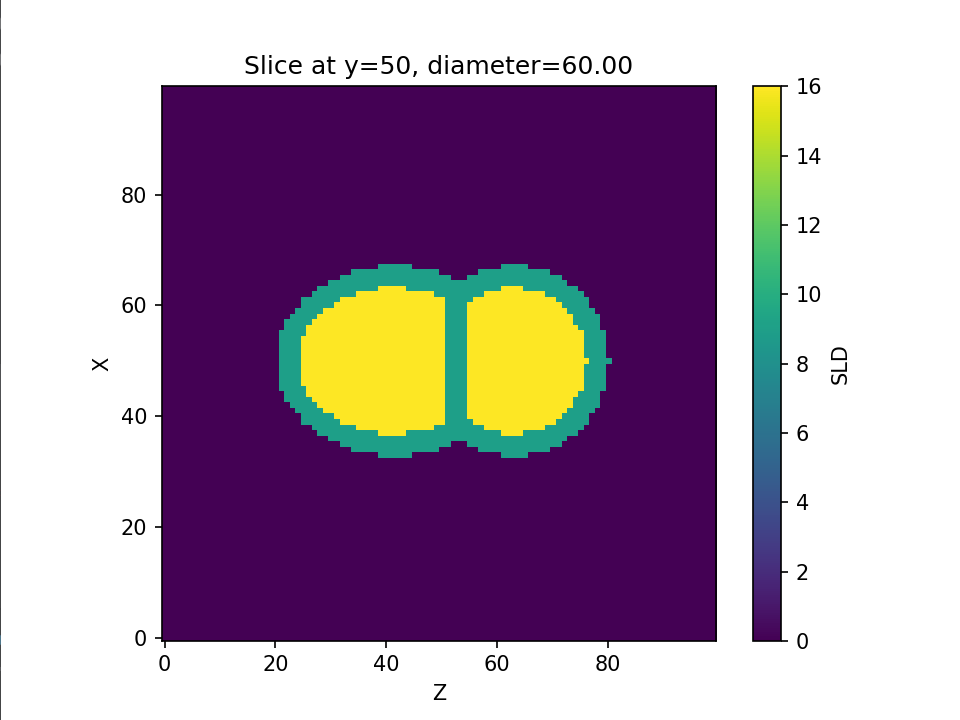
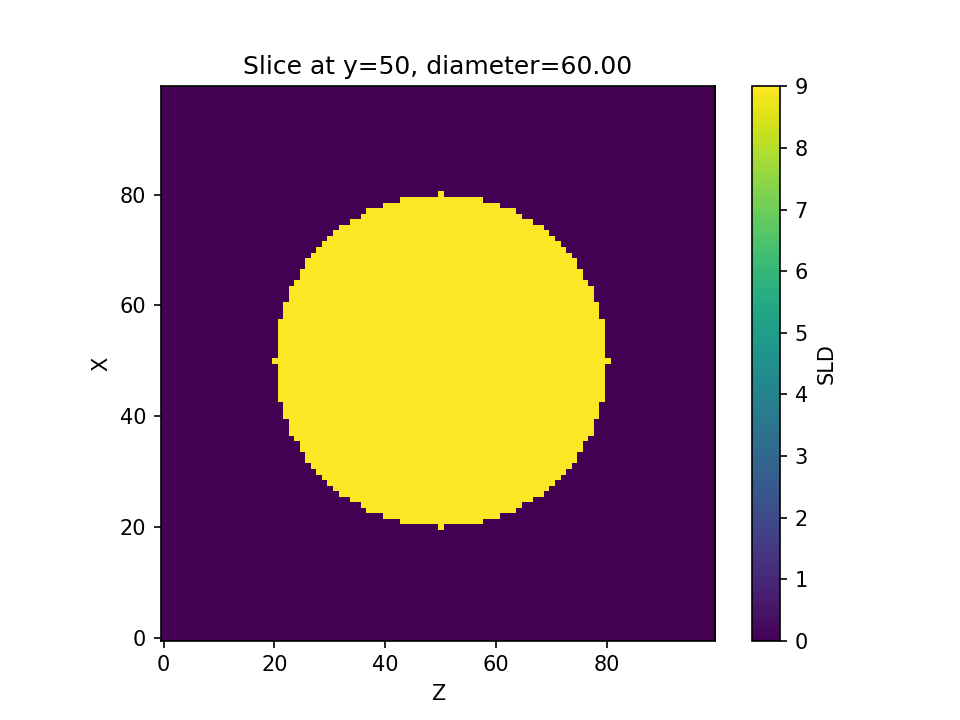
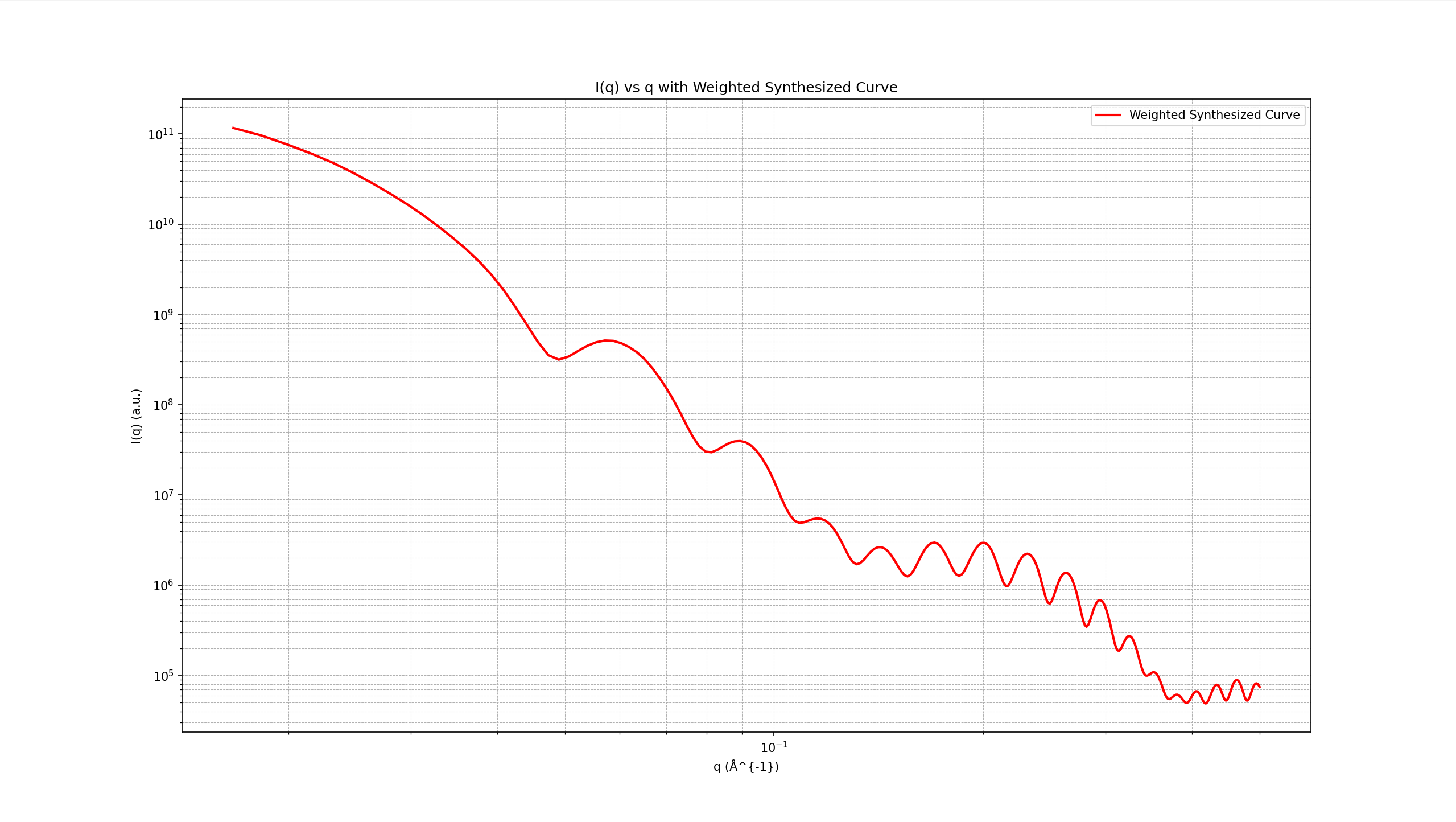
Model 3:



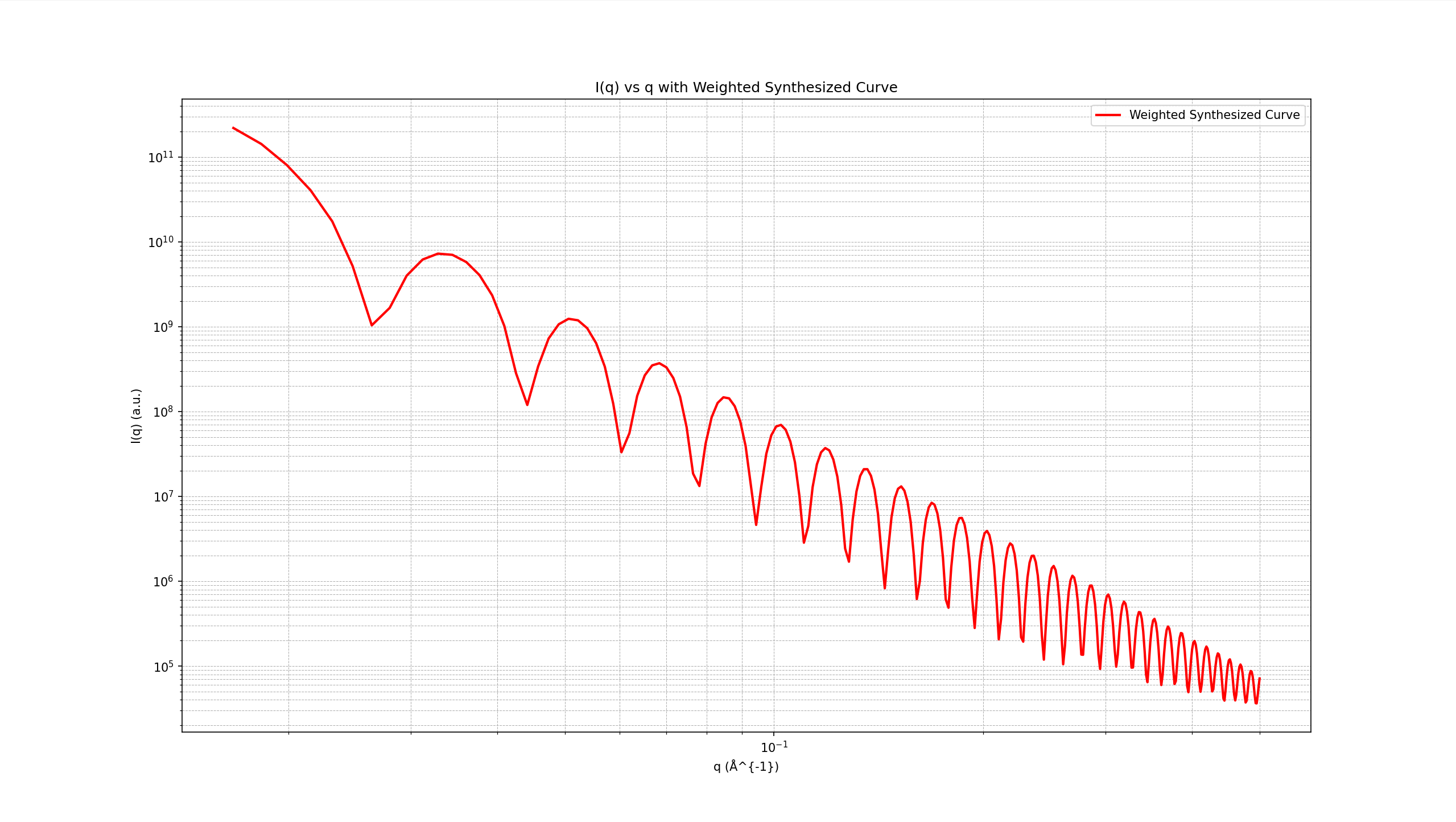
Model 4:



SAXS Model3 Size=60nm

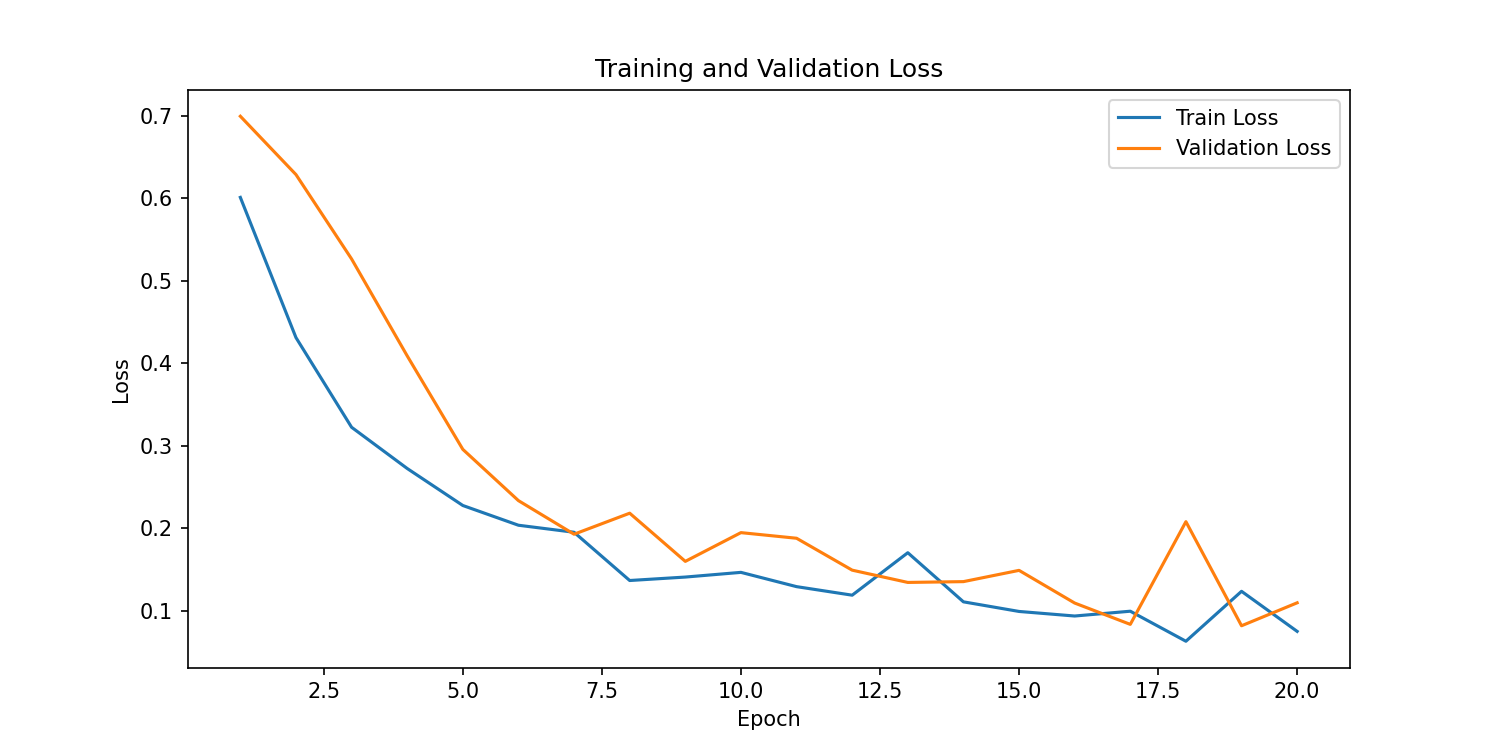


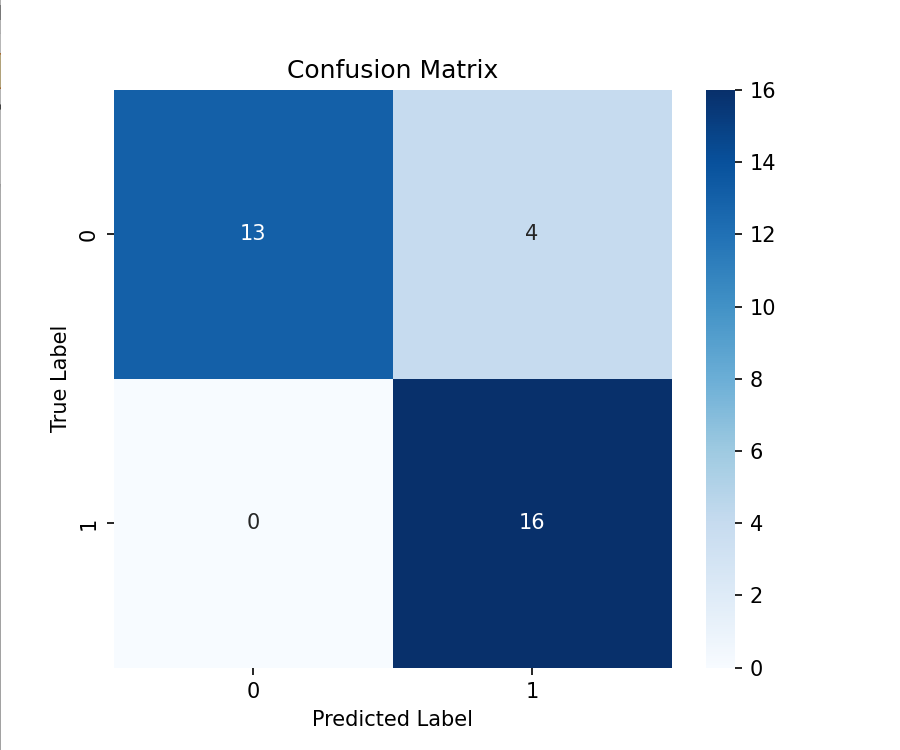
SAXS Model4 Size=60nm



V3.0 Clean data

Only two classes of LNP are involved, each class having 81 samples, 20, 21, …, 100nm. The dataset is segmented to 60%, 20%, 20% for training, validation, test dataset. The results are shown below: (test accuracy: 87.88%)

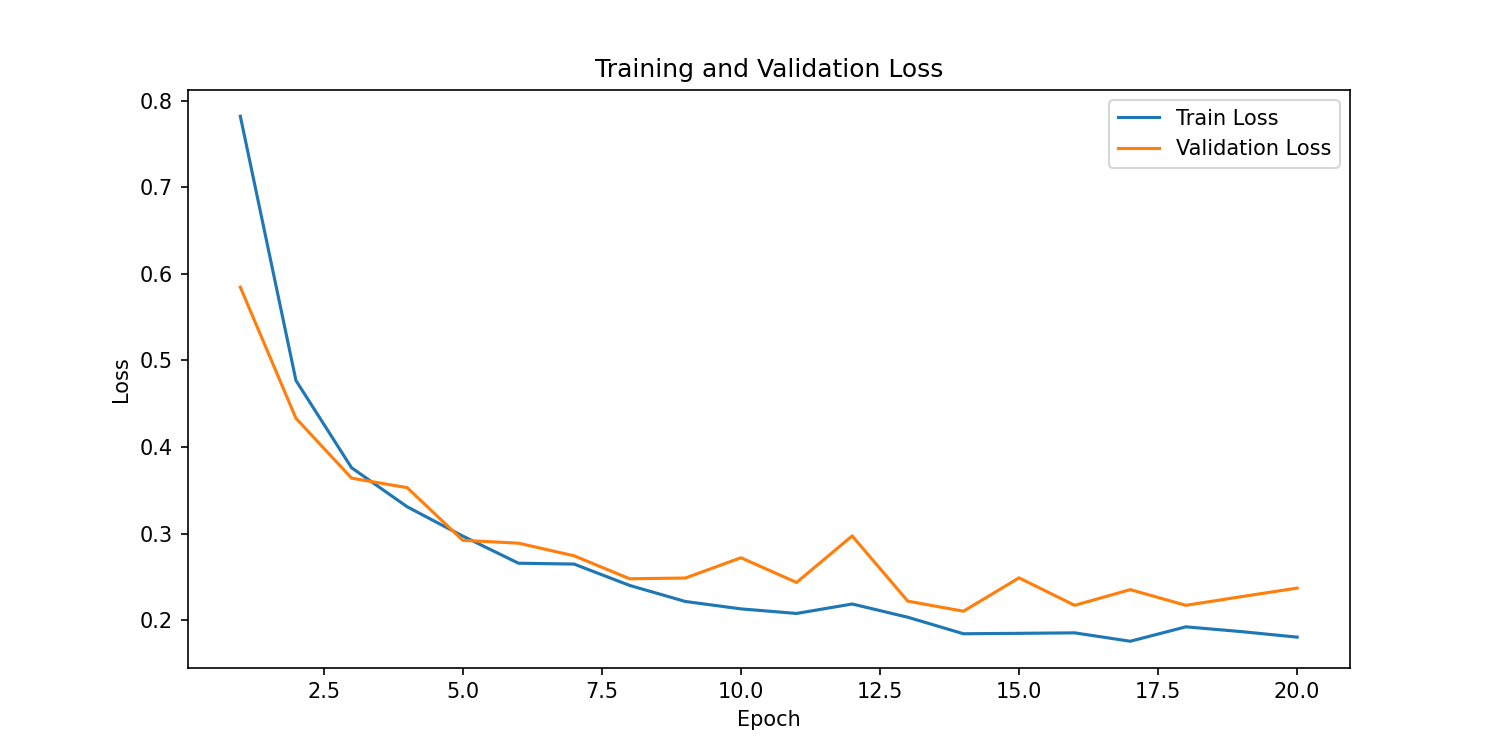


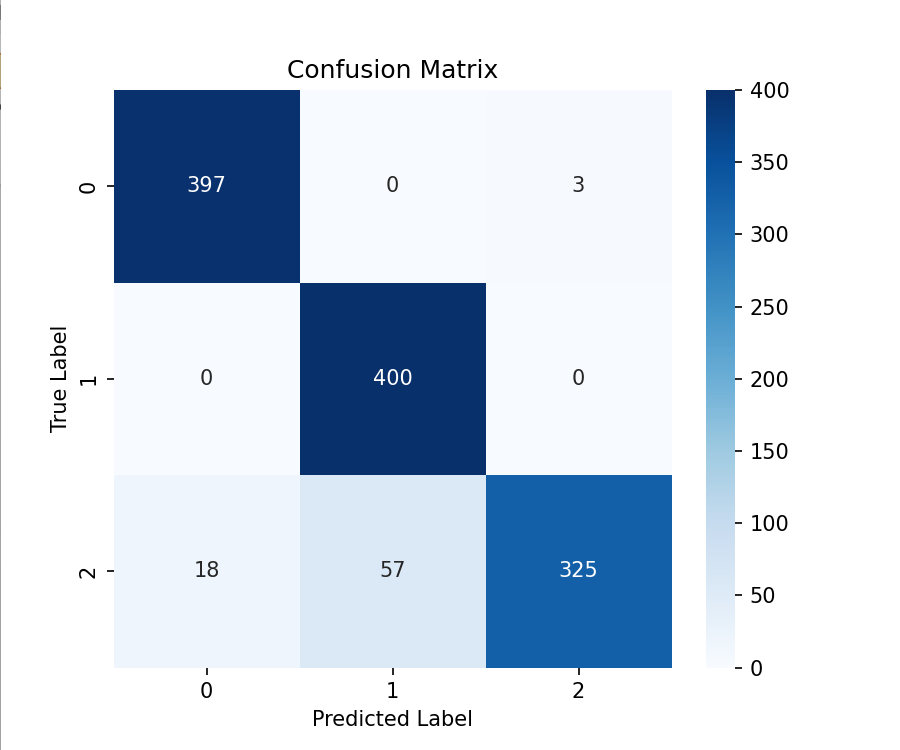


V3.1 Clean data

The above results show that the model can easily distinguish between these two types of LNPs based on SAXS data. Then, we can move to V3.1, where heterogeneous class and more data in each class are introduced.

