Article

MLMD: a programming-free AI platform to predict and design materials

Check for updates

Jiaxuan Ma[1,2,5], Bin Cao [3,5], Shuya Dong[1], Yuan Tian[1], Menghuan Wang[1], Jie Xiong [1,4] ✉ & Sheng Sun[1,2,4] ✉

Accelerating the discovery of advanced materials is crucial for modern industries, aerospace, biomedicine, and energy. Nevertheless, only a small fraction of materials are currently under experimental investigation within the vast chemical space. Materials scientists are plagued by time-consuming and labor-intensive experiments due to lacking efficient material discovery strategies. Artificial intelligence (AI) has emerged as a promising instrument to bridge this gap. Although numerous AI toolkits or platforms for material science have been developed, they suffer from many shortcomings. These include primarily focusing on material property prediction and being unfriendly to material scientists lacking programming experience, especially performing poorly with limited data. Here, we developed MLMD, an AI platform for materials design. It is capable of effectively discovering novel materials with high-potential advanced properties end-to-end, utilizing model inference, surrogate optimization, and even working in situations of data scarcity based on active learning. Additionally, it integrates data analysis, descriptor refactoring, hyper-parameters auto-optimizing, and properties prediction. It also provides a web-based friendly interface without need programming and can be used anywhere, anytime. MLMD is dedicated to the integration of material experiment/ computation and design, and accelerate the new material discovery with desired one or multiple properties. It demonstrates the strong power to direct experiments on various materials (perovskites, steel, high-entropy alloy, etc). MLMD will be an essential tool for materials scientists and facilitate the advancement of materials informatics.

Novel materials have a significant impact on our daily lives and modern industries such as the aerospace, biomedical, and energy sectors[1–5]. However, the conventional research and design (R&D) of novel materials utilizes a "trial-and-error" approach, which is challenging due to the complexity and diversity of materials. This approach incurs high costs, and the commercial implementation of novel materials generally takes decades. Fortunately, owing to the rapid advancements in artificial intelligence (AI) and machine learning (ML), materials R&D has evolved to a state-of-the-art data-driven paradigm[6–8]. The data-driven paradigm is expected to diminish the cost and duration of materials R&D by half, and thus expedite the materials development cycle from decades to a few years.

The key concept of the date-driven paradigm is the integration of AI techniques and materials science. Various AI techniques, especially ML algorithms, have been employed to uncover Composition-Process-Structure-Property (CPSP) relationships in materials science[9–13]. ML models trained on extensive data can aid in the discovery of innovative materials such as organic compounds[14,15], solar cells[16,17], alloys[18–20], and perovskites[21,22]. For example, Rao et al. proposed an active learning strategy to accelerate the exploration of high-entropy Invar alloys. With the assistance of ML, they characterized two novel Invar alloys from a plethora of potential combinations[23]. Raccuglia et al. utilized successful and unsuccessful historical reactions to train their ML model, discovering critical factors that affect chemical reactions and synthesizing new organic compounds based on this trained model[24].

Within the material and physical computation community, several notable AI platforms have emerged. Materials Cloud[25] offers an ensemble

[1]Materials Genome Institute, Shanghai University, Shanghai 200444, China. [2]Zhejiang Laboratory, Hangzhou 311100, China. [3]Guangzhou Municipal Key Laboratory of Materials Informatics, Advanced Materials Thrust, Hong Kong University of Science and Technology (Guangzhou), Guangzhou 510000, China. [4]Shanghai Frontier Science Center of Mechanoinformatics, Shanghai University, Shanghai 200444, China. [5]These authors contributed equally: Jiaxuan Ma, Bin Cao. ✉e-mail: xiongjie@shu.edu.cn; mgissh@t.shu.edu.cn

material simulation platform with a primary focus on ab initio computation to emulate material mechanisms. Similarly, the Materials Project[26] stands out as a quantum computation platform providing a valuable inorganic materials database and associated quantum properties, leveraging advanced ML acceleration methods such as ML-based potential fields: M3GNet[27], MEGNet[28], etc. AFLOW-ML[29] and JARVIS-ML[30] contribute significantly by offering crystal properties prediction tools based on DFT calculations or ML surrogate models, covering formation energies, exfoliation energies, bandgaps, and magnetic moments, among other attributes. MatMiner[31] and Magpie[32], popular ML tools in materials research, host various downstream ML libraries that greatly benefit the material community. In addition, the AI/materials community has also developed general AI toolkits and platforms that have broad applications in materials science[33–36]. For instance, a simple but powerful tool for non-data researchers has been developed by Peng[37], which possesses a command-line interface and a web-based GUI for classification and regression tasks. Another representative is AlphaMat, an emerging materials informatics hub developed by Wang[38]. This platform focuses on data preprocessing and downstream ML models, offering a streamlined workflow for materials designers. AI platforms like AlphaMat can automate and expedite the construction of accurate ML models for different materials, providing AI support for materials scientists.

Despite these advancements, utilizing the aforementioned general-purpose AI toolkits and platforms requires programming skills, which might be a barrier for material designers lacking programming experience. Additionally, existing platforms emphasize model construction and overlook inverse materials design. There is significant scope for enriching AI platforms in material R&D. Therefore, we have created the MLMD (machine learning for materials design) platform, providing a friendly interface and more comprehensive AI tools for material designers. MLMD distinguishes itself with an entirely code-free interface for seamlessly executing property inference and inverse material design. Crucially, MLMD not only considers inherent physical structure properties but also explores the impact of material defects and processing/testing technologies on material performance. It includes a range of ML algorithms for constructing models besides data analysis and descriptor refactoring, enabling end-to-end (data to new materials) novel materials discovery. MLMD can automatically construct classification or regression models based on user-uploaded data, with a strong emphasis on privacy protection (no data will be stored). Users can simply select an ML algorithm, and the associated hyper-parameters will be automatically tuned. In addition to model construction, MLMD also incorporates model inference[20,39], surrogate optimization[40–42], and active learning[43–48] techniques for materials inverse design. By utilizing model inference or surrogate optimization, users can screen out novel materials from a virtual search space based on the constructed models.

In addition, the MLMD framework has made significant strides in addressing the challenge of limited data availability. To address data scarcity in materials science, we develop a Bayesian toolkit for material design to integrate into active modules in MLMD. This feature incorporates nine utility functions that balance exploration and exploitation. Through the active learning module, MLMD enables the provision of novel advanced materials with single or multiobjective properties through an iterative design process. MLMD also integrates transfer learning into heuristic algorithms (TL-opt) to address small data problems in material design, demonstrating its applicability in both single and multiple objective material design, showcasing advantages in Al alloy design[42].

The effectiveness and robustness of the MLMD platform in both model construction and inverse design have been demonstrated through various datasets in the work, including perovskites, steel, high-entropy alloy, et al. This framework adheres to the material genome concept, thereby emphasizing the emerging paradigm of material design. We firmly believe that the MLMD platform has the potential to greatly enhance the accessibility and utility of AI techniques for materials communities.

Nomenclature

LR[49]	Logic regression	MLPR[49]	Multi-layer perception regression
SVC[49]	Support vector classification	RFR[49]	Random forest regression
BTC[49]	Bagging tree classification	XGBR[50]	XGBoost regression
RFC[49]	Random forest classification	LOO	Leave one out
XGBC[50]	XGBoost classification	R^2	Determination coefficient
CBC[51]	CatBoost classification	NSGA-II[52]	Non-dominated sorting genetic algorithm II
CV	Cross validation	GA[53]	Genetic algorithm
SVR[49]	Support vector regression	DE[54]	Differential evolution
KNNR[49]	K-nearest neighbor regression	PSO[55]	Particle swarm optimization
SA[56]	Simulated annealing	EI[57]	Expected improvement
AEI[57]	Augmented expected improvement	EQI[57]	Expected quantile improvement
REI[57]	Reinterpolation expected improvement	PES[58]	Predictive entropy search
POI[59]	Probability of improvement	LassoR[49]	Lasso regression
UCB[60]	Upper confidence bound	KG[61]	Knowledge gradient
EHVI[62]	Expected hypervolume improvement	LinearR[49]	Linear regression
PCA[49]	Principal component analysis	t-SEN[49]	t-distribution stochastic neighbor embedding
ABR[49]	AdaBoost regression	BR[49]	Bagging regression
CBR[51]	CatBoost regression	GPR[49]	Gaussian process regression
DTR[49]	Decision tree regression	GBR[49]	Gradient boosting regression
RidgeR[49]	Ridge regression	ABC[49]	AdaBoost classification
EIP[57]	Expected improvement with "plugin"	SHAP[63]	SHapley Additive exPlanations
GBC[49]	Gradient boosting classification	SMS-EMOA[52]	$\mathscr{L}$ metric selection evolutionary multiobjective optimization algorithm

