ACS POLYMERS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

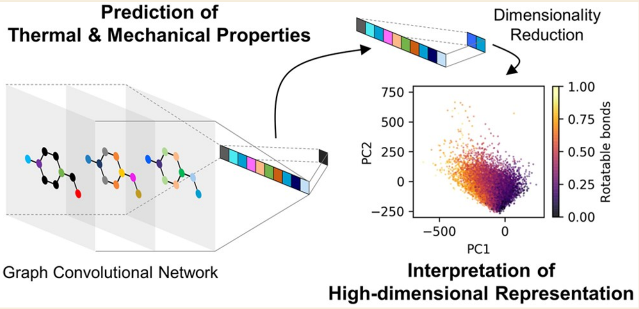
Prediction and Interpretation of Polymer Properties Using the Graph Convolutional Network

Jaehong Park, Youngseon Shim,* Franklin Lee, Aravind Rammohan, Sushmit Goyal, Munbo Shim, Changwook Jeong,* and Dae Sin Kim

Cite This: *ACS Polym. Au* 2022, 2, 213−222 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We present machine learning models for the prediction of thermal and mechanical properties of polymers based on the graph convolutional network (GCN). GCN-based models provide reliable prediction performances for the glass transition temperature (T_g), melting temperature (T_m), density (ρ), and elastic modulus (E) with substantial dependence on the dataset, which is the best for T_g ($R^2 \sim 0.9$) and worst for E ($R^2 \sim 0.5$). It is found that the GCN representations for polymers provide prediction performances of their properties comparable to the popular extended-connectivity circular fingerprint (ECFP) representation. Notably, the GCN combined with the neural network regression (GCN-NN) slightly outperforms the ECFP. It is investigated how the GCN captures important structural features of polymers to learn their properties. Using the dimensionality reduction, we demonstrate that the polymers are organized in the principal subspace of the GCN representation spaces with respect to the backbone rigidity. The organization in the representation space adaptively changes with the training and through the NN layers, which might facilitate a subsequent prediction of target properties based on the relationships between the structure and the property. The GCN models are found to provide an advantage to automatically extract a backbone rigidity, strongly correlated with T_g, as well as a potential transferability to predict other properties associated with a backbone rigidity. Our results indicate both the capability and limitations of the GCN in learning to describe polymer systems depending on the property.



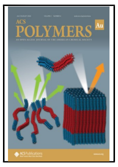
KEYWORDS: *machine learning, graph convolutional network, molecular featurization, backbone rigidity, polymer property prediction, neural network*

1. INTRODUCTION

Machine learning (ML) has been influencing many facets of materials sciences and chemistry by providing powerful techniques of leveraging data, which enables rapid prediction of properties, the discovery of novel molecules, and the route of synthesis.[1,2] An application of ML to polymer systems is promising because they exhibit interesting physical phenomena over extensive time and length scales in nonequilibrium glassy states for which experimental characterization and simulations are relatively costly and difficult.[3,4] This indicates also the challenges in the application of ML because the power of ML hinges on a high-quality database.[4] The promise of ML triggered a surge of studies on polymers to predict a range of physical properties with various learning algorithms,[5−9] to build a database of the prediction models,[10] and to augment the polymer structure database.[11] Also, the transfer learning and generative model were combined to overcome the problem of limited data such as thermal conductivity and to discover the novel polymers.[12]

ML models learn the complex patterns in data through the suitable representation of the data.[13] The conventional molecular representation is a list of features hand-crafted based on domain knowledge for the quantitative structure−activity relationships.[14] On the other hand, extended-connectivity circular fingerprints (ECFPs) are widely used methods that exploit the graph structures of molecules to encode the presence of particular substructures of chosen scale into discrete bit vector representations.[15] Deep learning-based approaches like the graph convolutional network (GCN) are the state-of-the-art learning method for graph-structured data and can be used to generate representations of molecules.[16] The GCN-based representation is promising compared to the hand-crafted descriptors or the ECFP representations because it can be optimized for data domains through the supervision.[17] Also, the GCN generates continuous representations

Received: November 2, 2021
Revised: January 6, 2022
Accepted: January 6, 2022
Published: January 21, 2022



which are desirable for the exploration of chemical space for the inverse design of molecules.[18] There have been comparative studies using different representations for benchmark datasets for small molecules or drug molecules, where the GCN showed favorable performance compared to other representations for many cases.[14,19−22] However, such benchmark studies on the polymers are very limited, and the application of the GCN to polymers is still elusive. A recent study explored different kinds of molecular representations including the ECFP and GCN in conjunction with various learning algorithms to predict polymeric properties, where the GNN was found to be much inferior to other representations.[8]

Thermal and mechanical properties of the polymers are known to strongly depend on the monomer structure, molecular interactions, and their morphological characteristic, such as the rubbery or glassy state of the amorphous phase and crystallinity.[23−30] For example, glass transition temperature was found to increase with the chain rigidity of the polymers,[23−26] where rigid rings in the polymer backbone increase steric hindrance because of their limited structural reorganization with high dihedral free energy.[30] One of the thermoplastic classes, polyamides, has higher melting points compared to high-density polyethylene, where the neighboring chains are networked to be well packed in an energetically more stable conformation via hydrogen-bonding between adjacent chains.[31] It should be understood how the polymer structure affects their macroscopic properties to design the next generation of functional materials. There have been few ML studies to understand the relationship between the molecular feature and property of polymer systems. Recently, a ML workflow was reported to predict melting points of small chemicals in which the graph attribution technique was used to analyze atom-level contributions to the predicted value of melting point.[9]

The increase of model complexity in ML has led to improvement of prediction performance but blurred the reasoning behind the prediction of the model.[13] Interpretability of ML is crucial to understand and improve behaviors of the model and provides the possibility to discover the new physical and chemical principles of polymer design.[32,33] The key issue we undertake in this article is how the GCN-based ML models work on the structural featurization to predict thermal and mechanical properties of polymers. To address this issue, we built the GCN-based models for the glass transition temperature (T_g), melting temperature (T_m), density (ρ), and elastic modulus (E) and analyze how the learned structure−property correlations manifest in the representation spaces. The outline of this paper is as follows: In Section 2, we give a brief description of the models and methods employed in this study. The GCN-based models to predict $T_{g,}$ T_m, ρ, and E of polyamides are investigated and compared with ECFP-based ones in Section 3. It is also scrutinized through the dimensional reduction how GCN representation captures the structural feature, strongly correlated with the structural rigidity, and associated polymer properties. Concluding remarks are offered in Section 4.

