npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01469-2

The ab initio non-crystalline structure database: empowering machine learning to decode diffusivity

Hui Zheng[1], Eric Sivonxay[2,3], Rasmus Christensen[1,4], Max Gallant[1,2], Ziyao Luo[1], Matthew McDermott[1,2], Patrick Huck[1], Morten M. Smedskjær[4] & Kristin A. Persson[1]

Non-crystalline materials exhibit unique properties that make them suitable for various applications in science and technology, ranging from optical and electronic devices and solid-state batteries to protective coatings. However, data-driven exploration and design of non-crystalline materials is hampered by the absence of a comprehensive database covering a broad chemical space. In this work, we present the largest computed non-crystalline structure database to date, generated from systematic and accurate ab initio molecular dynamics (AIMD) calculations. We also show how the database can be used in simple machine-learning models to connect properties to composition and structure, here specifically targeting ionic conductivity. These models predict the Li-ion diffusivity with speed and accuracy, offering a cost-effective alternative to expensive density functional theory (DFT) calculations. Furthermore, the process of computational quenching non-crystalline structures provides a unique sampling of out-of-equilibrium structures, energies, and force landscape, and we anticipate that the corresponding trajectories will inform future work in universal machine learning potentials, impacting design beyond that of non-crystalline materials. In addition, combining diffusion trajectories from our dataset with models that predict liquidus viscosity and melting temperature could be utilized to develop models for predicting glass-forming ability.

Amorphous materials are generally characterized by the lack of long-range order as a result of synthesis processes that freeze in a non-equilibrium, non-crystalline structure. Notably, such non-crystalline materials can manifest in structures characterized, e.g., as liquid, supercooled liquid, or glass[1]. Compared to the stringent synthesis requirements of crystalline materials for ordered atomic arrangement, synthesizing non-crystalline materials tends to be less energy-intensive as it is often done via low-temperature methods such as ball milling, vapor deposition, and sol-gel synthesis or through rapid cooling from the liquid state via the melt-quenching method[2]. Different synthesis methods generally lead to different structures. For glass produced via melt-quenching, the rate at which a material is cooled from a liquid to a solid state can influence the structure and, thereby, its properties significantly[1]. Such tunability can be leveraged to engineer a wide range of physical, chemical, and mechanical properties. As examples, bulk metallic glasses with unique magnetic properties are found in high-efficiency transformers[3,4], amorphous alkali-aluminosilicates were made famous as the world-leading cover glass for portable electronics[5], and amorphous 2D boron nitrides are proposed for the next-gen memory solutions due to their ultra-low dielectric constant combined with excellent electrical and mechanical properties[6,7].

Amorphous materials are also considered for various applications in energy storage. For example, amorphous anodes, particularly silicon and silicon-tin alloys, are pursued as high-capacity, lower-cost alternatives to graphite.[8–14] Furthermore, the conformal nature of amorphous materials proffers major advantages in electrode coating applications[15–18] and as electrolytes for all-solid-state batteries. While crystalline $Li_7La_3Zr_2O_{12}$ (LLZO) exhibits high Li-ion conductivity, it also allows for lithium dendrite growth through the grain boundaries[19–21], which presents safety concerns. In comparison, amorphous LLZO exhibits lower Li-ion diffusivity but shows marked improvement in safety and cyclability[21]. In contrast, amorphous lithium phosphorus oxynitride (LiPON) shows a higher Li-ion diffusivity than its crystalline counterpart[20–22], which indicates a possible design space where ionic conductivity and safety can be optimized within an amorphous phase space.

[1]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [2]Materials Science and Engineering, University of California, Berkeley, Berkeley, CA, USA. [3]Energy Technologies Area, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [4]Department of Chemistry and Bioscience, Aalborg University, Aalborg, Denmark. e-mail: kapersson@lbl.gov

Unfortunately, measuring ionic diffusivity in inorganic (amorphous or crystalline) materials is highly time-consuming, and more often than not, the inherent bulk diffusivity is masked by other factors, such as pellet densification. Therefore, current composition-property databases of glass materials, such as SciGlass[23] and INTERGLAD[24], only contain measurements of these properties for a limited number of compositions. It is possible to obtain an estimate of the ionic diffusivity through Ab Initio Molecular Dynamics (AIMD); However, unlike crystalline compounds, the atomic structures of non-crystalline materials are usually not well known, as complete structural mapping of non-crystalline materials is challenging to conduct in experiments[25]. Furthermore, non-crystalline structures require larger unit cells to capture sufficient and representative local environments compared to crystals. As of writing, the only reported structure database is the amorphous nanoporous materials database, which includes atomic configurations of 75 amorphous carbons, 119 polymers, and 16 kerogens. The database is curated from earlier literature[26] and covers a limited range of compositions. To meet the need to accelerate our discovery and design of non-crystalline materials with target functionality, in this study, we present an extensive, computed database of melt-quenched non-crystalline structures covering 4849 compositions and 79 elements generated through systematic AIMD calculations. Due to the large coverage of compositions, where the vast majority of melting and glass transition temperatures are unknown, we are unable to rigorously specify the phase of each composition. Therefore, we use the broad term "non-crystalline structure" to encompass all non-crystalline phases studied. We demonstrate one aspect of the database's applicability in training an efficient machine-learning model to rapidly and accurately predict Li diffusivity, providing a cost-effective alternative to density functional theory (DFT) calculations. We also anticipate a broader usefulness of the database as it opens up new possibilities for improving current directions in universal machine learning potentials by providing unique information about structure-energy-force relationships far from equilibrium configurations.

Results

Data scope

The synthesis method used to obtain a non-crystalline structure significantly impacts its final form. This variability poses a challenge when creating a non-crystalline structure database, as a single composition can yield a broad variety of non-crystalline structures. To address this, we have developed a self-consistent and computationally efficient methodology for generating non-crystalline structures across various chemical compositions, detailed in the Methods section. Our approach emulates the experimental melt-quenching technique, commonly used in simulations to produce non-crystalline structures. For computational efficiency, we employ instantaneous cooling to target temperatures, followed by volume relaxation using an equation of state approach.

The produced non-crystalline structure database includes two subset databases. The first one consists of 5120 compounds, which are melted at 5000K using the MPMorph workflow. Details about the workflow can be found in Methods section. This database is here denoted as the "5000K non-crystalline database", containing liquid structures of these compounds. A second lower temperature database is generated for 220 selected compounds by instantaneously quenching the last snapshot structures from the 5000K database to the target temperatures of 1000K, 1500K, 2000K, and 2500K and annealing them using the same MPMorph workflow. This database is denoted as the "multi-temperature non-crystalline database." As our database encompasses structures ranging from liquid to supercooled liquid and glassy states, we have classified it as the "non-crystalline structure database." This terminology captures the diversity of non-crystalline structures present, acknowledging their varying degrees of structural disorder.

Among the 4849 compositions in the 5000K non-crystalline database, 3533 compounds contain lithium (Supplementary Fig. 1). Figure 1 presents the proportion of each element's occurrence within the compositions in the 5000K non-crystalline database, compared to its occurrence within the Materials Project database. We note that the 5000K database exhibits extensive coverage, providing a similar representation of compositions compared to the Materials Project. The element occurrence of the compounds in the Materials Project is shown in Supplementary Fig. 2, where there are approximately 50,000 compounds containing Li. Similarly, the element occurrence in the compositions covered in the multi-temperature non-crystalline database is shown in Supplementary Fig. 3. Supplementary Fig. 4 shows the ratio of the element occurrence within the multi-temperature non-crystalline database compared to its occurrence within the Materials Project database. We find that the multi-temperature non-crystalline database also effectively captures a diverse range of material compositions, ensuring a comprehensive chemical representation.

