Predicting Ionic Conductivity of Polymer Electrolytes using Molecular Descriptors and Machine Learning

Young-Cheol Cho, Eric Giavedoni, Siri Phuangthong

ChBE 413 Final Project

1 Abstract

Polymer electrolytes are a class of materials widely studied for applications in energy storage (e.g. batteries and fuel cells) due to their ability to conduct ions while maintaining mechanical stability. Predicting their ionic conductivity, however, remains a significant challenge, as it depends on a complex interplay of polymer structure, ion–polymer interactions, and environmental conditions. To alleviate experimental efforts, molecular dynamics (MD) simulations have been employed to model polymer dynamics and estimate properties important to energy storage applications, but they are often computationally expensive and limited in their ability to explore large chemical spaces efficiently. To address these limitations, we aim to leverage molecular descriptors to capture relevant chemical and structural information for predicting the efficacy of these polymer ectrolytes, thereby enhancing predictive capabilities without resorting to MD simulations.

In this project, our goal is to optimize ionic conductivity and diffusivity predictions by enriching dataset features using RDKit-generated descriptors and polymer physics-based descriptors. We will also create machine learning models used for ionic conductivity prediction from previous works and iterate upon them via novel featurization and hyperparameter tuning to create a more robust model. Ultimately, we hope to determine not only which polymer electrolyte features are important for these predictions, but also which model architectures would be most suitable for property predictions that produce similar if not improved results compared to previous works.

2 Background

Polymer electrolytes are a class of polymers that either possess ionizable groups along their backbones (polyelectrolytes) or achieve ionic conduction through complexation with salts (polymer–salt complexes) [1]. These materials have garnered extensive research attention due to their potential in energy storage applications, particularly in batteries [2]. Key performance metrics for polymer electrolytes include ionic conductivity [3] (σ), cation transference number [4] (t_{cat^+}), and glass transition temperature [5] (T_g), all of which are critical for optimizing battery efficiency and reliability .

Traditional experimental approaches for developing polymer electrolytes rely heavily on trial-and-error synthesis, which is often costly and time-consuming. To address these challenges, computational simulations such as molecular dynamics (MD) have been employed to predict the transport properties and structural behaviors of polymer electrolytes prior to experimental validation. For instance, Borodin and Smith [6] used MD simulations to investigate lithium ion transport along poly(ethylene oxide) chains, Zhang et al. [4] predicted t_{cat^+} for lithium ions using MD-derived Onsager coefficients, and Allam and Jang [7] employed all-atom MD to study lithium solvation structures, diffusion, and phase morphology in polymer matrices.

While MD is highly effective for predicting critical polymer electrolyte properties, it is computationally expensive and scales poorly with the size of the explored chemical space. For example, Toyota Research Institute's database of amorphous polymer electrolytes encompasses 6,286 MD trajectories totaling 5.7 terabytes of data. This data was used to compute ionic diffusivities, cation transference numbers, and ionic conductivities [8]. Although these simulations provide a rich dataset, the predictive performance for ionic conductivity (σ) remains limited, highlighting the need for more efficient modeling approaches.

In this study, we aim to leverage the Toyota Research Institute (TRI) dataset by integrating molecular descriptors with polymer physics-informed features. Specifically, we employ RDKit, an open-source cheminformatics toolkit, to calculate a broad range of molecular descriptors, and where necessary, derive polymer physics-based parameters (e.g., effective radius, diffusion proxies) from RDKit outputs [9]. Descriptor selection is informed by literature and chemical intuition, ensuring relevance to ionic transport. Feature importance analyses are performed using Random Forest regression to identify descriptors most strongly correlated with σ. To further enhance predictive capability, we apply multiple machine learning models, including Random Forest Regressors (RF) and Graph Neural Networks (GNNs), chosen to enable direct comparison with prior work. Through this framework, we aim not only to uncover the most influential polymer features governing ionic conductivity but also to optimize machine learning strategies for materials property prediction, offering a scalable alternative to computationally intensive MD simulations.

3 Method

Toyota Research Institute Database [8] previously generated 6270 MD trajectories for simulations of polymer electrolytes at 353 K. Because 133 samples were simulated for 50 ns as opposed to the other 6137 samples that were simulated at 50 ns, they were removed for data consistency. This database contains values of σ and t_{cat}, molality, structural properties (monomer molecular weight, degree of polymerization, and density), as well as diffusion coefficients calculated from MD simulations that are likely used as intermediate parameters for the electrochemical properties.

To prepare the dataset, we first checked for the possibility of duplicated rows, missing values, and the legitimacy of the outliers. We then proceed to removing $[Cu]$ and $[Au]$ used as labels for polymerization points in the SMILES strings, and validate each SMILES string with RDkit. Because all the outliers can be justified through literature search and chemical understanding, and because all the SMILES strings were successfully validated, all 6137 molecules were processed as RDkit's Mol list.

