Predicting Transthyretin Activity Using Graph Convolutional and 3D Convolutional Neural Networks

Helen Henry

Independent Researcher

Abstract

Accurately predicting the activity of chemical compounds against biological targets is crucial in drug discovery and toxicological assessment. This study explores an approach that integrates graph convolutional neural networks (GCNNs) and 3D convolutional neural networks (3D-CNNs) to predict the activity of compounds against transthyretin (TTR) using their SMILES representations. We hypothesized that generating electrostatic potential (ESP) surfaces with a pretrained GCNN would provide richer information for activity prediction. The data exhibited a multimodal distribution with class imbalance and outliers, presenting challenges for modeling. Our final model achieved an RMSE of 31.8 on the validation set, which is suboptimal compared to the competition baseline of 21.8. This discrepancy may be due to the simplicity of the model and the limited voxel resolution used. While the results did not meet expectations, the study provides insights into the complexities of modeling molecular activity and highlights areas for future improvement. The methodologies, code, and voxelized datasets are available in our GitHub repository (`https://github.com/helenhenryz/tox24-challenge-project.git`), facilitating reproducibility and further research.

Predicting Transthyretin Activity Using Graph Convolutional and 3D Convolutional Neural Networks

Introduction

Predicting the biological activity of chemical compounds is fundamental in computational chemistry and drug discovery. Traditional experimental methods for assessing compound activity are time-consuming and resource-intensive, necessitating the development of computational models that can efficiently predict activity based on chemical structure alone (Ekins et al. (2019)). Transthyretin (TTR) is a transport protein implicated in amyloid diseases, making it a significant target for therapeutic intervention (Sekijima (2015)).

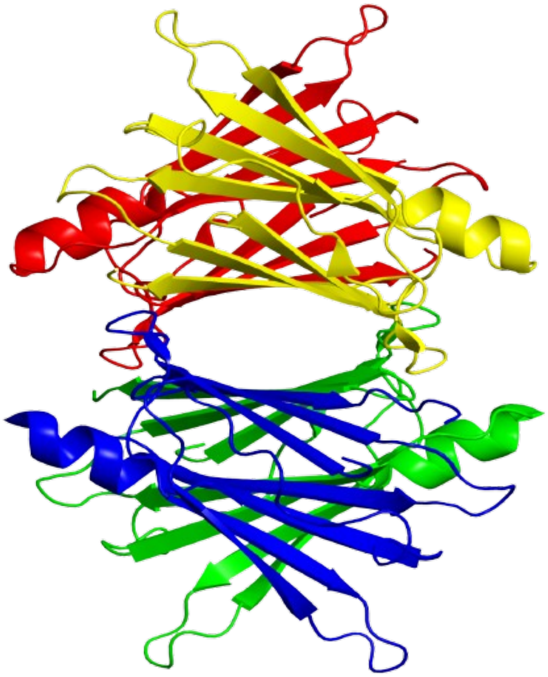

Figure 1. 3D Structure of Transthyretin (TTR) Protein

Figure 1 illustrates the 3D structure of the TTR protein, highlighting its quaternary structure and the binding sites relevant for drug interaction.

The Tox24 Challenge aims to evaluate computational methods for predicting *in vitro* activity of compounds using only their chemical structures (Tetko (2024)). Leveraging this opportunity, we explored a hybrid approach that combines graph convolutional neural networks (GCNNs) and 3D convolutional neural networks (3D-CNNs) to predict TTR activity from SMILES representations. We hoped that by generating electrostatic potential (ESP) surfaces using a pretrained GCNN, we could capture essential molecular features that contribute to biological activity.

Our approach addresses the challenge of predicting compound activity by incorporating spatial and electrostatic information, providing a foundation for further exploration in computational toxicology. This project reflects a commitment to understanding the complexities of molecular modeling and highlights the potential of advanced AI techniques in this field.

