REVIEW

National Science Review
10: nwad125, 2023
https://doi.org/10.1093/nsr/nwad125
Advance access publication 1 May 2023

MATERIALS SCIENCE

Data quantity governance for machine learning in materials science

Yue Liu[1,4], Zhengwei Yang[1], Xinxin Zou[1], Shuchang Ma[1], Dahui Liu[1], Maxim Avdeev[5,6] and Siqi Shi[2,3,*]

[1]School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; [2]State Key Laboratory of Advanced Special Steel, School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China; [3]Materials Genome Institute, Shanghai University, Shanghai 200444, China; [4]Shanghai Engineering Research Center of Intelligent Computing System, Shanghai 200444, China; [5]Australian Nuclear Science and Technology Organisation, Sydney 2232, Australia and [6]School of Chemistry, The University of Sydney, Sydney 2006, Australia

*Corresponding author. E-mail: sqshi@shu.edu.cn

Received 27 February 2023; **Revised** 14 April 2023; **Accepted** 26 April 2023

ABSTRACT

Data-driven machine learning (ML) is widely employed in the analysis of materials structure–activity relationships, performance optimization and materials design due to its superior ability to reveal latent data patterns and make accurate prediction. However, because of the laborious process of materials data acquisition, ML models encounter the issue of the mismatch between a high dimension of feature space and a small sample size (for traditional ML models) or the mismatch between model parameters and sample size (for deep-learning models), usually resulting in terrible performance. Here, we review the efforts for tackling this issue via feature reduction, sample augmentation and specific ML approaches, and show that the balance between the number of samples and features or model parameters should attract great attention during data quantity governance. Following this, we propose a synergistic data quantity governance flow with the incorporation of materials domain knowledge. After summarizing the approaches to incorporating materials domain knowledge into the process of ML, we provide examples of incorporating domain knowledge into governance schemes to demonstrate the advantages of the approach and applications. The work paves the way for obtaining the required high-quality data to accelerate materials design and discovery based on ML.

Keywords: machine learning, data governance, data quantity, materials science

INTRODUCTION

Data-driven machine learning (ML) is widely employed in development of novel materials due to its ability to quickly, accurately and cheaply reveal 'composition–structure–process–property' relationships [1]. Note that its performance depends heavily on the quality and quantity of the input data (i.e. samples). Because obtaining material samples relies on tedious experiments or labor-intensive acquisition, the sample size is commonly small. On the other hand, materials experts usually select multiple descriptors (i.e. features) to study complex structure–activity relationships, resulting in a high dimension of the feature space of samples. Hence, keeping the balance between the number of features and the sample size is vital for the discovery and design of novel materials using data-driven ML. Note that the above balance mainly concentrates on the research employing traditional ML models because this type of model has limited ability for data analysis and is profoundly affected by the number of features and the size of the samples. When it occurs in deep-learning (DL) models, thanks to their excellent ability for information extraction, the major concern will transfer to the balance between the sample size and the model scales (i.e. the number of model parameters).

Herein, we summarize 107 papers in materials science, containing 109 data sets, to illustrate the statistics on the data-set size, as shown in Supplementary Fig. S1. The size of ∼57% of the data sets is <500, ∼67% of data sets comprise <1000 samples and only ∼21% of data sets contain >2000 samples. Next, we analyse the ratio of the number of features to the sample size. In ∼35.8% of the data sets, the ratio exceeds 1/4 and for ∼8% of the data sets it exceeds >1. This indicates that in most cases, even though the sample size is four times larger than the number of features, the issue of

'high dimensionality of feature space vs. small data set size' is obviously encountered in ML applications for materials science. Meanwhile, as the model scale increases, the demand for training data will grow exponentially. Hence, the balance between the sample size and the number of model parameters should also be valued, for which the solutions fall into data augmentation, learning algorithm modification and the improvement of data representation patterns.

To alleviate the problem of the poor ratios of the feature space dimensionality to the data-set size, researchers usually focus on eliminating redundant features, e.g. feature selection [2] and dimensionality reduction [3]. This facilitates keeping a proper balance between the number of features and samples so that ML models can easily mine latent patterns in the context of small samples [4]. However, these methods mainly rely on statistical approaches, namely they cannot directly assess which features are most relevant. For example, the single-variable correlation analysis method can only identify features that are highly correlated with each other. However, it cannot determine which features should be retained due to the coupling of multiple features of the structure–activity relationships. Therefore, incorporating materials domain knowledge into the methods becomes necessary for the guidance of feature selection.

In addition, focusing on governing the number of samples can also be an effective approach to keeping a proper balance with the number of features or model parameters. This can be categorized into sample-oriented and model-oriented methods. The sample-oriented methods aim to augment or select samples to facilitate ML modeling, e.g. generative adversarial network (GAN) [5], auto-encoder (AE) [6] and active learning [7]. The model-oriented methods aim to change the learning patterns of the model that can capture the latent information in a few-shot context, e.g. ensemble learning [8], transfer learning (TL) [9].

Though the research related to data quantity governance is beginning to be conducted, there is still a lack of governance flow to systematically and comprehensively improve the volume of data in materials science. This study aims to introduce and review common data quantity governance methods for materials science dealing with the balance between data-set size and feature space dimensionality (or the number of model parameters), including feature reduction, sample augmentation and specific ML approaches. We also show that incorporating materials domain knowledge into the learning process allows the construction of ML models with explainability, robustness and generalization, and illustrates corresponding incorporating approaches. Following this, we propose a synergistic data quantity governance flow with the incorporation of materials domain knowledge, aiming to govern the data quantity in an accurate and explainable way. The examples are provided on the integration of domain knowledge into the data-driven ML model to further improve reliability and prediction accuracy.

DATA QUANTITY GOVERNANCE METHODS IN MATERIALS SCIENCE

In this section, the data quantity governance methods employed in materials science are detailed (shown in Fig. 1), separately focusing on sample and feature quantity governance.

Feature quantity governance

Feature definition is one of the foundations for ML modeling. However, defining a large number of features is not guaranteed to improve the performance of ML models. That is, in high-dimension feature space, redundant features with high correlation may produce a negative impact on the prediction performance and explainability of the ML model. Therefore, it is a critical issue to effectively govern features to reduce the feature quantity (i.e. the number of dimensions of feature space) while at the same time enabling the model to mine general patterns latent in the materials data. The common approaches focus on feature selection (FS) and feature transform (FT).

FS

FS is a critical preprocessing step to identify and rank the most relevant features, and can be generally categorized into filter, embedded, wrapper and hybrid methods. Here, we illustrate examples of the research on FS in materials science in Table 1.

Filter methods select the best features from a rank of the whole set of features by evaluating their intrinsic characteristics. Im *et al.* [10] employed the Pearson correlation coefficient (PCC) to select and confirm the 20 most important features for the bandgap. This enabled the root mean square error (RMSE) of the gradient-boosted regression tree to reach as low as 0.322. Note that Deng [11] employed PCC and the Spearman correlation coefficient (SCC) to explore the correlation between candidate descriptors and target property, whose results show that the valence state (descriptor) and lattice constants (target property) in cubic perovskite oxides had no correlation and thus the former were eliminated from the model. Agrawal *et al.* [12] leveraged an information-entropy-based (IE) metric to measure the features

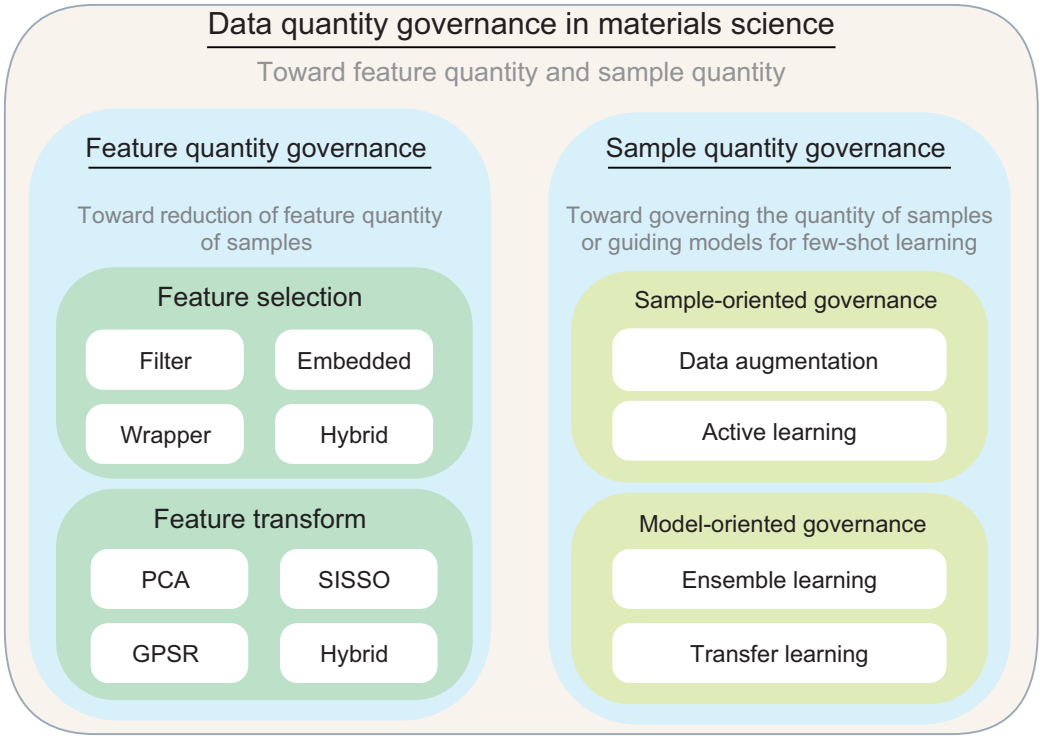

Figure 1. The main aspects of data quantity governance.

Table 1. Examples of feature selection.

Types	Methods	Materials	Number of raw features	Number of governed features	Number of samples	Results	Reference
Filter	PCC	Pb-free perovskites	32	20	540	RMSE: 0.322	[10]
	PCC and SCC	Cubic perovskite oxides	5	3	1376	R^2: 0.87	[11]
	IE	Steels	25	–	437	R^2: 0.98	[12]
	MIC	High-temperature alloys	466	21	166	Accuracy: >90%	[13]
Embedded	L1-Norm	Carbon fiber	6	4	500	RMSE/AV: 2.55%	[16]
	GBR	Nickel-based single-crystal superalloys	21	14	94	R^2: 0.90	[17]
	RF	High-temperature superconductors	145	5	∼16 400	Accuracy: ∼90%	[18]
	LASSO	Single-atom catalyst	333 932	75	91	RMSE: 0.4096	[19]
	FeaLect	FCC and HCP	34	10	–	AUC: 80.75	[20]
Wrapper	ANN	FCC solute	111	14	218	RMSE: 0.092	[22]
	RF	Crystalline materials	452	187	952	R^2: 0.92	[23]
	Cluster	Equiatomic ternary phases ABC	990	113	1556	Accuracy: 96.9%	[24]
	SVR	ABO_3 perovskites	15	6	128	RMSE: 1.08	[25]
	LR	Solid lithium-ion conductor materials	20	5	40	–	[26]
Hybrid	CA and GB	High-entropy alloys	54	3	162	–	[28]
	PCC and SC and LASSO	Chalcogenides	23 454	85	119	R^2: 0.9408	[29]
	PCC and LASSO	Functionalized MXene	47	8	70	RMSE: 0.14 eV	[30]
	Wrapper and PSO	Coal	27 620	114	840	R^2: 0.99	[31]

'–' indicates that this term is not mentioned. FCC, face-centered cubic materials; HCP, hexagonal close-packed materials; RMSE, root mean square error between the predicted value and actual value; AV, average of the actual value; RMSE/AV, a ratio that shows the overall deviation degree of the predicted sample; AUC, area under the receiver operating characteristic curve (ROC), a metric for binary classification.

with respect to the target variable and selected the proper features for modeling. The results show that, with the favor of it, the multivariate polynomial regression obtained the best prediction performance ($R^2 = 0.98$). Shin *et al.* [13] employed the maximal information coefficient (MIC) to calculate the correlations between descriptors, then 21 descriptors were selected from 466 candidates. The results show that better ML performances can be obtained based on MIC than those based on PCC, whose accuracy reaches >90%.