3.1 Molecular Featurization

Molecular descriptors chosen have been selected or produced due to literature search, polymer physics knowledge, or chemical intuition. The full list is presented in table 1 below:

Molecular Descriptors
Heteroatoms
Chain-Level Fragment: Topological diameter
Block-Level Fragment: Unique block fragments, Heavy atom block fragment
Atomic-Level Fragment Functional groups, rings, hybridizations, H-bond
Electrotopological State (EState)
Partial Charges through Gasteiger computation
Solvent Accessible Surface Area (SASA)/van der Waals surface area
Hydrophobicity (MolLogP)
Topological Polar Surface Area (TPSA)
Number of Rotatable Bonds
Carbon Hybridization Fraction

Table 1: Molecular Descriptors either obtained from or calculated using RDkit

We first analyze all elements present in the SMILES strings, excluding $[Cu]$ and $[Au]$, and then focus on the number of heteroatoms within each monomer. Heteroatoms, particularly those with lone pairs of electrons, play a crucial role in facilitating ion transfer between sites [10]. Among the elements present ('P', 'Cl', 'O', 'Si', 'C', 'F', 'N', 'S'), fluorine, nitrogen, and oxygen are of particular interest due to their high electronegativity and availability of lone pairs. To capture structural and functional information, we consider fragments at the chain, block, and atomic levels, which can be obtained using RDKit [11]. Chain-level fragments are characterized by the topological diameter, representing the maximum distance between two atoms along the chain [11]. Block-level fragments was performed using BRICS method [12] to include unique block structures

and the largest heavy-atom blocks, reflecting molecular complexity that may influence ion transport [11;13]. Atomic-level fragments encompass functional groups, ring counts, hybridization, and hydrogen-bond donors and acceptors [13]. These atomic features are particularly relevant: rigid ring structures can hinder ion mobility, whereas hydrogen-bonding sites can promote ion transfer, providing a mechanistic rationale for including these descriptors in our model.

While RDKit does not directly compute electronic properties critical to electron transfer, such as HOMO and LUMO energies, it provides descriptors that capture atomic charges and electrochemical characteristics relevant to ion transport. The Electrotopological State (EState) encodes the structural and electronic environment of each atom [14], reflecting how "electronically accessible" an atom is within a molecule. Although EState has primarily been applied in fields such as toxicity prediction [15], it can also inform ionic conductivity, as regions of high electronic accessibility may facilitate interactions with mobile ions. In our work, we capture this through the sum of EState values, the most and least reactive atoms (EState_max and `EState_min`), and the average and variation across the molecule (`EState_mean` and `EState_std`). In addition, we include partial atomic charges computed via the Gasteiger method [16], which provides an efficient approximation of electron distribution based on electronegativity equalization. These charges reflect the local electrostatic environment, where regions of high polarity or charge separation can promote ion dissociation and transport. Together, EState and Gasteiger-derived charges allow our descriptors to capture key electronic and electrostatic factors that influence ion mobility, providing valuable information for predicting ionic conductivity in polymer electrolytes.

Solvent Accessible Surface Area (SASA), or representation of van der Waals surface area, was also included to approximate the part of a molecule that is exposed to the surrounding environment [17]. This feature is particularly relevant for ionic conductivity as it reflects both the solvation characteristics and the accessibility of polymer segments to mobile ions. Larger exposed polar surfaces can allow the interaction between the polymer and ions, promoting ion dissociation and enhancing mobility. Consequently, molecules with higher SASA values, especially in polar regions, are more likely to support efficient ion transport.

We also include four physicochemical descriptors based on chemical intuition that can influence ion transport. Hydrophobicity (MolLogP) indicates the balance between polar and nonpolar regions, affecting polymer–ion interactions. Topological Polar Surface Area (TPSA) reflects the extent of polar regions exposed to the environment, which can facilitate ion dissociation and mobility. The number of rotatable bonds captures chain flexibility, with more flexible segments generally allowing easier ion migration. Carbon hybridization fraction provides insight into backbone rigidity and electronic environment, which can impact both polymer packing and ion accessibility.

Finally, in addition to chemical descriptors, we include several physics-informed features that capture polymer size, mobility, and segmental dynamics, which are directly relevant to ionic conductivity. The effective hydrodynamic radius (R_{eff}), estimated from the monomer molecular weight, provides a rough measure of polymer chain size, which influences how easily ions can navigate through the polymer matrix. Using R_{eff}, we estimate a pseudo Stokes–Einstein diffusion coefficient (D), representing the expected mobility of polymer chains in solution and how this may affect ion transport. The glass-transition temperature (T_g) is approximated as a linear combination of molecular weight, polar surface area (TPSA), and chain flexibility (number of rotatable bonds), reflecting the segmental dynamics of the polymer; higher T_g generally indicates stiffer chains and reduced ionic mobility. Finally, viscosity (η) is derived from T_g and backbone flexibility (fraction of sp^3-hybridized carbons), capturing the resistance to segmental motion. Together, these physics-inspired descriptors provide mechanistic insight into how polymer structure and dynamics govern ion transport, complementing the chemical and topological descriptors in our model.

