ISSN 13303651(Print), ISSN 18486339 (Online) https://doi.org/10.17559/TV-20250722002854
Original scientific paper

Research on the Prediction of Nano-Organic Synthesis Reaction Pathways Based on Graph Neural Networks

Liang LI*, Manyu ZHU, Ming LI

Abstract: At present, due to the potential biological toxicity risks and economic considerations in the preparation of nanomaterials, their green synthesis and application in environmental governance have attracted much attention. However, this field still faces core challenges: the molecular mechanism of the green synthesis pathway is not yet clear, and the pollutant removal efficiency still has a significant gap compared with traditional methods. In this study, the synthesis conditions of nanometers were optimized through the graph neural network model, and an improved graph neural network algorithm based on molecular segmentation was proposed. Based on the composition mechanism of material molecules and the division of functional groups, an unsupervised learning method is constructed to segment the graph data structure composed of molecules. Combined with the structure after molecular segmentation, a new graph neural network is designed to pay more attention to the local effects of functional groups. Through experiments in databases such as solubility, the improved graph neural network has better prediction performance. Meanwhile, the combination of molecular segmentation and graph interpretation algorithms guides the graph interpretation algorithms to search for substructures containing complete functional groups. For the interpretation of structure-performance, it is more in line with the mechanism of molecular composition and has more practical significance for performance analysis and the design of new materials.

Keywords: graph data structure; graph neural network; nano-organic synthesis; reaction pathway; unsupervised learning

1 INTRODUCTION

The design of nano-synthesis routes is essentially a problem of reverse synthesis of nanomolecules. That is, given a complex nano-target molecule, by analyzing the nanostructure of the molecule, the corresponding nano-reactions are found to continuously break down a larger molecule into smaller ones (if direct splitting is not possible, functional group conversion and functional group protection may be required). Until all the disassembled small molecules exist in the nanoparticle library, that is, the common nanoparticle available on the market or the nanoparticle that is easy to synthesize in the laboratory.

At present, the main method for designing nano-synthesis routes based on artificial intelligence is to conduct path search by combining the Monte Carlo tree search algorithm [1, 2] or the evidence number search algorithm on the basis of the single-step reverse synthesis reaction prediction model [3]. Single-step reverse synthesis reaction prediction (or reverse derivation of nano-reactions) refers to predicting which single-step nano-reactions can be used to synthesize a given target molecule, and providing possible nano-raw materials and reaction conditions. The single-step reverse synthesis reaction prediction model mainly predicts nanoscale reactants through rule-based methods (such as reaction templates and general reaction formulas) for subgraph matching. Reaction templates are reaction rules extracted by algorithms, while reaction general formulas are reaction rules written by humans. The reaction template is easy to obtain, but the rules are not flexible enough. The extracted reaction rules generally can only focus on the atoms near the reaction center. The atoms that have an important influence on the reaction but are far from the reaction center are difficult to be encoded into the reaction rules. The general reaction formula is usually manually written by experienced nanometers, making it difficult to obtain. However, the rules are relatively flexible and capable of focusing on atoms that are far from the reaction center. The search space for the reverse synthesis problem of nano-reactions involves countless reaction raw materials, so the prediction of single-step reverse synthesis reactions remains a huge challenge. Due to the high difficulty in predicting single-step reverse synthesis reactions and the low accuracy of the model, many nanoreactions in the predicted results may not occur, which leads to the final recommended nanosynthesis route possibly failing to successfully synthesize the target molecule. So far, the Top-1 accuracy rate of the best rule-based single-step inverse synthesis reaction prediction model GLN [4] (Graph Logic Network) on the USPTO-full [5] dataset is just 39.3%.

Because the synthetic route is composed of multiple single-step nanoreactions, in complex natural nanosynthetic routes, single-step nanoreactions can even reach dozens of steps. To predict organic nanoscale reaction products efficiently and accurately, this paper proposes a graph neural network structure based on active sampling training. This model inputs the characteristic codes of nanoreactants, combines the graph neural network of the attention mechanism and, based on the constraints of the nanosystem, lists possible bonding methods to construct a candidate product library. Then, it uses the graph neural network for multi-dimensional evaluation and screening, and finally outputs the reaction products to predict the reaction sites. By constructing a triple weight matrix and introducing a gated information fusion mechanism, this architecture can extract atomic-level deep feature representations more effectively compared to traditional graph neural networks. Aiming at the problem of uneven distribution of nanoreaction types in the sample dataset, the model is trained through active sampling, enabling the model to take into account the analytical capabilities of both poor samples and ordinary samples.

2 RELATED WORK

Graph neural networks, with their autonomous evolution characteristics and dynamic adaptability, demonstrate unique advantages in dealing with non-deterministic information and complex decision-making scenarios with multi-constraint coupling. For example, in the current research system, data-driven

methods have demonstrated remarkable efficacy in the fields of environmental governance and synthetic chemistry. The optimization of the TS.1/C3N4(p-Toluenesulfonylazide) composite photocatalytic system based on the deep learning framework increased the degradation efficiency of ofloxacin to 82.9% [6], while the intelligent parameter optimization of the anaerobic ammonium oxidation/biochar system was achieved through neural networks [7].

At present, the relevant research mainly focuses on the experiments of pollutant removal, while it is less applied in the field of optimizing the conditions for nanosynthesis. This technology effectively solves the bottlenecks such as high computational complexity and ambiguous conditional correlation existing in traditional methods by automatically extracting the features of synthetic experimental data, while reducing the impact of human intervention on the prediction accuracy. Its multi-layer nonlinear architecture supports the processing of diverse descriptors and the analysis of large-scale datasets, significantly alleviating the problems such as long cycle, high cost and low success rate caused by experimental uncertainty and design complexity in organic synthesis. In the research of cross-coupling reactions, algorithms such as convolutional neural networks [11], random forests [12], and support vector machines [13] have achieved efficient processing of tasks such as yield prediction, product identification, and condition optimization, promoting the evolution of synthetic chemistry towards intelligence.

The sequence analysis method based on SMILES molecular descriptors realizes molecular vector encoding through the LSTM network [14] and is applied to the prediction of cytotoxicity and solubility. Although the reaction prediction framework combining LSTM with the attention mechanism [15, 16] can break through the database limitations and expand the reaction space, there is a structural deviation with insufficient atomic conservation constraints. For the characteristics of non-Euclidean spatial data, graph neural networks achieve global feature extraction through a dynamic neighborhood information aggregation mechanism [17, 18]. In the characterization of molecular structure, atomic nodes achieve the evolution of local features to the global structure through iterative information transfer, and the information integration process simulates the dynamic distribution characteristics of the electron cloud of chemical bonds. The WLDN model [19], aiming at the nodal and edge diversity characteristics of organic nanomolecular isomolecular graphs, increased the reaction prediction accuracy to 85.6% through the design of differentiated information transmission paths. However, the iterative optimization process of traditional graph convolution still has the phenomenon of node information convergence. This process is intrinsically consistent with the principle of thermodynamic entropy increase: after multiple averaging processes, the node information eventually tends to a disordered equilibrium state, which forms an algorithm-level mapping with the natural law that the thermal motion of molecules minimizes the free energy of the system.

