Predictive Modeling of Diffusion Rates in Pure Solvent & Hydrogel Systems

Andrew Lefors, Joshua Yuan

Contents

Abstract

Hydrogel microneedle systems have gained increased attention for their unique structural properties, enabling them to reliably deliver medication without pain over sustained periods of time directly at the target site. To control the release rate of medications in these systems, diffusion rates for the compounds in the given system must be known a prior. To enable future studies, a predictive model was developed using a feed-forward deep neural network to predict diffusion rates in pure solvents, improving upon current machine learning models using Ensemble Tree methods and Support Vector Machines (SVM). A second model was developed to utilize the larger pure solvent dataset to predict diffusion rates in hydrogel systems.

Introduction

In recent years, hydrogels have garnered significant attention as potential delivery systems for therapeutics owing to their tunable properties and controllable dissolving rates[ref]. These micro systems can deliver therapeutics directly to the target site, reducing harm to healthy cells and reducing the overall load required for proper therapeutic levels in the tissues. The key to controlling the release rate of the desired compounds depends primarily on the diffusion rate of the given compound in that hydrogel system through mass transport.

The same tunable properties of hydrogel also make it more difficult to calculate the diffusion rates, as an increase in concentration of polymers, or different polymers altogether, can greatly alter the rate of diffusion by changing the network-mesh the solute must travel through. This in turn reduces the rate at which new designs can be tested, as the diffusion rates must first be experimentally determined and then incorporated into the design for validation. Thus, the cost for developing these novel devices may increase to a point where they are no longer feasible over more developed technologies in use.

The current study has two objectives:

1. Improve upon the Predictive model of diffusion coefficients in pure solvents using fine tree and Support Vector Machine (SVM) models.
2. Develop a model capable of generalizing to hydrogel systems by combining two datasets and feature engineering.

By developing these tools, we hope to reduce the time from research to patient time for novel biomedical devices and reduce human suffering.

Artificial Intelligence Model

To improve upon the existing models, a feed-forward deep neural network (FFDNN) architecture was designed with one (1) input layer for each of the 148 features, four (4) hidden layer, and one (1) output layer. A DNN was chosen because it can capture features of non-linear data sets, it can perform its own feature extraction and find meta-features, and it has the ability to generalize better than other methods. This reduces our need to design many features and utilize existing datasets, allowing the model to determine what is relevant. The hyperparameters for the model were tuned using the keras_tuner package, minimizing the loss for each epoch on the validation set until the best was found. The model uses the Adam optimizer and Relu/Tanh activation functions with a linear output function.

Model Training Method:

To train the feed forward deep neural network (FFDNN), several similar but distinct models were trained using the same data, differing mainly in the number of hidden layer and the number of neurons per layer, optimizing on the root mean squared error as the loss function. The keras_tuner package was used to iterate through various hyperparameters to arrive at optimal loss based on the validation set. Normalization was used on some training models, but in each instance the predictions tended to be almost all the same, resulting in a high degree of error. Using the raw data provided better performance overall, and thus those models trained on the raw data were utilized for the evaluation.

It was found that decreasing the number of neurons in the hidden layers to less than 20 improved the predictive capacity of the model, as did using a split of 80/10/10 for training/validation/testing. The models also performed best when the number of hidden layers was less than or equal to 5. In the end, two models were chosen for their ability to predict most samples within 25% of their actual value, optimizing their Pearson correlation coefficient. The two chosen models were named Delta_v3 and Delta_v4.

Data Collection/Engineering

Raw data

The raw data was collected from previous studies on the non-electrolytic organic compounds in pure solvent diffusion studies. The data was made available as one dataset from the previous predictive models using fine tree and SVM. This collection stems from dozens of research studies experimentally determining diffusion rates, and thus provides an excellent raw-data source from which to begin.

The second dataset utilized is much smaller but is the most important data-set available, as it relates to diffusion in hydrogels using a relatively simple fluorescently labeled dextran molecule, enable quick calculation of functional group contributions to add to the existing larger dataset. Unfortunately, due to the obfuscation of the functional groups in the given datasets and limited time to communicate with the corresponding author, the additional dataset was not able to be incorporated, and therefore had to be removed from the evaluation process.

Cleaned data

The larger dataset containing diffusion rates for the 4823 organic non-electrolytic compounds was cleaned by removing name identifiers and any other irrelevant information, such as originating dataset. The remaining data consisted of 148 feature columns and 1 label column with 4823 rows.

