

ARTICLE OPEN



AlphaMat: a material informatics hub connecting data, features, models and applications

Zhilong Wang^{1,2,3}, An Chen^{1,2,3}, Kehao Tao^{1,2,3}, Junfei Cai^{1,2}, Yanqiang Han^{1,2}, Jing Gao^{1,2}, Simin Ye^{1,2}, Shiwei Wang^{1,2}, Imran Ali^{1,2} and Jinjin Li^{1,2}✉

The development of modern civil industry, energy and information technology is inseparable from the rapid explorations of new materials. However, only a small fraction of materials being experimentally/computationally studied in a vast chemical space. Artificial intelligence (AI) is promising to address this gap, but faces many challenges, such as data scarcity and inaccurate material descriptors. Here, we develop an AI platform, AlphaMat, that can complete data preprocessing and downstream AI models. With high efficiency and accuracy, AlphaMat exhibits strong powers to model typical 12 material attributes (formation energy, band gap, ionic conductivity, magnetism, bulk modulus, etc.). AlphaMat's capabilities are further demonstrated to discover thousands of new materials for use in specific domains. AlphaMat does not require users to have strong programming experience, and its effective use will facilitate the development of materials informatics, which is of great significance for the implementation of AI for Science (AI4S).

npj Computational Materials (2023)9:130; <https://doi.org/10.1038/s41524-023-01086-5>

INTRODUCTION

Material science has developed rapidly in the twenty-first century, both theoretically and experimentally, such as the development of gas conversion catalytic materials, the discovery of energy harvesting and storage materials, the design of information functional materials, etc.^{1–3}. As an interdisciplinary subject of material science and computer science, computational material science is increasingly powerful due to the significant improvement of computing devices, and has become a bridge between theoretical prediction and experimental research^{3–5}. Computational material science not only frees theoretical work from the bondage of analytical derivation, but also carries on the fundamental reform to the experimental research methods, which is more conducive to researchers to reveal and confirm objective laws from experimental phenomena. Currently, the modern material-simulation toolkits (e.g., Vienna Ab Initio Simulation Package (VASP)⁶, Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)⁷, Quantum Espresso⁸, crystal structure analysis by particle swarm optimization (CALYPSO)^{9,10}, nonadiabatic molecular dynamics (Hefei-NAMD)¹¹, and defect and dopant ab-initio simulation package (DASP)¹², and user-friendly VASP-KIT¹³) have brought computational material science to the masses in form of useful practical tools, enabling experimentalists with little or no theoretical training to perform first-principles calculations (e.g., density functional theory (DFT) calculations^{14,15}). Consequently, high-throughput calculation (HTC) becomes a routine approach, and accelerates the development of databases with materials (organic and inorganic crystals, single molecules, and metal alloys) and properties (band gaps, formation energies, ionic conductivities, and elastic modulus^{16–18}). The Materials Genome Initiative (MGI) proposed in 2011 pushed computational material science into high gear^{19,20}, and many material databases and platforms sprung out, such as the Materials Project (MP)¹⁶, the Open Quantum Materials database (OQMD)²¹, the Novel Materials

Discovery (NOMAD)²², and various proprietary databases from the literatures. Afterwards, six application-focused areas were identified as important (health and consumer materials, information technology materials, etc.), and a new route was planned for the development of new materials. Further investment in MGI principles can generate extraordinary advances that can spark revolutionary new technologies and provide important opportunities for the next generation of advanced materials with transformative impact²³.

The establishment and sharing of these databases offer an opportunity for the emergence of the “fourth paradigm of science” and the “fourth industrial revolution”, i.e. the “data-driven material discovery”²⁴, the critical idea of which is the combination of big data, artificial intelligence (AI), and material science^{25–28}. The number of AI applications in material science is growing at an alarming rate, with notable success in many systems, such as batteries^{29–31}, solar cells^{32–34}, ecomaterials^{35,36}. Just like the implementation of quantum mechanical (QM) computing softwares, it is necessary to develop infrastructures that combine material science and AI in order to enable both AI researchers and material experts to design materials using AI methods (machine learning (ML), etc.). Several pioneering efforts have been launched in recent years to achieve this goal^{37,38}. Ward et al. developed a material data mining toolkit (Matminer), which offered one-stop access to multiple data sets and provided feature descriptors of components and structures for property prediction. This toolkit has become an important foundation for the joint use of AI and material data³⁹. However, Matminer does not contain AI routines itself, but instead processes data format in order to make various downstream AI libraries available for material science applications. The subsequent Automatminer pipeline can performs many AI steps (feature engineering, model selection, hyperparameter tuning, etc.), allowing the combined application of AutoML and Matminer to implement end-to-end material

¹National Key Laboratory of Science and Technology on Micro/Nano Fabrication, Shanghai Jiao Tong University, Shanghai 200240, China. ²Key Laboratory of Thin Film and Microfabrication of Ministry of Education, Department of Micro/Nano Electronics, Shanghai Jiao Tong University, Shanghai 200240, China. ³These authors contributed equally: Zhilong Wang, An Chen, Kehao Tao. ✉email: lijinjin@sjtu.edu.cn

modeling pipelines⁴⁰. Also, the Materials Simulation Toolkit for Machine Learning (MAST-ML) was proposed to broaden and accelerate the use of ML in material science, which lowered the barrier to entry for supervised learning (SL) modeling⁴¹. NOMAD AI toolkit, a web-browser-based platform for performing AI analysis of materials-science data was presented by Sbailò et al., which will bring the concept of reproducibility in material science to the next level⁴². Whereas, using Matminer/AutoMatminer requires the basic of programming, such as Python, which is unfriendly to material designers with little programming experience, and other ML modeling (transfer learning (TL), unsupervised learning (UL), etc) need to be integrated to MAST-ML. The list of more toolkits was provided by Morgan and Jacobs⁴³. Generally speaking, existing material informatic tools can still be improved further. It is necessary to establish a material informatic platform that supports all commonly used AI algorithms, requires no or minimal programming skill, and contains material databases. In addition, the lack of data of material properties, as well as inaccurate material descriptors, have become challenges for materials modeling.

