Oxford University Press., 2025, pp. 1–4

doi:
Paper

PAPER

Assessing the Reliability of AlphaFold3 Predictions for Protein-Ligand Affinity Prediction via Sfcnn

Guo Yu,[1] Yiming Wu[1] and Yiyang Tan[1]

[1]School of Information Science and Technology, ShanghaiTech University, 393 Middle Huaxia Road, 201210, Shanghai, China

Abstract

This study systematically evaluates the reliability of AlphaFold3 (AF3)-predicted protein structures for protein-ligand affinity (PLA) prediction tasks. The Sfcnn model, a 3D convolutional neural network (CNN) for PLA prediction, was reproduced using PyTorch. Model performance was validated on the PDBbind v2019 refined set for training and the CASF-2016 core set for testing. Subsequently, AF3-derived protein structures from the CASF-2016 core set were assessed and compared to experimentally determined structures using Sfcnn scores, to determine the suitability of AF3 predictions in PLA applications.

Key words: AlphaFold3, protein-ligand affinity, CNN scoring function, CASF-2016

Introduction

Sfcnn Overview

Sfcnn is a 3D convolutional neural network-based scoring function introduced by Wang et al. [3] in 2022, designed to provide accurate and reliable predictions of binding affinities for protein-ligand complexes.

Methods

Datasets

The Sfcnn network was trained using protein-ligand complexes from the PDBbind v2019 refined set [2], which includes experimentally determined binding affinities (pKa values). The model was evaluated on the CASF-2016 [1] core set, comprising 285 protein-ligand complexes. To prevent data leakage, 266 overlapping protein complexes between the training and test sets were excluded, resulting in 4,852 unique training complexes.

Data Augmentation

To increase the effective size of the training set, each protein-ligand complex was randomly rotated nine times using random rotation matrices, yielding ten variants per complex. All variants share the same PLA score, resulting in a total of 48,520 training samples.

Featurization

Protein-ligand complexes are represented as 3D grids of size 20 × 20 × 20, with each grid cell encoded as a one-hot vector of length 28. This vector comprises 14 protein atom types[1] and 14 ligand atom types. The resulting training tensor has shape $(48520, 20, 20, 20, 28)$.

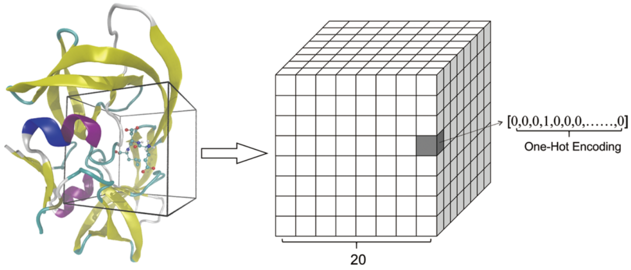

Fig. 1. Featurization of protein-ligand complexes. Example shown: PDB ID 1a30. Default resolution is 20 × 20 × 20 with 28 atomic categories.

Network Architecture

The original Sfcnn publication describes four network architectures and three featurization strategies. The architecture depicted in Figure 2, combined with the aforementioned featurization, achieved optimal validation performance.

[1] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04762-3

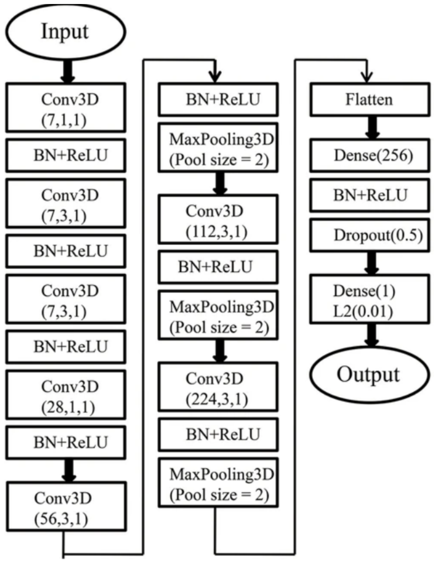

Fig. 2. Final CNN architecture for the Sfcnn network.