Data Analysis

Exploratory Data Analysis

The dataset provided by the Tox24 Challenge consists of:

- **Training Set**: 1,012 compounds with known activity against TTR.
- **Leaderboard Set**: 200 compounds for preliminary testing.
- **Blind Set**: 300 compounds for final evaluation.

Each compound is represented by its SMILES string. We focused on the training set for model development.

Activity Distribution

An initial analysis of the activity values revealed that the data is not normally distributed and exhibits a multimodal structure. There is noticeable class imbalance, and several outliers are present, which can complicate the modeling process.

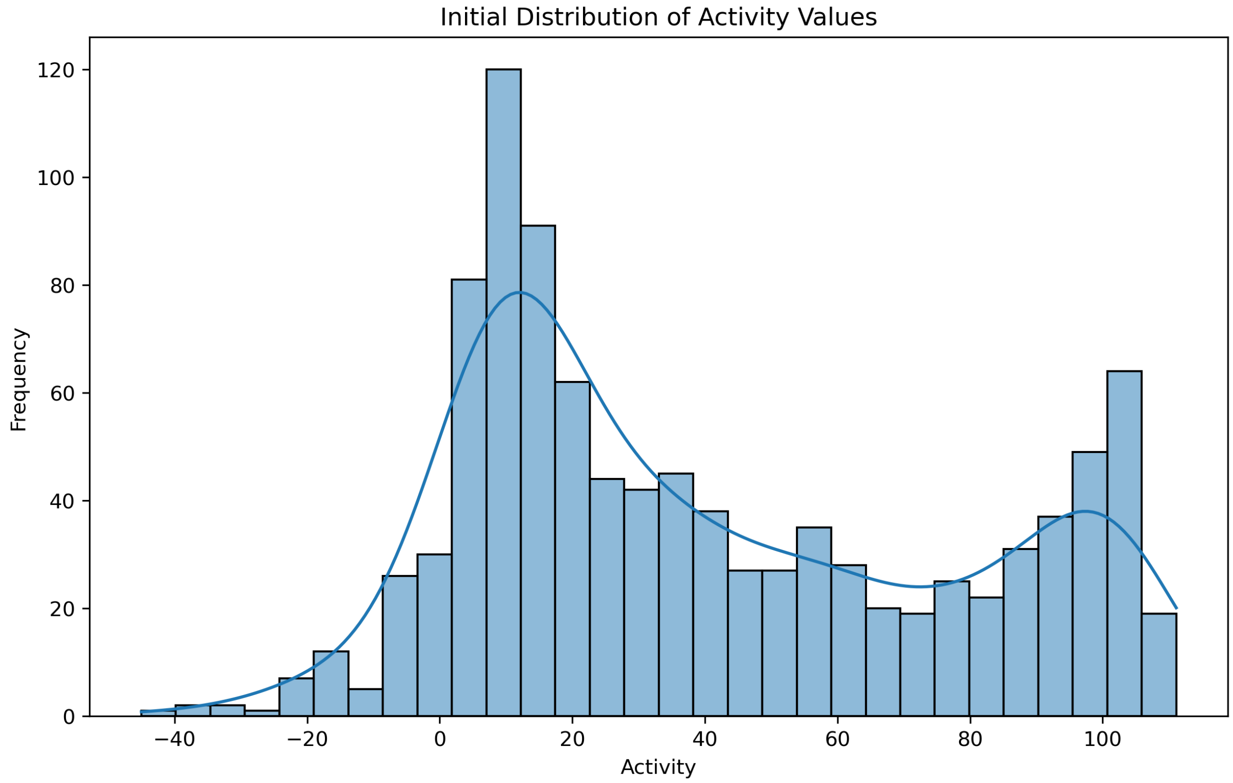

Figure 2. Distribution of Activity Values in the Training Set

Figure 2 shows the histogram of activity values, highlighting the skewness and the presence of multiple peaks.

Data Challenges

The multimodal distribution and class imbalance suggest that certain activity ranges are overrepresented, while others have fewer samples. Outliers may significantly influence model training, potentially leading to biased predictions. These factors necessitate careful consideration in model development and evaluation.