Here, we developed an AI platform, AlphaMat, that supports the whole life cycle of material modeling with over 90 functions (data collection → data preprocessing → feature engineering → model establishment → parameter optimization → model evaluation → result analysis). AlphaMat has a higher applicability in material modeling, benefiting from component and structural descriptors. AlphaMat is the first material informatic platform that possesses SL, TL, and UL simultaneously, and can tackle the tasks of material modeling without the limitation of data scale. In addition, AlphaMat has an interactive interface, runs locally, requires no programming experience. As typical cases, we collected 12 material property databases from experiments and HTC calculations, including formation energy, metal/semiconductor, phonon property, dielectric constant, ionic conductivity, thermal conductivity, optical property, magnetism, ferroelectric property, band gap, bulk modulus, and adsorption energy (covering a total of 19,488 materials). And then AI models were established, which can be used to enhance photoelectric conversion efficiency, improve

conductivity of metallic electrode materials, promote cycle performance of batteries, discover new solid-state electrolyte, inhibit the shuttle effect of Li-S batteries, develop high thermal conductivity materials, solve the heat dissipation of electronic devices, etc. Compared with the time cost of experiments or calculations used to construct the database, AlphaMat saved significant time cost and hardware cost in material discovery. The practical application in energy science demonstrates AlphaMat's ability to discover and design materials that it successfully identify 491 potential photovoltaic materials, 78 metallic electrode materials, 9 solid-state electrolytes, 58 thermal-conductivity materials, and 39 cathodes of Li-S batteries. By AlphaMat, users can directly search the database according to various needs; AI models can also be easily built on any data scale to discover and design materials. Following the principles of interaction, scalability, efficiency and intelligence, AlphaMat, together with many other toolkits built by the larger material community are expected to promote and accelerate the development of material science, computer science, physical and chemical science.

RESULTS

Overview and architecture

Considering the current challenges and requirements of material modeling, AlphaMat was developed with nine core elements (Fig. 1): (1) Proprietary databases. AlphaMat aims to build database of material properties from experiments, calculations, literatures, and open databases (e.g., databases of formation energy or band gap). (2) Data processing and analysis. The establishment of material data requires the unification of data format, the conversion of file format and the statistics of material properties. (3) Material descriptor design. AlphaMat can calculate suitable digitization vectors or matrices to represent materials, including component and structural descriptors. (4) Quantitative structure-property relationships (QSPR). Establishment of material-property QSPR through AI models is the most important goal and pursuit of AlphaMat. (5) New materials. Based on the well-trained QSPR, new

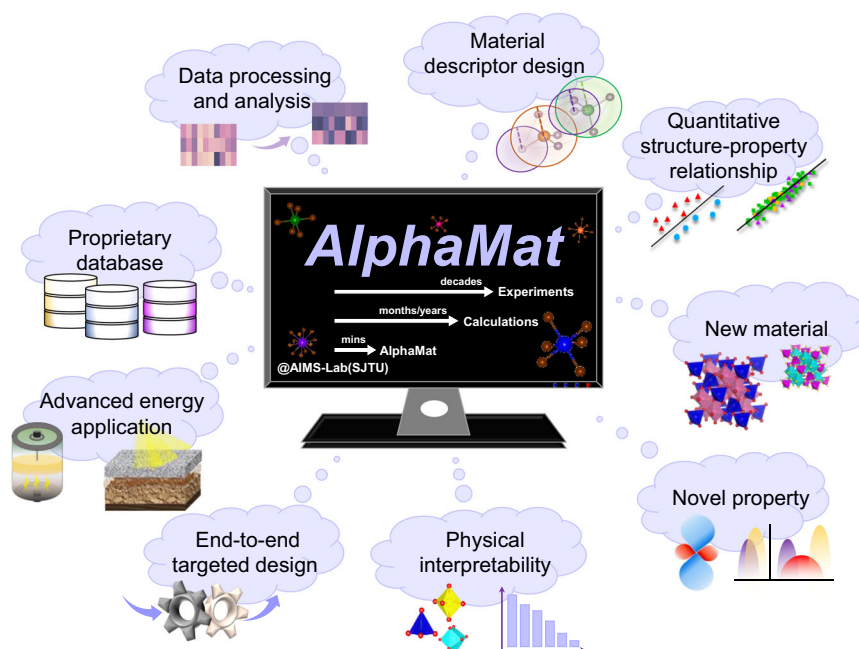


Fig. 1 Overview of AlphaMat. Focusing on nine elements of material informatics (proprietary databases, data processing and analysis, material descriptor design, QSPR, new materials, novel properties, physical interpretability, end-to-end targeted design, and advanced applications), AlphaMat aims to accelerate the development of materials, shorten the development cycle of materials, and reduce the cost of experiments and traditional computations.