3.2 Machine Learning Models

Two model architectures were used for this project for direct comparison to the original work: random forest regression and graph convolutional neural network.

3.2.1 Random Forest Regression

Random Forest Regression models were developed using the scikit-learn framework to predict ionic conductivity, with log scaling applied to the conductivity. This is because ionic conductivity values tend to

skew towards smaller values and does not represent larger values well. 5-fold cross-validation was performed on the training set to evaluate model performance using mean absolute error (MAE) and R-squared (R^2) as key metrics, facilitating comparison with the original work published by TRI[8]. An 80:20 train-test split of the full dataset was used to separate training and validation data prior to cross-validation, consistent with the approach described in the original work. Hyperparameter tuning was conducted for `n_estimators`, `max_features`, `max_depth`, and `max_leaf_nodes`, with the best-performing values, presented in Table 2, selected based on cross-validation results and subsequently applied to the held-out test set for final evaluation. Feature importance was also performed to see if descriptors most important show agreement with Pearson Correlation Matrix and, more importantly, chemical knowledge.

RF Regression Hyperparameters

Hyperparameter name	Value
`n_estimators`	750
`max_features`	0.3
`max_depth`	20
`max_leaf_nodes`	None

Table 2: Hyperparameters selected for Random Forest Regression Model

To investigate the impact of RDkit features on predictive performance, models were trained and evaluated on three types of datasets: one containing only RDkit-dependent features, one containing only molecular dynamics (MD)-dependent features, and another incorporating both MD-dependent and RDKit-generated molecular descriptors. This comparison was designed to assess whether RDKit descriptors alone could approximate the information captured by MD simulations, as well as to quantify the improvement in predictive accuracy achieved when augmenting the MD dataset with RDKit descriptors. By analyzing MAE and R^2 across these datasets, we can evaluate the contribution of molecular descriptors derived from RDKit relative to MD-only features, and determine the extent to which combining these complementary datasets enhances model performance, providing insights into how feature engineering and descriptor selection influence the accuracy of Random Forest models for ionic conductivity prediction.

3.2.2 Graph Convolutional Neural Network (GCNN)

A graph-based convolutional neural network was generated based on early work from TRI[18] that allows for unique molecular graphs to be generated for each molecule and passed into the same model. As a part of the data processing, the feature maps for the nodes and edges for the molecular graph are generated based on the training data, with a fallback 'UNK' token added to each feature set to account for unseen feature values. Node features (representing atoms) included atomic symbol, the degree of the atom (number of directly-bonded neighbors), the formal charge of the atom, the total number of hydrogens bonded to the atom, the hybridization of the atom, a boolean indicating whether an atom is part of an aromatic ring, and a boolean indicating whether an atom is part of a ring regardless of aromaticity. Edge features (representing physical bonds) included bond type, bond stereochemistry, and the conjugation of the bond. All node and edge features were generated using RDKit and one-hot encoded for each atom and bond, respectively.

The GCNN itself used linear layers for node and edge convolutions before global pooling. If included, the previously described molecular features (MF) would be concatenated with this pooled representation. The resulting vector is then passed through a variable number of hidden layers before outputting a value based on one of the five prediction tasks of the original work from TRI[8]: conductivity (σ), polymer diffusivity (D_{chain}), Li^+ diffusivity (D_{Li^+}), bis(trifluoromethanesulfonyl)imide diffusivity (D_{TFSI^-}), and cation transference number (t^+). Each convolutional layer and hidden layer was activated with a leaky ReLU function. During training, mean-squared error was used as the loss metric, and validation performance was calculated using said loss metric as well as mean absolute error and R^2. At the end of training, these same metrics were calculated upon the testing set.

