Systems & Control Transactions

Research Article
Foundations of Computer Aided Process Design (FOCAPD 2024)
Breckenridge, Colorado, USA. July 14-18, 2024
Peer Reviewed Conference Proceeding

PSE PRESS

Enhancing Polymer Reaction Engineering Through the Power of Machine Learning

Habibollah Safari[a] and Mona Bavarian[a*]

[a] Department of Chemical and Biomolecular Engineering, University of Nebraska-Lincoln, Lincoln, NE, 68588
* Corresponding Author: mona.bavarian@unl.edu.

ABSTRACT

Copolymers are commonplace in various industries. Nevertheless, fine-tuning their properties bears significant cost and effort. Hence, an ability to predict polymer properties a priori can significantly reduce costs and shorten the need for extensive experimentation. Given that the physical and chemical characteristics of copolymers are correlated with molecular arrangement and chain topology, understanding the reactivity ratios of monomers—which determine the copolymer composition and sequence distribution of monomers in a chain—is important in accelerating research and cutting R&D costs. In this study, the prediction accuracy of two Artificial Neural Network (ANN) approaches, namely, Multi-layer Perceptron (MLP) and Graph Attention Network (GAT), are compared. The results highlight the potency and accuracy of the intrinsically interpretable ML approaches in predicting the molecular structures of copolymers. Our data indicates that even a well-regularized MLP cannot predict the reactivity ratio of copolymers as accurately as GAT. This is attributed to the compatibility of GAT with the data structure of molecules, which are graph-representative.

Keywords: Reaction Engineering, Polymerization, Artificial Neural Network, Multilayer Perceptron, Graph Attention Network,

INTRODUCTION

Copolymers are widely used and have a variety of applications such as coatings, in electronic devices, the packaging industry, or pharmaceutical manufacturing[1-3]. Copolymers are often designed with the aim of creating materials that possess the characteristics of their constituent monomers.

The properties of a copolymer are basically determined by the paired monomers' sequence distribution in the constructed copolymers. This distribution is commonly defined by the reactivity ratios, presenting the ratio of each monomer's propensity to react with itself over the inclination to react with another monomer[4]:

$$r_1 = \frac{k_{p,11}}{k_{p,12}}$$

$$r_2 = \frac{k_{p,22}}{k_{p,21}}$$

Here, the $k_{p,xy}$ is the rate coefficient of propagation of radical x with monomer species y. For instance, poly (ethylene-co-vinyl acetate) or EVA is a commercial polymer in various industries and is constructed from ethylene and vinyl acetate. The reactivity ratios for these monomers are r_1=0.88≈1 and r_2=1.03≈1[5]. In a system where both r_1 and r_2 are close to one, the copolymerization tends to produce a random copolymer, and no specific sequence is expected. On the other hand, when $r_1 \ll 1$ and $r_2 \ll 1$, an alternating copolymer can be expected Like Styrene-Maleic Anhydride copolymer or SMA in which r_1=0.02 and r_2=0.003[5]. In this condition, each monomer has a preference to react with the other monomer instead of itself. As evident, varying values of reactivity ratios (r_1 and r_2) in copolymerization lead to different arrangements in the copolymer structure, subsequently influencing the final properties of the copolymer. Thus, having a reasonable prediction for the sequence of monomers in a copolymer chain facilitates the process of producing fit-for-purpose macromolecules. Traditionally, estimation of the reactivity ratio for a new polymer heavily depends on experimental work and the repeatability of the experiments, which are laborious, sluggish, and costly.

Numerous computational methods, such as Density Functional Theory (DFT), are used for predicting the reactivity ratio in copolymers. However, these methods generally incur significant computational costs, rendering them impractical for certain engineering applications[6].