Literature Review

Graph neural networks have emerged as powerful tools for modeling molecular structures due to their ability to capture the relational information inherent in chemical compounds (Duvenaud et al. (2015); Gilmer, Schoenholz, Riley, Vinyals, and Dahl (2017)).

GCNNs, in particular, have been effective in predicting molecular properties by operating directly on graph representations of molecules (Kearnes, McCloskey, Berndl, Pande, and Riley (2016)).

The use of electrostatic potential surfaces provides valuable insights into molecular interactions, which are critical for understanding biological activity (Hu, Lu, and Yang (2007)). Rathi, Ludlow, and Verdonk (2020) introduced a practical method for generating high-quality ESP surfaces using a GCNN, significantly reducing computational costs compared to traditional quantum mechanics (QM) calculations.

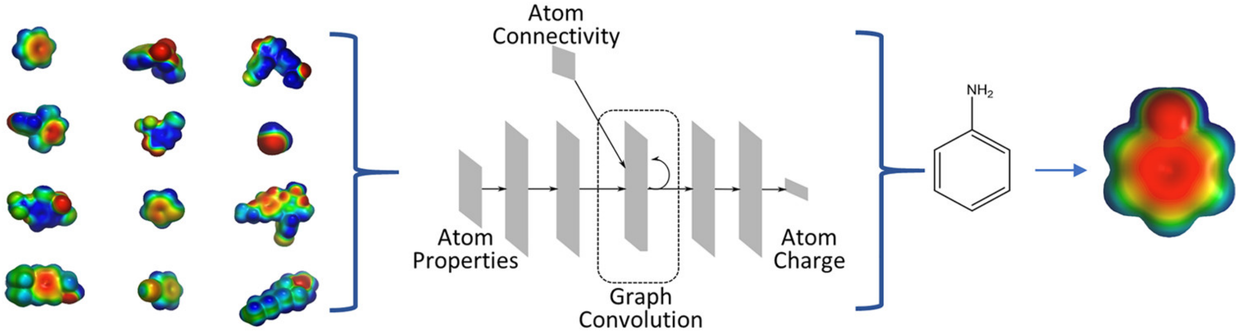

Figure 3. Illustration of ESP Surface Generation using GCNN (Adapted from Rathi et al. (2020))

Figure 3 shows the process of generating ESP surfaces using the GCNN model as described in the ESP-DNN paper.

3D convolutional neural networks have been successfully applied in various domains requiring spatial analysis, including molecular property prediction (Wallach, Dzamba, and Heifets (2015)). By processing voxelized 3D representations of molecules, 3D-CNNs can capture spatial features that are otherwise inaccessible in traditional 2D representations.

Methodology

Electrostatic Potential Surface Generation

SMILES to PDB Conversion. Using RDKit (Landrum et al. (2024)), we converted SMILES strings into 3D molecular structures in PDB format. This step generates initial 3D coordinates necessary for further processing. However, some SMILES strings failed to convert properly due to issues such as invalid structures or unsupported chemical features.

ESP-DNN Model Application. We utilized the pretrained ESP-DNN model provided by Rathi et al. (2020) to generate ESP surfaces. The process involves:

1. **PDB to PQR Conversion**: The PDB files were converted to PQR format, incorporating atomic charges and radii.

2. **ESP Surface Generation**: The ESP-DNN model predicts the ESP values on the molecular surface based on the input PQR files.

Similar to the SMILES to PDB conversion, some PDB files failed during PQR conversion or ESP-DNN processing. After accounting for these issues, we obtained ESP surfaces for 956 training compounds and 481 test compounds.

Data Alignment and Processing. To ensure consistency, molecules were aligned using principal axes calculated from atomic coordinates. The ESP values were scaled using Min-Max scaling to standardize the data.

Data Challenges and Corrections. During the data processing, we encountered several challenges related to molecule conversions and data integrity. The failures in conversions reduced the effective size of our dataset, which may have impacted the model's performance.