Correlations between Li^+ diffusivity and composition

Amorphous materials exhibit short-range ordering, which is strongly dependent on the composition. For example, amorphous Al_2O_3 exhibits a distribution of 4, 5, and 6-fold oxygen-coordinated Al^{3+} units, where the distribution depends on the synthesis or formation conditions. Since cationic diffusion in amorphous structures has been shown to be highly correlated and dependent on bond-formation/breaking events between the cation and its anionic environment[15,16], we anticipate that Li^+ diffusion in non-crystalline inorganic materials will correlate strongly to anion specie and composition. In the following section, we identify and analyze the correlations between the Li-ion diffusivity and (1) the composition of the materials, (2) the size of the anions and cations, and (3) the electronegativity difference between the compositional species.

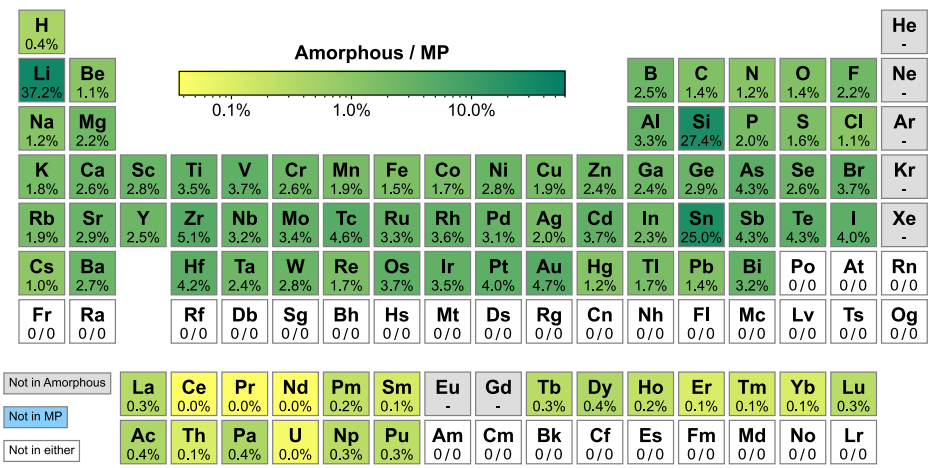


Fig. 1 | Elemental occurrence in the 5000K non-crystalline database compared to the Materials Project. Element occurrence ratios for compositions in the non-crystalline database are shaded by color scale.

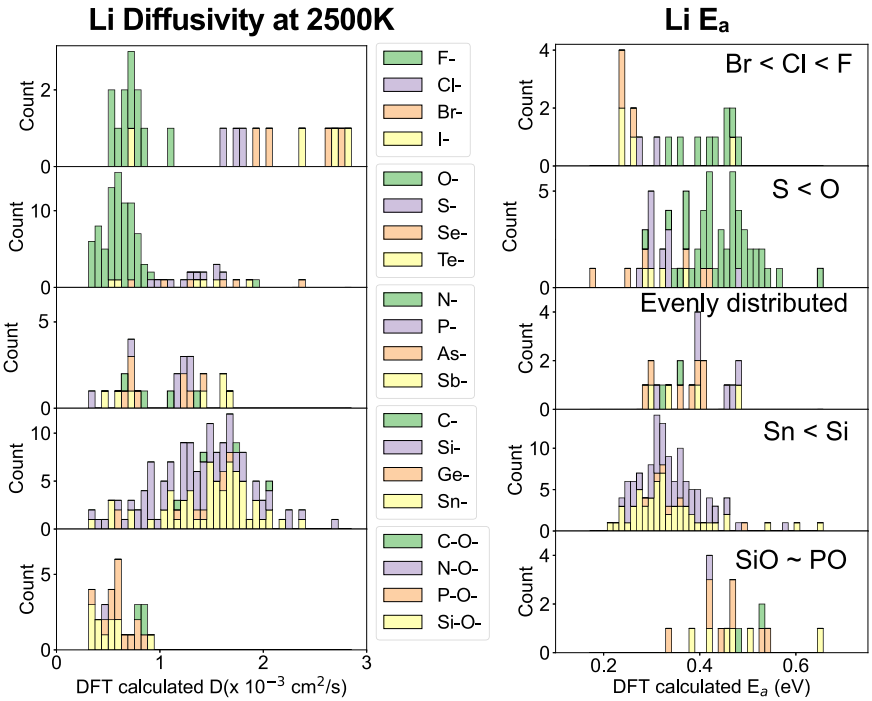


Fig. 2 | Li diffusivities and activation energies. Distributions of Li diffusivity, D (2500 K), and activation energy, E_a, calculated from the multi-temperature non-crystalline database. Compositions are sorted based on the anion element present in the system and collated by group on the periodic table. The annotation on the right panel shows the Li *Ea* order of the peak location from kernel density estimate (KDE) as shown in Supplementary Fig. 6.

In the context of data coverage, we emphasize that the samples obtained from the collected group may have different sizes. For instance, there is a higher number of compounds containing oxygen compared to those containing sulfur, selenium, and tellurium. Similarly, there is a greater presence of compounds with fluorine compared to compounds containing chlorine, bromine, and iodine. This is demonstrated in Fig. 2 and Supplementary Fig. 3. A small sample size may impact the accuracy in comparing different anion groups. Therefore, our analysis focuses on compounds with larger sample sizes, ensuring that the distributions are distinct enough to yield conclusive results.

Figure 2 and Supplementary Fig. 6 show the Li diffusivity and activation barrier distributions calculated from the multi-temperature non-crystalline database. We make several observations of Li-ion diffusivity trends within different groups of the Periodic Table. Within the halogen group, compositions that include fluorine (F) demonstrate significantly lower diffusivity and higher activation energy (E_a) compared to those containing chlorine (Cl), bromine (Br), or iodine (I). The general trend matches the order of the bond dissociation energy. i.e., Li-F has the highest bond strength of 577 kJ/mol, compared to 469 kJ/mol for Li-Cl, 423 kJ/mol for Li-Br, and 352 kJ/mol for Li-I.[27] Similarly, in the chalcogen group, compositions incorporating oxygen (O) exhibit lower diffusivity and a higher E_a when compared to those that include sulfur (S). Compounds containing elements from the VA group have been evenly distributed E_a due to the small sample size. Compounds with tin (Sn) have slightly lower E_a compared to those with silicon (Si). These trends all follow a similar pattern, such that a larger atomic radius of the anion species—corresponding to elements from a larger row number within the same group—results in lower Li activation energy (E_a) and, thus, higher Li diffusivity. Correspondingly, higher electronegativity or higher charge density leads to stronger bonding between Li and anions, resulting in higher E_a for Li diffusion. The even distribution of activation energies (E_a) among compounds containing oxyanions, as shown in the bottom distribution, may be attributable to the small sample size.

In addition to the correlation between anion species and Li diffusivity, the presence of other cations can also influence Li diffusivity. Supplementary Fig. 7 depicts the average Li diffusivity values from the 5000K non-crystalline database across compounds containing specific elements from the periodic table. Certain elements correspond to higher Li diffusivity than others. For example, there are two regions of elements that contribute to high Li diffusivity: the alkali/alkaline metals group (IA and IIA) and the right-hand side of the periodic table, encompassing groups IB, IIB, and IIIA through VIIA. For compounds containing elements from these groups, a trend is observable: with an increasing row number (and hence, larger atomic radius), Li diffusivity also increases. The presence of cations originating from groups IIIA to VIIA will likely result in polyanionic environments (carbonates, nitrates, phosphates, polyhalogens, etc.) within the non-crystalline material, which on average, leaves the Li^+ less directly coordinated to oxygen and hence more free to move. Supplementary Fig. 8 further illustrates the standard deviation (STD) of Li diffusivity for compounds containing specific periodic elements. Notably, compounds that incorporate elements from groups IIB and IIIA to VIA demonstrate smaller STDs when compared to compounds containing alkali elements.

Supplementary Figs. 9 and 10 present similar plots for the activation barrier (E_a) of Li^+, derived from the more limited multi-temperature non-crystalline dataset. Owing to the smaller sample size of the data, the distribution of elements associated with lower Li E_a is not as pronounced as the Li diffusivity distribution from the 5000K database displayed in Supplementary Fig. 7. Nevertheless, compounds containing Cl, Br, I, S, Se, Pb, Sr, Sn, In, Ba, Na, K, and Rb demonstrate lower E_a than other compounds. This observation aligns with cases of high Li diffusivity in the 5000K database, as depicted in Supplementary Fig. 7.