2. METHOD

2.1. Data Collection

We collected data for polyamides, which is an important class of engineering polymers exhibiting high thermal stability and mechanical strength. The dataset employed in this work contains a total of 2687 different structures of organic polyamides of which experimental data were manually collected from the PoLyInfo database.[34] The distributions of the experimental property data for T_g, T_m, ρ, and E are displayed in Figure 1 for 1388, 942, 390, and 306 homopolymer structures, respectively. Multiple experimental values of the property for the single polymer structure were reduced to their mean value. We excluded the polymers for which the standard deviations of the reported property values exceed the value σ_{max}, where the σ_{max} is set to be 30 K for T_g and T_m. We note that our E dataset shows an imbalance in their distribution, where the data in the region of high modulus (E > ∼6 GPa) are significantly scarce with only six polymer structures as shown in Figure 1d. For the E datasets, we trained the prediction models either excluding or including those six polymers.

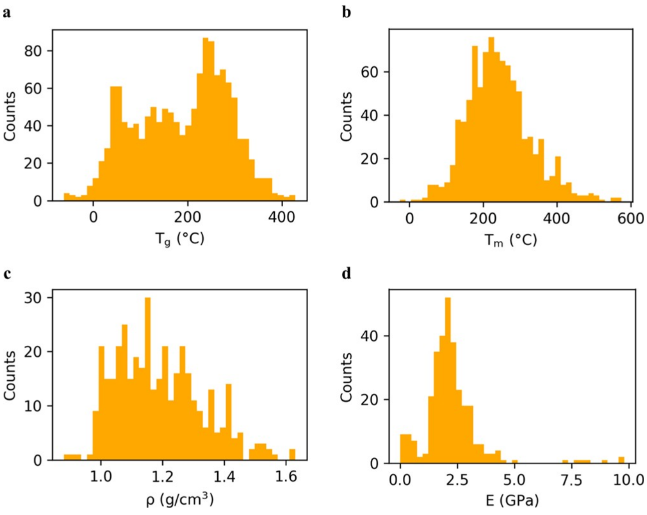


Figure 1. Data distributions of (a) glass transition temperature (T_g), (b) melting temperature (T_m), (c) density (ρ), and (d) elastic modulus (E) of 1388, 942, 390, and 306 polyamides, respectively, collected from the PoLyInfo database.

2.2. Molecular Representations

Different variants of the GCN have been developed to process the graph data.[16] We used the most popular and simple variant of the GCN devised by Kipf and Welling[35] A molecular graph is represented by $G(\nu, \varepsilon)$, where ν is the set of nodes (atoms) and ε is the set of edges, that is, connectivity of atoms. The attributes of each node (atomic feature) are represented by the vector $x_n^l \in \mathbb{R}^{m_l}$ for n-th node and are iteratively updated by the approximated spectral graph convolution $H^{l+1} = \sigma(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^lW^l)$, where m_l is the dimension of the atom feature vector at the l-th layer, $H^l = [x_1^l, x_2^l, ..., x_N^l]^T$ is a feature (activation) matrix, N is the number of atoms, $\tilde{A} = A + I$ is an adjacency matrix of the molecular graph added with the self-connection, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ is a degree matrix and $W^l \in \mathbb{R}^{m_l \times m_{l+1}}$ is a weight matrix, and σ is an activation function. Figure 2a shows schematics of the graph convolution process by which the feature vector of an atom is updated by aggregating features of neighboring atoms through the convolutional layers. The weight parameters in W^l for the convolution operation can be optimized through the training to adapt to data, which is one of the important advantages of the GCN. In essence, the convolution operations in the GCN act as a low-pass filter

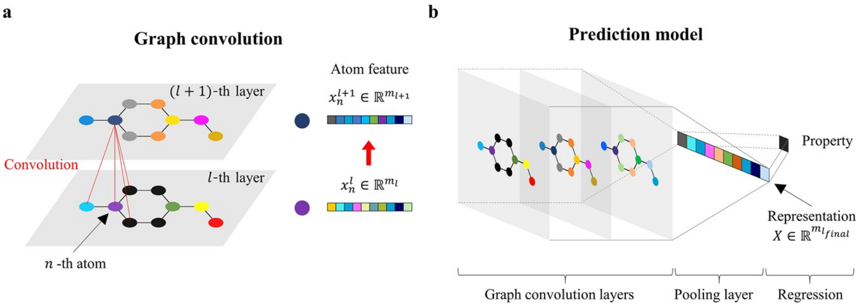


Figure 2. Property prediction model based on the GCN. (a) Feature vector of n-th atom x_n^l is updated iteratively through the l-th convolutional layer by the graph convolution that aggregates the features of neighboring atoms. (b) Graph-level molecular representation vector X is obtained by the pooling layer that sums up the feature vectors of all the atoms at the final convolution layer. The vector X is input to the regression algorithms (linear regression (LR) only shown here) for the property prediction.

Table 1. Prediction Performances of the ML Models for the Glass Transition Temperature (T_g), Melting Temperature (T_m), Density (ρ), and Elastic Modulus (E) Using Different Molecular Representations and Regression Algorithms[a]

		T_g		T_m		ρ		E	
featurization	regression	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
GCN	LR	34.09 (3.07)	0.87 (0.03)	44.81 (2.25)	0.70 (0.05)	0.073 (0.018)	0.58 (0.22)	0.49 (0.04)	0.47 (0.20)
	NN	**29.98** (2.16)	**0.90** (0.02)	**40.37** (2.94)	**0.76** (0.05)	**0.064** (0.013)	**0.70** (0.17)	**0.47** (0.06)	**0.54** (0.20)
ECFP	LR	33.23 (2.88)	0.87 (0.03)	44.28 (2.70)	0.67 (0.05)	0.074 (0.013)	0.64 (0.15)	0.53 (0.06)	0.35 (0.21)
	NN	31.11 (1.80)	0.89 (0.01)	44.85 (2.52)	0.69 (0.06)	0.068 (0.014)	0.67 (0.16)	0.51 (0.07)	0.42 (0.21)

[a] The values are mean and standard deviations (in the parenthesis) over 5-fold CV splits. The bold case indicates the best performances for each property dataset. The units of RMSE are K, g/cm^3, and GPa for T, ρ, and E, respectively.

which smooths out the node attributes by averaging them over neighboring nodes and facilitates the classification of nodes and graphs.[36,37] We set the initial atom feature vector by a concatenation of one-hot vectors containing information on the atom types, the number of attached hydrogen atoms, degree of an atom, implicit valence, and aromaticity. The initial feature vector consists of 25 binary components, $m_0 = 25$, in total (see the Supporting Information Table S1 for details of the atom features), whose dimension can change depending on the m_l in the subsequent convolutional layers.

