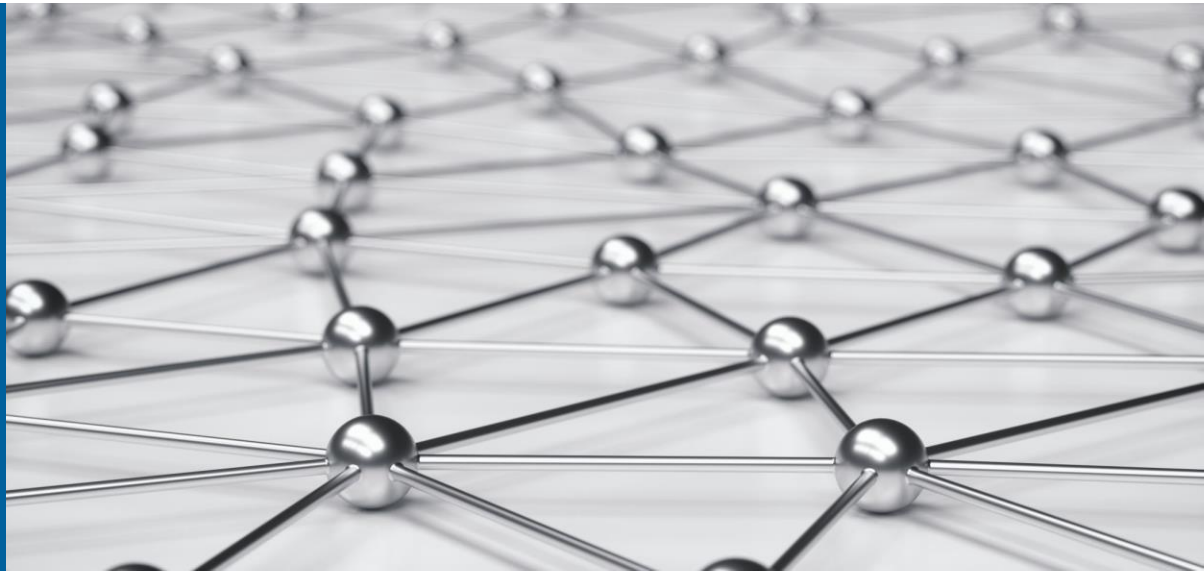
ABSTRACT

PREDICTING POLYMER PROPERTIES VIA GRAPH CONVOLUTIONAL NETWORK

Graph representations of molecules have been used to accurately predict the properties of polymers, in some cases with more success than typical feature-list molecular representations. The purpose of this project is to recreate the paper published by Park et al, Prediction and Interpretation of Polymer Properties Using the Graph Convolutional Network, in order to affirm that graph representations of molecules can be used to predict polymer properties. By means of recreating the Park study, the relationship between thermal and mechanical properties and structural characteristics of monomer units is explored by implementing a graph convolutional network (GCN) to model and predict the glass transition temperature, melting temperature, and density for polymers curated from the PolyInfo open-access database. This project focuses on building a comparable graph convolutional network machine learning model that can reproduce the accuracy of the Park model and extended to applications beyond the scope of the Park study. Additionally, improvements for Park's machine learning model are proposed in this project.

3 AUGUST 2022

PREDICTING POLYMER PROPERTIES VIA GRAPH CONVOLUTIONAL NETWORK (GCN)



CARLOS MATHERSON
SULI Intern, EVS Division
Email: camather@asu.edu

U.S. DEPARTMENT OF ENERGY
Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Virtual Presentation

BACKGROUND

Predicting Polymer Properties with GCN

- Graph representations can be used to predict polymer properties
- Bench-mark studies comparing polymer representations are limited
- Comparisons of typical feature-list and graph representations are inconclusive with graphs often being the inferior data representation
  - – the application of GCN models to polymers is elusive
- Park et al (2022) shows that GCN models can perform better than ECFP models

PURPOSE

Predicting Polymer Properties with GCN

- To recreate the results from the Park study using different, more available data