Our analysis encompasses compounds that range from binary to ternary, quaternary, and even quinary. Consequently, Li-ion diffusivity associated with one element often cross-correlates with other elements present in the same set of compounds. This necessarily results in some over-counting and cooperative effects on Li diffusivity or E_a. We focus on available binary LiX compounds (where X represents any species within the composition) to deconvolute these relationships. This approach allows us to clarify the correlations between Li diffusivity and other elements. Figure 3 illustrates the correlation between the activation barrier of Li (E_a) and the properties of the X species in LiX compounds. Panel a reveals a negative correlation between E_a and the Li fraction in the composition. This observation corresponds to a similar phenomenon found in crystalline solid-state electrolyte systems, denoted "Li stuffing", where increased Li content improves Li diffusion[28]. A similar trend is observed in the curated

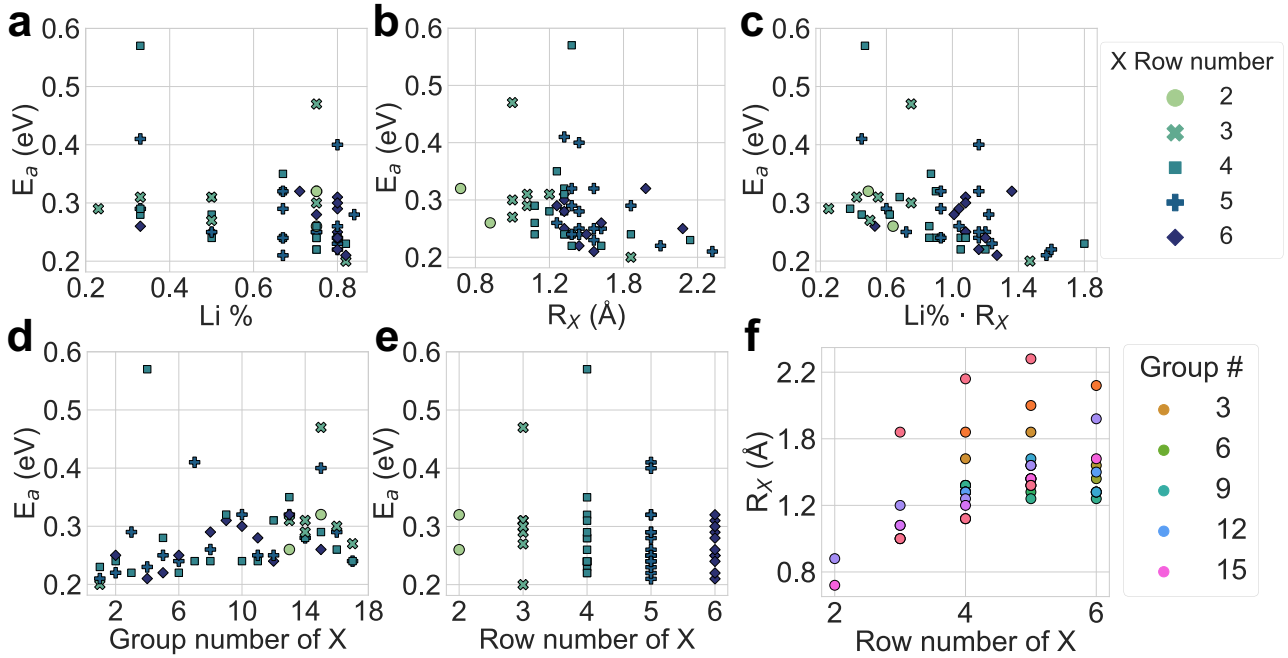


Fig. 3 | Effect of composition on Li activation energies in binary compounds. **a**–**e** show the correlations between the activation energy E_a of Li in binary compounds and the elemental properties of the coexisting species X and Lithium fraction (Li%) in the binary composition (LiX), such as atomic radius of element X (R_X), the multiplication of these two (Li% · R_X), group number of X, row number of X. Color and marker shapes signify the row number of element X. **f** Plots the correlation between the R_X and row number of X; colors are used to distinguish group numbers of element X.

experimentally measured Li conductivity in glasses, as reported by Hargreaves et al.[29] and visualized in Supplementary Fig. 22. SciGlass[23] database also reported limited Li diffusivity; a similar trend is observed in Supplementary Fig. 23. The parity plots between the experimental and DFT-calculated data of Li diffusivity and Li conductivity can be found in Supplementary Fig. 24. Figure 3b demonstrates a strong negative relationship between Li E_a and the atomic radius of X (R_X), showing that larger X species facilitates Li diffusion by providing more spacious frameworks and—in the case of anions, lowers the electronegativity. In Supplementary Fig. 11, the color gradient further clarifies why some large-radius points do not exhibit correspondingly small E_a: a lower Li percentage. Thus, for compounds with similar Li percentages, a larger atomic radius of X implies lower E_a. Figure 3c presents an approximately linear negative correlation between Li E_a and the product of Li fraction (Li%) and atomic radius of X R_X. Panels d and e explore the influence of X species' group and row numbers on Li E_a in LiX. As the group number of X and the corresponding electronegativity difference between Li and X increase, Li E_a tends to rise due to stronger Li-X bonds, thus inhibiting Li diffusion. Although an increase in X row number generally results in lower mean values of Li E_a, the impact is not pronounced due to the wide E_a distribution within species of the same row number. Panel f highlights the trend that larger X row numbers correlate with larger atomic radii, with the hue distinguishing X species from different groups.

Feature design for machine learning

As previously analyzed, both compositional and elemental properties of species correlate with Li diffusivity. However, other features also directly or indirectly impact Li diffusivity. Building on the work of Sendek et al.[30], who developed a feature set to differentiate high- and low-diffusivity materials, we have expanded the feature set to include more compositional features. This approach equips the machine learning models to learn the underlying correlations between the features and Li diffusivity more comprehensively. The expanded list of features, sorted by their Pearson correlation coefficient, is provided in Table 1. The compositional features added to the feature set here include Li percentage (Li%), which has a significant positive correlation with the Li diffusivity in different compositions. A similar trend has been observed in amorphous coating materials for Li-ion batteries as well as crystalline solid-state electrolyte materials[15,16,28]. The weighted average of cohesive energy $\overline{E_{coh}}$ is calculated from the cohesive energy of the ground state of the constituent elemental systems. The cohesive energy provides a useful metric for describing the average bond strength of the local units in the non-crystalline material. Specifically, the stronger the bonds, the harder it is for the activated bond-breaking process to occur, which underpins the diffusion process. Hence, E_{coh} negatively correlates with the Li diffusivity. In addition, a few other features show a negative correlation with Li diffusivity and can be explained in terms of the packing fraction of a set of non-lithium (non-Li) atoms that are in close proximity to a central lithium (Li) atom within a specified radius. We here denote these features as (1) set-of-non-Li-atoms packing fraction (SPF), (2) set-of-non-Li-atoms neighbor count (SNC), (3) Li neighbor count (LNC), (4) Li bond ionicity (LBI), and (5) density. These structural features also show a strong negative correlation with Li diffusivity, as when the non-Li atoms are parking closely and form a tight structural motif framework, it is harder for Li to diffuse. For more details on the quantitative definition of other features, such as weighted average bulk moduli ($\overline{B}$) and electronegativity ($\overline{X}$), please refer to Table 1, supplementary materials, and Supplementary Note of reference by Sendek et al.[30].