The second data set was first cleaned by removing the identifiers and removing solute size. The functional group contributions were calculated and added, using the size to estimate the additional group contributions. Since dextran is a polymer, the number of attached monomers will change the functional group contributions, making size an adequate proxy for this estimation.

Exploratory data analysis

Visualization techniques, such as scatter plots, histograms, and box plots, were used to identify trends and variations within the data. Correlation matrices were generated to assess the strength and direction of relationships between variables, highlighting potential dependencies. Grouped analyses were performed to discern patterns specific to distinct hydrogel types, shedding light on their unique diffusion characteristics. Additionally, Pearson correlation tests were run to compare to previous studies.

Results

Model 1: DV3

Delta_v3 (DV3) was trained with a 60/20/20 training/validation/testing split on the data. The results are as follows:

Figure 1. Displays the percent adherence the predictions have with the actual diffusion values. 78.4% of all predicted diffusion rates were within 25% of their actual values.

Figure 2. Displays the Training and Validation loss over the number of epochs run for the hyperparameter tuning. Since diffusion rates are on the scale of 1e-5 and smaller, even the smallest step can cause dramatic changes in the models' performance. Due to this, the next model was designed with a lower learning rate.

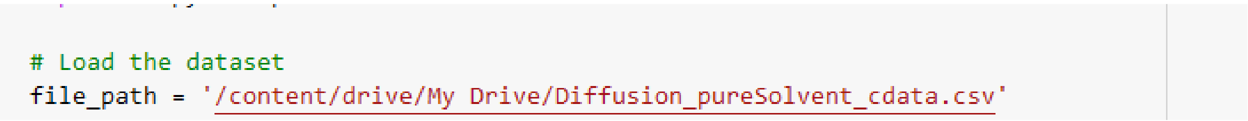
Figure 3. on the top left Is a scatter plot of the predicted values with the actual values along the diagonal. A Pearsons Correlation Regression was performed on the data to compare against other models and provide a metric for self-evaluating performance. This figure was made from data produced by our model.
Figure 4 and **Figure 5** are the models from the (Zhou et al, 2022) study. Our first model outperforms the SVM (**Figure 5**) model, but the falls short of the fine tree (**Figure 4**) method. To improve our results, the learning rate was reduced, and a new model was trained.

Model 2: DV4

Delta_v4 (DV4) was trained with an 80/10/10 split training/validation/testing split on the data. The results for the model are as follows:
Figure 6. Displays the precent adherence the new models predictions have with the actual diffusion values. 84% of all predictions lie within 25% of their actual values, an improvement over the previous model.
Figure 7. Shows the training and validation against increasing epochs. Again the variation is much less in this model compared to the previous, likely owing to the lower learning rate enabling finer tuning on a dataset where small changes have large impacts on the prediction.
Figure 8. Is the scatter plot of Predicted vs Actual diffusion rates, with a Pearson Correlation coefficient R of 0.84. Testing this model with a hydrogel dataset to see if it generalizes well is the next step

Code:

The code used to develop the models and figures for this report can be found at the following collab link:
https://colab.research.google.com/drive/1_vtpTrV749ULiM7cNYii8G33byiP7gBt?usp=sharing
Please note this is an editors link, please do not share. To run the code, download the included diffusion_pure-solvent.csv dataset, change the filepath to the location of the dataset:



```
# Load the dataset
file_path = '/content/drive/My Drive/Diffusion_pureSolvent_cdata.csv'
```

Ensure you have the requisite libraries installed, including keras_tuner, scikitlearn, and tensorflow. If you do not have any of the libraries, a simple pip install will suffice.

Once you have the data loaded, you can run the cell to split the data into train and test segments, and run the cell to define the model iteself. Each of the two distinct models have different hyerparameter tunings, and the configurations are included in separate files to download. Otherwise you can re-train using the same parameters, but this may take some time.

Discussion:

Due to the limited availability of diffusion data in hydrogels, a generalized model from diffusion in water was attempted to be improved upon and applied to a hydrogel dataset. The dataset was feature engineered to be added to the larger diffusion dataset, but due to hiding of the functional group categories in the database, the assignment of group contributions for the solutes in the hydrogel studies were not able to be determined.

