Deep Learning Based Approach to Predict Glass Transition Temperature of Polymers

Saket Thavanani
Department of Materials Science and Engineering
University of Toronto

Mohamad Danish Anis
Department of Mechanical and Industrial Engineering
University of Toronto

***Abstract*—Glass transition is a significant relaxation occurring in amorphous polymers at a wide temperature range depending on kinematic properties like transition method and pressure. Predicting this temperature (Tg) provides valuable insights into material properties whose synthesis may otherwise be costly and time consuming. Over the years, AI tools have been increasingly employed to quantify compound properties with higher accuracy. This paper proposes two models that use SMILES code of the polymer to predict its Tg with average relative error of 6 % and 5 %. The authors train and compare two neural networks for their accuracy in predicting Tg. The first neural network model is implemented using convolutional neural networks. In this model, Binary image features are engineered using SMILES into machine readable forms that serves as an input to the CNN. The second model is based upon fully connected Neural Networks. However in this model, 28 new ingenious features were extricated using the SMILES representation. Both the models show excellent level of generalization with second model performing slightly better.**

***Index Terms*—Property prediction, glass transition temperature, neural networks, deep learning, convolution neural networks**

I. Introduction

It is difficult to overstate the importance of developing quantitative structural property relationship models in the scope of material science. A decade of research has highlighted a literature gap in the need for identifying unknown materials from their predicted macroscopic properties. As little calculation is required for identifying polymers by their chemical structure, there is a need for more focus on property-based identification practices. One such thermo mechanical property, according to [1], which also happens to be the second most cited keyword in the history of glass science is Glass Transition Temperature (Tg). It marks the temperature range below which the atoms of a super cooled liquid are temporarily frozen (without crystallizing) upon cooling. As a well-documented guiding parameter that has several times been correlated with a monomer's physical quantities, Tg has been chosen to validate this study's approach to analyze structural information. The focus of this paper is to predict Tg of unknown polymers by synthesizing its chemical structure using an AI approach.

The existing literature includes theoretical and empirical approaches based on density function theory (DFT), molecular dynamics, evolutionary algorithms and even topological constraint theories to providing solutions for identifying compositions from their macro- properties. However, an empirical model employing the use of AI machine learning, particularly Artificial Neural Networks (ANN) is more suited to identifying such structure- property relationships. As powerful tools capable of fast-paced development with memory suitable for highly complex problem solving, ANN have produced good results for problems requiring DFT calculations. ANN based approaches have shown results in polymer studies related to atomization energy [2], mechanical properties [3], rheometric properties [4] and glass transition temperatures [5] , among other studies on extracting structural information from X-ray diffraction data [6]. Variants of ANN like convolutional neural networks (CNN) were introduced in the recent past to revolutionize the field of visual imagery analysis [7]. Extensive research has been carried out for their applications in medical image recognition [8], demonstrative robots [9], self-driven cars [10] and machine translation [11].

ANNs have been deemed fit for this study of predicting Tg identifying polymers due to their ability to mine critical feature information with their uniquely robust architecture that that includes fully connected (FC) layers. A novelty being proposed in this study is the prediction of Tg using binary images of a polymer's chemical composition fed into the neural network architecture. This is so that the resulting model will only require an image of the unknown compound's composition to predict its Tg, surpassing the otherwise tedious mathematical calculations required for material classification. As a benchmark to be improved, well documented studies employing CNN in image recognition were reviewed. Although CNN possess hugely popular FC layers and coupled with their backpropagation learning based training, they can modify weights that reduce overfitting. However, this study proposes to improve the performance with feature engineering that allows better representation of intra-molecular interactions within the dataset.

CNN based deep learning approaches were proposed by [12], [13] earlier to predict Tg of known compounds, however, there were some limitations that the present study aims to overcome; (i) The previous studies failed to consider the molecular structure and atomic interactions while making Tg predictions, such as accounting for the relative amount of each atom in a glass compound. (ii) Another limitation identified was the complexity of encoded data to be fed into the neural network. While working with a fully connected CNN, the chemical structure of a monomer needs to be introduced into the hidden layers in machine readable form. In this paper, Simplified Molecular Input Line Entry System (SMILES) encodes the chemical structure of monomer into a numeric version, making it computationally a lot cheaper than the previous studies. (iii) Lastly, a major improvement has been in the overall AI comprehension for the problem statement. The chosen model converts all composition and structural information contained in the SMILES code into an easily readable binary image in order to predict the Tg of different polymers.

