Machine learning facilitating the rational design of nanozymes

Yucong Li[#ab], Ruofei Zhang[#a], Xiyun Yan[*abc], Kelong Fan[*abc]

a CAS Engineering Laboratory for Nanozyme, Key Laboratory of Protein and Peptide Pharmaceutical, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China

b University of Chinese Academy of Sciences, Chinese Academy of Sciences, Beijing 100408, China

c Nanozyme Medical Center, School of Basic Medical Sciences, Zhengzhou University, Zhengzhou 450052, China

# These authors contributed equally to this work.

* Corresponding authors: yanxy@ibp.ac.cn (X. Yan); fankelong@ibp.ac.cn (K. Fan).

Abstract

As a component substitute for natural enzymes, nanozymes have the advantages of easy synthesis, convenient modification, low cost, and high stability, which are widely used in many fields. However, their application is seriously restricted by the difficulty of rapidly creating high-performance nanozymes. The use of machine learning techniques to guide the rational design of nanozymes has greatly overcome this difficulty. In this review, we introduce the recent progress of machine learning in assisting the design of nanozymes. Particular focus is given to the successful strategies of machine learning in predicting the activity, selectivity, catalytic mechanisms, optimal structures and other features of nanozymes. The typical procedures and approaches for conducting machine learning in the study of nanozymes are also highlighted. Moreover, we discuss in detail the difficulties of machine learning methods in dealing with the redundant and chaotic nanozyme data and provide an outlook on the future application of machine learning in the nanozyme field. We hope this review will serve as a useful handbook for researchers in related fields and promote the utilization of machine learning in nanozyme rational design and related topics.

1. Introduction

As an indispensable substance for lives, enzymes have attracted tons of attention for their efficient catalytic ability. Widely used in the fields of biomedicine and industrial catalysis, enzymes reduce the energy barrier of reactions, which thus makes them faster and more efficient.[1] However, the application of natural enzymes has always been limited by their defects, such as dependence on mild catalytic conditions, easy denaturation, and purification difficulties. Finding substitutes for enzymes with comparable catalytic abilities is an important topic that lots of scientists are focusing on.

Since the discovery of ferromagnetic nanoparticles that possess unexpected peroxidase (POD)-like activity by Yan's Group in 2007, nanozymes have gained a lot of attention. The number of nanozyme-related research publications has been increasing year by year.[2, 3] Nanozymes, or nanomaterials with intrinsic enzyme-like activities, have been known as promising alternatives to natural enzymes. Besides robust catalytic ability, nanozymes possess many other better properties than natural enzymes, such as easy preparation, recovery, storage, and transportation, suitable for modification, well catalytic activity, and multiple enzyme-mimicking activities. In addition, As a kind of nanomaterials, they possess unique physicochemical properties, such as photothermal properties, magnetic properties, fluorescence properties, and so on, making their application range greatly increase.[4, 5] At present, nanozymes have been widely used in many fields, including artificial organelles[6], biosensing[7, 8], environmental treatment, as well as the treatment of tumor[9], virus[10], bacterial[11] and other diseases. Moreover, making full use of their catalytic ability, nano-properties, and easy modification to make sophisticated multimodal probes is also worthy of attention. For example, utilizing the photothermal effect and peroxidase (POD)-like activity of bifunctional Au@Pt nanozyme, Jiao *et al*. created a novel multimodal enzyme-linked immunosorbent assay (M-ELISA).[12] Multimodal detection has a good diagnostic effect and has great advantages in the integration of diagnosis and treatment, but the complexity of their synthesis often limits their clinical application. As a kind of material with both enzymatic activities and nano-effects, nanozyme has become a promising tool for multimodal detection by simple synthesis and modification. In general, nanozymes have many advantages and play a vital role in improving human life, which would attract the attention of more and more researchers. It is of great significance to deeply understand and effectively develop nanozymes.

However, the difficulty of developing high-performance nanozymes has greatly hindered the development and application of nanozymes. More tools that may help us to design nanozymes with better performance are badly needed. Machine learning (ML) has attracted the attention of some researchers for its robust learning and design ability.

As a powerful and smart tool, ML has been developed rapidly in recent years. ML is an artificial intelligence technique that lets computers learn and analyze data, thus the resulting predictive models could better perform tasks or predict outcomes. By autonomously learning from data, computers will become more accurate and efficient at predicting specific outcomes.

The basic idea of ML is to learn and analyze existing data to get a model, and then use the model to predict or classify unknown data. In this process, ML algorithms constantly try different models and combinations of parameters to find the best model for solving a particular problem.

The core of ML is the algorithm that trains the model. ML algorithms can be divided into supervised learning, unsupervised learning and reinforcement learning. Supervised learning requires labeled data, that is, the data set contains the correct answers. Unsupervised learning does not require labeled data, and its goal is to find some inherent structure and pattern in the data. In addition, reinforcement learning is a way to train a model through a reward mechanism, learning how to maximize the expected reward through trial and error. Common ML algorithms include linear regression, decision tree, support vector machine, neural network, *etc*. These algorithms can be applied to various fields, showing different prediction effects.

ML has been widely used in the field of life science. One and the most famous application is AlphaFold, which allows scientists to predict protein structure with high accuracy.[13] AlphaFold is only a representative application of ML in protein structure prediction, which falls under the field of bioinformatics. In addition, ML is also widely used in evolutionary analysis[14], genomic data analysis[15, 16] and other bioinformatics research. When combined with medicine, ML has demonstrated exceptional capabilities in biomedical image analysis[17, 18], biomedical engineering[19], medical diagnostics[20-22], drug development[23, 24], healthcare [25], *etc*. When working with big and multimodal data, ML often offers superior performance over other tools. Biological information is often complex and high-throughput, which makes the application of ML in the field of life science logical.

Prediction of catalyst performance has become one of the most important applications of ML, which has generally performed well in established practices.[26, 27] When applied in the field of nanozymes, ML not only helps us to rationally design nanozymes and deepen our understanding of nanozymes but also helps us solve more problems that may be encountered in the design and application of nanozymes (Scheme 1).

2. Applications of ML in the rational design of nanozymes

In the process of the discovery and development of nanozymes, the related fields have experienced many breakthroughs. The early discovery of nanozyme came from an accidental study. Our group found that Fe_3O_4 nanoparticles, which are usually considered to be biochemically inert, exhibited a catalytic ability similar to that of horseradish peroxidase (HRP). Based on this discovery, we proposed that magnetite nanoparticles possess intrinsic enzyme-like activity, which inspired and initiated the research on nanozymes.[3] In the early stage, the study of nanozymes remained at such the level of a combination of chance, experience, trial and error, seriously hindering the development of high-performance nanozymes. The catalytic activity of the developed nanozymes was typically much lower than that of natural enzymes, which limits their practicability. Therefore, it is imperative to design nanozymes rationally instead of randomly synthesizing them.