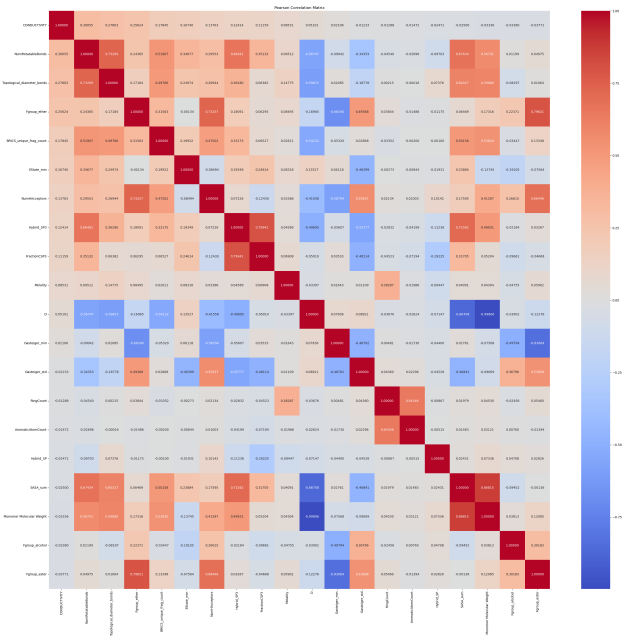

Figure 1: Pearson correlation matrix of RDkit features

4 Results

4.1 Molecular Featurization

Before incorporating additional molecular descriptors, we first examined how the MD-derived features included in the original database correlate with ionic conductivity. Various diffusivities predicted from molecular dynamics (MD) simulations exhibit strong correlation with experimentally reported ionic conductivity, with lithium-ion diffusivity showing the strongest correlation of approximately 0.9. This trend is expected because the ionic conductivity of lithium-ion electrolytes is fundamentally governed by the mobility of Lithium ions; thus, MD simulations that accurately capture Lithium ion transport naturally produce features that align closely with macroscopic conductivity measurements. Among the descriptors, bulk properties such as density show the weakest correlation of -0.01854. This is because density is a relatively coarse structural parameter that does not directly reflect ion transport mechanisms. While density can influence free volume and overall packing, it does not provide detailed information about ion-solvent interactions, hopping pathways, or dynamics that directly determine ionic mobility. As a result, density alone contributes limited predictive power compared to dynamical descriptors such as diffusivities, which more directly encode transport behavior relevant to conductivity.

When observing correlated features produced from RDkit to ionic conductivity, the 20 most correlated data can be seen in Pearson correlation matrix in Figure 1. Unfortunately, the RDkit descriptors show no strong correlation comparable to the MD-derived diffusivities, showing that polymer dynamics are very important in predicting ionic conductivity. However, among the RDKit-generated descriptors, the three features with the highest correlation to ionic conductivity are the number of rotatable bonds, chain-level fragments characterized by topological diameter, and ether functional group. These trends are reasonable with physical expectations: A higher number of rotatable bonds generally indicates a more flexible polymer backbone, which can facilitate easier segmental motion and improve ion transport. The topological diameter bond descriptor, which reflects the overall "reach" or connectivity span of the polymer structure, may relate to how open or extended the polymer network is potentially influencing pathways available for ion motion. Finally, the presence of ether groups also aligns with known polymer electrolyte behavior, as ether oxygens can coordinate with cations (e.g., Lithium ions), increasing local solvation and enabling hopping mechanisms that support higher conductivity.

In contrast, the three features with the weakest correlation are Gasteiger charge variations, ring count,

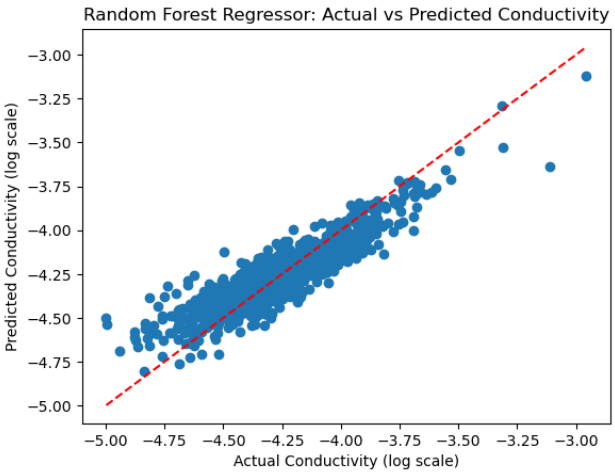

Figure 2: Plot of model prediction of ionic conductivity vs actual values in log scale

and aromatic atom count. These descriptors capture general electronic or structural characteristics that are not strongly connected to ionic transport in these systems. Our reasoning is that Gasteiger charge variation reflects coarse electronic distribution, while RingCount and AromaticAtomCount quantify the presence of rigid or aromatic structural motifs. Since these properties do not directly influence segmental mobility or ion coordination environments, they offer little predictive value for ionic conductivity in this dataset.

4.2 Random Forest Regression Model

We evaluated our models using the same metrics, MAE and R^2, as reported in the original study to enable a direct comparison. The model performance was visualized by plotting the predicted versus actual values from the TRI database[8], as shown in Figure 2.

Here, we can see that the data points generally cluster tightly near the ideal 1:1 relationship, indicating that the model provides a strong predictive capability for the majority of samples. The dense concentration of points around the center reflects that both the true and predicted conductivities are generally low, consistent with the overall distribution of the dataset. One thing to note is that there is a systematic underestimation of higher-conductivity samples, as evidenced by several points lying noticeably below the diagonal at the larger actual values. This suggests that the model performs reasonably for low-conductivity polymers but struggles to extrapolate to rarer high-conductivity cases, likely due to class imbalance and limited representation of those regions in the training data.