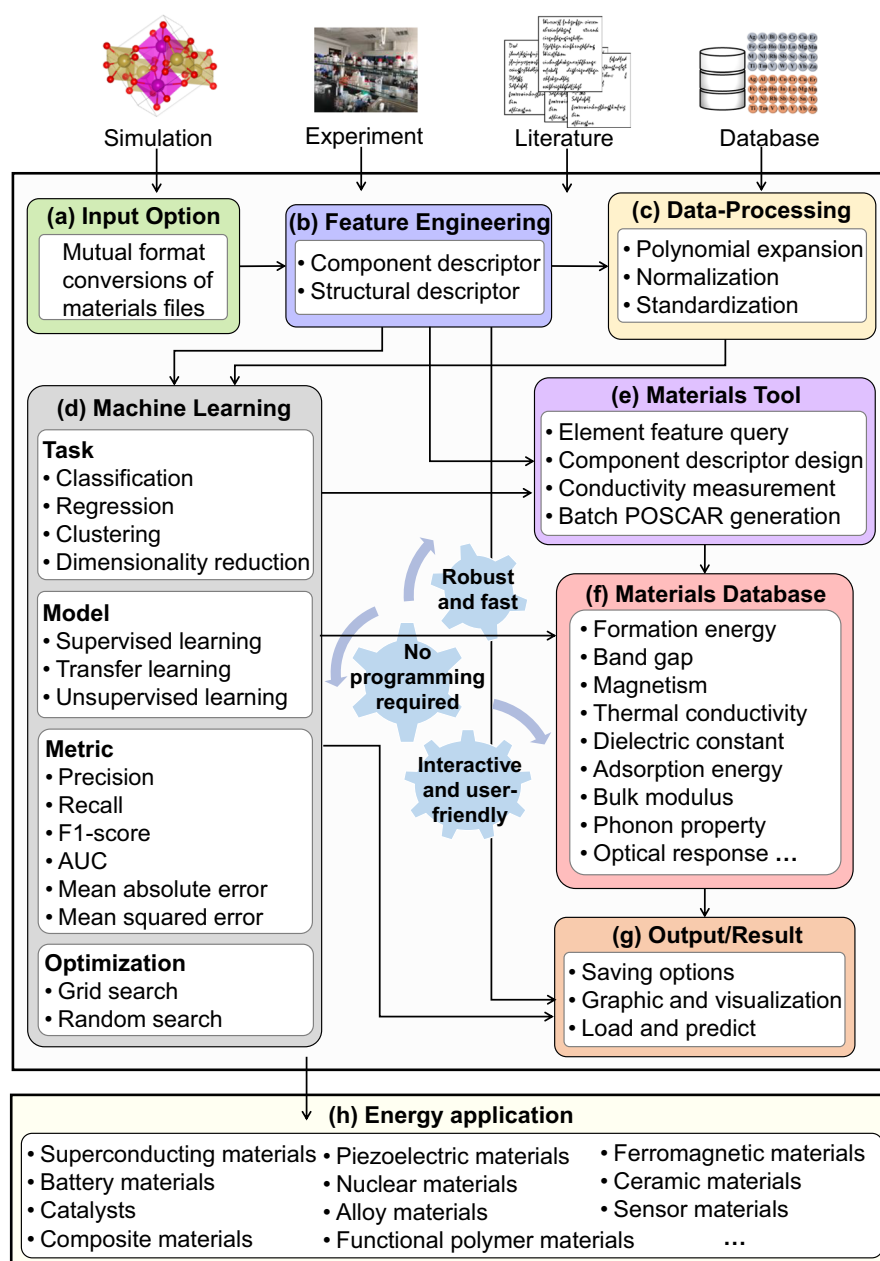


Fig. 2 Architecture of AlphaMat. **a** Input options, used to read and convert various material structure files. **b** Feature engineering, where component and structural descriptors are provided for materials representation. **c** Data-processing, which can preprocess the obtained features. **d** Machine learning, which covers almost all current AI modeling requirements. **e** Materials tool, which integrates a variety of convenient material data processing scripts to improve efficiency. **f** Materials database, also one of the main tasks of the software. It builds proprietary databases based on different material properties. **g** Output/result, which can be further analyzed with other various visualization tools (visual module is under development). **h** Application, the research areas/material systems to which the whole software can be applied.

materials with suitable properties can be explored and identified. (6) Novel properties of materials that have not been reported/studied before. (7) Physical interpretability to uncover the feature importance from AI models for the material design, which is the challenge and pursuit of material informatics. (8) End-to-end targeted design, which is closely related to physical interpretability and establish a pattern of input-to-output automation that facilitates practical applications. (9) Advanced applications. The ultimate goal of AlphaMat, is to promote the progress of various material systems (e.g., superconducting materials, battery

materials, piezoelectric materials) by discovering high-performance materials for applications.

The organization of AlphaMat abided by the data roadmap in the research field of material informatics, from data collection, data preprocessing, ML, and application, as shown in Fig. 2. More modeling process can be found in Supplementary Note 1. In AlphaMat (v0.0.7), over 90 functions have been designed, and some useful tools were used (e.g., Matminer³⁹, Python Materials Genomics (Pymatgen)⁴⁴, Scikit-Learn⁴⁵, extreme gradient boosting decision tree (XGBoost)⁴⁶, and Mendelev⁴⁷). Researchers can use AlphaMat to complete the entire process of AI and material

Table 1. Twelve case studies.

Material property	Task	Data scale	AlphaMat performance	Ref
E_f (eV/atom)	R ₁	3483 Cal. points	$P_c = 0.877$, MAE = 0.221, RMSE = 0.341, RMSE/S.D. = 0.482	16
E_g (eV)	R ₂	3895 Exp points	$P_c = 0.933$, MAE = 0.347, RMSE = 0.522, RMSE/S.D. = 0.360	68
B_{ASR} (cm ⁻¹)	R ₃	1515 Cal. points	$P_c = 0.878$, MAE = 2.278, RMSE = 4.381, RMSE/S.D. = 0.482	69
ϵ_{poly}	R ₄	1028 Cal. points	$P_c = 0.675$, MAE = 3.265, RMSE = 4.745, RMSE/S.D. = 0.765	70
K (GPa)	R ₅	373 Cal. points	$P_c = 0.895$, MAE = 15.853, RMSE = 21.248, RMSE/S.D. = 0.433	16
E_a (eV)	R ₆	1109 Cal. points	$P_c = 0.803$, MAE = 0.402, RMSE = 0.604, RMSE/S.D. = 0.596	71
κ (W m ⁻¹ K ⁻¹)	R ₇	128 Exp. points	$P_c = 0.829$, MAE = 0.293, RMSE = 0.465, RMSE/S.D. = 0.578	72
SHG responses (pm V ⁻¹)	R ₈	291 Cal. points	$P_c = 0.854$, MAE = 0.963, RMSE = 1.524, RMSE/S.D. = 0.550	73
Metal/Semiconductor	C ₁	6353 Exp. points	Precision = 0.93, recall = 0.93, F1-score = 0.93, AUC = 0.93	68
Ferro/Non-ferro	C ₂	1028 Cal. points	Precision = 0.87, recall = 0.87, F1-score = 0.87, AUC = 0.87	70
Strong/Weak ΔE	C ₃	65 Cal. points	Precision = 0.85, recall = 0.85, F1-score = 0.85, AUC = 0.85	66
FM/AFM	C ₄	220 Cal. points	Precision = 0.82, recall = 0.80, F1-score = 0.80, AUC = 0.80	18

The material features as ML inputs were calculated using component descriptors, and the classification models and regression models were implemented by training XGBoost models. The performance on testing data was selected to evaluate the AlphaMat performance. It is noted that these case studies are meant to provide macro guidance for the users to facilitate a complete employment of the available materials database.

R: Regression tasks.

C: Classification tasks.

Cal: Calculated data.

Exp: Experimental data.

P_c : Pearson correlation coefficient.

MAE: Mean absolute error.

RMSE: Root mean squared error.

S.D.: Standard deviation of the real data in testing set.

AUC: Area under receiver operator characteristic (ROC) curve.

modeling. The introduction of material descriptors, AI models, and analysis tools are provided in Supplementary Note 2–4.

Modeling cases

AlphaMat provides a complete process of the data collection → data preprocessing → feature engineering → model establishment → parameter optimization → model evaluation → result analysis. Therefore, AlphaMat will play a great role in calculating material descriptors, establishing QSPR, and material screening and mining.

Here, as case studies, we used AlphaMat to predict 12 typical material properties (containing eight regression tasks and four classification tasks, see Table 1) with 19,488 data points totally, and highlighted the advantages of AlphaMat in these works. The twelve material properties are formation energy (E_f), band gap (E_g), the maximum frequency of an acoustic mode at Γ (breaking of the ASR, B_{ASR}), dielectric constant (ϵ_{poly}), bulk modulus (K), ion migration activation energy (E_a), thermal conductivity (κ), second harmonic generation (SHG) responses, metals/semiconductors, ferroelectric/non-ferroelectric materials (Ferro/Non-ferro), strong/weak adsorption energy (ΔE), and ferromagnetic/antiferromagnetic materials (FM/AFM). It is worth noting that we chose the component descriptor of element property (a 120-element vector) as the material descriptor, which was defined by Meredig et al. and integrated in AlphaMat (instruction of 805)⁴⁸. XGBoost model was applied for model training, which was widely used in material science^{46,49,50}. The descriptions of XGBoost are shown in 12104 (for classification tasks) and 12204 (for regression tasks) (see <http://www.aims-lab.cn>). The data was split into training set (80%) and testing set (20%). As shown in Table 1 and Supplementary Figs. 1–12, the Pearson correlation coefficients (P_c) of eight regression models are from 0.675 to 0.933, with an average of 0.843, and the precision of four classification models are from 0.82 to 0.93, with an average of 0.868. More details and applications are provided in Supplementary Note 5–17. These typical case studies demonstrate the strong modeling ability of AlphaMat in material property predictions and material discovery.