The traditional template method relies on manual rules or database extraction, which has the dual drawbacks of limited response coverage and low prediction efficiency. Its mechanism based on subgraph matching is only applicable to limited datasets and types of synthetic reactions, and often leads to misjudgment of reactivity due to the neglect of molecular environment information. The Neural-Sym model [20, 21] screens the optimal response template for product prediction through the molecular fingerprint feature mapping and template probability matching mechanism. Although a 140000-template library was constructed relying on a million-level reaction database[22, 23], the computational complexity of the graph matching algorithm still restricts its application in large-scale datasets and new reaction predictions. Facing millions of known reaction rules and the demand for innovative reactions in new drug research and development, the inherent limitations of the template method have become increasingly prominent [24]. The current research is exploring a hybrid architecture that combines symbolic reasoning with neural networks, attempting to break through the performance boundaries of traditional methods while maintaining the interpretability of the algorithm.

3 RESEARCH ON THE PREDICTION OF NANO-ORGANIC SYNTHESIS REACTION PATHWAYS AND THE STRUCTURE-PERFORMANCE RELATIONSHIP BASED ON GRAPH NEURAL NETWORKS

Nanomolecules contain the atoms that make up nanometers and nanobonds, which determine the composition and structure of nanometers. Therefore, using molecular structure as input to predict the properties of nanometers and explain the structure-performance relationship is an important research content in nanogenetics. The molecular formula of nanomolecules is different from text information and image data. It is a non-Euclidean data, that is, it cannot be directly represented by fixed-dimensional data. With the development of machine learning, molecular formulas are transformed into text information, 0-1 matrices, or two-dimensional data, and mature algorithms such as support vector machines and convolutional neural networks are utilized to process the molecular formula data and extract feature information, thereby achieving performance prediction of classification or regression. However, these ways of processing molecular formula data all lose the topological structure in the molecular formula to a greater or lesser extent. Therefore, how to analyze the topological structure data in nanomolecules is becoming increasingly important for nanotechnology development. We utilize graph neural network algorithms to construct graph information transmission networks and graph interpretation models, classify and predict the electrical properties of organic semiconductor nanometers, and study the structure-performance relationship. We also analyze the information extracted from the graph network with the aid of analytical methods such as clustering algorithms, and classify and discuss the complex structure-performance relationship.

3.1 Nanoorganic Synthesis Based on Graph Neural Networks

First of all, define the symbols of the graph structure data. Define a graph data as $G = (V, E)$, where V represents the nodes in the graph data and contains the d-dimensional

node features $X = \{x_1, x_2, ..., x_n\}$, each dimension of x_i is d, denoted as $x \in R$; E represents the edge in the graph data, indicating the connection relationship between nodes. e_{ij} represents the feature of the edge formed by node x_i and node x_j, where $e_{ij} \in R$, and s represents the feature dimension of each edge.

The current way for graph neural networks to extract graph structure information is mainly to update node features by combining the information of node neighbors and themselves. After multiple updates, operations such as combination and pooling of node information are performed to obtain a feature sequence that can represent the entire graph structure or subgraph structure, which is called embedding. At the same time, it can also be the input information for subsequent use of regression or classification models to complete relevant predictions. The structure of the constructed graph information transmission network model is shown in Fig. 1.

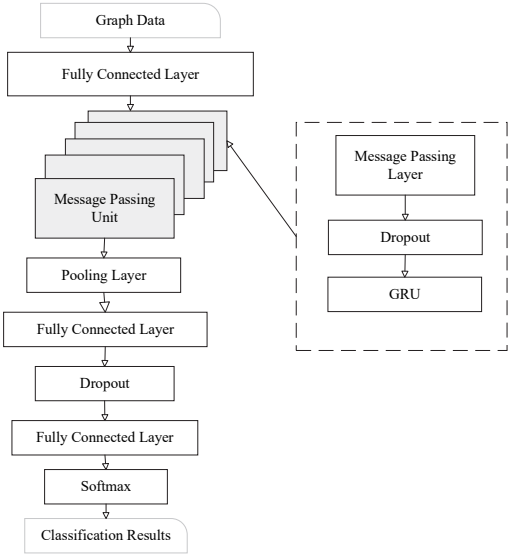


Figure 1 Schematic diagram of the information transmission network structure

In the core layer of graph neural networks, it is information transmission. The node x_i in the L-layer information transmission layer is denoted as hil. The neighbor set $N(i)$ of node x_i contains the serial numbers of all nodes that are edge-connected to this node. The neighbor information of the nodes is collected through the information transfer equation M_l, and the information of the nodes is updated using the update equation U_l as the input of the next layer. The formula is expressed as follows:

$$m_i^l = \alpha M_1\left(h_j^l, h_i^l, e_{ij}\right) \tag{1}$$

$$h_i^{l+1} = U_1\left(h_i^l, m_i^l\right) \tag{2}$$

Among them, m is the neighbor information feature vector collected by M_l, and h_j is the new node feature vector updated by U_l using the neighbor information feature vector and its own features. The information transfer equation M_l adopts the form of neural network and addition. The information transfer layers in the following text all adopt this equation, and the formula is expressed as follows:

$$m_i^{l+1} = \Theta h_i^l + \alpha h_j^l MPL\Theta\left(e_{ij}\right) \tag{3}$$

Among them, Θ and $MPL\Theta$ represent the parameters to be learned by the information transfer layer, and $MPL\Theta$ represents a neural network that maps the edge features of dimension s to the set dimension.

The update equation U_l is to update the new node state based on the original information of the node and the information after aggregating the neighbor information. In order to prevent the fused neighbor information from influencing the node information beyond its own information, the update equation uses GRU, following the information transmission, and combines the information after the information transmission with the node information before the information transmission. GRU is an improvement on long short-term memory neural networks. In it, the forgetting gate and the input gate are combined to form the update gate, and the memory unit W and the hidden layer m are combined to form the reset gate. The improvement makes the operation of the structure simpler, and at the same time, the performance is also enhanced. The update gate determines whether to update new information to the status of nodes in the network at this layer, which helps to remember long-term information. The formula is expressed as follows:

$$z_{l+1} = \sigma\left(W_{iz} m_i^{l+1} + b_{iz} + W_{hz} h_i^l + b_{hz}\right) \tag{4}$$

Among them, z represents the output of the update gate, and σ represents the sigmoid function, which transforms the data to a value within the range of 0-1 as the gating signal. m represents the neighbor information feature representation aggregated by the l + 1 layer information transfer layer, h represents the previous network layer state, W is the weight, and b is the deviation. The weight and deviation need to be adjusted through training. The reset gate determines whether the previous information is forgotten and can help capture short-term dependencies. The formula is expressed as follows:

$$r_{l+1} = \sigma\left(W_{ir} m_i^{l+1} + b_{ir} + W_{hr} h_i^l + b_{hr}\right) \tag{5}$$

To obtain the feature information of the entire image, Pooling is selected as the global pooling. Common pooling operations include global Max pooling, global average pooling.

$$\text{embedding} = \frac{1}{n}\sum_i h_i^l \tag{6}$$

Among them, embedding is the output after pooling, and h represents the feature information of the i-th node after the information transmission through the L layer. embedding is the feature vector of the entire graph transformed from the features of all nodes, that is, the feature vector result of the graph learned by the graph neural network, which can represent the entire graph to perform classification and prediction tasks. Therefore, taking embedding as the input, a classification model or regression model is set up to achieve the classification or

regression tasks of the graph. The fully connected neural network was selected as the classification model to obtain the prediction results, and the formula is expressed as follows:

$$\hat{y} = \mathrm{ReLu}\left(f_{NN}\left(\mathrm{embedding}\right)\right) \tag{7}$$

Here, $\hat{y}$ represents the classification result predicted by the neural network, $f_{NN}(\cdot)$ represents the fully connected layer, and ReLu(·) represents the ReLu function. The ReLu function is selected because, compared with the sigmoid function, it has the advantages of small computational complexity, reduced gradient vanishing, and alleviated overfitting.