Compared with filter methods, embedded methods perform FS during the process of model construction [14,15] and greatly simplify the process of FS, which is limited to specific ML models (decision-tree-based and linear models). Qi *et al.* [16] employed the L1 norm method for FS to avoid overfitting in the regression tree model for the prediction of the carbon fiber mechanical properties. Zeng *et al.* [17] utilized gradient boosting regression to quantify the feature importance, then 14 features were chosen out of 21 candidate features, whose regression model achieved high accuracy ($R^2 = 0.90$). Stanev *et al.* [18] employed random forest (RF) to quantify the feature importance, whose results show that RF using only the five most informative features can still achieve almost 90% accuracy. O'Connor *et al.* [19] utilized least absolute shrinkage and selection operator (LASSO) regression to select 75 descriptors for the accurate prediction of binding energy from 333 932 candidate features, achieving an RMSE of 0.4096. Mangal *et al.* [20] employed several FS methods to determine which microstructural characteristics can cause stress to build up in certain grains during uniaxial tensile deformation. The results show that the FeaLect algorithm, an improvement over the LASSO algorithm, was the most suitable and enabled feature rankings for physical interpretations.

Wrapper methods are model-dependent as well and are generally used in combination with a specific ML model and a meta-heuristic algorithm to identify the best feature subset without sacrificing prediction accuracy [21]. This method can be used with any ML model and can find the feature subset that enables the ML model to achieve optimal or near-optimal prediction performance in a huge feature space through search techniques. Wu *et al.* [22] employed different ML models to investigate the best combinations of descriptors for face-centered cubic solute diffusion predictions. The results show that artificial neural networks (ANNs) gained the lowest RMSE (0.092), selecting 14 optimal descriptors from 111 candidate descriptors. Furmanchuk *et al.* [23] employed RF to select the proper number of descriptors from 452 candidate descriptors. The results show that the RMSE of RF in the context of the 187 selected descriptors (0.92) surpassed that of RF based on all the descriptors (0.81). Oliynyk *et al.* [24] utilized a cluster resolution FS to select the best combination of variables in conjunction with a support vector machine (SVM). The results show that only 113 features out of 990 were sufficient to gain optimal classification performance of the SVM (Accuracy = 96.9%). Xu *et al.* [25] utilized three ML models to explore the best subset of features from 15 candidate features, whose results show that support vector regression (SVR) was able to achieve the lowest RMSE (1.08) tested on the data set with six features. Sendek *et al.* [26] performed FS via logistic regression (LR) from 20 candidate features. As a result, a five-feature model with a minimal cross-validated misclassification rate was constructed to screen out 21 structures that showed promise as electrolytes from 12 831 candidate materials.

The hybrid methods have the advantages of both the filter and the wrapper methods [27]. The filter approach is first employed to identify the best relevant features of the data sets. Then, the wrapper method is employed to verify the previously identified relevant feature subsets by using a method that gives higher classification accuracy rates. Wen *et al.* [28] employed a hybrid method combining correlation analysis and a gradient boosting algorithm to remove 14 features from the original pool of 54 features. Then, physics-based models and several ML models were employed to perform FS so that three features were determined to build a new model for the accurate prediction of solid solution strengthening. Wang *et al.* [29] employed hybrid methods combining PCC, Spearman correlation (SC) and LASSO to select 85 descriptors from 23 454 candidates and developed a stacking regression model (GBDT) using 119 compounds that efficiently predicted the band gaps of diamond-type-structure chalcogenides. Rajan *et al.* [30] employed PCC to quantify the feature correlation, then LASSO was used to further reduce the feature quantity down to eight features, which yielded a Gaussian process regression (GPR) with the lowest RMSE of 0.14 eV. Yan *et al.* [31] proposed a hybrid FS approach, combined with wrapper and particle swarm optimization (PSO), to select an optimal subset to model Laser-induced breakdown spectroscopy data. The results show that this method could select the optimal feature subset (114) from the original 27 620 candidates.

FT

FT generates new features via mapping original features into a space with lower dimensions,

Table 2. Examples of feature transform.

Methods	Materials	Number of raw features	Number of governed features	Number of samples	Results	Reference
PCA	Battery	5	1	34	Accuracy: 99.7%	[34]
	Binary metallic alloys	114	9	55	RMSE: 50 meV	[35]
SISSO	Binary systems	6	2	299	Accuracy: ∼99%	[36]
	Perovskite oxides and halides	9	2	576	Accuracy: 91%	[37]
	Catalytic materials	18	8	211	RMSE: 0.18	[38]
	Inorganic crystalline solids	19	3	309	RMSD: 61 meV/atom	[39]
GPSR	Oxide perovskite catalysts	1808	1	∼8640	–	[40]
	MXene materials	25	16	25	–	[41]

'–' indicates that this item is not mentioned. RMSD, root mean square deviation.

i.e. dimensionality reduction. Dimensionality reduction originates from the latent semantic index model [32]. It aims to transform feature values into a certain pattern, change the spatial relations of original features and obtain new features by analysing the relations between the original ones. Examples of FT in materials science are presented in Table 2.

Principal component analysis (PCA) [33] is an unsupervised linear dimensionality reduction method that employs covariance as a measure to remove noise and redundant features to the greatest extent. Sturlaugson *et al.* [34] transformed five initial features into one feature via PCA to simplify the Bayesian network (BN) model of diagnostics on lithium-ion batteries. The results show that the average accuracy of the BN model with PCA achieved $\geq$10% improvement over the model without PCA. Curtarolo *et al.* [35] employed PCA to transfer 114 features into nine dimensions, which enabled partial least squares regression to describe the energies with a RMSE of 50 meV.

Sure independence screening (SIS) based on correlation learning is effective for the dimensionality reduction of ultra-high-dimensional feature spaces, which scores each feature with a correlation magnitude and keeps only the top-ranked. The SIS and sparsifying operator method called SISSO autonomously finds the optimal N-dimensional descriptor through sparsifying operators after SIS. Ouyang *et al.* [36] utilized SISSO to construct a 2D descriptor from a six-dimension descriptor, which enabled SVM to classify metal and nonmetal materials with a training accuracy of ∼99.0%. Bartel *et al.* [37] used SISSO to obtain a tolerance factor descriptor to predict the stability of perovskites by using only the atomic oxidation states and ionic radius, achieving an overall accuracy of 91%. Andersen *et al.* [38] employed SISSO for descriptor identification and obtained 8 of the most proper features from 18 candidates. The results show that SISSO with eight descriptors achieved the lowest RMSE (0.18 eV). Bartel *et al.* [39] used SISSO to find a physical descriptor for the inorganic crystalline solids Gibbs energy, of which the simple descriptor based only on the atomic volume, reduced mass and temperature reached a root mean square deviation of 61 meV/atom.

In addition, generating new features by employing basic mathematical functions alongside the original features can be regarded as an FT approach as well, such as genetic programming-based symbolic regression (GPSR). Weng *et al.* [40] generated effective descriptors using GPSR, which identified the combination of the tolerance factor (t) and the octahedral factor (μ), μ/t, as a primary descriptor for predicting the oxygen evolution reaction activity of oxide perovskites quantitatively. He *et al.* [41] employed symbolic regression to select a new relevant descriptor for a fast and efficient method to perform the classification of MXene materials and design new descriptors.

Sample quantity governance

There have been some efforts towards data accumulation to build large databases to capture results of high-throughput computation and high-throughput experiments. However, in many materials studies, especially of novel materials, the problem of insufficient data to construct reliable ML models is still often faced [42]. To this end, sample quantity governance is utilized. This type of method focuses on treating original samples (i.e. sample-oriented method) or modifying the learning algorithms that enable ML models to extract useful information from small samples (i.e. model-oriented method).

Sample-oriented method

Data augmentation. Data augmentation commonly refers to data expansion of the original small sample data set with the help of auxiliary data or information [43]. Typically, researchers add Gaussian or

Table 3. Examples of data augmentation.

Methods	Fields	Number of features	Number of raw samples	Number of governed samples	Results	Reference
GAN	2D materials	8*85	291 840	2 650 264	AUC: 0.96	[45]
	Inorganic materials	8*85	251 368 (OQMD) 57 530 (MP)25 323 (ICSD)	1 831 648 (OQMD) 1 969 633(MP) 1 983 231 (ICSD)	–	[46]
	Polycrystalline materials	400*400	47	136	MAP: 0.586	[47]
VAE	Inorganic materials	32*32*32	10 981 (MP)	∼20 000 (MP)	–	[48]
	3D molecules	30*30*30	46 744	–	–	[49]

'–' indicates that this item is not mentioned. AUC, area under the receiver operating characteristic curve (ROC), a metric for binary classification; MAP, mean average precision; OQMD, Open Quantum Materials Database; MP, Materials Project platform; ICSD, Inorganic Crystal Structure Database.

impulse noise into the samples to increase the diversity of the initial samples [44]. However, this can limit the diversity of the augmented samples with numerous samples without any practical value generated. With the development of DL, neural-network-based methods such as the variational auto-encoder (VAE) and GAN have been successfully applied to data augmentation and have achieved better performance in materials science, examples of which are summarized in Table 3. Thereunto, Song *et al.* [45] employed GAN to generate 2.65 million 2D material samples, then discovered 26 489 new potential 2D materials, among which 1 485 2D materials had an area under the curve (AUC) of >0.95. Dan *et al.* [46] employed GAN to generate novel hypothetical inorganic materials that are not recorded in existing databases to enable the inverse design of such materials. Ma *et al.* [47] utilized GAN to expand the original 47 images of polycrystalline iron to 136, of which the model performance on the data sets composed of generated data and 35% real data were comparable to those on all real data. This demonstrates the feasibility of combining generated data with real experimental data. As for VAE, Noh *et al.* [48] employed VAE to perform an inverse design of solid-state materials. The results show that ∼20 000 hypothetical solid-state materials are generated via VAE, leading to several completely new metastable V_xO_y materials that may be synthesizable. Hoffmann *et al.* [49] utilized VAE trained on >120 000 3D positions of atoms in a molecule to generate new materials.

Although the data-driven methods achieved impressive results, they have limitations. For example, the VAE-based methods compress the input data via an encoder to extract the concept of data. Then, the decoder restores the input data according to the concept. In this process, the output data are slightly different from the initial data and can be regarded as new data. However, the augmented data may drop details during the process of encoding because the concept only contains the most representative eigenvectors. Moreover, the VAE-based method fails to learn the physical information from the materials data, which is also a common drawback of the neural-network-based method. GAN-based methods generate data from Gaussian noise via a generator, then the discriminator distinguishes the generated data and initial data. By this means, GAN can learn the complicated rules from the original data, then apply the learned rules to generate new samples with the target properties. Note that the obvious limitation of GAN is that only a Gaussian distribution can be leveraged to mimic the initial data. However, in reality, the materials data are complicated so the generation process may not be valid for a given material system. Moreover, differently from VAE-based methods that generate data according to the feature vector of the data, the data generation of GAN is from zero (noise) to one (valid data), which may result in low robustness.

Active learning. Active learning is a method that allows an efficient iterative search in the search space to identify candidate objects. In materials science, active learning employs pre-built ML models to sample candidate chemistry spaces iteratively and adaptively. It provides the most valuable candidate samples for costly calculations or experimental validation to accelerate the screening for novel high-performance materials. Therefore, for cases in which the data set is sample-scarce or difficult to generate, it can effectively overcome the problem of poor predictive power due to the limited number of samples, and thus effectively explore materials with the target conditional properties [50]. Examples of using active learning in materials science are shown in Table 4. Min *et al.* [51] employed active learning to screen inorganic ABO_3 perovskite materials. The results show that through active learning with 30% of the whole data set (5218), >90% of the screening

Table 4. Examples of active learning.

Fields	Number of features	Number of candidate samples	Number of selected samples	Results	Reference
Inorganic perovskite	159	5218	79	–	[51]
Ni Ti-based shape memory alloy	4	256	15	R^2: 0.85	[52]
Homobenzylic ether	30	112 000	42	–	[53]
Layered material	12	500	–	–	[50]

'–' indicates that this item is not mentioned.

Table 5. Examples of ensemble learning.

Fields	Number of features	Number of samples	Results	Reference
Hydroxyapatite	2	900	R^2: 0.9983	[58]
High-performance concrete	9	1030	R^2: 0.92	[59]
Strip steel	22	3534	MAP: 0.0252	[60]
Strip steel	11	2278	AUC: 0.92	[61]

AUC, area under the receiver operating characteristic curve (ROC), a metric for binary classification; MAP, mean average precision.

materials were satisfactory. Pruksawan *et al.* [52] employed active learning to select 15 new samples from 256 possible experimental conditions. The results show that, compared with the initial data set, the R^2 of the gradient boosting model had an improvement of 25% (0.85). Doan *et al.* [53] employed active learning based on Bayesian optimization (BO) to perform the efficient identification of the desired homobenzylic ethers (HBEs), which screened 42 optimal HBE candidates from an unseen data set of 112 000 HBEs. Bassman *et al.* [50] used BO with a surrogate GPR from 500 candidate samples to find layered materials with desired properties.

