementary Fig. 1. The interface was created to present a single entry from the database with its composition, associated

conductivity at a specific temperature, and source paper. For each entry, the researcher was tasked with evaluating the conductivity at that specific temperature, making note of any mistakes with the composition, and reported conductivity or temperature from the source. Positive feedback to researchers was provided through the presentation of a unique compliment provided by a GPT-2 transformer based language generation model[25,26], displayed to the researcher after evaluating and recording each entry.

Database overview

A database was created with 820 entries collected from 214 sources; each entry contains the ionic conductivity of a chemical composition at a specific temperature, ranging from 5 to 873 °C, with an expert-assigned structural label. There are 434 different entries (Table 1) in the database for ionic conductivities experimentally measured at room temperature (15–35 °C). For a further 31 materials, the room temperature conductivities are extrapolated from measurements above room temperature, to obtain a dataset of 465 entries, with 403 unique compositions, as 37 room temperature compositions have conductivities extracted from multiple reports. The room temperature conductivities span the range of 5.00×10^{-16} to 2.50×10^{-2} S cm^{-1}, with a mean $\log_{10}(\sigma)$ of −5.01 and median of −4.41 (Fig. 1). The distribution of conductivities in this dataset and the associated standard deviation are estimated by optimizing the parameters of many probability distribution functions using the Fitter library (github.com/cokelaer/fitter); the distribution which fits the data with the lowest error is an asymmetric Laplace distribution. The interquartile range (50% of the data; materials from the 25th to the 75th centile of $\log_{10}(\sigma)$ in the dataset) spans from −7.30 to −3.03.

Table 1. The number of experimentally reported conductivities contained within this dataset.

Description	Count
ACIS measured conductivities at any temperature	789
ACIS measured conductivities at room temperature (15–35 °C)	434
Room temperature conductivities extrapolated from higher temperature	31
Total number of conductivities at room temperature	465
Total number of conductivities at any temperature	820
Number of unique compositions with a conductivity at any temperature	455
Number of unique compositions with a conductivity at room temperature	403

During database construction, each material in the dataset was manually allocated a label, based on the structural prototype the material belongs to. If the material structure was not discussed directly in the text and its family could not be deduced with reasoning, then this composition was assigned the structural label of *Other*. The breadth of structural chemistry encompassed by this dataset is shown by the fifteen unique families present in this set of expert-curated labels (Supplementary Table 1), which can be used to partition this database and expose trends that have been reported in the literature.

In Fig. 2 the distribution of $\log_{10}(\sigma)$ for each structural family for which room temperature data is available, has been created by fitting a density kernel to the conductivities. This consists of placing a Gaussian distribution of fixed height and width at the x co-ordinate for each conductivity, and summing these together to approximate the probability density, allowing us to estimate the spread of reported conductivities. Irregular distributions with long tails are observed for some structural families. As the majority of these sets contain fewer than 50 reported materials, reports of materials with higher conductivities in the literature will lead to anthropogenically biased distributions[27]. Anthropogenic bias is inescapable when constructing a dataset of experimentally measured properties from the literature. The reduced scientific interest in undertaking the lengthy characterization of materials with little importance to electrolyte chemistry, has meant that materials with very low or negligible conductivity are underreported. Distributions will be skewed towards conductivities of interest, and thus not truly representative of the underlying chemistry.

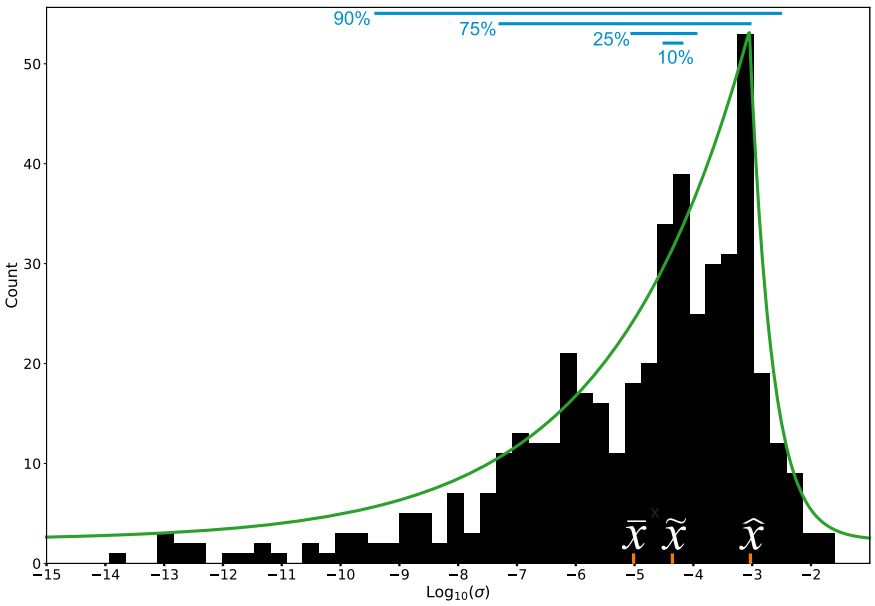

Fig. 1 Distribution in room temperature conductivities for materials in the dataset. A histogram displaying the 465 room temperature conductivities (in units of S cm^{-1}) from materials contained in this dataset and the relative distribution of their $\log_{10}(\sigma)$. The mean ($\bar{x}$) value of −5.01, the median ($\tilde{x}$) value of −4.41, and the mode ($\hat{x}$) value of −3.05 are marked on the x-axis. An asymmetric Laplace distribution has been fit to this data, overlaid in green. The count of each bar is given on the y-axis, with the percentage of materials falling within each percentile range around the median overlaid on the top axis.

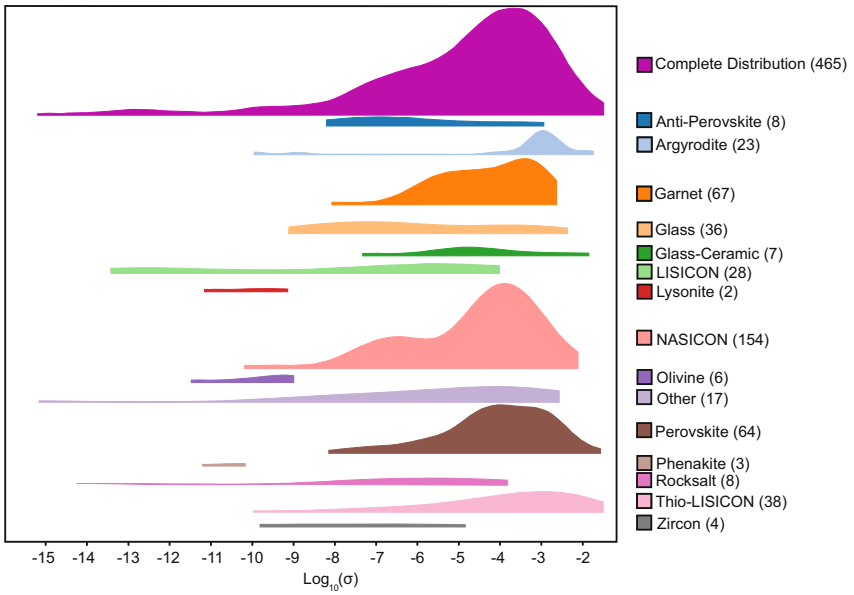

Fig. 2 Distribution of room temperature conductivities across expert-curated structural families. Fitted distribution functions of the room temperature $\log_{10}(\sigma)$ for all materials within the database separated into expert-curated structural families and scaled by the number of entries within each family, given in brackets.

Table 2. Chemistries of the materials in the database of 465 room temperature Li ion conductivities based on anions: pure oxides, oxides containing at least one other anion, pure sulfides, sulfides containing at least one other anion, pure halides, and other (which contains materials such as $LiBH_4$, Li_3P, and $Li_2Ca(NH)_2$).

Materials	No. entries	σ_{min} (S cm^{-1})	σ_{max} (S cm^{-1})
Oxides	346	5.00×10^{-16}	6.31×10^{-3}
Oxide with other anion(s)	18	1.00×10^{-10}	9.38×10^{-4}
Sulfides	55	1.60×10^{-10}	1.70×10^{-2}
Sulfide with other anion(s)	32	8.13×10^{-9}	2.50×10^{-2}
Halides	7	1.18×10^{-14}	1.51×10^{-6}
Other	7	2.00×10^{-9}	1.00×10^{-3}

The minimum and maximum Li ion conductivities at room temperature are given for each group.

The room temperature dataset predominantly consists of NASICON, garnet, perovskite, glass, thio-LISICON, and LISICON type materials, each with more than 27 members. The anion chemistries of the materials are provided in Table 2, showing that 75% of the materials in the database are pure oxide compounds (consisting of 44% NASICON, 19% garnet, 18% perovskite, and 8% LISICON type materials), 12% are pure sulfides, and 2% are pure halide compounds. Mixed anion materials (oxyhalides, oxysulphides, etc.) make up 11% of the materials included (46% of these are argyrodites such as Li_6PS_5Cl, and 16% are antiperovskites such as Li_3OCl). In general, materials containing sulfur as an anion exhibit higher minimum and maximum conductivities which is supportive of the outlook that is commonly encountered in the literature that sulfides exhibit the highest Li ion conductivities.