As the study progresses, researchers have made exploratory attempts in many directions. For example, by studying the structure-activity relationship, it was revealed that the catalase (CAT)-like activity of ferrihydrite nanozyme was mainly determined by the number of abundant surface iron-associated hydroxyl groups, which provides a reference for the improvement of catalytic activity of ferrihydrite.[28] The activity of the new nanozyme was predicted based on the density functional theory (DFT) calculations developed in the field of materials, which was verified by experiments. With the screening conditions calculated by DFT, nanozymes with good activity were found with computer-aided screening.[29] Moreover, some scientists are working on using chemical theories to predict the catalytic activity of nanozyme. For instance, metal-organic framework (MOF) nanozymes were found to follow a Hamment-type structure-activity relationship, and their oxidase (OXD)-mimicking activities could be enhanced by modulating the Hammett σm value.[30] The relationship between the number of e_g occupancy and the catalytic activity of perovskite oxide-based POD-like nanozyme was also revealed using Sabatier's principle. [31] Apart from these aspects, the design inspired by natural enzymes is also an important and effective way to improve the catalytic capacity of nanozymes. Some groups aim to design highly active nanozymes based on simulating the structures of natural enzymes, including the active centers, amino acid microenvironments, and coordination structures.[6, 32, 33] By mimicking natural enzymes, researchers prepared a series of single-atom nanozymes (SANs), which are known as nanozymes with isolated single metal atoms on each active site. SANs provide enough sites for reactants to bind and react and significantly improve the atomic utilization efficiency of nanozyme, making it possible to greatly enhance the catalytic efficiency of nanozymes.[10] SANs have been paid more and more attention because of their superior properties and have become an important research direction of the current nanozyme research.

In the development of SANs, the activity enhancement process of iron-based SANs is particularly noteworthy. Imitating heme iron, the bioinspired single-atom iron nanozymes typically have a $Fe-N_4$ coordination structure, which tends to exhibit a variety of enzyme-mimicking activities that include POD-like and OXD-like activities (Figure 1a).[6] Other researchers found that $Fe-N_5$ SAN has superior OXD-like catalytic activity (Figure 1b).[34] When one of the four N atoms coordinated with Fe was replaced by P, the kinetic constant of $Fe-N_3P$ nanozyme was an order of magnitude higher than those of the natural HRP ($k_{cat}/K_M \approx 2.34\times10^6$ $M^{-1}s^{-1}$), which made it a better artificial enzyme (Figure 1c).[35] It can be seen that the structure of the active center is a determined characteristic of the catalytic activity of SANs, and precise engineering of the active center greatly improves the catalytic ability. As a kind of nanozymes with a regularly distributed active center, the type and distribution of metal atoms in the active center of SANs can be simply represented by numbers, so they are particularly suitable to be predicted by ML. SANs were also one of the first nanozymes to be used for ML prediction. In 2020, ML as well as the database obtained by DFT calculation were used by Zhou *et al.* to screen the SANs with better performance. They predicted the adsorption Gibbs free energy change of H using the XGBR algorithm and screened catalysts for CO_2 electroreduction with potential optimal performance from 1060 designed SANs. Moreover, a method for nanozyme discovery through ML and DFT database was established.[36] This method is relatively complete and mature and is still widely used.

The application of ML in the field of nanozymes is still in the development stage, but it has preliminarily shown its powerful performance. The introduction of ML into the field of rational design of nanozymes is one of the major events in such field, which lays a foundation for the development of various high-performance nanozymes, and even could contribute to a deeper understanding of the features and catalytic mechanism of nanozymes.

2.1 Evaluate the kinetic properties of nanozymes

ML can be involved in the screening of nanozymes by evaluating their kinetic properties and intuitively evaluating their activity and selectivity. A variety of enzyme kinetics descriptors, such as K_M, V_{max}, k_{cat}, k_{cat}/K_M, *etc*., can be conveniently predicted by ML.

Wei *et al.* used the ten most important features of nanozymes (i.e., enzyme type, metal type, metal proportion, metal value, shape, *etc.*) to predict the four commonly used kinetic parameters of nanozymes (K_M, V_{max}, k_{cat}, k_{cat}/K_M), and obtained the nanozymes with optimal catalytic activity. The accuracy of the prediction was 90.6% and the maximum R^2 was 0.80 (Figure 2a). The predicted results were in agreement with the experimental results, indicating that using ML to predict the kinetic constants of nanozymes is reliable. By using SHapley Additive exPlanations (SHAP) to evaluate the contribution of input data, the metal type was found to have the greatest influence on the activity of nanozymes. This work deepened the understanding of researchers on metal-based nanozymes.[37]

2.2 Evaluate the thermodynamic properties of nanozymes

ML can also be used to predict the energy barrier and other thermodynamic properties of nanozyme-catalyzed reactions, and evaluate the difficulty of catalyzing these reactions. These thermodynamic data are usually obtained by theoretical calculation such as DFT rather than experiment. However, as a theoretical method that needs a large amount of computation, DFT calculation usually requires the assistance of a high-performance computer and its long-time operation, which poses a great challenge to its wide application.[38] Fortunately, ML can be used to replace DFT calculation, which can estimate the energy barrier with higher efficiency and accuracy to screen out the catalytic materials with the best activity. [9] The prediction accuracy of ML models trained with precise electron-related quantum data or experimental data may also be higher than that of mixed DFT.[39] Another research found that after being trained on a database of thousands of DFT calculations, the resulting ML model predicted the thermodynamic data of material components in six orders of magnitude less time than direct DFT calculations (Figure 2b).[40] This is a relatively mature and efficient method, which has been widely used in several related studies.

Yu *et al.* performed ML-assisted DFT calculations using two atom descriptors (electronegativity, dopant atom radius) and three model descriptors (dopant atom location, dopant atom concentration, and band gap) to predict the maximum energy barrier and energy consumption steps of the reactions catalyzed by metal-doped graphdiyne-based nanozymes. The authors evaluated and screened the performance of different models by using the database to get the best performance model. The obtained extreme gradient boosting (XGB) model had excellent performance (R^2=0.781), and the predicted results were consistent with the experimental results (Figure 2c).[41]

A similar approach was also used in the study conducted by Zhang *et al.* Based on materials information as well as DFT calculation, they predicted the Gibbs reaction free energy of superoxide dismutase (SOD)-like reaction steps catalyzed by different thiophosphate coordinated by different transition metal atoms ($M_xP_yS_z$; x=1-7, y=1-4, z=1-29). The obtained $MnPS_3$ showed the strongest free radical scavenging and SOD-like activity both in human skin fibroblast cells and in a mouse model of Androgenetic alopecia (AGA).[42]

2.3 Analysis of the enzyme-mimicking activities of nanozymes

ML is also widely used to predict the types of enzymatic activities of nanozymes. One approach is to digitize the different catalytic activities and predict them directly using representative intrinsic and extrinsic factors of nanozymes. In the study conducted by Huang's group, the prediction accuracy for the POD-like performance of nanozymes using ML was 94.5%, which proved that the prediction accuracy of the group with more training data would be significantly higher.[37]

Another method is to predict the enzymatic types and their specificity with the physical descriptors of enzyme-like activities. For example, Gao *et al.* found a criteria through physical analysis, that is, the total energy change for the adsorption of a hydroxyl radical (OH•) or hydrogen radical (H•) onto the material surface in the gas phase ($E_{ads,OH}$ and $E_{ads,H}$, respectively) can be used to predict the enzymatic types of nanomaterials (i.e. the nanozymes exhibit single CAT-like or combined POD-like activity). Therefore, with the $E_{ads,OH}$ and $E_{ads,H}$ database of nanozymes calculated by DFT, ML-assisted mining of 2D nanozymes with exclusive CAT-like activity was realized (Figure 3a).[43]

At present, many nanozymes with different enzyme-mimicking activities have been discovered, including oxidoreductase-like, hydrolase-like, lyase-like, and isomerase-like nanozymes.[4] The activity types of nanozyme are mostly oxidoreductase-like and hydrolase-like, which limits the practical application of nanozymes to a greater extent. With the help of physics and chemistry, ML may help the scientific community to better understand the complex and specific relationship between the enzyme-mimicking activities of nanozymes and their structure (including surface lattice and composition, *etc*.). Thus, the current dilemma of lacking catalytic types of nanozymes will be overcome.