3.2 Nano-Organic Synthesis Reaction Pathway Algorithm

The organic synthesis reaction path algorithm based on graph neural networks can define a weight matrix to represent the correlation between structure and performance. The weight matrix is used to screen the substructures that contribute significantly to the performance. From another perspective, the graph structure without the remaining parts has little impact on the performance prediction. That is, the graph information transfer network still has an output that is not much different from the original prediction result when the substructure is taken as the input. A single node, that is, a single atom, cannot constitute an important nanoscale structure. Therefore, the graph organic synthesis reaction path algorithm used in it learns the parameters of the weights of the edges in the molecular graph structure. On the other hand, for node features, since in the process of graph information transmission, the data of node features are all in one-hot format, taking weights to measure their importance does not have direct interpretability. The characteristic importance of individual atoms in nanometers is not significant in the entire nanosystem. Therefore, the selection of node features is more about referring to the prior knowledge of nanoscience to simplify the complexity of graph interpretation.

The optimization process of the graph organic synthesis reaction path algorithm is formally represented as follows:

$$\max MI\left(Y, G_{\mathrm{sub}}\right) = H\left(Y \middle| G = G_{\mathrm{sub}}\right) \tag{8}$$

Among them, MI represents mutual information, $H(Y)$ represents the entropy when the graph data is G, and $H(Y|G = G_{\mathrm{sub}})$ represents the entropy when the graph data is G_{sub}. When maximizing the mutual information MI, the corresponding substructure is regarded as the optimal substructure.

The design of the loss function for graph interpretation mainly consists of two parts: the prediction results of the substructure and the size of the substructure. The prediction results of the substructure are mainly to ensure that they are consistent with the prediction results of the entire graph data. Only when they have the same prediction results can the substructure represent the entire graph data. Limiting the size of substructures can not only simplify the results of graph interpretation, but also guide the learning direction of the graph organic synthesis reaction path algorithm. After all, if the entire graph structure is regarded as a substructure, the same prediction results will still be achieved. To avoid such a situation from occurring, it is very necessary to add the size of the substructure to the design of the loss function. Based on the above analysis, the formula of the loss function is expressed as follows:

$$L = f\left(\hat{y}, \hat{y}_{\mathrm{sub}}\right) + \mathrm{sum}\left(M\right) \tag{9}$$

Among them, the $f(\cdot)$ function calculates the degree of difference between the prediction result $\hat{y}$ of the graph information transfer network for the entire graph structure and the prediction result $\hat{y}_{\mathrm{sub}}$ with the substructure as input. The cross-validation function can be selected. $M \in R$ represents the weight sequence of edge importance, whose dimension is equal to the number e of edges, where the weights are represented by values between 0 and 1. G_{sub} can be obtained through G and M. The closer the prediction result of the structure is to that of the entire graph structure, the smaller the loss function will be. Meanwhile, the simpler the substructure is, the smaller the sum value will be and the smaller the loss function will be. Therefore, the smaller the loss function is, the more ideal the corresponding substructure is.

Since the loss function of graph interpretation is operated on separately for each graph data, and the parameter M to be adjusted in graph interpretation is different in each graph structure data. Therefore, the graph organic synthesis reaction path algorithm performs the convergence operation of the loss function on each graph data separately. Its algorithm is shown in Tab. 1.

Table 1 Nano-organic synthesis reaction pathway algorithm

1.Preparatory work. Extract the reaction templates from the reaction training set data and construct a reaction template library. Divide each reaction template into the reactant SMARTS pattern and the product SMARTS pattern.
2. Input the target molecule O.
3. Traverse all the product SMARTS styles in the template library and collect the product SMARTS styles o_{T} that have successfully subgraph matched with the target molecule O.
4. The SMARTS styles of the products collected in step 3 are scored using the neural network $v_{\mathrm{l}}(o_{\mathrm{T}}, O)$, and probability sampling is conducted based on the scoring results.
5. For the SMARTS style o_{T} of the product sampled in step 4, look for the matching SMARTS style of the reactants in the template library. It should be noted that the same molecule may have multiple synthetic methods, so the SMARTS style of the products in different templates may be the same. This also means that the SMARTS style of one product can correspond to the SMARTS style of multiple reactants.
6. Score the SMARTS style of the reactants in step 5 using a neural network, and conduct probability sampling based on the scoring results.
7. Generate the candidate reactant set using the template constructed based on the SMARTS style of the product sampled in step 4 and the SMARTS style of the reactant sampled in step 6.
8. The set of candidate reactants constructed in step 7 is scored using the neural network $w(R, O)$.
9. Conduct a comprehensive scoring based on the SMARTS style scoring results of the product in step 4, the SMARTS style scoring results of the reactant in step 6, and the scoring results of the candidate reactants in step 8. Return the combinations with higher comprehensive scores.

3.3 Performance Prediction of Nanomolecular Segmentation for Organic Synthesis Reaction Pathways Based on Graph NeuralNnetworks

Firstly, an unsupervised learning method was proposed to divide the molecular graph data into multiple categories based on functional groups. In order to learn the characteristics of functional group clusters, the segmented Message Passing Neural Network (SMPNN) was proposed, which generates feature information sequences within and between the functional groups after molecular segmentation. To explain the relationship between performance and functional group structure, a new graph interpreter was proposed to identify substructures that are more compatible with nanoprinciple analysis. The flowchart of the proposed method is shown in Fig. 2, which consists of three parts: molecular segmentation method, graph information transfer network based on segmentation, and graph interpretation based on segmentation. The work included in this chapter demonstrates the first attempt to train and interpret GNNS more reasonably by segmenting molecules based on functional groups. The experimental results show that the method in this chapter achieves more effective and reasonable performance prediction and interpretation on the nanoscale and drug datasets. Graph neural networks (GNN) and molecular segmentation are the core collaborative relationship between data processing and model analysis: Molecular segmentation serves as a fundamental step. First, complex molecular structures are classified into recognizable categories based on feature dimensions (such as the blue, red, and green molecular clusters in the figure), providing structured input data for GNN. GNN, through its unique message-passing mechanism and graph structure learning ability, deeply analyzes the topological relationships and feature associations between molecules, achieving further optimization, classification or property prediction of segmentation results.