Machine learning

With a database of materials gathered, unsupervised or supervised machine learning (ML) may be applied to these compositions to extract chemical trends. Unsupervised learning involves the application of embedding and clustering techniques based on the elements in the material, with no further knowledge of chemical properties such as conductivity required. Unsupervised techniques are beneficial as they do not require time-intensive labelling, and may highlight trends and similarities that may not be immediately apparent from a large collection of data in a table. Unsupervised clustering has successfully been applied in previous investigations to cluster electrolyte materials[8] based on crystal structure through hierarchical clustering applied to the anionic frameworks of 528 lithium containing structures from the ICSD. Conversely, supervised techniques attempt to fit a predictive function for a property to chemical descriptors such that the property can be predicted for a new material by statistical learning from known examples in a given training set. Machine learning is applied to compositional descriptors to predict each material's room temperature lithium ion conductivity (a regression task), or to predict whether each material possesses a room temperature lithium ion conductivity $\log_{10}(\sigma) \geq 4$ (a classification task).

In our previous work, we introduce the Element Movers Distance (ElMD)[28] as a metric to quantify the similarity between two chemical formulae. This is demonstrated to be an expressive measure of chemical similarity that aligns with domain expert judgement. This metric can be incorporated with unsupervised dimensionality reduction and automated clustering to present chemical composition data to those who study these spaces. This brings high-dimensional compositional spaces into concise structured representations, such as maps, that can be interpreted by humans. In doing this the landscape of known compositions can be categorized according to our knowledge of related materials. Following the methods described previously with the EIM2D plotting library (github.com/lrcfmd/ElM2D), we construct a distance matrix of ElMD scores between the compositions in the ICSD (2021)[29] and the compositions contained within the ionic conductors database here. This metric space is reduced to two dimensions with principle component analysis (PCA) (Fig. 3). A Gram centred matrix[30] is first obtained from the given distance matrix, and then singular value decomposition of the Gram matrix carried forward to obtain the coordinates of each point projected to the first two principle components. PCA linearly scales each metric distance to maximally preserve each of the interpoint relationships across the dataset, which has previously been shown to closely reflect the true structure of the metric space[28]. Figure 3

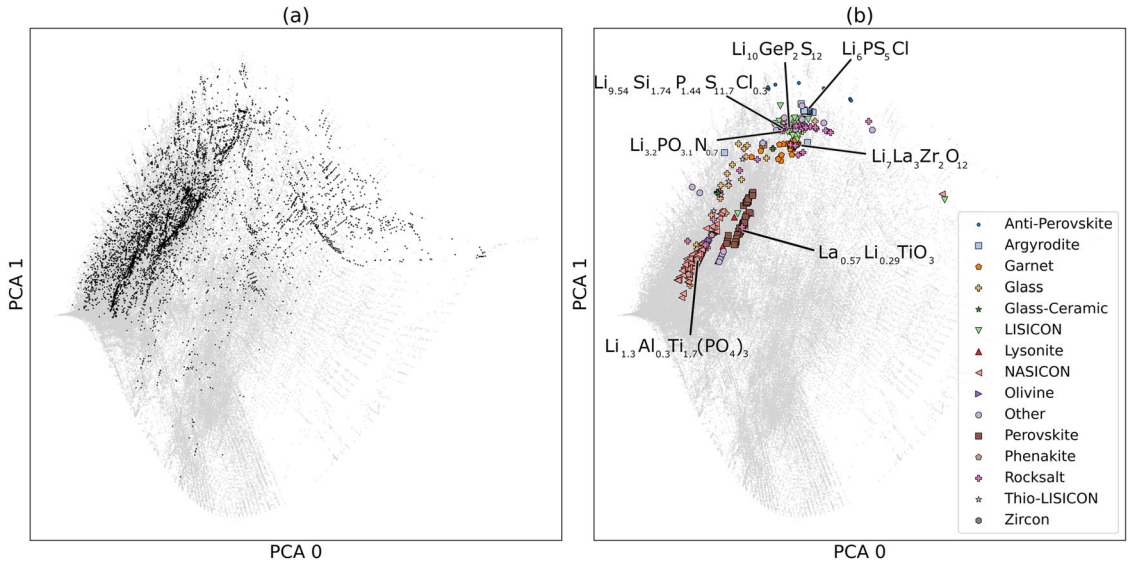

Fig. 3 An embedding of the 127,638 unique compositions (grey) from the ICSD database (2021) with respect to ElMD similarity between compounds, embedded to 2 principle axes with PCA. a 6972 of these compositions contain lithium (black), and **b** 455 unique compositions from this dataset with an experimentally measured conductivity at any temperature. The expert-curated structural label that each composition belongs to is indicated by the colour scheme given in the legend, with a selection of representative compositions and their embedded coordinates indicated.

thus represents the distribution of this dataset in the compositional space of the materials constituting the ICSD.

Each of the lithium-containing compounds of the ICSD are highlighted against other compositions of the ICSD and the 455 unique compositions from our entire database (i.e., compositions with data recorded at any temperature) in Fig. 3a, with the expert-curated labels of the structural families included in the lithium conductors database in Fig. 3b. Though structure has not been included in the initial representation, expert-identified structural families are seen to tend to cluster in this compositional embedding, reflecting the connection between composition and structure. Perovskites (Supplementary Fig. 2), NASICONs (Supplementary Fig. 3), thio-LISICONs, and garnets are found in distinct areas of the compositional map; each of these structural families are grouped tightly on the map, despite the absence of structural information (Fig. 3b). The lithium ion conducting materials in the database are found in the same regions of compositional space as known lithium compounds, and can be seen to match the diversity of lithium chemistry that has been explored to date reasonably well. This reflects the anthropogenic bias intrinsic to the research process, as much of the work devoted to discovering new lithium-containing materials has been driven by applications in battery technologies. There are a number of areas of accessible lithium-based chemistry (compounds seen on the right-hand side of Fig. 3a) where known materials appear underexplored with regard to ionic conductivity. This compositional space should be considered in the search for new families of lithium ion conductors.

Previous work has shown that, while PCA gives an accurate realization of compositional space with respect to ElMD[28], it is not the best representation for further processing with automated clustering techniques. The compact and concentric patterns that these clusters follow are difficult to unravel both visually and algorithmically, particularly when framed against the noise of so many unrelated compounds. We find that non-linear dimension reduction techniques attain a much clearer separation of the space into distinct regions of compositional similarity, which can be clustered more consistently (Fig. 4). Uniform manifold approximation and projection (UMAP) draws apart the points of a space by first forming a neighbourhood graph of points in the metric space then embedding this graph to a two-dimensional plane of projection via Laplacian Eigenmaps to capture global information[31]. These 2D distances are then refined through a ball and spring model[32] to capture the local intricacies of the metric space.

UMAP (Fig. 4a, b) and PCA (Fig. 4c, d) are applied to evaluate the reduced space of the 403 compositions of room temperature solid state lithium ion conductors in the database reported here. The UMAP plot contains several clear regions, which can be separated into nine distinct clusters using the density-based spatial clustering of applications with noise (DBSCAN) algorithm[33] with an epsilon radius of 4 (Fig. 4a). The epsilon value determines the radius of disks that are overlaid on every point in the two-dimensional plot, which are then used to classify the points into different clusters. If two points cover each other with overlapping disks, then these will be assigned the same cluster label. DBSCAN has the ability to capture dense regions of an embedding, but if epsilon is too large then the output will fail to separate disjoint clusters. In this study, epsilon was chosen manually to maximize consistency between automated clusters and the clusters that can be visually observed.