As active learning is performed based on the BO algorithm, its performance is determined by using surrogate models and utility functions of BO. Moreover, as per the 'no free lunch theorem' [54], there is no combination of surrogate models and utility functions that can be equally suitable for all materials research and thus additional optimization of the method is required for particular models. Each iteration of BO needs to update the probabilistic agent model. Therefore, updating the probabilistic model is computationally expensive and often cannot be employed for practical tasks requiring real-time operation when the problem dimension is high or there is a large amount of data. Note that in the case of a high-dimension search space, BO has low effective dimensionality, namely there are only a few dimensions that determine the objective function while the rest have little or no influence.

Model-oriented method

Ensemble learning. Ensemble learning [8] is an algorithm that combines base classifiers to achieve better prediction performance. The main idea is to learn by cascading multiple weak learners and combining these test results together according to a certain strategy, which can obtain a better pattern recognition effect than a single classifier. According to the types of generation methods of individual learners, ensemble learning can be divided into the serial ensemble (e.g. boosting [55]) and parallel ensemble [56]. The boosting algorithm is prone to overfitting in the case of training data with noise. The bagging algorithm reduces variance by averaging the results of multiple models, which can effectively alleviate the overfitting problem. Therefore, bagging is commonly employed in materials science. Bagging utilizes the bootstrap algorithm [57] to realize the operation of sampling. The bootstrap algorithm belongs to the put-back sampling method aiming to obtain the distribution of statistics and confidence intervals. RF is built based on the decision-tree-based learner for bagging integration and random attribute selection is introduced in its training process. This can effectively achieve dimensionality reduction of the features while obtaining higher accuracy.

Here, we summarize and analyse the research on ensemble learning in materials science (shown in Table 5). Thereunto, Okafor *et al.* [58] compared different ensemble learning approaches (RF and XGBoost) in the study of the transmission prediction of infrared radiation from hydroxyapatite samples. The results show that RF outperformed XGBoost, but its computational cost was higher. Farooq *et al.* [59] proposed the use of the ensemble learning approach to predict the strength of high-performance concrete prepared from waste materials. This study compared several ensemble learning

Table 6. Examples of transfer learning.

Fields	Number of features	Number of samples in source domain	Number of samples in target domain	Results	Reference
Materials science	86	321 140 (OQMD)	28 171 (JARVIS)	MAE: 0.0708	[63]
Tactile material	1000*16	30 classes (100 samples per class)	6 classes (100 samples per class)	Accuracy: 90.3%	[64]
Non-ferrous metal scrap	416*416	COCO data set [100]	920 multi-target image samples	Accuracy: 95.3%	[65]
Semiconductor material	–	1439 PBE band gaps	64 HSE06 band gaps	R^2: 0.98	[66]
Gas adsorption material	5	13 506 MOFs of H_2	1000 MOFs of CH_4	Accuracy: 99.1%	[67]

'–' indicates that this item is not mentioned. OQMD, Open Quantum Materials Database; JARVIS, Joint Automated Repository for Various Integrated Simulations; PBE, Perdew–Burke–Ernzerhof; HSE06, Heyd–Scuseria–Ernzerhof; MOFs, metal—organic frameworks; MAE, mean absolute error.

methods and individual models, of which the results show that the utilization of bagging- and boosting-based algorithms can improve the responsiveness of individual models, and the RF- and bagging-based models have better robustness. Yang *et al.* [60] investigated the problem of screening the influencing factors of steel mechanical properties via RF, which gains a high accuracy of an average absolute percentage error of 2.52% and a RMSE of 21.65 MPa. Ji *et al.* [61] established a hot-rolled strip steel product defect prediction model based on the ensemble learning method (improved RF) to extract the key process parameters affecting the product quality. The results show that the prediction of defects in hot-rolled strip steel is significantly improved for the improved RF compared with a single RF method.

TL. TL is the application of knowledge learned in solving one problem (source domain) to a different but related problem (target domain). This algorithm aims to enable the model to obtain a better learning effect in new tasks [62]. Note that the precondition for TL to succeed is that the source domain is relevant to the target domain. In this process, the knowledge (namely parameters of ML models and prior distribution of source data) is transferred into models. According to the transferred knowledge, models can learn the latent patterns and gain faster convergence, higher robustness and high accuracy from small new samples. Examples of TL are presented in Table 6. Gupta *et al.* [63] transferred the deep knowledge representation from the abundant streaming potential data (from OQMD) to efficiently guide the design of materials with limited experimental data (from JARVIS). The results show that the TL model had the lowest mean absolute error (MAE) (0.0708) for the prediction of formation energy, which reduced the MAE by 26% compared with the model trained from scratch (0.0964). Bäuml *et al.* [64] used TL to classify the tactile material. The results show that in 10-shot learning the accuracy of the deep model reached 90.3%, outperforming classification without TL by >40%. Chen *et al.* [65] utilized TL on YOLOv3, a DL model for multi-target detection, to eliminate the overfitting. The results show that TL models gained 95.3% accuracy, which was ∼5% higher than the model trained from scratch. Wang *et al.* [66] trained a DL model to accurately predict HSE06 band gaps. The results show than the prediction performance of the TL model ($R^2 = 0.98$) outperformed that of the ordinary DL model ($R^2 = 0.89$). Ma *et al.* [67] transferred knowledge from the source task to H_2 adsorption at 100 bar and 130 K (one target task) via TL, which enabled the average predictive accuracy on the target tasks to be improved from 0.960 (direct training) to 0.991 (TL).

In conclusion, the efforts above provide an accessible approach to governing the data quantity by reducing the number of features, augmenting the number of samples or leveraging specific learning mechanisms. However, the governance results are limited by the one-side pattern. Hence, how to synergistically govern the data quantity is a more important issue to be considered.

A SYNERGISTIC DATA QUANTITY GOVERNANCE FLOW WITH INCORPORATION OF MATERIALS DOMAIN KNOWLEDGE

The application of data quantity governance methods to reconcile the contradiction between small data and high dimension has been a hot research topic in computer science [43] but so far has received little attention in materials science. In addition, the existing governance methods are purely data-driven, which often leads to contradictions between ML model prediction results and materials domain knowledge. To resolve these contradictions, the framework of Machine Learning Embedded with Materials Domain Knowledge [68] (Fig. 2) is proposed. In particular, this framework aims to perform symbolic representation of the materials domain knowledge and then embed it into the three key elements of ML (i.e. Model, Strategy and Algorithm). In this way, materials domain knowledge can be effectively embedded into the whole process of ML so that new ML models with high

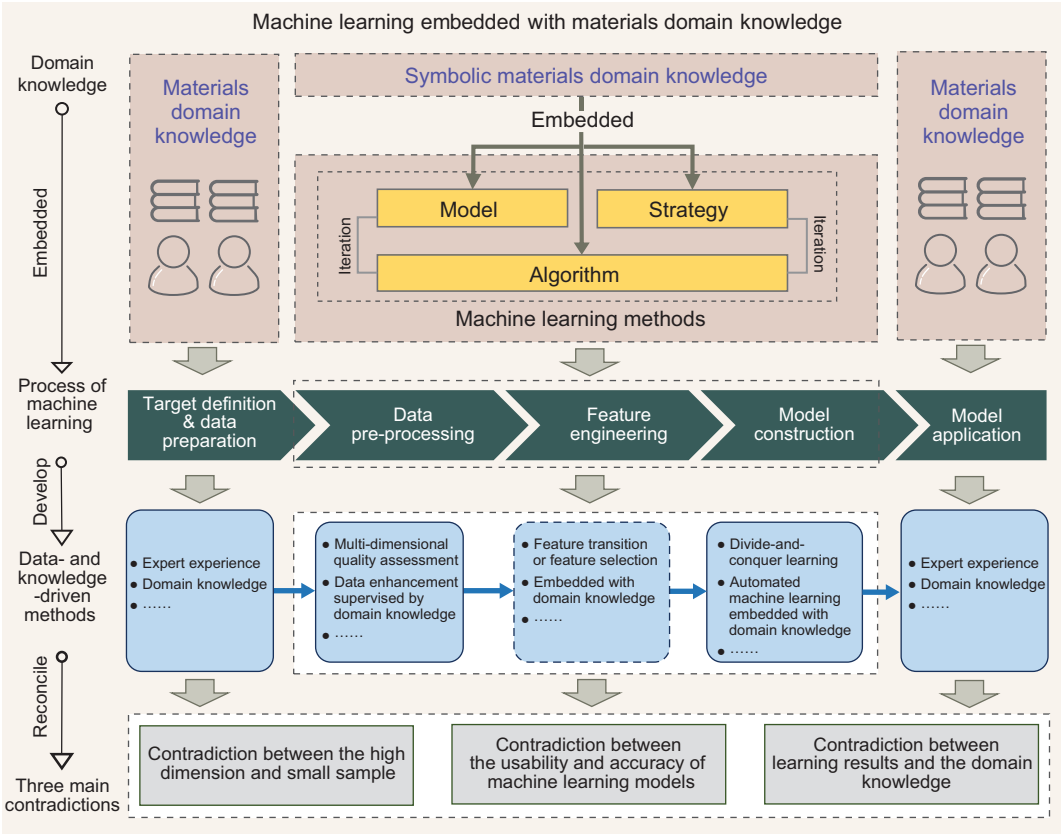

Figure 2. Framework of machine learning embedded with materials domain knowledge [68].

accuracy and explainability can be constructed and a high-coupled synergistic state can be realized in the paradigm of the 'AI for Science (AI4Science)' [69,70] in materials science. Based on this framework, we propose a general synergistic data quantity governance flow with the incorporation of materials domain knowledge, focusing on integrating materials domain knowledge into both feature and sample quantity governance, to ensure the interpretability, reliability and prediction accuracy of the resulting model.

Domain-knowledge acquisition and representation

In conventional ML, materials domain knowledge mainly participates in data preprocessing or feature engineering. This method of participation is deeply intertwined with the learning process and thus cannot be employed as an independent source or through separated representations but is rather employed with adaption and as required. Nowadays, many studies focus on mining materials domain knowledge from materials science literature [71], while few studies focus on the integration of mined knowledge and the ML process.

Here, we summarize the entire process of the approaches to domain-knowledge acquisition and representation (Fig. 3) that consists of knowledge acquisition, knowledge representation, knowledge incorporation and data-driven ML model layers. In the knowledge acquisition layer, multi-source domain knowledge can be extracted through an information filter [72] or approaches based on natural language-processing technologies such as entity extraction [73], relation extraction [74] and entity–relation extraction [75]. Then, the knowledge representation layer represents the extracted knowledge in the form of feature importance [76], relation rules [77], a physics model [78] or a knowledge graph [79]. Concretely, the feature importance analysis is quantified through domain experts, which can be combined with importance quantifying analysis technologies such as LASSO, RF, SHAP, etc. Relation rule can be employed to describe the correlation between descriptors and construct the conditions to eliminate redundant descriptors or assist the models in identifying the latent correlation relations. Based on the paradigm of 'AI4Science', performing the formulation representation of physical models and incorporating them into the learning algorithms of ML models