Future work will seek to establish a database for hydrogel diffusion studies and incorporate these with the larger pure-solvent datasets by adding a feature for percent polymer volume fraction. For pure solvent, this feature is simply 0, hopefully enabling a more robust and generalizable model, as diffusion through pure solutes still occurs in hydrogels.

Accomplishments:

Developed a model capable of predicting diffusion in pure solvents, similar in capability to those published. Datasets were collected and are being feature engineered to generalize the model for hydrogel systems. The model for hydrogel systems was not developed, but the groundwork has begun to make use of what is available. We also learned it is much more difficult to find highly specialized data, as even when it’s available, portions can be obfuscated, challenging the process.

Figures:

Figure 1: Percent Adherence Histogram Model: DV3

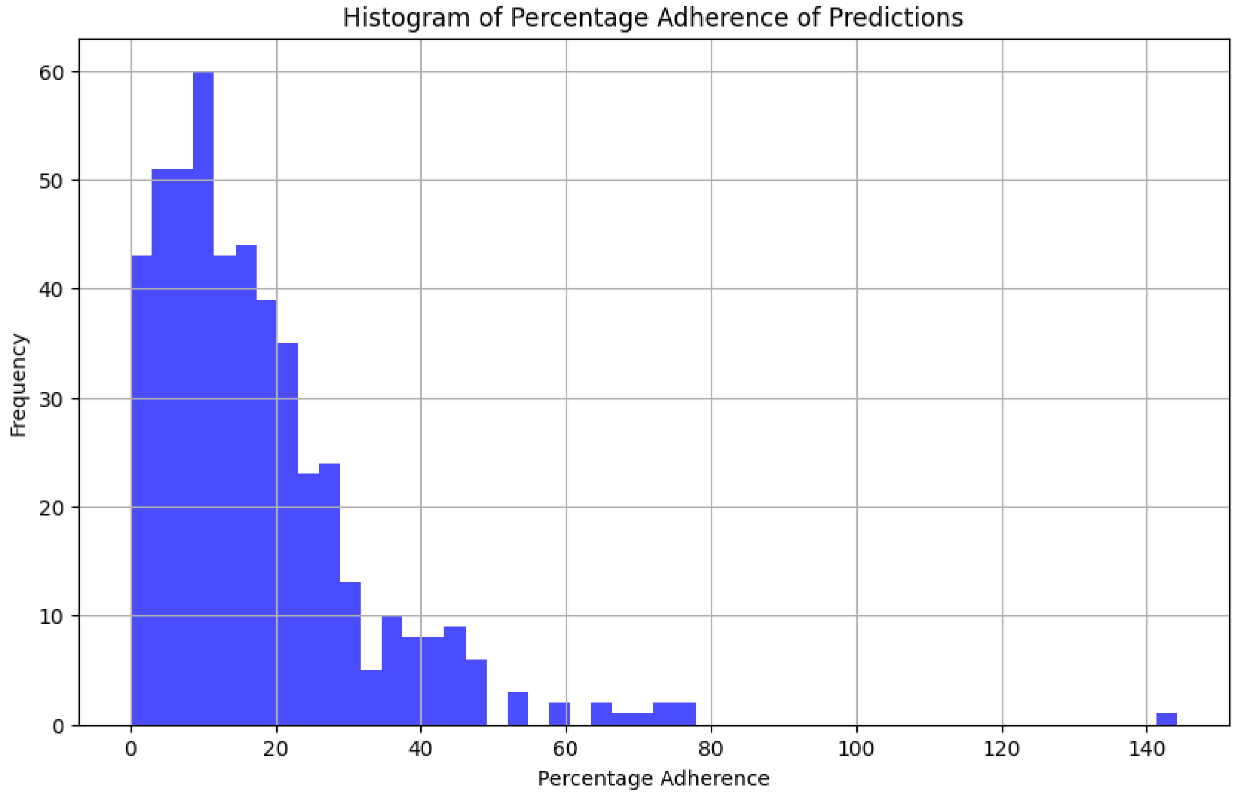