Next, we assessed the performance of RDKit descriptors in comparison to the original paper and MD-derived descriptors on this model to investigated whether 1. RDkit-only descriptors with our RF model can be compared to the results of the original papers, and 2. integrating both feature sets leads to a substantial improvement in predictive accuracy. The results are presented in Table 3.

Model	**Performance**	
	MAE	R^2
Original Paper	0.120	0.532
RDKit Only	0.123	0.559
MD Only	0.088	0.749
RDKit + MD (Full)	**0.781**	**0.806**

Table 3: Model Performance Using RDKit, MD, and Combined Feature Sets

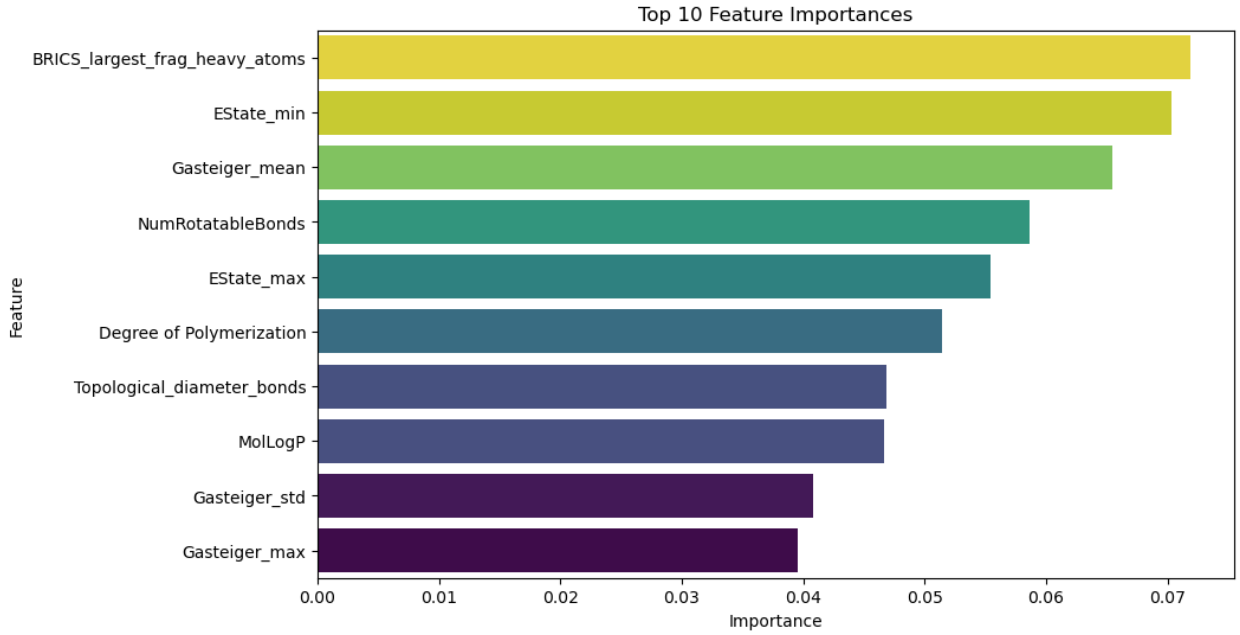

Figure 3: Top 10 feature importances from Random Forest regression model

Model	**Performance**	
	MAE	R^2
Original Paper	0.120	0.532
RDKit Only	0.123	0.559
MD Only	0.088	0.749
RDKit + MD (Full)	**0.781**	**0.806**

Table 3: Model Performance Using RDKit, MD, and Combined Feature Sets

The results show that the RDKit descriptors on their own can predict ionic conductivity reasonably well, but they are not as strong as the MD-derived descriptors. The MD features, which contain information about how molecules behave dynamically, give noticeably better accuracy by themselves. However, the best performance appears when we combine both RDKit and MD features. In this case, the model reaches its lowest error and highest R^2, meaning the predictions match the real values more closely.

Finally, we plotted the top 10 molecular feature importances to identify which descriptors contribute most strongly to the model, as shown in Figure 3.

Our results indicate that the largest heavy atom in block-level fragment of the molecule, as defined by the BRICS scheme (`BRICS_largest_frag_heavy_atoms`), is the most important descriptor, followed by the atom with the lowest electrotopological state (`EState_min`) and the mean Gasteiger partial charge (`Gasteiger_mean`). `BRICS_largest_frag_heavy_atoms` reflects molecular size at block-level and structural complexity, which influence polymer chain mobility and the ability of ions to diffuse through the matrix. `EState_min` captures the electronic environment at the least electron-rich atom, indicating potential sites for ion coordination or interaction. `Gasteiger_mean` represents the average partial charge distribution across the molecule, highlighting the overall electrostatic environment that governs ionic transport. Note that these feature importances show limited agreement with the Pearson correlation matrix, suggesting that nonlinear and interaction effects captured by the tree-based model shift the prominence of certain descriptors. While `NumRotatableBonds` remains influential, features such as topological diameter, and ether group count are less critical in this context.