Machine-learning models

Three different diffusivity-prediction models are trained on the non-crystalline database. Two ensemble learning models, Random Forest (RF)[31] and Extreme Gradient Boosting (XGBoost)[32], were employed to learn the temperature-dependent diffusivity. The parity plots comparing the DFT calculated Li diffusivity and ML-predicted Li diffusivity for both training and test data are shown in Fig. 4. We observe that both algorithms achieve coefficients of determination (R^2) close to 1, very low mean absolute error (MAE), and root mean squared error (RMSE). Fivefold cross-validations have been used to assess the performance and the generalization ability of these two models via Scikit-learn[33]. Both RF and XGBoost measure feature

Table 1 | Pearson correlation coefficients between various features and the Li diffusivity, obtained from 5000 K AIMD calculations

Feature	Feature description	Pearson *r*	Unit
$\overline{E_{coh}}$	weighted cohesive energy	−0.81	eV
SPF	set-of-non-Li-atoms packing fraction	−0.76	1
density	weight/volume	−0.67	$g \cdot cm^{-3}$
LNC	Li neighbor count	−0.64	1
$\overline{B}$	weighted bulk modulus	−0.64	GPa
$\overline{X}$	weighted electronegativity	−0.64	1
SNC	set-of-non-Li-atoms neighbor count	−0.63	1
LBI	Li bond ionicity	−0.61	1
AFC	anion framework coordination	−0.54	1
$\overline{E_{coh}^{Li}}$	weighted cohesive energy exclude Li	−0.53	eV
SLPE	straight-line path electronegativity	−0.51	1
PF	packing fraction	−0.46	1
SDLI	standard deviation of Li bond ionicity	−0.34	1
$\overline{m}$	weighted atomic mass	−0.32	kg
$\overline{G}$	weighted shear modulus	−0.29	GPa
ENS	electronegativity of set-of-non-Li-atoms	−0.29	1
$\overline{X^{Li}}$	weighted electronegativity exclude Li	−0.29	GPa
$\overline{B^{Li}}$	weighted bulk modulus exclude Li	−0.28	GPa
LLSD	Li-Li separation distance	−0.27	Å
RBI	ratio of LBI and SBI	−0.24	1
$\overline{G^{Li}}$	weighted shear modulus exclude Li	−0.10	GPa
SDLC	standard deviation in Li neighbor count	−0.08	1
$\frac{\overline{R^{Li}}}{\overline{R}}$	the ratio of the average radius without Li and with Li	−0.04	1
$\bar{R}^2 \cdot \sqrt{\frac{\overline{B \cdot R}}{\overline{m}}}$-Li	synthetic feature (exclude Li)	−0.01	m^2/s
Li%	Li percentage	0.77	1
SLPW_pp	average straight-line path width (point-to-point)	0.59	Å
AAV	average atomic volume	0.58	$Å^3$
$\overline{R}$	weighted average atomic radius	0.56	Å
$\overline{X_{others}} - \overline{X_{Li}}$	weighted average electronegativity difference	0.54	1
LLB	Li-Li bonds per Li	0.49	1
LASD	Li-anion separation distance	0.49	Å
SLPW	average straight-line path width	0.48	Å
VPA	volume per anion	0.44	$Å^3$
AASD	minimum anion-anion separation distance	0.36	Å
$\overline{R^{Li}}$	weighted average atomic radius (exclude Li)	0.35	Å
$\sqrt{\frac{\overline{E_{coh}} \cdot \bar{R}^2}{\overline{m}}}$	synthetic feature	0.29	m^2/s
SBI	set-of-non-Li-atoms bond ionicity	0.26	1
$\overline{m^{Li}}$	average mass (exclude Li)	0.21	kg
$\bar{R}^2 \cdot \sqrt{\frac{\overline{B \cdot R}}{\overline{m}}}$	synthetic feature	0.18	m^2/s
$\bar{R}^2 \cdot \sqrt{\frac{\overline{G \cdot R}}{\overline{m}}}$	synthetic feature	0.06	m^2/s
RNC	the ratio of LNC and SNC	0.01	1
$\bar{R}^2 \cdot \sqrt{\frac{\overline{G \cdot R}}{\overline{m}}}$-Li	synthetic feature (exclude Li)	0.01	m^2/s

The coefficients are categorized into two bins—positive and negative—and the values within each bin are sorted by the Pearson correlation coefficients.

importance, which indicates the relative significance of a particular feature in diffusivity prediction. The top 11 most relevant features identified from RF and XGBoost models are shown in Fig. 4c and f, respectively. We also use the SHAP method to analyze the feature importance as listed in Supplementary Figs. 13, 14. Somewhat trivially, both models rank temperature as the most important feature for predicting the diffusivity at different temperatures. Further, while the orders of the important features predicted from RF and XGBoost may differ, both models share similar highly ranked features: average atomic volume (AAV), the ratio (RBI) of average Li bond ionicity (LBI) with average bond ionicity of set-of-non-Li-atoms (SBI), average Li neighbor count (LNC), Li percentage in the compositions (Li%), etc. The definitions of features can be referenced in Table 1, Supplementary Note, and ref. 30 for details.

Finally, we developed a descriptor for Li^+ diffusivity using the sure independence screening and sparsifying operator (SISSO) method[34]. The SISSO model training and prediction results are shown in Fig. 5. The model successfully captures the relationship similar to the Arrhenius equation between features and Li diffusivity. Specifically, the temperature term $\frac{1}{k_B T}$ is present in each model shown in Fig. 5. As the dimensionality (n) of the model increases from 1 to 6, the RMSE decreases monotonically and converges around when $n = 4$ for the training set. The parity plots for $n = 5$ and $n = 6$ are omitted due to the marginal improvement observed for both training and test datasets; therefore, the $n = 4$ model is selected as the final SISSO model. The analytical equation for the four-dimensional model is as follows:

$$\widehat{lnD_{Li}} = -0.33 \times \frac{1}{k_B T} - 0.11 \times LNC + 0.11 \times LLB - 1.95 \times PF - 3.55 \quad (1)$$

The first term of this equation resembles the Arrhenius relationship $lnD = lnD_0 - \frac{E_a}{k_B T}$, predicting a linear correlation between the natural logarithm of D_{Li} and the inverse temperature ($\frac{1}{T}$), where k_B is the Boltzmann constant. This term captures that elevated temperatures tend to increase Li diffusivity. Although the SISSO model predicts that lnD_{Li} scales linearly with $\frac{1}{T}$, this behavior cannot be guaranteed, as diffusivity may be influenced by phase changes, particularly when crossing the glass transition.

The remaining terms in the equation reveal a negative correlation between lnD_{Li} and both the Li neighbor count (LNC) and the structure's packing fraction (PF), consistent with the negative Pearson correlation coefficients. This aligns with the intuition that if Li is bonded with more neighbors and the structure is more densely packed, it becomes harder for Li to diffuse, resulting in a lower D_{Li}. Conversely, the positive correlation between the number of Li-Li bonds per Li (LLB) and D_{Li} suggests that when Li atoms are surrounded by more Li atoms, the D_{Li} increases.