Each of these unsupervised ML-derived clusters from Fig. 4a are chemically reasonable, with clear stoichiometric substitutions or structural similarities connecting their constituents. This becomes apparent from comparison with the expert-derived structural family labelling in Fig. 4b, d. For example, Clusters 0 and 8 from the automated clustering are predominantly populated by NASICONs, perovskites are exclusively found in Clusters 5 and 6, whereas Cluster 4 is almost exclusively garnet structure materials. In addition to the practical benefits automated embedding and classification provides to rationally organize materials with minimal human bias, these clusters have further application in supervised training. As some data must be withheld from training and retained to test the performance of a trained model, each DBSCAN-derived cluster will be used as a testing set in a process referred to as Leave One Cluster Out Cross Validation (LOCO-CV). These clusters range in size from 6 materials to 93 materials, with the training set then typically containing 85–90% of the available data to train each model. The distributions of $\log_{10}(\sigma)$ for each LOCO cluster have been plotted in Supplementary Fig. 4, with basic statistics given in Supplementary Table 2, where many of the

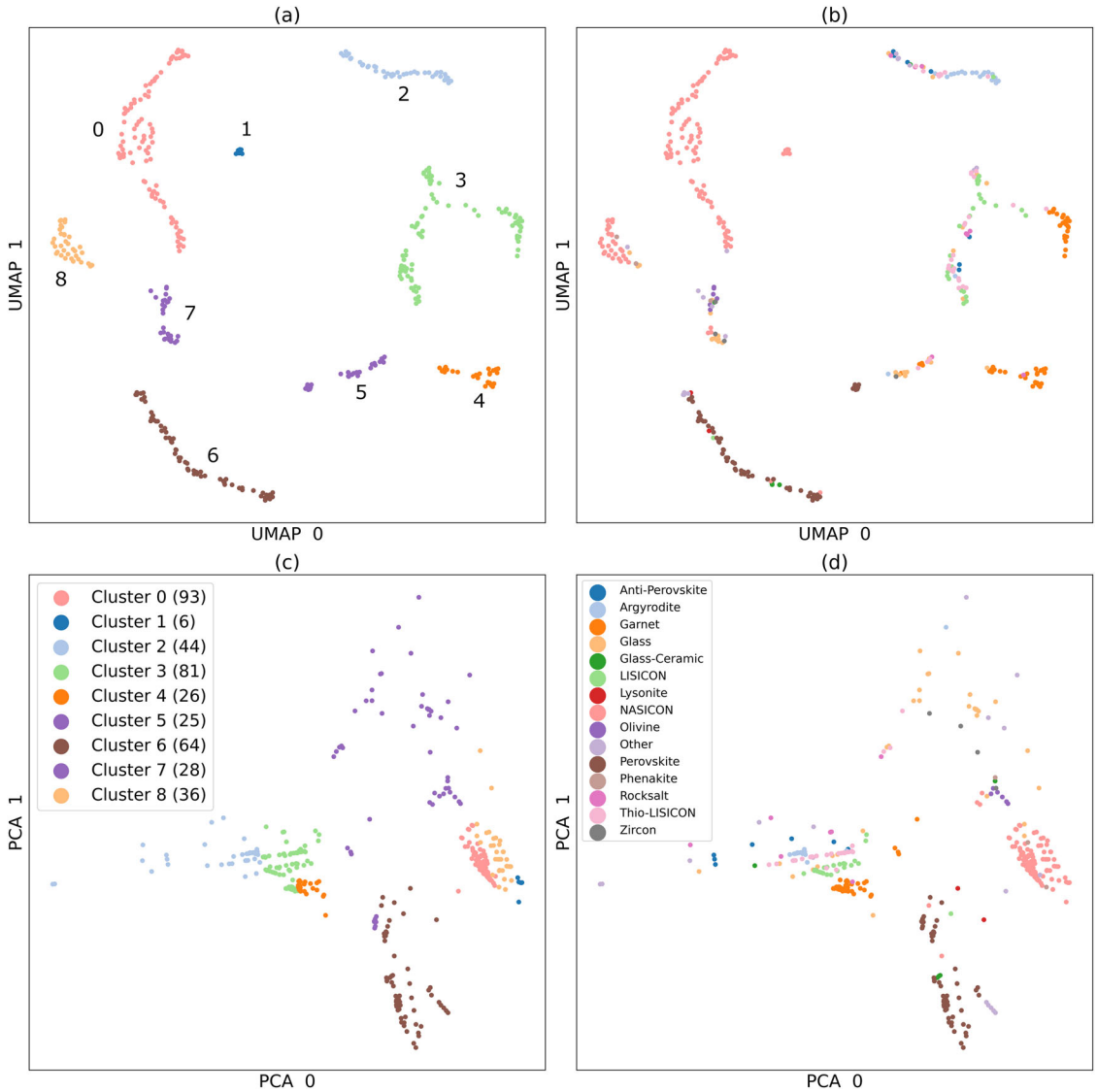

Fig. 4 Embeddings of the 403 unique room temperature solid state electrolytes compositional data. a, **b** Show the coordinates obtained from the UMAP embedding algorithm, whilst (**c**) and (**d**) arise from PCA. The cluster labels in (**a**) and (**c**) are obtained from the DBSCAN clustering algorithm applied to the UMAP embedded points in (**a**), with the number of materials in each cluster given in brackets. Cluster labels in (**b**) and (**d**) were assigned from expert review to classify each material under a structural prototype.

clusters span similar ranges of conductivity. Given the intra-cluster chemical consistency and inter-cluster dissimilarity, these assessments are a measure of how each model performs at predicting the ionic conductivities of materials that are chemically dissimilar from those on which the model has been trained.

Supervised learning

A dataset of 403 entries is constructed, where compositions with duplicate room temperature conductivities from differing sources have been represented by the median of these multiple reported conductivities. With this dataset in hand, we apply the best available ML models that can be implemented with minimal modification, i.e., off the shelf. This is done with traditional statistical learners (ensemble models) with mat2vec[14] composition-based feature vectors[34], and deep learning techniques (CrabNet). For statistical learners, we wish to ensure the best models and associated hyperparameters are chosen, so that we do not simply overfit to one portion of the data. A simple model with fixed hyperparameters is not guaranteed to give good predictions on unseen compounds. Such models may overfit to the training data, leading to poor predictions on unseen compositions, or give exceptional performance on certain subsets of the data with poor performance on the rest. Some of the issues of overprediction can be remedied by surveying a range of statistical models[35]. State of the art techniques for predicting materials properties through composition apply this principle by training an ensemble of models, in the belief that each model will learn to focus on a different set of features. The predictions of each individual model are combined, which tends to give more robust predictions across the entire domain. In statistical models, the ensemble approach is notably used in the random forest (RF) algorithm[36], where large ensembles of decision trees are randomly constructed and kept or discarded depending on their predictive quality. The resulting quality of RF predictions depends on the values of each hyperparameter chosen when initializing the model, and poor choices can lead to very poor models. To alleviate this, best practice has traditionally focussed on trialling a

range of hyperparameters in combination with one another, but this is time consuming and does not guarantee that the optimal configuration will be found. More recent AutoML approaches[37] improve on this by framing the choice of statistical model and its associated hyperparameters as a meta-problem to be solved. Many separate algorithms and hyperparameters can be trialled and assessed in combination, with the measured performance used to update a selection policy for future trials until optimal combinations are found.

In AutoSklearn[38], many types of models and data pre-processing stages from the scikit-learn library are chained together to form data processing pipelines. The supplied training data is shuffled into *k*-folds cross-validation sets and used to assess each pipeline, with the performance noted. This performance is used to update the parameters of a tree-based Bayesian optimization selection policy, which will decide the models and hyperparameters to choose in future iterations, alternating between exploring untried combinations, and exploiting relationships known to give good results. Given that RFs return more robust predictions through ensembling many weaker models together, we would expect an ensemble of effective models to give even stronger predictions. As simple models are quick to train, thousands of pipelines can be evaluated during the AutoSklearn training process. After the allotted training time of ten minutes, the 50 pipelines with the highest performance are selected to form a trained ensemble which can be used to predict unseen data.

In comparison, Compositionally Restricted Attention Based Networks (CrabNets)[39] are an implementation of the transformer model[40] of deep learning. Here, self-attention is employed to learn how relationships between each of the elemental vectors in a composition are aligned with a target property. The transformer's positional encoder is repurposed as a fractional encoder to capture the ratio of each element in the composition, which enables CrabNets to capture similarities and small variations in stoichiometry with precision. This is particularly relevant for ionic conductors, where minor substituents (e.g., those controlling the exact lithium content) can significantly influence the ionic conductivity because they determine the defect concentrations and associated local structure that can govern ionic motion.

One shortcoming of deep neural networks such as CrabNets is that they require large quantities of training data which are typically unavailable for materials science problems. This limitation can be alleviated by transfer learning, which involves pretraining networks on much larger datasets of compounds and their associated properties, such as the computed energy of formation. The trained parameters of this network can be exported to initialize future models for different properties, as opposed to initializing all of these values randomly. The desired benefit of pretraining the network on a wider range of compositions and their associated formation energies, is that the knowledge of chemical relationships absent in our training set can be extrapolated to future predictions. By transferring this knowledge from another domain, the most salient chemical relations are intended to be well represented in the network. This typically leads to a faster convergence to the optimal value when training the neural network on the desired property, and can lead to improved predictive performance in the target domain. This has been demonstrated in other investigations[41,42], where the application of transfer learning and neural networks has achieved state of the art for materials property prediction. In this work we compare the performance of AutoSklearn ensembles, randomly initialized CrabNets, and CrabNets that have been pretrained on compositions and their formation energies from the OQMD[43].