To this end, using of Machine Learning (ML) approaches can be a pragmatic solution for addressing the reaction engineering problems specifically for copolymer synthesis. From a process design perspective, ML models serve as a pivotal tool in predicting the chain topology of copolymers derived from novel monomer pairs. This application, termed the **Forward Design Problem**, utilizes predictive analytics to determine the reactivity ratios of two monomers, thereby identifying the final chain topology of the resultant copolymers. Such predictive capability enables researchers to envision the structural and functional attributes of new materials before their synthesis, streamlining the development process and enhancing the efficiency of material discovery. Conversely, the **Backward Design Problem** represents an equally vital application of ML in polymer science. By integrating ML models with optimization frameworks, it is possible to reverse-engineer the design process to identify monomer structures that yield specific reactivity ratios[7]. This is particularly crucial when aiming for copolymers with precise structural configurations, such as those with alternating monomer sequences achieved when the reactivity ratios of two monomers are significantly less than one. Through optimization, the model identifies 'monomer fingerprints' that are most likely to result in the desired topology, thus guiding the synthesis towards copolymers with predetermined properties and applications. This dual approach—predicting copolymer topologies through forward design and deducing optimal monomer structures for targeted outcomes via backward design illustrates the notable potential of ML in polymer design. This not only accelerates the material development cycle but also opens new realm for the tailored synthesis of copolymers, marking a significant advancement in the field of polymer reaction engineering. The organization of the rest of this paper is as follows. The methodology section presents the data preprocessing and implementation of the Multilayer Perceptron (MLP) and Graph Attention Network (GAT). The result and discussion describe the models' performance in predicting the monomers' reactivity ratios. Eventually, we present some concluding remarks.

METHODOLOGY

Data Curation and Preprocessing

In the large view, the development of an ML model constitutes 4 phases including Data Mining, Data Cleaning, Model Construction, and Performance Assessment. Figure 1 presents a general overview of the development of an ML approach for reactivity ratio prediction in copolymers. In the development of machine learning models, our investigation is positioned within the realm of supervised learning, where each data point is labelled. Our dataset comprises pairs of monomer names, with the associated label being their reactivity ratio during the copolymerization process. Given that the output variable, the reactivity ratio, is a continuous value, our problem is identified as a regression task within the supervised learning framework. A significant challenge in this context is the effective introduction of monomer pairs to the machine learning model (particularly in the development of Multi-Layer Perceptron).

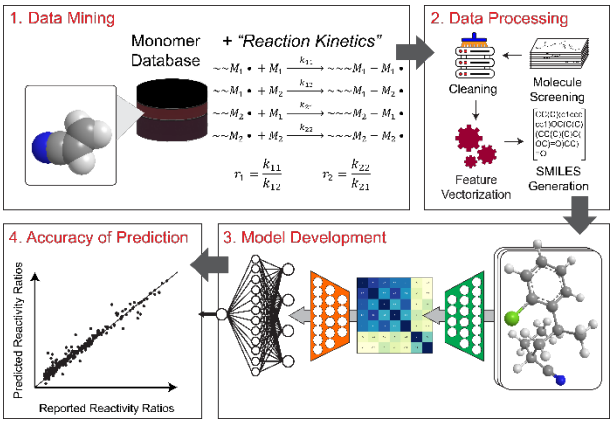

Figure 1. Schematic Workflow for Presentation of an ML Approach for Reactivity Ratio Prediction.

After the collection of reactivity ratios of monomers from different references, we evaluated the data in terms of value and distribution. It is generally observed that extracting patterns from a dataset with a distribution close to normal is more effective. To ameliorate the training process of the ML model, we applied a square root transformation to the raw data, aiming to achieve a distribution that is more closely aligned with a normal distribution. Figure 2. represents the original reactivity ratio (r_1 and r_2) distribution and also schemes the data distribution when we impose a square root transformation on our data ($\sqrt{r_1}$ and $\sqrt{r_2}$).

In the next phase, we convert the monomers into the Simplified Molecular-Input Line-Entry System (SMILES) format using open-source cheminformatics toolkits. This allows us to transform the monomers into a machine-readable language, suitable for machine learning development.

Multi-Layer Perceptron

Here, a Multilayer Perceptron (MLP) for the prediction of the reactivity ratio in copolymers is considered. For converting the SMILES to a numerical vector that is ANN-compatible, a Morgan fingerprint with 2048 bits for each monomer was employed. In the