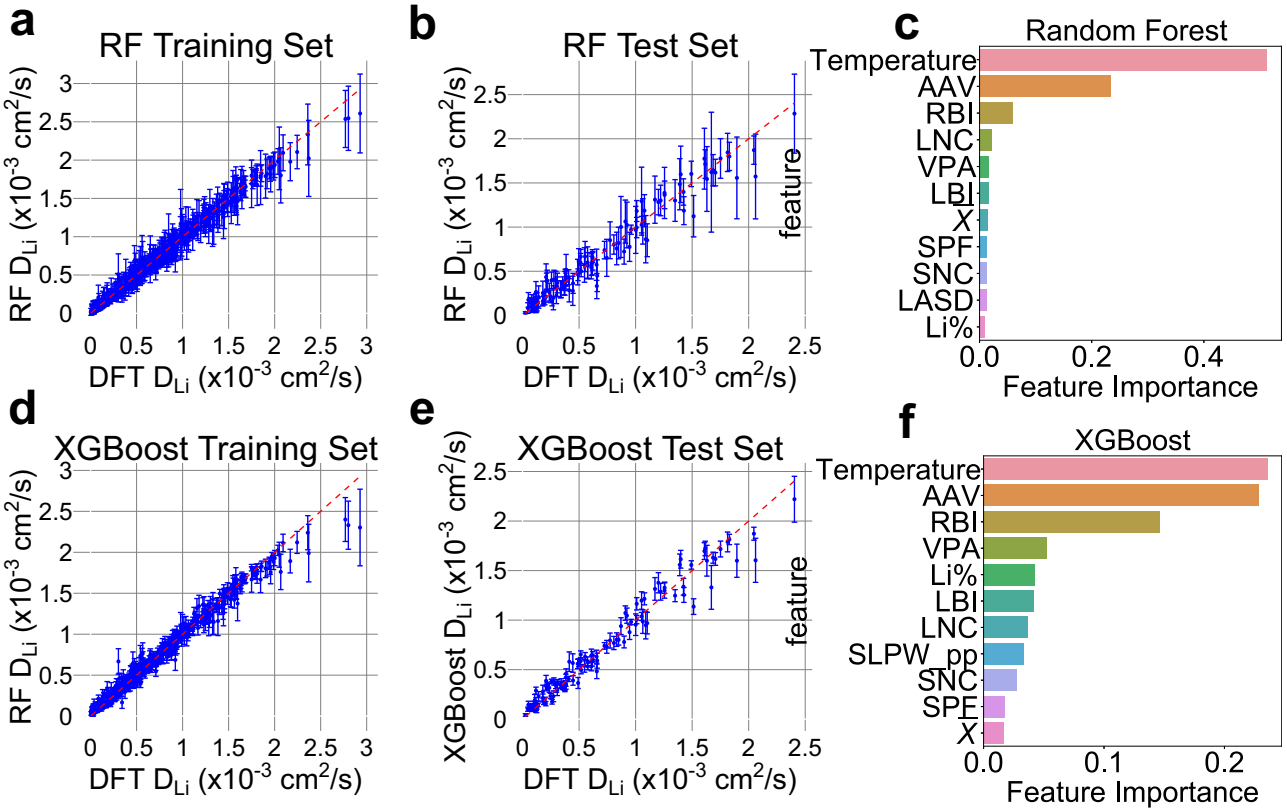


Fig. 4 | Performance of ensemble learning models in diffusivity prediction. The random forest and XGBoost models are used to predict the Li diffusivity. Parity plots between the DFT calculated Li diffusivity, and the random forest model (**a** for training and **b** for testing) and XGBoost model (**d** for training and **e** for testing) predicted Li diffusivity, respectively. **c**, **f** Ranked the top 11 important features analyzed from the random forest and XGBoost models, respectively. The performance of these two models was assessed using the average fivefold cross-validation (CV) root mean squared error (RMSE). The average fivefold CV RMSE for the random forest is 1.41×10^{-4} cm^2/s. For the XGBoost model, the average fivefold CV RMSE was 1.23×10^{-4} cm^2/s. The error bars in the prediction of the random forest model are achieved by calculating the mean and standard deviation of predictions from all individual trees in the forest. For the XGBoost model, the bootstrapping method is used for 100 XGBoost models on different bootstrap samples to calculate the mean and standard deviation of these predictions. For both models, Li-metal and Li-Si alloys (e.g., $SrLi_4$, $RbLi_2$, $Na_9Li_3Sn_4$, Na_2Li_9, $Li_4(SiI_3)_3$ etc.) present higher error bars.

Application of universal machine learning potentials

Here, we explore whether the universal interatomic potential M3GNet[35], and CHGNet[36] can be used as the surrogate for AIMD calculations to generate the non-crystalline structures of any composition at a specific temperature. Figure 6a shows the pairwise RDF comparison between AIMD and M3GNet calculated structures for $LiCuSi_2$ as an example; the parity plot is shown in Fig. 6b, with the R^2 equals to 0.99, very close to 1. Additional comparisons of oxide glass between DFT-calculated and M3GNet-calculated RDFs for LiO_2, LiSiO, and $LiSi_2O$ can be found in the Supplementary Information, specifically in Supplementary Figs. 17–19. For the rest of the compounds, RDFs are available in the additional attachment file and on the GitHub repository https://github.com/Tinaatucsd/DFT_amorphous_structure. For 245 samples of non-crystalline compositions, the distribution of R^2 of RDF comparison is plotted in Fig. 6c, where 91% of samples exhibit an $R^2 > 0.85$, 85% of samples show $R^2 > 0.9$, and 68% of samples manifest $R^2 > 0.95$. The parity plot of the structure feature comparison is shown in Fig. 6d, with a decent R^2 of 0.95. The scales of the structure features have a wide range; the parity plots with different scales can be found in Supplementary Fig. 15. At all ranges, the M3GNet MD calculations can reproduce the AIMD-calculated structures. Therefore, we find that M3GNet is able to generate reasonable non-crystalline structures and calculate structure features as inputs for the SISSO model to predict Li diffusivity. However, while M3GNet is able to reproduce the Li diffusivity, as shown in Supplementary Fig. 16a, decently well for the temperature range from 1000K to 2500K, it fails to reproduce AIMD-calculated Li diffusivity at high temperatures (5000K). Therefore, it is suggested that M3GNet be used as a surrogate for AIMD calculations to generate non-crystalline structures and then used to calculate structure features for the SISSO model to predict Li diffusivity. By employing M3GNet-based molecular dynamics (MD), the calculations achieve a significant speedup of ~2000 times (in CPU hours) compared to traditional AIMD methods for diffusivity calculation.

Discussion

We have developed a comprehensive database for non-crystalline structures, employing precise but computationally intensive AIMD calculations for 4849 compositions, spanning from binary alloys Na_2Li_9, $RbLi_2$, $SrLi_4$, Li_4Zr, and Li_4Ta to ternary, quaternary compounds like $Li_4(SiI_3)_3$, $Sr_2Li_{12}Sn$, $Li_5La_3Nb_2O_{12}$ and $Li_{20}Si_2NiSn_2$. This database provides a robust platform for various streamlined machine-learning models, enabling rapid and accurate predictions of ionic diffusivity, here demonstrated for Li^+ and relevant for applications such as protective coatings and solid-state electrolytes. Universal potentials such as M3GNet and CHGNet, which are predominantly trained on crystalline relaxation trajectories, significantly accelerate calculations compared to traditional AIMD methods and are found to perform well in structure generation but less so for providing ionic diffusivity data. The publication of this database provides unique information about structure-energy-force relationships far away from equilibrium configurations, and we anticipate that it will be a valuable asset in the pursuit of superior universal potentials applicable to non-crystalline materials. Our database also provides a comprehensive resource for mapping