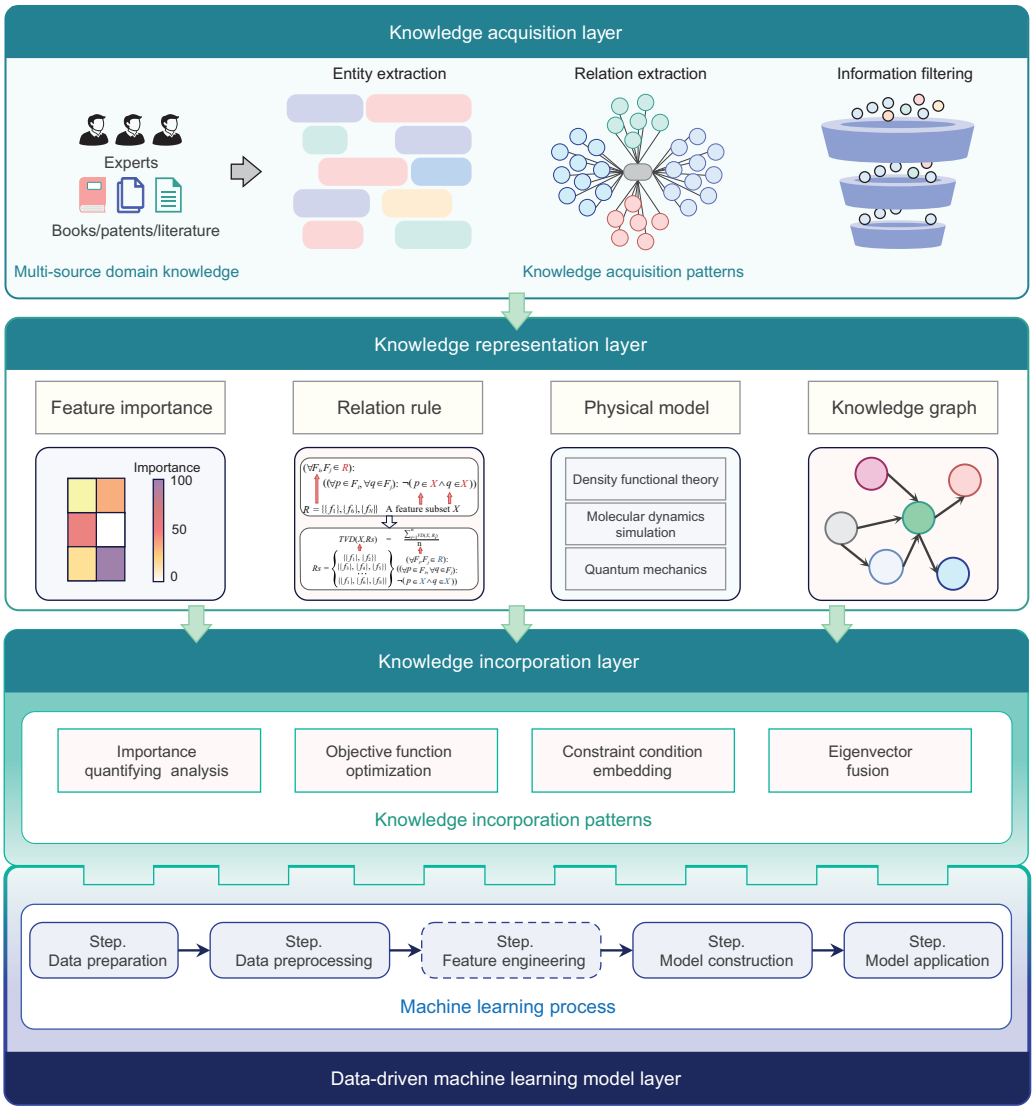

Figure 3. Schematic diagram of domain knowledge acquisition and representation. Feature engineering can be omitted in deep-learning schemes.

can facilitate alternately optimizing the parameters of the neural network and the coefficients of the algebraic terms of the equation. Therefore, neural networks with high generalization can be trained to provide accurate derivative estimation, and finally ensure that the valuable information latent in the materials data can be mined. The knowledge graph can be extracted and regarded as rules [80]. This has been utilized in the neural-network-based models to encode the knowledge graph as an eigenvector and concatenate with original data [81]. By this means, ML models can be guided to mine the latent information in the original data. Note that one knowledge representation pattern can be modified to other patterns, determined by the demand for specific tasks [82]. To effectively embed mined domain knowledge into the process of ML models (i.e. data-driven ML model layer), the knowledge incorporation layer is constructed to provide various proper approaches. We here summarize these approaches into importance quantifying analysis [76], objective function optimization [83], constraint condition embedding [84] and eigenvector fusion [81,85]. Through the incorporation of various representations of domain knowledge into each step of data-driven ML, the domain knowledge can be effectively transformed and can participate the model training process so that the overall flow can achieve more reliable and accurate analysis results. More details of applications can be seen in Section S2 of the Supplementary data.

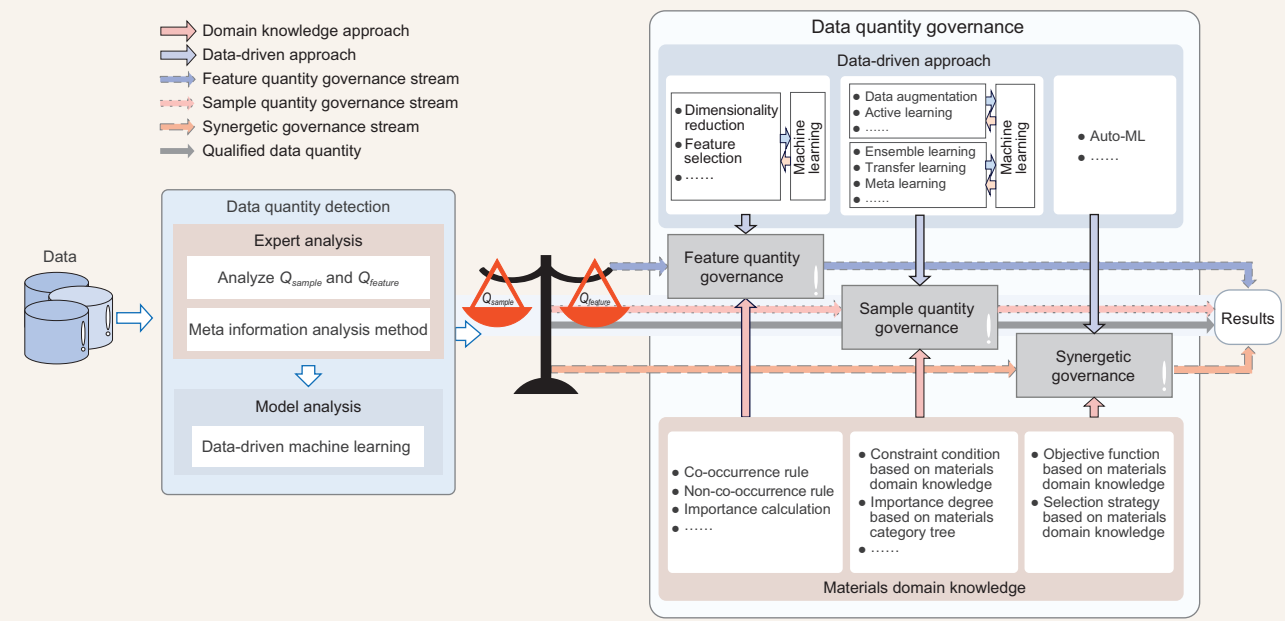

Figure 4. Data quantity governance flow embedded with materials domain knowledge. Q_{Sample} represents the number of samples. $Q_{Feature}$ represents the number of features. Data quantity detection guides the relevant data quantity governance (feature quantity, sample quantity and synergistic governance). If the data quantity qualifies, the learning results of data-driven machine-learning models are directly outputted. Note that for DL models, Q_{Sample} matters only; thus, data quantity governance merely focuses on the sample level.

Data quantity governance flow with incorporation of materials domain knowledge

As mentioned earlier, the sample size and the number of features jointly affect the performance of ML models. Therefore, we consider keeping the balance between sample size and feature space dimensionality as the data quantity governance objective. Concretely, on the one hand, high-dimensional materials data usually contain redundant features, but only pursuing the reduction of feature quantity may result in the loss of key features, which may aggravate the demand for more samples. Moreover, adding valid features into the data can be regarded as an effective approach to improving the performance of ML models [86], which requires abundant training samples. On the other hand, theoretically, the more samples provided, the more accurate the ML models will become. However, each sample, in fact, also adds noise or error information and thus may increase computation costs without providing insights. Therefore, data quantity governance should aim not at the sample and feature quantities separately but rather at the feature-to-sample ratio.

As shown in Fig. 4, the data quantity governance flow with the incorporation of materials domain knowledge consists of data quantity detection and data quantity governance. The former aims to assess whether the data sets need to be governed from the perspectives of domain-knowledge and data-driven aspects. The latter performs targeted governance according to the detection results and is divided into feature quantity governance, sample quantity governance and synergistic governance.

Data quantity detection

To circumvent redundant operation, the data quantity detection module is constructed to effectively confirm whether the data need quantity governance. Note that the relevant materials domain knowledge that a researcher possesses affects whether the data quantity requires governance or not. Expert analysis is first performed based on materials domain knowledge to roughly estimate the characteristics of the data. Meta-information (metadata) not only contains basic information (e.g. the number of features, the number of samples and rules), but also reflects the statistical characteristics. Therefore, the stage of meta-information analysis here is performed, namely experts prejudge the metadata of data sets to estimate whether the number of samples can support the accurate analysis of ML models according to the number of features. Then, model analysis is performed, as shown in Supplementary Fig. S2. At that stage, if the number of samples is sufficient to drive DL models, then DL models can be regarded as detection tools directly, otherwise data-driven methods that are

suitable for the analysis of high-dimension and small samples (e.g. SVR) are adopted. Depending on the analysis results, a suitable governance scheme is selected. Specifically, if the researchers are not familiar with the target tasks, it is correct to select the synergistic governance scheme to circumvent manual intervention.

Feature quantity governance with incorporation of materials domain knowledge

This module allows data-driven models to verify the effectiveness of feature quantity governance, and the governance methods can be further adjusted according to the analysis result of the data-driven models. Incorporating material domain knowledge into the process of feature quantity governance can help models to eliminate redundant features and construct more simple but efficient ML models. As the FT methods focus on feature space transformation, it is hard to embed existing symbolic domain knowledge into them. Herein, only FS methods are considered to illustrate the incorporation patterns of domain knowledge.

Liu *et al.* [76] proposed FS embedded with materials knowledge including data-driven multi-layer FS embedded with domain expert knowledge that combines the materials expert assessment of the importance of descriptors with data-driven measurement results, as shown in Supplementary Fig. S3. This method combines filter and wrapper methods to remove sparse, irrelevant and redundant features automatically. Then, the importance of descriptors is introduced into the FS process via domain expert knowledge to eliminate the risk of the deletion of key features. This has been shown to be effective in selecting descriptors with high prediction accuracy and consistently with expert knowledge.

In addition, materials domain knowledge can help to find correlation among descriptors. For example, Liu *et al.* [87] transferred the materials domain knowledge for the relationships between descriptors into non-co-occurrence rules (NCOR) and embedded NCOR into the process of FS. Then, a feature selection method is proposed to reduce correlations among features by embedding domain knowledge (NCOR-FS), as shown in Supplementary Fig. S4. The correlation between various factors affecting the ion transport performance of a solid electrolyte is transformed into NCOR and embedded into the objective function of the FS. In the ion activation energy prediction task of a NASICON-type electrolyte, the descriptor set with a lower internal correlation is successfully selected and a regression prediction model with better generalization performance and stability is constructed.

Sample quantity governance with incorporation of materials domain knowledge

Through sample quantity governance, the number of samples can be optimized, which effectively facilitates the application of data-driven methods. Its performance can be also verified through data-driven methods that are the same as the feature quantity governance module. Note that, in this module, either sample-oriented or model-oriented types are essentially driven by the ML model. Therefore, it can alleviate the contradiction of the small sample size and high dimensionality of the feature space (or model parameter space) to some extent without embedding materials domain knowledge. To further circumvent the model uncertainty, low interpretability as well as low robustness [88], sample quantity governance with incorporation of materials domain knowledge is proposed.

Data augmentation. Materials domain knowledge can be regarded as the constraint conditions for generative models to limit the distribution of generated samples with a reasonable value range. Concretely, for VAE-based methods, materials domain knowledge can be used in the form of formulas to be embedded into the training process. This facilitates the model catching the most important features vector, retaining the details of the data. Note that VAE and its variants are commonly utilized to ensure the explainability of input data via analysing the concept generated from the encoder [89]. Hence, materials domain knowledge can be also transformed into the physical or statistical formulas to preserve details of the input data. As for GAN-based methods, the Gaussian mixture model [90] can be employed to investigate the most representative Gaussian distribution of material data. In this way, GAN-based methods can generate valid samples according to the representative distribution. Then, materials domain knowledge is transformed into constraint conditions to eliminate the unreasonable samples. Besides, the original data can be first encoded by a neural network under the guidance of materials domain knowledge, whose form is the same as modified VAE-based methods. Besides, there are some approaches that can also provide insights for physical-informed data augmentation, e.g. the addition of perturbation physical information to original samples [91] and physical-based approaches (e.g. *ab initio* MD calculations) or simulation approaches for density functional theory (DFT) (e.g. projector-augmented wave) [92]. More details of applications can be seen in Section S3 of the Supplementary data.

Ensemble learning. The key idea of ensemble learning is combining several weak learners to obtain a comprehensive strong learner, which can

circumvent the difficulty of directly constructing a strong learner. Based on this thought, Liu *et al.* [93] proposed the machine-learning-based method of 'divide-and-conquer', which can be regarded as a novel type of ensemble learning, to predict the creep rupture life of Ni-based single-crystal superalloys in the context of a small sample size and high dimension of feature space (shown in Supplementary Fig. S5). Concretely, to conquer the complicated distribution of the Ni-based single-crystal superalloys data set, this method leveraged the K-Means cluster method to divide the data set into different clusters according to their statistical characteristics. Then, appropriate models were automatically selected for each according to their individual characteristics. The results show that the divide-and-conquer method could be successful for a small sample size (266 samples) and high dimension of feature space (27 descriptors) and achieved a predictive accuracy of 91.76%, which surpassed that of ML models based on the original undivided data set. In the divide-and-conquer process, materials domain knowledge can guide cluster methods to group data sets according to chemical composition, processing conditions, etc. By this means, the data space with complicated influence mechanisms of materials can effectively divide the overall data set into data subspaces, which facilitates modeling for each of them. Finally, a simple ML model with high prediction performance and explainability will be obtained.