Training CrabNets involves iteratively updating many model parameters of the network on the same dataset multiple times; each iteration is called a training epoch. Once an iteration has completed, the millions of model parameters will have been more finely tuned to align the data with the target property, which should give a better model than the previous iteration. When model training begins, we expect poor performance when predicting properties of materials in the test set, but as the model is further biased by training data after several epochs, more robust predictions should be attained. In general, when training neural networks, the training error steadily decreases over time, as the parameters of the model get more aligned with the input. After prolonged training, however, these parameters begin to overfit to the training data, and the model gets steadily worse at predicting anything outside the training set[44].

The training and testing performance at each epoch can be plotted on a training curve, which characterizes how performance evolves with the number of training epochs (Supplementary Figs. 5–8). A training curve can be used to determine the optimal training time (e.g., number of epochs). Model parameters can be exported from the training epoch that displays best performance at test set predictions. Training for sufficiently long time (to see degradation in test set performance) and then reverting to an earlier state in training is referred to as early stopping, in contrast to a priori deciding the number of training epochs, or training indefinitely. Early stopping across 500 training epochs is applied in this study, with each model taking the optimal set of training weights, giving a reasonable measure of how CrabNets with and without transfer learning perform using standard hyperparameters (discussed in Supplementary Note 1).

The performance of AutoSklearn and CrabNet regression and classification models at predicting the conductivities of the materials in this dataset is evaluated through four methods: control studies, parity plots, scoring metrics, and cross-validation techniques. We then use the best approach from this assessment to train final regression and classification models on all available data.

To give some measure of the worst-case performance, we provide two control experiments. In the first control experiment, we take the reported conductivity of each material, shuffle these labels, and treat the average of five of these shuffled values as an ensemble prediction from a poor model. This has the effect of providing a quasi-random prediction that demonstrates how ensembles can bring predictions closer to the mean (Fig. 5a). In the second control experiment, we demonstrate how a model which simply predicts the mean will perform. We take the mean of all of the room temperature conductivities (-5.02 in $\log_{10}(\sigma)$) and treat these as the output prediction for each material, giving the same prediction for every entry. The true conductivities are plotted against each of these control predictions to observe the performance (Fig. 5b).

Plots are an effective method to directly confirm the performance of a statistical model. For regression tasks, we plot the actual conductivities of each material against the predicted conductivities of a trained model. An ideal model would give each prediction perfectly on the leading diagonal. Dense pointclouds can be difficult to visually interpret, so errors of each prediction ($y_{\text{pred}} - y_{\text{true}}$) are calculated and plotted via histogram to quantify this distribution of errors. A Student's t-distribution is fitted to the errors of all repetitions (without averaging) to provide intervals for how many predictions are within certain bounds of error for each model. The shuffled control has a zero-centred gaussian distribution of errors on the histogram with a standard deviation of 2.34 (Fig. 5c). The mean control has an error of -0.44 below the true value on average, with 68% of the predictions having an error within -1.99 to 1.10 of the true $\log_{10}(\sigma)$ (Fig. 5d). Given this worst-case performance, we may demonstrate how the best compositional models perform at predicting new compositions.

When we have many plots for different models, it becomes difficult to visually confirm the best performing model. To quantify which of these models are best performing, we must use statistical metrics to rank the quality of the output predictions for each

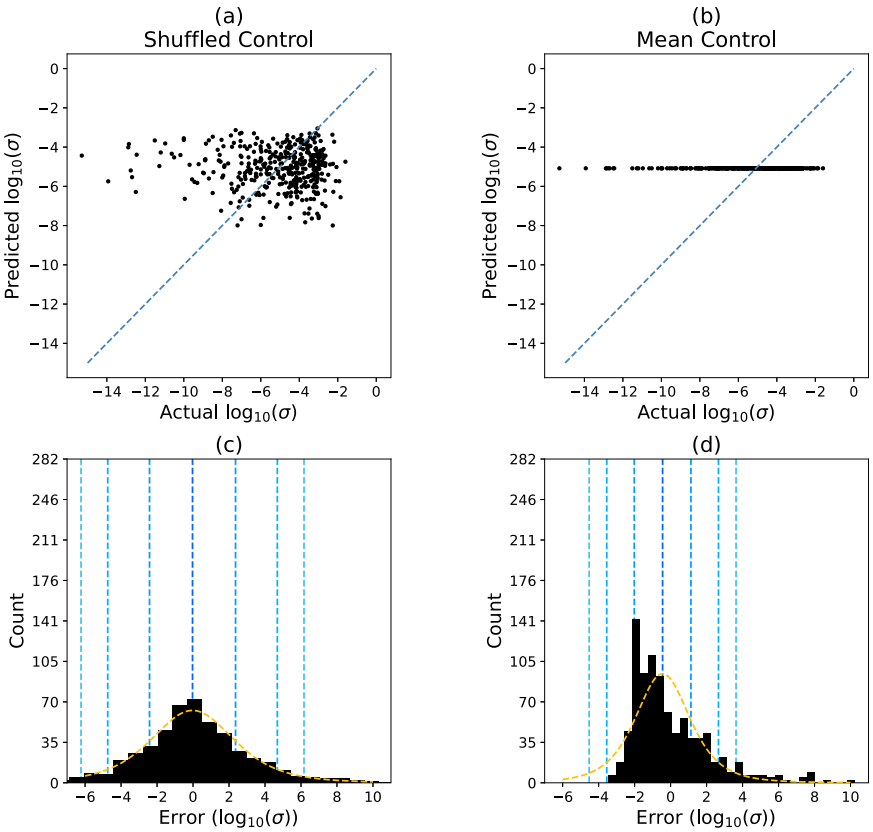

Fig. 5 Parity plots and error distribution for two control studies. a The shuffled control parity plot demonstrates each materials actual conductivity plotted against an average of five randomly selected values across the dataset. **c** The distribution of errors across all experiments (without averaging) demonstrates the maximal error bounds we would expect from a poor statistical model, with 68% of predictions falling between −2.36 to 2.31 away from the true values. **b** The mean control experiment demonstrates the expected predictions for a model which has simply learnt the mean value of the dataset. Correspondingly, the distribution of errors (**d**) is simply a reflection of the distribution of conductivities around the mean value, and models which form predictions close to the mean will resemble this distribution. A Student's t-distribution (orange) is fit to the underlying data, with the mean of this distribution (dark blue), and the first, second, and third standard deviations away from this mean (light blue) overlaid in (**c**) and (**d**). A good model should have a mean of zero, with tight error bounds.

model. Regression models are often scored via Mean Absolute Error (MAE) and Pearsons R^2 score. The MAE returns the average difference between each prediction and its known value, where values closer to 0 reflect stronger model performance. The R^2 score shows the correlation between the true and predicted values, where a 1 is a perfect score, and anything below zero indicates that on average model predictions perform worse than simply returning the mean of the test set for all inputs.

For classification tasks, the performance may be demonstrated via a confusion matrix. This is a 2×2 matrix that compares the predictions made by the classification model against the true classification labels. An ideal result would have leading values (True Positives and True Negatives) and zeros elsewhere, but in reality, many predictions will be False Positives and False Negatives. For simplicity, however, the most frequently reported score for classification is accuracy. The accuracy score is defined as the number of true predictions divided by the total count of values in the testing set:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

On heavily imbalanced datasets with few negative class instances, the accuracy can return a high score for poor classifiers that output a single classification. This is due to the small number of negative instances, which do not significantly alter the denominator even if they are heavily misclassified (Eq. 1). To prevent misleading reporting, the MCC[45] can be taken as a more informative score[46] by considering the proportion of each class in the confusion matrix:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FP)}} \tag{2}$$

The MCC is calculated by taking the difference of the product of true predictions and the product of false predictions, and dividing by the geometric mean of all entries in the confusion matrix. This returns a value from 1 for perfect classifications to −1 for entirely incorrect classifications. The MCC provides more weighting to the score for any misclassified values, allowing us to judge the outcome of the confusion matrix succinctly. By themselves, isolated scores do not convey the strength of a model and these must be compared against a known point of reference, such as a control study, to understand the significance of a particular result.

As an aim of machine learning models is to predict the behaviour of as-yet unknown materials, it is important to distinguish between performance in interpolation between materials that have similar chemistries, where similar structure-property-composition relationships would be expected, and in extrapolation to materials characterized by structure and bonding that is not found in the training set. For example, predicting performance within a solid solution family with some members in the training set used would be interpolation, whereas evaluating the conductivity from a material with a new structure type would be extrapolation.

Table 3. Regression Performance Metrics Average results of each regression model, judged by Mean Absolute Error and Pearsons R^2 metric under both dataset cross-validation regimes.