2.4 Assisting in the structure study of nanozymes

It is also feasible for ML to be directly used in the structural analysis of nanozymes. With ML based on a global exploration algorithm of configuration space and DFT calculations, the structure of $MgCl_2$ heterogeneous nanozyme coated by $TiCl_4$ was optimized and predicted by Toshiaki's group without prior knowledge (Figure 3b). The stable and metastable structures of the nanozyme were obtained. The existence of a variety of heterogeneous structures indicates that the nanozyme can be interconvertible to a large extent. This study successfully constructed realistic models of complex nanozymes through ML and increased the understanding of the properties of corresponding nanozymes.[44] Using the combination of stochastic surface walking (SSW) global optimization with the neural network (NN) method (SSW-NN) invented by Liu's Group in 2017[45], the stable structures and their transition states of the MoS_2-supported Cu single-atom nanozyme with sulfur vacancies (Cu@MoS_2-Vs) with robust POD-like activity was successfully explored. Based on the simulation results, it is also proved that the increase of Cu loading and sulfur vacancies, and the decrease of microenvironment pH value can significantly improve the performance of Cu@MoS_2-Vs.[46]

2.5 Assisting in the exploration of structure-activity relationships of nanozymes

Another approach is to use ML to reveal the relationship between the structure of nanozymes and their catalytic activity. In addition to the direct prediction of nanozyme catalytic activities and enzyme-mimicking types, ML also has good performance for atomic simulation and evaluation of candidate catalyst structures. Li *et al.* applied stochastic surface walking based on the global neural network potential (SSW-NN) method and DFT calculations to fast global potential energy surface (PES) exploration. By this method, they determined the phase diagrams for the bulk and surface of PdAg catalyst under reaction conditions, revealing the effects of different phase compositions and surface structures on the catalytic activity (Figure 3c). Based on their findings on the influence of *in situ* surface structures on PdAg catalyst, they designed a rutile-supported Pd_1Ag_3 catalyst that showed the best catalytic performance with > 96% conversion and > 85% selectivity.[47]

2.6 Assisting in the discovery of new nanozymes through energy computational acceleration

Traditional ML relies heavily on training data sets. The quantity and quality of data directly determine whether the prediction performance is good or not. However, by combining the ML, DFT, and discovery algorithms, we can get rid of this dependence, realize unbiased data generation, and greatly accelerate the discovery of nanozymes with similar properties. For example, Jennings *et al.* discovered the composition and structure of Pt_xAu_{147-x} icosahedral nanoparticles with an optimal activity using ML-accelerated genetic algorithms (GA) (Figure 4a). Working as an energy predictor, ML reduces the amount of computation required by GA and DFT calculations by 50 times, showing a good effect for accelerating search algorithms.[48]

2.7 Analyzing the influencing features

In the process of using ML to screen nanozymes, some methods can be used to conveniently analyze the main factors affecting their catalytic activity.

One approach is to construct ML using different factors and compare the performance of these models to determine the most important factors. However, this approach is not optimal due to its complexity and may lead to misjudgment. More often, the interpretation module of ML is used to provide the interpretation of the models.

Some models evolved from the base model to predict performance gains while retaining the strong interpretability of the base model, which makes them more widely used. For instance, the random forest can analyze the importance of multiple factors in the input data and output the influence of different factors, so that we can understand which factors have the greatest influence on the activity of nanozymes, and then help us rationally design them. For example, the random forest model was used to predict the ammonia conversion for RuMK (M is the secondary metal) catalysts, which showed that the reactor temperature was the most important factor affecting the predicted activity (Figure 4b). This result is consistent with previous knowledge, proving that the random forest model has a robust feature ranking ability.[49]

For the black-box models that are difficult to explain, some external interpretation modules that do not rely on ML models to explain them are quite important. A reliable choice is Local Interpretable Model- Agnostic Explanations (LIME). LIME is a powerful interpretation technique that can interpret any model, providing solutions to model selection, model improvement, and model trust issues. The basic interpretation idea is to get the prediction results by locally perturbing the sample input, and then use the new data points obtained in this way to train the decision tree and other white-box models so that the black-box algorithm can get some interpretation with the help of trees. It can explain the local structure and results of the model with good interpretation performance.[50] As an improved method of LIME, SHAP is more commonly used. SHAP calculates a Shapley value for each feature to measure its contribution to the predicted value.[51] By applying SHAP, Zhang *et al.* studied the influence of different atomic radii and lattice factors on the formability of hybrid organic perovskite (HOIPs) (Figure 4c).[52] Analysis of the influencing factors of POD/OXD activity by SHAP showed that metal factors had the greatest influence on it, which was the same as the result obtained by the classification model.[37]

2.8 Deepen our understanding of nanozymes

On the one hand, ML combined with interpretation modules can explore the features that have the greatest influence on nanozymes, which is conducive to understanding the structure-activity relationships of nanozymes.[53] The assistance of ML can avoid the subjective bias of researchers that may hinder experimental design and the discovery of important factors.[54] On the other hand, various structural motifs and functional groups within the surface or structure of nanozymes can be explored through ML combined with materials science calculations, such as the calculation of interatomic potentials.[55] This approach has a level of detail that DFT simulations cannot achieve. This will enhance our understanding of the mechanisms by which nanozymes exert their catalytic effects. All of these suggest that further development of ML in the field of nanozyme will be of great benefit to our understanding of nanozymes.

2.9 Helping with other aspects in the process of nanozyme design and application

What should be noticed is that in practice ML could be used for purposes other than those mentioned above: classify the candidates for the potential nanozymes[56], directly learn and output some data related to the catalytic performance of nanozymes, clarify the actual catalytic mechanism, and predict reactive trajectories (Figure 4d) [57], *etc*. In addition, how to make use of ML for image processing and powerful visualization to facilitate the development and application of nanozymes is still being explored.

3. Key ingredients affecting the predictive power of ML for nanozyme design

3.1 Collecting and processing of data set

The quantity and quality of samples have a great impact on the predicted results. Powerful data sets are the basis of the prediction accuracy of ML models, which largely determines the prediction performance of models. ML research generally requires 100 to 10000 training data points.[58] Such a huge number of instances and a high degree of consistency put forward high requirements for data collection. Current data set collection methods mainly include sorting out a large number of experimental data in the literature and obtaining highly consistent data through First-principles calculations. Both of them require a lot of work and almost inevitably introduce errors. It is also a promising approach to create more high-throughput nanozyme synthesis and characterization techniques through the aid of flow reactors, robots, or computer vision to construct large experimental data sets with similar operating conditions, which can simultaneously overcome both statistical and computational errors.