Results

Overview and architecture

The primary objective of MLMD is to make ML programming free and empower materials scientists with an end-to-end approach to materials design. To train a prediction model on MLMD, users are required to upload a CSV-formatted data file containing feature matrix X and target variable Y. The feature matrix X includes information regarding material components and processes, and the target variable Y incorporates one or more material properties. MLMD was developed with six core modules, as depicted in Fig. 1, and seven functionalities:

(1) Database. MLMD provides materials scientists with databases containing material data (e.g., polycrystalline ceramic, HEAs, ferroelectric perovskites) generated from experiments or collected from literatures. These databases are downloadable and serve various purposes, including serving as source domains for transfer learning. Additionally, MLMD offers outlier detection algorithms like DBSCAN[64], IsolationForest[65], LocalOutlierFactor[66], and One Class SVM[67] to identify data points that deviate significantly from the rest. Outlier detection can greatly enhance the generalization of ML models.

(2) Data visualization. MLMD offers an initial data overview, including the distribution of features and targets, along with the statistics derived from the data.

(3) Feature engineering. Material compositions and processes significantly influence the structure, properties, and performance of materials. These components are commonly used as the feature descriptors in ML, determining the performance limits of prediction models. Thus, MLMD integrates feature engineering, encompassing handling missing and duplicate values, assessing feature correlation, and ranking feature importance. Additionally, MLMD also provides transformation functions to transform the composition descriptors to atomic descriptors, such as atomic radius, band gap, and valences.

(4) Quantitative CPSP relationships (QCPSP). The establishment of QCPSP in material through ML is fundamental to material design. MLMD supports nearly all widely utilized regression and classification algorithms, such as linear analysis, sparse kernel machine, probability model, neural network, transfer learning, and ensemble learning. The most suitable model can be selected for training on the data and making inference.

(5) Surrogate optimization. Integrating predictive models into numerical optimization algorithms can accelerate the attainment of optimal material compositions, processes, and other relevant features that align with desired properties. Subsequently, the discovered advanced materials will undergo experimental verification.

(6) Active learning. Achieving a high-accuracy prediction model is a challenge in materials science due to the limited data. Consequently, sampling-based material design strategies are provided within the active learning module in MLMD platform. Global optimization by Bayesian-based active learning has gained prominence for the ability to address data scarcity and reduce material discovery costs. Typically, this approach explores the design space using an optimal policy that balances exploitation and exploration to identify the global optimum. In active learning, a probability surrogate model, Gaussian process (GP), is constructed using initial input-output data obtained from costly experiments or simulations.

(7) Interpretable ML. Achieving physical interpretability is a significant challenge and goal in material informatics. Interpretable ML can enhance materials scientists' understanding of the CPSP relationships of materials. MLMD platform also provides the Shapley Additive Explanations (SHAP) method to facilitate model explanation.

As illustrated in Fig. 2, there are three primary flowcharts for materials design within the MLMD platform. These include model inference, surrogate optimization, and active learning. The efficiency of model inference and surrogate optimization relies on the robustness of the prediction model (surrogate model), and the model performance is limited by available data. In surrogate optimization, the well-trained prediction model will be integrated into stochastic optimization algorithms to accelerate materials design. Active learning in MLMD employs a sampling strategy grounded in Bayesian principles. It balances exploration and exploitation in order to formulate an optimal material design strategy. The active learning module will recommend the next experiment via Bayesian global optimization under limited data. The recommended experiment can be conducted, and the new results of the experiment will validate the ML prediction and simultaneously are fed back to the dataset for the following cycle of iterations in the active learning loop. In the materials design flowcharts within MLMD platform, some useful tools were used, including Streamlit, Scikit-Learn[49], Pymoo[52], extreme gradient boosting decision tree (XGBoost)[50], Scikit-Opt, and Bgolearn.

Classification module

Here, eight materials datasets labeled as C1-C3 and R1-R5 were used as case studies to showcase the reliability and effectiveness of four commonly used modules within our MLMD platform. Details of the datasets can be found in Table 1.

The proposed classification module aims to address classification issues in materials science. It merely necessitates uploading a dataset in the CSV format and the programming-free selection of an MLMD-implemented algorithm to complete model construction. Moreover, the user can straightforwardly adjust or auto-optimize hyper-parameters to further refine constructed MLMD classification models to enhance their accuracy (metric details are provided in Supplementary Note 1). The performances of six MLMD-implemented classification algorithms (LR, SVC, BTC, RFC, XGBC, and CBC), were accessed and compared with the baseline models across three distinct classification issues, described as follows. C1: Identify the crystalline structure of a polycrystalline ferroelectric ceramic, categorizing it as either a perovskite or a non-perovskite structure. C2: Categorize an alloy into one of three classes: crystalline alloy (CRA), ribbon metallic glass (RMG), or bulk metallic glass (BMG). C3: Discriminate between solid-solution HEAs and classify them as hexagonal close-packed (HCP), body-centered cubic (BCC), face-centered cubic (FCC), or mixed solid solution (MSS). The baseline models were SVC implemented in R[75], RFC implemented in Java[13], and RFC implemented in python[76] for C1, C2 and C3, respectively.