Model	MAE (*k*-folds)	R^2 (*k*-folds)	MAE (LOCO)	R^2 (LOCO)
Shuffled Control Study	2.31 (0.06)	−0.99 (0.13)	2.43 (0.08)	−2.6 (0.4)
Mean Control Study	1.71 (0.0)	0 (0)	1.72 (0)	−0.46 (0)
AutoSklearn 2.0	1.10 (0.04)	0.46 (0.05)	1.62 (0.08)	−0.4 (0.2)
Randomly Initialized CrabNet	0.96 (0.02)	0.55 (0.03)	1.131 (0.006)	0.15 (0.03)
Transfer CrabNet	**0.85 (0.02)**	**0.62 (0.02)**	**0.99 (0.03)**	**0.33 (0.02)**

The average value of the training performance across the test sets is first calculated for each metric, and then averaged across each of the five repetitions; standard deviation shown in brackets. Values in bold represent the best performing model under each metric.

Table 4. Classification Performance Metrics Average results of each classification model predicting whether materials possess $\log_{10}(\sigma) \geq 4$, judged by Matthews Correlation Coefficient (MCC) and accuracy under both dataset cross-validation regimes.

Model	MCC (*k*-folds)	Accuracy (*k*-folds)	MCC (LOCO)	Accuracy (LOCO)
Shuffled Control Study	−0.02 (0.03)	0.50 (0.02)	0.00 (0.07)	0.52 (0.03)
Mean Control Study	0 (0)	0.58 (0)	0 (0)	0.64 (0)
AutoSklearn 2.0	0.46 (0.04)	0.74 (0.01)	0.10 (0.05)	0.63 (0.03)
Randomly Initialized CrabNet	0.57 (0.01)	0.786 (0.006)	0.36 (0.01)	0.62 (0.01)
Transfer CrabNet	**0.633 (0.002)**	**0.814 (0.009)**	**0.38 (0.01)**	**0.71 (0.01)**

The average value of the training performance across the test sets is first calculated for each metric, and then averaged across each of the five repetitions; standard deviation shown in brackets. Values in bold represent the best performing model under each metric.

This question naturally arises when evaluating ML model performance. Here, it is important that the data being tested have not been previously used to train the model, but in and of itself, this does not directly address interpolation versus extrapolation ability. The standard method of splitting data is via *k*-folds cross-validation, where the dataset is split into *k* equal sets, and one of these sets is used to test the model. In this report we take $k = 5$, where the model is trained on four of these subsets (80% of the data) and then tested on the fifth (20% of the data). This process is repeated for each set, and the mean score across all test sets is used as the final measure of performance. As many of the compounds in this dataset possess some similarity with one another, we expect the model should be able to interpolate relationships between known compositions.

Ideally, we want predictive models to be able to extrapolate beyond known materials, and statistically infer future chemical relationships from observed compositions. To test this, we utilize the DBSCAN labels assigned in Fig. 4 as Leave One Cluster Out (LOCO) labels to separate the 403 unique room temperature conductors into testing sets. As the compositions within each cluster have been confirmed to share chemical similarity, and to have dissimilarity from other clusters, using each cluster shown in Fig. 4a as a testing set provides a better estimate of the ability of a model to screen novel compositions than the *k*-folds approach, which will entail greater chemical similarity between the training and testing sets.

Both of these cross-validation techniques are applied to train AutoSklearn and CrabNet regressors and classifiers, with the average of five repetitions of each experiment taken as the final score. We collate the performance of the two control studies and the ML models for regression and classification, in Tables 3 and 4 respectively. Plots of all regression models performance can be found in Supplementary Figs. 9, 10.

The two control studies give the highest MAE and lowest R^2 scores between the actual and the predicted values under each cross-validation scheme. These numbers are important to consider when evaluating any improvement in predictive performance. All models perform better than these controls, and under *k*-folds cross-validation, and AutoSklearn models perform comparably to randomly initialized CrabNet models. However, under LOCO-CV, the AutoSklearn model fails to fit a suitable decision boundary to predict unseen materials; performance metrics reveal no significant improvement over the mean control. CrabNet models are better than AutoSklearn models at the extrapolatory LOCO task, and these see improved performance in both MAE and R^2 correlation. CrabNet models with transfer learning outperform all other models across each metric and cross-validation scheme. The ~10% increase in performance of transfer learning regression models over those initialized randomly suggests that pretraining in other domains has given the model a clear advantage when inferring unseen chemical relationships. To demonstrate this further, three of the regression models parity plots and distribution of errors are given in Fig. 6. These plots allow us to visually judge models against one another, and to assess each model's performance at predicting materials similar to those within the training dataset (*k*-folds) as opposed to materials with unseen chemistry (LOCO-CV).

The AutoSklearn regression model under LOCO-CV (Fig. 6a) demonstrates tighter prediction error bounds than the shuffled control, but still leads to predictions with an error of −0.68 on average and a standard deviation of 1.55 (Fig. 6d). An ML model which typically achieves predictions of ionic conductivity within two orders of magnitude could be interpreted as a positive outcome. However, comparison to the mean control demonstrates that this model has not learned a meaningful representation for extrapolating beyond the chemistries within the training set. The AutoSklearn error distribution is not an improvement over the mean control, which has an average error of −0.44 and a standard deviation of 1.54 (Fig. 5d). CrabNets with and without transfer initialization output a range of predictions closer to the real values, with tighter error bounds than AutoSklearn models. The CrabNet regression models with transfer learning trained under LOCO-CV (Fig. 6b) are not as consistently skewed as AutoSklearn, with an average error of −0.02 and a standard deviation of 0.811 (Fig. 6e). These models typically return

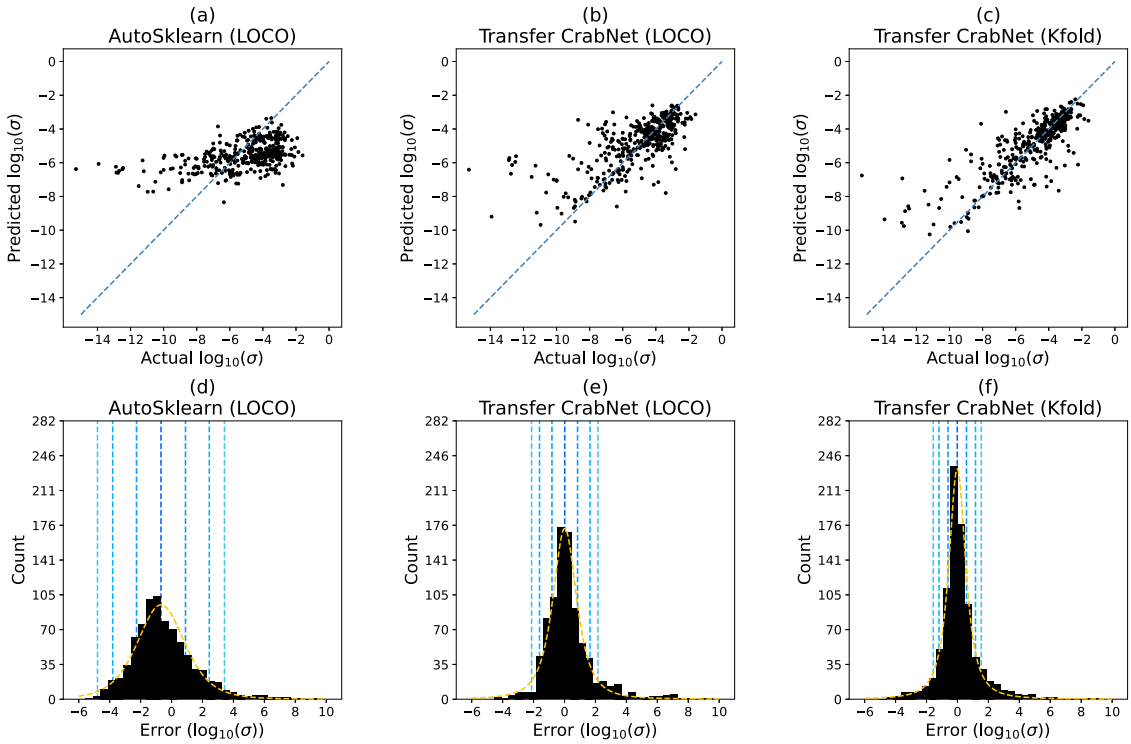

Fig. 6 Parity plots and error distributions for three regression models. AutoSklearn models assessed under LOCO-CV (**a**), (**d**) share the most similarity to the controls in Fig. 5, and are thus judged to be the least effective ML model under investigation. Under LOCO-CV, CrabNet models with transfer learning offer improved performance, which can be visually confirmed by the spread of points falling closer to the leading diagonal (**b**), and the distribution of errors being centred around 0 with a smaller standard deviation (**e**). ML models give a tighter distribution of errors when validated with k-folds, with transfer learned CrabNets possessing the most favourable actual vs. true characteristics (**c**) and distribution of errors (**f**).

predictions with less error for high and medium conductivity materials, but often fail to capture the outlying low conductivity regions. This highlights the complexity of predicting exact materials properties when there has been little exposure to these unexplored chemistries. The best regression performance is achieved using CrabNet models with transfer learning under *k*-folds cross-validation (Fig. 6c), which leads to a distribution of errors centred around −0.01, and a standard deviation of 0.58 (Fig. 6f). As LOCO-CV forces each model to extrapolate future predictions, it is expected that the figures of merit will be less attractive than under *k*-folds cross-validation. Whereas regression models achieve only a modest improvement to the bounds set by the respective control studies, this is not the case for each of the classification models, which we turn to now.