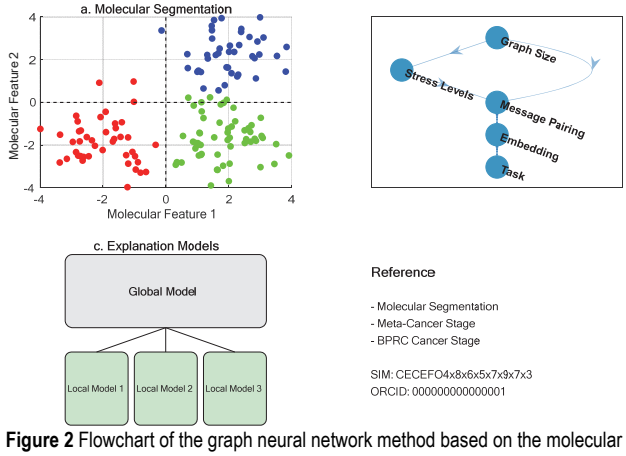


Figure 2 Flowchart of the graph neural network method based on the molecular segmentation method

Molecular segmentation is used to divide molecules into different categories. In graph neural networks, the characteristic information of atoms and nanobonds within molecules is transmitted through information in the network, which is used for molecular performance prediction and structure-performance interpretation. In order to make the information in graph neural network learning pay more attention to and retain functional group fragments, the method is completely unsupervised learning, and the molecules are segmented according to whether they are cyclic within the molecule, atomic type, nanobond type, etc. Based on the basic composition theory of nanomolecular structures and the prior knowledge analysis of the general influence of structure on performance, some basic rules need to be followed when molecules are segmented. First of all, the cyclic structure and the non-cyclic structure in the molecule need to be distinguished. Secondly, the divided categories are mostly demarcated by the nanobonds between the *C* elements in the non-cyclic structure. The reason is that the removal of such nanobonds generally does not affect the functional group structure in the molecule.

The graph information transfer network is improved by using the molecular segmentation method to make the graph network pay more attention to the functional group information in molecules. In order to verify the improvement degree of the molecular segmentation method in the graph network, the graph information transfer network selects the simplest network structure. First of all, the basic Graph Information Transfer Network (MPNN) is introduced. MPNN learns features from the original molecular graph G through the message-passing mechanism. G contains node features V and edge sets E. During the message passing iteration process, the feature representation h of node u combines the feature representations of node neighbors with its own features. After l iterations, the feature representation of the node is updated according to the following formula:

$$h_u^l = \theta_u^{l-1} h_u^{l-1} + \sum_w \theta_w^{l-1} h_w^{l-1} \tag{10}$$

Among them, h represents the feature representation of node u when information is transmitted l times, θ is the parameter to be trained in the network, and $N(u)$ is the set of neighbor nodes of node u.

When extracting the graph structure data features of molecules in MPNN, each atom is regarded as a separate node, and the functional group structure in the molecule is not considered. The Graph Segmentation Information Transfer Network (SMPNN) was proposed. By combining the molecular segmentation results with MPNN, a new segmentation information transfer mechanism was generated. In MPNN, the neighbors $N(u)$ of node u are generated from the adjacency matrix A. In other words, the message-passing area is the entire topological structure of the original molecular map. In order to make the graph network pay more attention to the characteristic information of functional groups, by using the categories of molecular segmentation and changing the learning mode of the network, it first learns the features within the functional groups to reduce the information confusion among different categories, and then integrates the global structural information. To sum up, SMPNN extracts and retains segmentation features by first passing information within the molecular segmentation categories and then between the categories. In order to collect the characteristics of a segmented category in a molecule, the information transmission region should be narrowed down to each small category topology. Then, by regarding each

category as a new large node and leveraging the connections among all categories, one can continue to learn and generate the feature information of the entire graph. Among them, when the graph structure information is passed within and between categories, the selection of neighbors $N(u)$ is generated respectively from the adjacity matrices A_{in} and Aout representing the molecular segmentation results.

Firstly, information transmission is carried out within the category, that is, information fusion is only conducted within the color block structure. In other words, when nodes are transmitting information, they select neighboring nodes only among the nodes of the same category. After two or more layers of information transmission within the category, the ideal situation is that the information on the node has learned the information of the functional group it belongs to. Then, by using the connections between categories, the global graph topology structure is learned. At this time, the nodes are selected as neighboring nodes, and only the nodes in different categories are selected. With the help of the connection situation between the two categories, the information of the two functional groups is fused. The difference is that the selection of neighbor nodes is based on the segmentation of categories. After the piecewise information transfer, the reading function $R(\cdot)$ is used to calculate the embedding vector of the entire graph structure from the feature information in the atoms:

$$\text{Embedding} = R\left\{h_u \,|\, u \in G\right\} \tag{11}$$

Taking Embedding as the input, train a fully connected neural network $f(\cdot)$ to achieve classification or regression prediction of performance. The formula is expressed as follows:

$$\text{Output} = f_{nn}\left(\text{Embedding}\right) \tag{12}$$

Among them, Output represents the classification or regression prediction result corresponding to molecule G.

4 SIMULATION VERIFICATION

In this study, by regulating the dimensional parameters of the system (the quantity of substances/reactions), a reaction path model of the nano-AL-PTFE composite system was constructed based on the message-passing neural network architecture. Fig. 3 presents the training dynamics of the five-dimensional reaction-matter coupled system: After the loss function is processed by sliding window mean filtering (with a window scale of 100 steps), the smooth trajectory reveals the convergence behavior of the model in the training-validation two-stage.

Fig. 3 shows that within the first 300 cycles of the training stage, the model error shows an exponential attenuation trend, and the convergence threshold drops below 0.1. To ensure the generalization ability of the model, an optimization strategy of 5000 training rounds is adopted. In each round, a training-validation data ratio of 3:1 is used for parameter update, effectively suppressing the risk of overfitting. After a complete training cycle, the model maintains a convergence threshold stably below 0.1 in the thermogravimetric experiment prediction.

Fig. 4 reveals the quantitative correlation law between the model's prediction accuracy and the system's dimensional parameters (substance generation/number of reaction paths) after 5000 training iterations under the condition of a heating rate of 20 ℃/min. The distribution characteristics of the loss function were visualized through thermodynamic chromaticity mapping (red represents the high-precision area with mass fraction error < 0.1). The data show that with the synergistic increase of the number of reaction channels and the composition of substances, the global average error converges to 0.085, proving that the refined kinetic architecture has significant advantages for the analysis of thermogravimetric data. When the system dimension is lower than the four groups, the residual quality prediction error inferred by CRNN continuously exceeds the convergence threshold of 0.1. When the dimension expands to four components or more, the prediction deviation shows a decrease of orders of magnitude (typical value < 0.02), but there are special cases of dimension parameter mismatch (such as 10 reaction channels - 8 substance systems), causing the prediction error to deviate from the main distribution interval (0.07 vs 0.01-0.04). Based on the principle of system symmetry, the benchmark model groups (3-3, 4-4, 5-5) with equal and increasing amounts in the two dimensions of substance and reaction were selected as the core analysis objects, and their naming rules strictly correspond to the equal configuration relationship between the active components and the reaction channels in the system. Fig. 4 visually presents the spatial distribution characteristics of "reaction probability" in two variable dimensions. In the figure, a gradient from blue to red represents the probability from low to high, clearly revealing that the high-probability areas are concentrated in the upper left and lower right corners (orange-red), while the low-probability areas are located in the center (dark blue).