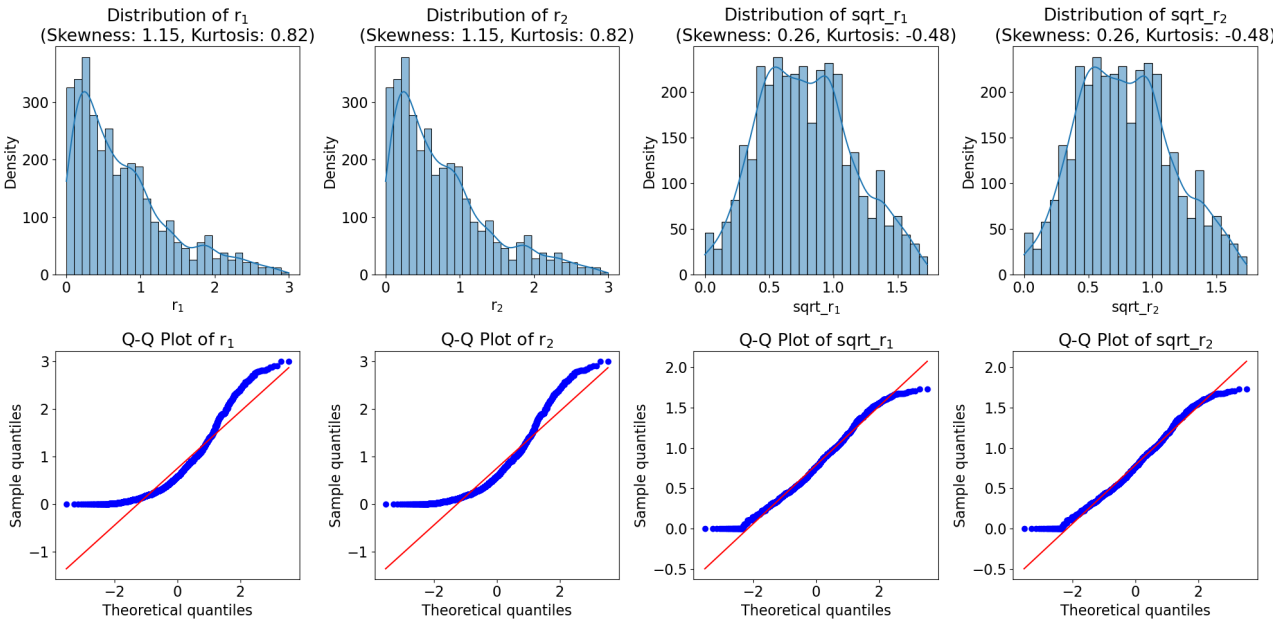

Figure 2. Distribution of original data (r_1 and r_2) and distribution of data with applied square root transformation ($\sqrt{r_1}$ and $\sqrt{r_2}$). As shown in the figure, the skewness of data reduces significantly and make data close to normal distribution. In the quantile-quantile (Q-Q) plot, the degree to which our data aligns with the reference line (y = ax+b) provides an indication of how closely the data conform to a normal distribution. A close alignment suggests that the data are approximately normally distributed.

following, the fingerprint vectors of the two monomers were concatenated, and the concatenated vector was considered as the input of the MLP. So, the input of the model is a one-dimensional vector ($X \in R^{4096}$), and the outputs of the model are $\sqrt{r_1}$ and $\sqrt{r_2}$ ($y \in R^2$).

In defining the structure of the model, a proposed MLP with 4096 inputs, 80 neurons in the first hidden layer, 40 neurons in the second hidden layer, and 2 outputs is used[8]. Regarding the full connectivity of the neurons in all layers, the MLP performance was improved using the dropout technique. **Figure 3** represents a schematic overview of the multilayer perceptron used for reactivity ratio estimation. For a better training process, it was found that the implementation of a regularization technique can improve the Mean Square Error (MSE) significantly. In this regard, a Grid Search Optimization method was implemented to find L_1 Regularization, L_2 Regularization, and Dropout Rate. It was found that using Dropout Rate = 0.65 can significantly ameliorate the MLP performance. Using the optimum dropout rate in the MLP could significantly improve the reported MSE in the modified model (0.1 in the test dataset and 0.08 in the training dataset) compared to the reference model[8].

For the training of our optimized Multilayer Perceptron (MLP), the dataset was allocated as follows: 10% was reserved for testing, while the remaining 90% was utilized for both training and validation. Specifically, of the data allocated for model training and validation, 90% was used for actual training purposes, and the remaining 10% served as validation data. This approach of incorporating a validation subset within the training data allows for regular assessment of the model's performance against overfitting. By doing so, we ensure that the model not only learns from the training data but also generalizes well to unseen data, thereby enhancing its reliability and applicability in real-world scenarios.

Graph Attention Network

In the realm of reaction engineering, the structural intricacies of molecular data present unique challenges and opportunities for computational analysis. A promising approach for addressing these challenges is the adoption of graph-based machine learning approaches in which molecules are represented as graphs, atoms as nodes, and chemical bonds as edges. This approach appears to be an encouraging solution for various problems, including reactivity ratio prediction in polymer science. Graph Attention Networks (GAT) are a specific type of Graph Neural Network (GNN) that incorporate attention mechanisms to specify different weights to different nodes in a graph[9,10].