As depicted in Fig. 3a, c, e, the default MLMD-implemented SVC, RFC, and XGBC models achieved a 10-fold CV accuracy exceeding 80% for all three issues. The results demonstrate that the default MLMD-implemented model can provide satisfactory classification accuracy without requiring any other operations. In addition, the optimization of hyper-parameters can significantly enhance the performance of all MLMD-implemented models. The MLMD platform also offers a user-friendly hyper-parameter tuning feature that can achieve improved models without requiring programming skills. In the three cases, the recommended MLMD-implemented models are the tuned XGBC model (CV-accuracy = 86.5%), tuned RFC model (CV-accuracy = 87.4%), and tuned RFC model (CV-accuracy = 92.6%) for C1, C2, and C3, respectively. The recommended MLMD-implemented model performed comparably to the baseline model for C2, and outperformed the baseline model for C1 and C3, indicating the robust classification capability of our platform. MLMD platform also provides a confusion matrix for each recommended model in the classification module, as illustrated in Fig. 3b, d, f. The confusion matrix is used to observe the performance of the classification model in each category, and is able to calculate the other classify performance metrics, such as precision and recall. (raw confusion matrix plots are provided in Supplementary Fig. 1). According to the CPSP relationship, the property of a material significantly relies on its microstructure. Therefore, identifying the microstructure based on composition and process is very important for material design. For instance, the BCC HEAs are much harder than FCC HEAs, and a HEA that belongs to BCC class should be designed for wear resistance. Researchers generally modify microstructures based on experience in conventional materials design paradigm, while our platform provides a convenient tool for identifying the microstructure through classification.
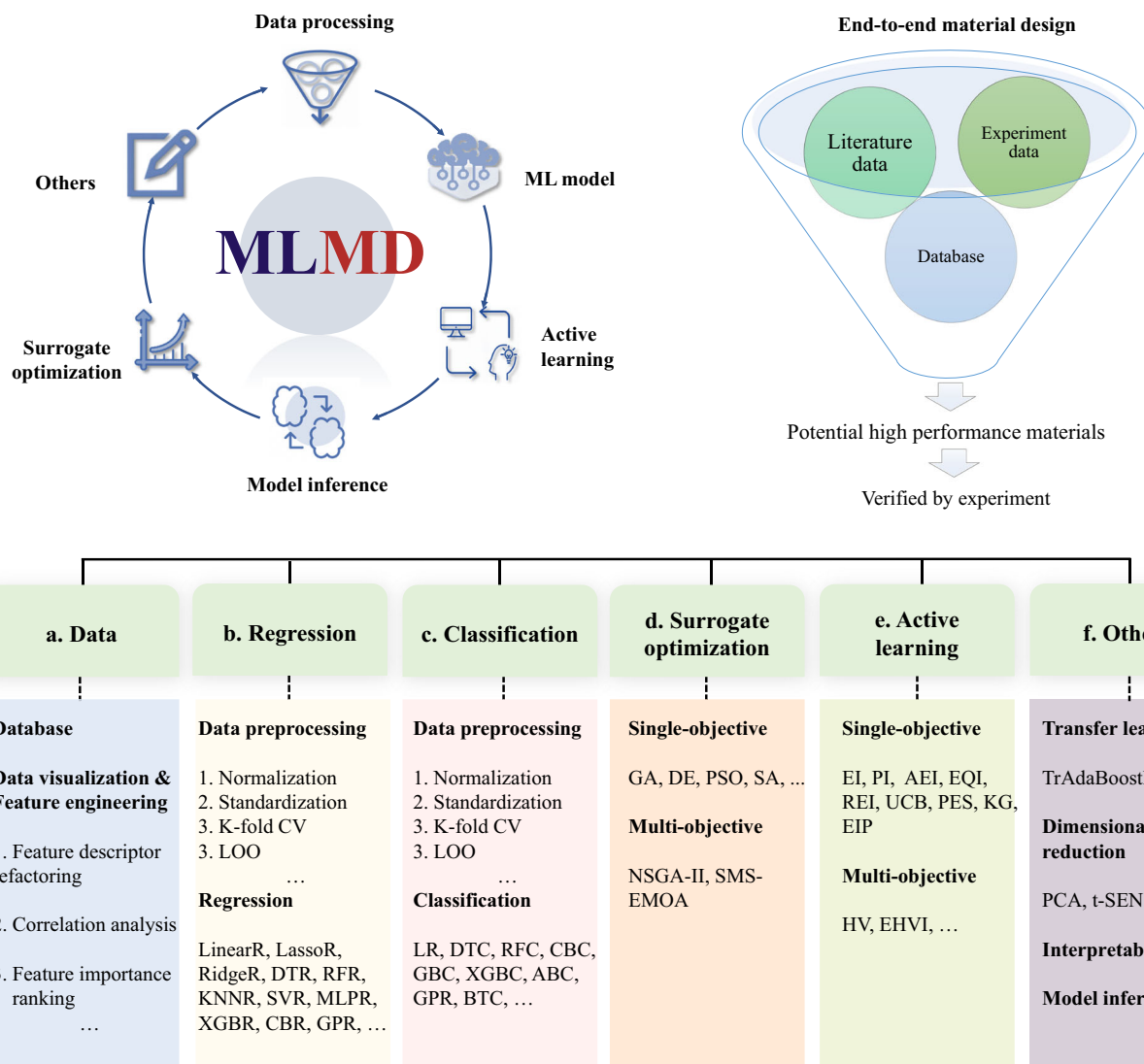
Fig. 1 | **Overview and architecture of MLMD. a** Data module, which encompasses material databases, data visualization, and feature engineering. **b** Regression module, which comprises a group of ML regression algorithms. These algorithms can be further utilized in the surrogate optimization module. **c** Classification module, which involves a group of ML classification algorithms. These algorithms can be further leveraged in the surrogate optimization module. **d** Surrogate optimization module, where the ML model is incorporated into numerical algorithms to accelerate materials design. **e** Active learning module, sampling methods based on Bayesian are provided to search the material composition space and discover novel materials, particularly under limited available data. **f** Other module, which provides advanced ML algorithms such as transfer learning, dimensionality reduction, and interpretable ML.

Regression module

Similar to the Classification Module, the regression module only requires a CSV-format dataset for constructing predictive models. It is flexible to select various regression algorithms and adjust corresponding hyper-parameters without programming. We compared the performance of six MLMD-implemented regressors (SVR, KNNR, MLPR, RFR, XGBR, and CBR) to the baseline model for predicting the fracture stress of low-alloy steels (R1), the Curie temperature of ferroelectric perovskite (R2), and the flow stress of FGH98 superalloys under hot deformation (R3). The baseline models utilized here were RFR implemented in Java[77], SVR implemented in R[75], and GPR implemented in python[78] for R1, R2 and R3, respectively. As can be seen in Fig. 4a, c, e, the 10-fold CV-R^2 (metric details are provided in Supplementary Note 1) of the recommended XGBR model for R1, SVR model for R2, and CBR model for R3 are 0.9427, 0.8480, and 0.9828, respectively. The recommended MLMD-implemented models outperform the baseline model for all three regression problems. The properties predicted from the recommended MLMD-implemented regressors have been plotted against the experiment measurement, as shown in Fig. 4b, d, f.

Notably, the data points clustering near the diagonal line exhibit that our MLMD platform delivers satisfactory performance across diverse regression problems (raw plots are provided in Supplementary Fig. 2). Different from the classification, regression is commonly employed to predict material performance in relation to properties like strength, elongation, and hardness, among others. Researchers can leverage well-trained regression models to surrogate time-consuming trial-and-error experiments and design advanced materials at low cost. The regression module within MLMD offers a convenient tool for the experiment researchers lacking programming skills.

In summary, the classification and regression module within MLMD platform allow the acquisition of accurate models through programming-free algorithm selection and hyper-parameter tuning. Additionally, well-trained prediction models can be preserved for various other applications.

Surrogate optimization module

The well-trained regression model serves as a surrogate model, which can be integrated into the efficient numerical optimization algorithm to accelerate
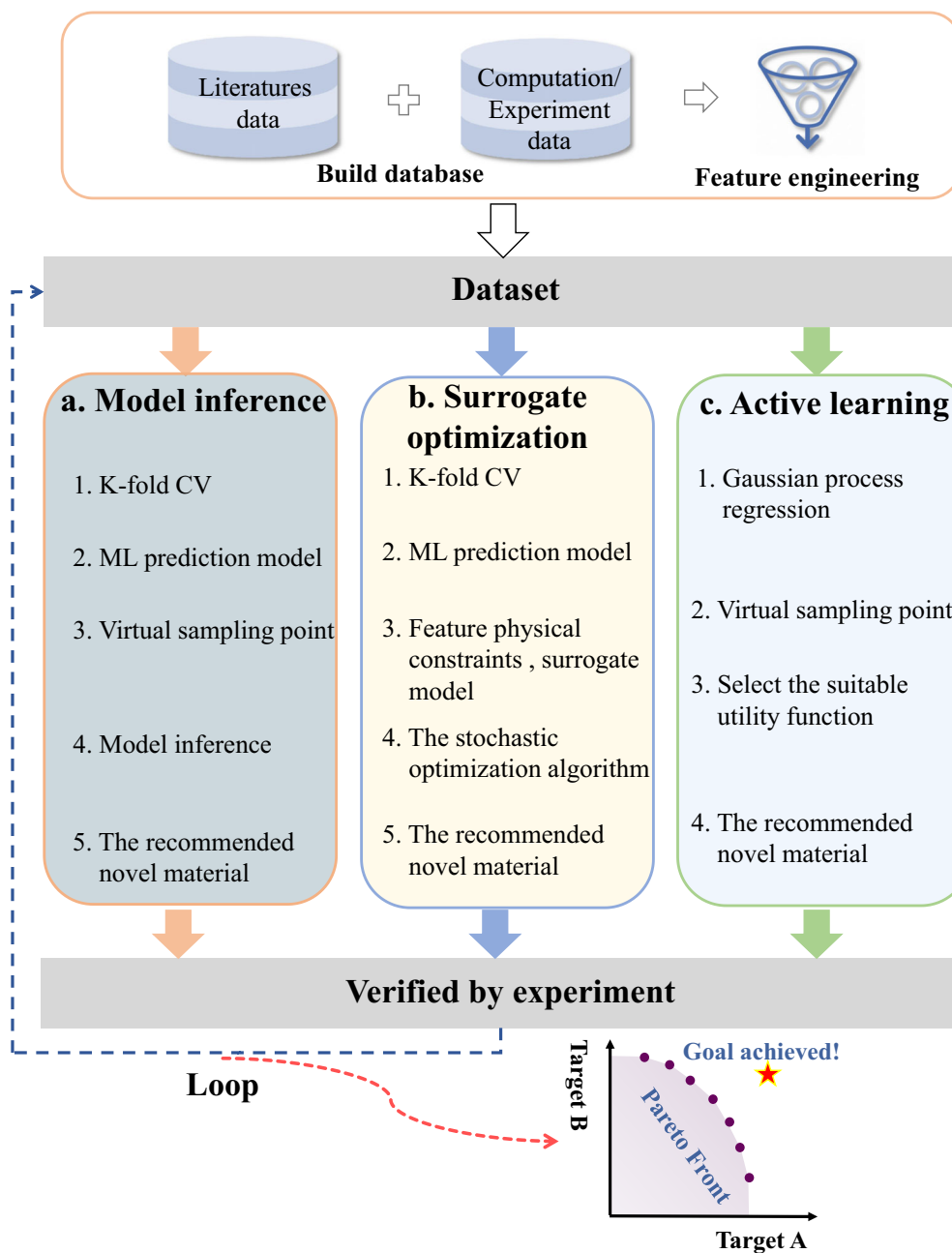
Fig. 2 | Flowcharts of materials design in MLMD platform. a Model inference involves establishing an ML model, generating visual samples, executing model inference, and subsequently verifying the recommended new materials through experiment. **b** Surrogate optimization entails establishing ML models, incorporating feature physical constraints, selecting heuristic optimization algorithms, and subsequently verifying the recommended new materials through experiment. **c** Active learning involves utilizing GPR model, creating a virtual sample space, selecting a suitable utility function, and ultimately verifying the recommended new materials through experiment. Notably, the results from experiments on the new materials of the three methods can be iterated into the next loop for building more better ML model.