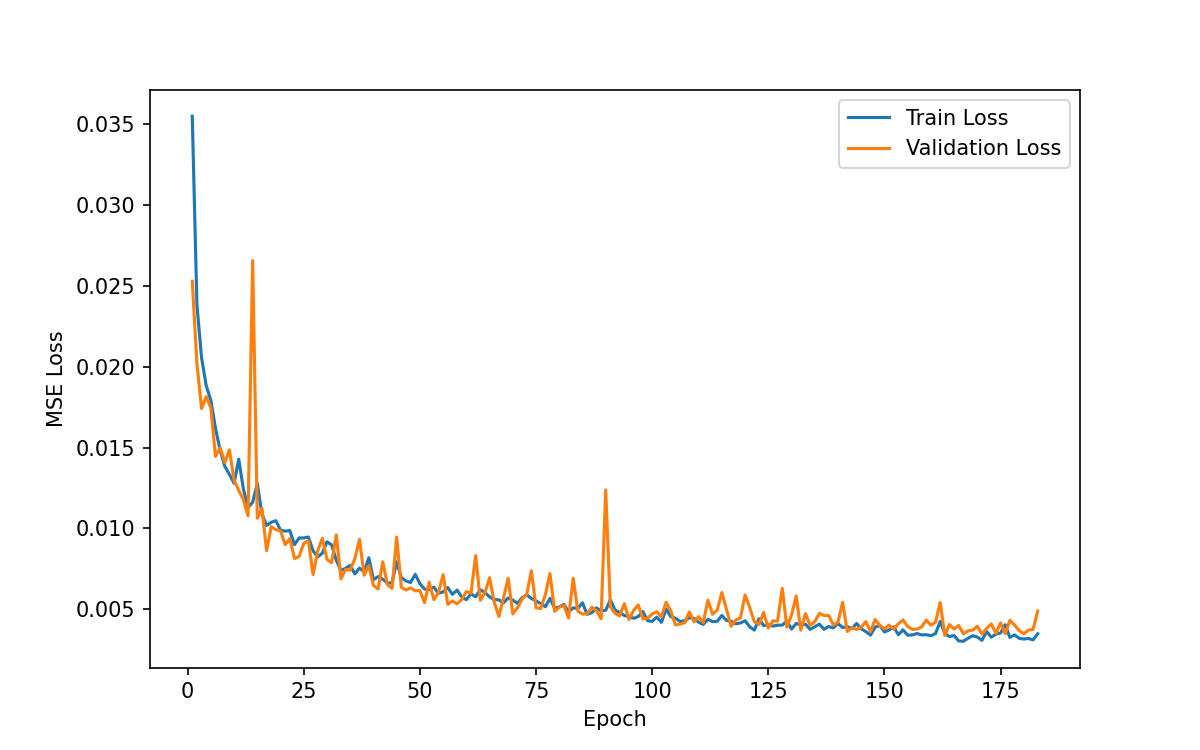
In my data assembly process, I first load data from two separate HDF5 files containing Type3 and Type4 samples, respectively—stacking all keys (except for 'q\_fixed') vertically to form complete datasets for each type. For each class, if the number of samples exceeds the desired count, I randomly down-sample the data; if it falls short, I augment the dataset by generating new samples via linear interpolation between randomly selected rows within the same class. Additionally, I generate a heterogeneous class by randomly selecting one sample from each of the two types and blending them using a random weight. Finally, I merge these three balanced datasets—assigning labels 0, 1, and 2 to Type3, Type4, and the heterogeneous data respectively—resulting in a unified dataset that is ready for model training and evaluation. The performance is shown below: (test accuracy: 93.50%)