The 19,488 data points currently used for modeling are just the tip of the iceberg in the vast material space, and there are hundreds of millions of materials and properties to be explored. For example, in the MP database¹⁶, the E_f and K of 144,595 materials data can be modeled by AlphaMat, saving significant computational cycles. In addition, the E_g prediction model established by AlphaMat can predict the E_g of ~68,000 materials (with unknown E_g in the MP database¹⁶) at the experimental level (each takes 64 h⁵¹), which will save a lot of experimental cycle. It can be seen that, based on the existing experimental/computational data, modeling based on AlphaMat can greatly shorten the experimental/computational cycle for new material discovery. We can foresee that AlphaMat will be an important part in existing material informatics software, accelerating the deployment of material engineering.

Practical applications in high-performance materials

Twelve case studies demonstrate AlphaMat's capabilities in material modeling. Here, with different material property modeling, we presented several practical examples about electrode materials, photoelectric materials, solid-state electrolyte materials, and thermal-conductivity materials, etc.

Practical applications based on E_g . E_g is a key characteristic of electronic materials. For example, in perovskite solar cells, the hole transport layer (HTL) and the electron transport layer (ETL) should have appropriate E_g (0.9–1.6 eV) to ensure the efficient transmission of holes and electrons and the implement of optimal optical conversion efficiency^{52,53}. Electrode materials generally have high electronic conductivity, i.e., $E_g = 0$ ^{54,55}, while solid-state electrolytes require extremely low electronic conductivity, i.e., $E_g > 3.5$ eV^{51,56,57}. Thus, accurately determining E_g is the key to select functional materials and accelerate their development.

MP database contains 144,595 data entries¹⁶, among which the studies of mono-element compounds are quite mature, while the laboratory synthesis of multi-element compounds is challenging.

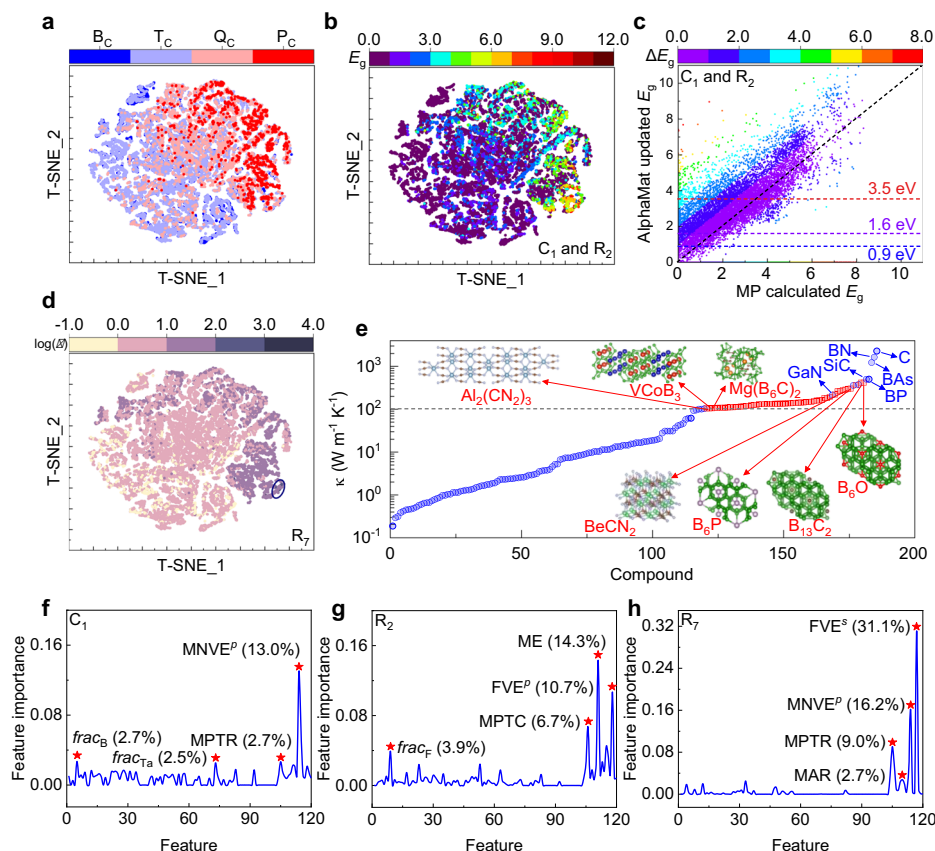


Fig. 3 Material discovery through C_1 , R_2 , and R_7 models established by AlphaMat. **a** t-SNE plot, where the sites are colored with their compound type. **b** t-SNE plot, where the sites are colored with their E_g . **c** Correlation of predicted E_g and calculated E_g , the color bar denotes the E_g changes. **d** t-SNE plot, where the sites are colored with their thermal conductivity, the oval mark determines the compounds with high thermal conductivities. **e** Discovered compounds with high thermal conductivities (red squares) and the known compounds (blue circles). **f** Embedded feature importance of C_1 . **g** Embedded feature importance of R_2 . **h** Embedded feature importance of R_7 .

Therefore, binary (B_C), ternary (T_C), quaternary (Q_C), and pentabasic (P_C) compounds were selected from MP to establish the initial data set. In addition, thermal stability is the most basic property of materials, so we excluded materials with convex hull energy (E_{hull}) greater than 0, leaving 32,858 materials in the end (5039 B_C , 19,257 T_C , 7287 Q_C , 1275 P_C). Among the 12 case studies in Table 1, C_1 can distinguish metals ($E_g = 0$) and semiconductors ($E_g > 0$), R_2 can predict the E_g for semiconductors. By using element property as the material descriptor (805 in AlphaMat), we made use of well-trained C_1 and R_2 models for searching new materials.

As shown in Fig. 3a, using t-distributed stochastic neighbor embedding (t-SNE) method, the sites are colored with their compound types, and compounds with different number of element types can be distinguished, as the sites of P_C , Q_C , T_C , and B_C are stacked on top of each other. In MP database, the E_g values of 32,858 compounds were calculated based on semi-empirical or low-precision functional, which deviate greatly from the experimental values (the deviation is 1.0–2.0 eV generally) and are difficult to be directly used in the actual screening of materials^{18,58}. By using band gap-based models C_1 and R_2 , we can rapidly predict (or update) the E_g of 32,858 compounds. Our well-trained C_1 has a prediction accuracy of 93% for identifying metals and semiconductors, and the P_C between the E_g predicted by R_2 and the experimental value is 0.933, and the MAE is only 0.347 eV (see Table 1). Therefore, the two models are of great significance to update and reuse materials in MP database. As shown in Fig. 3b, the sites are colored with their E_g values predicted by C_1 and R_2 , where the sites with large values (> 3.0 eV) are mainly concentrated on the right side of the t-SNE plot. This

phenomenon can be associated with Fig. 3a, as the types of compound elements increase, the new elements introduced are mainly non-metallic elements, such as O, S, F, Cl, Br, etc., leading to the weakening of the electronic conductivity of the material.