Active learning. BO constructs prior distribution through sampling parts of samples, thus the results depend on the sample quality and on how close the assumed sample distribution is to the true sample distribution. To improve the effectiveness of BO, it is possible to map the high-dimension search space into a lower one. For example, Wang *et al.* [94] employed BO to replace solving an ∼1 billion-dimensional problem with a large number of low-dimensional problems (the effective dimension is 2), which can be solved much faster. Li *et al.* [95] proposed a BO method for a high-dimension space that uses a projected-additive Gaussian process to solve high-dimensional problems. Moreover, material domain knowledge can be transformed into rules of distribution of materials data to improve its sampling procedures. For example, Xue *et al.* [96] embedded the knowledge of the morphotropic phase boundary providing a temperature-independent d_{33} piezoelectric property for the compounds into the active learning loop. Then, the optimal solid solution composition was predicted and validated, which shows good temperature reliability. Yuan *et al.* [97] utilized active learning based on physical insights from the composition–temperature phase diagram to shrink the size of the virtual space from ∼9 million to 700 000. Note that there has been interest in the use of text mining and natural language-processing techniques to build data sets from the materials science literature for guiding materials synthesis and design [98]. Such advances can have key implications for the adaptive learning approach, especially in rapidly constructing training data sets for building surrogate models. Hence, integrating text mining for training set construction and active learning can have an impact in the accelerated search for novel materials.

TL. For incorporating materials domain knowledge into TL, we believe that calculating meta-information of the target domain and various candidate source domains ensures their relevance. All candidate source domains can be constructed as a tree based on materials domain knowledge. This enables each source candidate domain to get its scores from materials experts according to the need for the target source. By this means, the candidate source domain can be comprehensively analysed from the perspective of domain knowledge and statistics. As for the network structure, though DL can extract multi-scale and deeper features from materials data, the inexplicability of the model and difficulties in learning physical and geometric information are inevitable. Hence, the incorporation of materials domain knowledge can not only guide models to learn the latent pattern of materials data rapidly (i.e. fast convergence speed during model training), but also help materials experts to comprehend the learning process of the model.

Synergistic governance with incorporation of materials domain knowledge

Synergistic governance aims to balance the sample and feature quantity (or model parameter quantity) automatically. Automated machine learning (Auto-ML) [99] can meet this demand in that it can perform the steps of data processing, feature engineering and model construction automatically according to the specific tasks of materials design and discovery. That is, the procedure of synergistic governance is consistent with that of Auto-ML because sample quantity governance can be regarded as data processing. Meanwhile, feature quantity governance can be set as the step of feature engineering. Moreover, Auto-ML can reduce human intervention in the steps of model selection, optimization and implementation. According to the 'no free lunch' theorem [54], it is impossible for a single algorithm to be universally superior to any other algorithm. This implies that ML algorithm selection and hyperparameter setting differ in data sets with different

characteristics (e.g. data size and distribution). Therefore, Auto-ML can help researchers who have low familiarity with the downstream tasks to gain the intended results. However, the conventional Auto-ML is purely data-driven, and knowledge incorporation can substantially improve the accuracy and explainability of the constructed models via synergistic governance. The knowledge incorporation pattern for synergistic governance can be seen in Supplementary Fig. S6 and the details of knowledge incorporation can be seen in Section 3.1.

Although the Auto-ML method based on meta-learning can automatically select the ML algorithm with ideal prediction performance by learning the historical modeling experience and greatly reduce the combined algorithm selection and hyperparameter optimization (CASH) time and calculation cost, the tedious training process of Auto-ML can hinder wide employment in materials science. To this end, a feasible Auto-ML scheme with materials domain knowledge is proposed here, to accelerate this process and improve its accuracy. Concretely, a new meta-feature is constructed to enhance the similarity measurement between data sets. The candidate ML algorithms suitable for materials property prediction are selected by investigating materials literature related to ML. As for limited materials data sets and the lack of domain knowledge, a collaborative learning mechanism with materials domain knowledge is employed to integrate ML public data sets and materials data sets for meta-learning. This alleviates the overfitting problem and improves the interpretability and reliability of ML models.

CONCLUSIONS AND OUTLOOK

As ML models are widely employed in the field of materials science, the contradiction of a high dimensionality of feature space and a small sample size is increasingly encountered. The misbalance in data quantity (i.e. the number of samples vs. the number of features or model parameters) limits the performance of ML models both in prediction accuracy and the ability to mine latent patterns in the materials data. Here, we reviewed the efforts for data quantity governance in materials science such as feature quantity reduction, sample quantity augmentation and specific ML approaches. Then, a synergistic data quantity governance flow with incorporation of materials domain knowledge is proposed and corresponding approaches to knowledge representation and incorporation for ML are summarized. Benefitting from the incorporated materials domain knowledge, this flow can construct a high-quality data foundation to facilitate ML modeling. As for the development of a general generative artificial intelligence model (e.g. ChatGPT), it is promising that researchers can more effectively embed various pieces of domain knowledge into a sample generation process and materials samples can effectively be generated following their instructions, which can facilitate the development of AI4Science in materials science. Moreover, further development of ensemble tools for data quantity governance with the incorporation of materials domain knowledge (e.g. the flow or framework of data quality governance) can expedite progress in materials science. Through systematically governing the data quantity, researchers can conduct reliable and reproducible data analysis in an orderly manner, and then monitor the materials data quality and construct high-quality samples and high-accuracy ML models.

SUPPLEMENTARY DATA

Supplementary data are available at *NSR* online.

ACKNOWLEDGEMENTS

We appreciated the High-Performance Computing Center of Shanghai University and Shanghai Engineering Research Center of Intelligent Computing System for providing the computing resources and technical support.

FUNDING

This work was supported in part by the National Natural Science Foundation of China (92270124 and 52073169), the National Key Research and Development Program of China (2021YFB3802101) and the Key Research Project of Zhejiang Laboratory (2021PE0AC02).

Conflict of interest statement. None declared.

REFERENCES

1. Liu Y, Zhao T and Ju W *et al.* Materials discovery and design using machine learning. *J Materiomics* 2017; **3**: 159–77.
2. Cai J, Luo J and Wang S *et al.* Feature selection in machine learning: a new perspective. *Neurocomputing* 2018; **300**: 70–9.
3. Van Der Maaten L, Postma E and Van den Herik J. Dimensionality reduction: a comparative. *J Mach Learn Res* 2009; **10**: 66–71.
4. Agrawal A and Choudhary A. Perspective: materials informatics and big data: realization of the 'fourth paradigm' of science in materials science. *APL Mater* 2016; **4**: 053208.
5. Goodfellow I, Pouget-abadie J and Mirza M *et al.* Generative adversarial networks. *Commun ACM* 2020; **63**: 139–44.
6. Hinton GE and Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006; **313**: 504–7.

7. Lookman T, Balachandran PV and Yuan R *et al.* Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Comput Mater* 2019; **5**: 21.
8. Zhou Z. *Ensemble Learning*. In: *Machine Learning*. Singapore: Springer Singapore, 2021, 181–210.
9. Torrey L and Shavlik J. Transfer learning. In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. Hershey: IGI Global, 2010, 242–64.
10. Im J, Lee S and Ko T-W *et al.* Identifying Pb-free perovskites for solar cells by machine learning. *npj Comput Mater* 2019; **5**: 37.
11. Deng Q and Lin B. Exploring structure-composition relationships of cubic perovskite oxides via extreme feature engineering and automated machine learning. *Mater Today Commun* 2021; **28**: 102590.
12. Agrawal A, Deshpande PD and Cecen A *et al.* Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters. *Integr Mater Manuf Innov* 2014; **3**: 90–108.
13. Shin D, Yamamoto Y and Brady MP *et al.* Modern data analytics approach to predict creep of high-temperature alloys. *Acta Mater* 2019; **168**: 321–30.
14. Genuer R, Poggi J-M and Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett* 2010; **31**: 2225–36.
15. Rodriguez-Galiano VF, Luque-Espinar JA and Chica-Olmo M *et al.* Feature selection approaches for predictive modelling of groundwater nitrate pollution: an evaluation of filters, embedded and wrapper methods. *Sci Total Environ* 2018; **624**: 661–72.
16. Qi Z, Zhang N and Liu Y *et al.* Prediction of mechanical properties of carbon fiber based on cross-scale FEM and machine learning. *Compos Struct* 2019; **212**: 199–206.
17. Zeng Y, Li Q and Bai K. Prediction of interstitial diffusion activation energies of nitrogen, oxygen, boron and carbon in bcc, fcc, and hcp metals using machine learning. *Comput Mater Sci* 2018; **144**: 232–47.
18. Stanev V, Oses C and Kusne AG *et al.* Machine learning modeling of superconducting critical temperature. *npj Comput Mater* 2018; **4**: 29.
19. O'Connor NJ, Jonayat ASM and Janik MJ *et al.* Interaction trends between single metal atoms and oxide supports identified with density functional theory and statistical learning. *Nat Catal* 2018; **1**: 531–9.
20. Mangal A and Holm EA. A comparative study of feature selection methods for stress hotspot classification in materials. *Integr Mater Manuf Innov* 2018; **7**: 87–95.
21. Tekin Erguzel T, Tas C and Cebi M. A wrapper-based approach for feature selection and classification of major depressive disorder-bipolar disorders. *Comput Biol Med* 2015; **64**: 127–37.
22. Wu H, Lorenson A and Anderson B *et al.* Robust FCC solute diffusion predictions from ab-initio machine learning methods. *Comput Mater Sci* 2017; **134**: 160–5.
23. Furmanchuk A, Saal JE and Doak JW *et al.* Prediction of Seebeck coefficient for compounds without restriction to fixed stoichiometry: a machine learning approach. *J Comput Chem* 2018; **39**: 191–202.
24. Oliynyk AO, Adutwum LA and Rudyk BW *et al.* Disentangling structural confusion through machine learning: structure prediction and polymorphism of equiatomic ternary phases ABC. *J Am Chem Soc* 2017; **139**: 17870–81.
25. Liu X, Lu W and Peng C *et al.* Two semi-empirical approaches for the prediction of oxide ionic conductivities in ABO_3 perovskites. *Comp Mater Sci* 2009; **46**: 860–8.
26. Sendek AD, Yang Q and Cubuk ED *et al.* Holistic computational structure screening of more than 12 000 candidates for solid lithium-ion conductor materials. *Energy Environ Sci* 2017; **10**: 306–20.
27. Aziz R, Verma C and Srivastava N. Dimension reduction methods for microarray data: a review. *AIMS Bioeng* 2017; **4**: 179–97.
28. Wen C, Wang C and Zhang Y *et al.* Modeling solid solution strengthening in high entropy alloys using machine learning. *Acta Mater* 2021; **212**: 116917.
29. Wang X, Xu Y and Yang J *et al.* ThermoEPred-EL: robust bandgap predictions of chalcogenides with diamond-like structure via feature cross-based stacked ensemble learning. *Comput Mater Sci* 2019; **169**: 109117.
30. Rajan AC, Mishra A and Satsangi S *et al.* Machine-learning-assisted accurate band gap predictions of functionalized MXene. *Chem Mater* 2018; **30**: 4031–8.
31. Yan C, Liang J and Zhao M *et al.* A novel hybrid feature selection strategy in quantitative analysis of laser-induced breakdown spectroscopy. *Anal Chim Acta* 2019; **1080**: 35–42.
32. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 2001; **42**: 177–96.
33. Wold S, Esbensen K and Geladi P. Principal component analysis. *Chemom Intell Lab Syst* 1987; **2**: 37–52.
34. Sturlaugson LE and Sheppard JW. Principal component analysis preprocessing with Bayesian networks for battery capacity estimation. In: *Proceedings of the 2013 IEEE International Instrumentation and Measurement Technology Conference, Minneapolis, USA*. New York: IEEE Press, 2013, 98–101.
35. Curtarolo S, Morgan D and Persson K *et al.* Predicting crystal structures with data mining of quantum calculations. *Phys Rev Lett* 2003; **91**: 135503.
36. Ouyang R, Curtarolo S and Ahmetcik E *et al.* SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys Rev Mater* 2018; **2**: 083802.
37. Bartel CJ, Sutton C and Goldsmith BR *et al.* New tolerance factor to predict the stability of perovskite oxides and halides. *Sci Adv* 2019; **5**: eaav0693.
38. Andersen M, Levchenko SV and Scheffler M *et al.* Beyond scaling relations for the description of catalytic materials. *ACS Catal* 2019; **9**: 2752–9.
39. Bartel CJ, Millican SL and Deml AM *et al.* Physical descriptor for the Gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry. *Nat Commun* 2018; **9**: 4168.
40. Weng B, Song Z and Zhu R *et al.* Simple descriptor derived from symbolic regression accelerating the discovery of new perovskite catalysts. *Nat Commun* 2020; **11**: 3513.
41. He M and Zhang L. Machine learning and symbolic regression investigation on stability of MXene materials. *Comput Mater Sci* 2021; **196**: 110578.
42. Wu YJ, Fang L and Xu Y. Predicting interfacial thermal resistance by machine learning. *npj Comput Mater* 2019; **5**: 56.
43. Zhao K, Jin X and Wang Y. Survey on few-shot learning. *J Softw* 2021; **32**: 349–69.
44. Yang Z, Gao J and Wang S *et al.* Synergetic application of E-tongue and E-eye based on deep learning to discrimination of Pu-erh tea storage time. *Comput Electron Agri* 2021; **187**: 106297.
45. Song Y, Siriwardane EMD and Zhao Y *et al.* Computational discovery of new 2D materials using deep learning generative models. *ACS Appl Mater Interfaces* 2021; **13**: 53303–13.
46. Dan Y, Zhao Y and Li X *et al.* Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput Mater* 2020; **6**: 84.
47. Ma B, Wei X and Liu C *et al.* Data augmentation in microscopic images for material data mining. *npj Comput Mater* 2020; **6**: 125.
48. Noh J, Kim J and Stein HS *et al.* Inverse design of solid-state materials via a continuous representation. *Matter* 2019; **1**: 1370–84.
49. Hoffmann J, Maestrati L and Sawada Y *et al.* Data-driven approach to encoding and decoding 3-d crystal structures. arXiv:1909.00949.