Herein, In the second part, a Graph attention Network with an Attentive Fingerprint was utilized. This approach employs a Recurrent Neural Network (RNN) and an Attention Mechanism for the extraction of the

most important features from the input in a molecule structure. The attention mechanism in the GAT operates based on the three which are mechanism-alignment, weighting, and context operation[11]:

The alignment equation is represented as follows:

$$e_{vu} = (W.[h_v, h_u]) \quad (1)$$

For weighting, the softmax function is applied:

$$a_{vu} = softmax(e_{vu}) = \frac{\exp(e_{vu})}{\sum_{u \in N(v)} \exp(e_{vu})} \quad (2)$$

The context vector is then calculated using the Exponential Linear Unit (ELU) function:

$$C_v = elu(\sum_{u \in N(v)} a_{vu}.W.h_u) \quad (3)$$

Here, $'v'$ denotes a specific target node or atom, $'N(v)'$ its neighboring nodes, $'h_v'$ the states vector of node $'v'$, $'h_u'$ the states vector of a neighboring atom or node $'u'$, and $'W'$ the learnable weight matrix indicating relationships between the target node and its neighbors. The alignment scores are calculated using Equation 1 and then normalized (Equation 2). Subsequently, the context vector is formulated using the ELU function, allowing a non-zero gradient for negative inputs[11]. Features close to a score of 1 exert more influence on the output, while those closer to 0 are deemed less significant, and negative values result in feature exclusion.

The second stage involves constructing a viable model incorporating a Gate Recurrent Unit (GRU), which enhances the RNN by adding reset and update gates. This allows the GRU to selectively retain and disregard information, thereby augmenting RNN's memory capabilities. The GRU functions in two phases within the model: messaging and readout, described mathematically in Equations (4) and (5).

$$C_v^{k-1} = \sum_{u \in N(v)} M^{K-1}(h_v^{k-1}, h_u^{k-1}) \quad (4)$$

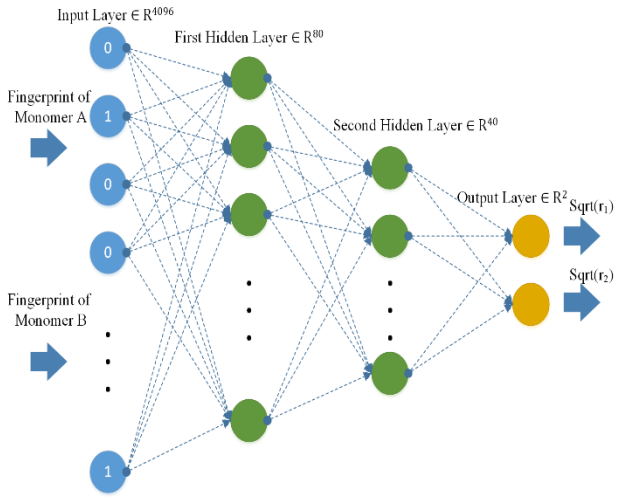

Figure 3. Representation of MLP Architecture

$$h_v^k = GRU^{k-1}(C_v^{k-1}, h_v^{k-1}) \quad (5)$$

In the messaging stage of the GAT model, the message function, denoted as M^{k-1} plays a crucial role. This function operates at the $k-1$ iteration, where it aids in assimilating the learned features of the nodes. During this phase, the representation of the nodes within the molecules is compiled. The message function aggregates details from neighboring nodes on the graph for each target node. This process is pivotal as the graph attention mechanism focuses on collating data from all adjacent nodes in the messaging phase to effectively update their state in the subsequent read-out phase. In this context, the GRU acts by integrating inputs from the previous state vector h_v^{k-1} of the target node and the attention context C_v^{k-1} from its neighboring nodes. In the readout phase, the GRU updates the current hidden state of the target node by employing information obtained from the messaging phase and the node's prior hidden state.