Notably, after our data processing was completed, the competition organizers released a data correction on June 24, 2021, stating:

> *Data correction 21.06.24: We reviewed and corrected several structures following remarks of challenge participants. The corrections were done mainly for mixtures and salts. Structure check was extended to prioritize structures as provided by ACS CAS service. In cases when CAS suggestions were ambiguous,*

we used structures as retrieved from PubChem. The structural information and descriptors were also updated.

Since our processing was done before this correction, our dataset may include compounds with structural issues that were later corrected by the organizers. This could have contributed to conversion failures and affected the overall performance of our model.

Voxelization and 3D Grid Creation

The ESP surfaces were converted into voxel grids suitable for input into a 3D-CNN:

1. **Mesh Exporting**: Surface meshes were exported in TMESH format using tools from the ESP-DNN repository.

2. **Voxelization**: Open3D (Zhou, Park, and Koltun (2018)) was used to voxelize the surface meshes into 3D grids with a resolution of 64 × 64 × 64.

3. **Data Augmentation**: Random rotations and flips were applied to the voxel grids to enhance model robustness.

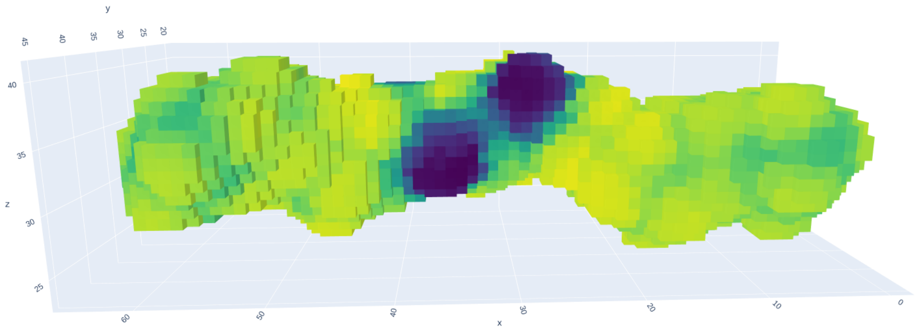

Figure 4. Voxelized Representation of a Molecule's ESP Surface

Figure 4 illustrates the voxelized representation of a molecule's ESP surface, showing the spatial distribution of electrostatic potential. The molecule depicted is defined by its original SMILES string OC(CCCC(C=C)=C)(C)C, which was used to generate the voxelized ESP surface.

Model Development

3D Convolutional Neural Network Architecture. Our 3D-CNN model was designed with multiple convolutional layers, each followed by batch normalization and max-pooling layers. The architecture is relatively simple due to computational constraints but aims to capture essential spatial features.