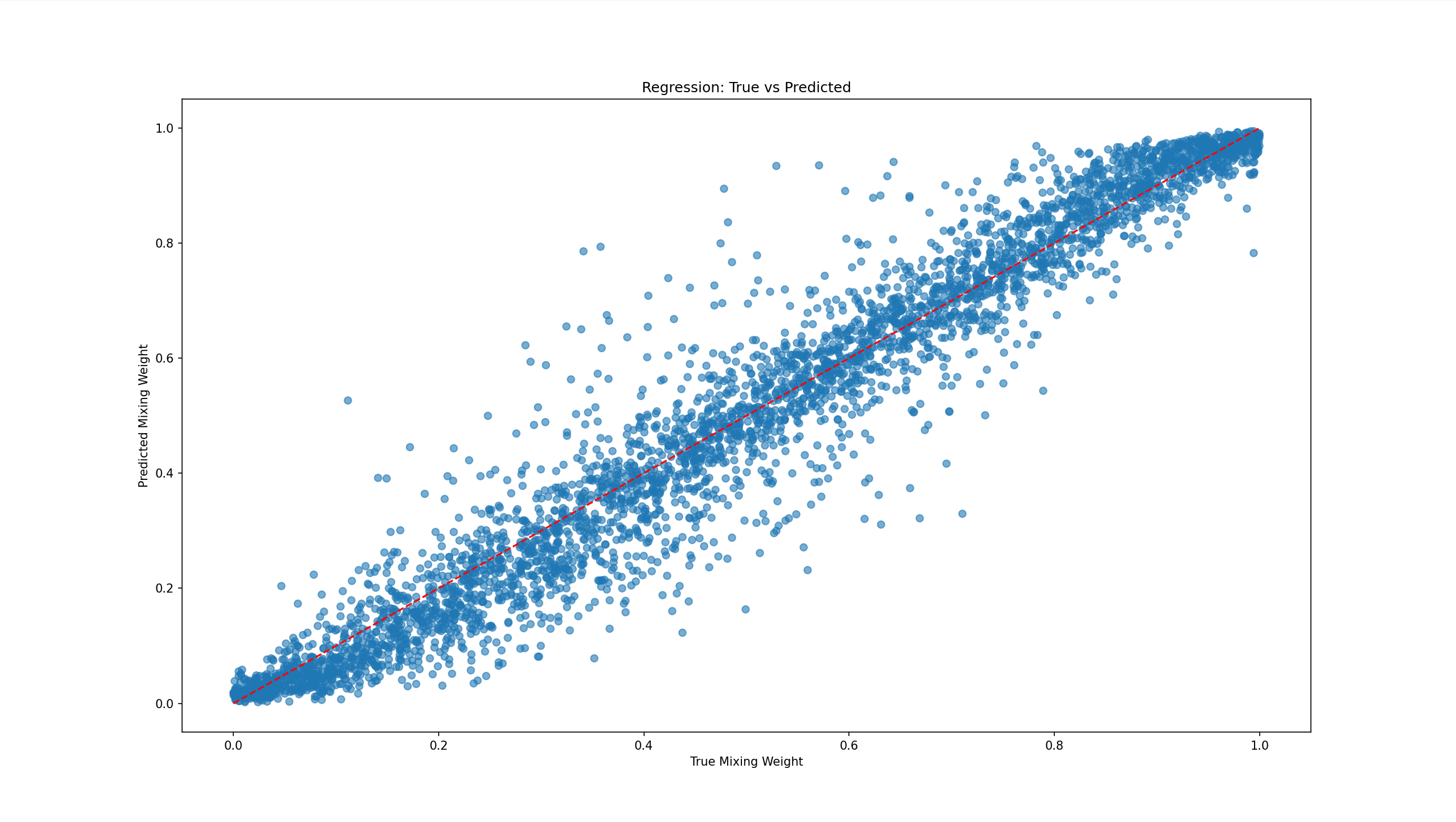




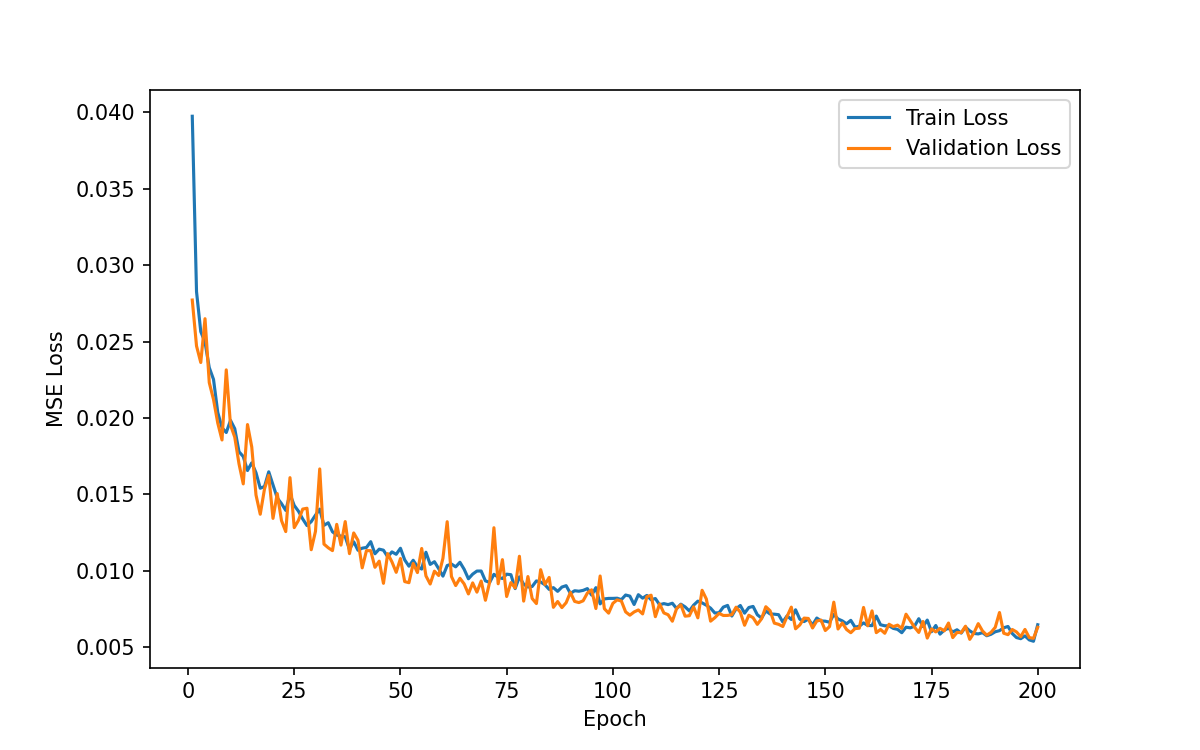
V3.2 Clean data

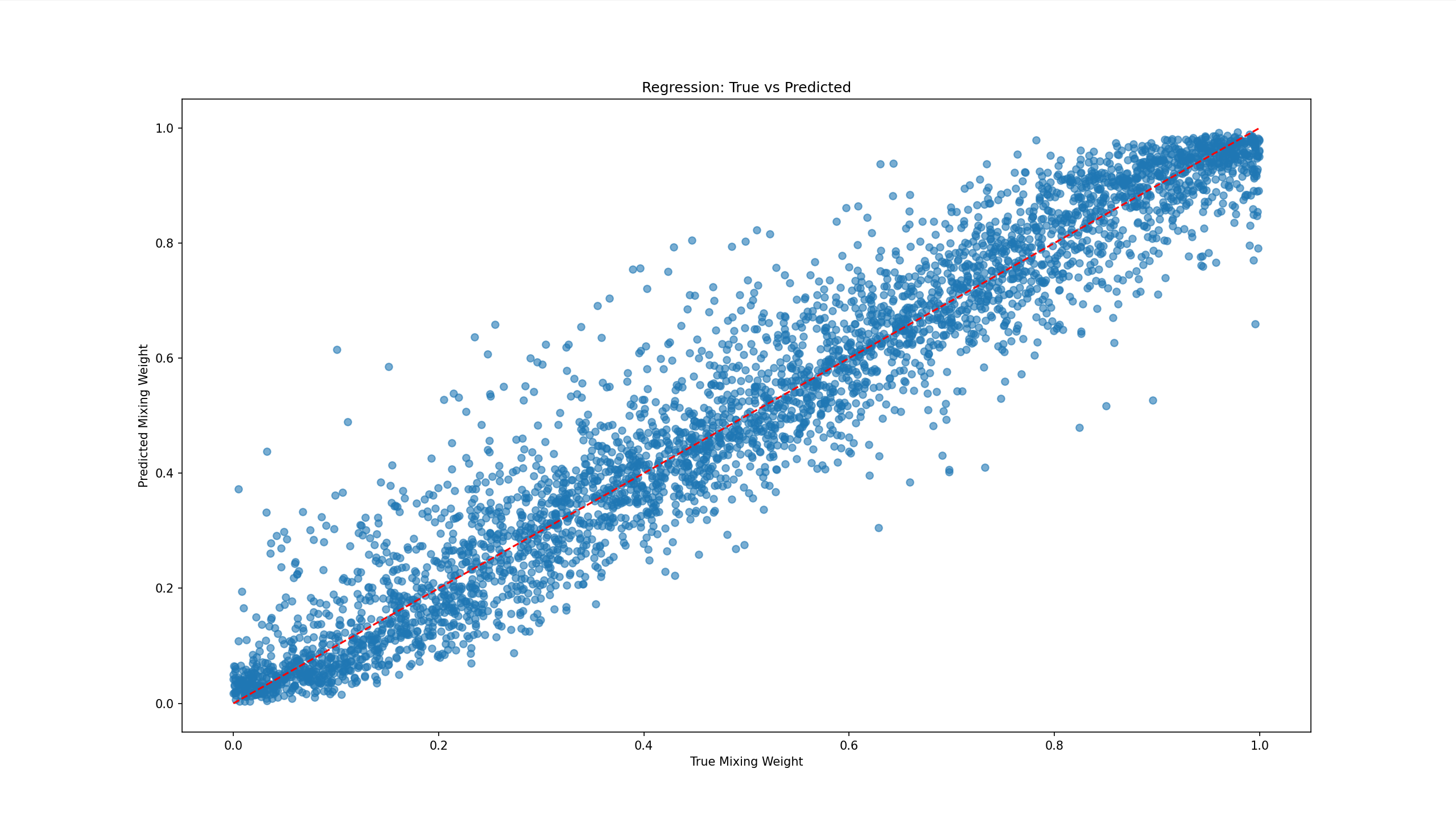
By observing the results shown above, we can basically assume the model can tell the difference between homogeneous LNPs and distinguish heterogeneous LNP solution among homogeneous one. Thus, we can try to predict the random weight (ratio) we used to synthesize the heterogeneous data. In this version, all the data is synthesized by applying random ratio. Therefore, the task is to predict the fraction of certain type of LNP in the heterogeneous solution. The regression performance is shown below (R2 score: 0.936470):





I also tested whether more points for interpolation will provide better performance. The model performance with 600 interpolation points is given below (R2: 0.920068):





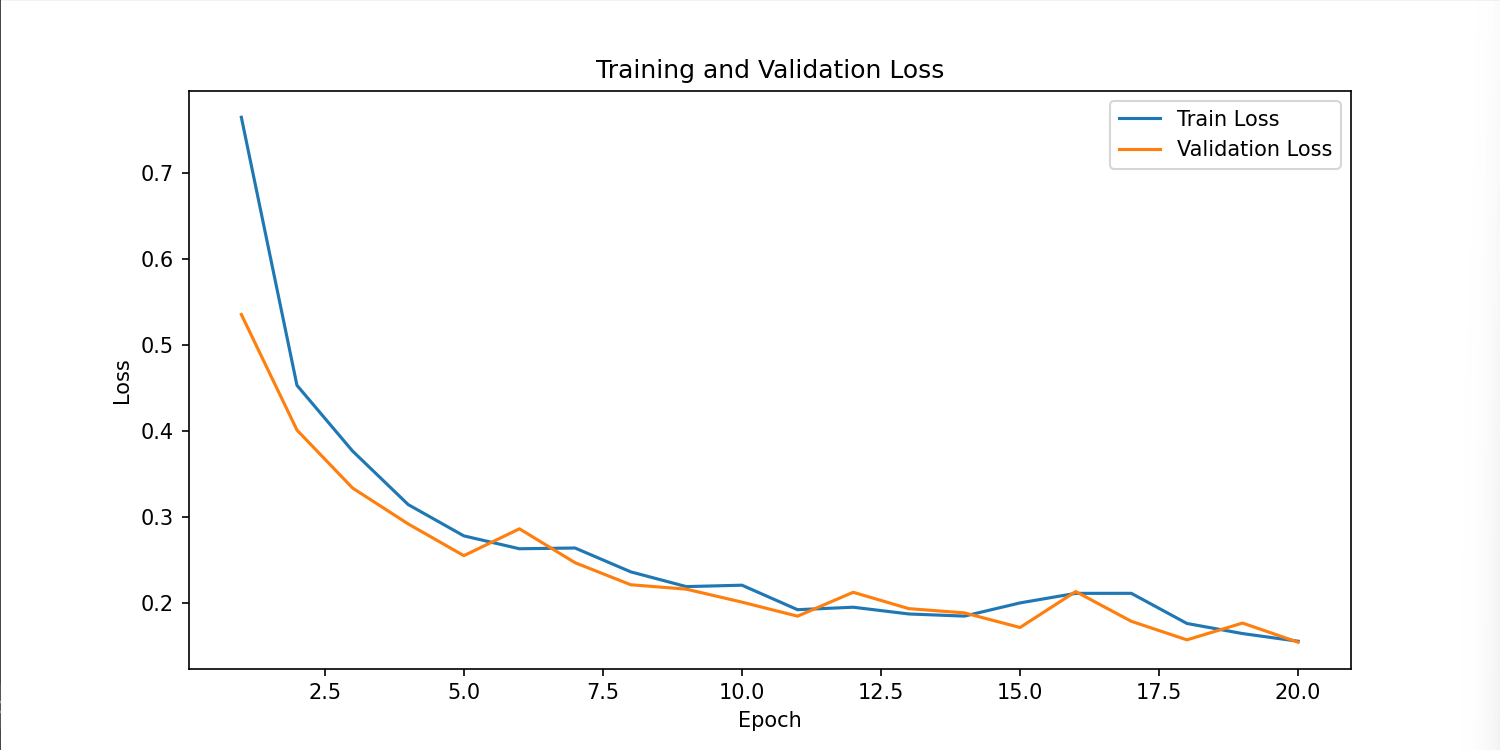
It can be seen that there is no big difference. Thus, the original 300 points scheme will be adopted.