the materials design. The surrogate optimization module within MLMD platform necessitates an experimental or simulation dataset, corresponding ML prediction models, and boundary constraints for feature variable. The surrogate optimization module provided a convenient tool that quickly finds out a reasonable combination in the search space of compositions and processing parameters according to the targeted properties.

Reduced activation ferritic-martensitic (RAFM) steels developed from conventional 9Cr-1Mo steels have been regarded as promising candidate structural materials for fusion reactors, owing to their good thermo-physical, thermomechanical, and irradiation-resistant properties compared to those of austenitic steels[68,69]. In the section, the surrogate optimization module was utilized to design RAFM steels with enhanced strength and excellent ductility. Initially, two ML prediction models were constructed via
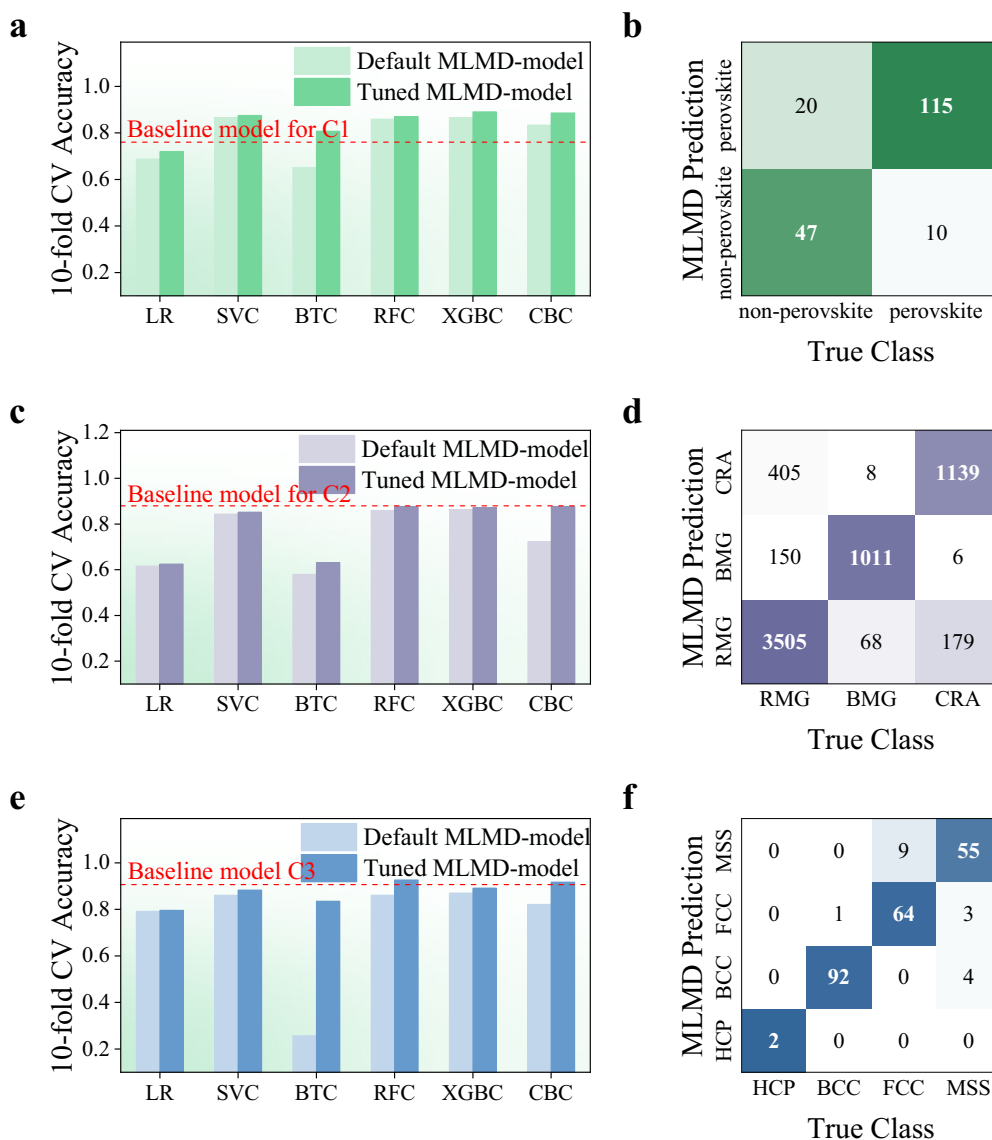
the regression module in MLMD platform based on dataset R4. It is worth noting that an accurate prediction model is an essential prerequisite for optimization. The CV-ρ (metric details are provided in Supplementary Note 1) of the regressor for predicting UTS and total elongation (TE) is 0.9912 and 0.8816, respectively, as shown in Fig. 5a, b. Hence, these two ML prediction models can be employed to discover novel RAFM steels with given physical constraints (details are provided in Supplementary Table 1). Figure 5c, d depict the tensile properties of RAFM steels in R4-dataset and MLMD-recommended steels at 600 °C and 300 °C, respectively, and the Pareto front of RAFM steel is significantly pushed forward, especially at 300 °C. Li et al.[70] proposed an intelligent design model, which consists of the forward model that maps composition and processing to properties and the reverse model that maps properties to composition and processing. The

intelligent design model is capable of implementing property-oriented composition and processing design. As illustrated in Fig. 5c, the optimal

Table 1 | Eight case material datasets

Dataset	Materials	Property
C1[75]	polycrystalline ceramic	formability
C2[13]	alloy	glass-forming ability
C3[76]	high-entropy alloy	solid-solution structure
R1[77]	low-alloy steel	fracture strength
R2[75]	ferroelectric perovskites	Curie temperature
R3[78]	FGH98 superalloy	flow stress
R4[70]	reduced activation ferritic-martensitic steel	ultimate tensile strength and total elongation
R5[74]	AlCoCrCuFeNi high-entropy alloy	hardness

