5-fold cross-validated performance on train set and test set

|  |  |  |  |
| --- | --- | --- | --- |
| **Performance metric** | **rmse** | **pearson\_r** | **kendall\_tau** |
| Train\_fold\_1\* | 1.344 | 0.486 | 0.302 |
| Train\_fold\_2 | 1.405 | 0.406 | 0.258 |
| Train\_fold\_3 | 1.619 | 0.477 | 0.189 |
| Train\_fold\_4 | 2.431 | 0.029 | 0.06 |
| Train\_fold\_5 | 1.597 | 0.172 | 0.102 |
| Test | 1.720 | 0.088 | 0.304 |

\*indicate selected fold for evaluating on test set.

The DTIGN model was rebuilt based on the original paper, aiming to replicate the architecture and practical deviations and design decisions were made:

* Cosine Alignment Loss Omitted:

In the original study, the native and aggregated posture representations are aligned using a cosine similarity loss. Because the dataset lacked native pose labels and structural references, this loss was not included in the calculation. As a results, the total loss for training was only calculated on bioactivity attention.

* Data Compatibility Fixes:

The .pyg files were incompatible with torch\_geometric.data.Batch, which led to missing “ptr” fields and erroneous pooling. To verify graph-level pooling worked as intended, we reconstructed the datasets using standard Data(...) objects and resaved them.

Observations During Training and Evaluation

The model converged slowly during early epochs, especially when the attention mechanism was improperly configured due to shape mismatches. After correcting data format and dimensionality issues, training loss decreased steadily and validation metrics (RMSE, Pearson r) became stable after a few epochs. In overall, folds 1–3 show **moderate positive correlations** and reasonably low RMSE while fold 4 performs **significantly worse,** suggesting **data quality issues or outlier bias**. The test RMSE is comparable to training folds, indicating reasonable generalization. However, the test Pearson r is low (0.088), suggesting the model has poor linear correlation with true pEC50 values. Despite that, Kendall’s tau is relatively better (0.304) meaning that model may preserve relative ordering rather than exact values.

In the absence of native pose supervision, the model still learned pose-aware representations, though with slightly lower peak performance than reported in the original paper.

Failure Modes and Limitations

Loss Mismatch Warnings:

When the model's output shape did not match batch.y due to pooling errors, PyTorch produced shape mismatch warnings. These were resolved after correcting ptr usage and ensuring graph-level batch structure.

Low test Pearson r indicates: Possible model underfitting or lack of discriminative power in learned graph embeddings. To handling this issue, some improvemts can be applied such as add BatchNorm, deeper MLP, or global pooling layer in DTIGN; Model parameters, such as learning rate, batch size and number of hidden dimensions can also still be optimized.