This architecture employs 3D convolutional layers with batch normalization and ReLU activation. L2 regularization is applied to the output layer to mitigate overfitting and enhance generalization.

Reproduction Study

Dataset and Featurization

The reproduction pipeline utilizes the same datasets and featurization approach, yielding a training tensor of shape $(48520, 20, 20, 20, 28)$ and a test tensor of shape $(285, 20, 20, 20, 28)$.

Data Storage

The original Sfcnn implementation stored data as concatenated arrays in a single `.pkl` (pickle) file, requiring all data to reside in memory, which is impractical for extremely large datasets. The format of `.h5` (HDF5) via `h5py` is adopted for incremental writing and efficient storage. The resulting training grid occupies 40.1 GiB.

Network Implementation

The PyTorch implementation closely mirrors the original TensorFlow version, with two key differences:

1. Due to PyTorch's `Conv3D` API, input tensors are permuted to shape (batch_size, 28, 20, 20, 20).
2. PyTorch lacks a direct L2 regularization API; instead, weight decay is applied to the final fully connected layer to approximate this effect.

Training Procedure

Training and validation sets are partitioned as in the original study, with the validation set comprising indices 41,000 to 48,520. The resulting splits are: training $(41,000, 20, 20, 20, 28)$, validation $(7,520, 20, 20, 20, 28)$, and test $(285, 20, 20, 20, 28)$, corresponding to a ratio of $84.00\% : 15.42\% : 0.58\%$. Note that the original hyperparameters did not yield convergence in our PyTorch experiments, both hyperparameters are summarized in the the following table.

Table 1. Original and Reproduced Hyperparameters

Parameter	Original	Reproduced
Learning rate	0.004	0.0015
Batch size	64	32
Dropout rate	0.5	0.15
L2 regularization / FC weight decay	0.01	0.01
Epochs	200	200

Reproduced Results

Evaluation Metrics

Sfcnn performance is evaluated using the following metrics:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{\text{predict}} - y_{\text{true}})^2}$$

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_{\text{predict}} - y_{\text{true}}|$$

$$\text{SD} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}((ay_{\text{predict}} + b) - y_{\text{true}})^2}$$

$$\text{R} = \frac{\mathbb{E}[(y_{\text{predict}} - \mu_{y_{\text{predict}}})(y_{\text{true}} - \mu_{y_{\text{true}}})]}{\sigma_{y_{\text{predict}}}\sigma_{y_{\text{true}}}}$$

where a and b are the slope and intercept of the linear regression between predicted and measured values, $\mathbb{E}[\cdot]$ denotes expectation, and μ and σ represent means and standard deviations, respectively.

Table 2. Performance Metrics on CASF-2016 Core Set

Metric	Reproduced Sfcnn	Original Sfcnn
Pearson R	0.7286	0.7928
RMSE	1.5481	1.3263
MAE	1.2579	1.0277
SD	1.4892	1.3252

Although the original Sfcnn reports superior metrics, its training process did not converge in the reproduction, raising concerns regarding the reproducibility of the reported performance. Due to the lack of access to the original training data and the absence of author response to data requests on GitHub[2], the original training process is deemed irreproducible. The training, validation, and testing results for the four metrics are presented in Figure 3 and Figure 4.

[2] https://github.com/bioinfocqupt/Sfcnn/issues/1

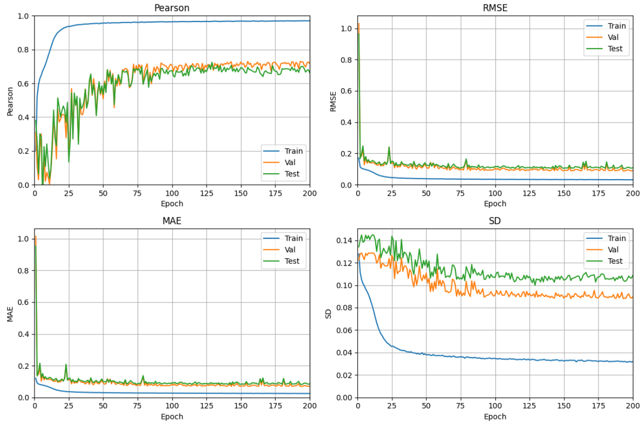

Fig. 3. Training curve for reproduced hyperparameters.