RAFM steel at ambient temperature 600 °C achieved by Li is denoted by the blue star, and the RAFM steels designed by MLMD are highlighted by the red star. It is evident that the properties of the designed materials are very close. Furthermore, the compositions and processing of the materials are also similar, as shown in Table 2. Besides, MLMD can also discover other advanced materials with interesting properties that lie on Pareto Front, and the novel material number depends on the simple hyper-parameter setting. These materials can be applied in various scenarios according to specific requirements (the composition and processes at 600 °C are provided in Supplementary Table 2). Meanwhile, we also design RAFM steels at temperature 300 °C to further demonstrate the convenience and effectiveness of material design in MLMD platform, as shown in Fig. 5d. The recommended properties exhibit a sharp improvement compared to the report[71], with a UTS of 643 MPa and TE of 14.64%. The selected material on Pareto Front, highlighted with a red star at 723.10 MPa and a TE of 20.7%, shows a 12.5% improvement in UTS and a 41.4% improvement in TE. An experiment will be conducted to validate the finding in a future study (the composition and
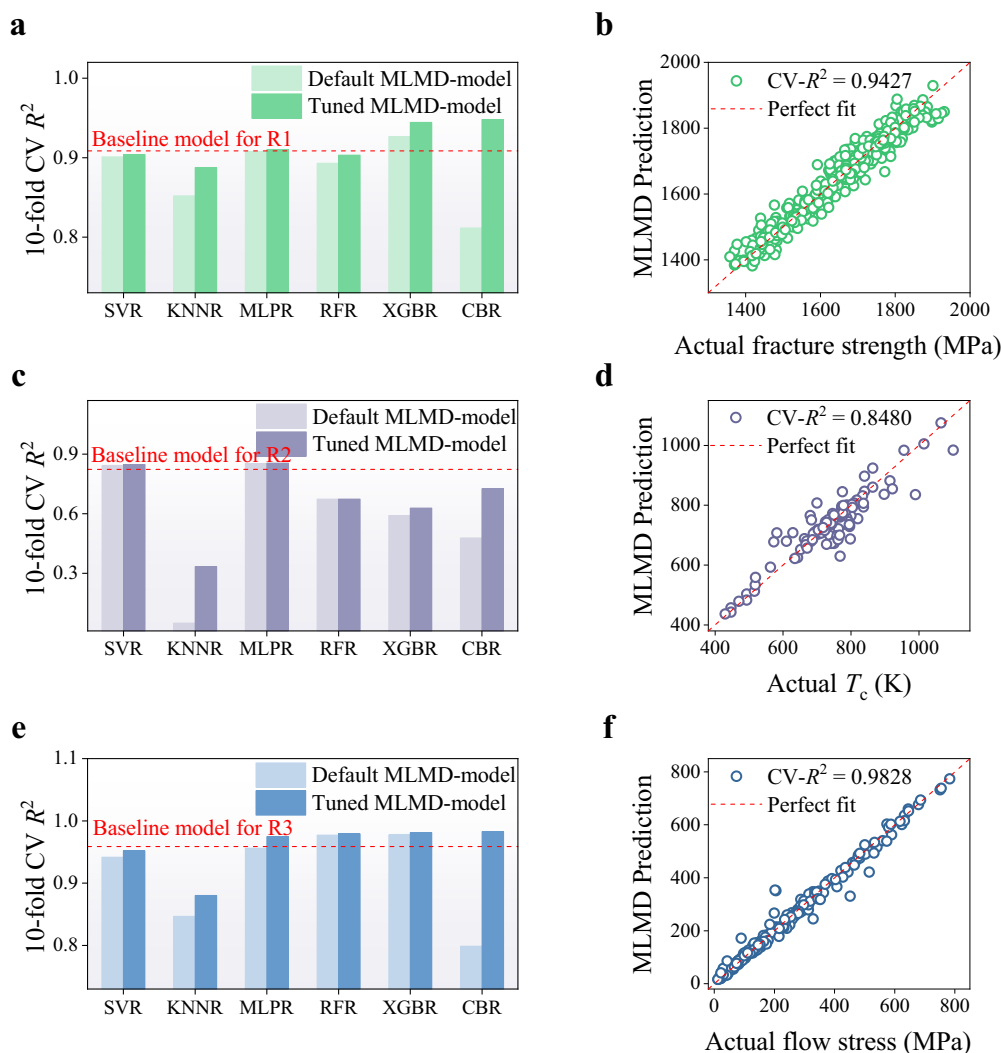

Fig. 3 | Cross-validation results of six ML models through classification module within our MLMD platform. a The 10-fold CV accuracy for classifying perovskite formability of ferroelectric perovskites (Case C1). **b** The confusion matrix of the tuned MLMD classification model of perovskite formability of ferroelectric perovskites (Case C1). **c** The 10-fold CV accuracy for classifying the glass-forming ability of alloys (Case C2). **d** The confusion matrix of the tuned MLMD classification model of the glass-forming ability of alloys (Case C2). **e** The 10-fold CV accuracy for classifying solid-solution structures of HEAs (Case C3). **f** The confusion matrix of the tuned MLMD classification model of solid-solution structures of HEAs (Case C3). In each sub-figure, the lighter bar, darker bar, and red dashed line represent the default MLMD model, tuned MLMD model, and baseline model, respectively.

Fig. 4 | Cross-validation results of six ML models through regression module within our MLMD platform. a The 10-fold CV R^2 for regressing fracture strength of steels (Case R1). **(b)** The prediction of tuned MLMD regression model of perovskite fracture strength of steels (Case R1). **c** The 10-fold CV R^2 for regressing high ferroelectric Curie temperature of perovskite (Case R2). **d** The prediction of tuned MLMD regression model of high ferroelectric Curie temperature of perovskite (Case R2). **e** The 10-fold CV R^2 for regressing flow stress of FGH98 superalloys under hot deformation. (Case R3). **f** The prediction of tuned MLMD regression model of flow stress of FGH98 superalloys under hot deformation (Case R3). In each sub-figure, the lighter bar, darker bar, and red dashed line represent the default MLMD model, tuned MLMD model, and baseline model, respectively.

processes at 300 °C are provided in Supplementary Table 3). In summary, these results demonstrate the powerful ability of the surrogate optimization module within the MLMD platform to effectively accelerate the discovery of high performance materials (raw plots are provided in Supplementary Fig. 3).

Active learning module

As previously mentioned, an accurate predictive model is necessary for implementing surrogate optimization, but it often requires a large quantity of data, while the data in materials science are typically limited. Consequently, we provide a sampling-based material design strategy within the active learning module of MLMD platform. The active learning module necessitates only two datasets: the experimental/simulation data and the virtual sample data. Novel materials with desired properties can be discovered via Bayesian sampling implemented in MLMD platform from the virtual sample space.

HEAs possess excellent properties, including cryogenic toughness, strength, and thermal stability at elevated temperatures, as well as good corrosion and wear resistance[72,73]. Wen et al. designed a strategy combining SVM with experimental design algorithms to search for