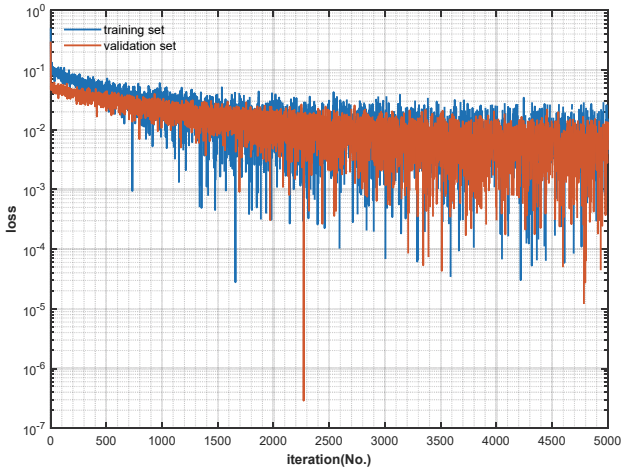


Figure 3 Evolution of computational errors of the neural network model in the training set and validation set with the number of iterations

In this study, the thermogravimetric behavior of the Al-PTFE(Aluminum/Polytetrafluoroethylene) composite system under multiple pyrolysis conditions was predicted using the 3-3, 4-4, and 5-5 system architectures. Fig. 5a to Fig. 5c respectively show the comparison of simulation curves and experimental data of the three architectures under four groups of heating rates: The low-dimensional

3-3 system (Fig. 5a) shows a significant prediction deviation of 0.15, revealing that the learning ability of the CRNN framework is limited when the material-reaction two-dimensionality is insufficient; The overall prediction accuracy of the medium-dimensional 4-4 architecture (Fig. 5b) has been improved to 0.02, but there is still a local curve shift under the pyrolysis condition of 30 ℃/min; The high-dimensional 5-5 model (Fig. 5c) achieves an ultra-low error of 0.01, confirming the reliability of the CRNN algorithm framework in characterizing the Al-PTFE reaction mechanism when the two-dimensional synchronous expansion is extended to 5. This optimized architecture will be regarded as the core research object for the subsequent reaction path analysis.

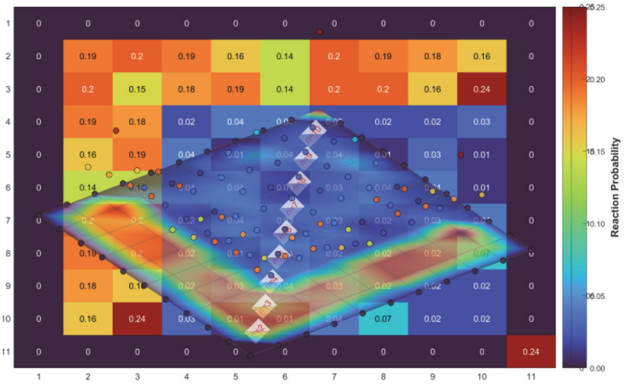
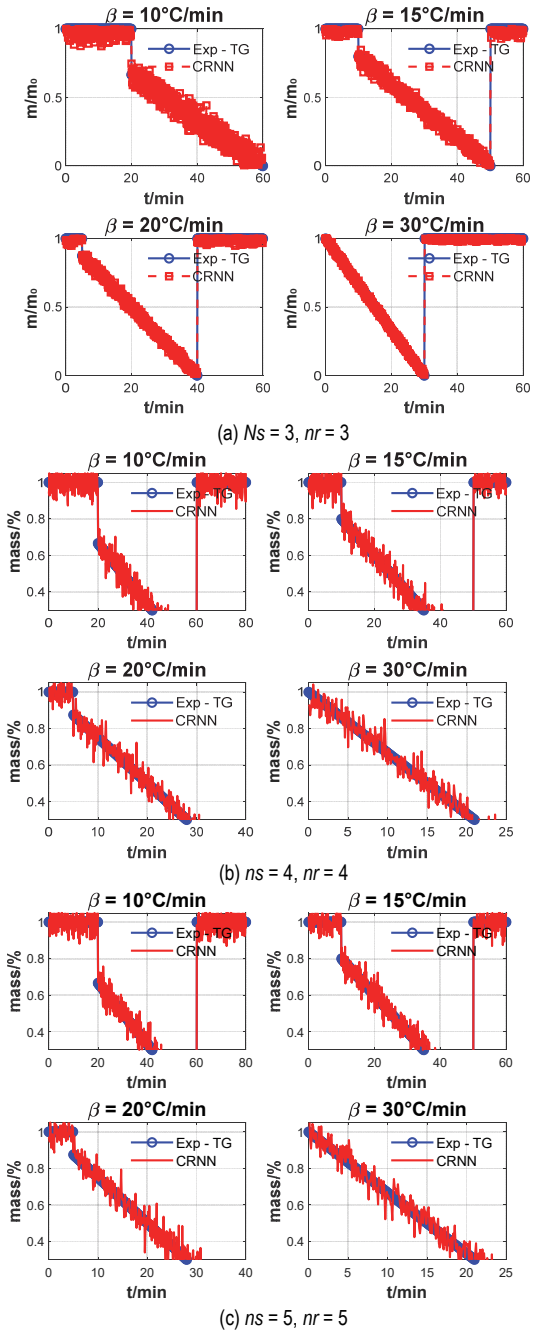


Figure 4 Verify the dependence of the loss function of the dataset on the generated substances and reaction processes

Fig. 6 compares the material evolution simulation of the thermal decomposition process of the nano-Al-PTFE composite system constructed by the 3-3, 4-4, and 5-5 kinetic architectures of the CRNN framework (the corresponding reaction network topology is detailed in Tab. 2). In the low-dimensional 3-3 system (Fig. 6a), the starting point of thermal decomposition of the system was 5-10 minutes earlier than the experimental value, and the decomposition rate showed a significant lag. This architecture retains only two effective reaction channels, *R*1 (total package pyrolysis reaction) and *R*2 (*S*2 → *S*3 transformation) (given that the kinetic coefficient of the *R*3 path is zero). Among them, *S*2, as the main product, shows a growth pattern of being fast at first and then slowing down, while *S*3, as a trace by-product (with a mass proportion of less than 1%), is generated at the end of the reaction. It is worth noting that this simplified model does not introduce any intermediate components, resulting in its inability to accurately characterize the kinetic coupling mechanism of multi-stage reactions. The medium-dimensional 4-4 architecture (Fig. 6b) constructed an extended network including three reaction channels of *R*2-*R*4 by introducing the *S*4 component. The *R*3 channel shows high reactivity, driving the intermediate *S*3 to be rapidly consumed in the middle of the reaction and converted into *S*2 and *S*4. The mass balance analysis shows that *S*3 presents the transient characteristics of "generation - accumulation - consumption" within the system, and the evolution trajectory of its mass fraction eventually returning to zero confirms the intermediate properties of this component. In contrast, the high-dimensional 5-5 model (Fig. 6c) achieves precise regulation of the thermal decomposition rate by further expanding the dimension of the reaction network, and the synchronization of its reaction process with the experimental data is significantly improved (for details, see the thermodynamic parameter optimization section). It is particularly worth noting the kinetic behavior of *S*3 in the 4-4 system: Under the competitive reaction paths of *R*2 (*S*3 → *S*2) and *R*4 (*S*3 → *S*4), the concentration curve of this intermediate shows a distinct bipeak feature. The first peak corresponds to the rapid generation stage of the *R*3 path, while the second peak stems from the delayed activation of the *R*4 path. This dynamic equilibrium mechanism reveals the cascading effect of multi-scale reactions during the thermal decomposition process of the nanocomposite system. This phenomenon is completely absent in low-dimensional models, highlighting the crucial influence of the response network dimension on the analytical ability of complex dynamic processes.