Including MD-derived descriptors alters the importance landscape: with the most important being Lithium Diffusitivity, `MolLogP` emerges as a dominant feature, followed by `BRICS_largest_frag_heavy_atoms` and `Gasteiger_mean`. This shift likely reflects the ability of MD parameters to capture mesoscale transport

behavior, allowing the model to re-weight features toward those most strongly governing local charge environment, chain flexibility, and solvation dynamics.

4.3 Graph Convolutional Neural Network

These same performance metrics were measured for our GCNN, with MAE and R^2 calculated after training to predict each aforementioned prediction task as well as hyperparameter optimization The search range for the models' hyperparameters (with and without molecular features) as well as the computed ideal values after 10 rounds of Bayesian optimization are shown in Table 4 and Table 5, respectively. For most of the models, the hyperparameter optimization yielded a high number of epochs and a lower batch size, which intuitively would allow for more opportunities for information to be passed in more granular pieces, leading to potentially more nuanced relationships being learned. However, exceptions to these trends as well as the lack of consistency for most of the other features indicates the models are relatively agnostic to many hyperparameters for this architecture.

GCNN Regression Hyperparameters (With Molecular Features)

Hyperparameter	Search range	σ	D_{Li^+}	D_{TFSI^-}	D_{chain}	t^+
batch_size	$32 \sim 512$	58	58	44	389	253
num_epochs	$64 \sim 256$	134	179	195	215	147
fea_dim	$32 \sim 128$	39	67	40	74	110
n_hidden	$2 \sim 8$	2	7	3	4	4
n_layers	$2 \sim 8$	7	7	3	4	7
learning_rate	$10^{-6} \sim 10^{-3}$	2.0×10^{-4}	4.2×10^{-5}	7.0×10^{-4}	5.6×10^{-4}	1.7×10^{-5}
weight_decay	$10^{-8} \sim 10^{-3}$	6.7×10^{-4}	3.8×10^{-6}	7.5×10^{-8}	2.2×10^{-7}	5.6×10^{-6}

Table 4: Hyperparameters selected for GCNN with molecular features. Each column after the "Search range" column represent the hyperparameters used for GCNN to predict certain label. Optimal hyperparameters are found with Bayesian optimization.

GCNN Regression Hyperparameters (Without Molecular Features)

Hyperparameter	Search range	σ	D_{Li^+}	D_{TFSI^-}	D_{chain}	t^+
batch_size	$32 \sim 512$	59	34	95	55	35
num_epochs	$64 \sim 256$	220	177	76	209	252
fea_dim	$32 \sim 128$	111	94	85	46	108
n_hidden	$2 \sim 8$	4	5	5	7	7
n_layers	$2 \sim 8$	7	2	5	4	8
learning_rate	$10^{-6} \sim 10^{-3}$	2.0×10^{-5}	3.1×10^{-4}	1.6×10^{-4}	1.3×10^{-4}	8.7×10^{-4}
weight_decay	$10^{-8} \sim 10^{-3}$	3.8×10^{-5}	1.2×10^{-5}	4.0×10^{-8}	3.9×10^{-7}	7.4×10^{-4}

Table 5: Hyperparameters selected for GCNN without molecular features. Each column after the "Search range" column represent the hyperparameters used for GCNN to predict certain label. Optimal hyperparameters are found with Bayesian optimization.

The results for the GCNN with and without molecular features are plotted in Figure 4 and Figure 5, respectively. While relatively noisy across all epochs, the models appear to converge upon stable values for each prediction task, indicating that the model would likely not benefit from further training epochs and would lead to overfitting

Table 6 displays the MAE and R^2 values for the original compared work as well as our GCNN model with and without the added molecular features after graph pooling. Notably, while both iterations of our GCNN yielded better R^2 values compared to the original work, they also yielded consistently higher MAE. While we are unsure as to why this happens in our model specifically, it indicates that while our models have a strong correlation with the ground truth, the predictions likely have a constant scalar offset from the true value. Furthermore, despite incorporating more information, the GCNN with molecular features performed similarly if not consistently worse than without those same features. This may be due to the fact that these features only enter the model after the molecular graph information is pooled, but the added noise appears to have at best a neutral impact on the model performance.