Table 4 displays the average MCC and accuracy score for each model's test set performance across five runs, where it is seen that control models may seem initially reasonable when judged by accuracy. A complete table of results under standard metrics may be found in Supplementary Tables 3 and 4 for comparison, although we consider MCC to carry the strongest judgement of model performance. CrabNet models with transfer learning return the highest MCC of 0.63 under *k*-folds cross-validation, and CrabNets without transfer learning return a slightly lower score of 0.57. AutoSklearn models do not give as strong performance, with an MCC of 0.46, but this is clearly a step improvement on the MCC scores of the control studies, with accuracy also seen to improve by some margin when comparing each model to the controls. As with the regression models, classification models trained under LOCO-CV return lower scores. This is highlighted by the AutoSklearn model, which has a particularly poor MCC (close to the MCC of zero of the two controls) of 0.10 when classifying LOCO test set materials, despite a promising accuracy score. The highest scoring LOCO classification model is the CrabNet with transfer learning; an MCC of 0.38 indicates more of the high conductive materials are correctly classified as having $\log_{10}(\sigma) \geq 4$ than misclassified, which is supported by the high test set accuracy of 0.73.

The two distinct cross-validation techniques have been applied to rank these statistical models against one another. However, interpolation between related materials within known chemistries (defined as known structure and bonding) should be considered independently from extrapolating into unknown chemistries beyond the training data. Accordingly direct comparison should not be drawn between the metrics for the two different cross-validation protocols, as these assess different aspects of the performance of the ML models trained against the dataset. We are forced to use the data in our possession to assess the quality of each model. The data arise from the efforts of researchers in the field, and thus reflect various research trends and foci that have emerged, rather than directly expressing the possibilities for structure, bonding, and performance for materials drawn from element combination at the level of the periodic table. Given this anthropogenic bias, there will be consistencies and trends within each chemical family of the dataset.

By separating the materials of the database into clusters by chemical similarity and testing under LOCO-CV, the reduced performance compared to validation by *k*-folds highlights the challenge of extrapolating known compositional relationships to other chemical families that may span different ranges of conductivity. Comparatively, under *k*-folds cross-validation, each material in the testing set has a greater likelihood of having corresponding materials with similar elemental composition to their own in the training set. The model under assessment thus has more opportunities to interpolate between compositions in the training data, allowing it to make stronger predictions as it has

Table 5. Final regression and classification model predictions of the experimental holdout set.

Composition	Measured conductivity ($\log_{10}(\sigma)$)	CrabNet regression prediction ($\log_{10}(\sigma)$)	CrabNet classifier prediction ($\log_{10}(\sigma) \geq 4$)
$Li_{10.35}Ge_{1.35}P_{1.65}S_{12}$[50]	−1.85	−3.60	1
$Li_{10.35}[Sn_{0.27}Si_{1.08}]P_{1.65}S_{12}$[51]	−1.96	−3.50	1
$Li_{10}GeP_2S_{11.7}O_{0.3}$[52]	−1.99	−3.06	1
$Li_{10}GeP_2S_{11.4}O_{0.6}$[52]	−2.07	−3.07	1
$Li_{10}[Si_{0.3}Sn_{0.7}]P_2S_{12}$[51]	−2.09	−2.66	1
$Li_{9.42}Si_{1.02}P_{2.1}S_{9.96}O_{2.04}$[53]	−3.49	−3.67	1
$Li_{3.35}P_{0.93}S_{3.5}O_{0.5}$[48]	−4.04	−2.67	1
$Li_{3.3}SnS_{3.3}Cl_{0.7}$[54]	−4.49	−3.62	0
$Li_{4.3}AlS_{3.3}Cl_{0.7}$[49]	−5.09	−7.14	0
$Li_3P_5O_{14}$[55]	−6.04	−7.73	0
$LiAlP_2O_7$[56]	(Very low)	−6.32	0

CrabNets with transfer learning are trained on all 403 unique compositions and the associated $\log_{10}(\sigma)$ or classification target at room temperature. The experimentally measured $\log_{10}(\sigma)$ of each of the 11 materials in the holdout set are given alongside a predicted $\log_{10}(\sigma)$ and conductivity class for the material from the final models, the boundary against which the classification is performed has been marked in black.

to some extent been presented with similar examples during the training, rather than having them deliberately withheld.

This emphasizes the strength of structure-property-composition relationships in lithium ion transport. It is reasonable to assume that ion transport takes place by local hopping through barriers governed by physical models that are closely connected in their physiochemical origin across all materials in the dataset regardless of structure and bonding. However, the changes in structure and bonding between these machine-identified materials clusters in which lithium transport occurs by similar, unifying diffusion mechanisms are sufficient to hinder extrapolation of performance from one set of chemistry to another, despite no fundamental change in mechanism taking place between the clusters. This contrasts with the situation prevailing for example in superconductivity, where entirely different mechanisms may govern high-temperature superconductivity in cuprates and low temperature superconductivity in elemental and alloy systems that pair by weak-coupling BCS. This mechanistic difference has been shown to undermine attempts to extrapolate with machine learning from superconductors with one pairing mechanism to another[47], whereas for lithium ion transport it is the chemistry (the structure and bonding) that controls performance even under a unified physical mechanism. Nevertheless, CrabNet models with transfer learning are seen to consistently outperform both the control studies and AutoSklearn models at predicting ionic conductivity. This is shown statistically across all cross-validation schemes and metrics in both classification and regression models, and can be visually attested from the parity plots. As such further discussion will assume these models as the focus unless stated otherwise.

The Final Models

When screening compositions with machine learning we want to use the best possible model to increase the likelihood of making robust predictions. Model performance is typically improved by using the most training data available, and choosing an optimal training time. As discussed earlier, the optimal training time can be determined by assessing the performance vs. epoch training curve to decide which set of model parameters to use (i.e., early stopping). An important practical consideration is that any model to predict ionic conductivity would be most valuable when screening new materials. Accordingly, to assess the ability of our ML models to estimate the ionic conductivities of unstudied materials or novel chemistries, we train a final classifier and a final regressor on the entire initial database of unique room temperature conductivities and test it against eleven newly reported materials that have not been included in the initial database. We refer to this new set of materials as the experimental holdout set. These are selected to represent a range of chemistries and also conductivities, which matches the situation facing the experimentalist targeting new families of ion-transporting materials: it is desirable to understand the likely lithium conductivity of a particular composition in order to aid the selection of specific new chemistries for investigation.

We select CrabNet with transfer learning as the architecture for these two models, as *k*-folds and LOCO-CV assessment show that it offers the best interpolation and extrapolation performance based on the considerations above. The final CrabNet models are trained on all unique entries of the initial database presented here. In the earlier validation investigations, early stopping could be employed by using the test data to select the set of network weights at the best performing training epoch on the training curve. In our final models, a fixed number of training epochs are determined a priori by assessing the training curves of CrabNets with transfer learning under LOCO-CV and selecting a training time which typically attains optimal performance (Supplementary Note 2). Final models are trained on all unique compositions with room temperature conductivity (i.e., all 9 LOCO clusters), with the classification model trained for 98 epochs, and the regression model trained for 323 epochs.

The performance of these neural networks at classifying or predicting the $\log_{10}(\sigma)$ of a selection of recently reported materials is assessed across a range of reported conductivities. The individual performance for each material in the holdout set is given in Table 5. As there are more training data available than in the validation investigations, the final models should have similar or improved performance to the results observed through cross-validation. The final classification model predicts whether the compounds of the experimental holdout set possess high ($\log_{10}(\sigma) \geq 4$) or low ionic conductivity with an accuracy of 0.91 and a MCC of 0.83. The final regression model achieves an MAE of 1.34 on the holdout set, with an R^2 score of 0.51. The performance of the final model against this necessarily small holdout set is consistent with the more robust performance indicators obtained from the previous validation investigations.

Despite the disparity in chemistries between the majority oxide training set and the more varied experimental holdout set (Supplementary Fig. 11), it appears from these metrics and also from consideration at the level of individual materials, that the regressor predicts properties reasonably. Compositions with exceptionally high conductivity are underestimated by the regression model. For nine of the eleven materials, the conductivity has been correctly predicted within two orders of magnitude, which would be expected for materials related to $Li_{10}GeP_2S_{12}$, as this is contained in the training data. However, for the non-oxide materials of the holdout set that are dissimilar to those in the training set, performance is reasonable even when these materials have crystal structures that differ from other materials included in the training set. $Li_{3.3}SnS_{3.3}Cl_{0.7}$ is the first lithium ion conducting defect stuffed wurtzite based on hexagonal close packed S^{2-} anions[48]. $Li_3P_5O_{14}$ has an ultraphosphate crystal structure defined by extended anionic layers, and is also structurally distinct from materials included in the training set[49]. Given that these are structurally differentiated materials, the ionic conductivities have been reasonably predicted (within 1.69

of the true $\log_{10}(\sigma)$) by a regression model that is based purely on composition. These models can be used as screening tools to motivate the further study of candidate materials and phase fields, and assist in the prioritization of resource commitment for experimental synthetic work.