50. Bassman L, Rajak P and Kalia RK *et al.* Active learning for accelerated design of layered materials. *npj Comput Mater* 2018; **4**: 74.
51. Min K and Cho E. Accelerated discovery of potential ferroelectric perovskite via active learning. *J Mater Chem C* 2020; **8**: 7866–72.
52. Pruksawan S, Lambard G and Samitsu S *et al.* Prediction and optimization of epoxy adhesive strength from a small dataset through active learning. *Sci Technol Adv Mater* 2019; **20**: 1010–21.
53. Doan HA, Agarwal G and Qian H *et al.* Quantum chemistry-informed active learning to accelerate the design and discovery of sustainable energy storage materials. *Chem Mater* 2020; **32**: 6338–46.
54. Wolpert DH and Macready WG. No free lunch theorems for optimization. *IEEE Trans Evol Comput* 1997; **1**: 67–82.
55. Breiman L. Bagging predictors. *Mach Learn* 1996; **24**: 123–40.
56. Biau G and Scornet E. A random forest guided tour. *TEST* 2016; **25**: 197–227.
57. Efron B and Tibshirani RJ. *An Introduction to the Bootstrap*. Boca Rato: CRC Press, 1994.
58. Okafor E, Obada DO and Dodoo-Arhin D. Ensemble learning prediction of transmittance at different wavenumbers in natural hydroxyapatite. *Sci Afr* 2020; **9**: e00516.
59. Farooq F, Ahmed W and Akbar A *et al.* Predictive modeling for sustainable high-performance concrete from industrial wastes: a comparison and optimization of models using ensemble learners. *J Cleaner Prod* 2021; **292**: 126032.
60. Yang W, Li W and Zhao Y *et al.* Mechanical property prediction of steel and influence factors selection based on random forests. *Iron and Steel* 2018; **3**: 44–9.
61. Ji Y, Yong X and Liu Y *et al.* Random forest based quality analysis and prediction method for hot-rolled strip. *J Northeastern Univ (Nat Sci)* 2019; **40**: 11–5.
62. Pan S and Yang Q. A survey on transfer learning. *IEEE Trans Knowl Data Eng* 2009; **22**: 1345–59.
63. Gupta V, Choudhary K and Tavazza F *et al.* Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data. *Nat Commun* 2021; **12**: 6595.
64. Bäuml B and Tulbure A. Deep n-shot transfer learning for tactile material classification with a flexible pressure-sensitive skin. In: *International Conference on Robotics and Automation (ICRA), Montreal, Canada*. New York: IEEE Press, 2019, 4262–8.
65. Chen S, Hu Z and Wang C *et al.* Research on the process of small sample non-ferrous metal recognition and separation based on deep learning. *Waste Manage* 2021; **126**: 266–73.
66. Wang Z, Wang Q and Han Y *et al.* Deep learning for ultra-fast and high precision screening of energy materials. *Energy Storage Mater* 2021; **39**: 45–53.
67. Ma R, Colon YJ and Luo T. Transfer learning study of gas adsorption in metal–organic frameworks. *ACS Appl Mater Interfaces* 2020; **12**: 34041–8.
68. Liu Y, Zou X and Yang Z *et al.* Machine learning embedded with materials domain knowledge. *J Chin Cera Soc* 2022; **50**: 863–76.
69. Stevens R, Taylor V and Nichols J *et al.* AI for Science. *Tech Rep* 2020. Argonne National Lab (ANL), Argonne, US. doi: 10.2172/1604756.
70. Weinan E. Machine learning and computational mathematics. arXiv: 2009.14596.
71. Kononova O, He T and Huo H *et al.* Opportunities and challenges of text mining in materials research. *iScience* 2021; **24**: 102155.
72. Pouran Ben Veyseh A, Meister N and Dernoncourt F *et al.* Improving keyphrase extraction with data augmentation and information filtering. arXiv: 2209.04951.
73. Weston L, Tshitoyan V and Dagdelen J *et al.* Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J Chem Inf Mode* 2019; **59**: 3692–702.
74. Al-Moslmi T, Ocaña MG and Opdahl AL *et al.* Named entity extraction for knowledge graphs: a literature overview. *IEEE Access* 2020; **8**: 32862–81.
75. Xu J, Zhang Z and Wu Z. Review on techniques of entity relation extraction. *Data Anal Knowl Discov* 2008; **24**: 18–23.
76. Liu Y, Wu J and Avdeev M *et al.* Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties. *Adv Theory Simul* 2020; **3**: 1900215.
77. Yuan J, Wang Q and Li Z *et al.* Domain-knowledge-oriented data preprocessing and machine learning of corrosion-resistant γ-U alloys with a small database. *Comput Mater Sci* 2021; **194**: 110472.
78. Chen Z, Liu Y and Sun H. Physics-informed learning of governing equations from scarce data. *Nat Commun* 2021; **12**: 6136.
79. Nie Z, Zheng S and Liu Y *et al.* Automating materials exploration with a semantic knowledge graph for Li-ion battery cathodes. *Adv Funct Mater* 2022; **32**: 2201437.
80. Ji S, Pan S and Cambria E *et al.* A survey on knowledge graphs: representation, acquisition, and applications. *IEEE Trans Neural Netw Learning Syst* 2021; **33**: 494–514.
81. Zhang N, Bi Z and Liang X *et al.* Ontoprotein: protein pretraining with gene ontology embedding. arXiv:2201.11147.
82. Von Rueden L, Mayer S and Beckh K *et al.* Informed machine learning: a taxonomy and survey of integrating knowledge into learning systems. *IEEE T Knowl Data En* 2021; **35**: 614–33.
83. Gasteiger J, Groß J and Günnemann S. Directional message passing for molecular graphs. arXiv:2003.03123.
84. Jia X, Willard J and Karpatne A *et al.* Physics-guided machine learning for scientific discovery: an application in simulating lake temperature profiles. *ACM/IMS Trans Data Sci* 2021; **2**: 1–26.
85. Deng S, Zhang N and Zhang W *et al.* Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In: *Companion Proceedings of the 2019 World Wide Web Conference, San Francisco, USA*. New York: IEEE Press, 2019; 678–85.
86. Zhang Y and Ling C. A strategy to apply machine learning to small datasets in materials science. *npj Comput Mater* 2018; **4**: 25.
87. Liu Y, Zou X and Ma S *et al.* Feature selection method reducing correlations among features by embedding domain knowledge. *Acta Mater* 2022; **238**: 118195.
88. Liu Y, Ge X and Yang Z *et al.* An automatic descriptors recognizer customized for materials science literature. *J Power Sources* 2022; **545**: 231946.
89. Tjoa E and Guan C. A survey on explainable artificial intelligence (XAI): toward medical XAI. *IEEE Trans Neural Netw Learning Syst* 2020; **32**: 4793–813.
90. Reynolds DA. Gaussian mixture models. *Encyclopedia biometr* 2009; **741**: 659–63.
91. Gibson J, Hire A and Hennig RG. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. *npj Comput Mater* 2022; **8**: 211.
92. Li H, Wang Z and Zou N *et al.* Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation. *Nat Comput Sci* 2022; **2**: 367–77.
93. Liu Y, Wu J and Wang Z *et al.* Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Mater* 2020; **195**: 454–67.

94. Wang Z, Zoghi M and Hutter F *et al*. Bayesian optimization in high dimensions via random embeddings. In: *International Joint Conference On Artificial Intelligence (IJCAI), Beijing, China*. AAAI Press: Washington, DC, 2013, 1778–84.
95. Li C, Kandasamy K and Póczos B *et al*. High dimensional Bayesian optimization via restricted projection pursuit models. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, Cadiz, Spain*. New York: PMLR, 2016, 884–92.
96. Xue D, Balachandran PV and Hogden J *et al*. Accelerated search for materials with targeted properties by adaptive design. *Nat Commun* 2016; **7**: 11241.
97. Yuan R, Tian Y and Xue D *et al*. Accelerated search for $BaTiO_3$-based ceramics with large energy storage at low fields using machine learning and experimental design. *Adv Sci* 2019; **6**: 1901395.
98. Kim E, Huang K and Saunders A *et al*. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem Mater* 2017; **29**: 9436–44.
99. Hutter F, Kotthoff L and Vanschoren J. *Automated Machine Learning: Methods, Systems, Challenges*. Berlin: Springer Nature, 2019, 3–5.
100. Lin T, Maire M and Belongie S *et al*. Microsoft COCO: common objects In context. In: *13th Proceedings of the European Conference On Computer Vision (ECCV), Zurich, Switzerland*. Springer: Berlin, 2014, 740–55.

Supplementary data for

data quantity governance for machine learning in materials science

Yue Liu[1,4], Zhengwei Yang[1], Xinxin Zou[1], Shuchang Ma[1], Dahui Liu[1], Maxim Avdeev[5,6] and Siqi Shi[2,3,*]

[1]*School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China;*

[2]*State Key Laboratory of Advanced Special Steel, School of Materials Science and Engineering, Shanghai University, Shanghai 200444, China;*

[3]*Materials Genome Institute, Shanghai University, Shanghai 200444, China;*

[4]*Shanghai Engineering Research Center of Intelligent Computing System, Shanghai 200444, China;*

[5]*Australian Nuclear Science and Technology Organisation, Sydney 2232, Australia;*

[6]*School of Chemistry, The University of Sydney, Sydney 2006, Australia*

[*]**Corresponding author.** E-mail: sqshi@shu.edu.cn

S1. Supplementary Figures

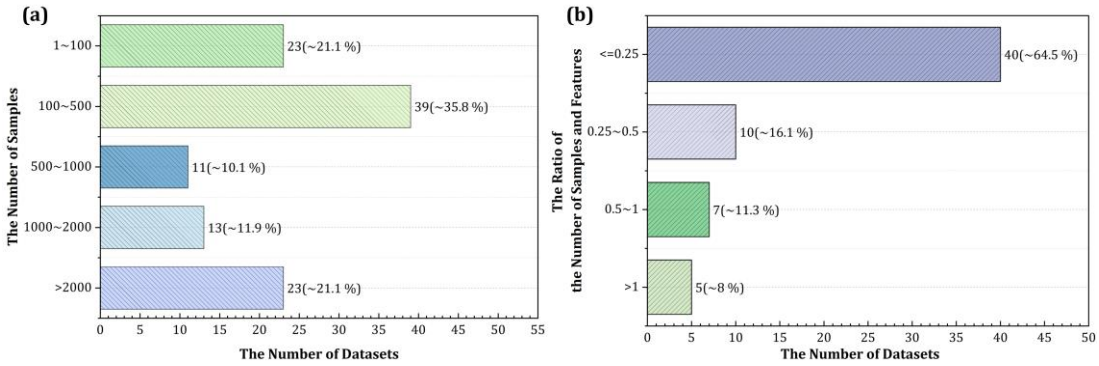

Figure S1. The Number of Samples and Features in 107 Papers in Materials Machine Learning. (a) The dataset size distribution. (b) The ratio of dataset size to the number of features