The representations of the target nodes, once learned, are then employed in the read-out phase to predict molecular properties. Detailed explanations of the functioning and application of these processes and equations (1-5) within the GAT framework can be found in the cited references[11] The model further refines its accuracy by using features such as atom symbols, neighboring atoms, atom masks, bond types, and neighboring bonds to effectively differentiate each target node from its neighbors. In our study, the Multi-Input-Multi-Output Graph Attention Network (MIMO GAT) was used as an advanced version of the Graph Attention Network. This new model includes a special multimodal fusion block, making it different from the Attentive FP model. This network was employed to predict the reactivity ratios of monomers, using SMILES notations of monomers and copolymers. Leveraging multi-task learning, we first converted molecular structures into graph representations using RDKit for feature extraction. These features are then encoded and processed through individual Graph Attention Modules within MIMO GAT. Each module incorporated attentive-layer embedding for both atom and full molecule levels. After that, the outputs from these modules were concatenated and fed to fully connected layers for final prediction. The details of this approach can be found in the reference[10].

RESULT AND DISCUSSION

Traditional Artificial Neural Networks (ANNs) often employ a black-box methodology for problem-solving, where their primary goal is to identify patterns or relationships within raw input data, without an explicit focus on underlying physical laws or domain-specific knowledge. While these approaches prove effective in numerous scenarios, they may fail to capture crucial

insights tied to the intrinsic properties or governing principles of the system in question.

In contrast, there is an emerging interest in the development of specialized ANNs tailored for particular data types and tasks. For example, Graph Attention Networks (GATs) have shown superior performance compared to Multi-Layer Perceptron (MLPs) in handling graph-structured data. This advancement is largely due to GATs' ability to more accurately represent and process-relational and structural information inherent in such data. Given that chemical science data fundamentally relates to molecular structures, which are naturally representable as graphs, Graph Neural Networks, including GATs, emerge as particularly promising for tackling challenges in polymer science. Their capacity to intuitively map and interpret the complex, interconnected nature of molecular structures positions them as a fitting choice for this field.

In this study, an MLP model was developed for the prediction of the reactivity ratio of paired monomers in copolymerization. It was observed that a regularized MLP predicted the reactivity ratios with improved accuracy, yet the Mean Square Error (MSE) could not be reduced to less than 0.1 in the evaluation of the test dataset. Although this MSE is substantially lower than those reported in other references[8], the model does not seem to be entirely appropriate for accurate prediction in this context.

Figure 4. compares the R^2 score on the test dataset when MLP and GAT are used as the reactivity ratio estimator. As shown in the figure, GAT significantly outperforms in comparison to MLP in the prediction of r_1 and r_2. While the R^2 score could hardly be more than 10% in using MLP, the GAT could achieve 85±5%[10].

It is evident that there is a substantial enhancement in the prediction of reactivity ratios for copolymerization when using GATs. This suggests that employing a graph-based machine-learning approach can outperform conventional ANNs. The key advantage lies in the graph-based model's ability to incorporate the natural intrinsic structure of the data into the training process, enabling it to capture patterns more effectively than other machine learning models, which may treat each model completely as a black box without considering the system's inherent nature.