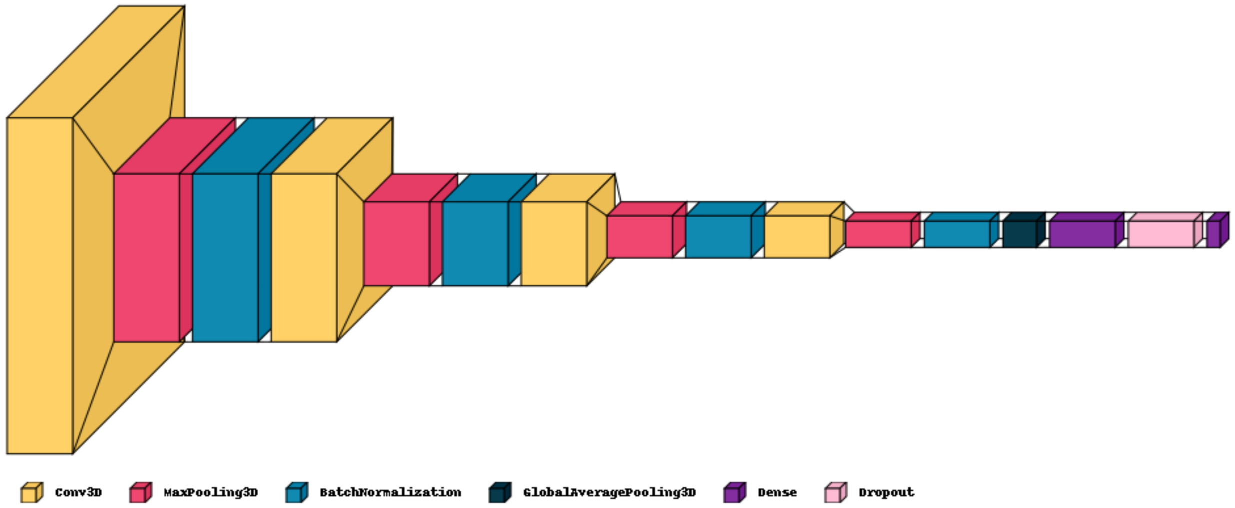

Figure 5. 3D-CNN Model Architecture

Training Parameters. The model was trained using the Adam optimizer (Kingma and Ba (2014)) with an initial learning rate of 0.001. Mean squared error (MSE) was used as the loss function. Early stopping and learning rate reduction were implemented to prevent overfitting and improve convergence..

Evaluation Metrics

We evaluated the model using Root Mean Squared Error (RMSE) between the predicted and actual activity values. RMSE provides a straightforward measure of prediction accuracy, with lower values indicating better performance.

Results

Model Performance

The final model achieved an RMSE of 31.8 on the validation set, which is higher than the competition baseline of 21.8. Figure 6 shows the training and validation loss over epochs.

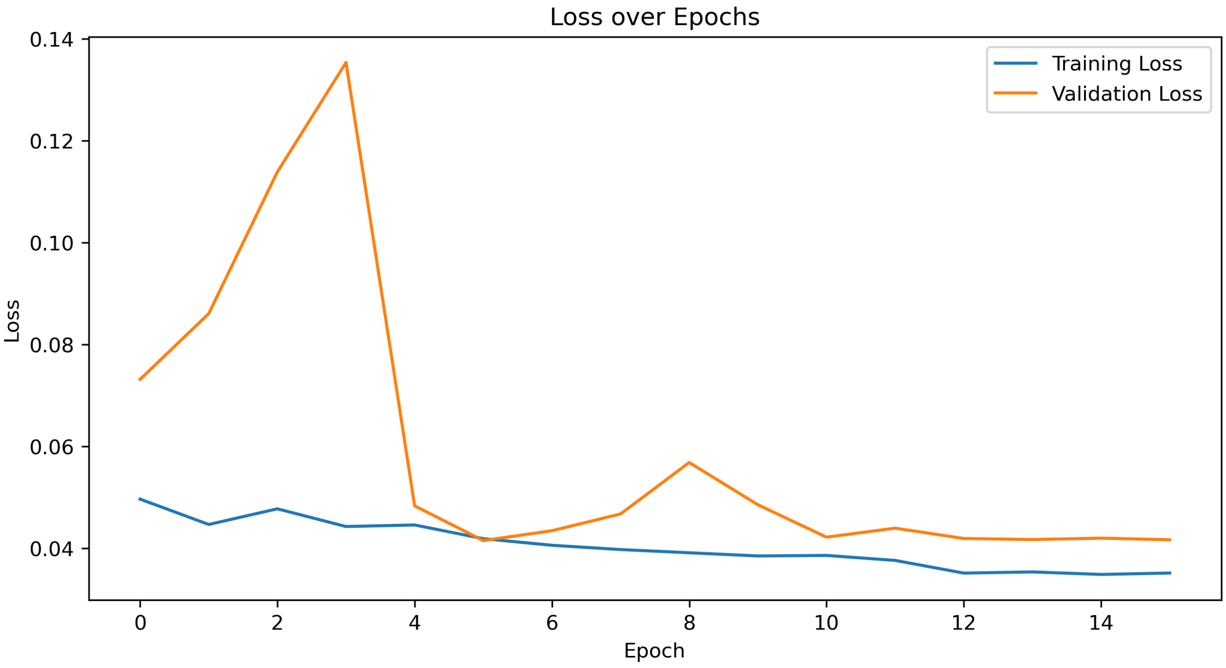

Figure 6. Training and Validation Loss over Epochs

The training and validation losses decreased over epochs but exhibited fluctuations, suggesting that the model may have struggled to generalize from the training data.