$Al_xCo_yCr_zCu_uFe_vNi_w$ HEAs with large hardness[74]. After seven iterations by iterative loops of AI-dominated and knowledge-dominated methods, the 10 new alloys with high hardness were achieved and synthesized, and their compositions are listed in Table 3. To assess the effectiveness and convenience of the active learning module within the MLMD platform, we use the identical samples as those utilized in Wen' report[74], which encompassed 155 synthesized HEAs as well as the supplementary virtual sample space, to design the HEAs with high hardness. The concentrations of six-component alloys in the virtual sample space are constrained within $34 < x < 47$, $5 < y < 33$, $8 < z < 34$, $0 < u < 13$, $5 < v < 20$, and $0 < w < 16$ at.%. Various utility functions are utilized for HEAs design within active learning module, towards desired properties, including EI, EIP, AEI, EQI, REI, UCB, POI, and PES. The distinction between utility functions lies in their varying emphasis on exploration or exploitation during the sampling process. Furthermore, the appropriate utility function may exhibit varying levels of efficiency, necessitating a case-by-case evaluation. Fortunately, the selection of the utility functions can be easily changed due to the programming-free nature of MLMD. Here, the results of three samples designed using EI, REI, and UCB in active module are illustrated in Fig. 6a–c, d–f, g–i, respectively (the results of remaining utility functions are provided in Supplementary
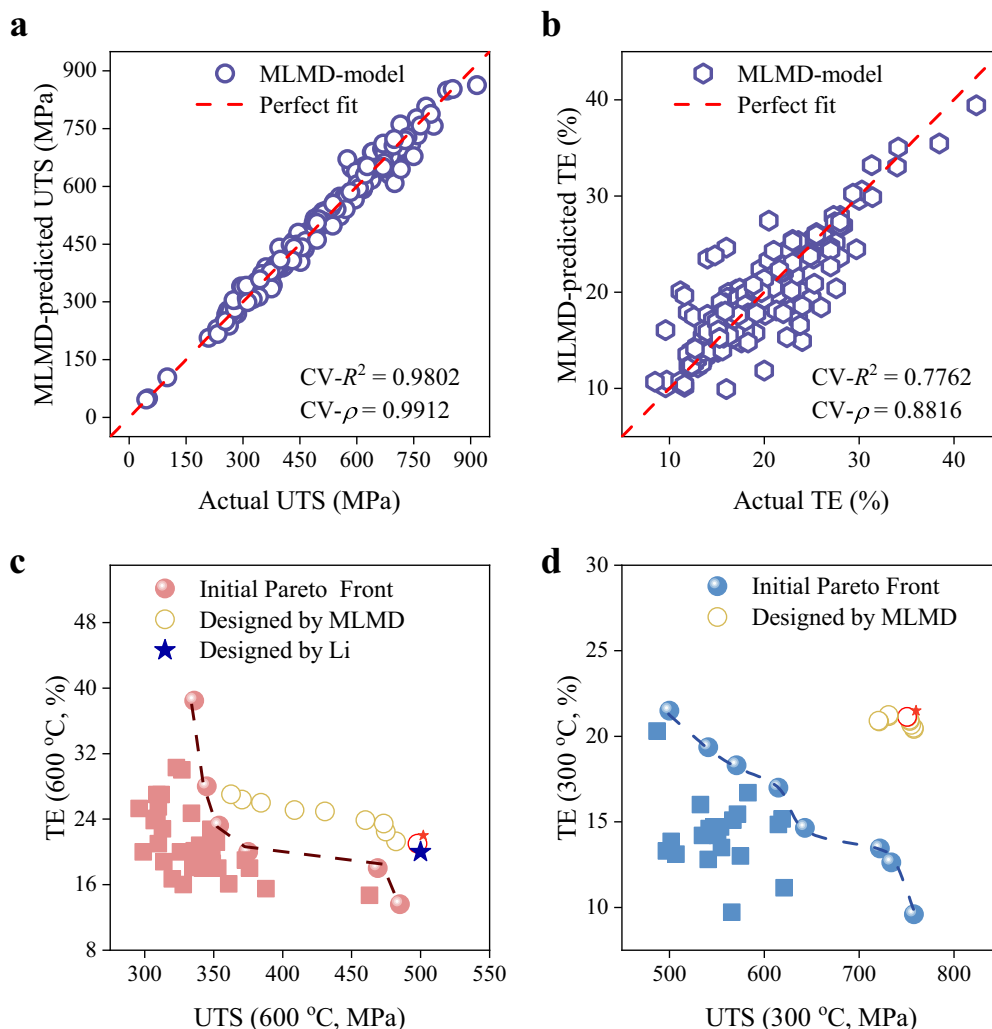
Fig. 5 | The RAFM steels design process through surrogate optimization module in MLMD. a The prediction of UTS via the tuned MLMD regression model. **b** The prediction of TE via the tuned MLMD regression model. **c** The Pareto front of RAFM steels in original data is represented by red circles, and the pushed forward Pareto front of RAFM steels by NSGA-II algorithm with surrogate optimization module are represented by yellow circles at ambient temperature 600 °C. The blue Pentasta represents the material designed by the original work. **d** The Pareto front of RAFM steels in original data are represented by blue circles, and the pushed forward Pareto front of RAFM steels by NSGA-II algorithm with surrogate optimization module are represented by yellow circles at ambient temperature 300 °C. The red Pentasta represents the optimal material designed by MLMD at an ambient temperature 300 °C.

Table 2 | Comparison among the AI-model-based designs and experiment results

	UTS (MPa)	TE	T$_{test}$	Alloy elements (wt%)												NT	Nt	TT	Tt
		(%)	°C	C	Cr	W	Si	Mn	V	Ta	Zr	Y	Ti	N	(°C)	(min)	(°C)	(min)
Designed by work[70]	500	20.0	600	0.13	9.5	1.7	0.49	0.67	0.25	0.20	0.002	0.03	0.005	0.021	1010	36	660	76
Designed by MLMD	498	21.0	600	0.11	9.7	1.6	0.41	0.27	0.49	0.20	0.004	0.009	0.010	0.039	1045	49	658	75
Experiment in work[70]	539	20.6	600	0.13	9.5	1.7	0.49	0.67	0.25	0.20	0.002	0.03	0.005	0.021	1010	36	660	76

The designed data are provided in original work[70]. The targeted properties are UTS of 500 MPa and TE of 20% at 600 °C.

Note 2). The compositions of three samples recommended by the three sampling approaches closely resemble the alloys designed in the original work[74] through multiple iterations. The detailed composition of HEAs through the active learning module is presented in Table 4. Consequently, these alloys designed using MLMD demonstrate a high potential to posses high hardness. A similar design strategy can be applied to optimize other properties, such as light HEAs with high strength and parameters of HEA coatings. The active learning module within MLMD platform can also be extended to bulk metallic glasses, superalloys, and other materials.

Discussion

In this work, MLMD, a cutting-edge AI platform for material design was developed, aiming to accelerate the discovery of advanced materials. MLMD provides a user-friendly programming-free interface. MLMD enables efficient end-to-end materials design with one or more desired properties. Simultaneously, to tackle the issues of limited data availability in materials science, MLMD also provides the sampling-based active learning method to recommend new materials efficiently. It also integrates commonly utilized functions in AI platform such as data analysis, descriptors refactoring, and

prediction modules. The outcomes of regression, classification, surrogate optimization, and active learning modules within MLMD in the study demonstrate the strong power of material design.

Notably, the approach to material design employed in MLMD differs from that outlined in the original research work[70,74]. In the surrogate optimization and active learning module, MLMD employs a distinct method for

recommending novel materials, albeit yielding highly consistent results within previous research. Additionally, the user-friendly interface of MLMD streamlines the design process, ensuring that researchers without extensive programming skills can focus their efforts on conducting experiments, analyzing mechanisms and characterizing novel materials.

Finally, we will continue to commit to ongoing enhancements and releases of MLMD, aiming to address the challenges frequently encountered in material design, including the development of more efficient ML algorithms, more frontier tools, and visualization interfaces to enhance the efficiency of processing material data. We believe that MLMD has the potential to become an indispensable tool for material design, especially friendly for researchers who are unfamiliar with programming, and facilitate the advancement of material informatics.
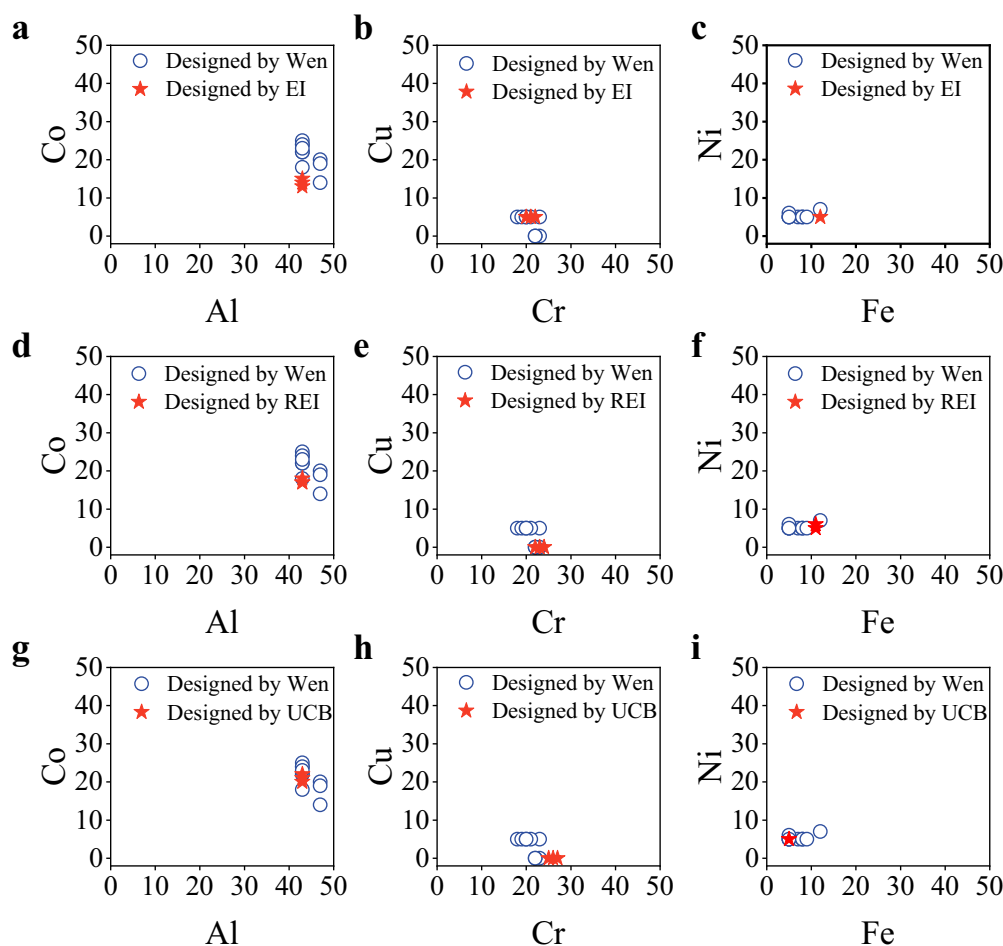
Table 3 | The top 10 newly predicted and synthesized alloys after seven iterations by iterative loops of AI-dominated and knowledge-dominated methods in original work[74]