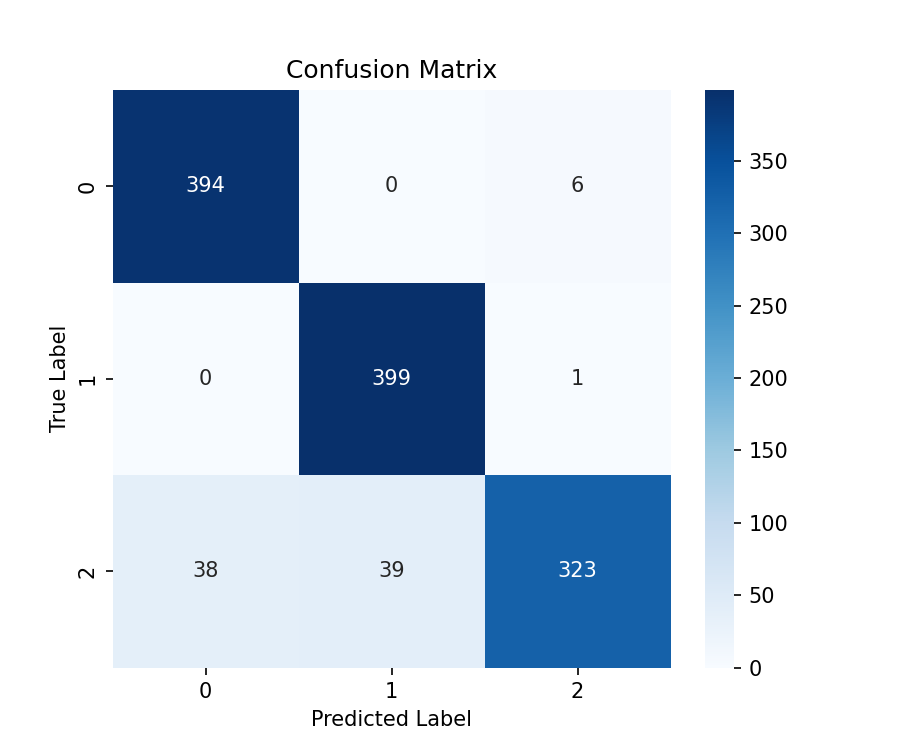
To further test the feasibility of this simulation method, according to [1], a noise model inspired by the Poisson distributed photon counting nature of the data acquisition. Moreover, similar to [2], the scattering intensity normalization can also be applied before the data is sent to the ML model. Different from a simple Gaussian noise applied to the scattering intensity, the lognormal noise model will not produce a negative scattering intensity value. In [2], the author mentioned “Note that whereas we account for measurement noise, we do not account for the finite instrument resolution which would yield a slight blurring of the SAXS curves.”

Then, repeat the experiments done in V3.1 and V3.2. The results are shown:

V3.1 modified data

In this experiment, all the model and hyperparameters remain the same at V3.1 with clean SAXS data. Test accuracy = 93%