We used identical convolutional layers to have the same m_l through $1 \leq l \leq l_{\text{final}}$.We used the ReLU for the activation function σ and set the initial weight parameters W^l in the convolutional layers with the random values using the scheme proposed by He et al.[38] The final graph-level molecular representation is obtained by the pooling operation which sums up all the atom feature vectors at the final convolutional layer as $X = \sum_{n=1}^{N} x_n^{l_{\text{final}}} \in \mathbb{R}^{m_{l_{\text{final}}}}$, which is invariant under the permutation of atoms. The molecular representation X is input to the regression algorithms to predict the target properties as shown in Figure 2b. The ECFP has been also known as the Morgan fingerprint to update the features of an atom by aggregating the information of neighboring atoms iteratively within a radius r around the atom, in analogy to the GCN.[17] Particular substructures are then encoded to the molecular-level representation in the form of a fixed-length bit vector by the hash function.[15] It should be noted that both the GCN and ECFP take a monomer structure as an input and are of the monomer-level representations of the polymers. We used RDKit[39] to process the molecular graph data and to obtain the ECFP.[15]

2.3. Model Selection and Assessment

We used LR and a fully connected neural network (NN) for the property predictions based on the molecular representations. We use the NN with two hidden layers. The 5-fold cross-validation (CV) was used to estimate the performance of the prediction models. We split the data into the 5-fold training/test sets for which model selections and assessment were performed instead of training/validation/test sets considering the relatively small size of data.[40] We optimized hyper-parameters such as the depth and width of the GCN layers, the width of the NN layers, learning rate, weight decay for L2 regularization, and epochs through the grid search for each dataset. For the ECFP, the radius and the number of bits were optimized. We choose the best models based on the loss function averaged over the 5-fold validation sets for each dataset. The performance of models was measured by two metrics such as root-mean-square error (RMSE) and coefficient of determination R^2 to evaluate the best models. We used Pytorch to build the models and adaptive moment estimation (ADAM) algorithm for the training.[41]

3. RESULTS AND DISCUSSION

We begin by presenting the performances of the GCN-based models, compared to the ECFP-based models. Table 1 and Figure 3 show the prediction accuracies of our ML models with different molecular representations and regression algorithms. We notice that results with NN regression show clear improvements compared to those with the LR algorithm, regardless of the molecular representations and the types of property. Figure 4 shows scatter plots for the GCN combined with the NN (GCN-NN) and the ECFP combined with NN (ECFP-NN) models comparing the experimental values and predicted values. As for the T_g, both the GCN-NN and the ECFP-NN provide very excellent prediction accuracies to be RMSE ~30 K and R^2 ~ 0.9 shown in Figure 3, Figure 4a,e, and Table 1, while the LR models give rise to RMSE ~ 35 K. This indicates that both GCN and ECFP representations desire to be nonlinearly processed for better fitting of the data. The

https://doi.org/10.1021/acspolymersau.1c00050

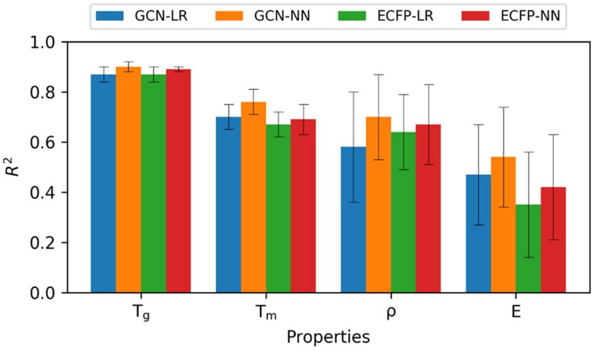


Figure 3. Bar chart for the comparison of the performances across the properties measured in the R^2 score. The error bar indicates the standard deviations over 5-fold CV splits.

GCN combined with LR (GCN-LR) models show the performance metrics comparable to those of the ECFP combined with LR (ECFP-LR) models within the range of standard deviations. Another noteworthy point is that the GCN-NN models slightly outperform the ECFP-based models for all the properties. In particular, as for the T_m dataset, the GCN-NN model outperforms the ECFP-NN model with a considerable margin, that is, ∼10% increase in R^2. The GCN-NN model predicts T_m with RMSE ∼40 K smaller by ∼5 K than the ECFP-NN model. Although the model performances for the properties such as ρ and E are degraded compared to those for T_g and T_m, GCN representations hold their performances better than the ECFP ones. In Figure 3, the prediction accuracy becomes lower in the order of T_g, T_m, ρ, and E for all the models, where the R^2 scores obtained with GCN-NN models are 0.90, 0.76, 0.70, and 0.54, respectively. The prediction accuracy of E is the lowest and significantly degraded when we extend their dataset into the region up to ∼10 GPa. We turn to this later.

The best models to predict the given properties of polyamides are GCN-based ones, attributed to the fact that the GCN learns the representations from data.[16,17,35] The result is in contrast to the recent study[8] which presented the ML prediction models trained on the PoLyInfo datasets with similar sizes for the T_g (1034 polymers), T_m (641 polymers), and ρ (318 polymers). They investigated various learning algorithms such as the random forest, support vector machine, and the NN regression in conjunction with the ECFP and GCN where the predictions by their GCN models were significantly inferior to those by the ECFP-based models. While the best performing models in the study based on the ECFP showed performances comparable to our work (R^2 ∼ 0.85 for T_g, R^2 ∼ 0.64 for T_m, and R^2 ∼ 0.56 for ρ), the GCN-based models showed poor performances (R^2 ∼ 0.71 for T_g, R^2 ∼ −0.15 for T_m, and R^2 ∼ 0.26 for ρ). The direct comparison of model performances with their results is not straightforward because the adopted GCN architecture,[42] methods of model evaluations, and datasets are different in details. However, we have observed consistently for our datasets that the GCN representations outperformed the ECFP ones at least with a small margin provided that the models are properly optimized. We note that hyperparameters largely affect the performance and optimal ones vary with datasets. The prediction performances depending on the hyperparameters such as the number of nodes and the number of layers are given in Figure 5. The prediction performances for T_g and T_m show nonmonotonic behaviors with hyperparameters of the GCN in Figure 5a,b. The best GCN-NN model for T_g has six convolutional layers with hundred nodes, while that for T_m has three convolutional layers in the GCN. On the other hand, the performance of the ρ prediction model is monotonically degraded with increasing the model complexity and the best with the single convolutional layer with the hundred nodes. The model for E does not show a clear trend, and the best one has five convolutional layers with the three hundred nodes (see the Supporting Information Figure S1 for the ECFP, and Tables S2 and S3 for the details of the hyperparameters).