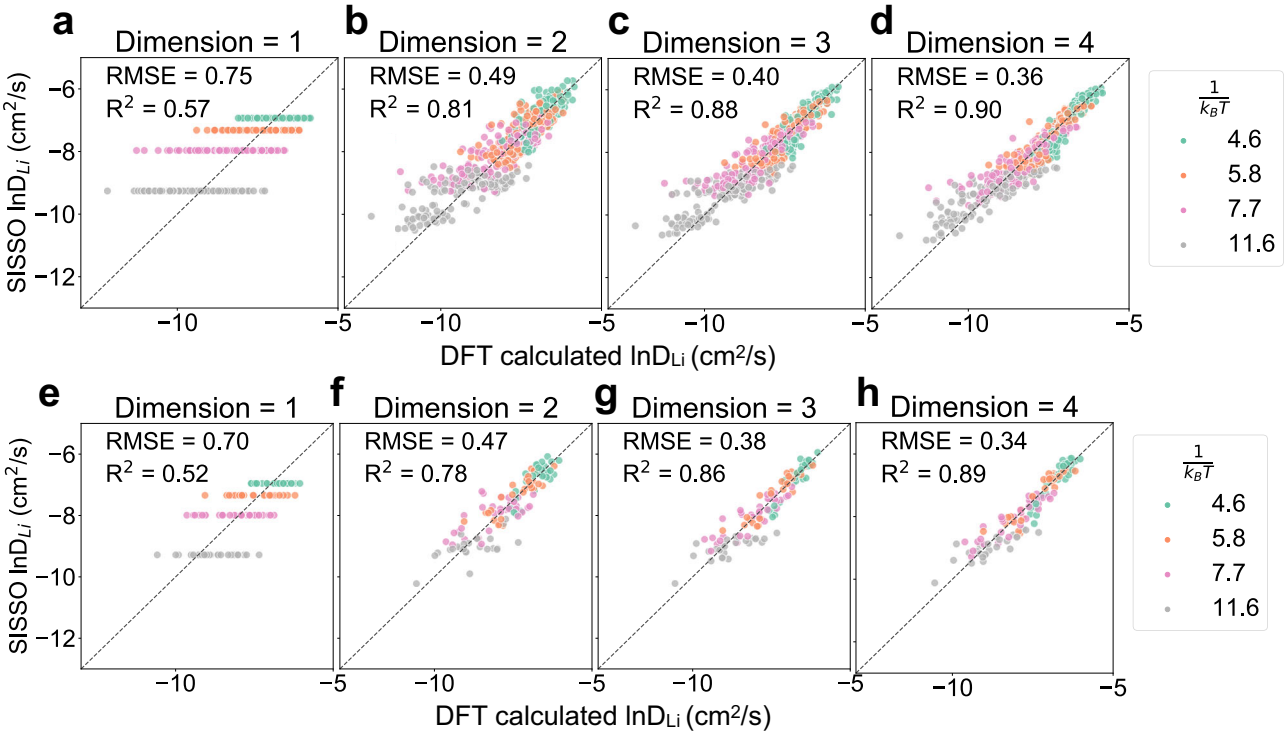
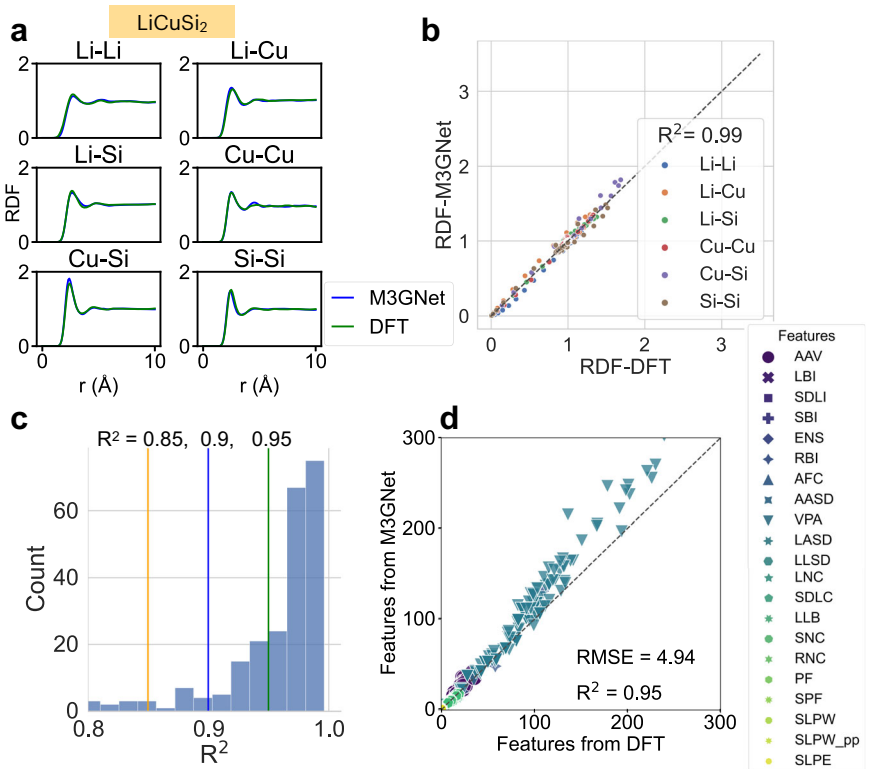


Fig. 5 | Performance of SISSO models with different complexities in diffusivity prediction. Parity plots between the DFT calculated Li diffusivity and the SISSO model predicted Li diffusivity. **a**–**d** Correspond to the parity plots for training data. The colors indicate the values of the first descriptor in the SISSO model, specifically the inverse of the product of the Boltzmann constant k_B and the temperature term $\frac{1}{k_B T}$. Subfigures **e**–**h** are the parity plots for the test data. The complexity of the SISSO model increases from (**a**–**d**), i.e., the dimension of the descriptor (a hyperparameter) increases from one to four. The $n = 4$ model is selected as the final SISSO model. n-dimensional descriptor means the set of features selected by the n nonzero components of the solution vector **c**[34].

Fig. 6 | The performance of M3GNet in reproducing the non-crystalline structures from 5000 K MD. a Pairwise radial distribution function (RDF) of $LiCuSi_2$, **b** parity plot comparing DFT-calculated RDF against M3GNet-calculated RDF, with the coefficient of determination regression score (R^2) annotated in the legend, **c** distribution of R^2 across the 245 samples of the composition, **d** parity plot for structure features calculated from AIMD and M3GNet at 5000 K.



experimental diffraction patterns to non-crystalline atomic structures, addressing the challenge of interpreting diffraction data for materials lacking long-range order. By including pairwise radial distribution functions (RDFs) across diverse compositions and temperatures, we enhance the analysis of short- and medium-range order in non-crystalline materials, facilitating the identification and characterization of non-crystalline phases.

Methods

DFT workflow

The database is generated through a combination of well-benchmarked[37] AIMD and MPMorph workflows (see Fig. 7a, b), which are designed to generate a series of samples of non-crystalline structures and their respective dynamic behavior at a range of temperatures. The structure sample generation uses PACKMOL[38] to approximate an initial random structure for a given composition of interest. Subsequently, the MPMorph workflow (Fig. 7b) scales the volumes to 0.8 and 1.2 times the initial volume, performing a 4 ps NVT AIMD run to fit the equation of state at the specified temperature. A tentative volume is then used to execute another 4 ps NVT AIMD run, ensuring the energy and density have converged. If convergence is achieved, a 20 ps AIMD "production" run is conducted using this volume. If not, the workflow iteratively rescales the volume until a value that ensures energy and density convergence is identified. This converged volume is then employed for the 20 ps production run. As shown in Fig. 7a, the 5000K NVT runs have so far generated non-crystalline structures of 4849 compositions. The database corresponding to the 20 ps 5000K NVT run trajectories is denoted the "5000K non-crystalline database". The last snapshot structure from the 5000K run is used as the input structure for MPMorph workflow at 1000K, 1500K, 2000K, and 2500K to generate the multi-temperature non-crystalline database.

The AIMD simulations at 5000K are performed to ensure that each material reaches its molten state, regardless of its unknown melting point. This elevated temperature accelerates the attainment of equilibrium liquid structures, which would otherwise be slower at temperatures closer to the melting point. The materials in our database exhibit a wide range of melting points (Tm) and glass transition temperatures (Tg), many of which are not specifically determined. Calculating these temperatures for each composition individually would be computationally prohibitive. Consequently, we opted to systematically measure diffusivity at four intermediate temperatures (1000K, 1500K, 2000K, and 2500K) to facilitate efficient high-throughput AIMD calculations. It is crucial to note that the machine