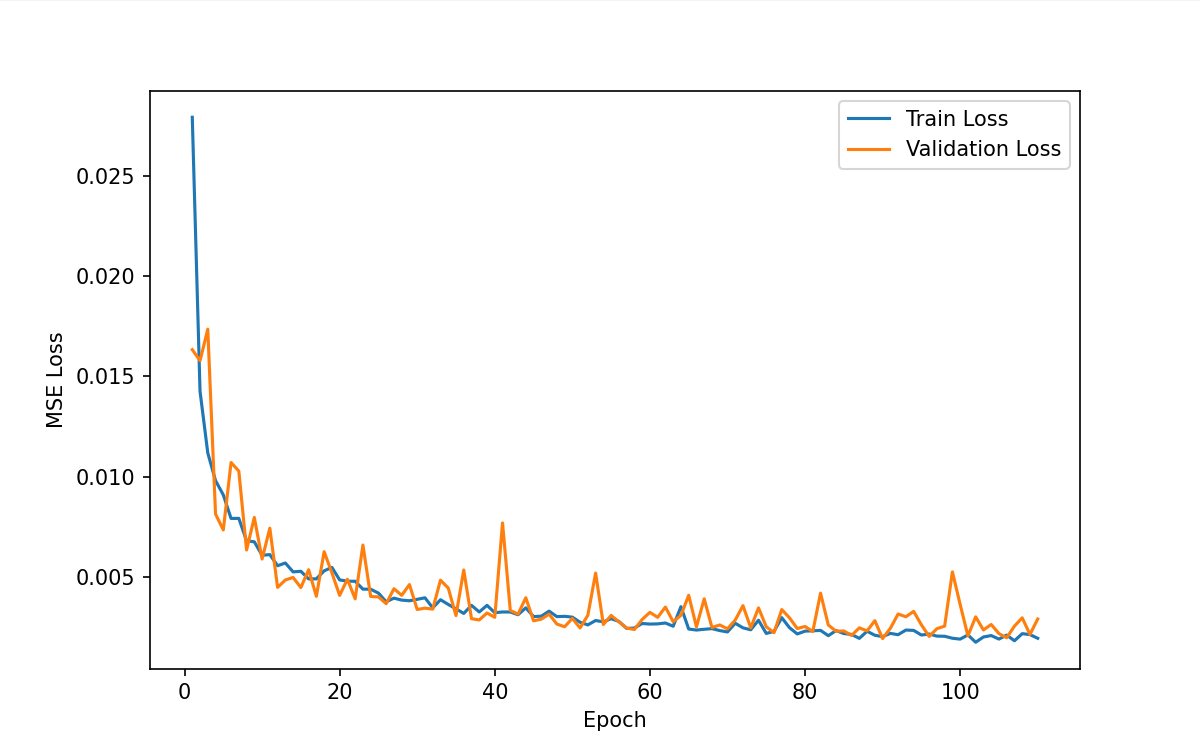


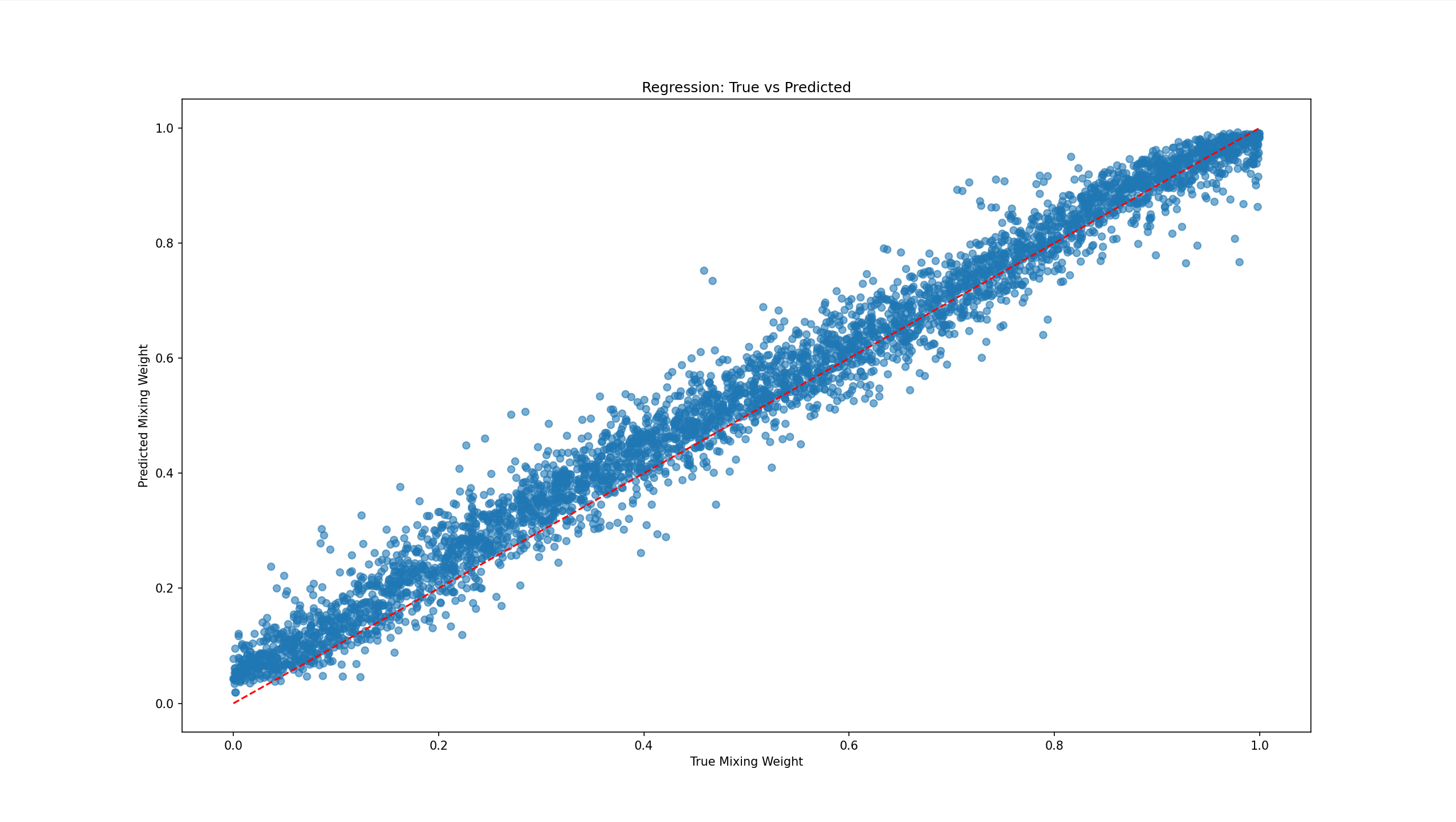
V3.2 Modified data

With all the parameters fixed, the model even provided a higher R2 score, 0.966008. To verify whether this result is only caused by randomness, several experiment results are recorded as listed below:

V3.2 Modified data: R2: 0.962928 / 0.979331 / 0.963520

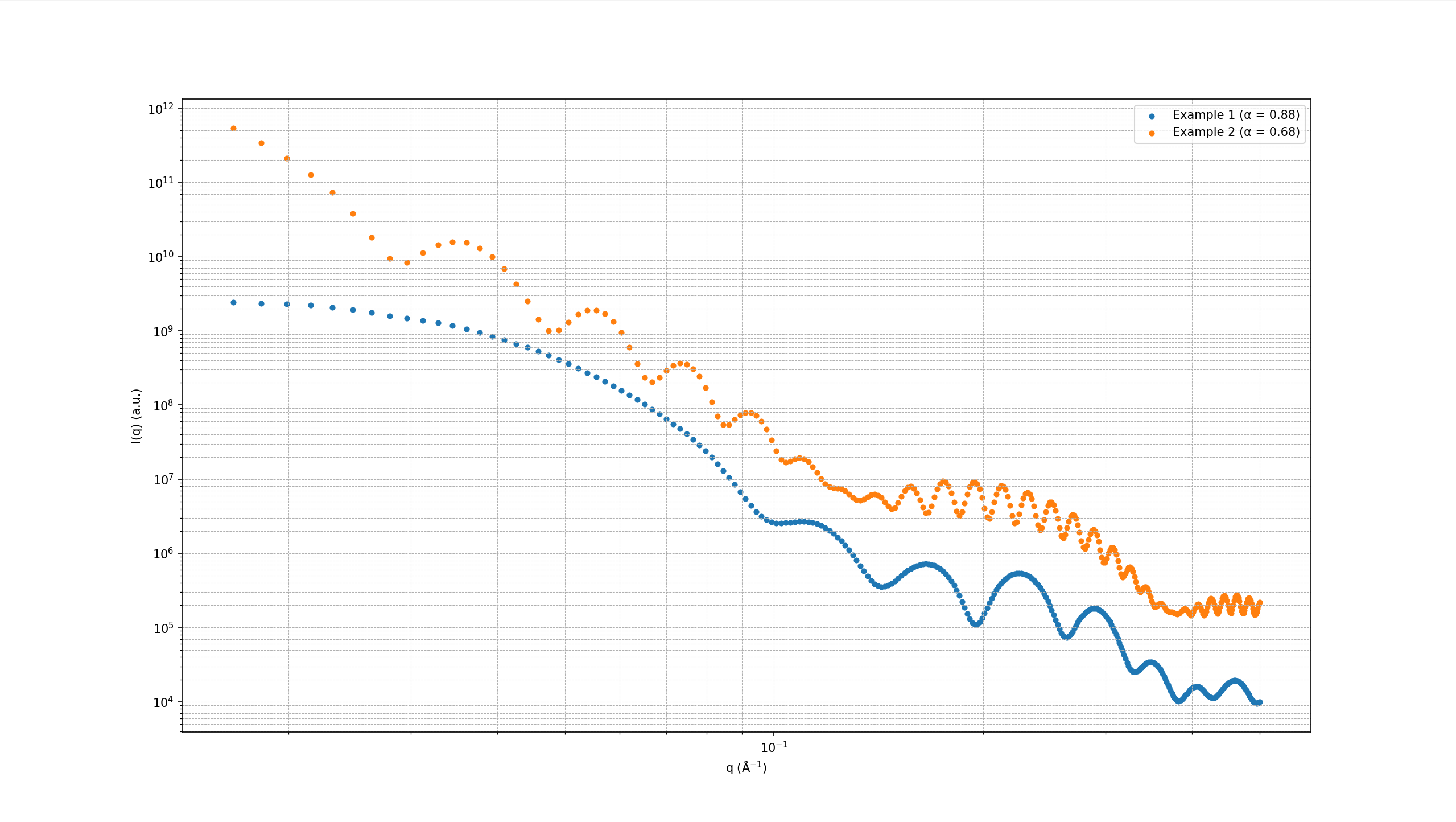
V3.2 Clean data: R2: 0.939722 / 0.932174 / 0.945876