To validate our prediction accuracies of T_g and T_m, we address how the final prediction performance is affected by initial uncertainties in the employed dataset. The irreducible error accounts for the unremovable portion of the prediction error, originating from the noise of the target value (measured

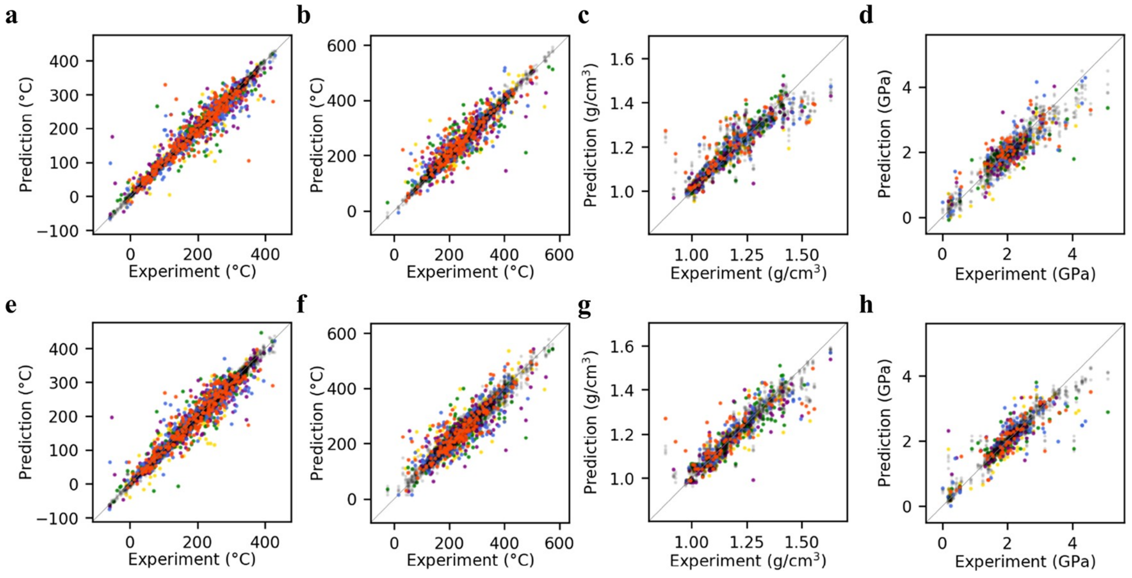


Figure 4. Scatter plots for the ground truth and prediction using (a−d) GCN-NN and (e−h) ECFP-NN models for T_g, T_m, ρ, and E, respectively; 5-fold validation data points are colored by each fold, while training data points are colored by gray.

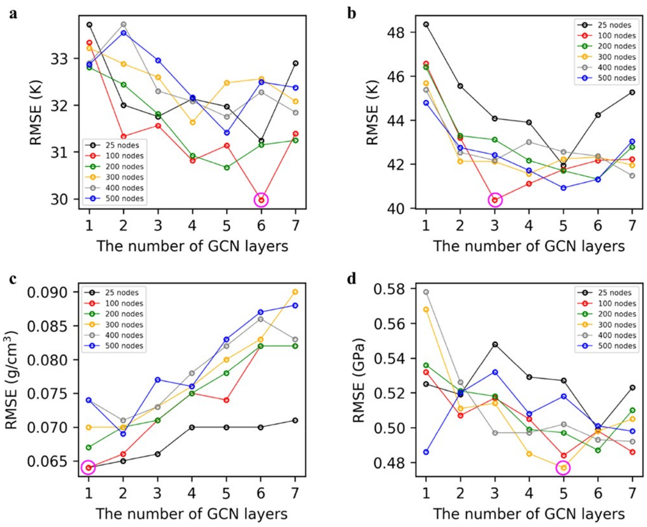


Figure 5. Prediction performances of the GCN-NN models depending on the hyperparameters of GCN for (a) T_g, (b) T_m, (c) ρ, and (d) E datasets, respectively. The best models are marked by magenta circles.

values of T_g and T_m) around the true value in the data. The irreducible errors in our T_g and T_m datasets are estimated by ∼30 K which is the maximum value of the standard deviations of the measured values over the polymers. The uncertainties of datasets arise from not only the method of experimental measurements but also the complex polymer properties associated with morphological characteristics, crystallinity, cross-linking, chain length, and molecular weight. For example, the T_g is not considered to be a thermodynamic state variable and highly depends on various experimental processing conditions such as a cooling rate, but the detailed information is not available within the database. It might be possible for the models to make more sophisticated predictions in the presence of an available database of relationships between the T_g and cooling rates. It is recently pointed out that improvements of the machine prediction accuracies should be intertwined with the database that provides retrieval of relevant information.[4,43] Eventually, the important factors beyond what is considered here should be included and reflected as a feature to improve the model performance for the polymer design.

The prediction performances of all the models for E indicate considerable difficulties in learning the mechanical property. The size of the E dataset (306 polymers) is the smallest compared to the other datasets, which can be thought of as a reason for the degraded performance because it is generally improved with increasing the data size.[44] To check the effect of the dataset size, we trained the GCN-NN models with the reduced size of data of T_g (random sampling of 306 polymers) and observed the R^2 scores consistently larger than ∼0.8. The result indicates that the size of data is not the major factor, which is also corroborated by the performances (R^2 ∼ 0.7) of the ρ models with a similar data size (390 polymers). On the other hand, it can be attributed to irreducible errors in the dataset arising from the divergence in the experimental measurements. Moreover, when we include the six data points into datasets with high E beyond ∼6 GPa in Figure 1d, the RMSE of the GCN-NN (ECFP-NN) model dramatically decreases from ∼0.47 (0.51) GPa to ∼1.0 (1.0) GPa with negative R^2 scores. Different fabrication processes of polymer materials can dramatically change morphological features and properties. For example, the elastic moduli of rigid aromatic poly(*p*-phenylene terephthalamide) and poly(*m*-phenylene isophthalamide) are in the wide range of 10−180 GPa and 3−90 GPa, respectively.[34] This reveals the possibilities of higher irreducible errors from the deviations in the measured values for high-modulus polymers. Another reason could be that the imbalance in the data distribution of E leads to performance degradation. For the classification tasks, imbalanced learning provides several strategies that balance either the data distribution itself or the cost functions for the rare observations on a minor class.[45,46] However, the study for the regression task is yet an emerging field.[47]

The limitation of the GCN in describing various levels of structural information required to learn the structure−property correlations could be an important factor. For instance, the