Figure 3c shows the correlation of MP calculated E_g and predicted E_g . It can be seen that the E_g changes of most materials are less than 2.0 eV (blue and purple dots), and the updated E_g is general larger (green, yellow, and red dots), which is consistent with the conclusion that the E_g calculated by low-precision functional is seriously underestimated^{18,59,60}. Then, we identified 832 materials with E_g of 0.9–1.6 eV for photoelectric materials (HTLs, ETLs, photocatalysts, etc.), 13 materials containing Li^+ with $E_g > 3.5$ eV for solid-state electrolytes. In addition, for searching the electrode materials, excellent electronic conductivity with $E_g = 0$ is necessary, as well as high mechanical properties. Referring to the shear modulus and bulk modulus of commercialized materials $\text{LiNi}_{0.3}\text{Mn}_{0.3}\text{Co}_{0.3}\text{O}_2$ (NMC333), $\text{LiNi}_{0.4}\text{Mn}_{0.4}\text{Co}_{0.2}\text{O}_2$ (NMC442), $\text{LiNi}_{0.5}\text{Mn}_{0.3}\text{Co}_{0.2}\text{O}_2$ (NMC532), $\text{LiNi}_{0.6}\text{Mn}_{0.2}\text{Co}_{0.2}\text{O}_2$ (NMC622), $\text{LiNi}_{0.8}\text{Mn}_{0.1}\text{Co}_{0.1}\text{O}_2$ (NMC811)^{61,62}, we further selected 95 materials with shear modulus > 67 GPa and bulk modulus > 85 GPa as candidates for electrode materials. Moreover, from economic and environmental considerations, some materials containing rare precious metal elements or radioactive elements were excluded, resulting in 491, 9, and 78 materials, respectively (see Supplementary Tables 1–3).

Practical applications based on κ . κ is an important thermal property of electronic materials and devices. Materials with high κ (e.g., C, 2235 $\text{W m}^{-1} \text{K}^{-1}$; BN, 1600 $\text{W m}^{-1} \text{K}^{-1}$) can be used to

solve the heat dissipation problem of electronic products, and the development of new thermal conductivity materials will provide strong support for future space exploration activities and ocean exploration activities^{63,64}. Among the 12 case studies in Table 1, R_7 can predict the κ of given materials. By using element property as the material descriptor, we made use of well-trained R_7 models for searching materials. As shown in Fig. 3d, the t-SNE plot shows that most materials have very small κ ($< 100 \text{ W m}^{-1} \text{ K}^{-1}$), and materials with $\kappa > 100 \text{ W m}^{-1} \text{ K}^{-1}$ are very concentrated (see the oval mark). Figure 3e shows the discovered materials (red squares) with high κ , such as B_6O ($408.7 \text{ W m}^{-1} \text{ K}^{-1}$), B_{13}C_2 ($407.7 \text{ W m}^{-1} \text{ K}^{-1}$), B_6P ($355.0 \text{ W m}^{-1} \text{ K}^{-1}$), and BeCN_2 ($296.0 \text{ W m}^{-1} \text{ K}^{-1}$). The new thermal-conductivity materials can be comparable to the famous GaN ($210.0 \text{ W m}^{-1} \text{ K}^{-1}$), which have a broad prospect in the application of optoelectronics, high temperature high power devices and high frequency microwave devices (see Supplementary Table 4).

The predicted E_g and κ of 32,858 materials at experimental level may be of wide interest to the experimental community in multiple areas of research (batteries, catalysis, electronics, etc.). In addition to establish the high-precision QSPR, AlphaMat also provides the interpretability of the model, which is a unique feature. Figure 3f–h shows the embedded feature importance of C_1 , R_2 , and R_7 , respectively. For C_1 , the mean number of valence electrons of p orbitals (MNVE^p , 13%) in compounds and the mean of periodic table rows (MPTR, 2.7%) play a key role in distinguishing metals from semiconductors. This has guiding significance for the design of corresponding materials. The fraction of B (frac_B , 2.7%) and Ta (frac_{Ta} , 2.5%) are also important due to the compounds containing B in the data are mainly semiconductors, while those containing Ta are metals in training data set. For R_2 , the mean electronegativity (ME, 14.3%), the fraction of valence electrons of p orbitals (FVE^p , 10.7%), the mean of periodic table columns (MPTC, 6.7%), and the fraction of F (frac_F , 3.9%) are relatively important for predicting E_g values. For R_7 , the fraction of valence electrons of s orbital (FVE^s , 31.1%) and MNVE^p (16.2%) are particularly important for thermal conductivity prediction, which is consistent with the phenomenon that heat conduction is mainly the diffusion of free electrons from the high end to the low end, resulting in heat flow. These key features are of great significance for further directed design of functional materials⁶⁵.

Practical applications based on ΔE . UL methods are based on unlabeled data, can completely overcome the obstacle of scarce material attributes. However, UL module is still a gap in many existing material informatic platforms. In above case studies, the data scale of ΔE between AB_2 -type 2D materials and Li_2S_6 is few (only 65 entries)⁶⁶, which is not conducive to establish the QSPR. The search for materials with strong adsorption ($|\Delta E| > 1.0 \text{ eV}$) for Li_2S_6 is helpful to discover new cathode materials for lithium-sulfur (Li-S) batteries and inhibit the “shuttle effect”. Here, we demonstrated an UL method for discovering new cathodes for Li-S batteries. Total 826 stable AB_2 -type compounds were selected from the 2DMatPedia database, of which 65 materials have known adsorption energies with Li_2S_6 , and the remaining 761 were unknown⁶⁷. Figure 4a shows the bottom-up tree diagram (dendrogram) by using the agglomerative hierarchical clustering (AHC) algorithm in AlphaMat, where a suitable partition line was selected and the 826 AB_2 -type compounds were classified into seven groups (see Supplementary Fig. 13, from G_1 , G_2 , ..., to G_7). We mapped 65 known ΔE to the dendrogram, and compounds marked by green, orange and red are promising according to ideal thresholds (-1.0 eV)⁶⁶. The clustering of AB_2 -type compounds provides physical insights into understanding of compounds exhibiting proper adsorption energies for Li_2S_6 . Figure 4b gives the statistic of known and unknown compounds each group, G_4 has the most compounds of 319, while G_3 has the fewest

compounds of 29, indicating that a targeted study of these groups would significantly narrow down the initial scope (761 unknown compounds). Figure 4c shows the ratio of known compounds (black line) and the ratio of desired compounds (blue line) of each group. Notably, in G_1 , G_3 , and G_5 , the ratio of desired compounds to known compounds is 100%, which is much higher than that in other groups. This phenomenon can also be observed in Fig. 4a. These suggest that the unknown compounds in G_1 , G_3 , and G_5 are worthy of further investigation (142 compounds in total), and that they may also be potential cathode materials for Li-S batteries. The violin plots of the known ΔE shown in Fig. 4d further reveal that G_5 is of high research value because of its higher average absolute adsorption energy value (1.62 eV). As a result, the scope of exploration narrowed from 761 compounds to 84 compounds in G_5 . Moreover, compounds containing rare precious metal elements or radioactive elements were excluded, resulting in 39 compounds finally, respectively, as provided in Supplementary Figs. 14, 15. More details about the position of the partition line are discussed in Supplementary Note 18.