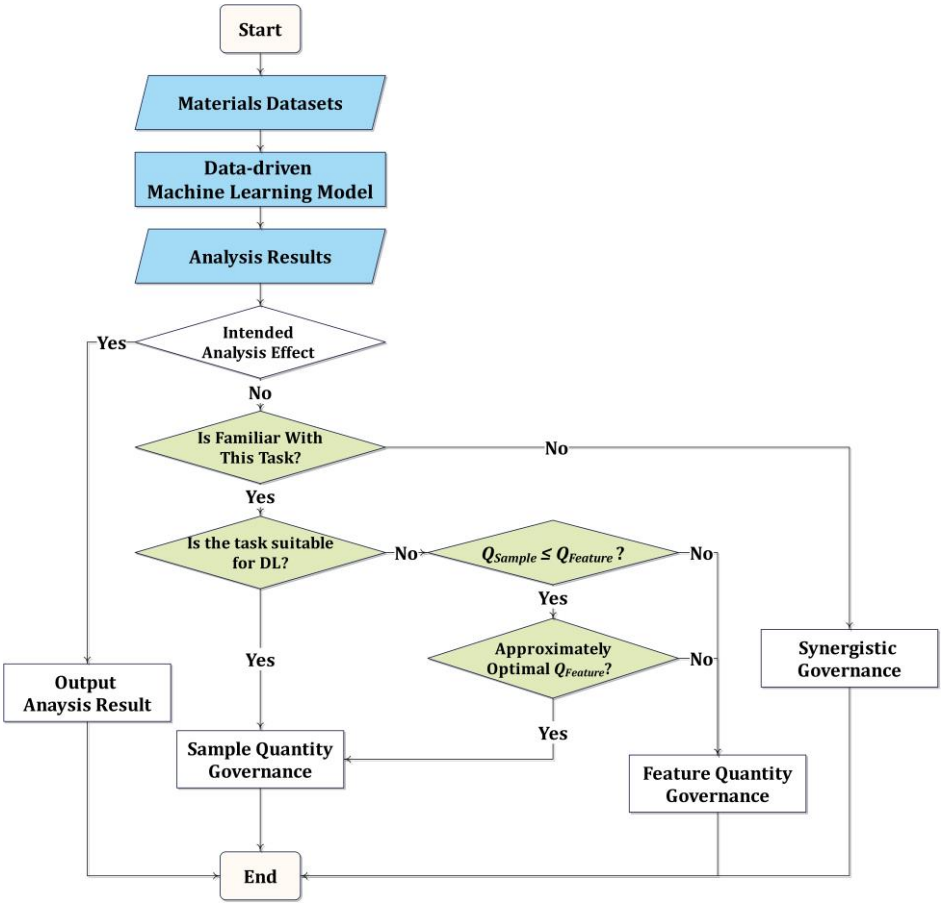

Figure S2. The Flow of Model Analysis Stage. The green background represents this step requires domain knowledge; The blue background represents this step is data relevant. $\boldsymbol{Q_{Sample}}$ represents the sample quantity; $\boldsymbol{Q_{Feature}}$ represents the feature quantity.

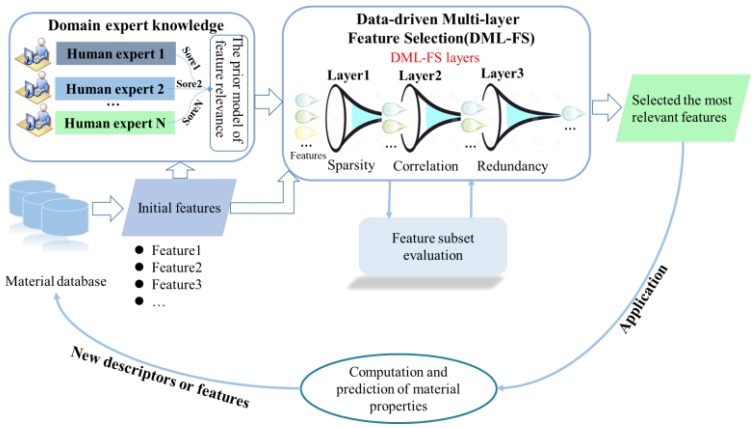

Figure S3. Data-driven Multi-layer Feature Selection Method Incorporating Domain Expert Knowledge [1].

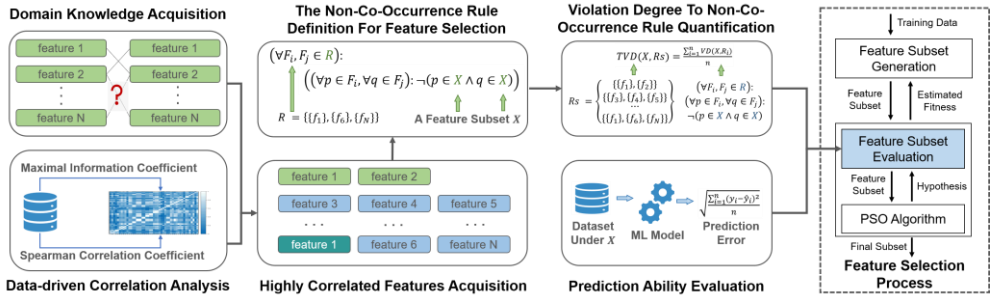

Figure S4. Flow diagram of NCOR-FS method. R and Rs represent a NCOR and the set of all NCORs, respectively. $f_1, \cdots, f_n$ represent features constructed by materials experts [2].

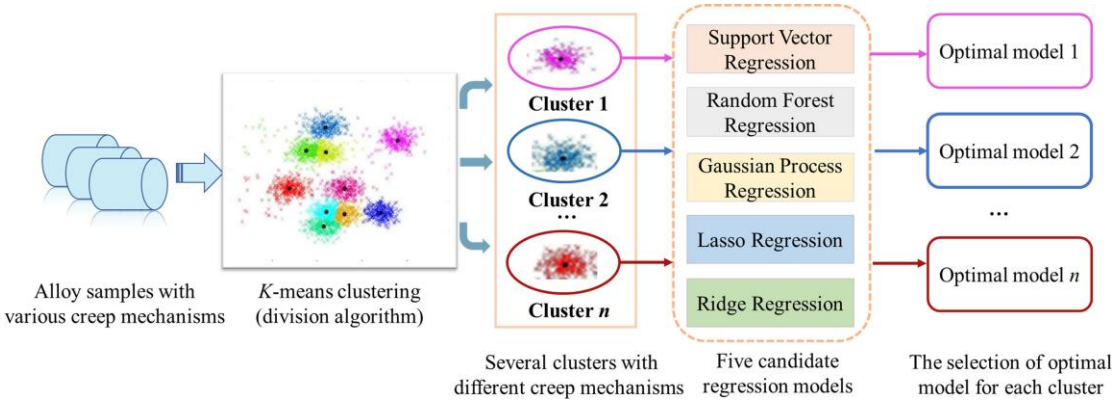

Figure S5. The procedure of divide-and-conquer self-adaptive (DCSA) learning method for modeling the creep rupture life [3].

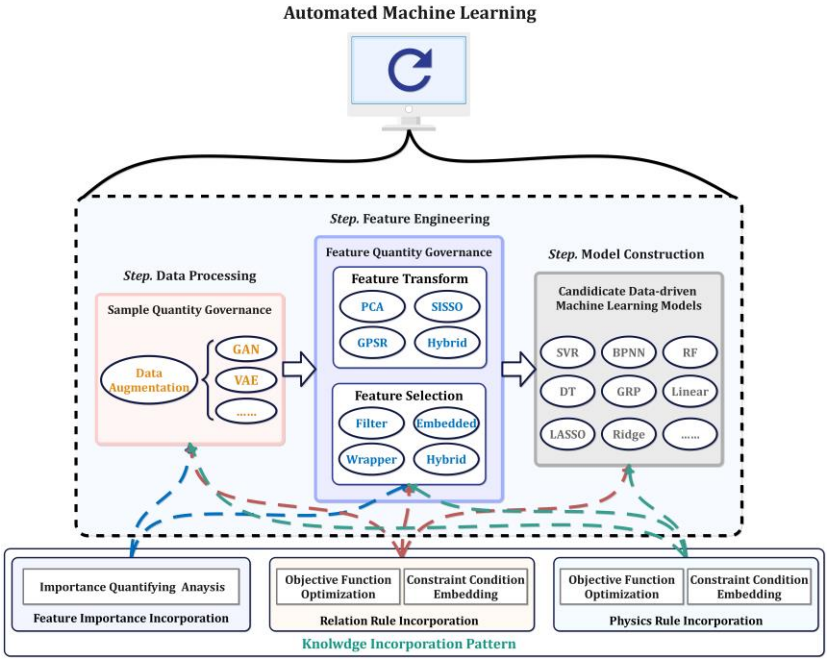

Figure S6. Synergistic Governance with Incorporation of Materials Domain Knowledge. The dummy line represents this incorporation pattern has not been used in this step.

S2. Examples of Knowledge Acquirement and Representation

This section details examples of knowledge acquirement and representation, which aims to facilitate the readers from broad and different background to comprehend the content of Section 3.1.

S2.1 Knowledge Acquirement

Ceder et al. [4] proposed an unsupervised approach to efficiently encode materials science knowledge present in the published literature as information-dense word embeddings. The results show that this method can recommend materials for functional applications several years before their discovery, suggesting that latent knowledge regarding future discoveries is to a large extent embedded in past publications, which points a novel path to extracting materials knowledge and relationships from scientific literature.

On this basis, Weston et al. [5] achieved automatic extraction of large-scale inorganic material information and solid-state synthesis information by manually annotating a large amount of supervised data and then training a deep learning NER model (BiLSTM-CRF), shown in **Figure S7**. The results show that their proposed model can effectively extract summary-level information from materials science documents, including inorganic material mentions, sample descriptors, phase labels, material properties and applications, as well as any synthesis and characterization methods used, which achieves 87% of identification accuracy.

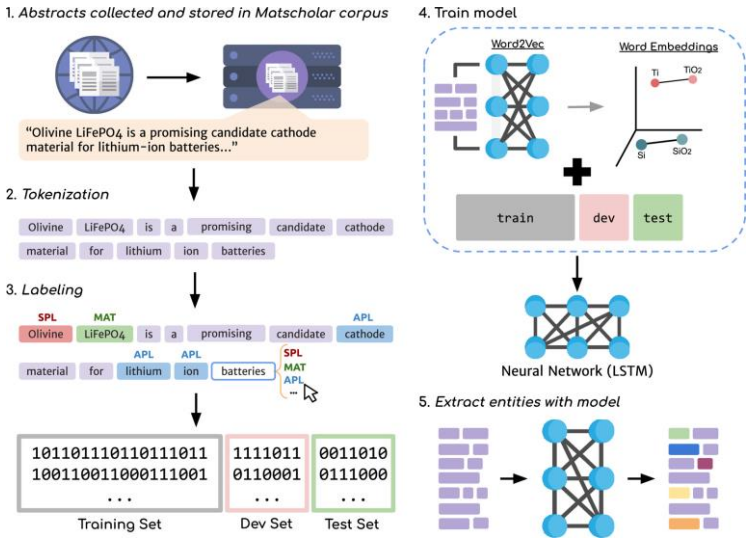

Figure S7. Workflow for named entity recognition [5].

Liu et al. [6] proposed an automatic descriptors recognizer based on natural language processing to mine latent descriptors (**Figure S8**), which can realize data augmentation with embedded domain knowledge from text data and filter task-related descriptors from coarse-grained to fine-grained. The results show that the proposed model can fully capture the contextual semantic features of the material text, classify words or phrases, and then use them for automatic recognition of descriptors. Finally, using the filtered descriptors, two datasets are constructed as activation energy predictions for ML models. These models have achieved good prediction results, demonstrating the effectiveness of automatic descriptor recognizers.

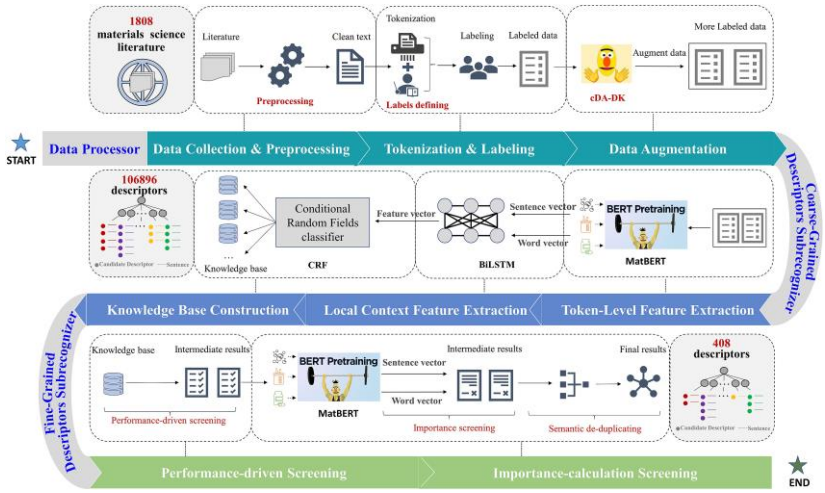

Figure S8. The overall pipeline of automatic descriptors recognizer [6].