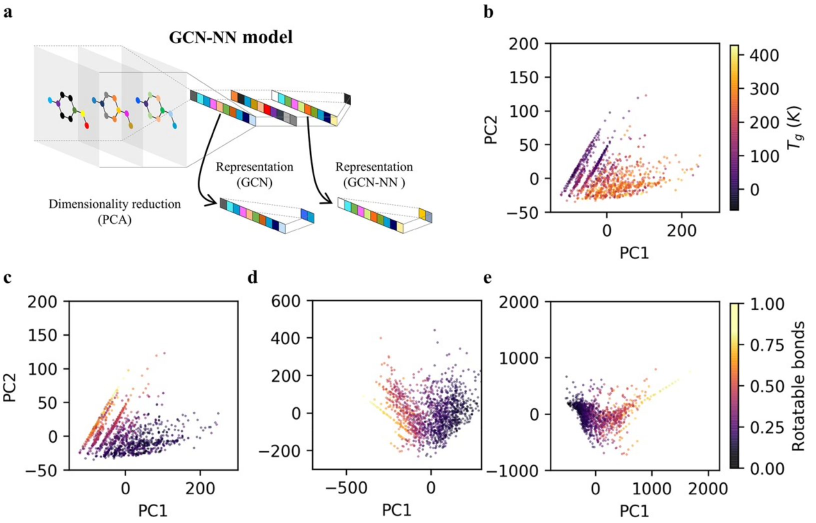


Figure 6. (a) Schematics of the GCN-NN model for T_g prediction. The high-dimensional representations by the GCN layer and GCN-NN layer are reduced to two-dimensional subspaces spanned by the PC1 and PC2 for the analysis. Distributions of (b) experimental T_g values and (c−e) fraction of rotatable bonds of polymers in the PoLyInfo dataset on the representation spaces. Polymers are mapped by (b, c) GCN without training (random weight), (d) trained GCN, and (e) trained GCN-NN to get their representations.

https://doi.org/10.1021/acspolymersau.1c00050

properties of polymers are strongly influenced by intermolecular interactions, polymer chain conformations, and associated morphologies, including the orientational ordering as well as the intermolecular π-π stacking of the polymers.[27] A crystalline form is necessary for the polyamide like poly(p-phenylene terephthalamide) to acquire high strength through the dense π-π stacking between conjugated rings, and the crystallinity can be significantly affected by the experimental fabrication process.[28,29] Our ML models have a limitation in describing the morphological characteristic because the present GCN representation only contains the information limited in the monomer. Also, the GCN cannot account for the stereochemistry which requires information on the three-dimensional configurations of the atoms and the techniques to process the molecular graph.[48−50] For the data domains for which the factors like the morphology at the mesoscopic scales or the stereochemistry come into play, the GCN-based ML models should be extended to consider such three-dimensional information to learn the structure−property relationships.

To gain more insights into the GCN-based prediction models, we analyze how the GCN captures the structural features relevant to the property predictions. The high-dimensionality and nonlinearity of the GCN make a direct interpretation of the meaning of representations almost infeasible. To alleviate the problem, we need to associate the data points in the representation space with a proxy label that provides semantic information. As a representative structural feature, the fraction of rotatable bonds in the monomer is chosen to account for the atomic scale rigidity arising from the π bondings,[51] based on the fact that the fraction of rotatable bonds is related to the flexibility of the backbone chain which plays an important role in determining various physical properties.[23−26,30] Another important reason is that the rotatability of bonds is a structural feature defined solely from the information on the neighboring atoms within the monomer structure. The GCN might be able to classify them by aggregating the information of neighboring atoms, such as the types of atoms and the number of attached hydrogen atoms. We used RDKit[39] to obtain the quantities for all the polymers in the datasets.

We explore the characteristic of the GCN representation spaces with the T_g dataset because of the well-known correlation between the T_g and backbone rigidity. The effect of supervision of the GCN-NN models with the T_g dataset on the structural featurization is investigated by tracing the evolution of representation spaces. Figure 6a shows schematics of analysis of the GCN-NN model using the principal component analysis to examine the high-dimensional representation spaces with few hundred vector components (100 components for the GCN representation and 300 components for the latent representation through the NN layer). In Figure 6b−e, the distributions of the polymers are projected on the two-dimensional subspace spanned by the first two principal components PC1 and PC2, respectively. Figure 6b,c shows the distributions of experimental T_g values and the fraction of rotatable bonds, respectively, in the GCN representation space before the training with the T_g dataset. Figure 6d displays the distribution of the fraction of rotatable bonds in the trained GCN representation space followed by their distributions in the latent space (Figure 6e) of the NN layers in the trained GCN-NN model. Note that the latent space corresponds to the activations in the final layer of NN layers before the regression in the GCN-NN model as shown in Figure 6a.

Figures 6b,c shows GCN representations without any supervision, for which the weight parameters are randomly initialized.[38] The existence of gradients in the maps of both T_g and backbone rigidity indicates that polymers tend to be organized in their spaces. The organization is ascribed to the structure of GCN in which the graph convolution operation is designed to capture a great deal of molecular features to predict T_g prior to any learning. This is in analogy with the convolutional NN which shows good performances over various tasks on the image based on its architectural design even without training.[52] Their gradients in the maps of GCN representation spaces are anticorrelated to each other in Figure 6b,c. The organization of the data distribution in the representation space will facilitate the subsequent regression task to predict T_g. As we employ the GCN-NN models of which only the NN layers are allowed to be trained by fixing weight parameters for the GCN by the initial random values, the prediction performance for T_g is degraded compared to that of the fully trained GCN-NN model but still reliable to be RMSE ∼38 K and R^2 ∼ 0.83, respectively. The effect of training on the GCN layers becomes more significant for T_m, where the RMSE dramatically increases from ∼40 to ∼53 K without training of GCN layers. Although the distribution is not unique because of the randomness of the weight parameters of the GCN, the PC1 is consistently observed to account for the majority (∼85%) of the total variance. A noteworthy point is that the directions of the gradient in T_g and the fraction of rotatable bonds are not aligned to the PC1.

We show how the training processes affect the GCN layers in Figure 6c,d. As we train the GCN, the polymers are redistributed such that the directions of gradients the fraction of rotatable bonds rotate. Accordingly, the PC1 and PC2 account for ∼36 and ∼18% of the total variance of data, respectively. A notable thing is that the directions of the gradients tend to be aligned to the principal axis PC1. One can speculate that the change of distribution in the representation space would be coupled to the improvement of prediction performances of the GCN-NN model with the training of GCN layers (RMSE ∼ 30 K in Table 1), compared to that without the training of the GCN layer (RMSE ∼ 38 K). In the GCN-NN models, the GCN representations will be further processed through the subsequent NN layers resulting in the latent representations as shown in Figure 6e. It is also found that the directions of the gradients continue to be aligned to the principal axis PC1. Better performance of the GCN-NN models compared to the GCN-LR models implies that the latent space representation through the NN layer would have improved linearity with respect to $T_{g,}$ compared to the GCN representation. We note that the ECFP also shows similar behaviors in the organization through the ECFP space and their latent space, obtained with the ECFP-NN model (Figure S2). In particular, the gradients of T_g and rotatable bonds are aligned to the PC1 axis in the latent space representation through the NN layer.