- Explore the relationship between the properties and structure of monomer units using a graph convolutional network (GCN) to predict the polymer's
  - glass transition temperature (Tg),
  - melting temperature (Tm),
  - and density (D)

- To build a foundation for a future graph convolutional model that can be used to predict thermal and mechanical properties for a wide domain of polymers

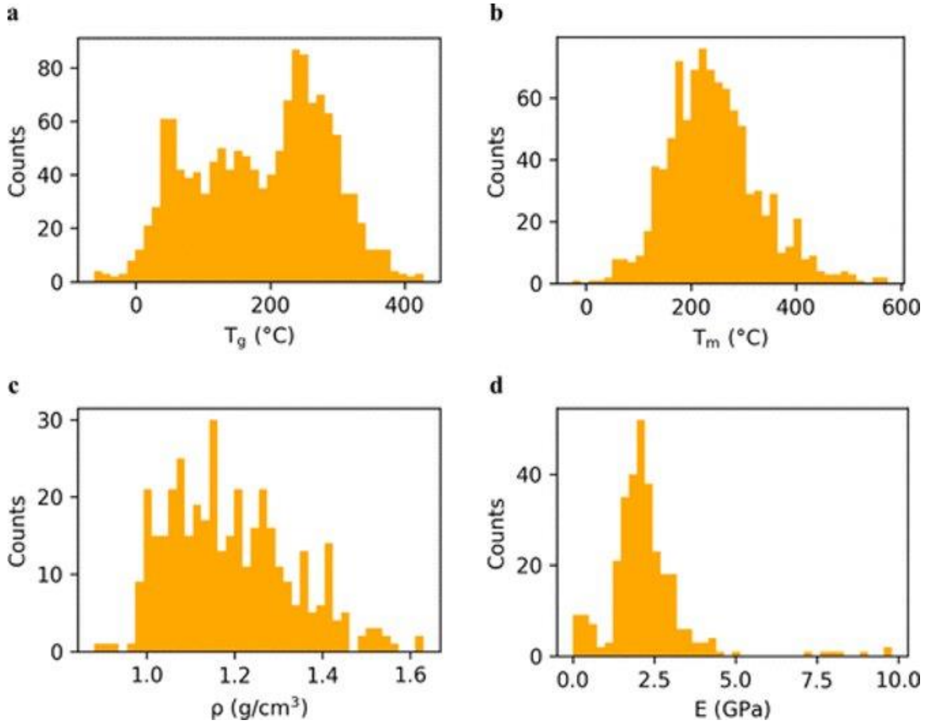
METHODS

Data Collection

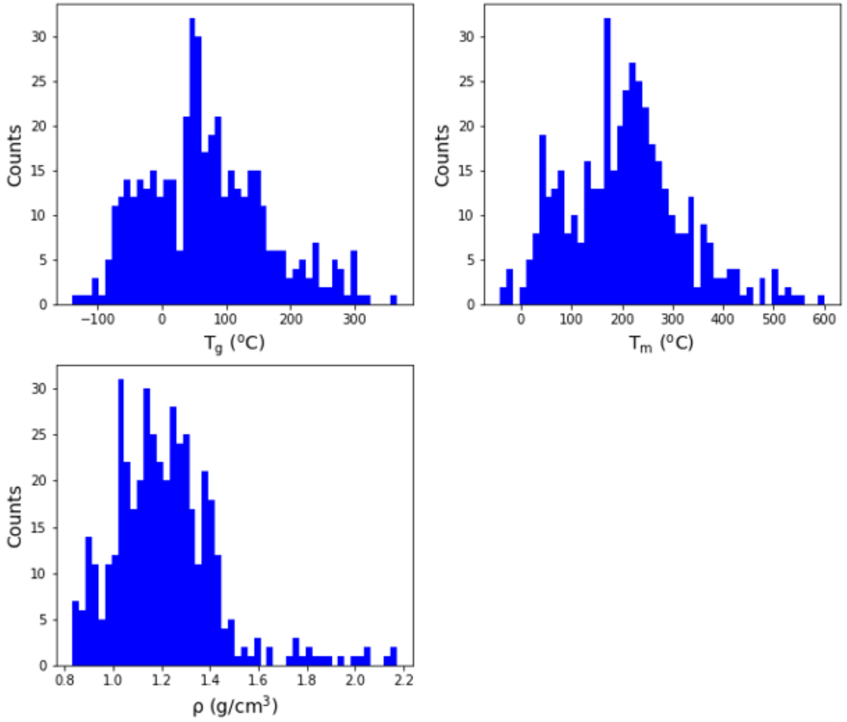
- Datasets used in Park study are unavailable for study recreation
- Substituted Park data with data from Liu et al (2022)
- Both datasets were gathered from PolyInfo open-access database
- Filtered Liu dataset, curated set of 445 compounds with SMILES and all three properties
  - Park study datasets contained 1388, 942, 390 compounds for Tg, Tm, D