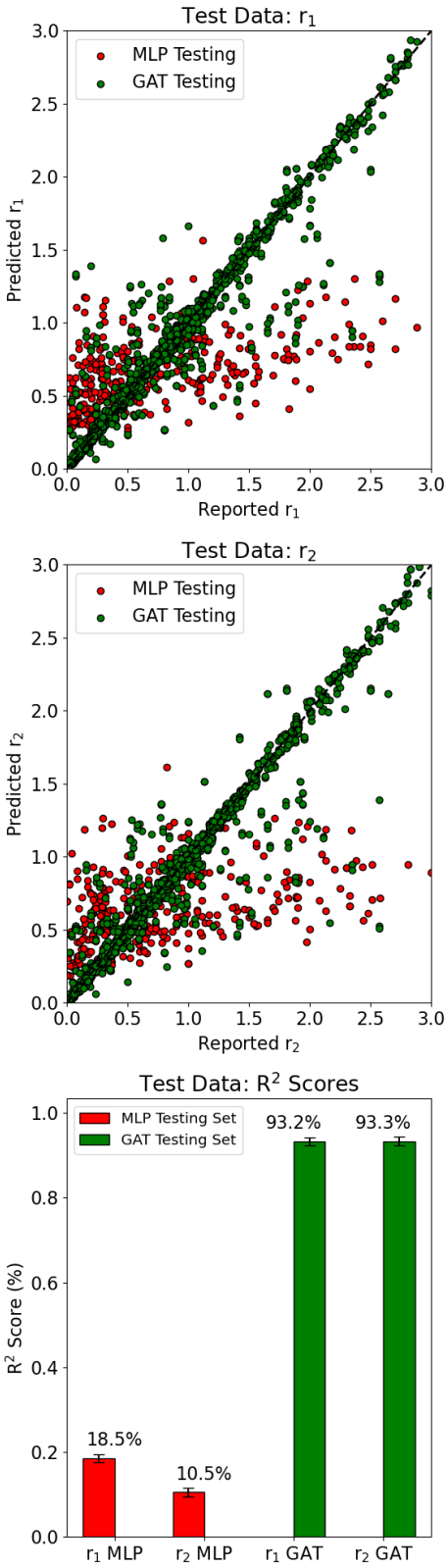

Figure 4. Comparing the R^2 score in test data for Grah Attention Network (GAT) and Multilayer Perceptron(MLP)

CONCLUSION

Here, this study clearly demonstrates the superiority of Graph Attention Networks (GATs) over traditional Multilayer Perceptron (MLPs) in the context of predicting reactivity ratios for copolymers. The substantial improvement in prediction accuracy, as evidenced by the R^2 scores, underscores the effectiveness of GATs in handling complex, graph-structured data inherent in chemical science. By integrating the intrinsic structural information of molecular data into their learning process, GATs not only outperform conventional ANNs but also pave the way for more nuanced and accurate models in polymer science. This comparison highlights the potential of graph-based machine learning approaches to revolutionize data analysis in fields where understanding the interconnected nature of data is critical.

ACKNOWLEDGMENT

The authors acknowledge the support of the National Science Foundation (NSF) under the Award Number 2238147.The authors would like to extend their thanks to the Holland Computing Center at the University of Nebraska-Lincoln for the essential computational resources and support provided, which played a crucial role in the conduct of this research.

REFERENCES

1. Z Dang, F Guo, Z Wu, K Jin, J Hao: Interface Engineering and Device Applications of 2D Ultrathin Film/Ferroelectric Copolymer P(VDF-TrFE). Adv Phys Res, 2:1 (2023), doi: 10.1002/apxr.202200038.
2. N Politakos: Block Copolymers in 3D/4D Printing: Advances and Applications as Biomaterials. Polymers, 15:2 (2023), doi: 10.3390/polym15020322.
3. M Gigli, N Lotti, M Gazzano, V Siracusa, L Finelli, A Munari, M Dalla Rosa: Biodegradable aliphatic copolyesters containing PEG-like sequences for sustainable food packaging applications. Polym Degrad Stab, 105:1, 96-106 (2014), doi: 10.1016/j.polymdegradstab.2014.04.006.
4. FR Mayo, C Walling: Chem Rev, 1950, 46, 191-287.
5. KI Takahashi, H Mamitsuka, M Tosaka, N Zhu, S Yamago: CoPolDB: a copolymerization database for radical polymerization. Polym Chem, 15:10, 965-971 (2024), doi: 10.1039/d3py01372c.
6. M Dossi, D Moscatelli: A QM Approach to the Calculation of Reactivity Ratios in Free-Radical Copolymerization. Macromol React Eng, 6:2-3, 74-84 (2012), doi: 10.1002/mren.201100065.
7. T McDonald, C Tsay, AM Schweidtmann, N Yorke-Smith: Mixed-integer optimization of graph neural networks for computer-aided molecular design. Comput Chem Eng, 185 (2024), doi: 10.1016/j.compchemeng.2024.108660.
8. K Farajzadehahary, X Telleria-Allika, JM Asua, N Ballard: An artificial neural network to predict reactivity ratios in radical copolymerization. Polym Chem, 14:23, 2779-2787 (2023), doi: 10.1039/d3py00246b.
9. W Sha, Y Li, S Tang, J Tian, Y Zhao, Y Guo, W Zhang, X Zhang, S Lu, Y-C Cao, S Cheng: Machine learning in polymer informatics. InfoMat, 3:4, 353-361 (2021), doi: 10.1002/inf2.12167.
10. T Nguyen, M Bavarian: Machine learning approach to polymer reaction engineering: Determining monomers reactivity ratios. Polymer, 275 (2023), doi: 10.1016/j.polymer.2023.125866.
11. P Veličković, G Cucurull, A Casanova, A Romero, P Liò, Y Bengio: Graph Attention Networks. [Online]. Available: http://arxiv.org/abs/1710.10903 (2017).