Figure 2: Training and Validation Loss DV3

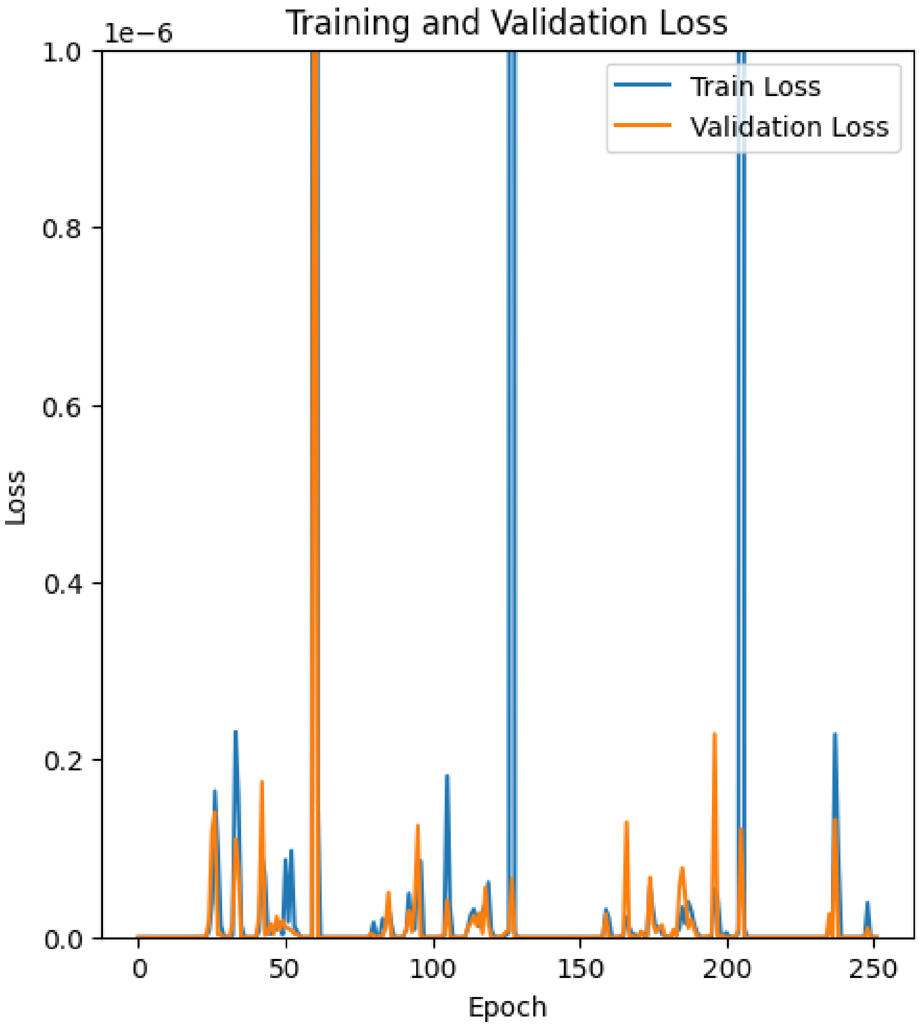


Figure 3: Scatter Plot of Predicted vs Actual Diffusion Values

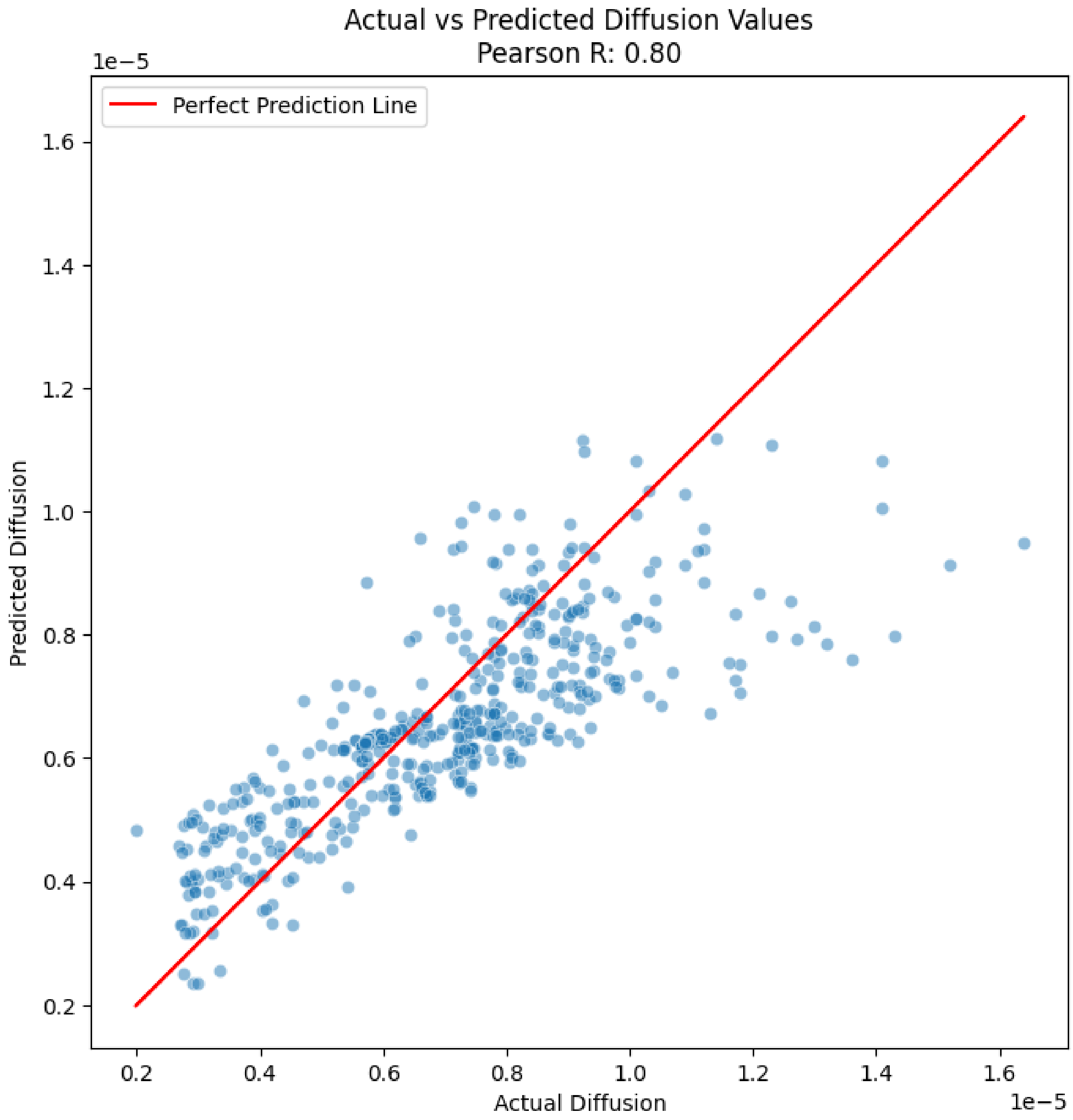


Figure 4: Zhou et al. Fine Tree Model performance

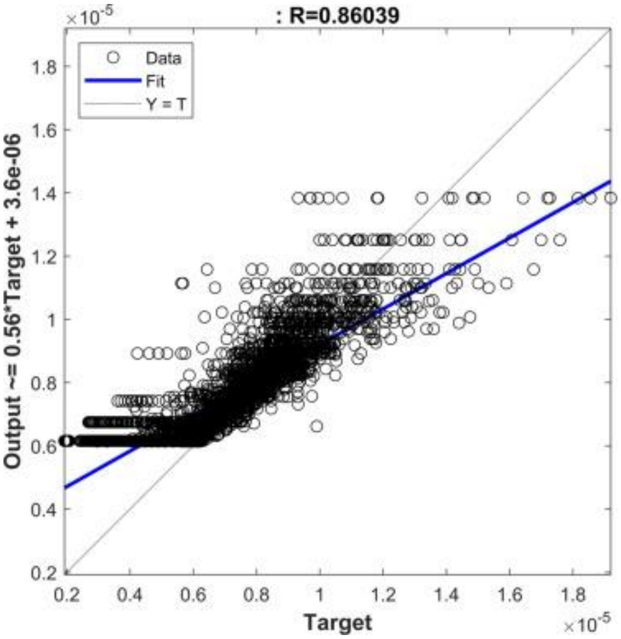