This study proposes a comparative analysis of two AI techniques on data taken from a database consisting of known novel unknown polymers for the purpose of validation to incorporate all inter-atomic interactions, chemical composition as well as the chemical structure. A detailed discussion of the project plan, methodology and description of data, along with the neural network architecture will follow in the upcoming sections.

II. Methodology

Along with a high level summary of the proposed model's implementation plan, this section details the following: the dataset used to train and test the neural network, the architectures for the two employed approaches, the feature engineering employed to make sure that the data is adequately machine readable and the tuned hyper-parameters.

Project Plan: As highlighted earlier, the existing literature on ANN based Tg prediction primarily relies on polymer structure and chemical composition. This study proposes a comparative analysis of benchmark image processing CNN a feature-engineered ANN. Building on the benchmark's inadequacies, the proposed ANN model would consider chemical composition, molecular structure and all kinds of intramolecular interactions that govern the movement of carbon chains as in the compound during glass transition. As a function of varying molecular weights, degree of branching and various other newly designed features, we aim to develop a new empirical relationship using ANNs. The hidden layer in the proposed ANN will be able to better highlight when the composition starts to experience a transition, as opposed to the complexities of benchmark model's convolutions. From an algorithm point of view, feature engineering will play an important role in this model, being reflective of how well the transition is learned.

Dataset: The dataset used in our study was gathered from a popular polymer database [14]. Similar databases have been explored and later used in their studies by [5], [15], [16]. The dataset for this study comprises of 351 polymers along with their smiles codes, molecular names as input attributes and glass transition temperatures as the output variable. Subsets of 300 polymers and their Tg values were used for training validating the dataset, whereas the rest 51 unseen polymers were used to test the results for both the models, the CNN and the proposed ANN. The dataset was manually explored using the pandas library in python and was classified into eight different classes of polymers- acrylates, styrenes, amides, alkenes, ether, amides, carbonates and others. Fig 1 and Fig 2 show the pie plot and the box plot corresponding to the mentioned 8 different classes of polymers.

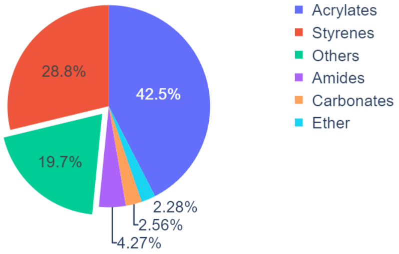


Fig. 1. Polymer Classes

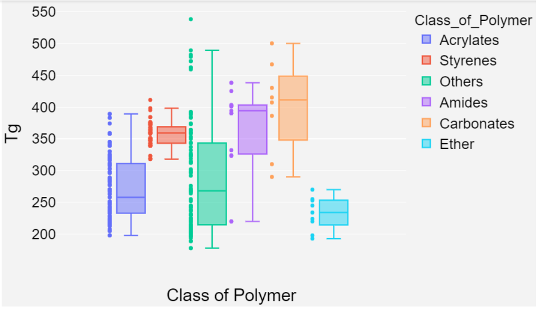


Fig. 2. Box Plot showing Polymer Classes

The pie plot in Figure 1 shows the exact composition of the dataset, with acrylates and styrenes being the highest contributors. The box plot in Figure 2 is used to show the scatter plot of the underlying Tg distributions for each class of polymers. It can be seen that styrenes tend to have higher Tg where as acrylates have a fairly mixed distribution.

Featurization in CNN: Engineering features has been vital tin preparing the data for modelling and presenting attributes in machine readable form. As per the problem statement, the Tg prediction was to be based on images of the polymer chemical structure fed into the CNN architecture in encoded form using SMILES line notations. SMILES consists of an inbuilt dictionary of characters for every chemical notation.

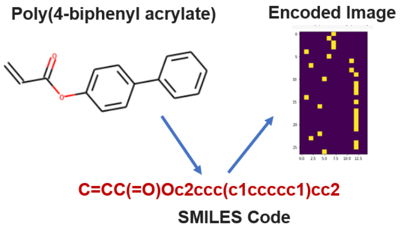


Fig. 3. Molecular Strutcure to Image Encoding

The linear strings thus generated are 1-hot encoded in machine readable form through binary images by employing this dictionary of SMILES characters. The resulting transformation is an *n-dimensional* matrix consisting of binary images that can be fed into the CNN architecture. Each binary image is a matrix of the dimensions m × n, where *n* represents the number of characters in the SMILES dictionary and *m* is the number of characters present in polymer with longest smiles code.