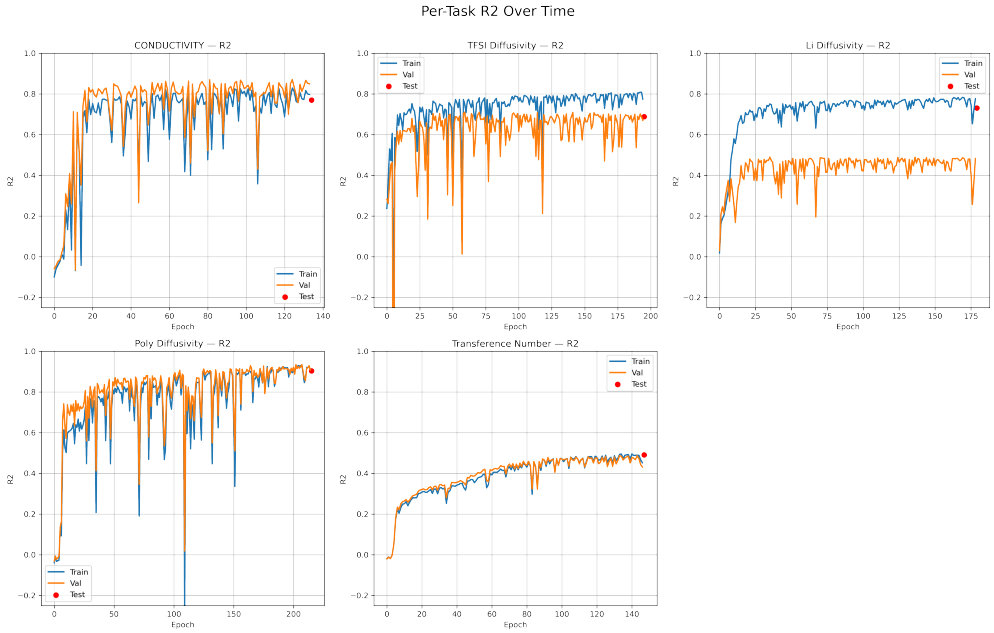

Figure 4: Model Training Performance (R^2, with molecular features)

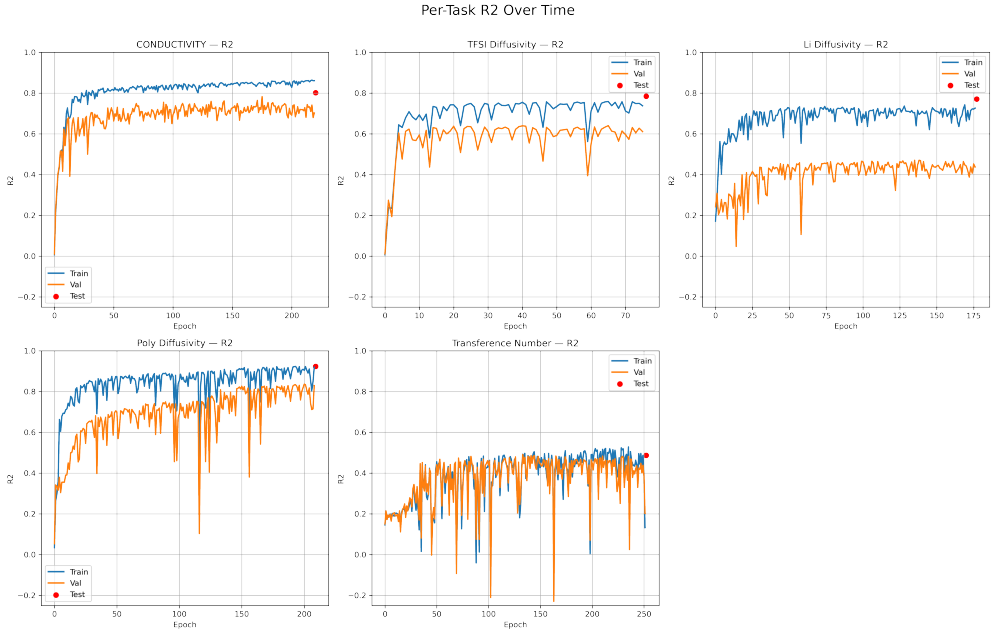

Figure 5: Model Training Performance (R^2, without molecular features)

GCNN Regression Result

Task	Metric	Original Work	GCNN (w/ MF)	GCNN (w/o MF)
σ	MAE	**0.115**	0.310	0.312
	R^2	0.573	0.770	**0.802**
D_{Li^+}	MAE	**0.115**	0.395	0.370
	R^2	0.492	0.731	**0.771**
D_{TFSI^-}	MAE	**0.100**	0.351	0.374
	R^2	0.650	0.688	**0.784**
D_{chain}	MAE	**0.082**	0.216	0.197
	R^2	0.832	0.903	**0.923**
t^+	MAE	**0.159**	0.556	0.561
	R^2	**0.491**	**0.491**	0.487

Table 6: Model Performance for GCNN, with and without molecular features after the graph pooling. Values in bold indicate the best performance of a given task's metric between the three models.

5 Conclusion

Our results suggest that the RDKit descriptors provide useful chemical and structural information to the RF model that the MD descriptors do not fully capture. Even though the MD features are generally more powerful on their own, adding the RDKit features helps the model understand subtle variations in molecular structure, charge distribution, and functional groups. As a result, the combined feature set gives a more complete picture of each molecule, which improves the overall prediction quality.