Predicted vs. Actual Activity

A scatter plot of predicted versus actual activity values demonstrates the model's predictive capability (Figure 7).

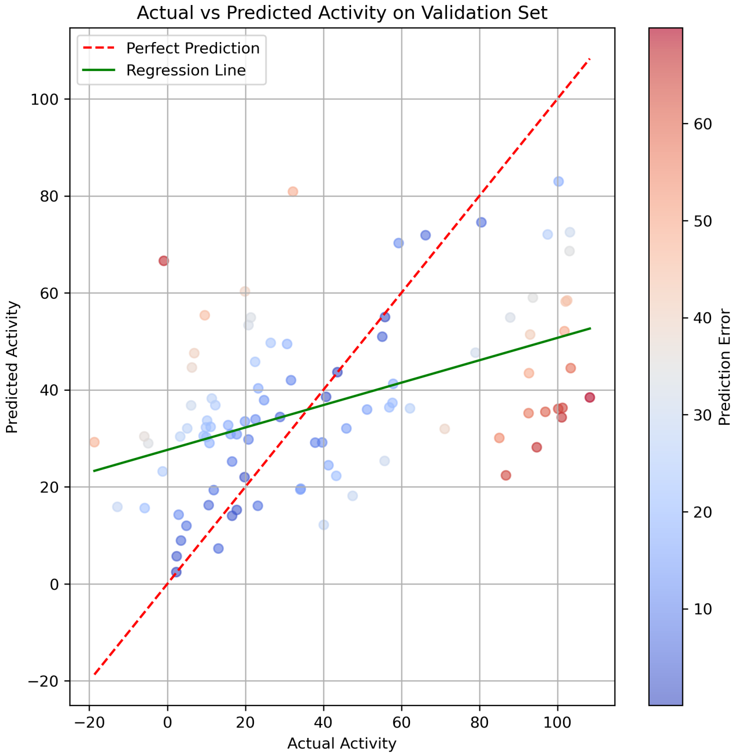

Figure 7. Predicted vs. Actual Activity Values

The scatter plot indicates a moderate correlation between the predicted and actual values but also highlights significant deviations, particularly for higher activity values.

Discussion

Our exploration of using ESP surfaces generated by a GCNN and processed through a 3D-CNN provided valuable insights, although the predictive performance was suboptimal. Several factors may have contributed to the higher RMSE:

Data Challenges

The small size of the dataset posed significant challenges for model training and evaluation. With only 956 training compounds after accounting for conversion failures, the model had limited data to learn complex patterns and generalize effectively. The multimodal distribution, class imbalance, and presence of outliers further complicated the situation. With fewer samples, especially in underrepresented activity ranges, the model may not have received enough information to capture the underlying relationships.

Techniques such as data augmentation, synthetic data generation, or transfer learning could potentially alleviate the limitations imposed by the small dataset size. Additionally, data normalization, outlier removal, or specialized loss functions (e.g., robust losses) might help mitigate the impact of data imbalance and outliers.

Model Complexity

The simplicity of the 3D-CNN architecture may have limited its ability to capture complex relationships in the data. A deeper network with more parameters might better model the nuances of molecular interactions but would require more computational resources and careful regularization to prevent overfitting.

Voxel Grid Resolution

Using a voxel grid resolution of $64 \times 64 \times 64$ may have been insufficient to capture finer spatial details of the ESP surfaces. Higher resolutions could provide more detailed representations but at the cost of increased computational demands.

Future Directions

Future work could involve:

- Experimenting with more complex neural network architectures, potentially incorporating pretrained weights from similar tasks to accelerate convergence and improve generalization.