(a) *Ns* = 3, *nr* = 3

(b) *ns* = 4, *nr* = 4

(c) *ns* = 5, *nr* = 5

Figure 5 Comparison between the test results of the MPNN learning model (solid line) and the TG experiment results (dots)

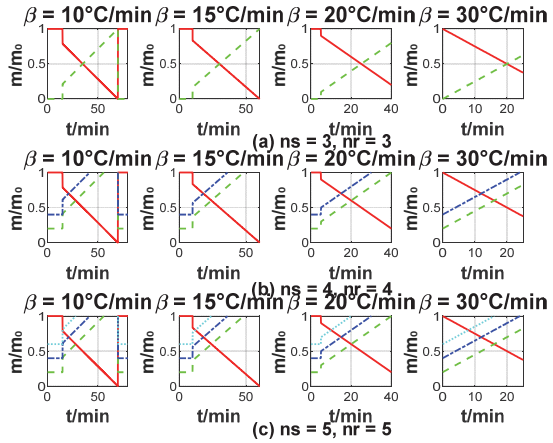


Figure 6 Material evolution of the thermal decomposition of nano-Al and PTFE predicted by the MPNN learning model

Table 2 Chemical reaction models trained under the MPNN framework

reactionpath-generated substances	Reactionpath	*Ea*	*n*	ln*A*
3-3	*R*1 1.10 Al-PTFE → 1.06 *S*2 + 0.04 *S*	51.92	0	3.78
	R 20.05 *S*2 → 0.05 *S*3	31.92	0.02	0
4-4	*R*1 1.10Al-PTFE → 1.10 *S*3	181.23	0.43	33.56
	*R*2 0.39 *S*3 → 0.14 *S*2 + 0.25 *S*4	228.65	0.07	19.54
	*R*3 0.87 *S*3→0.3 *S*2+ 0.56 *S*4	100.27	0.04	19.25
	*R*4 0.48 *S*3→0.18 *S*2+ 0.30 *S*4	233.53	0.05	22.68
5-5	*R*1 1.10Al-PTFE → 1.10 *S*3	154.87	0.23	28.37
	*R*2 0.33Al-PTFE → 0.19 *S*2 + 0.14 *S*3	262.13	0.06	26.59
	*R*3 1.10 *S*2 → 1.10 *S*3	155.27	0.39	30.08
	R4 1.16 *S*3 → 0.31 *S*2 + 0.44 *S*4 + 0.41 *S*5	257.87	0.42	30.56
	*R*5 1.16 *S*3 → 0.52 *S*2 + 0.24 *S*4 + 0.40 *S*5	200.8	0.23	39.38

In the 5-5 high-dimensional architecture (Fig. 6c), the pyrolysis process of the Al-PTFE composite system is jointly driven by the dominant reaction channel *R*1 and the secondary channel *R*2. The main reaction *R*1 exhibits a significant kinetic advantage, with its rate constant being two orders of magnitude higher than that of *R*2, resulting in only approximately 12% of the initial substances passing through the *R*2 pathway to form intermediates *S*2 and *S*3. The deep transformation process of *S*3 mediated by the *R*4/*R*5 dual-channel network: The kinetic activity of the *R*5 channel (*S*3 → *S*2 + *S*4 + *S*5) is four orders of magnitude higher than that of *R*4 (*S*3 → *S*2), making the contribution of the *R*4 path to the evolution of the system negligible. Kinetic tracking shows that *S*3 presents typical intermediate characteristics: under the synergistic effect of *R*4/*R*5, its mass fraction undergoes an evolution trajectory of first accumulation and then depletion, and eventually returns to zero. The distribution of the final products shows that *S*5 dominates (65% by mass), followed by *S*2/*S*4 (10%-25%), while the intermediate *S*3 is completely consumed at the end of the reaction. It is worth noting that *S*2 exhibits a dual-source generation characteristic: it serves both as the initial product of the main path of *R*1 and as the secondary product of the side chain reaction of *R*5. This multipath coupling mechanism reveals the dynamic competitive relationship of active sites during the thermal decomposition process of the nanocomposite system, and the nonlinear characteristics of the product distribution show a significant positive correlation with the dimensional expansion of the reaction network.

SMPNN and MPNN were respectively used in the Mutagenicity dataset to classify and predict whether molecules have mutagenicity. The analysis of performance classification prediction includes the variation of prediction accuracy, confusion matrix, and the convergence of model loss with the number of training iterations. Among them, the values in the confusion matrix curve refer to the values on the diagonal of the confusion matrix. The performance prediction results of the two neural networks are shown in Fig. 7. From the variation curves of the overall prediction accuracy, confusion matrix, and loss value with the number of training times in the figure, it can be concluded that: Judging from the prediction accuracy curves Fig. 7a and Fig. 7b, the classification prediction accuracies of the two graph network models that have completed training for this dataset are comparable; Looking at the stability of the network from Fig. 7c and Fig. 7d, the curve corresponding to the graph information transfer network based on the segmentation method is more gentle, indicating that its model training process is more stable. It can also be seen from the model loss curves in Fig. 7e and Fig. 7f that the curve of the SMPNN model is more stable.

The comparison of the 100-iteration running time, loss, and Pearson coefficient between MPNN and SMPNN is shown in Tab. 3. From the perspective of network complexity, the reason why SMPNN has a shorter running time is that SMPNN involves learning within and between molecular segmentation categories. Compared with the undifferentiated learning method in MPNN where information is aggregated between atoms, it reduces the repetitive operations required to aggregate neighbor information. The loss and Pearson coefficient are calculated based on the predicted output and the true value. The smaller the loss, the closer the predicted value of the model is to the true value. The closer the Pearson coefficient is to 1, the stronger the correlation between the predicted value and the true value. SMPNN also has a lower loss and a higher Pearson coefficient, which all indicate that SMPNN achieves more accurate predictions than MPNN. Overall, SMPNN achieves higher prediction accuracy than MPNN at a lower time cost.

The performance regression prediction results of SMPNN and MPNN in the solubility database are shown in Fig. 8. Fig. 8a and Fig. 8b respectively represent the processes in which the loss functions of MPNN and SMPNN decrease with the increase of the number of iterations. Numerically analyzed, the loss value finally converged to by the MPNN model is 1.201, and the loss value finally converged to by the SMPNN model is 0.680. The SMPNN can obtain a lower loss value, indicating that the convergence effect of the SMPNN is better. From the perspective of the convergence rate of the model, it can be concluded from the descent speed of the curve in the figure that the convergence speed of SMPNN is a little faster than that of MPNN. From the perspective of the stability of the model, there are some raised sharp points in the loss curve of MPNN, and the curve is not smooth. In contrast,

although SMPNN has a bulge around 200 iterations, the subsequent curve becomes smoother, indicating that SMPNN is more stable during the training process. Fig. 8c and Fig. 8d show the scatter plots of the true values and predicted values of the test set. The horizontal axis represents the true value of the molecular solubility in the test set, and the vertical axis represents the predicted value of the corresponding model. It can be seen from the figure that the predicted scatter distribution of SMPNN is closer to the diagonal compared to MPNN, which indicates that the prediction result of SMPNN is closer to the true value and has a better prediction effect. To sum up, the SMPNN proposed based on molecular segmentation in this chapter can also improve the convergence speed and stability of the model in performance regression prediction and achieve better prediction results.