The GCN is found to have an inherent capability of organizing the polymers by the structural rigidity which can be further improved through the training. To check the reliability of the organization over the unexplored region, we revealed an extensive region of the representation space by mapping 100,000 polymers provided by the PI1M dataset.[11] The PI1M database provides an extensive dataset of polymer structures without any property labels, obtained by a generative model trained with the PoLyInfo database over the various polymer

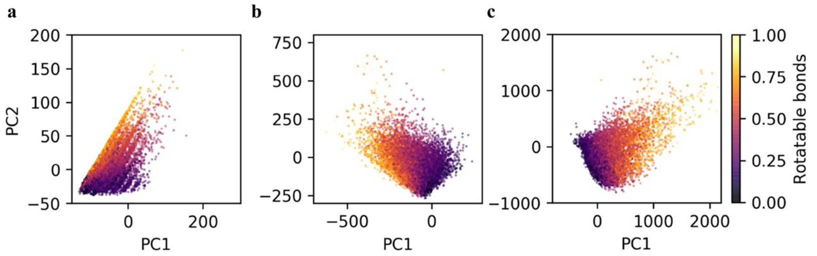


Figure 7. Distributions of the fraction of rotatable bonds of polymers in the PI1M dataset on the representation spaces obtained by the GCN-NN model for T_g prediction. Polymers are mapped by (a) GCN without training (random weight), (b) trained GCN, and (c) trained GCN-NN to get representations.

classes. Figure 7 shows the distributions of the fraction of rotatable bonds on the three representation spaces using the untrained GCN (Figure 7a), trained GCN (Figure 7b), and the latent space through the NN layers (Figure 7c), corresponding to those in Figure 6c−e. Note that the same principal axes are used for more comprehensive visualization. Although no structures overlap with our T_g dataset, the data distributions for the PI1M clearly show that the GCN indeed organizes the polymer by the backbone rigidity with gradients whose directions are shown to be similar to those in Figure 6c−e. This confirms that the GCN reliably captures the backbone rigidity with a generalization capability beyond the training data. Therefore, our GCN models might be transferred to predict other properties governed by the structural feature associated with a backbone rigidity.

To examine how different structural features other than the rotatable bonds manifest in the representation space, we consider four more features such as the fractions of sp^3 carbon atoms, aromatic rings, amide bonds, and hydrogen bonds in the monomer structures, where the number of hydrogen bonds corresponds to the sum of the numbers of hydrogen bond acceptors and donors (see Supporting Information Table S4 for details of the structural features). We calculated the Pearson correlation coefficients R between the PC1 and structural features on the representation spaces studied in Figures 6 and 7. The Pearson correlation coefficients for different features are shown in Figure 8a. The coefficient is largest for the rotatable bonds which increases from $R \sim -0.52$ with the GCN without training to $R \sim -0.87$ with the trained GCN and −0.90 through the NN layers in the GCN-NN model. The fraction of aromatic rings and sp^3 carbon atoms show relatively lower correlations as the $R \sim 0.76$ and ∼0.61 in the absolute values, respectively, for the trained models. Figure 8b,c shows correlation plots for the T_g and the rotatable bonds with the PC1 on the GCN representation space after the GCN training, where the PC1 shows strong correlations for both T_g ($R \sim 0.88$) and rotatable bonds ($R \sim -0.87$) (see the Supporting Information Figures S3−S5 for the correlation plots for all the structural features). The existence of a vector component in the representation space, linearly correlated with the given structural features, further supports that our GCN models extract a backbone rigidity relevant to the prediction of target property T_g.

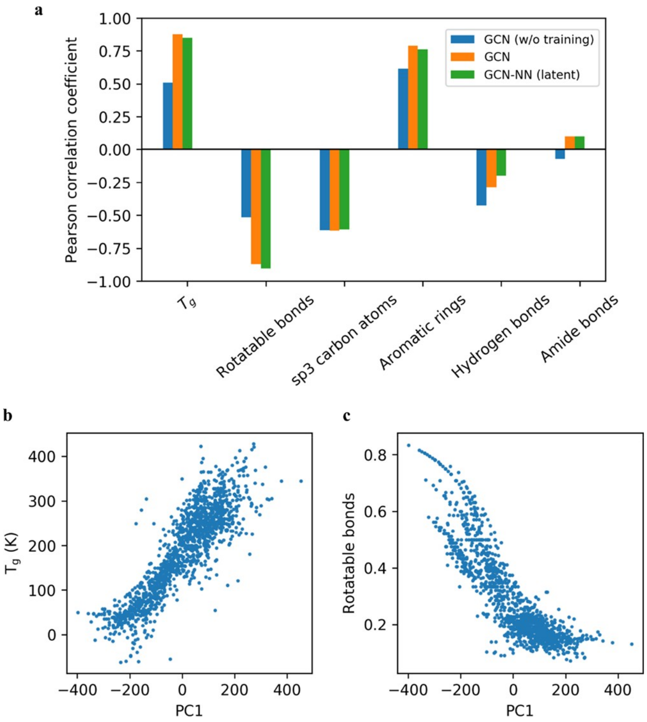


Figure 8. (a) Pearson correlation coefficients between PC1 and various structural features. Data for the GCN (w/o train), GCN (w/ train), and GCN-NN (w/ train) correspond to those in Figures S3−S5, respectively. Plots for the correlations of the PC1 with (b) T_g and (c) rotatable bonds using the trained GCN representation are shown.

Finally, we show the extent of the correlations between the thermomechanical properties and the fraction of rotatable bonds in our datasets in Figure 9. In Figure 9a, the T_g with relatively large data points most clearly shows the anti-correlation with the fraction of rotatable bonds among all the properties, where their correlation with backbone rigidity has been studied.[23−26,30] Therefore, training the GCN with the T_g dataset could induce the organization in the representation space to better capture the correlations for the prediction. The strength of the correlations becomes relatively lower for the T_m and ρ datasets and nearly diminishes for E datasets. The T_m is also known to increase with the backbone rigidity,[53] but systematic experimental studies on the effect of the backbone rigidity on ρ are limited. While the rigidity would frustrate the dense packing during cooling,[23] the aging of the glass will result in volume relaxations compensating the nonequilibrium effect.[54] Although the incorporation of conjugated rings into the backbone chain to enhance the rigidity is a generic strategy to increase the elastic moduli of polymer materials,[28,29] the diminished correlations in the E dataset are consistent with the fact that interchain interactions through the crystallization are also critical. As the correlation between the backbone rigidity and the target property decreases, the prediction performance of our GCN models is also reduced. It is feasible for the GCN model to automatically extract a backbone rigidity of polymers for high prediction performance of their properties.