Figure 5: Zhou et al. SVM Model performance

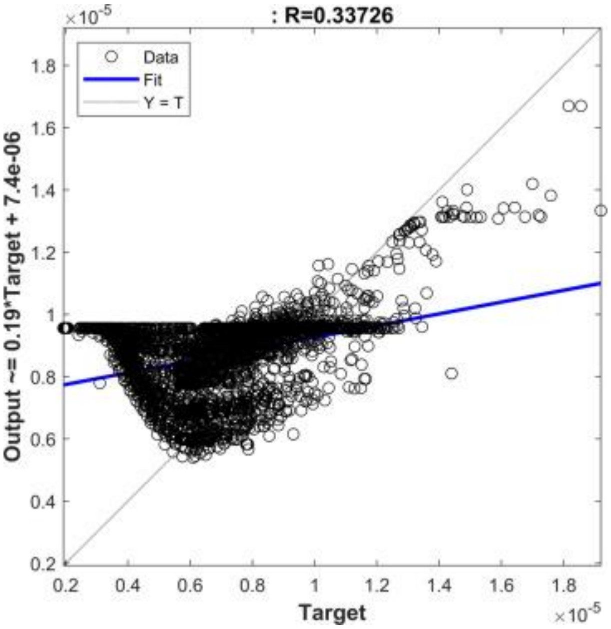


Figure 6: Percent Adherence Histogram Model DV4

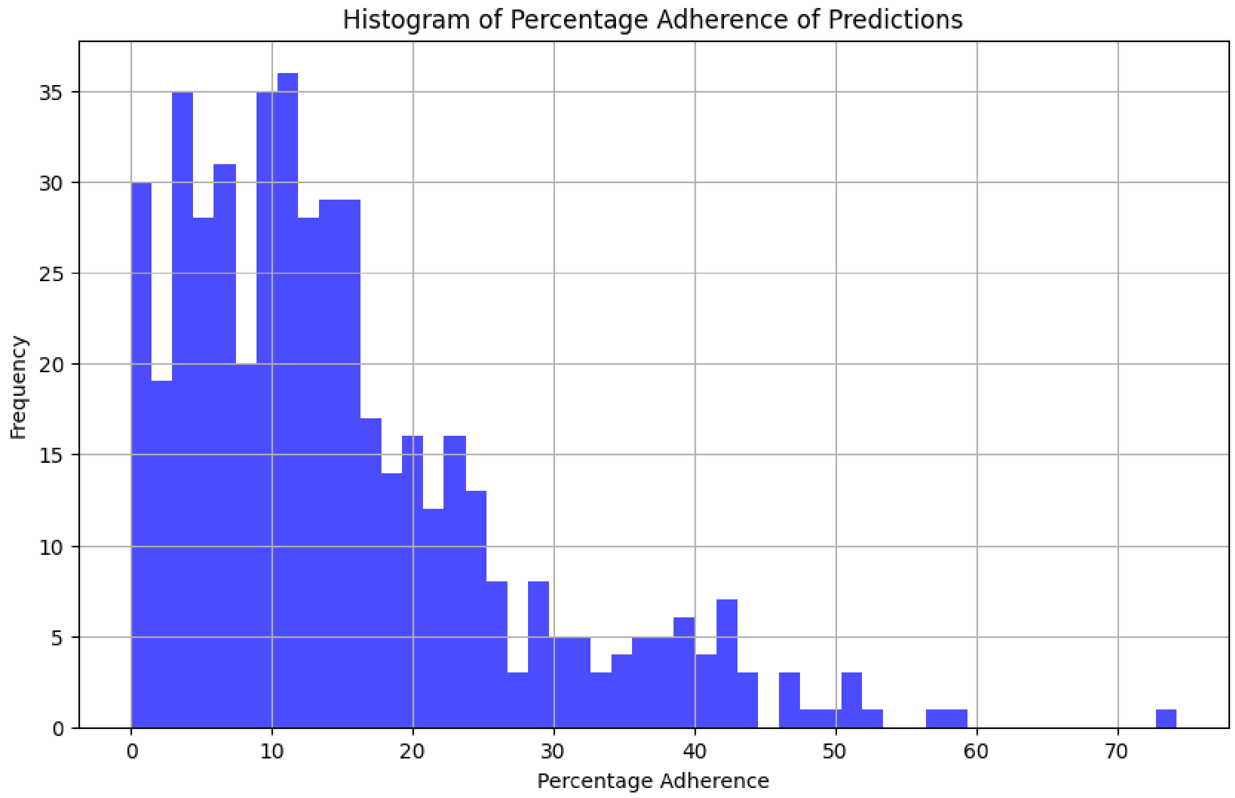


Figure 7: Training and Validation loss over Epochs Model DV4

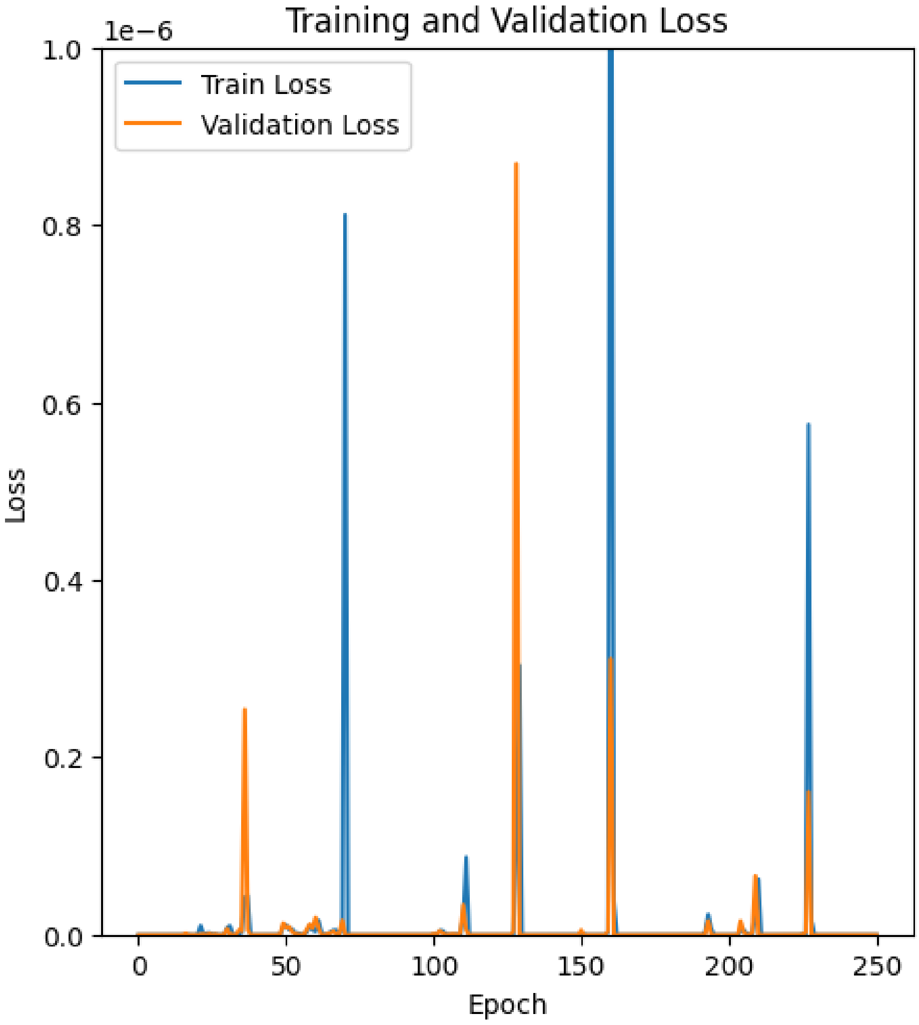
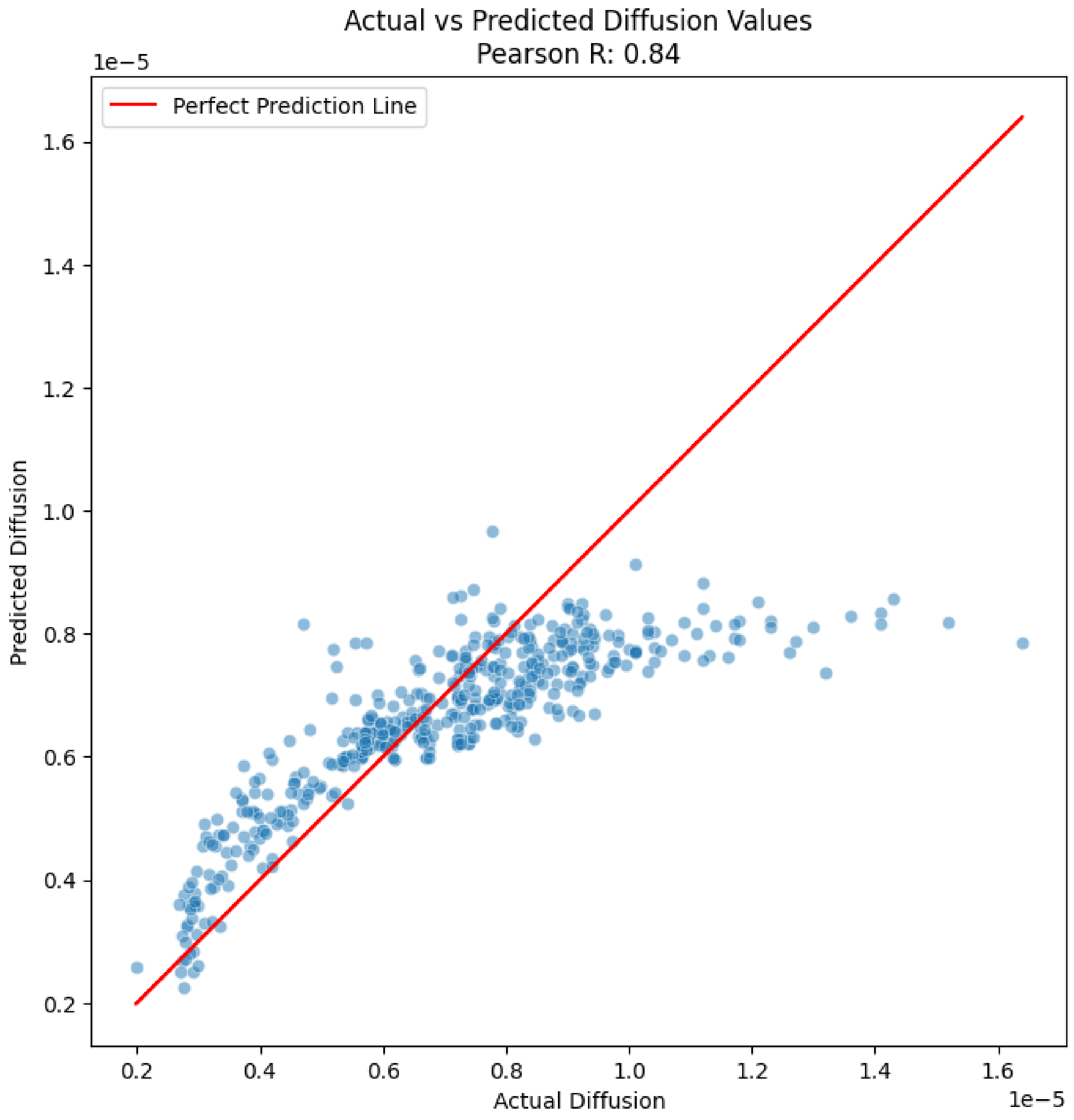