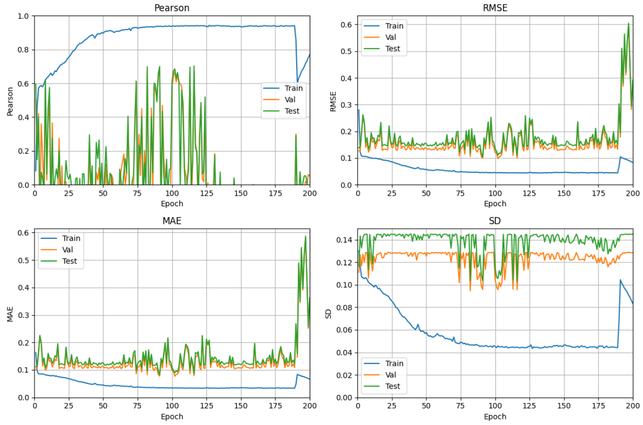

Fig. 4. Training curve for original hyperparameters.

Since the original hyperparameters significantly deviate from the norm and did not reach proper comvergence, the reproducted results are considered the baseline for subsequent AF3 result assessment.

AlphaFold3 Structure Assessment

Dataset

The assessment utilizes the CASF-2016 core set, excluding six proteins with structural complexity beyond AlphaFold3's predictive capacity, resulting in 279 proteins. Proteins with more than five isomorphic or heterogeneous chains were excluded, as detailed in Table 3.

Table 3. Excluded Complex Protein Structures

PDB ID	Number of Chains
2xb8	12
2ymd	10
3n76	12
3n7a	12
3n86	12
4ciw	12

Workflow

Structure Generation

Protein structures were generated using the *Chai-1 online server*[3]. The AlphaFold3 online server[4] was not used due to its inability to accept specific ligand SMILES codes. The MSA (Multiple Sequence Alignment) option was enabled with the MMseqs2 algorithm for all generations.

Docking and File Handling

Server outputs were downloaded as zip archives containing multiple ranked structures and associated metrics. The top-ranked structure (`pred.rank_0.cif`) was selected for analysis. To avoid conversion errors between file formats (.cif, .pdb, .mol2), structures were parsed directly using the `MMCIFParser` from Bio.PDB, followed by featurization and grid mapping. Atoms or isotopes not included in the 14 predefined atom types are categorized as `other`.

Scoring

Predicted structures were scored using the reproduced Sfcnn network, loaded with pre-trained weights (Pearson R = 0.728). Detailed analysis of PLA results and metrics is provided in the subsequent section.

Result Analysis

[Detailed analysis of the results should be inserted here.]

Conclusion

This study presents a systematic evaluation of AlphaFold3-predicted protein structures for protein-ligand affinity prediction using a reproduced Sfcnn model. The findings highlight the importance of reproducibility in deep learning models for structural biology and demonstrate the current capabilities and limitations of AF3 in the context of PLA prediction. [Further conclusions and implications should be added here.]

External Libraries

- **PyTorch**: Custom neural network implementation.
- **Pandas**: Data storage and analysis.
- **Numpy**: Data processing and manipulation.
- **Matplotlib**: Visualization of training loss curves.

Author Contributions

G.Y., L.T., and J.C. conceived, conducted, and analyzed the experiments, and contributed to manuscript preparation and review.

Acknowledgments

The authors thank the anonymous reviewers for their valuable feedback. This work was supported in part by the National Science Foundation (NSF: #1636933 and #1920920).

[3] `https://lab.chaidiscovery.com/dashboard`

[4] `https://alphafoldserver.com/`

References

1. Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913, 2019.
2. Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: Methodologies and updates. *Journal of Medicinal Chemistry*, 48(12):4111–4119, 2005.
3. Yu Wang, Zhengxiao Wei, and Lei Xi. Sfcnn: a novel scoring function based on 3d convolutional neural network for accurate and stable protein–ligand affinity prediction. *BMC Bioinformatics*, 23(1):222, 2022.