	Al	Co	Cr	Cu	Fe	Ni
No.1	43	22	23	0	7	5
No.2	47	20	18	5	5	5
No.3	43	22	22	0	8	5
No.4	47	19	19	5	5	5
No.5	43	24	22	0	5	6
No.6	43	25	22	0	5	5
No.7	43	24	23	5	5	5
No.8	43	18	20	5	12	7
No.9	43	23	21	5	8	5
No.10	47	14	20	5	9	5

Methods
Architecture
MLMD offers a complete materials design wrokflow, encompassing data collection, data preprocessing, feature engineering, model establishment, parameters optimization, material discovery, and experiment validation, as illustrated in Fig. 2. This process can be executed through a user-friendly interface. Various material data can be sourced from simulations, experiments, literature (manually collected data from published papers and patents), and open databases. In addition, MLMD provides material databases (e.g., polycrystalline ceramic, HEAs, ferroelectric perovskites). The feature engineering module involves descriptor refactoring, correlation

Fig. 6 | The atomic percentage distribution of novel $Al_xCo_yCr_zCu_uFe_vNi_w$ HEAs designed by Wen and MLMD. a–c The HEAs designed by Wen through iterations are represented by blue circles, while those designed by the EI strategy in active learning in MLMD are denoted by red Pentastar. **d–f** The HEAs designed by Wen through iterations are represented by blue circles, while those designed by the REI strategy in active learning in MLMD are denoted by red Pentastar. **g–i** The HEAs designed by Wen through iterations are represented by blue circles, while those designed by the UCB strategy in active learning in MLMD are denoted by red Pentastar.

Table 4 | The 9 newly designed alloys in one iteration by EI, REI, and UCB in MLMD

	Al	Co	Cr	Cu	Fe	Ni
EI	43	13	22	5	12	5
EI	43	15	20	5	12	5
EI	43	14	21	5	12	5
REI	43	17	24	0	11	5
REI	43	18	22	0	11	6
REI	43	17	23	0	11	6
UCB	43	22	25	0	5	5
UCB	43	21	26	0	5	5
UCB	43	20	27	0	5	5

analysis, and feature importance ranking. The classification and regression modules are used to predict the material properties, which integrate different AI models. The surrogate optimization module encompasses GA, PSO, DE, SA, and NSGA-II to accelerate the material discovery with single or multiple objectives. The active learning module offers various material design strategies such as EI, PI, AEI, UCB, and EHVI to direct experiments. The core elements and detailed architecture can be found in Fig. 1.

Data availability

More raw details and tutorials are also available from Supplementary Information. The program and source codes of the MLMD platform are available (https://github.com/Jiaxuan-Ma/MLMD).

References

1. Gnanasekaran, R. K., Shanmugam, B., Raja, V. & Kathiresan, S. Multi-disciplinary optimizations on flexural behavioural effects on various advanced aerospace materials: a validated investigation. *Mater. Plast.* **59**, 223–242 (2022).
2. Barile, C., Casavola, C. & De Cillis, F. Mechanical comparison of new composite materials for aerospace applications. *Compos. Part B: Eng.* **162**, 122–128 (2019).
3. Hench, L. L. & Polak, J. M. Third-generation biomedical materials. *Science* **295**, 1014–1017 (2002).
4. Zhang, G., Yi, Z., Cheng, G., Yang, W. & Yang, H. Polynitro-functionalized azopyrazole with high performance and low sensitivity as novel energetic materials. *ACS Appl. Mater. Interfaces* **14**, 10594–10604 (2022).
5. Pang, S.-Y. et al. Universal strategy for hf-free facile and rapid synthesis of two-dimensional mxenes as multifunctional energy materials. *J. Am. Chem. Soc.* **141**, 9610–9616 (2019).
6. Chen, C. et al. A critical review of machine learning of energy materials. *Adv. Energy Mater.* **10**, 1903242 (2020).
7. Hart, G. L. W., Mueller, T., Toher, C. & Curtarolo, S. Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).
8. Guo, K., Yang, Z., Yu, C.-H. & Buehler, M. J. Artificial intelligence and machine learning in design of mechanical materials. *Mater. Horiz.* **8**, 1153–1172 (2021).
9. Cao, B., Yang, S., Sun, A., Dong, Z. & Zhang, T.-Y. Domain knowledge-guided interpretive machine learning: Formula discovery for the oxidation behavior of ferritic-martensitic steels in supercritical water. *J. Mater. Inform.* **2**, 4 (2022).
10. Wang, Z.-L., Funada, T., Onda, T. & Chen, Z.-C. Knowledge extraction and performance improvement of bi2te3-based thermoelectric materials by machine learning. *Mater. Today Phys.* **31**, 100971 (2023).
11. Wu, P.-Y., Sandels, C., Johansson, T., Mangold, M. & Mjörnell, K. Machine learning models for the prediction of polychlorinated biphenyls and asbestos materials in buildings. *Resour. Conserv. Recycl.* **199**, 107253 (2023).
12. Zhang, Y. et al. Discovering the ultralow thermal conductive a2b2o7-type high-entropy oxides through the hybrid knowledge-assisted data-driven machine learning. *J. Mater. Sci. Technol.* **168**, 131–142 (2024).
13. Xiong, J., Shi, S.-Q. & Zhang, T.-Y. A machine-learning approach to predicting and understanding the properties of amorphous metallic alloys. *Mater. Des.* **187**, 108378 (2020).
14. Hyttinen, N., Pihlajamäki, A. & Häkkinen, H. Machine learning for predicting chemical potentials of multifunctional organic compounds in atmospherically relevant solutions. *J. Phys. Chem. Lett.* **13**, 9928–9933 (2022).
15. Vigneau, E., Courcoux, P., Symoneaux, R., Guérin, L. & Villière, A. Random forests: a machine learning methodology to highlight the volatile organic compounds involved in olfactory perception. *Food Qual. Prefer.* **68**, 135–145 (2018).
16. Mahmood, A. & Wang, J.-L. Machine learning for high performance organic solar cells: current scenario and future prospects. *Energy Environ. Sci.* **14**, 90–105 (2021).
17. Li, F. et al. Machine learning (ml)-assisted design and fabrication for solar cells. *Energy Environ. Mater.* **2**, 280–291 (2019).
18. Geng, X. et al. A hybrid machine learning model for predicting continuous cooling transformation diagrams in welding heat-affected zone of low alloy steels. *J. Mater. Sci. Technol.* **107**, 207–215 (2022).
19. Liu, H.-X. et al. Machine-learning-assisted discovery of empirical rule for inherent brittleness of full heusler alloys. *J. Mater. Sci. Technol.* **131**, 1–13 (2022).
20. Yang, C. et al. A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness. *Acta Mater.* **222**, 117431 (2022).
21. Xu, P. et al. Search for abo$_3$ type ferroelectric perovskites with targeted multi-properties by machine learning strategies. *J. Chem. Inf. Model.* **62**, 5038–5049 (2022).
22. Kim, C., Pilania, G. & Ramprasad, R. Machine learning assisted predictions of intrinsic dielectric breakdown strength of abx$_3$ perovskites. *J. Phys. Chem. C* **120**, 14575–14580 (2016).
23. Rao, Z. et al. Machine learning–enabled high-entropy alloy discovery. *Science* **378**, 78–85 (2022).
24. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
25. Talirz, L. et al. Materials cloud, a platform for open computational science. *Sci. Data* **7**, 299 (2020).
26. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
27. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
28. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
29. Gossett, E. et al. Aflow-ml: a restful api for machine-learning predictions of materials properties. *Comput. Mater. Sci.* **152**, 134–145 (2018).
30. Choudhary, K. et al. The joint automated repository for various integrated simulations (jarvis) for data-driven materials design. *npj. Comput. Mater.* **6**, 173 (2020).
31. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
32. Ward, L. A general-purpose machine learning framework for predicting. *npj. Mater.* **2**, 1602 (2016).
33. Zhao, X.-G. et al. Jamip: an artificial-intelligence aided data-driven infrastructure for computational materials informatics. *Sci. Bull.* **66**, 1973–1985 (2021).
34. Wang, G. et al. Alkemie: an intelligent computational platform for accelerating materials discovery and design. *Comput. Mater. Sci.* **186**, 110064 (2021).