DISCUSSION

The challenges of material informatics prompt us to develop an advanced computational infrastructure. In this work, we presented an AI platform that supports the whole life cycle of material modeling, including data analysis, feature engineering, model establishment and optimization, evaluation to result analysis. The proposed AlphaMat integrates supervised SL, TL, and UL simultaneously, which can tackle the tasks in material science more comprehensively. Furthermore, AlphaMat establishes proprietary databases with more than 117,000 material-property entries (see <http://www.aims-lab.cn>). Since AlphaMat runs locally, the training of its AI models is not limited by the scale of data sets (from 10^1 to 10^6). Consequently, AlphaMat will accelerate the innovative discovery of new materials, new functions, and new principles, compared to the trial-and-error experiments and high-throughput calculation methods. 12 case studies of material modelings (formation energy, band gap, magnetism, adsorption energy, thermal conductivity, and ionic conductivity, etc) demonstrate the effectiveness and usefulness of AlphaMat, and the practical application in searching high-performance materials demonstrates AlphaMat's ability to mine and design materials that it successfully identify new materials for use in various systems (photonics, batteries, catalysis, and capacitors, etc.) from the large inorganic compound databases. Using AlphaMat, users can either directly retrieve our database or easily build AI models to discover and design materials.

It should be mentioned that ML is only as good as the data it is trained on and predictions using data outside the training set are likely to fail dramatically. Therefore, the prediction results of the ML model will be uncertain to a certain extent. In the face of more complex problems, traditional computational methods or experimental methods are also needed for further verification. But at the very least, ML offers specific candidates to speed up material development. Further, we will continue to improve and release AlphaMat to address the challenges commonly encountered in material modeling: (1) continuously expand the databases according to the material systems (fullerenes, nanocomposites, metamaterials, etc.) and properties (superconductivity, optical coefficient, etc.) to alleviate the challenge of data scarcity and make it available to more scientists in different material subfields, with the using of AI methods (e.g., natural language processing, generative model); (2) integrate more popular component and structural descriptor, and innovate new descriptors to represent the materials, improve model accuracy and make models interpretable; (3) combine frontier AI algorithms timely to cope with more material modeling tasks; (4) add more convenient tools and visualization interface to improve the efficiency for processing

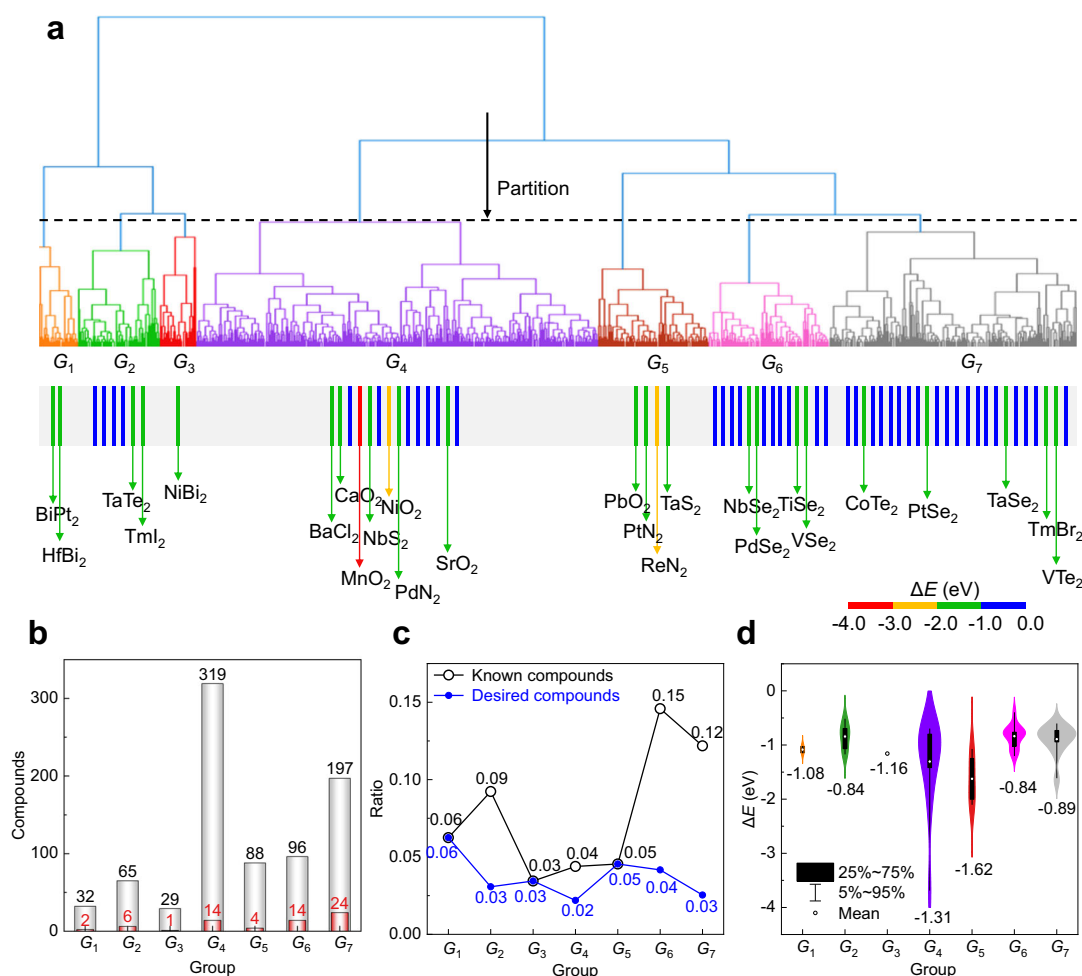


Fig. 4 Unsupervised discovery of AB₂-type compounds as cathodes for Li-S batteries. **a** Dendrogram generated by the AHC method in AlphaMat. The dashed line shows the position where all compounds are partitioned into seven groups, marked as G₁–G₇ from left to right and distinguished by different colors. **b** Statistic of known compounds in each group. Gray bars represent the number of compounds in each group after clustering, and red represents the number of known compounds in each group. **c** Ratio of known compounds (black line) and the ratio of desired compounds (blue line) of each group. **d** Violin plots of ΔE in seven groups. The outer shells of the violins bound all data, narrow horizontal lines bound 90% of the data, thick horizontal lines bound 50% of the data, and white dots represent means.

material data. We hope that the continually released AlphaMat will deeply unite material science and AI approaches, and become an essential tool in science researches.