METHODS

Comparing Datasets



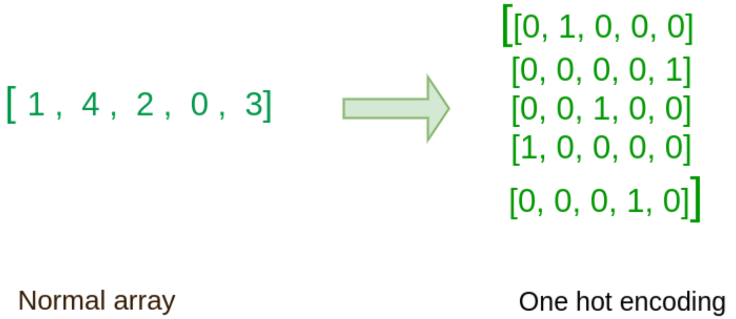
- Distribution of data used in Park study
- From PolyInfo Database



- Distribution of data used in this project
- Used in Liu et al (2022), From PolyInfo

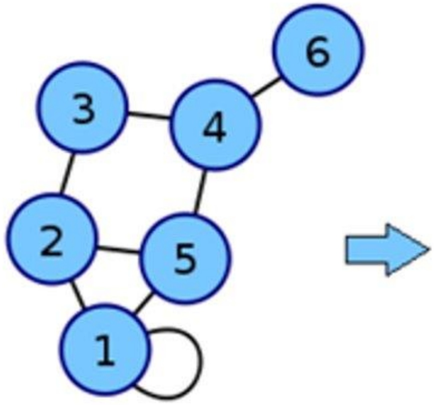
METHODS

Featurization



Normal array

One hot encoding



$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

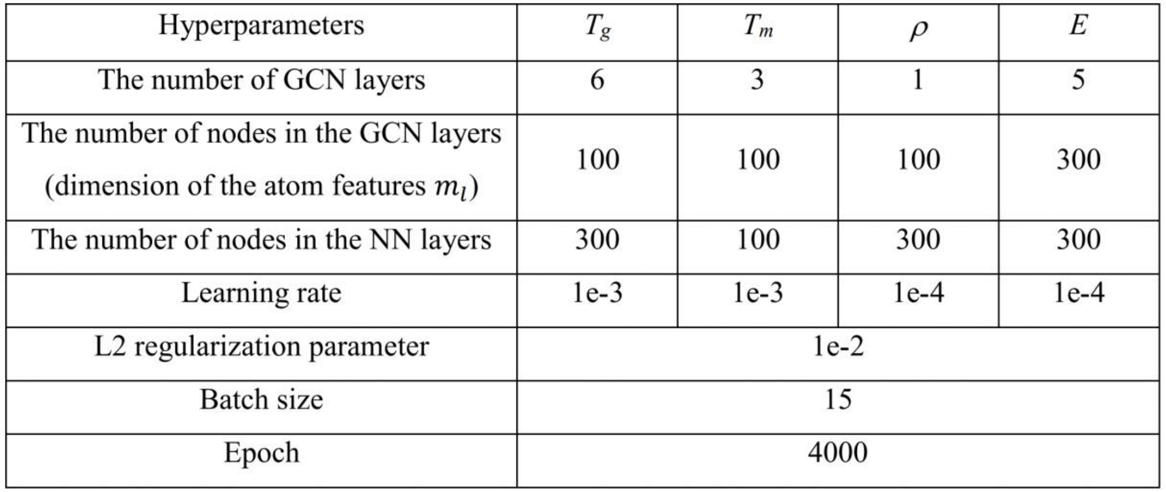
- 5 input node features were extracted from monomer unit SMILES using RDKit and one-hot encoding
  - Atom Type
  - Number of H atoms
  - Implicit Valence
  - Degree
  - Aromaticity
- Adjacency matrices used to represent bonds of molecule