S2.2 Knowledge Representation

Zhang et al. [7] proposed OntoProtein (**Figure S9**), a framework that integrates external knowledge graphs into protein pre-training, and proposed novel contrastive learning with knowledge-aware negative sampling to jointly optimize the knowledge graph and protein embedding during pre-training. The results demonstrate that efficient knowledge injection helps understand and uncover the grammar of life. Meanwhile, the proposed model is compatible with the model parameters of lots of pre-trained protein language models, which means that users can

directly adopt the available pre-trained parameters on OntoProtein without modifying the architecture.

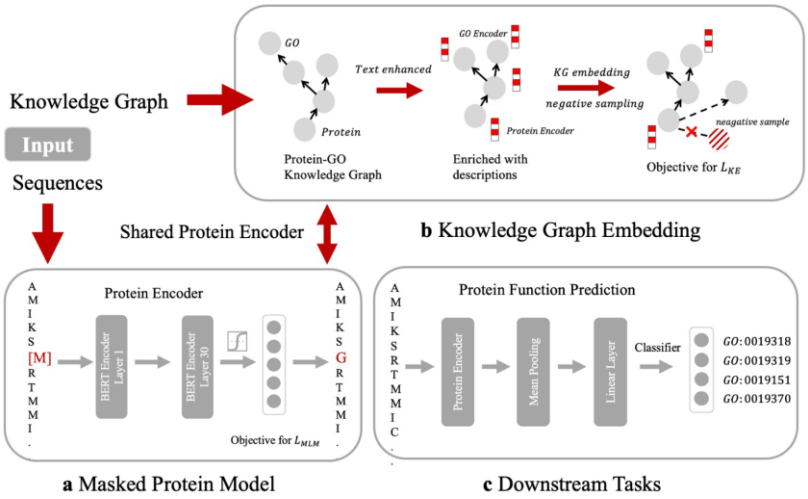

Figure S9. Overview of our proposed OntoProtein [7].

To discover governing partial differential equations from scarce and noisy data for nonlinear spatiotemporal systems, Chen et al [8] proposed a novel approach, physics-informed neural network with sparse regression (PINN-SR), shown in **Figure S10**, where objective function is comprised of data loss for descripting the difference between the measurement data and the corresponding DNN-approximated solution, physics loss for the description of the physical law, and regularization for accelerating the convergence. The results show that the proposed model can accurately discovering the exact form of the governing equation(s), even in an information-poor space where the multi-dimensional measurements are scarce and noisy.

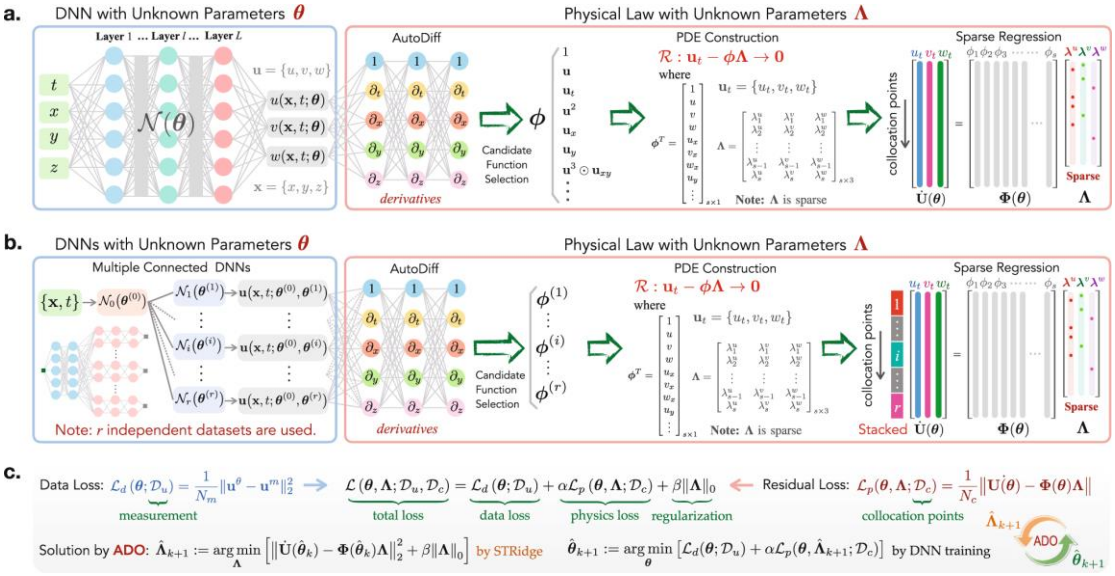

Figure S10. The framework of PINN-SR for data-driven discovery of PDE(s) [8].

To depict the heat energy fluxes and the lake thermal energy, Jia et al. [9] proposed PGRNN, which is embedded constraint condition as knowledge-based loss terms into the learning objective function. Moreover, generic general lake model (GLM) is employed to generate physical simulation data to pretrain the proposed model, aiming to leverage physical knowledge to help

inform the initialization of the weights, thus accelerating model training. The results show that the PGRNN can effectively model spatial and temporal physical processes while incorporating energy conservation.

Deng et al. [10] proposed knowledge-driven temporal convolutional network (KDTCN) for accurate stock trend prediction and explanation. Concretely, they extracted structured events from financial news and utilizes external knowledge from knowledge graph to obtain event embeddings, thus combining event embeddings and price values together to forecast stock trend. The results show that the proposed model not only can more accurately forecast stock trend with abrupt changes than present deep models but make explanation on prediction results with abrupt changes.

S3. Examples of Data Quantity Governance with Incorporation of Materials Domain Knowledge

This section details examples of knowledge acquirement and representation, which aims to facilitate the readers from broad and different background to comprehend the content of Section 3.2.

Gómez-Bombarelli et al. [11] applied VAE to the design of drug-like molecules, of which framework can be seen in **Figure S11**. By using an encoder network to convert a discrete molecular representation into a continuous vector in the latent space, and then performing simple operations on the latent continuous vector, such as perturbing known chemical structures, or interpolating between molecules. Then the modified vector can be converted back into a new discrete molecular representation through a decoder. Finally, a predictor module is used to predict the chemical properties of the new molecular representation, thus effectively searching for candidate materials with higher target performance.

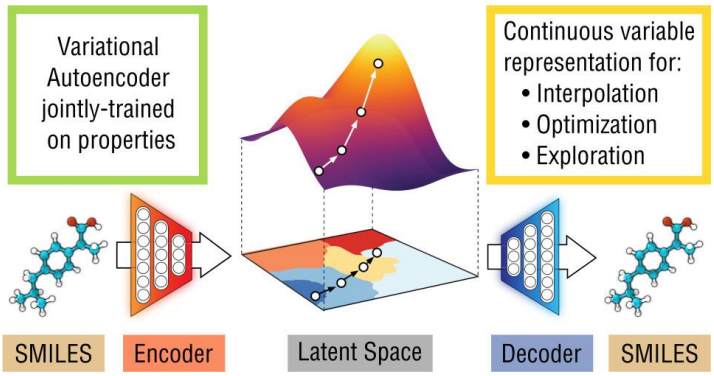

Figure S11. A diagram of VAE used for molecular design, including the joint property prediction model [11].

To improve the machine learning of this effective PES by better sampling the configuration space, Gibson et al. [12] add a perturbed structure for every relaxed structure and map it to the same energy as the relaxed structure, thus representing an additional point for a given basin of attraction of the energy landscape (shown in **Figure S12**). The results show that the prediction MAE of the CGCNN and CGCNN-HD were reduced from 251 meV/atom and 172 meV/atom to 86 meV/atom and 82 meV/atom, respectively, compared to training on only relaxed structures, which shows the surprising effectiveness of a relatively simple method of augmentations that outperformed the

current state of the art in formation energy prediction of unrelaxed structures.

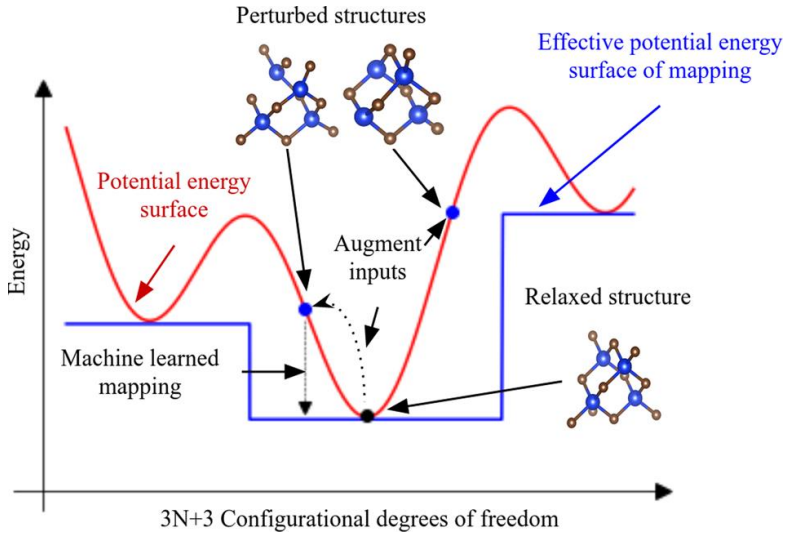

Figure S12. Data augmentation for learning the potential energy surface (PES) [12]. The red line denotes a 2D representation of the continuous PES of materials. The blue line illustrates the effective PES, which describes the energy of a relaxed structure for a given unrelaxed input structure. The black circle means the relaxed structures contained in the dataset, and the blue circles symbolize artificially generated structures for the data augmentation.

Moreover, to drive the accurate prediction of deep learning models, Li et al [13] generate plenty of monolayer graphene and monolayer MoS2 supercells by ab initio MD calculations and the Simulations are performed with the projector-augmented wave [14] pseudopotentials and the GGA parameterized by Perdew, Berke and Ernzerhof (PBE) [15].

Reference:

[1] Liu Y, Wu J and Avdeev M *et al*. Multi-layer feature selection incorporating weighted score-based expert knowledge toward modeling materials with targeted properties. *Adv Theor Simul* 2020; **3**: 1900215.

[2] Liu Y, Zou X and Ma S *et al.* Feature Selection Method Reducing Correlations among Features by Embedding Domain Knowledge. *Acta Mater* 2022; **238**: 118195.

[3] Liu Y, Wu J, Wang Z *et al.* Predicting creep rupture life of Ni-based single crystal superalloys using divide-and-conquer approach based machine learning. *Acta Mater* 2020; **195**: 454-67.

[4] Tshitoyan V, Dagdelen J and Weston L *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* 2019; **571**: 95-8.

[5] Weston L, Tshitoyan V and Dagdelen J *et al*. Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. *J Chem Inf Model* 2019; **59**: 3692-702.

[6] Liu Y, Ge X, Yang Z, *et al.* An automatic descriptors recognizer customized for materials science literature. *J Power Sources* 2022: **545**: 231946.

[7] Zhang N, Bi Z and Liang X, *et al.* Ontoprotein: Protein pretraining with gene ontology embedding. arXiv:2201.11147.

[8] Chen Z, Liu Y and Sun H. Physics-informed learning of governing equations from scarce data. *Nat Commun* 2021; **12**: 6136.

[9] Jia X, Willard J and Karpatne A *et al.* Physics-guided machine learning for scientific discovery: An application in simulating lake temperature profiles. *ACM/IMS Trans on Data Sci*; **2**:1-26.

[10] Deng S, Zhang N, Zhang W, et al. Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In: *Companion Proceedings of The 2019 World Wide Web Conference, San Francisco, USA*. New York: IEEE Press, 2019, 678-685.

[11] Gomez-Bombarelli R, Wei J and Duvenaud D *et al*. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* 2018; **4**: 268-76. https://pubs.acs.org/doi/10.1021/acscentsci.7b00572.

[12] Gibson J, Hire A, and Hennig R G. Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures. *npj Comput Mater* 2022; **8**: 211.

[13] Li H, Wang Z and ZOU N *et al.* Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation. *Nat Comput Sci* 2022; **2**: 367-77.

[14] Kresse G, Joubert D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys Rev B* 1999; **59**: 1758.

[15] Perdew J P, Burke K and Ernzerhof M. Generalized Gradient Approximation Made Simple. *Physical Rev Lett* 1996; **77**: 3865.