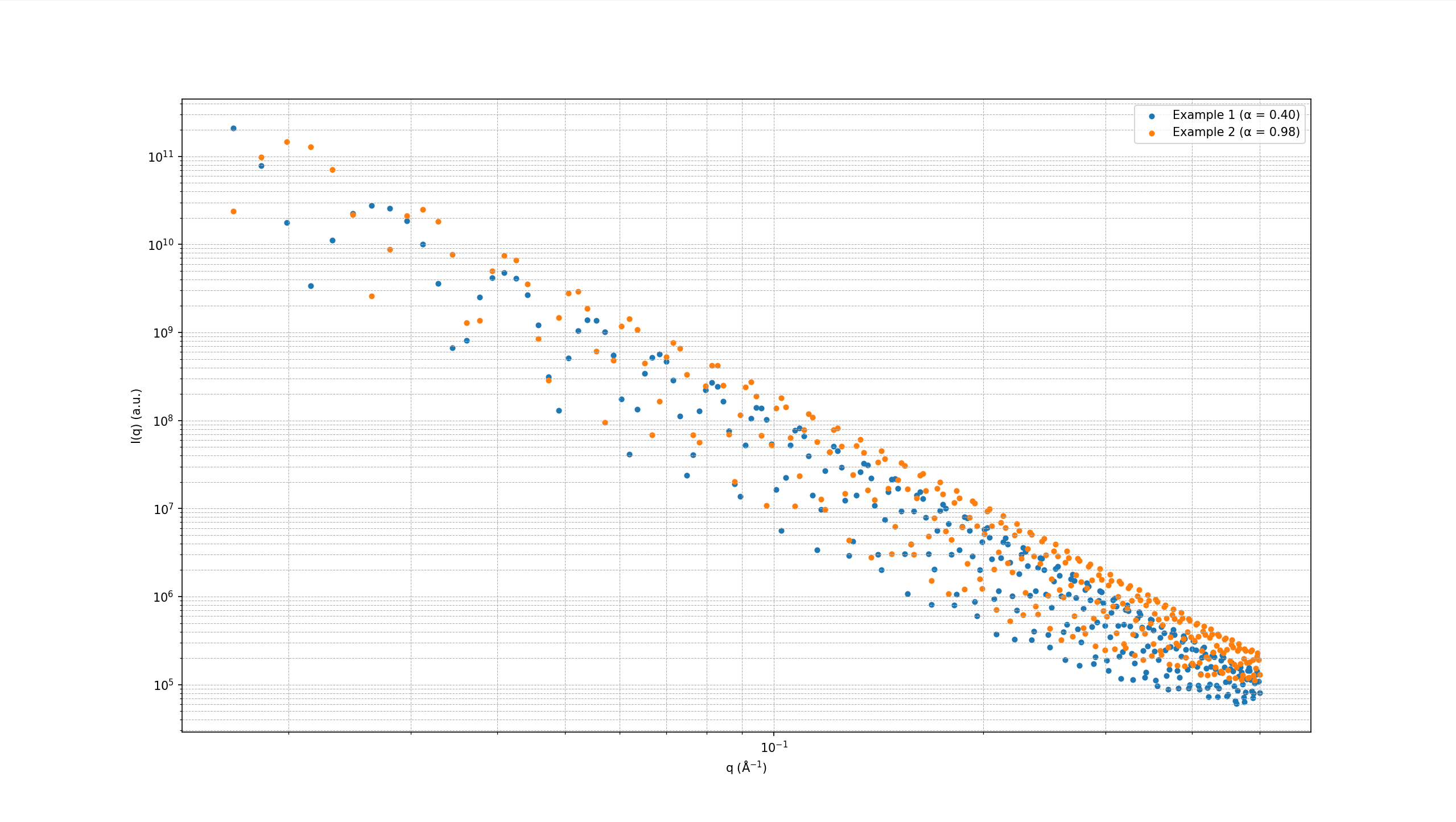




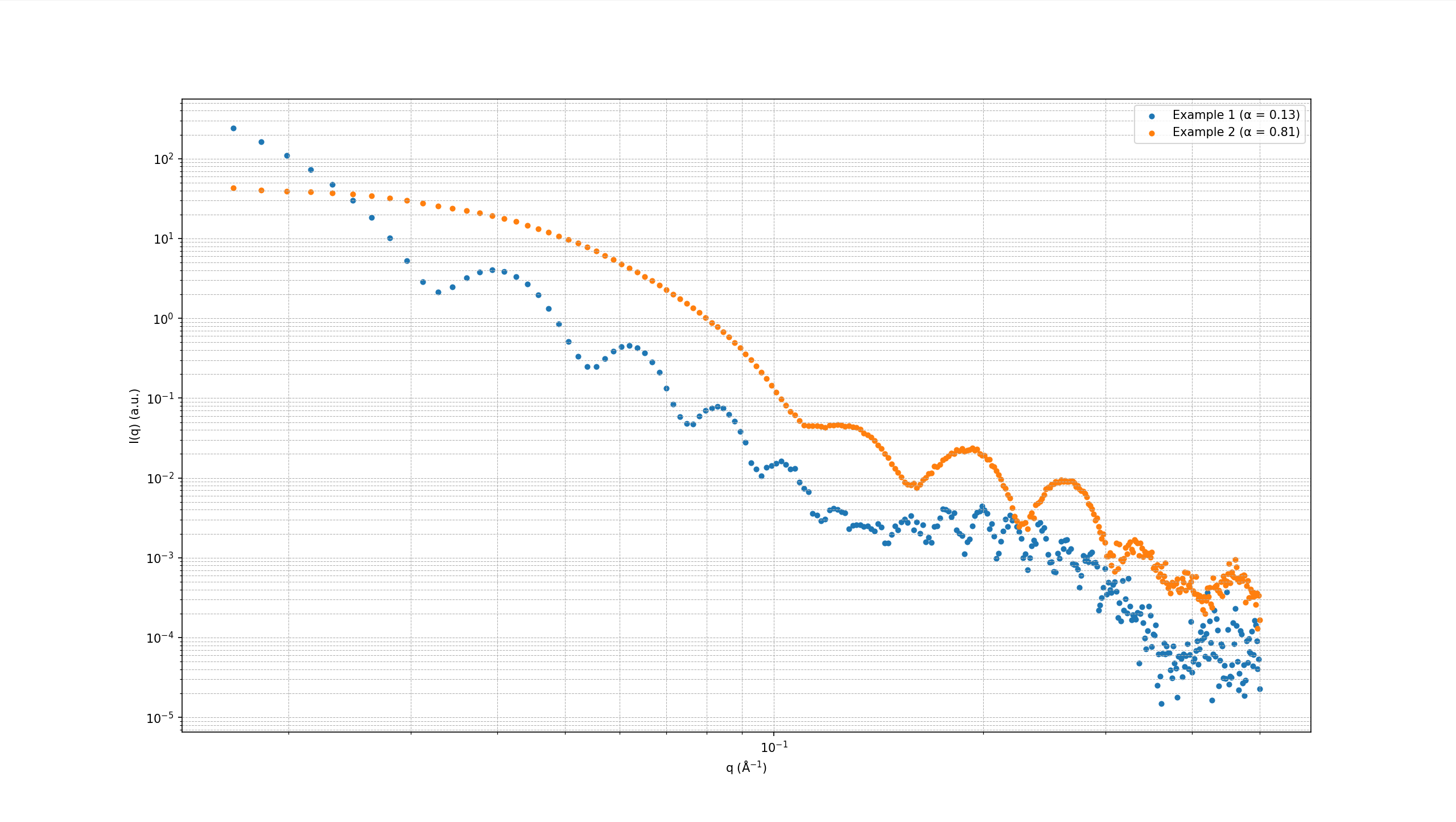
Example Clean/Modified data

Clean data

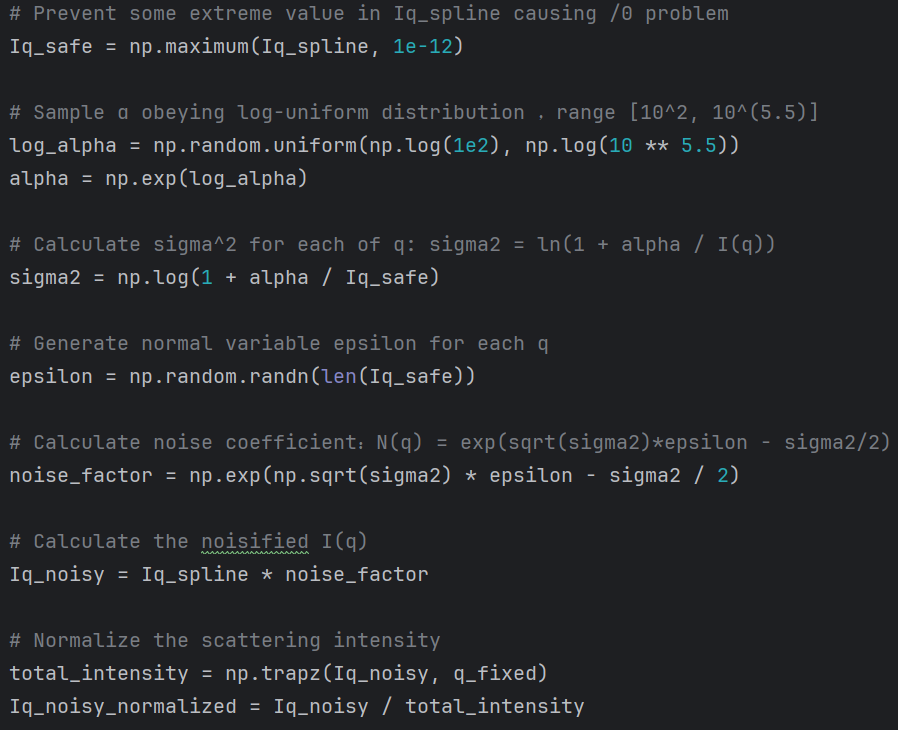




Modified data



As shown above, the SAXS curves are not modified to the extend of real-world experiment. Thus, I will try to modify the data to obtain more realistic results. The code for preprocessing (noisifying) is shown below:



[1] Predicting permeability via statistical learning on higher-order microstructural information.

[2] Machine learning-accelerated small-angle X-ray scattering analysis of disordered two- and three-phase materials