In terms of the GCNN, we were able to achieve improved correlations between the actual and predicted values for each prediction task (albeit with a scalar offset), indicating the strength of graph-based neural networks in predicting bulk chemical properties by utilizing molecules' unique molecular structures. However, in contrast to our results for the RF model, incorporating molecular information in addition to the molecular graph information appeared to add noise that did not improve if not hindered model performance.

Thus, it follows that more calculated chemical descriptors can allow for more nuanced distinctions/relationships between molecules and their electropolymer-relevant properties in the RF model architectures, but graph-based networks are not as receptive to such information. The architecture itself may be to blame for this lack of performance due to the relative position of where molecular information is typically passed into a graph-based architecture. As such, improved performance may be seen if molecular information is somehow passed into the model in parallel with the graph-based information, but that falls well outside the scope of this study. Additionally, it should go without saying that determining the cause of and eliminating this apparent prediction offset in the GCNN would lead to a high-performing model for the predictions tasks discussed. Regardless, it is evident that predicting properties relevant to developing novel electropolymers remains a challenge, highly dependent on data featurization, model architecture, and data quantity/quality.

6 References

References

[1] P. G. Bruce and C. A. Vincent. Polymer electrolytes. *J. Chem. Soc., Faraday Trans.*, 89:3187–3203, 1993.

[2] Jayeeta Chattopadhyay, Tara Sankar Pathak, and Diogo M. F. Santos. Applications of polymer electrolytes in lithium-ion batteries: A review. *Polymers*, 15(19), 2023.

[3] Hui Yang and Nianqiang Wu. Ionic conductivity and ion transport mechanisms of solid-state lithium-ion battery electrolytes: A review. *Energy Science & Engineering*, 10(5):1643–1671, 2022.

[4] Yunqi Shao, Harish Gudla, Daniel Brandell, and Chao Zhang. Transference number in polymer electrolytes: Mind the reference-frame gap. *Journal of the American Chemical Society*, 144(17):7583–7587, 2022. PMID: 35446043.

[5] Harish Gudla and Chao Zhang. How to determine glass transition temperature of polymer electrolytes from molecular dynamics simulations. *The Journal of Physical Chemistry B*, 128(43):10537–10540, 2024. PMID: 39433295.

[6] Oleg Borodin and Grant D. Smith. Mechanism of ion transport in amorphous poly(ethylene oxide)/litfsi from molecular dynamics simulations. *Macromolecules*, 39(4):1620–1629, 2006.

[7] Omar Allam, Seung Soon Jang, et al. Molecular insights into lithium-ion coordination and morphology in carbonate polymer electrolytes. *Chemistry of Materials*, 37(17), 2015.

[8] T. Xie et al. A cloud platform for sharing and automated analysis of raw data from high-throughput materials screening. *APL Materials Letters*, 1(4):046108, 2023.

[9] Greg Landrum. Rdkit: Open-source cheminformatics. *Journal of Chemical Information and Modeling*, 50(5):871–884, 2010.

[10] A. Dey, Y. Li, X. Chen, et al. Heteroatom doping: An effective way to boost sodium ion storage. *Advanced Energy Materials*, 10(27), 2020.

[11] Harikrishna Sahu, Hongmo Li, Lihua Chen, Arunkumar Chitteth Rajan, Tran Doan Huan, Natalie Stingelin, and Ramamurthy Ramprasad. An informatics approach for designing conducting polymers. *ACS Applied Materials Interfaces*, 13(45):53314–53322, 2021.

[12] S. Jinsong, J. Qifeng, C. Xing, and et al. Molecular fragmentation as a crucial step in the ai-based drug development pathway. *Communications Chemistry*, 7:20, 2024.

[13] Chiho Kim, Anand Chandrasekaran, Tran Ngoc Huan, Deya Das, and Rampi Ramprasad. Polymer genome: A data-powered polymer informatics platform for property predictions. *The Journal of Physical Chemistry C*, 122(31):17575–17585, 2018.

[14] Maximilian Schalenbach, Yasin Emre Durmus, Hermann Tempel, Hans Kungl, and Rüdiger-A. Eichel. Ion transport and limited currents in supporting electrolytes and ionic liquids. *Scientific Reports*, 12(1):6215, 2022.

[15] L. B. Kier and L. H. Hall. An electrotopological-state index for atoms in molecules. *Pharmaceutical Research*, 7(8):801–807, 1990.

[16] Johann Gasteiger and Mario Marsili. A new model for calculating atomic charges in molecules. *Tetrahedron Letters*, 19(34):3181–3184, 1978.

[17] Seungho Choe, Karen A. Hecht, and Michael Grabe. A continuum method for determining membrane protein insertion energies and the problem of charged residues. *The Journal of General Physiology*, 131(6):563–573, 2008.

[18] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.*, 120:145301, Apr 2018.