Different data collection methods face different problems. The data sets obtained from published literature are typically from real experiments, with high credibility but poor consistency. The experimental conditions or statistical recording methods inevitably introduce inestimable errors in the data. Counting such a large amount of data and recording it in the same format requires also a lot of work. Although it is better and easier to use existing databases, there is still a lack of relevant databases in the field of nanozymes. First-principles calculations (or DFT) is also a common and useful approach for data set collection, but it is computationally intensive and the calculation number they need increases dramatically as feature dimensions increase. In addition, DFT is an approximate algorithm, of which the calculation results generally have a gap with the real-world data. The accumulation of errors in the calculation process may lead to more problems and even affect the final prediction results.[59]

Whether the data is obtained from experiments, literature, database, or DFT calculation, the samples should be sorted out, screened, and classified. First, the data should include features that have an important impact on catalytic activity. Otherwise, the predictive effect of the models may be significantly weakened. Common factors include the characteristics of the active sites (*e.g.*, atomic number, atomic radius, d/f orbital electron number, d-band, *etc*.), atomic location factors (such as the active site of atoms doped site), microenvironment characteristics (such as the characteristics of carrier material, effective coordination number, *etc*.), the substrate binding energy, dissociation energy, *etc*. These factors deeply affect the catalytic ability of nanozymes and play an important role in predicting their catalytic activity. These features should be easily obtained from experimental data, mature databases, or DFT calculations. Moreover, selecting the most important features is also the key to producing reliable data sets. Extensive features may contain some redundant ones that are useless and may hinder or even bias the prediction results.[60] For example, they may cause model overfitting and reduced model predictive power.[61] Notably, some ML models could be used for feature selection when researchers are unable to determine the most important features. For example, random forest classifier has the optimal performance for feature selection and are widely used.[62] In addition, using a correlation matrix to select less relevant features is a sensible way. Zhang *et al*. select seven important features including the Gibbs reaction free energy gotten from DFT through the application of correlation matrix as well as Pearson's correlation coefficient, which is helpful to the discovery of SOD-like nanozymes.[42] It is also a good choice to select the best one by getting mature models according to different features and compare their performance, which is conducive to the selection of a better model, but relatively more time-consuming and laborious.

All samples for ML should be presented in the form of numerical values. For some non-quantitative instances, researchers need to manually score them or use some algorithms, such as LabelEncoder, to assign scores to them.[63] Bad data creates problems that no good algorithm can remedy. When collecting data, we should also attach great importance to their credibility, usability, and consistency, which may affect the accuracy of prediction results.

3.2 Data set division

For one-shot ML, the evaluation and selection of models are typically done using one data set, which contains a limited number of instances. Therefore, the data set needs to be properly processed to get a training set and a testing set, so that training and testing can be performed at the same time. There are many ways to partition a data set, the simplest of which is called "hold-out". "Hold-out" directly divides a data set into two mutually exclusive sets. This method is limited by many aspects, such as data generalization errors. The more commonly used method is cross-validation. A traditional type of cross-validation, called hold-out cross-validation, divides the data set into a training set, validation set, and testing set with a proportion of typically 60%, 20%, and 20%. Then the model is trained with the training set, and different models and parameters are repeatedly tested under the validation set to find a satisfactory one. Finally, the testing set is used for error evaluation. At present, the k-fold cross-validation method is more commonly used, that is, the data set is equally divided into k-folds with consistent data distribution without replacement, among which the k-1 group is used for model training, and the remaining groups are used for model capability evaluation. By repeating this process k times, k training and testing sets are obtained to fully evaluate the model and its performance. By comparing the k trained models, the relatively optimal parameters can be selected or calculated, and then the new model can be trained on the whole data set. Moreover, with the testing dataset, the performance differences between different models can be compared horizontally, which helps to select the appropriate model. Sun *et al.* used 20-fold cross-validation to train all of the models in their work, getting the predicted results that had fewer outliers and good accuracy.[64] 10-fold cross-validation was used to test the predictive power of the sure independence screening sparsifying operator (SISSO) model and find the optimal dimensionalities for each of the 3 descriptors.[61] Through grid-search optimization combined with cross-validation, Ding *et al.* optimized 54 classification models and 18 regression models built from 9 ML algorithms and got the feature importance distribution results that affect the chemical properties of catalysts.[54]

3.3 Application of suitable evaluation indicators

For different models, evaluation, and purpose requirements, different evaluation indicators need to be applied to measure their performance to indicate their predictive ability. The frequently used error rate and accuracy could help us measure performance, which respectively represents the proportion of the number of samples with correct and incorrect predictions in the total number of samples. However, these indicators alone are far from sufficient.

There are a variety of evaluation indicators and evaluation methods available. For regression models, such as those predicting catalytic constants for nanozymes, the most commonly used performance metric is mean squared error. In particular, in specific models, the mean squared error is directly applied as part of themselves. For example, for Deep Neural Network (DNN), a commonly used ML model, there is a unique performance evaluation method, namely the loss function. Common loss functions include the coefficient of determination value and mean squared error, *etc*. The DNN model is constantly improved through real-time feedback, which is also the reason for its strong performance.[65]

For specific purposes and contexts, such as predicting the type of catalytic activity of nanozymes, we can use classification models for which several performance measures have been invented. Most of these performance measures are based on confusion matrix. Confusion matrix can be used for intuitive observation of the relationship between the predicted results of the model and the real results, and is commonly used in research. Confusion matrix columns true positive, false positive, true negative, and false negative as matrices. By observing which categories are prone to confusion, the model can be modified specifically to improve the overall prediction accuracy of the model. The good confusion matrix of the data is also one of the powerful pieces of evidence to prove the classification and prediction ability of models.[37]

Based on the confusion matrix, we can calculate many commonly used metrics, such as precision, recall rate, specificity, and F1-measure that comprehensively evaluates performance precision and recall. Taking recall as abscissa and precision as ordinate, P-R curve can be made to visually show the recall and precision of the learner in the sample population. In addition, the Break-Even Point can also be obtained at the position of the curve where recall is equal to the precision. The larger its value is, the better the learner is. F1 measure is the harmonic mean of precision and recall, which has the same function as the Break-Even Point but is more commonly used.

The Receiver Operation Characteristic (ROC) curve, which takes a false Positive Rate as X-axis and a true Positive Rate as Y-axis, is also one of the most commonly used evaluation methods. A greater radian of the ROC curve (the area under the ROC curve) indicates a higher sensitivity, a lower false positive rate, and a better prediction performance of the model. The ROC curve can ignore the sample imbalance, so it is very commonly used in model evaluation. Using k-means combined with Area Under ROC Curve (AUC), researchers evaluated and compared the influence of different structural and dynamic drivers on the catalytic activities of nanozymes. The factors that had the most obvious influence (maximum AUC) were screened out, proving that conformational descriptors are sufficient to predict reactivity.[57]

It is important to note that different performance measures can lead to different results when comparing the capabilities of different models. Any model has its advantages and disadvantages. Therefore, when evaluating, we need to select the most appropriate performance indicators to evaluate the model according to the data, algorithms, and requirements.

3.4 Experimental verification

ML has powerful computing power, which can achieve high-throughput screening of a large number of materials. However, no matter how ML develops and how its predictive ability is enhanced, it cannot replace the position of researchers in the experimental design, screening, and validation of nanozymes. Any prediction methods using ML are only estimation and prediction, but not substitutes for experimental results. After the prediction of nanozymes by ML, adequate experimental verification is still needed, which is also an essential step for model evaluation.