The convenience of generating such matrices as inputs to the CNN architecture also relies on the software environment being used. The open source *RDKit* [17] python package was used to visualize the molecular structure from dataset into drawings.

Featurization in Proposed ANN: Innovative Feature engineering is a crucial part of building any machine learning model. For the proposed ANN, 28 innovative features were extracted using every polymer’s SMILES code. In other words, the (351,3) dimensional dataset was converted to (351,31) using critical thinking domain expertise. Each polymer’s SMILES code was analyzed at character level to understand the coding terminology. For instance, (C) signifies that a single carbon atom is branched to the polymer in the monomer structure. The table given below shows the list of 28 novel features along with the feature extraction technique.

Features	Extraction
Number of C Atoms in Ring	count(’c’)
Number of C Atoms in Chain	count(’C’)
Number of O Atoms in Ring	count(’o’)
Number of O Atoms in Chain	count(’O’)
Number of Double Bonds in Ring	count(’c’)/2
Number of Double Bonds in Chain	count(’=’)
Number of N Atoms	count(’N’)
Number of F Atoms	count(’F’)
Number of Cl Atoms	count(’Cl’)
Number of Br Atoms	count(’Br’)
Number of S Atoms	count(’S’)
Molecular Weight of the Monomer	MW Function
Degree of Branching	count(’(’)+count(’)’)/2
Number of Single Bonds	len(Smiles)
Number of Double Bonds	len(Smiles)-count(’=’)
Number of Triple Bonds	count(’’)
Number of branched C Atoms	count(’(C)’)
Number of branched F Atoms	count(’(F)’)
Number of branched Cl Atoms	count(’(Cl)’)
Number of branched Br Atoms	count(’(Br)’)
Number of branched =O	count(’(=0)’)
Number of CC Single Bonds Chain	count(’CC’)
Number of CC Single Bonds Ring	count(’cc’)
Number of CC Double Bonds	count(’C=C’)
Number of Rings	count(’c1ccccc1’)
Benzene Ring Variation	count(’c1cc’)
Benzene Ring Second Variation	count(’1cc’)
Is Aromatic	if (’c’) Yes

TABLE I
INNOVATIVE FEATURES ALONG WITH EXTRACTION METHODOLOGY

Proposed Architecture: The structure of the proposed model that accepts image inputs has been widely discussed in the literature [18]–[20]. As a rule of thumb, a CNN needs to analyze the influence of nearby pixels by moving a filter across the binary image to learn the different features. This operation is meant to allocate the number of weights that the neural network must learn accordingly and give importance to relevant parts of an image (pixels). In this study however, a frame of 0’s (or zero padding) has been added to every image to ensure that the pixels weigh equally, and the ANN is trained as required. As explained in literature, ANN layers extract the hidden features from the data using a ReLu activation function and ADAM optimizer. Batch normalization is finally appointed for layers to re-adjust the weights and normalize the outputs, all the while the filters and window sizes are kept varying. In addition, by normalizing the output of the neurons, the activation function will only receive inputs close to zero, thus ensuring a non-vanishing gradient. The data tensor is flattened passed through a FC layer of 100 neurons. Overfitting is reduced by the dropout algorithm spreading weights equally through all neurons with a dropout algorithm being employed. Using the ReLu activation function, the FC layer is connected to a single neuron responsible for making the Tg prediction. The weighting is kept constant throughout the training process using the error between actual predicted values called loss function, mathematically represented as:

$$loss = \frac{100}{mx} \sum_{i=1}^{mx} \left(\frac{Actual(Tg) - Predicted(Tg)}{Actual(Tg)} \right) \quad (1)$$

Optimization: Employing an ADAM optimizer with learning rates 0.005, 0.01 and 0.05, there was convergence observed in mean absolute error (MAE) values for different epochs with four different batch sizes, 64, 100, 128 and 256. Different combinations of hyper-parameters were employed to find the best configuration of hyper-parameters for the both the proposed models.

III. Model Implementation

CNN: This model was implemented using the Keras library which serves as an Application Programming Interface (API) for Tensorflow. The choice of the final hyper-parameters has been made by incorporating various combinations of all the different hyper-parameters. The best observed configuration uses filter size of 64 with a window size of (5,5) in the first layer and size (3,3) with 32 filters in the second layer. This is followed by a max pooling layer with a window size of (3,3). Post the max pooling layer, we have three dense layers with 32, 10 and 1 neurons respectively, with the final dense layer being the output of our proposed ANN model. ReLu activation function was used by all layers with l2 regularization. The model achieved its best generalization by training upto 180 epochs with a batch size of 64 and learning rate of 0.03. A validation split of 0.1 and drop out probability of 0.1 was used in training the network to perform cross validation. Fig 4 and Fig 5 show the experimental and the predicted values of the glass transition temperatures for the training and the testing sets.