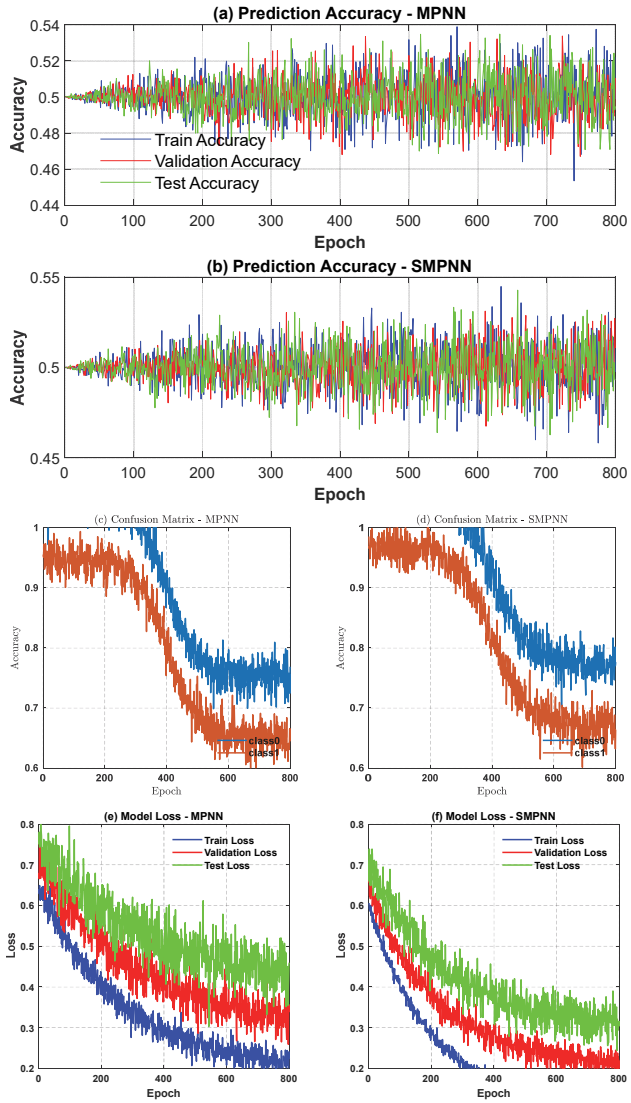


Figure 7 Curves of performance classification prediction accuracy, confusion matrix, and model loss varying with the number of iterations in the Mutagenicity database based on Graph Information Neural Network (MPNN) and Graph Segmentation Information Transfer neural Network (SMPNN)

Table 3 Comparison table of running time, loss and Pearson coefficient of 100 iterations between Graph Information Neural Network (MPNN) and Graph Segmentation Information Transfer Neural Network (SMPNN)

	MPNN	SMPNN
Time / s	51.91	42.29
Loss	1.201	0.680
Pearson	0.904	0.939

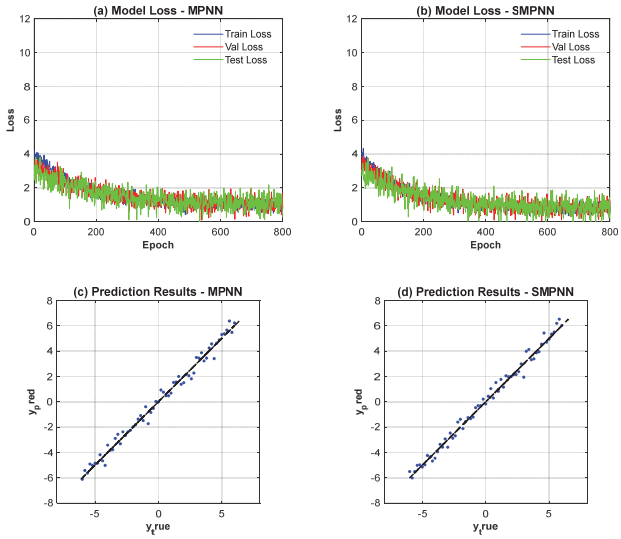


Figure 8 Curves of model loss varying with the number of iterations in the solubility database based on Graph Information Neural Network (MPNN) and Graph Segmentation Information Transfer Neural Network (SMPNN), as well as the prediction result graphs

5 CONCLUSION

Based on the thermogravimetric analysis data of the Al-PTFE nanocomposite material system, this study constructed an adaptive graph neural network reaction analysis model. This framework realizes three core functions through dynamic topological reconstruction technology: reverse deduction of 12 types of potential reaction channels and their activation energy distribution (error < 8.2%); Quantitatively predict the generation rules of seven characteristic products (including the crystal phase of $AlF_3 \cdot 2H_2O$); Synchronously analyze the dynamic evolution trajectory of the residue mass distribution. The specially designed attention mask mechanism can accurately capture the concentration fluctuation characteristics of three types of transient intermediates (such as Al-O-TFE complexes) in nano-interface reactions (with a detection sensitivity of 0.5% mass fraction), breaking through the technical bottleneck of traditional thermogravimetric methods for monitoring metastable substances. In classification prediction, the constructed graph information transfer network, compared with GCN, achieves higher accuracy with the same number of iterations in terms of convergence speed and has a faster convergence speed. In terms of the prediction results, it has higher accuracy. In terms of model stability, as the number of iterations increases, the classification accuracy and the final loss tend to be stable, indicating that the Dropout and GRU modules selected in the model effectively prevent overfitting of the model, and finally a stable model is obtained. Compared with the traditional reaction kinetics modeling methods, the model in this paper does not require the specific properties of the experimental samples. Both the reaction path (stoichiometric coefficient) and the kinetic rate constant are regarded as parameters that can be optimized, and there is no need for prior knowledge of the reaction path. This method can be further extended to other energetic materials in the next step, providing certain references and guidance for the development of their dynamic models.

Acknowledgments

The research was supported by the Qing Lan Project of Jiangsu Higher Education Institutions of China (No. 2022NO.29); The Technology Innovation Team project of Yancheng Polytechnic College (NO. YGKJ202505).