4. Conclusions and prospects

The rational design of nanozymes is the basis of improving the performance of nanozymes and promoting their applications. However, as a kind of artificial enzyme with unclear catalytic mechanism, it is not easy to design their structure reasonably. Fortunately, we are in an era of big data with rapid computer development, which gives us powerful tools like ML. ML allows us to avoid repeated trial and error and directly obtain reasonable predictions of the performance of new nanozymes. As a powerful data mining tool, ML is applied to the rational design of nanozymes and other related fields. It will help us to better understand nanozymes, thus promoting their rapid and better development. Of course, ML still has some inevitable problems, such as over-reliance on a certain amount of reliable data, greatly influenced by model selection and refinement, and difficult to match predictive performance and explanatory ability. However, these problems may be gradually solved in the development of ML applications in related fields. For the combination of nanozymes and ML, there are still some other questions we should pay attention to in the future:

(1) Establishment and development of nanozyme databases. ML is a high-throughput research method that relies entirely on a huge amount of training data. The training data set guarantees the accuracy of ML models. The quantity and accuracy of data directly determine the quality of prediction results, and too few or poor data may lead to model bias. Moreover, although there have been some mature DFT databases that could be utilized for nanozyme design,[66] the nature of DFT calculation as a kind of approximate estimation method possess certain errors that are inevitable in the progress, which would affect the prediction accuracy. Studies have shown that the prediction

performance error of the ML models trained on experimental data is less than that of the models trained on the DFT database, suggesting that the ML models trained on experimental data or more accurate thermodynamic data may bring greater benefits, while the DFT database training inevitably increases the error.[39] Therefore, it is important and indispensable to develop databases with enough available information on nanozymes. It is worth expecting that currently some nanozyme databases have been successfully constructed (e.g., http://cdeyun.com:6063/ and DiZyme[67]) and are in the process of rapid maturation and improvement. The construction and maturation of the nanozyme database require the joint efforts of researchers in all related fields, together with the development of more efficient, more accurate, and broader coverage methods. A high-quality database can be built only with high-quality and high-consistency experimental data. Therefore, some standard and unified detection and representation methods of activity, structure, and structure-activity relationship, etc., are needed to guarantee the quality of data and database. The harmonization of standards and the development of methods are critical and will affect the future of the entire field of nanozyme rational design. These databases would be used to design new and better nanozymes, which will ultimately benefit every nanozyme researcher.

(2) Design, establishment, remolding, and application of more ML models. ML is an approach based on training data, but the algorithms and methods it employs determine the way and efficiency it extracts information from data sets. Only an appropriate and suitable algorithm could improve the performance of ML and reduce the probability of data misunderstanding. Therefore, the development and application of ML cannot be separated from the development of related algorithms. In the 1980s, the decision tree using the ID3 algorithm, a simple but interpretable algorithm, appeared. In the 1990s, with the development of algorithms, other algorithms for decision trees building such as CART and C4.5 were invented successively, and the classification effect of decision trees was gradually optimized.[68-70] In 1995, the AdaBoost algorithm, which integrates weak classifiers to obtain strong classifiers, was born, and the performance of the decision tree integrated with the AdaBoost idea was greatly improved.[70] In 2001, the random forest was invented. As a new integrated ML algorithm, it randomly trains multiple simple decision trees to obtain an excellent algorithm with high performance and low complexity, which is still widely used today.[71] The development of new algorithms will never stop, and researchers interested in nanozyme rational design and other relative topics should keep pace with the times and continue to try and improve new algorithms and new interpretation models to improve the efficiency, performance, and adaptability of ML in the field of nanozyme. Of course, adding some auxiliary methods to an ML algorithm can significantly improve its efficiency or performance. For example, SWWS and similar methods can assist researchers in the automatic exploration of nanozyme optimal structures.[46] GA, an evolving method that finds optimal solutions based on evaluation, is often combined with ML algorithms such as NN to help researchers to find the best-performing materials.[49, 72] In addition, through manual labeling and filtering of data points and thus continuous cycle training of the model, the active learning method, which could reduce the dependence on data and enhance ML performance, is also increasingly common.[73, 74] These methods, combined with the conventional ML models, make the performance of the models far beyond their upper limit and are well worth trying and applying in the field of nanozymes.

(3) The combination of ML and the theoretical knowledge of nanozyme-related disciplines. At present, ML has been used for direct screening or design of nanozymes, but its accuracy and efficiency still have a lot of room to improve. This is related to the lack of development of the

nanozyme database, but more related to the characteristics of ML, such as learning and simulation. The combination of DFT makes ML overcome the dependence on experimental data, and because of this feature, it is widely used and mature.[74] DFT calculations were originally developed in the field of materials, so we thought that with more knowledge of materials science and biology, ML applications might be more flexible and accurate. For example, Routh *et al.* combined supervised ML with the X-ray absorption near-edge structure (XANES) spectra and found a potential characterization of the relationship between XANES spectra and the physicochemical properties of materials. This opened a new avenue for scientists to understand nanozymes.[75]

(4) Other things we can do with the assistance of ML. ML can do a lot more than aid the rational design of nanozymes. First of all, we could study the *in vivo* fate and biological toxicity of nanozymes with ML.[76, 77] Next, the help of ML allows us to predict the internalization ability of nanozymes to different cells and thus assess the toxicity or bioactivity of nanozymes. [77] In addition, ML combined with nanozymes can also have a much broader application prospect. For example, it can be used to predict the effect of additives on nanozyme activity during the synthesis and application process. In one study, Guo *et al.* developed a predictive study on the effects of additives on Cu catalysts selectivity in the ML-guided way.[78] A high-resolution holographic image acquisition method was constructed and the catalyst nanoparticles were statistically analyzed.[79] It also shows robust prediction ability in the aspect of nanomaterial self-assembly, and more and more related applications will be developed.[80]

High-throughput theoretical screening is a robust method for nanozyme rational design. A large amount of computation required determines the irreplaceable role of computer assistance in it. The powerful data processing, understanding, learning, and prediction performance of ML makes it a powerful tool for the rational design of nanozymes. Using ML for nanozyme design is an effective way to enhance the reactivity, efficiency, and selectivity of nanozymes, avoid duplicate experiments, and save energy and resources.[27] This would become one of the ideal directions for the rational design of nanozymes.

Author contributions

Yucong Li: writing - original draft preparation. Ruofei Zhang: writing - review and editing. Xiyun Yan: supervision. Kelong Fan: conceptualization, supervision, writing - review and editing.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China (No. 82122037), Youth Innovation Promotion Association of Chinese Academy of Sciences (2019093), CAS Interdisciplinary Innovation Team (JCTD-2020-08).

References

1. J. Chapman, A. E. Ismail and C. Z. Dinu, *Catalysts*, 2018, **8**, 238.
2. H. Wei and E. Wang, *Chemical Society Reviews*, 2013, **42**, 6060-6093.
3. L. Gao, J. Zhuang, L. Nie, J. Zhang, Y. Zhang, N. Gu, T. Wang, J. Feng, D. Yang and S. Perrett, *Nature nanotechnology*, 2007, **2**, 577-583.

4. C. Hong, X. Meng, J. He, K. Fan and X. Yan, *Particuology*, 2022, **71**, 90-107.

5. R. Zhang, X. Yan and K. Fan, *Accounts of Materials Research*, 2021, **2**, 534-547.

6. J. Xi, R. Zhang, L. Wang, W. Xu, Q. Liang, J. Li, J. Jiang, Y. Yang, X. Yan and K. Fan, *Advanced Functional Materials*, 2021, **31**, 2007130.

7. Q. M. Chen, X. D. Zhang, S. Q. Li, J. K. Tan, C. J. Xu and Y. M. Huang, *Chemical Engineering Journal,* 2020, **395**, 125130.

8. R. G. Mahmudunnabi, F. Z. Farhana, N. Kashaninejad, S. H. Firoz, Y. B. Shim and M. J. A. Shiddiky, *Analyst*, 2020, **145**, 4398-4420.