Features	Description	Set	Length	RDKit method
Atom type	Element symbols	{Br, C, Cl, F, Li, N, O, P, S, Si, *}	11	*GetSymbol*
The number of H atoms	The number of H's on the atom	{0, 1, 2, 3}	4	*GetTotalNumHs*
Implicit Valence	The number of implicit H's on the atom	{0, 1, 2, 3}	4	*GetImplicitValence*
Degree	The number of directly-bonded neighbors	{1, 2, 3, 4}	4	*GetDegree*
Aromaticity	Whether the atom consists in an aromatic ring	{True, False}	2	*GetIsAromatic*

METHODS

Model Hyperparameters

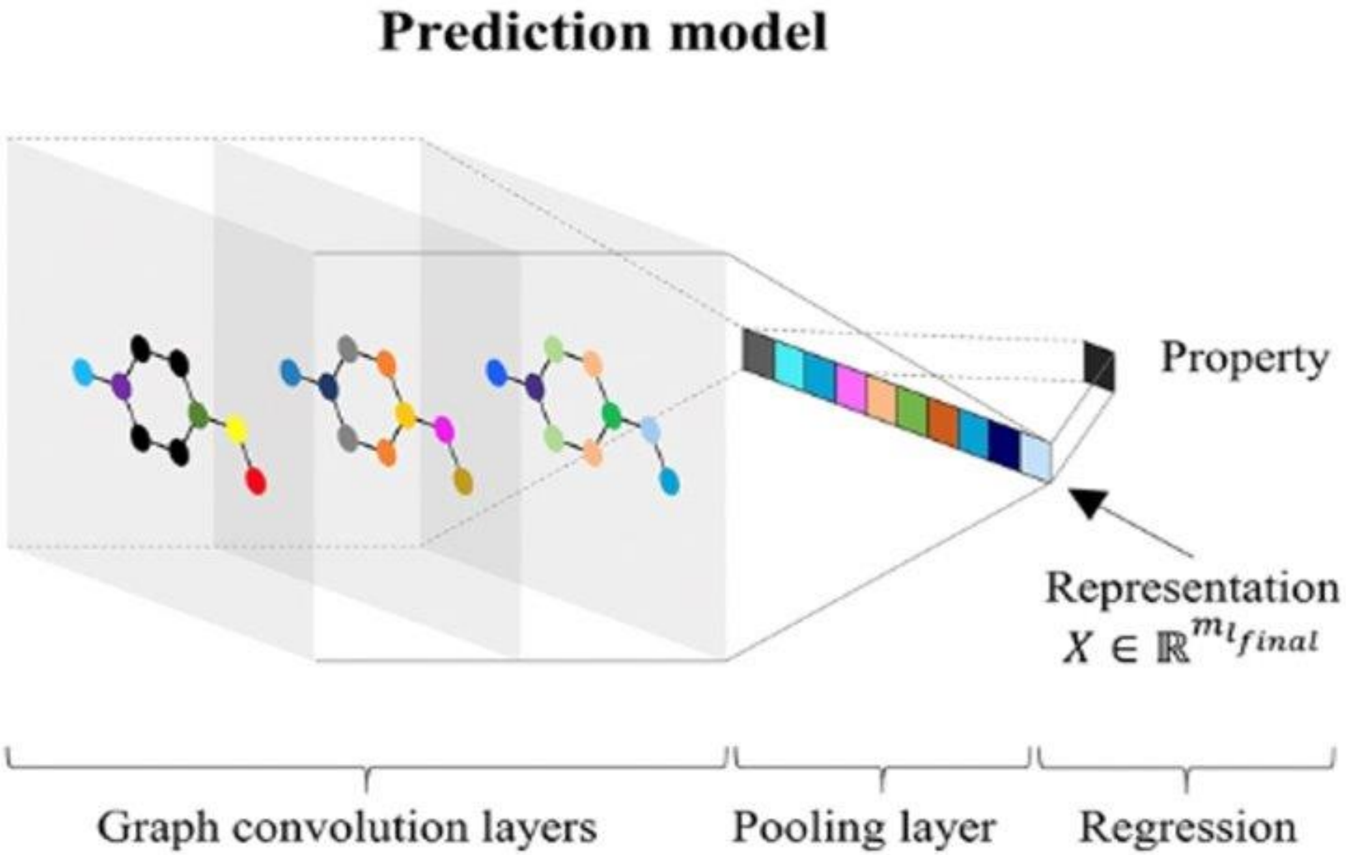
- Recreated model using tensorflow instead of pytorch
- Each model (3 total, 1 per property) follows set-up described in Park et al.
- Dual model: GCN with neural network (NN) and GCN with 1 dense output (SD)