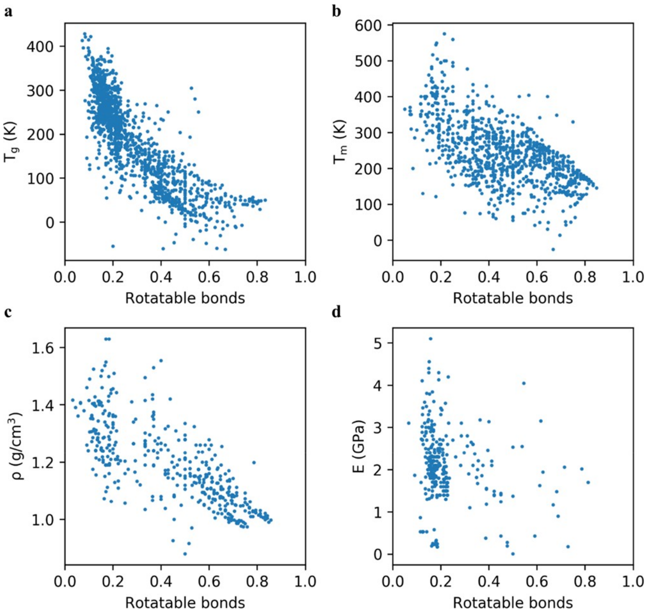


Figure 9. Correlations between the fraction of rotatable bonds in monomer structures and properties of (a) T_g, (b) T_m, (c) ρ, and (d) E in the PoLyInfo datasets.

4. CONCLUSIONS

We studied the ML models based on the GCN representations to predict thermal and mechanical properties of polymers such as T_g, T_m, ρ, and E. Performances of the GCN-based models were comparable to those of the ECFP-based models, and both the representations perform better with nonlinear NN regression rather than LR. The GCN-based models showed a high accuracy of predictions with the R^2 score ∼0.9 for the T_g data which exhibit strong anticorrelations with the backbone rigidity of the polymers. On the other hand, the prediction performance was significantly degraded to be R^2 score ∼0.5 for the E data without any noticeable correlations with the rigidity. Our results indicate the applicability of GCN representations, which extracts the bonding characteristics based on the information limited in the monomer structure, depending on the data domains. The GCN representation space exhibits a prominent organization of polymer data by the backbone rigidity. The organization of the data points in the representation space is optimized by training with T_g datasets for the prediction, which can be ascribed to their correlation, and is shown to be extended beyond the regions occupied by the polymers in the training set. This work provides an insight into how the GCN learns and predicts the polymer properties as well as a transferability to other properties associated with a backbone rigidity. Our ML models also have a limitation in describing the morphological characteristic because our polymer featurizations are conducted at the monomer level. For practical polymer design by ML, the featurizations for polymers should be directed toward reflecting the information beyond the monomer features and capturing the complex polymer properties. Improvements of the machine prediction accuracies should be accompanied by developing the database that provides retrieval of relevant information. In the future, it is also worthwhile to extend this study into developing transfer learning models to predict other properties with limited data.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acspolymersau.1c00050.

List of atom features used in the GCN, optimized hyperparameters for the prediction models, and structural features. Prediction performances of the ECFP-NN models and data distributions in ECFP representation spaces. Correlations of structural features with T_g on the principal axis of the GCN representation space (PDF)

AUTHOR INFORMATION

Corresponding Authors

Youngseon Shim − *Innovation Center, Samsung Electronics Co., Ltd., Hwaseong-si, Gyeonggi-do 18448, Korea;* orcid.org/0000-0002-5450-6328; Email: ys1231.shim@samsung.com

Changwook Jeong − *Innovation Center, Samsung Electronics Co., Ltd., Hwaseong-si, Gyeonggi-do 18448, Korea;* Email: chris.jeong@samsung.com

Authors

Jaehong Park − *Innovation Center, Samsung Electronics Co., Ltd., Hwaseong-si, Gyeonggi-do 18448, Korea*

Franklin Lee − *Science and Technology Division, Corning Incorporated, Corning, New York 14831, United States*

Aravind Rammohan − *Science and Technology Division, Corning Incorporated, Corning, New York 14831, United States*

Sushmit Goyal − *Science and Technology Division, Corning Incorporated, Corning, New York 14831, United States*

Munbo Shim − *Innovation Center, Samsung Electronics Co., Ltd., Hwaseong-si, Gyeonggi-do 18448, Korea*

Dae Sin Kim − *Innovation Center, Samsung Electronics Co., Ltd., Hwaseong-si, Gyeonggi-do 18448, Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/acspolymersau.1c00050

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

J.P. and Y.S. contributed equally to this work. Computing resources were provided by the Supercomputing Center of Samsung Electronics.

REFERENCES

(1) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5*, 83.

(2) Moosavi, S. M.; Jablonka, K. M.; Smit, B. The Role of Machine Learning in the Understanding and Design of Materials. *J. Am. Chem. Soc.* **2020**, *142*, 20273−20287.

(3) Gartner, T. E., III; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules* **2019**, *52*, 755−786.

(4) Audus, D. J.; de Pablo, J. J. Polymer Informatics: Opportunities and Challenges. *ACS Macro Lett.* **2017**, *6*, 1078−1082.

(5) Doan Tran, H.; Kim, C.; Chen, L.; Chandrasekaran, A.; Batra, R.; Venkatram, S.; Kamal, D.; Lightstone, J. P.; Gurnani, R.; Shetty, P.; Ramprasad, M.; Laws, J.; Shelton, M.; Ramprasad, R. Machine-

Learning Predictions of Polymer Properties with Polymer Genome. *J. Appl. Phys.* **2020**, *128*, 171104.

(6) Tao, L.; Chen, G.; Li, Y. Machine Learning Discovery of High-Temperature Polymers. *Patterns* **2021**, *2*, No. 100225.

(7) Chen, G.; Tao, L.; Li, Y. Predicting Polymers' Glass Transition Temperature by a Chemical Language Processing Model. *Polymer* **2021**, *13*, 1898.

(8) Ma, R.; Liu, Z.; Zhang, Q.; Liu, Z.; Luo, T. Evaluating Polymer Representations via Quantifying Structure−Property Relationships. *J. Chem. Inf. Model.* **2019**, *59*, 3110−3119.

(9) Sivaraman, G.; Jackson, N. E.; Sanchez-Lengeling, B.; Vázquez-Mayagoitia, Á.; Aspuru-Guzik, A.; Vishwanath, V.; de Pablo, J. J. A Machine Learning Workflow for Molecular Analysis: Application to Melting Points. *Mach. Learn. Sci. Technol.* **2020**, *1*, No. 025015.

(10) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717−1730.

(11) Ma, R.; Luo, T. PI1M: A Benchmark Database for Polymer Informatics. *J. Chem. Inf. Model.* **2020**, *60*, 4684−4690.

(12) Wu, S.; Kondo, Y.; Kakimoto, M.; Yang, B.; Yamada, H.; Kuwajima, I.; Lambard, G.; Hongo, K.; Xu, Y.; Shiomi, J.; Schick, C.; Morikawa, J.; Yoshida, R. Machine-Learning-Assisted Discovery of Polymers with High Thermal Conductivity Using a Molecular Design Algorithm. *npj Comput. Mater.* **2019**, *5*, 66.