Figure 8: Scatter Plot of Predicted vs Actual Diffusion Values Model DV4



References:

Drug Diffusion in Biomimetic Hydrogels: Importance for Drug Transport and Delivery in Non-Vascular Tumor Tissue." *ScienceDirect*, Oliver Degerstedt, Johan Gråsjö, Anton Norberg, Erik Sjögren, Per Hansson, Hans Lennernäs, 1 May 2022, https://www.sciencedirect.com/science/article/pii/S0928098722000355. Accessed 16 Oct. 2023.

Lavrentev, Filipp V, et al. "Diffusion-Limited Processes in Hydrogels with Chosen Applications from Drug Delivery to Electronic Components." *Molecules (Basel, Switzerland)*, U.S. National Library of Medicine, 7 Aug. 2023, www.ncbi.nlm.nih.gov/pmc/articles/PMC10421015/.

Axpe, Eneko, et al. "A multiscale model for solute diffusion in Hydrogels." *Macromolecules*, vol. 52, no. 18, 2019, pp. 6889–6897, https://doi.org/10.1021/acs.macromol.9b00753.

Vigata, Margaux, Christoph Meinert, Dietmar W. Hutmacher, and Nathalie Bock. 2020. "Hydrogels as Drug Delivery Systems: A Review of Current Characterization and Evaluation Techniques" *Pharmaceutics* 12, no. 12: 1188. https://doi.org/10.3390/pharmaceutics12121188

Zhou, J., Chupradit, S., Ershov, K. S., Suksatan, W., Marhoon, H. A., Alashwal, M., Ghazali, S., Algarni, M., & El-Shafay, A. S. (2022). Prediction of molecular diffusivity of organic molecules based on group contribution with tree optimization and SVM models. Journal of Molecular Liquids, 353, 118808. https://doi.org/10.1016/j.molliq.2022.118808