Hyperparameters	T_g	T_m	ρ	E
The number of GCN layers	6	3	1	5
The number of nodes in the GCN layers (dimension of the atom features m_l)	100	100	100	300
The number of nodes in the NN layers	300	100	300	300
Learning rate	1e-3	1e-3	1e-4	1e-4
L2 regularization parameter	1e-2			
Batch size	15			
Epoch	4000			

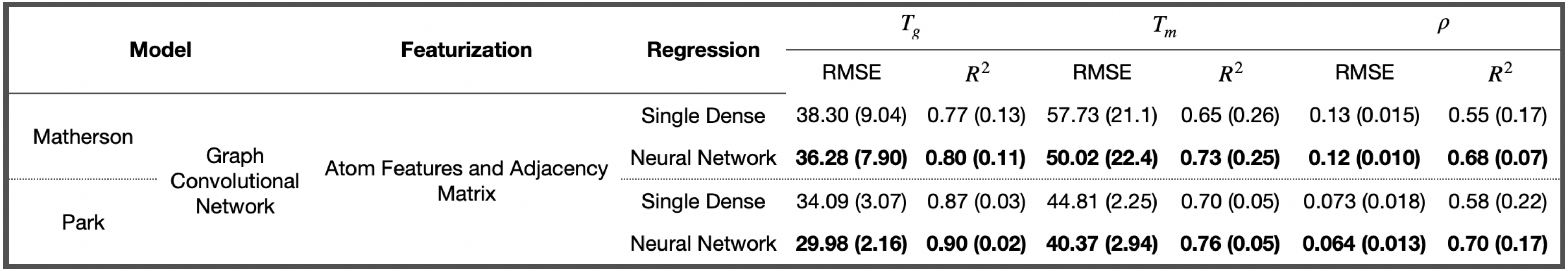
METHODS

Architecture



RESULTS

Model Performance

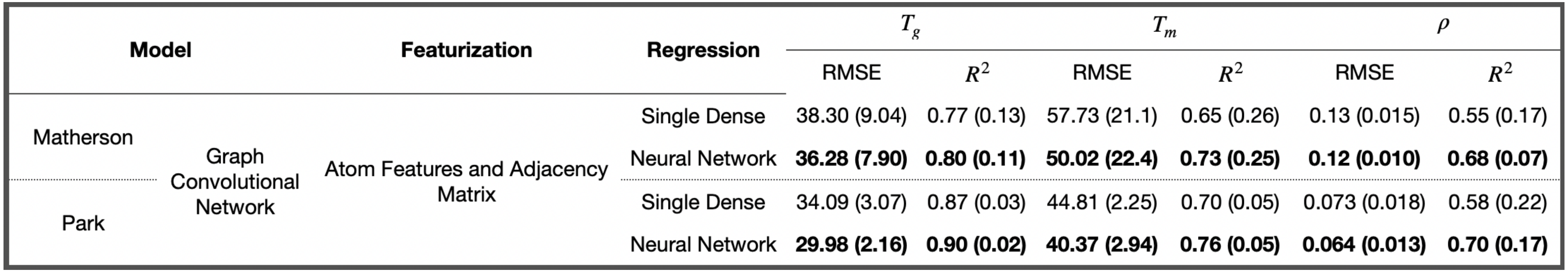


Model		Featurization	Regression	T_g		T_m		ρ	
				RMSE	R^2	RMSE	R^2	RMSE	R^2
Matherson	Graph Convolutional Network	Atom Features and Adjacency Matrix	Single Dense	38.30 (9.04)	0.77 (0.13)	57.73 (21.1)	0.65 (0.26)	0.13 (0.015)	0.55 (0.17)
			Neural Network	**36.28 (7.90)**	**0.80 (0.11)**	**50.02 (22.4)**	**0.73 (0.25)**	**0.12 (0.010)**	**0.68 (0.07)**
Park			Single Dense	34.09 (3.07)	0.87 (0.03)	44.81 (2.25)	0.70 (0.05)	0.073 (0.018)	0.58 (0.22)
			Neural Network	**29.98 (2.16)**	**0.90 (0.02)**	**40.37 (2.94)**	**0.76 (0.05)**	**0.064 (0.013)**	**0.70 (0.17)**

- Park model recreation positioned above original Park model
- The values are mean and (standard deviation) over 5-fold CV splits
- The bold case indicates the best performance per model for each property
- The units of RMSE are K for temperatures and g/cc for density.

RESULTS

Model Performance



Model		Featurization	Regression	T_g		T_m		ρ	
				RMSE	R^2	RMSE	R^2	RMSE	R^2
Matherson	Graph Convolutional Network	Atom Features and Adjacency Matrix	Single Dense	38.30 (9.04)	0.77 (0.13)	57.73 (21.1)	0.65 (0.26)	0.13 (0.015)	0.55 (0.17)
			Neural Network	**36.28 (7.90)**	**0.80 (0.11)**	**50.02 (22.4)**	**0.73 (0.25)**	**0.12 (0.010)**	**0.68 (0.07)**
Park			Single Dense	34.09 (3.07)	0.87 (0.03)	44.81 (2.25)	0.70 (0.05)	0.073 (0.018)	0.58 (0.22)