ANN: Using an ANN to predict Tg, 28 newly designed features were used as inputs to this model. The best configuration for this neural network uses three hidden layers of 30, 10, 30 neurons. The final layer of 1 neuron is the output of the neural network. This model was trained for 180 epochs

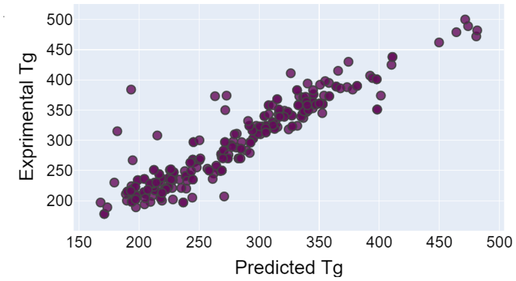


Fig. 4. Real vs Predicted for Training set for CNN

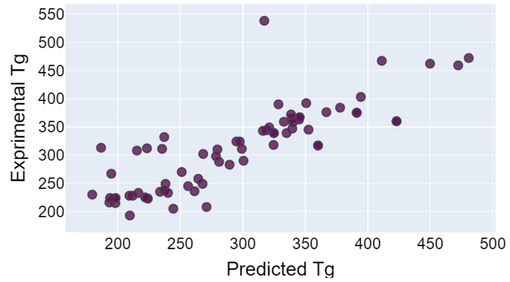


Fig. 5. Real vs Predicted for Test set for CNN

with a batch size of 100 and learning rate of 0.05. The cross validation was employed with a validation split of 0.1 on the training set. Figures 6 and 7 show the experimental and the predicted Tg values for the training and the testing sets. A relative training error of approximately **5%** was observed on both training and unseen testing sets. This minute difference in the error between test and train sets promises an excellent level of generalization.

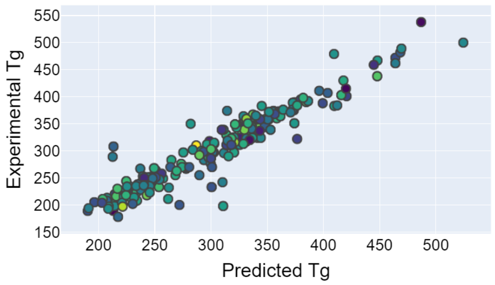


Fig. 6. Real vs Predicted for Training set for ANN

From Figure 6 and Figure 7, it can be perceived that most of the examples show very accurate prediction when compared to the real Tg values. However, there are a few polymers contributing to a significant level on uncertainty in prediction due to their lack of representation in the training set. These polymers belong to the minority classes of esters and ethers and due to insufficient training for either, the Tg of these polymers is not being learned effectively.

A. Conclusions

In this paper, the feasibility of using both CNN and a fully connected ANN has been demonstrated. CNN is able to predict Tg with mean relative error of 6% for both training and test sets. Fully connected ANN model attains ground breaking results with mean relative error of 5% along with an excellent level of generalization for both train and test sets. This is clearly evident in the Tg predictions obtained with CNN ANN models for 4 unknown polymers in Figures 8 9. ANN produces results a lot more closer to the original values as compared to the benchmark CNN model. As stated previously, a more interesting contribution of the proposed model is that it relies only on the chemical structure of the monomer unit of the corresponding polymer. Hence, this research can become a beneficial tool for materials scientists and polymer industries to discover new polymeric structures at a faster pace.

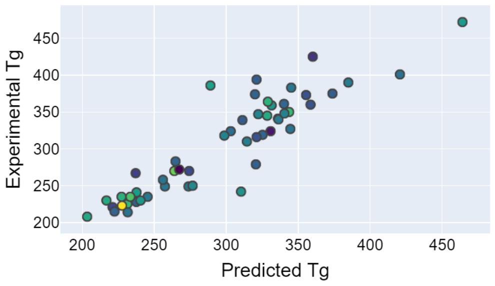
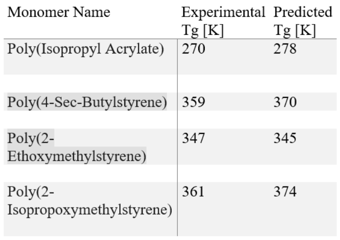