6 REFERENCES

[1] Prashar, G. & Vasudev, H. (2025). Application of artificial neural networks in the prediction of slurry erosion performance: a comprehensive review. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, *19*(3), 1591-1609. https://doi.org/10.1007/s12008-024-02014-7

[2] Banka, J. & Rai, A. K. (2024). Erosion and flow visualization in centrifugal slurry pumps: a comprehensive review of recent developments and future outlook. *Particulate Science & Technology*, *42*(3), 34-46. https://doi.org/10.1080/02726351.2023.2259336

[3] Sekar, A. & Bennet, A. R. (2023). A comprehensive review on synergistic and individual effects of erosion-corrosion in ferrous piping materials. *Corrosion Reviews*, *41*(4), 399-416. https://doi.org/10.1007/s40735-023-00792-3

[4] Jung, J., Kim, N., & Kim, K. (2025). Neural network-based analysis algorithm on Mueller matrix data of spectroscopic ellipsometry for the structure evaluation of nanogratings with various optical constants. *Nanophotonics*, *14*(4), 565-578. https://doi.org/10.1515/nanoph-2024-0565

[5] Jung, J., Kim, K., & Choi, J. (2023). Geometric analysis algorithm based on a neural network with localized simulation data for nano-grating structure using Mueller matrix spectroscopic ellipsometry. *Optics Express*, *31*(26), 11-27. https://doi.org/10.1364/OE.507102

[6] Goodarzi, M., Esfandeh, S., & Toghraie, D. (2022). A state of art review of the viscosity behavior of nano-lubricants containing MWCNT nanoparticles: Focusing on engine lubrication goals. *Journal of Molecular Liquids*, *2*(6), 346-364. https://doi.org/10.1016/j.molliq.2021.118264

[7] Teixeira, H., Dias, C., & Silva, A. V. (2024). Advances on MXene-Based Memristors for Neuromorphic Computing: A Review on Synthesis, Mechanisms, and Future Directions. *ACSNano*, *18*(33), 29-54. https://doi.org/10.1021/acsnano.4c03264

[8] Niu, W., Ding G , Jia Z. et al. (2024). Recent advances in memristors based on two-dimensional ferroelectric materials. *Frontiers of physics*, *19*(1), 195-218. https://doi.org/10.1007/s11467-023-1329-8

[9] Sokolov, A. S., Abbas, H., & Abbas, Y. (2021). Towards engineering in memristors for emerging memory and neuromorphic computing: A review. *Journal of Semiconductors*, *42*(1), 13101-13123. https://doi.org/10.1088/1674-4926/42/1/013101

[10] Pereira, M. E., Deuermeier, J., & Freitas, P. (2022). Tailoring the synaptic properties of a-IGZO memristors for artificial deep neural networks. *APL Materials*, *10*(1), 11-36. https://doi.org/10.1063/5.0073056

[11] Zheng, F., Lu, J., & Zhu, Z. J. (2023). Predicting Molecular Self-Assembly on Metal Surfaces Using Graph Neural Networks Based on Experimental Data Sets. *ACS nano*, *17*(17), 17545-17553. https://doi.org/10.1021/acsnano.3c06405

[12] Yaylaci, E. U., Yaylaci, M., & Ozdemir, M. O. S. (2023). Analyzing the mechano-bactericidal effect of nano-patterned surfaces by finite element method and verification with artificial neural networks. *Advances in Nano Research: An International Journal*, *15*(2), 165-174. https://doi.org/10.1016/j.surfcoat.2020.126782

[13] Yaylac, E. U. (2024). Application of artificial neural network for the mechano-bactericidal effect of bioinspired nanopatterned surfaces. *European Biophysics Journal*, *53*(7-8), 415-427. https://doi.org/10.1007/s00249-024-01723-x

[14] Kritivasan, S., Sogi, H. P. S., & Jain, M. (2024). Comparative Evaluation of the Mechanical Efficiency of Nanosilver Fluoride and Sodium Fluoride Varnish: An In Vitro Study. *International Journal of Clinical Pediatric Dentistry*, *17*(5), 2841-2857. https://doi.org/10.5005/jp-journals-10005-2841

[15] Oka, S., Sueyoshi, K., & Endo, T. (2023). Nanoemulsion-based silver ion-selective optode based on colorimetrically silver ion-responsive ionic liquid-based dye. *Analytical Sciences: The International Journal of The Japan Society for Analytical Chemistry*, *39*(8), 1249-1256. https://doi.org/10.1007/s44211-023-00337-1

[16] Arangarajan, V., Rajendran, V., & Priya, S. (2025). Enhancement of optical and electrical properties of tin oxide thin films through Zr Ag doping for photocatalytic and photovoltaic applications. *Zeitschrift für Physikalische Chemie*, *239*(4), 610-623. https://doi.org/10.1515/zpch-2024-0610

[17] Amor, N., Noman, M. T., & Petru, M. (2021). Prediction of functional properties of nano TiO_2 coated cotton composites by artificial neural network. *Scientific Reports*, *11*(1), 12235-12249. https://doi.org/10.1038/s41598-021-91733-y

[18] Allangawi, A., Ayub, K., Gilani M. A., Imran M., & Mahmood, T. (2024). Transition metal loaded carbon penta-belt SACs for hydrogen and oxygen evolution reactions and identification of systematic DFT method to characterize the main interacting orbital for non-periodic system. *International journal of hydrogen energy*, *53*, 989-998. https://doi.org/10.1016/j.ijhydene.2023.11.333

[19] Liu, Q., Li, Z., & Zou, C. (2025). A novel four-modal nano-sensor based on two-dimensional Mxenes and fully connected artificial neural networks for the highly sensitive and rapid detection of ochratoxin A. *Talanta*, *283*, 157-178. https://doi.org/10.1016/j.talanta.2024.127157

[20] Wei, X., Xia, D., Li, Y. et al. (2025). Attention-based spatial-temporal synchronous graph convolution networks for traffic flow forecasting. *Applied Intelligence*, *55*(7), 6341-6345. https://doi.org/10.1007/s10489-025-06341-4

[21] Luo, M., Zhang, X., & Long, T. (2023). Modeling and optimization study on degradation of organic contaminants using nZVI activated persulfate based on response surface methodology and artificial neural network: a case study of benzene as the model pollutant. *Frontiers in Chemistry*, *11*(1), 4-30. https://doi.org/10.3389/fchem.2023.1270730

[22] Jung, J., Kim, K., & Choi, J. (2023). Geometric analysis algorithm based on a neural network with localized simulation data for nano-grating structure using Mueller matrix spectroscopic ellipsometry. *Optics Express*, *31*(26), 11-23. https://doi.org/10.1364/OE.507102

[23] Tabatabaei, S. K., Pham, B., & Pan, C. (2022). Expanding the Molecular Alphabet of DNA-Based Data Storage Systems with Neural Network Nanopore Readout Processing. *Nano letters*, *22*(5):1905-1914. https://doi.org/10.1021/acs.nanolett.1c04203

[24] Said, G. E., Tarek, M., & Zen, A. A. (2024). Novel Cobalt/vitamin B_3 metal-organic framework as nano-catalyst in synthesis of some new bis-indole derivatives with staking validation towards Salmonella DNA. *Journal of Organometallic Chemistry*, *1008*(001), 9-18. https://doi.org/10.1016/j.jorganchem.2024.123074

Contact information:

Liang LI
(Corresponding author)
School of Medicine and Health,
Yancheng Polytechnic College, 224005, China
E-mail: liangli@ycpc.edu.cn

Manyu ZHU
School of Medicine and Health,
Yancheng Polytechnic College, 224005, China

Ming LI
School of Medicine and Health,
Yancheng Polytechnic College, 224005, China