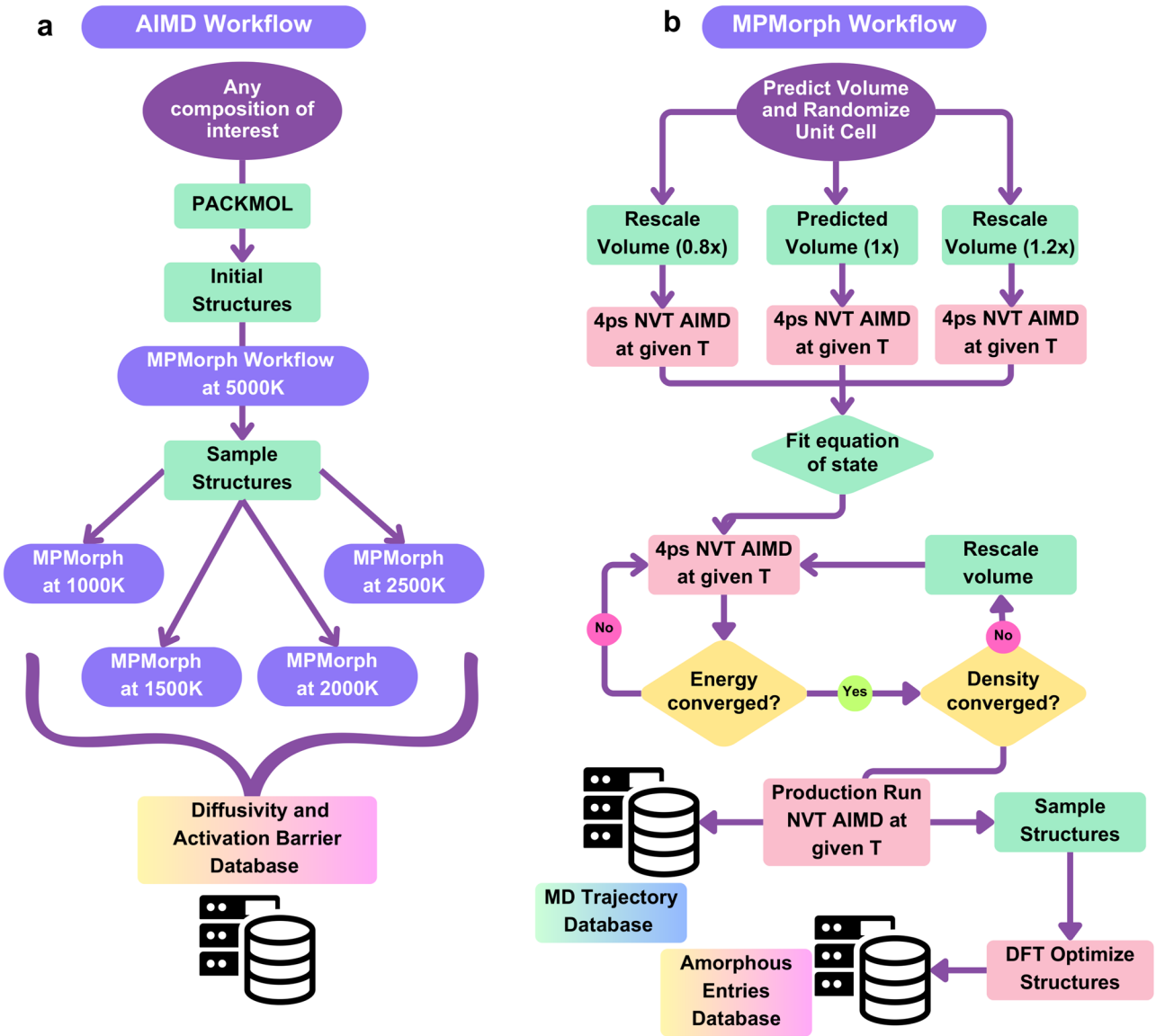


Fig. 7 | Non-crystalline database workflows. a Overview of the process used to generate the 5000K non-crystalline database and the multi-temperature non-crystalline database. **b** The MPMorph workflow procedure, which involves identifying the equilibrium volume using the NVT ensemble prior to executing the production run.

learning models discussed in this manuscript are exclusively trained on diffusivity data derived from these intermediate temperature AIMD simulations, rather than from the 5000K simulations. We posit that data obtained from these intermediate temperatures more accurately capture the dynamics of different non-crystalline states such as liquids, supercooled liquids, or even glasses.

The functional used in the database is projector-augmented-wave (PAW)[39] PBE[40]. The selection of functional is consistent with that of the Materials Project[41], such that the database can be used to assess the synthesizability of crystals based on the method used in ref. 37. The time step used is 2 fs; the Nose-Hoover thermostat was used for AIMD calculations. The distribution of the number of atoms in compositions in the 5000K database and multi-temperature database is added in Supplementary Fig. 5. Most of the structures used in the simulations are described by unit cells of around 100 atoms. Periodic boundary conditions are considered along all three directions.

We employ the NVT method, fitting the equation of state (EOS) over NPT due to several advantages. NVT simulations are more computationally efficient and stable, especially at high temperatures, where NPT's volume fluctuations can introduce instabilities. The NVT thermostat provides better temperature control, directly regulating particle kinetic energy. Additionally, NVT ensures consistency across simulations, facilitating easier comparisons with other studies. Finally, NVT achieves faster equilibration, avoiding the extra time needed to stabilize volume fluctuations in NPT simulations. These factors make NVT the preferred method for our work. The same procedures were used to generate the structure for both 5000K and multi-temperature databases.

The non-crystalline diffusivity database is made accessible to the public via the Material Project's MPContribs[42] website https://contribs.materialsproject.org/projects/amorphous_diffusivity and advanced application programming interface (API) with a dedicated Python client[43]. Ten structures are sampled every 2 ps for DFT relaxation and static calculation, with their corresponding formation energy serving as the amorphous limit to predict synthesizability, following the method proposed by ref. 37.

Machine-learning potential workflow

The M3GNet model developed by Chen et al.[35] offers an alternative surrogate model for AIMD computations, enabling the generation of non-crystalline structures and computation of Li diffusivities across various compositions. This surrogate model has been incorporated into the MPMorph workflow, serving as the calculator for energy and force. The implemented version can be accessed at https://github.com/Tinaatucsd/mpmorph/blob/chgnet_fm_refactor-pv-extract/src/mpmorph/flows/md_flow.py.

Random forest and XGBoost models

For model development, we utilized random sampling to divide the multi-temperature dataset into training and test sets. The training data constitutes 85% of the total data, with the remaining 15% reserved for testing. A fixed random state of 62 was used for consistency. To evaluate model performance, we experimented with different numbers of estimators (trees) for both the Random Forest and XGBoost models, assessing error reduction as more trees were added. For XGBoost, we used the `eval_set` parameter in XGBRegressor[32] to track training and testing errors during the training process. The corresponding plots are presented in Supplementary Fig. 12. Based on the loss curve in Supplementary Fig. 12, we selected `n_estimators=100` for both the Random Forest and XGBoost models. For hyperparameter optimization of the XGBoost model, a grid search was conducted over a predefined set of hyperparameters. The parameter grid included variations in the number of estimators (50, 100, 150), learning rate (0.01, 0.1, 0.5), and maximum tree depth (2, 4, 6). The optimal hyperparameters were determined to be a learning rate of 0.1, a maximum depth of 4, and `n_estimators=150`. Despite this, the loss curve indicated that the model had converged at `n_estimators=100`. Consequently, we selected `n_estimators=100` for the final XGBoost model used in this study. The specific settings for XGBoost models are `n_estimators=100`, `max_depth=4`, `n_jobs=6`, and cross-validation score = negative root mean squared error. Default values are used for all other hyperparameters of the XGBoost model and the Random Forest (RF) model implemented in the scikit-learn package.

SISSO model

A number of SISSO models[34] with increasing complexities were trained, and their prediction performances are shown in Fig. 5. Some of the key input settings for SISSO training include the dimension of the descriptor `desc_dim=4`; the Number of scalar features is 43; The parameters used to control the feature complexity `fcomplexity=7`; The metric root mean square error (RMSE) is used for model selection; The operator set considered include addition (+), subtraction (−), multiplication (*), division (/), exponentiation such as (^{2}), (^{3}), (^{6}),($^{-1}$),(exp),(exp-),(log).

Data availability

The non-crystalline structures and diffusivity are made accessible to the public via the Material Project's MPContribs[42] website https://contribs.materialsproject.org/projects/amorphous_diffusivity and advanced application programming interface (API) with a dedicated Python client[43]. The zipped `json` files are also available at Figshare https://figshare.com/s/30601968f9244d8dffaa.