Fig. 7. Real vs Predicted for Test set for ANN

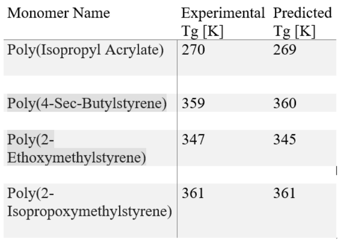


Monomer Name	Experimental Tg [K]	Predicted Tg [K]
Poly(Isopropyl Acrylate)	270	278
Poly(4-Sec-Butylstyrene)	359	370
Poly(2-Ethoxymethylstyrene)	347	345
Poly(2-Isopropoxymethylstyrene)	361	374

Fig. 8. CNN Prediction on Unseen test set

B. Acknowledgement

We would like to thank Professor C.V. Singh and Gurjot Dhaliwal for giving us great guidance throughout the project and course. We were able dig deeper into the algorithms and were able to make this project more meaningful.



Monomer Name	Experimental Tg [K]	Predicted Tg [K]
Poly(Isopropyl Acrylate)	270	269
Poly(4-Sec-Butylstyrene)	359	360
Poly(2-Ethoxymethylstyrene)	347	345
Poly(2-Isopropoxymethylstyrene)	361	361

Fig. 9. Fully Connected Neural Network Prediction on Unseen Test set

References

[1] E. Zanotto and J. Mauro, “The glassy state of matter: its definition and ultimate fate,” *Journal of Non-Crystalline Solids*, vol. 471, 2017.

[2] Z. Zhang and K. Friedrich, “Artificial neural networks applied to polymer composites: a review,” *Composites Science and technology*, 2003.

[3] N. Steiner, D. Hissel, P. Moçoteguy, and D. Candusso, “Diagnosis of polymer electrolyte fuel cells failure modes (flooding drying out) by neural networks modeling,” *International journal of hydrogen energy*, vol. 36, pp. 3067–3075, 2011.

[4] G. Schwartz, “Prediction of rheometric properties of compounds by using artificial neural networks,” *Rubber chemistry and technology*, vol. 74, no. 1, 2001.

[5] W. Liu and C. Cao, “Artificial neural network prediction of glass transition temperature of polymers,” *Colloid and Polymer Science*, vol. 284, no. 7, pp. 811–818, 2009.

[6] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[7] S. Lo, H. Chan, J. Lin, H. Li, M. Freedman, and S. Mun, “Artificial convolution neural network for medical image pattern recognition,” *Neural networks*, vol. 8, pp. 1201–1214, 1995.

[8] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” *Advances in neural information processing systems*, pp. 3844–3852, 2016.

[9] T. Li, P. Kuo, T. Tsai, and L. PC, “Cnn and lstm based facial expression analysis model for a humanoid robot,” *IEEE Access*, vol. 7, pp. 93998–94011, 2019.

[10] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Mobilenets, “Efficient convolutional neural networks for mobile vision applications,” *Efficient convolutional neural networks for mobile vision applications*, vol. arXiv preprint arXiv, p. 1704.04861, 2017.

[11] Z. Wu, S. Pan, G. Chen, F.and Long, C. Zhang, and S. Philip, “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[12] D. Cassar, A. de Carvalho, and E. Zanotto, “Predicting glass transition temperatures using neural networks. acta materialia,” *Acta Materialia*, vol. 159, pp. 249–256, 2018.

[13] S. Joyce, D. Osguthorpe, J. Padgett, and G. Price, “Neural network prediction of glass-transition temperatures from monomer structure,” *Journal of the Chemical Society, Faraday Transactions*, vol. 91, no. 16.

[14] CROW, “Polymer properties database,” *https://polymerdatabase.com/home.html*, 2019.

[15] G. Nadeau, H. Santelli, and Z. Dong, “Construction of a post-injury medicated ankle brace,” 2020.

[16] G. Wypych, *Handbook of polymers*. Elsevier, 2016.

[17] G. Landrum *et al.*, “Rdkit: cheminformatics and machine learning software,” *RDKIT. ORG*, 2013.

[18] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 1097–1105, Curran Associates, Inc., 2012.

[19] M. D. Zeiler and R. Fergus, “Computer vision–eccv 2014: 13th european conference, zurich, switzerland, september 6-12, 2014,” *Proceedings, Part I*, vol. 818, p. 833, 2014.

[20] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.