METHODS

Architecture

Various material data can be generated/collected from simulations, experiments, literatures (manually collect data from published papers), and open databases. For software-generated material structure files containing atomic information, batch conversion tasks between files needed to be completed first, as shown in Fig. 2a. The material descriptors are then constructed based on the component and structural information (Fig. 2b). For the data in plain text format, the data preprocessing module can be directly carried out (Fig. 2c). Four main learning tasks (classification, regression, clustering, and dimensionality reduction), three type models (supervised learning, transfer learning, and unsupervised learning), and different AI models have been designed and integrated in AlphaMat (Fig. 2d). Furthermore, considering the importance of hyper-parameters in AI models,

AlphaMat also provides two commonly used optimization methods to search the optimal hyper-parameters. In addition to the ease of AI modeling, we integrated various portable material tools in AlphaMat (Fig. 2e). Moreover, AlphaMat aims to comprise all kinds of material databases and categorize them according to the material properties (Fig. 2f). Finally, data, features and models can be automatically saved in the current directory for further visual analysis (Fig. 2g). The whole development and use process of AlphaMat closely combines the components, structures, and properties of materials with AI (data, features, and models), which is expected to be widely used in various material systems (superconducting materials, battery materials, alloy materials, etc.; Fig. 2h).

The core elements and architecture can be found in Fig. 1 and Fig. 2. Python was used as the AlphaMat primary back-end programming language to complete each function. More implementation details are provided in Supplementary Information.

DATA AVAILABILITY

More details and tutorials on using AlphaMat are also available from Supplementary Information and (<http://www.aims-lab.cn>). All proprietary databases are available from

out website, and can be easily obtained with a form submission for academic use. More database construction will be incorporated in the future release of AlphaMat, please visit our website: <http://www.aims-lab.cn>.

CODE AVAILABILITY

The codes to run AlphaMat are available from our website: <http://www.aims-lab.cn>.

Received: 27 March 2023; Accepted: 17 July 2023;

Published online: 26 July 2023

REFERENCES

- Daehn, K. et al. Innovations to decarbonize materials industries. *Nat. Rev. Mater.* **7**, 275–294 (2021).
- Marzari, N., Ferretti, A. & Wolverton, C. Electronic-structure methods for materials design. *Nat. Mater.* **20**, 736–749 (2021).
- Louie, S. G., Chan, Y.-H., da Jornada, F. H., Li, Z. & Qiu, D. Y. Discovering and understanding materials through computation. *Nat. Mater.* **20**, 728–735 (2021).
- Correa-Baena, J.-P. et al. Accelerating materials development via automation, machine learning, and high-performance computing. *Joule* **2**, 1410–1420 (2018).
- Hammes-Schiffer, S. & Galli, G. Integration of theory and experiment in the modelling of heterogeneous electrocatalysis. *Nat. Energy* **6**, 700–705 (2021).
- Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
- Thompson, A. P. et al. LAMMPS – a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
- Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
- Wang, Y., Lv, J., Zhu, L. & Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **82**, 094116 (2010).
- Wang, Y., Lv, J., Zhu, L. & Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **183**, 2063–2070 (2012).
- Zheng, Q. et al. Ab initio nonadiabatic molecular dynamics investigations on the excited carriers in condensed matter systems. *WIREs Comput. Mol. Sci.* **9**, e1411 (2019).
- Huang, M. et al. DASP: defect and dopant ab-initio simulation package. *J. Semi-cond.* **43**, 042101 (2022).
- Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. VASPKIT: A user-friendly interface facilitating high-throughput computing and analysis using VASP code. *Comput. Phys. Commun.* **267**, 108033 (2021).
- Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
- Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
- Jain, A. et al. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
- He, B. et al. High-throughput screening platform for solid electrolytes combining hierarchical ion-transport prediction algorithms. *Sci. Data* **7**, 151 (2020).
- Kim, S. et al. A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Sci. Data* **7**, 387 (2020).
- de Pablo, J. J., Jones, B., Kovacs, C. L., Ozolins, V. & Ramirez, A. P. The Materials Genome Initiative, the interplay of experiment, theory and computation. *Curr. Opin. Solid State Mater. Sci.* **18**, 99–117 (2014).
- The materials genome initiative at the national science foundation: a status report after the first year of funded research. *JOM* **66**, 336–344 (2014).
- Kirklin, S. et al. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
- Draxl, C. & Scheffler, M. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
- de Pablo, J. J. et al. New frontiers for the materials genome initiative. *npj Comput. Mater.* **5**, 41 (2019).
- Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
- Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
- Chen, A., Zhang, X. & Zhou, Z. Machine learning: Accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553–576 (2020).
- Han, Y. et al. Machine learning accelerates quantum mechanics predictions of molecular crystals. *Phys. Rep.* **934**, 1–71 (2021).
- Wang, Z., Han, Y., Cai, J., Chen, A. & Li, J. Vision for energy material design: a roadmap for integrated data-driven modeling. *J. Energy Chem.* **71**, 56–62 (2022).
- Zou, X. et al. Machine learning analysis and prediction models of alkaline anion exchange membranes for fuel cells. *Energy Environ. Sci.* **14**, 3965–3975 (2021).
- Zhang, H., Wang, Z., Ren, J., Liu, J. & Li, J. Ultra-fast and accurate binding energy prediction of shuttle effect-suppressive sulfur hosts for lithium-sulfur batteries using machine learning. *Energy Stor. Mater.* **35**, 88–98 (2021).
- Jiang, B. et al. Bayesian learning for rapid prediction of lithium-ion battery-cycling protocols. *Joule* **5**, 3187–3203 (2021).
- Lyu, R., Moore, C. E., Liu, T., Yu, Y. & Wu, Y. Predictive design model for low-dimensional organic–inorganic halide perovskites assisted by machine learning. *J. Am. Chem. Soc.* **143**, 12766–12776 (2021).
- Wang, Z., Cai, J., Wang, Q., Wu, S. & Li, J. Unsupervised discovery of thin-film photovoltaic materials from unlabeled data. *npj Comput. Mater.* **7**, 128 (2021).
- Miyake, Y. & Saeki, A. Machine learning-assisted development of organic solar cell materials: issues, analyses, and outlooks. *J. Phys. Chem. Lett.* **12**, 12391–12401 (2021).
- Batra, R., Song, L. & Ramprasad, R. Emerging materials intelligence ecosystems propelled by machine learning. *Nat. Rev. Mater.* **6**, 655–678 (2021).
- Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* **604**, 662–667 (2022).
- Wang, G. et al. ALKEMIE: An intelligent computational platform for accelerating materials discovery and design. *Comput. Mater. Sci.* **186**, 110064 (2021).
- Zhao, X.-G. et al. JAMIP: an artificial-intelligence aided data-driven infrastructure for computational materials informatics. *Sci. Bull.* **66**, 1973–1985 (2021).
- Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
- Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automaterminer reference algorithm. *npj Comput. Mater.* **6**, 138 (2020).
- Jacobs, R. et al. The Materials Simulation Toolkit for Machine learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research. *Comput. Mater. Sci.* **176**, 109544 (2020).
- Sbailo, L., Fekete, Á., Ghiringhelli, L. M. & Scheffler, M. The NOMAD Artificial-Intelligence Toolkit: turning materials-science data into knowledge and understanding. *npj Comput. Mater.* **8**, 250 (2022).
- Morgan, D. & Jacobs, R. Opportunities and Challenges for Machine Learning in Materials Science. *Annu. Rev. Mater. Res.* **50**, 71–103 (2020).
- Ong, S. P. et al. Python materials genomics (pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
- Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Chen, T. & Guestrin, C. XGBoost. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.* 785–794 <https://doi.org/10.1145/2939672.2939785> (2016).
- Mentel, Ł. mendeleev – A Python resource for properties of chemical elements, ions and isotopes. <https://github.com/lmmmentel/mendeleev> (2014).
- Meredig, B. et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
- Wang, Z., Zhang, H. & Li, J. Accelerated discovery of stable spinels in energy systems via machine learning. *Nano Energy* **81**, 105665 (2021).
- Cai, J., Wang, Z., Wu, S., Han, Y. & Li, J. A machine learning shortcut for screening the spinel structures of Mg/Zn ion battery cathodes with a high conductivity and rapid ion kinetics. *Energy Stor. Mater.* **42**, 277–285 (2021).
- Wang, Z. et al. Harnessing artificial intelligence to holistic design and identification for solid electrolytes. *Nano Energy* **89**, 106337 (2021).
- Wang, Y., Schwartz, J., Gim, J., Hovden, R. & Mi, Z. Stable unassisted solar water splitting on semiconductor photocathodes protected by multifunctional GaN nanostructures. *ACS Energy Lett.* **4**, 1541–1548 (2019).
- Zhang, Y. et al. Synthesis and characterization of spinel cobaltite (Co₃O₄) thin films for function as hole transport materials in organometallic halide perovskite solar cells. *ACS Appl. Energy Mater.* **3**, 3755–3769 (2020).
- He, X. et al. The passivity of lithium electrodes in liquid electrolytes for secondary batteries. *Nat. Rev. Mater.* **6**, 1036–1052 (2021).
- Wang, Z. et al. Computational screening of spinel structure cathodes for Li-ion battery with low expansion and rapid ion kinetics. *Comput. Mater. Sci.* **204**, 111187 (2022).
- Balaish, M. et al. Processing thin but robust electrolytes for solid-state batteries. *Nat. Energy* **6**, 227–239 (2021).
- Chen, Y.-T. et al. Fabrication of high-quality thin solid-state electrolyte films assisted by machine learning. *ACS Energy Lett.* **6**, 1639–1648 (2021).
- Borlido, P. et al. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. *npj Comput. Mater.* **6**, 96 (2020).
- Borlido, P. et al. Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **15**, 5069–5079 (2019).