(13) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521*, 436−444.

(14) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *Aust. J. Chem.* **2021**, *13*, 12.

(15) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(16) Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P. S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4−24.

(17) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. In *Advances in Neural Information Processing Systems*; Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; *28*.

(18) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.

(19) Hop, P.; Allgood, B.; Yu, J. Geometric Deep Learning Autonomously Learns Chemical Features That Outperform Those Engineered by Domain Experts. *Mol. Pharmaceutics* **2018**, *15*, 4371−4377.

(20) Korolev, V.; Mitrofanov, A.; Korotcov, A.; Tkachenko, V. Graph Convolutional Neural Networks as "General-Purpose" Property Predictors: The Universality and Limits of Applicability. *J. Chem. Inf. Model.* **2020**, *60*, 22−28.

(21) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(22) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9*, 513−530.

(23) Kumar, R.; Goswami, M.; Sumpter, B. G.; Novikov, V. N.; Sokolov, A. P. Effects of Backbone Rigidity on the Local Structure and Dynamics in Polymer Melts and Glasses. *Phys. Chem. Chem. Phys.* **2013**, *15*, 4604−4609.

(24) Kunal, K.; Robertson, C. G.; Pawlus, S.; Hahn, S. F.; Sokolov, A. P. Role of Chemical Structure in Fragility of Polymers: A Qualitative Picture. *Macromolecules* **2008**, *41*, 7232−7238.

(25) Ngai, K. L.; Roland, C. M. Chemical Structure and Intermolecular Cooperativity: Dielectric Relaxation Results. *Macromolecules* **1993**, *26*, 6824−6830.

(26) Xie, R.; Weisen, A. R.; Lee, Y.; Aplan, M. A.; Fenton, A. M.; Masucci, A. E.; Kempe, F.; Sommer, M.; Pester, C. W.; Colby, R. H.; Gomez, E. D. Glass Transition Temperature from the Chemical Structure of Conjugated Polymers. *Nat. Commun.* **2020**, *11*, 893.

(27) Tashiro, K. Molecular Theory of Mechanical Properties of Crystalline Polymers. *Prog. Polym. Sci.* **1993**, *18*, 377−435.

(28) Ahmed, D.; Hongpeng, Z.; Haijuan, K.; Jing, L.; Yu, M.; Muhuo, Y. Microstructural Developments of Poly (p-Phenylene Terephthalamide) Fibers during Heat Treatment Process: A Review. *Mater. Res.* **2014**, *17*, 1180−1200.

(29) Bigg, D. M. A Review of Techniques for Processing Ultra-High Modulus Polymers. *Polym. Eng. Sci.* **1976**, *16*, 725−734.

(30) Chantawansri, T. L.; Yeh, I.-C.; Hsieh, A. J. Investigating the Glass Transition Temperature at the Atom-Level in Select Model Polyamides: A Molecular Dynamics Study. *Polymer* **2015**, *81*, 50−61.

(31) Li, X.; He, Y.; Dong, X.; Ren, X.; Gao, H.; Hu, W. Effects of Hydrogen-Bonding Density on Polyamide Crystallization Kinetics. *Polymer* **2020**, *189*, No. 122165.

(32) Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138−52160.

(33) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 22071−22080.

(34) Otsuka, S.; Kuwajima, I.; Hosoya, J.; Xu, Y.; Yamazaki, M. PoLyInfo: Polymer Database for Polymeric Materials Design. In *2011 International Conference on Emerging Intelligent Data and Web Technologies*; IEEE, 2011; 22−29.

(35) Kipf, T. N.; Welling, M. *Semi-Supervised Classification with Graph Convolutional Networks. ArXiv:1609.02907* 2016.

(36) Li, Q.; Han, Z.; Wu, X.-M. Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning. *32nd AAAI Conf. Artif. Intell. AAAI 2018* 2018, 3538−3545.

(37) Balcilar, M.; Renton, G.; Héroux, P.; Gaüzere, B.; Adam, S.; Honeine, P. Analyzing the Expressive Power of Graph Neural Networks in a Spectral Perspective. *Proc. Int. Conf. Learn. Represent. ICLR* 2021, 1−26.

(38) He, K.; Zhang, X.; Ren, S.; Sun, J. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. ArXiv:1502.01852* 2015.

(39) RDKit: Open-source cheminformatics http://www.rdkit.org. (accessed 2021-12-11)

(40) Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer series in statistics New York, 2001; *1*.

(41) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.* 2014, 1−15.

(42) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.

(43) Chen, L.; Pilania, G.; Batra, R.; Huan, T. D.; Kim, C.; Kuenneth, C.; Ramprasad, R. Polymer Informatics: Current Status and Critical next Steps. *Mater. Sci. Eng. R Rep.* **2021**, *144*, No. 100595.

(44) Bishop, C. M. *Pattern Recognition and Machine Learning*; Springer: New York, 2006.

(45) Haibo He; Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263−1284.

(46) Krawczyk, B. Learning from Imbalanced Data: Open Challenges and Future Directions. *Prog. Artif. Intell.* **2016**, *5*, 221−232.

(47) Yang, Y.; Zha, K.; Chen, Y.-C.; Wang, H.; Katabi, D. Delving into Deep Imbalanced Regression ArXiv:2102.09554 Cs. 2021.

(48) Flam-Shepherd, D.; Wu, T.; Friederich, P.; Aspuru-Guzik, A. Neural Message Passing on High Order Paths. *ArXiv:2002.10413 Cs Stat* 2020. https://arxiv.org/abs/2002.10413. (accessed 2021-12-30).

(49) Pattanaik, L.; Ganea, O.-E.; Coley, I.; Jensen, K. F.; Green, W. H.; Coley, C. W. *Message Passing Networks for Molecules with Tetrahedral Chirality* ArXiv:2012.00094. 2020.

(50) Cho, H.; Choi, I. S. Enhanced Deep-Learning Prediction of Molecular Properties via Augmentation of Bond Topology. *ChemMedChem* **2019**, *14*, 1604−1609.

(51) Albright, T. A.; Burdett, J. K.; Whangbo, M.-H. *Orbital Interactions in Chemistry*; John Wiley & Sons, 2013.

(52) Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. *Int. J. Comput. Vis.* **2020**, *128*, 1867−1888.

(53) Balani, K.; Verma, V.; Agarwal, A.; Narayan, R. Physical, Thermal, and Mechanical Properties of Polymers. In *Biosurfaces*; John Wiley & Sons, Inc: Hoboken, NJ, USA, 2015; 329−344.

(54) Pan, P.; Zhu, B.; Inoue, Y. Enthalpy Relaxation and Embrittlement of Poly(l-Lactide) during Physical Aging. *Macromolecules* **2007**, *40*, 9664−9671.