9. Q. Liang, J. Xi, X. J. Gao, R. Zhang, Y. Yang, X. Gao, X. Yan, L. Gao and K. Fan, *Nano Today*, 2020, **35**, 100935.

10. D. Wang, B. Zhang, H. Ding, D. Liu, J. Xiang, X. J. Gao, X. Chen, Z. Li, L. Yang, H. Duan, J. Zheng, Z. Liu, B. Jiang, Y. Liu, N. Xie, H. Zhang, X. Yan, K. Fan and G. Nie, *Nano Today*, 2021, **40**, 101243.

11. Z. Jia, X. Lv, Y. Hou, K. Wang, F. Ren, D. Xu, Q. Wang, K. Fan, C. Xie and X. Lu, *Bioactive materials*, 2021, **6**, 2676-2687.

12. L. Jiao, L. Zhang, W. Du, H. Li, D. Yang and C. Zhu, *Nanoscale*, 2019, **11**, 8798-8802.

13. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek and A. Potapenko, *Nature*, 2021, **596**, 583-589.

14. J.-x. Fan, S.-z. Shen, D. H. Erwin, P. M. Sadler, N. MacLeod, Q.-m. Cheng, X.-d. Hou, J. Yang, X.-d. Wang and Y. Wang, *Science*, 2020, **367**, 272-277.

15. C. Weinreb, A. Rodriguez-Fraticelli, F. D. Camargo and A. M. Klein, *Science*, 2020, **367**, eaaw3381.

16. C. Y. Huang, C. J. Cassidy, C. Medrano and J. T. Kadonaga, *Nature*, 2020, **585**, 459-463.

17. Y. Zhou and Q. Meng, *Journal of Clinical Oncology*, 2020, **38**, e16095.

18. N. Dolensek, D. A. Gehrlach, A. S. Klein and N. Gogolla, *Science*, 2020, **368**, 89-94.

19. G. M. Anand, H. C. Megale, S. H. Murphy, T. Weis, Z. Lin, Y. He, X. Wang, J. Liu and S. Ramanathan, *Cell*, 2023, **186**, 497-512.

20. C. J. Lynch and C. Liston, *Nature medicine*, 2018, **24**, 1304-1305.

21. K. Swanson, E. Wu, A. Zhang, A. A. Alizadeh and J. Zou, *Cell*, 2023, **186**, 1772-1791.

22. S. Warnat-Herresthal, H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers and N. A. Aziz, *Nature*, 2021, **594**, 265-270.

23. J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah and M. Spitzer, *Nature reviews Drug discovery*, 2019, **18**, 463-477.

24. M. Eisenstein, *Nature biotechnology*, 2022, **40**, 1303-1305.

25. K. Y. Ngiam and W. Khor, *The Lancet Oncology*, 2019, **20**, e262-e273.

26. D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher and A. G. Doyle, *Science*, 2018, **360**, 186-190.

27. A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow and S. E. Denmark, *Science*, 2019, **363**, eaau5631.

28. R. Zhang, L. Chen, Q. Liang, J. Xi, H. Zhao, Y. Jin, X. Gao, X. Yan, L. Gao and K. Fan, *Nano Today*, 2021, **41**, 101317.

29. Z. Wang, J. Wu, J.-J. Zheng, X. Shen, L. Yan, H. Wei, X. Gao and Y. Zhao, *Nature Communications*, 2021, **12**, 6866.

30. J. Wu, Z. Wang, X. Jin, S. Zhang, T. Li, Y. Zhang, H. Xing, Y. Yu, H. Zhang and X. Gao, *Advanced Materials*, 2021, **33**, 2005024.

31. X. Wang, X. J. Gao, L. Qin, C. Wang, L. Song, Y.-N. Zhou, G. Zhu, W. Cao, S. Lin, L. Zhou, K. Wang, H. Zhang, Z. Jin, P. Wang, X. Gao and H. Wei, *Nature Communications*, 2019, **10**, 704.

32. J. Wang, R. Huang, W. Qi, R. Su, B. P. Binks and Z. He, *Applied Catalysis B: Environmental*, 2019, **254**, 452-462.

33. M. Li, J. Chen, W. Wu, Y. Fang and S. Dong, *Journal of the American Chemical Society*, 2020, **142**, 15569-15574.

34. L. Huang, J. X. Chen, L. F. Gan, J. Wang and S. J. Dong, *Science Advances*, 2019, **5**, eaav5490.

35. S. F. Ji, B. Jiang, H. G. Hao, Y. J. Chen, J. C. Dong, Y. Mao, Z. D. Zhang, R. Gao, W. X. Chen, R. F. Zhang, Q. Liang, H. J. Li, S. H. Liu, Y. Wang, Q. H. Zhang, L. Gu, D. M. Duan, M. M. Liang, D. S. Wang, X. Y. Yan and Y. D. Li, *Nature Catalysis*, 2021, **4**, 407-417.

36. A. Chen, X. Zhang, L. Chen, S. Yao and Z. Zhou, *The Journal of Physical Chemistry C*, 2020, **124**, 22471-22478.

37. Y. Wei, J. Wu, Y. Wu, H. Liu, F. Meng, Q. Liu, A. C. Midgley, X. Zhang, T. Qi, H. Kang, R. Chen, D. Kong, J. Zhuang, X. Yan and X. Huang, *Advanced Materials*, 2022, **34**, e2201736.

38. Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nature communications*, 2017, **8**, 1-7.

39. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley and O. A. Von Lilienfeld, *Journal of chemical theory and computation*, 2017, **13**, 5255-5264.

40. B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary and C. Wolverton, *Physical Review B*, 2014, **89**, 094104.

41. Y. Yu, Y. Jiang, C. Zhang, Q. Bai, F. Fu, S. Li, L. Wang, W. W. Yu, N. Sui and Z. Zhu, *ACS Materials Letters*, 2022, **4**, 2134-2142.

42. C. Zhang, Y. Yu, S. Shi, M. Liang, D. Yang, N. Sui, W. W. Yu, L. Wang and Z. Zhu, *Nano Letters*, 2022, **22**, 8592-8600.

43. X. J. Gao, J. Yan, J. J. Zheng, S. Zhong and X. Gao, *Advanced Healthcare Materials*, 2023, **12**, 2202925.

44. G. Takasao, T. Wada, A. Thakur, P. Chammingkwan, M. Terano and T. Taniike, *Acs Catalysis*, 2019, **9**, 2599-2609.

45. S. Ma, C. Shang and Z.-P. Liu, *The Journal of Chemical Physics*, 2019, **151**, 050901.

46. D. Xu, W. Yin, J. Zhou, L. Wu, H. Yao, M. Sun, P. Chen, X. Deng and L. Zhao, *Nanoscale*, 2023, **15**, 6686-6695.

47. X.-T. Li, L. Chen, C. Shang and Z.-P. Liu, *Journal of the American Chemical Society*, 2021, **143**, 6281-6292.

48. P. C. Jennings, S. Lysgaard, J. S. Hummelshøj, T. Vegge and T. Bligaard, *npj Computational Materials*, 2019, **5**, 1-6.

49. T. Williams, K. McCullough and J. A. Lauterbach, *Chemistry of Materials*, 2019, **32**, 157-165.

50. M. T. Ribeiro, S. Singh and C. Guestrin, *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, 1135-1144.

51. S. M. Lundberg and S.-I. Lee, *Advances in neural information processing systems*, 2017, **30**, 4768–4777.

52. S. Zhang, T. Lu, P. Xu, Q. Tao, M. Li and W. Lu, *The Journal of Physical Chemistry Letters*, 2021, **12**, 7423-7430.