Given the intended purpose as a screening tool, and the more favourable metrics demonstrated by the classification model, a reliable classification of high conductivity materials is more helpful than an absolute estimate of the ionic conductivity from the regressor. There are fewer materials with exceptionally high or low conductivity in the database, and as such there will be greater uncertainty when predicting a specific conductivity for materials in these extrema. Training on classification features gives a more balanced distribution of positive and negative class labels, which gives the model a less skewed dataset for judging its composition-based decision boundary, as reflected in the more favourable performance scores of the classification models. Although there is identified anthropogenic bias present in the dataset, the MCC score under LOCO-CV improves in comparison to each control. This leads us to conclude that these classification models predict with sufficient reliability whether a material has a $\log_{10}(\sigma) \geq 4$ for these to be further employed to screen candidate ionic conductors (e.g., the material contains lithium and is likely to have low electronic conductivity). This does not replace expert chemical knowledge and judgement, instead providing a complementary numerical insight based on the evaluation of data at a scale hard for human experts to assimilate.

Here we present a dataset of experimentally reported lithium SSEs. This dataset includes the composition, structural type, conductivity, and measured temperatures of 789 ACIS measured conductivities, with 403 unique compositions with an associated ionic conductivity near room temperature. Multiple stages of data validation were carried forward by a team of domain experts to ensure that all data are correctly imported from the literature. The creation of a reliable database is a task that is particularly difficult to carry forward with automated tools due to the wide inconsistencies in how data is reported in the field of ionic conductors, necessitating lengthy human validation. Automated scraping would be a viable strategy if all future reports were to prominently state in the abstract a well-defined composition, ionic conductivity in common and clearly stated units (e.g., S cm^{-1}), the temperature at which it was measured (e.g., 298 K) and the technique used to measure it (e.g., ACIS). With this in mind, we encourage researchers and journal editors to consider reporting core findings in this manner, which will enable materials science researchers to leverage tools from the NLP community to gather even larger datasets in the future.

The dataset represents the diversity of chemistry spanned by lithium-containing materials, with a numerical preponderance of oxide-based examples. There are 15 structural families represented at room temperature, including oxides, sulfides, halides, and mixed anion materials. These room temperature compositions are visualized and clustered with the ElM2D package to partition the dataset into nine chemically distinct clusters for leave one cluster cross-validation (LOCO-CV) assessment of the performance of machine learning models.

Supervised statistical (AutoSklearn) and deep learning (CrabNet) models have been applied to this dataset to predict the ionic conductivity of a material from its elemental composition alone. Regression and classification models have been evaluated with standard statistical metrics under different cross-validation regimes to assess their performance at predicting the ionic conductivities of novel materials. The ionic conductivity of a material is the product of many chemical and structural considerations, and also depends on external factors such as temperature. Further, the measured conductivity can also strongly depend on sample preparation, the presence of impurity phases, and crystallite size distribution, which are often discussed collectively under the nebulous term, 'sample quality' This makes ionic transport a difficult property to reliably predict from limited and anthropogenically biased compositional data. Given this challenge, we go beyond standard statistical metrics by designing control studies to investigate the models more thoroughly. We show that CrabNets with transfer learning demonstrate the best performance under both *k*-folds and LOCO cross-validation.

We present a classification model that is able to estimate whether a material has high or low conductivity with reasonable reliability. This is a practical tool to aid experimentalists in their decisions to prioritize candidates for further investigation as lithium ion conductors. Predictions from this model for chemistries dissimilar to those contained in the database are likely to be less reliable than those of closer chemistries, and materials that may have received a low conductivity prediction from these models may still be of interest. This emphasizes the importance of reporting newly synthesized materials with distinct chemistry and their measured properties. This should be encouraged even if said property is not seen as being "exceptional" in comparison to heavily investigated and optimized materials families that have seized the attention of many researchers.

Acquiring new data is the only route to improving the performance of supervised models in outlier conductivity regions. Diversification of the structure and bonding within studied ionic conductors expands the predictive utility of these models because the database on which they are trained is more representative. This experimental synthetic exploration of uncharted chemical (composition and structure) space to generate new examples is thus of foundational importance, regardless of the absolute performance of the arising material. Each qualitatively distinct material in terms of differentiated structure and bonding assists our understanding of where high performing materials may be located in chemical space. This distinguishes the generation of materials closely related to existing examples—which is valuable for optimization—from studies that explore distinct parts of the relevant chemical space. The model performance here reinforces the importance of exploratory discovery synthesis coupled with definition of structure-property-composition relationships for lithium ion transport.

METHODS

Database construction

A visual interface was developed using the python library Streamlit 0.60.0. Data is read into the application using pandas 1.0.1, with interface fields to select the researcher and currently presented data entry. The pdfs of each paper, which had been downloaded during earlier validation stages, were presented to each researcher on each page by dynamically updating the file address in an embedded iframe, and running a python 3.7 http server in the pdf folder. Fields for comments were included in the application, which were stored in a csv file and updated manually after each round of validation.

Unsupervised learning

The PCA map of the ICSD was created by using the numpy 1.21.2 singular value decomposition implementation, applied to a centred 32-bit floating point ElMD[28] 0.4.15 kernel matrix, to project the distances of each point to two-dimensional co-ordinates. UMAP[32] embeddings were generated using umap-learn 0.5.2 with an increased spread value of 5, a random seed of 5, and default parameters otherwise.

Supervised learning

LOCO and *k*-folds cross validation methods were applied (discussed previously) using AutoSklearn[38] and CrabNet[39] models. AutoSklearn 0.14.5 models were trained on 128 vCPUs (dual AMD EPYC 7502)

with default hyperparameters and a timeout of 600 s. CrabNet (commit 6296be6b06dde24a5d32e3a42657ef0ba0339344) models were generated using a batch size of 512, a RobustL1 loss function, a Lamb lookahead optimizer with stochastic weight averaging, a cyclic learning rate from 1×10^{-4} to 6×10^{-3}, and a Leaky ReLU activation function. CrabNet models were trained as discussed previously on an Nvidia Quadro RTX 4000. Experiments may be replicated using the code provided in the "Code availability" section.

DATA AVAILABILITY

The dataset is freely available for academic use, at http://pcwww.liv.ac.uk/~msd30/lmds/LiIonDatabase.html in the form of a csv file. Please contact the corresponding author for all commercial enquiries, as per the licence of the dataset.

CODE AVAILABILITY

Final models and supporting code have been provided as part of this report; neural network parameter weights are available at https://github.com/lrcfmd/LiIonML.

Received: 5 May 2022; Accepted: 10 December 2022;
Published online: 16 January 2023