35. 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).

36. Liu, Y., Wang, S., Yang, Z., Avdeev, M. & Shi, S. Auto-matregressor: liberating machine learning alchemists. *Sci. Bull.* **68**, 1259–1270 (2023).

37. Peng, J., Lee, S., Williams, A., Haynes, J. A. & Shin, D. Advanced data science toolkit for non-data scientists – a user guide. *Calphad* **68**, 101733 (2020).

38. Wang, Z. et al. Alphamat: A material informatics hub connecting data, features, models and applications. *npj Comput. Mater.* **9**, 130 (2023).

39. Karthikeyan, M. et al. Machine learning aided synthesis and screening of her catalyst: present developments and prospects. *Catal. Rev.* **0**, 1–31 (2022).

40. Wang, Y. et al. Accelerated design of fe-based soft magnetic materials using machine learning and stochastic optimization. *Acta Mater.* **194**, 144–155 (2020).

41. Feng, X., Wang, Z., Jiang, L., Zhao, F. & Zhang, Z. Simultaneous enhancement in mechanical and corrosion properties of al-mg-si alloys using machine learning. *J. Mater. Sci. Technol.* **167**, 1–13 (2023).

42. Jiang, L. et al. A rapid and effective method for alloy materials design via sample data transfer machine learning. *npj Comput. Mater.* **9**, 26 (2023).

43. Botella, R., Fernández-Catalá, J. & Cao, W. Experimental ni3teo6 synthesis condition exploration accelerated by active learning. *Mater. Lett.* **352**, 135070 (2023).

44. Chen, S., Cao, H., Ouyang, Q., Wu, X. & Qian, Q. Alds: An active learning method for multi-source materials data screening and materials design. *Mater. Des.* **223**, 111092 (2022).

45. Khatamsaz, D. et al. Multi-objective materials bayesian optimization with active learning of design constraints: Design of ductile refractory multi-principal-element alloys. *Acta Mater.* **236**, 118133 (2022).

46. Yuan, R. et al. Accelerated discovery of large electrostrains in batio$_3$-based piezoelectrics using active learning. *Adv. Mater.* **30**, 1702884 (2018).

47. Li, H. et al. Towards high entropy alloy with enhanced strength and ductility using domain knowledge constrained active learning. *Mater. Des.* **223**, 111186 (2022).

48. Khatamsaz, D. et al. Bayesian optimization with active learning of design constraints using an entropy-based approach. *npj Comput Mater.* **9**, 49 (2023).

49. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

50. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. KDD '16, 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). https://doi.org/10.1145/2939672.2939785.

51. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features.

52. Blank, J. & Deb, K. pymoo: multi-objective optimization in python. *IEEE Access* **8**, 89497–89509 (2020).

53. Sivaraj, R. & Ravichandran, T. A review of selection methods in genetic algorithm. *Int. J. Eng. Sci. Technol.* **3**, 3792–3797 (2011).

54. Pant, M., Zaheer, H., Garcia-Hernandez, L. & Abraham, A. Differential evolution: a review of more than two decades of research. *Eng. Appl. Artif. Intell.* **90**, 103479 (2020).

55. Kennedy, J. & Eberhart, R. Particle swarm optimization. *In: Icnn95-international Conference on Neural Networks* (1995).

56. Kirkpatrick, S., Gelatt, C. D., Vecchi, M. P. Optimization by simulated annealing. *Science* (1983).

57. Picheny, V., Wagner, T. & Ginsbourger, D. A benchmark of kriging-based infill criteria for noisy optimization. *Struct. Multidiscip. Optim.* **48**, 607–626 (2013).

58. Hernandez-Lobato, J. M., Hoffman, M. W. & Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions.

59. Kushner, H. J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *J. Basic Eng.* **86**, 97–106 (1964).

60. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).

61. Frazier, P., Powell, W. & Dayanik, S. The knowledge-gradient policy for correlated normal beliefs. *In*: INFORMS journal on computing 21 (2009).

62. Couckuyt, I., Deschrijver, D. & Dhaene, T. Fast calculation of multiobjective probability of improvement and expected improvement criteria for pareto optimization. *J. Glob. Optim.* **60**, 575–594 (2014).

63. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *In:* Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 4768–4777 (Curran Associates Inc., Red Hook, NY, USA, 2017).

64. Fang-Ming, B., Wei-Kui, W. & Long, C. Dbscan: density-based spatial clustering of applications with noise. *J. Nanjing Univ.(Nat. Sci.)* **48**, 491–498 (2012).

65. Liu, F. T., Ting, K. M. & Zhou, Z.-H. Isolation forest. *In:* 2008 Eighth IEEE International Conference on Data Mining, 413–422 (2008).

66. Breunig, M. M., Kriegel, H. P., Ng, R. T. & Sander, J. Lof: Identifying density-based local outliers. *In:* Acm Sigmod International Conference on Management of Data (2000).

67. Schlkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J. & Platt, J. C. Support vector method for novelty detection. *In:* Advances in Neural Information Processing Systems 12, NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999 (1999).

68. Kano, S. et al. Precipitation of carbides in f82h steels and its impact on mechanical strength. *Nucl. Mater. Energy* **9**, 331–337 (2016).

69. Williams, C. A., Hyde, J. M., Smith, G. D. & Marquis, E. A. Effects of heavy-ion irradiation on solute segregation to dislocations in oxide-dispersion-strengthened eurofer 97 steel. *J. Nucl. Mater.* **412**, 100–105 (2011).

70. Li, X., Zheng, M., Yang, X., Chen, P. & Ding, W. A property-oriented design strategy of high-strength ductile rafm steels based on machine learning. *Mater. Sci. Eng.: A* **840**, 142891 (2022).

71. Qiu, G. et al. Effects of y and ti addition on microstructure stability and tensile properties of reduced activation ferritic/martensitic steel. *Nucl. Eng. Technol.* **51**, 1365–1372 (2019).

72. Gludovatz, B. et al. A fracture-resistant high-entropy alloy for cryogenic applications. *Science* **345**, 1153–1158 (2014).

73. Chou, Y., Wang, Y., Yeh, J. & Shih, H. Pitting corrosion of the high-entropy alloy co1.5crfeni1.5ti0.5mo0.1 in chloride-containing sulphate solutions. *Corros. Sci.* **52**, 3481–3491 (2010).

74. Wen, C. et al. Machine learning assisted design of high entropy alloys with desired property. *Acta Mater.* **170**, 109–117 (2019).

75. Balachandran, P. V., Kowalski, B., Sehirlioglu, A. & Lookman, T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **9**, 1668 (2018).

76. Xiong, J., Shi, S.-Q. & Zhang, T.-Y. Machine learning of phases and mechanical properties in complex concentrated alloys. *J. Mater. Sci. Technol.* **87**, 133–142 (2021).

77. Xiong, J., Zhang, T. & Shi, S. Machine learning of mechanical properties of steels. *Sci. China Technol. Sci.* **63**, 1247–1255 (2020).

78. Xiong, J., He, J.-C., Leng, X.-S. & Zhang, T.-Y. Gaussian process regressions on hot deformation behaviors of fgh98 nickel-based powder superalloy. *J. Mater. Sci. Technol.* **146**, 177–185 (2023).

Acknowledgements

Author contributions

Jiaxuan Ma designed the MLMD workflow, developed the MLMD, and wrote the paper. Bin Cao developed the Bgolearn toolkit, and edited the paper. Jiaxuan Ma and Bin Cao contributed equally to this work. Shuya Dong plotted the figures, edited the paper. Yuan Tian provided guidance on the multiobjective Bayesian optimization. Menghuan Wang tested MLMD, and provided some suggestions. Jie Xiong tested the MLMD, wrote the paper, and supervised the overall work. Sheng Sun designed the MLMD workflow, edited the paper, secured funding, and supervised the overall work.

Competing interests

The authors declare that they have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at
https://doi.org/10.1038/s41524-024-01243-4.

Correspondence and requests for materials should be addressed to Jie Xiong or Sheng Sun.

Reprints and permissions 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.