Code availability

The code used to generate a non-crystalline structure and run workflow is accessible from https://github.com/materialsproject/mpmorph. The code used to analyze the non-crystalline structure dataset and train machine learning models is available from https://github.com/Tinaatucsd/DFT_amorphous_structure.

Received: 27 January 2024; Accepted: 6 November 2024;
Published online: 19 December 2024

References

1. Debenedetti, P. G. & Stillinger, F. H. Supercooled liquids and the glass transition. *Nature* **410**, 259–267 (2001).
2. Hu, Z.-Q., Wang, A.-M. & Zhang, H.-F. in *Modern Inorganic Synthetic Chemistry* (eds Xu, R. & Xu. Y.) Ch. 22 (Elsevier, 2017).
3. Najgebauer, M. Advances in contemporary soft magnetic materials - a review. In *2023 10th International Conference on Electrical, Electronic and Computing Engineering* (*IcETRAN*) 1–10 (IEEE, 2023).
4. Khan, M. M. et al. Recent advancements in bulk metallic glasses and their applications: a review. *Crit. Rev. Solid State Mater. Sci.* **43**, 233–268 (2018).
5. Gomez, S., Dejneka, M. J., Ellison, A. J. & Rossington, K. R. A in *Ceramic Engineering and Science Proceedings* (ed. Drummond, C. H.) (Wiley, 2011).
6. Hong, S. et al. Ultralow-dielectric-constant amorphous boron nitride. *Nature* **582**, 511–514 (2020).
7. Khot, A. C. et al. Amorphous boron nitride memristive device for high-density memory and neuromorphic computing applications. *ACS Appl. Mater. Interfaces* **14**, 10546–10557 (2022).
8. Yoon, I., Larson, J. M. & Kostecki, R. The effect of the SEI layer mechanical deformation on the passivity of a Si anode in organic carbonate electrolytes. *ACS Nano* **17**, 6943–6954 (2023).
9. Majeed, M. K. et al. Silicon-based anode materials for lithium batteries: recent progress, new trends, and future perspectives. *Crit. Rev. Solid State Mater. Sci.* **49**, 1–33 (2023).
10. Zhang, C. et al. Challenges and recent progress on silicon-based anode materials for next-generation lithium-ion batteries. *Small Struct.* **2**, 2100009 (2021).
11. Zhang, Y. et al. Silicon anodes with improved calendar life enabled by multivalent additives. *Adv. Energy Mater.* **11**, 2101820 (2021).

12. Hasa, I. et al. Electrochemical reactivity and passivation of silicon thin-film electrodes in organic carbonate electrolytes. *ACS Appl. Mater. Interfaces* **12**, 40879–40890 (2020).
13. Beaulieu, L. Y., Hatchard, T. D., Bonakdarpour, A., Fleischauer, M. D. & Dahn, J. R. Reaction of Li with alloy thin films studied by in situ AFM. *J. Electrochem. Soc.* **150**, A1457 (2003).
14. McDowell, M. T., Lee, S. W., Nix, W. D. & Cui, Y. 25th anniversary article: understanding the lithiation of silicon and other alloying anodes for lithium-ion batteries. *Adv. Mater.* **25**, 4966–4985 (2013).
15. Cheng, J., Sivonxay, E. & Persson, K. A. Evaluation of amorphous oxide coatings for high-voltage Li-ion battery applications using a first-principles framework. *ACS Appl. Mater. Interfaces* **12**, 35748–35756 (2020).
16. Cheng, J., Fong, K. D. & Persson, K. A. Materials design principles of amorphous cathode coatings for lithium-ion battery applications. *J. Mater. Chem. A* **10**, 22245–22256 (2022).
17. Sivonxay, E., Aykol, M. & Persson, K. A. The lithiation process and Li diffusion in amorphous SiO 2 and Si from first-principles. *Electrochim. Acta* **331**, 135344 (2020).
18. Sivonxay, E. & Persson, K. A. Density functional theory assessment of the lithiation thermodynamics and phase evolution in si-based amorphous binary alloys. *Energy Storage Mater.* **53**, 42–50 (2022).
19. Han, F. et al. High electronic conductivity as the origin of lithium dendrite formation within solid electrolytes. *Nat. Energy* **4**, 187–196 (2019).
20. Grady, Z. A., Wilkinson, C. J., Randall, C. A. & Mauro, J. C. Emerging role of non-crystalline electrolytes in solid-state battery research. *Front. Energy Res.* **8**, 218 (2020).
21. Sastre, J. et al. Blocking lithium dendrite growth in solid-state batteries with an ultrathin amorphous Li-La-Zr-O solid electrolyte. *Commun. Mater.* **2**, 76 (2021).
22. Lacivita, V., Artrith, N. & Ceder, G. Structural and compositional factors that control the Li-ion conductivity in LiPON electrolytes. *Chem. Mater.* **30**, 7077–7090 (2018).
23. Glass, S. Sci glass - the scientific glass database https://sciglass.uni-jena.de/ (2024).
24. Interglad. Interglad - database of optical glass https://www.interglad.jp/interglad6/ (2024).
25. Zaki, M. et al. Natural language processing-guided meta-analysis and structure factor database extraction from glass literature. *J. Non-Cryst. Solids* **15**, 100103 (2022).
26. Thyagarajan, R. & Sholl, D. S. A database of porous rigid amorphous materials. *Chem. Mater.* **32**, 8020–8033 (2020).
27. Zakarian, A. Bond dissociation energy. https://labs.chem.ucsb.edu/zakarian/armen/11---bonddissociationenergy.pdf (2011).
28. Xiao, Y. et al. Lithium oxide superionic conductors inspired by garnet and NASICON structures. *Adv. Energy Mater.* **11**, 2101437 (2021).
29. Hargreaves, C. J. et al. A database of experimentally measured lithium solid electrolyte conductivities evaluated with machine learning. *npj Comput. Mater.* **9**, 9 (2023).
30. 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).
31. Ho, T. K. Random decision forests. In *Proc. 3rd International Conference on Document Analysis and Recognition* 278–282 (IEEE, 1995).
32. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, 785–794 (ACM, 2016).
33. Buitinck, L. et al. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning* 108–122 (2013).
34. Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M. & Ghiringhelli, L. M. SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.* **2**, 083802 (2018).
35. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
36. Deng, B. et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **5**, 1031–1041 (2023).
37. Aykol, M., Dwaraknath, S. S., Sun, W. & Persson, K. A. Thermodynamic limit for synthesis of metastable inorganic materials. *Sci. Adv.* **4**, eaaq0148 (2018).
38. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL : a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164 (2009).
39. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
40. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
41. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
42. Huck, P. et al. User applications driven by the community contribution framework MPContribs in the Materials Project. *Concurr. Comput.* **28**, 1982–1993 (2016).
43. Huck, P. mpcontribs-client https://pypi.org/project/mpcontribs-client/ (2024).

Acknowledgements

This research was intellectually led by the Materials Project program (Contract No. DE-AC02-05-CH11231, KC23MP), supported by the US Department of Energy, Office of Basic Energy Sciences. This study utilized the facilities of the National Energy Research Scientific Computing Center (NERSC), a User Facility of the U.S. Department of Energy Office of Science located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05-CH11231.

Author contributions

H.Z.: Data acquisition and analysis, methodology, software, writing, editing, and reviewing; E.S.: Data acquisition, software, review and editing, and conceptualization; R.C.: Writing, review, and editing, retrieval of experimental data, and validation; M.G.: Software, analysis, writing, and editing; Z.L.: Software, data, writing, and editing; M.M.: Software, review, and editing; P.H.: Software and review; M.M.S.: Editing and supervision; K.A.P.: Conceptualization, writing-review and editing, supervision, project administration, and funding acquisition.

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/s41524-024-01469-2.

Correspondence and requests for materials should be addressed to Kristin A. Persson.