60. Wang, Z. et al. Deep learning for ultra-fast and high precision screening of energy materials. *Energy Stor. Mater.* **39**, 45–53 (2021).
61. Sun, H. & Zhao, K. Electronic structure and comparative properties of $\text{LiNi}_x\text{Mn}_y\text{Co}_z\text{O}_2$ cathode materials. *J. Phys. Chem. C* **121**, 6002–6010 (2017).
62. Chakraborty, A. et al. Layered cathode materials for lithium-ion batteries: review of computational studies on $\text{LiNi}_{1-x-y}\text{Co}_x\text{Mn}_y\text{O}_2$ and $\text{LiNi}_{1-x-y}\text{Co}_x\text{Al}_y\text{O}_2$. *Chem. Mater.* **32**, 915–952 (2020).
63. Kim, T., Drakopoulos, S. X., Ronca, S. & Minnich, A. J. Origin of high thermal conductivity in disentangled ultra-high molecular weight polyethylene films: ballistic phonons within enlarged crystals. *Nat. Commun.* **13**, 2452 (2022).
64. Zhou, Y., Dong, Z.-Y., Hsieh, W.-P., Goncharov, A. F. & Chen, X.-J. Thermal conductivity of materials under pressure. *Nat. Rev. Phys.* **4**, 319–335 (2022).
65. Wang, Z. et al. IonML: A physically inspired machine learning platform to directed design superionic conductors. *Energy Stor. Mater.* **59**, 102781 (2023).
66. Zhang, H., Wang, Z., Cai, J., Wu, S. & Li, J. Machine-learning-enabled tricks of the trade for rapid host material discovery in Li-S battery. *ACS Appl. Mater. Interfaces* **13**, 53388–53397 (2021).
67. Zhou, J. et al. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Sci. Data* **6**, 86 (2019).
68. Zhuo, Y., Mansouri Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
69. Petretto, G. et al. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 (2018).
70. Petousis, I. et al. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci. Data* **4**, 160134 (2017).
71. Zhang, L. et al. A database of ionic transport characteristics for over 29 000 inorganic compounds. *Adv. Funct. Mater.* **30**, 2003087 (2020).
72. Zhu, T. et al. Charting lattice thermal conductivity for inorganic crystals and discovering rare earth chalcogenides for thermoelectrics. *Energy Environ. Sci.* **14**, 3559–3566 (2021).
73. Yu, J. et al. Finding optimal mid-infrared nonlinear optical materials in germanates by first-principles high-throughput screening and experimental verification. *ACS Appl. Mater. Interfaces* **12**, 45023–45035 (2020).

ACKNOWLEDGEMENTS

Our work is supported by the National Key Research & Development Program of China (No. 2021YFC2100100), and the Shanghai Science and Technology Project (No. 21JC1403400). We also acknowledge the important contributions to the development of AlphaMat code from the following AIMS-Lab members in Shanghai Jiao Tong University: Lin Zhang, Xirong Lin, Sicheng Wu, Zehao Yu, Jiequn Tang.

AUTHOR CONTRIBUTIONS

Z.W., A.C., K.T.: Conceptualization, Methodology, Visualization, Data curation, Writing - original draft. J.C., Y.H.: Methodology, Visualization. S.Y., S.W., I.A.: Data curation. J.L.: Conceptualization, Methodology, Supervision, Resources, Writing - review & editing. All authors commented on the manuscript. These authors contributed equally: Z.W., A.C., and K.T.

COMPETING INTERESTS

The Software Copyrights have been obtained in China for AlphaMat with No. 2022SR0405168 and No. 2022SR1364423.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-023-01086-5>.

Correspondence and requests for materials should be addressed to Jinjin Li.

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

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



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

© The Author(s) 2023