			Neural Network	**29.98 (2.16)**	**0.90 (0.02)**	**40.37 (2.94)**	**0.76 (0.05)**	**0.064 (0.013)**	**0.70 (0.17)**

- Trends seen in both models:
  - GCN+NN improves from GCN+SD in every case
  - Results show strongest correlation between Tg and structural features

- Park model performed better on average for each property
- Discrepancy in model performance is likely due to difference in dataset

RESULTS

Outcomes

- Results show correlation between thermal and mechanical properties and structural information of monomer unit
- Results fairly similar to results seen in Park study
- Model can be improved in future work

FUTURE WORK

Next Steps

- Optimize model
  - Model is not tuned for the training set
  - Tune model by tuning hyperparameters not tuned in Park Study
- Train model on larger dataset
  - Dataset used in project is small compared to data used in Park study
- Build 1 all-encompassing model
  - Individual models have marginal utility
  - A model that can used to predict all endpoints is more robust

REFERENCES

- Park, Jaehong, et al. “Prediction and Interpretation of Polymer Properties Using the Graph Convolutional Network.” *ACS Polymers Au*, Jan. 2022. *ACS Publications*, https://doi.org/10.1021/acspolymersau.1c00050.

- Gang Liu, Tong Zhao, Jiaxin Xu, Tengfei Luo, Meng Jiang. 2022. Graph Rationalization with Environment-based Augmentations. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD ’22), August 14–18, 2022, Washington, DC, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539347

QUESTIONS

PREDICTING POLYMER PROPERTIES VIA GRAPH CONVOLUTIONAL NETWORK

U.S. DEPARTMENT OF ENERGY

Argonne National Laboratory is a
U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC.

Argonne NATIONAL LABORATORY