- Increasing the voxel grid resolution or exploring alternative representations.
- Applying techniques to address data imbalance and outliers.
- Incorporating additional molecular descriptors or features.

Conclusion

This study explored the prediction of TTR activity using an approach that integrates GCNNs and 3D-CNNs, starting from SMILES representations. While the results did not meet our initial expectations, the process provided valuable lessons about the complexities involved in modeling molecular activity. The challenges faced highlight the importance of model complexity, data representation, and handling of challenging data distributions.

Our findings contribute to the ongoing exploration of AI techniques in computational chemistry and underscore the potential for future advancements with refined methodologies.

Acknowledgments

I would like to express my heartfelt gratitude to my brother, Kevin Henry, for his unwavering support throughout this project. His deep knowledge of coding, combined with his constant encouragement and hands-on guidance, played a pivotal role in overcoming the many challenges we faced. This project became a true brother-sister collaboration, and I couldn't have accomplished it without him by my side.

I also wish to sincerely thank Chris Lumen Ben for his generosity in providing access to his GPU, which was crucial for running the computational experiments required for this study.

References

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of the 28th international conference on neural information processing systems - volume 2* (p. 2224–2232). Cambridge, MA, USA: MIT Press.

Ekins, S., Puhl, A. C., Zorn, K. M., Lane, T. R., Russo, D. P., Klein, J. J., … Clark, A. M. (2019). Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials*, *18*(5), 435–441. Retrieved from `https://doi.org/10.1038/s41563-019-0338-z` doi: 10.1038/s41563-019-0338-z

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th international conference on machine learning - volume 70* (p. 1263–1272). JMLR.org.

Hu, H., Lu, Z., & Yang, W. (2007). Fitting molecular electrostatic potentials from quantum mechanical calculations. *Journal of Chemical Theory and Computation*, *3*(3), 1004-1013. Retrieved from `https://doi.org/10.1021/ct600295n` (PMID: 26627419) doi: 10.1021/ct600295n

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, *30*(8), 595–608. Retrieved from `https://doi.org/10.1007/s10822-016-9938-8` doi: 10.1007/s10822-016-9938-8

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, *abs/1412.6980*. Retrieved from `https://api.semanticscholar.org/CorpusID:6628106`

Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., … strets123 (2024, August). *rdkit/rdkit: 2024_03_6 (q1 2024) release.* Zenodo. Retrieved from `https://doi.org/10.5281/zenodo.13469390` doi: 10.5281/zenodo.13469390

Rathi, P. C., Ludlow, R. F., & Verdonk, M. L. (2020). Practical high-quality electrostatic potential surfaces for drug discovery using a graph-convolutional deep neural network. *Journal of Medicinal Chemistry*, *63*(16), 8778-8790. Retrieved from `https://doi.org/10.1021/acs.jmedchem.9b01129` (PMID: 31553186) doi: 10.1021/acs.jmedchem.9b01129

Sekijima, Y. (2015). Transthyretin (attr) amyloidosis: clinical spectrum, molecular pathogenesis and disease-modifying treatments. *Journal of Neurology, Neurosurgery & Psychiatry*, *86*(9), 1036–1043. Retrieved from `https://jnnp.bmj.com/content/86/9/1036` doi: 10.1136/jnnp-2014-308724

Tetko, I. V. (2024). Tox24 challenge. *Chemical Research in Toxicology*, *37*(6), 825-826. Retrieved from `https://doi.org/10.1021/acs.chemrestox.4c00192` (PMID: 38769907) doi: 10.1021/acs.chemrestox.4c00192

Wallach, I., Dzamba, M., & Heifets, A. (2015). Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *ArXiv*, *abs/1510.02855*. Retrieved from `https://api.semanticscholar.org/CorpusID:16690451`

Zhou, Q.-Y., Park, J., & Koltun, V. (2018). Open3D: A modern library for 3D data processing. *arXiv:1801.09847*.

Code Availability

The code and voxelized datasets are publicly available in our GitHub repository: `https://github.com/helenhenryz/tox24-challenge-project.git`