53. S. Mazurenko, Z. Prokop and J. Damborsky, *ACS Catalysis*, 2020, **10**, 1210-1223.

54. R. Ding, Y. Chen, P. Chen, R. Wang, J. Wang, Y. Ding, W. Yin, Y. Liu, J. Li and J. Liu, *ACS Catalysis*, 2021, **11**, 9798-9808.

55. V. L. Deringer, M. A. Caro and G. Csányi, *Advanced Materials*, 2019, **31**, 1902765.

56. M. Sun, T. Wu, Y. Xue, A. W. Dougherty, B. Huang, Y. Li and C.-H. Yan, *Nano Energy*, 2019, **62**, 754-763.

57. B. M. Bonk, J. W. Weis and B. Tidor, *Journal of the American Chemical Society*, 2019, **141**, 4108-4118.

58. J. A. Esterhuizen, B. R. Goldsmith and S. Linic, *Nature Catalysis*, 2022, **5**, 175-184.

59. H. Bhattacharjee, N. Anesiadis and D. G. Vlachos, *Scientific reports*, 2021, **11**, 1-10.

60. V. Kumar and S. Minz, *SmartCR*, 2014, **4**, 211-229.

61. Z.-K. Han, D. Sarker, R. Ouyang, A. Mazheika, Y. Gao and S. V. Levchenko, *Nature communications*, 2021, **12**, 1-9.

62. J. O. Sinayobye, K. S. Kaawaase, F. N. Kiwanuka and R. Musabe, *2019 IEEE/ACM Symposium on Software Engineering in Africa (SEiA)*, 2019, 1-10.

63. J. Guo, A. Nomura, R. Barton, H. Zhang and S. Matsuoka, *Supercomputing Frontiers: 4th Asian Conference, SCFA 2018, Singapore, March 26-29, 2018, Proceedings 4*, 2018, 179-198.

64. M. Sun, A. W. Dougherty, B. Huang, Y. Li and C. H. Yan, *Advanced Energy Materials*, 2020, **10**, 1903949.

65. M. Tamtaji, H. Gao, M. D. Hossain, P. R. Galligan, H. Wong, Z. Liu, H. Liu, Y. Cai, W. A. Goddard III and Z. Luo, *Journal of Materials Chemistry A*, 2022, **10**, 15309-15331.

66. J. E. Saal, S. Kirklin, M. Aykol, B. Meredig and C. Wolverton, *Jom*, 2013, **65**, 1501-1509.

67. J. Razlivina, N. Serov, O. Shapovalova and V. Vinogradov, *Small*, 2022, **18**, 2105673.

68. J. Quinlan, *Machine learning*, 1986, **1**, 81-106.

69. J. R. Quinlan, *Morgan Kaufmann Publishers Inc*, 1992.

70. Y. Freund, *Information and computation*, 1995, **121**, 256-285.

71. L. Breiman, *Machine learning*, 2001, **45**, 5-32.

72. Z. Rao, P.-Y. Tung, R. Xie, Y. Wei, H. Zhang, A. Ferrari, T. Klaver, F. Körmann, P. T. Sukumar and A. Kwiatkowski da Silva, *Science*, 2022, **378**, 78-85.

73. W. Wang, T. Yang, W. H. Harris and R. Gómez-Bombarelli, *Chemical Communications*, 2020, **56**, 8920-8923.

74. L. Wu, T. Guo and T. Li, *Journal of Materials Chemistry A*, 2020, **8**, 19290-19299.

75. P. K. Routh, Y. Liu, N. Marcella, B. Kozinsky and A. I. Frenkel, *The Journal of Physical Chemistry Letters*, 2021, **12**, 2086-2094.

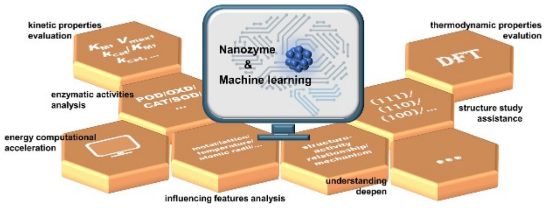
76. A. V. Singh, M. H. D. Ansari, D. Rosenkranz, R. S. Maharjan, F. L. Kriegel, K. Gandhi, A. Kanase, R. Singh, P. Laux and A. Luch, *Advanced Healthcare Materials*, 2020, **9**, 1901862.

77. H. Shi, Y. Pan, F. Yang, J. Cao, X. Tan, B. Yuan and J. Jiang, *Molecules*, 2021, **26**, 2188.

78. Y. Guo, X. He, Y. Su, Y. Dai, M. Xie, S. Yang, J. Chen, K. Wang, D. Zhou and C. Wang, *Journal of the American Chemical Society*, 2021, **143**, 5755-5762.

79. F. Ichihashi, A. Koyama, T. Akashi, S. Miyauchi, K. i. Morooka, H. Hojo, H. Einaga, Y. Takahashi, T. Tanigaki and H. Shinada, *Applied Physics Letters*, 2022, **120**, 064103.

80. E. Vargo, J. C. Dahl, K. M. Evans, T. Khan, P. Alivisatos and T. Xu, *Advanced Materials*, 2022, **34**, 2203168.


kinetic properties evaluation
K_M, V_{max}, k_{cat}/K_M, k_{cat}, ...
Nanozyme & Machine learning
thermodynamic properties evalution
DFT
enzymatic activities analysis
POD/OXD/ CAT/SOD/ ...
{111}/ {110}/ {100}/ ...
structure study assistance
energy computational acceleration
metal/lattice/ temperature/ atomic radii/ ...
structure-activity relationship/ mechanism
...
understanding deepen
influencing features analysis

Scheme 1 The rational design of nanozymes can be facilitated by machine learning in many aspects.

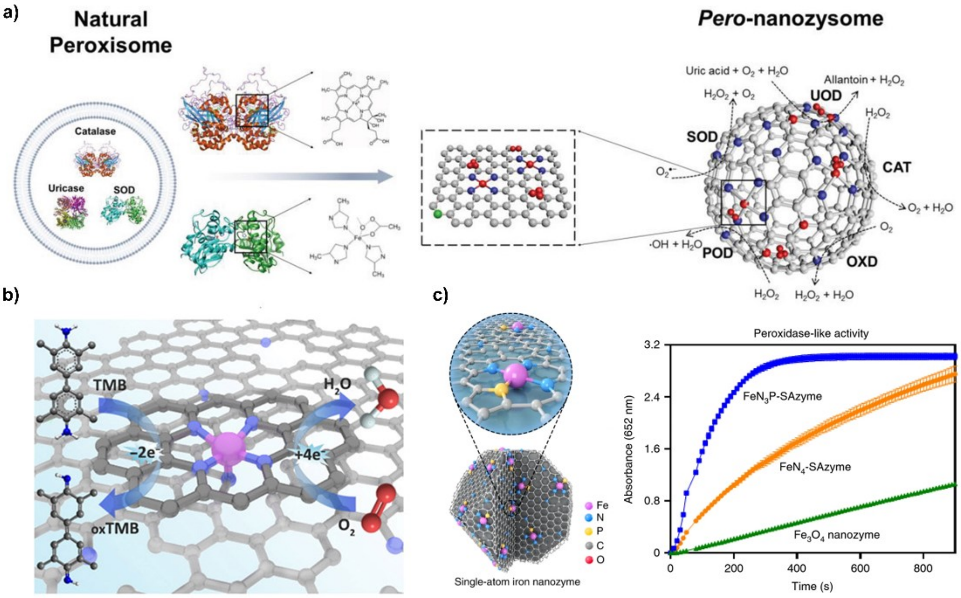


Figure 1

The activity enhancement process of iron-based SANs. a) Mimicking the structure of natural enzyme, the nanozyme exhibits multiple catalytic activities, including POD-like and OXD-like activities. Reproduced with permission.[6] Copyright 2021, Wiley-VCH. b) FeN_5 SAN exhibits OXD-like characteristics. Reproduced with permission.[34] Copyright 2019, AAAS. c) FeN_3P SAN shows better POD-like activity than FeN_4 SAN. Reproduced with permission.[35] Copyright 2021, Springer Nature.