REFERENCES

1. Goodenough, J. B. Rechargeable batteries: challenges old and new. *J. Solid State Electrochem.* **16**, 2019–2029 (2012).
2. Knauth, P. Inorganic solid Li ion conductors: an overview. *Solid State Ion.* **180**, 911–916 (2009).
3. Janek, J. & Zeier, W. G. A solid future for battery development. *Nat. Energy* **1**, 1–4 (2016).
4. Wang, Y. et al. Design principles for solid-state lithium superionic conductors. *Nat. Mater.* **14**, 1026–1031 (2015).
5. Bachman, J. C. et al. Inorganic solid-state electrolytes for lithium batteries: mechanisms and properties governing ion conduction. *Chem. Rev.* **116**, 140–162 (2016).
6. Lombardo, T. et al. Artificial intelligence applied to battery research: hype or reality? *Chem. Rev.* https://doi.org/10.1021/acs.chemrev.1c00108 (2021).
7. Sendek, A. D., Cheon, G., Pasta, M. & Reed, E. J. Quantifying the search for solid Li-ion electrolyte materials by anion: a data-driven perspective. *J. Phys. Chem.* **124**, 8067–8079 (2020).
8. Zhang, Y. et al. Unsupervised discovery of solid-state lithium ion conductors. *Nat. Commun.* **10**, 5260 (2019).
9. Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
10. Haghighatlari, M., Shih, C.-Y. & Hachmann, J. Thinking globally, acting locally: on the issue of training set imbalance and the case for local machine learning models in chemistry. Preprint at https://chemrxiv.org/engage/chemrxiv/article-details/60c745c4337d6cef32e2704f (2019).
11. De Breuck, P.-P., Evans, M. L. & Rignanese, G.-M. Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet. *J. Phys. Condens. Matter* **33**, 404002 (2021).
12. 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).
13. Irvine, J. T. S., Sinclair, D. C. & West, A. R. Electroceramics: characterization by impedance spectroscopy. *Adv. Mater.* **2**, 132–138 (1990).
14. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
15. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5**, 180111 (2018).
16. Sendek, A. D. et al. Holistic computational structure screening of more than 12,000 candidates for solid lithium-ion conductor materials. *Energy Environ. Sci.* **10**, 306–320 (2017).
17. Fergus, J. W. Ceramic and polymeric solid electrolytes for lithium-ion batteries. *J. Power Sources* **195**, 4554–4569 (2010).
18. Rossbach, A., Tietz, F. & Grieshammer, S. Structural and transport properties of lithium-conducting NASICON materials. *J. Power Sources* **391**, 1–9 (2018).
19. Stramare, S., Thangadurai, V. & Weppner, W. Lithium lanthanum titanates: a review. *Chem. Mater.* **15**, 3974–3990 (2003).
20. Zhang, Z. et al. New horizons for inorganic solid state ion conductors. *Energy Environ. Sci.* **11**, 1945–1976 (2018).
21. Chen, C. & Du, J. Lithium ion diffusion mechanism in lithium lanthanum titanate solid-state electrolytes from atomistic. *Simul. J. Am. Ceram. Soc.* **98**, 534–542 (2015).
22. Xiang, Y.-X. et al. Toward understanding of ion dynamics in highly conductive lithium ion conductors: some perspectives by solid state NMR techniques. *Solid State Ion.* **318**, 19–26 (2018).
23. Nolan, A. M., Zhu, Y., He, X., Bai, Q. & Mo, Y. Computation-accelerated design of materials and interfaces for all-solid-state lithium-ion batteries. *Joule* **2**, 2016–2046 (2018).
24. Manawan, M., Kartini, E. & Avdeev, M. Visualizing lithium ions in the crystal structure of Li_3PO_4 by in situ neutron diffraction. *J. Appl. Crystallogr.* **54**, 1409–1415 (2021).
25. Radford, A. et al. Language Models are Unsupervised Multitask Learners. OpenAI Blog https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf (2019).
26. Wolf, T. et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* 38–45 (Association for Computational Linguistics, 2020).
27. Ling, C. A review of the recent progress in battery informatics. *Npj Comput. Mater.* **8**, 33 (2022).
28. Hargreaves, C. J., Dyer, M. S., Gaultois, M. W., Kurlin, V. A. & Rosseinsky, M. J. The Earth mover's distance as a metric for the space of inorganic compositions. *Chem. Mater.* **32**, 10610–10620 (2020).
29. Levin, I. NIST Inorganic Crystal Structure Database (ICSD) https://doi.org/10.18434/M32147 (2020).
30. Krzanowski, W. *Principles of Multivariate Analysis* Ch. 2 (Oxford University Press, 2000).
31. Kobak, D. & Linderman, G. C. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat. Biotechnol.* **39**, 156–157 (2021).
32. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
33. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* 226–231 (1996).
34. Murdock, R. J., Kauwe, S. K., Wang, A. Y. T. & Sparks, T. D. Is domain knowledge necessary for machine learning materials properties? *Integrating Mater. Manuf. Innov.* **9**, 221–227 (2020).
35. Wang, A. Y.-T. et al. Machine learning for materials scientists: an introductory guide toward best practices. *Chem. Mater.* **32**, 4954–4965 (2020).
36. Ho, T. K. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition* Vol. 1, 278–282 (1995).
37. He, X., Zhao, K. & Chu, X. AutoML: A survey of the state-of-the-art. *Knowl. -Based Syst.* **212**, 106622 (2021).
38. Feurer, M. et al. Efficient and robust automated machine learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems* Vol. 2, 2755–2763 (2015).
39. 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).
40. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30**, 6000–6010 (2017).
41. Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nat. Commun.* **11**, 6280 (2020).
42. Kong, S., Guevarra, D., Gomes, C. P. & Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **8**, 021409 (2021).
43. Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Npj Comput. Mater.* **1**, 1–15 (2015).
44. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
45. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta BBA - Protein Struct.* **405**, 442–451 (1975).
46. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020).
47. Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *Npj Comput. Mater.* **4**, 1–14 (2018).
48. Suzuki, K. et al. Synthesis, structure, and electrochemical properties of crystalline Li–P–S–O solid electrolytes: novel lithium-conducting oxysulfides of $Li_{10}GeP_2S_{12}$ family. *Solid State Ion.* **288**, 229–234 (2016).
49. Gamon, J. et al. $Li_{4.3}AlS_{3.3}Cl_{0.7}$: a sulfide–chloride lithium ion conductor with highly disordered structure and increased conductivity. *Chem. Mater.* **33**, 8733–8744 (2021).
50. Jiang, Y., Hu, Z., Ling, M. & Zhu, X. A comparative study of $Li_{10.35}Ge_{1.35}P_{1.65}S_{12}$ and $Li_{10.5}Ge_{1.5}P_{1.5}S_{12}$ superionic conductors. *Funct. Mater. Lett.* **13**, 2050031 (2020).

51. Sun, Y., Suzuki, K., Hori, S., Hirayama, M. & Kanno, R. Superionic conductors: $Li_{10+\delta}[Sn_ySi_{1-y}]_{1+\delta}P_{2-\delta}S_{12}$ with a $Li_{10}GeP_2S_{12}$-type structure in the Li_3PS_4–Li_4SnS_4–Li_4SiS_4 quasi-ternary system. *Chem. Mater.* **29**, 5858–5864 (2017).
52. Sun, Y. et al. Oxygen substitution effects in $Li_{10}GeP_2S_{12}$ solid electrolyte. *J. Power Sources* **324**, 798–803 (2016).
53. Hori, S., Suzuki, K., Hirayama, M., Kato, Y. & Kanno, R. Lithium superionic conductor $Li_{9.42}Si_{1.02}P_{2.1}S_{9.96}O_{2.04}$ with $Li_{10}GeP_2S_{12}$-type structure in the Li_2S–P_2S_5–SiO_2 pseudoternary system: synthesis, electrochemical properties, and structure–composition relationships. *Front. Energy Res.* **4**, 38 (2016).
54. 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).
55. Han, G. et al. Extended condensed ultraphosphate frameworks with monovalent ions combine lithium mobility with high. *Comput. Electrochem. Stab. J. Am. Chem. Soc.* **143**, 18216–18232 (2021).
56. Shoko, E. et al. Polymorph of $LiAlP_2O_7$: combined computational, synthetic, crystallographic, and ionic conductivity study. *Inorg. Chem.* **60**, 14083–14095 (2021).

ACKNOWLEDGEMENTS

This work was supported by the University of Liverpool (studentship to C.J.H.), by the Faraday Institution (SOLBAT, grant number FIRG007), and by EPSRC under EP/V026887 and EP/R018472/1. The authors thank the Leverhulme Trust for funding this research via the Leverhulme Research Centre for Functional Materials Design (RC-2015-036). This work was undertaken on Barkla, part of the High-Performance Computing facilities at the University of Liverpool, UK. K.T., B.-E.P., C.A.C., J.G., G.H., B.T.L., A.J.P., A.R., O.R., P.M.S., W.J.T., A.V., and L.W. thank the UK Engineering and Physical Sciences Research Council (EPSRC) for funding under EP/N004884. We acknowledge the ICSF Faraday Challenge projects "SOLBAT – The Solid-State (Li or Na) Metal-Anode Battery" [grant number FIRG007] and "All-Solid State Lithium Anode Battery 2" [grant number FIRG026] for funding Y.D., A.M., C.M.C. and E.S., including partial support of a studentship to B.B.D., who is also supported by the University of Liverpool. V.A.K. thanks the Royal Academy of Engineering for their fellowship support [ref IF2122\186]. We acknowledge the ICSF Faraday Institution projects "CATMAT – Next Generation Li-Ion Cathode Materials" [grant number FIRG016] for funding M.S.

AUTHOR CONTRIBUTIONS

M.W.G. and M.S.D. conceived the project and led the initial collaboration with E.J.W., data gathering and validation stages were continued by C.J.H., M.W.G., M.M., and L.M.D. This included the efforts of Y.D., R.M., A.M., K.T., M.A.W., B.E.P., F.B., C.M.C., C.A.C., B.B.D., J.E., J.G., G.H., G.T.L., H.N., A.J.P., A.R., O.R., P.M.S., E.S., M.S., W.J.T., A.V., and L.W. Further validation was carried forward by C.J.H., M.W.G., L.M.D., M.M., Y.D., R.M., A.M., K.T., M.A.W., and M.S.D.; C.J.H. carried forward the ML investigation under the supervision of M.W.G. and M.S.D.; C.J.H., M.W.G., L.M.D., M.J.R., and M.S.D. co-wrote the manuscript, review copies were shared to all authors for feedback.

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/s41524-022-00951-z.

Correspondence and requests for materials should be addressed to Matthew S. Dyer.

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.