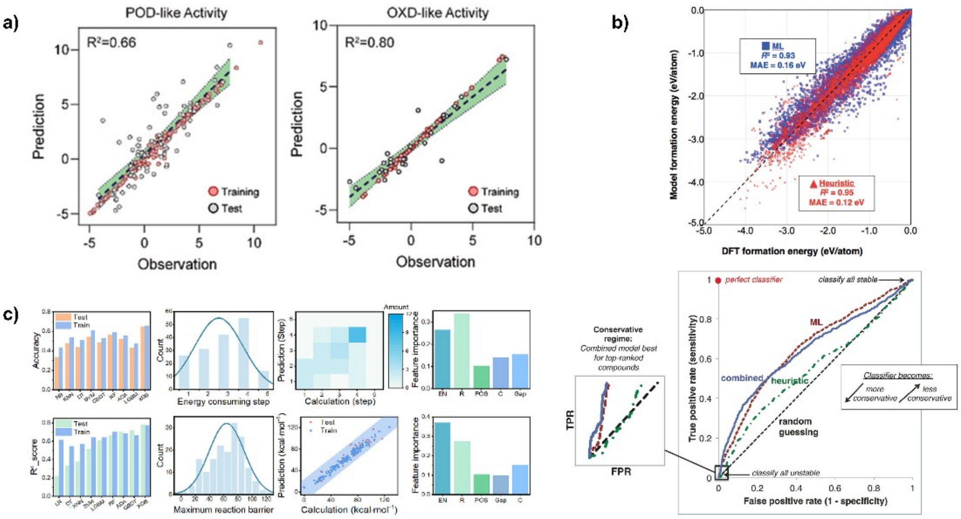


Figure 2

ML assisted evaluation of kinetic properties and thermodynamic properties of nanozymes. a) Prediction accuracy of ML on the catalytic activities of POD and OXD-like nanozymes. Reproduced with permission.[37] Copyright 2022, Wiley-VCH. b) ML exhibits excellent DFT replacement abilities and predicts material properties with high accuracy.[40] Copyright 2014, American Physical Society. c) Using ML to predict the energy consumption step and maximum energy barrier of nanozyme. Reproduced with permission.[41] Copyright 2022, American Chemical Society.

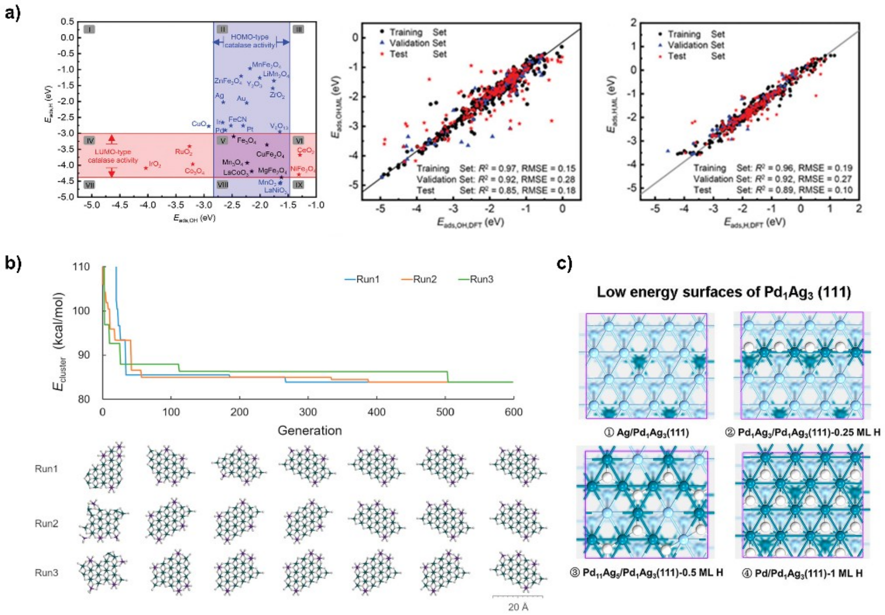


Figure 3

ML assisted evaluation of enzyme-mimicking activities and structure study of nanozymes. a) Physical descriptors were applied to predict the activity types of nanozymes. Reproduced with permission.[43] Copyright 2022, Wiley-VCH. b) Through multiple calculations, ML helps with the structure determine of nanoplates. Reproduced with permission.[44] Copyright 2019, American Chemical Society. c) ML predicts the stable surface configurations of nanozymes under certain reaction conditions.[47] Copyright 2021, American Chemical Society.

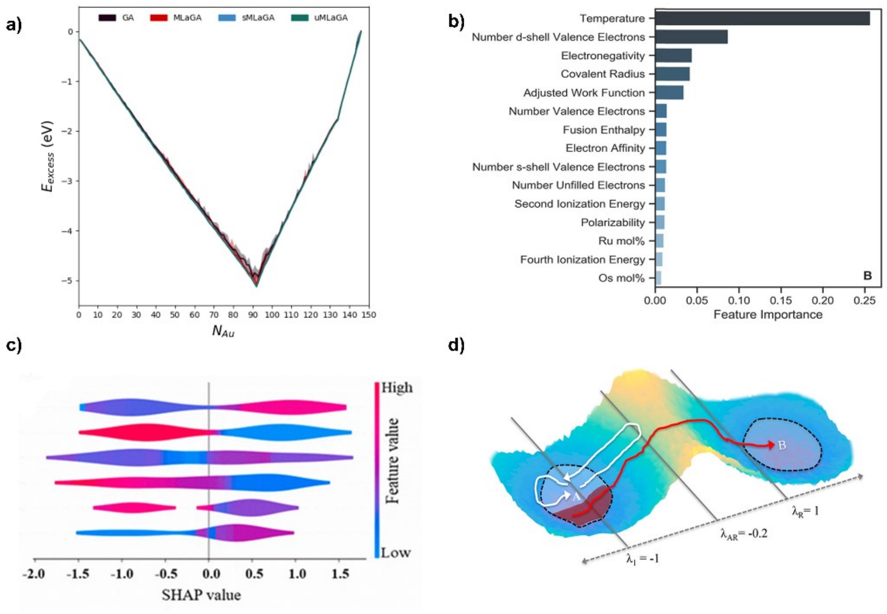


Figure 4

ML assisting in the energy calculation acceleration, influencing features analysis and reaction trajectories prediction of nanozymes. a) ML assists the acceleration of energy calculation algorithm. Reproduced with permission.[48] Copyright 2019, Springer Nature. b) Through ML, 15 important features that affect the catalytic activity of specific nanozymes most were directly output. Reproduced with permission.[49] Copyright 2019, American Chemical Society. c) SHAP assisted ML in ranking the importance of features. Reproduced with permission.[52] Copyright 2021, American Chemical Society. d) ML predicts catalytic reaction trajectories of nanomaterials.[57] Copyright 2019, American Chemical Society.