Oxford University Press., 2025, pp. 1–6

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, its 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

Background

AlphaFold3 (AF3), DeepMind's latest AlphaFold model, predicts protein and protein–ligand structures with high accuracy. It extends AlphaFold2 by adding explicit ligand modeling, enhanced multimer assembly support, and optimized multiple sequence alignments (MSAs) via deep neural networks trained on extensive sequence and structural data [1].

Sfcnn is a 3D convolutional neural network-based scoring function introduced by Wang et al. [4] in 2022, designed to provide accurate and reliable predictions of binding affinities for protein-ligand complexes.

Objective

The primary objective of this study is to evaluate the reliability of AlphaFold3-predicted protein-ligand complex structures for protein-ligand affinity (PLA) prediction. Using the Chai-1 server for AF3 structure generation, we utilized its support for custom ligands and robust MSA construction. The resulting AF3 structures are assessed with the reproduced Sfcnn model, and predicted affinities are compared to those from experimentally determined structures. This enables a direct evaluation of AF3's suitability for PLA prediction and highlights its current strengths and limitations.

Materials and Methods

Datasets

The Sfcnn network was trained using protein-ligand complexes from the PDBbind v2019 refined set[3], which includes experimentally determined binding affinities (pKa values). The model was evaluated on the CASF-2016 [2] 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 $20 \times 20 \times 20$, 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 $(48520, 20, 20, 20, 28)$.

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

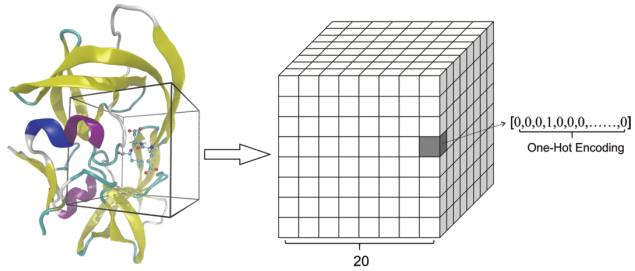

Fig. 1. Featurization of protein-ligand complexes. Example shown: PDB ID 1a30. Resolution is 20 × 20 × 20 with 28 atom types.

Sfcnn Network Architecture and Implementation

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.

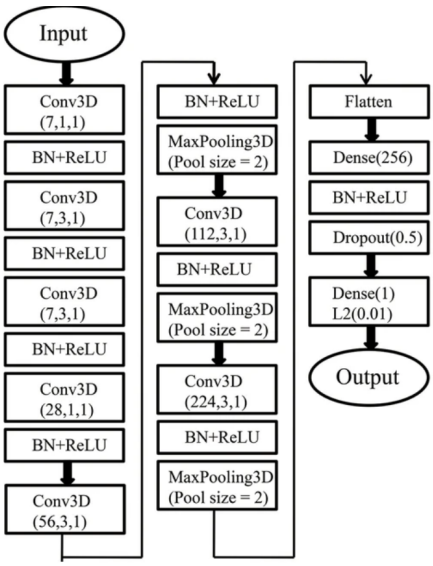

Fig. 2. Final CNN architecture for the Sfcnn network.

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.

Implementation Details

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.

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 HDF5 format (`.h5`) via `h5py` was adopted for incremental writing and efficient storage. The resulting training grid occupies 40.1 GiB.

Training Procedure

Training and validation sets are partitioned using a 7-fold cross-validation approach on the entire training dataset of 48,520 samples. The dataset was randomly shuffled and divided into 7 folds, maintaining consistent train-validation splits across experiments. For each fold, approximately 85% of the data (around 41,242 samples) is used for training and 15% (around 7,278 samples) for validation. The test set comprises 285 samples. The cross-validation framework enables comprehensive evaluation of model stability and generalization capability while providing statistical confidence intervals for performance metrics.

Note that the original hyperparameters did not yield convergence in our PyTorch experiments. Both sets of hyperparameters are summarized below.

Table 1. Original and Reproduced Hyperparameters

Parameter	Original	Reproduced
Learning rate	0.004	0.00068
Batch size	64	32
Dropout rate	0.5	0.15
L2 regularization / FC weight decay	0.01	0.01
Epochs	200	150
Train/validation split	85%/15%	85%/15%

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. Highest Performance Metrics on CASF-2016 Core Set

Metric	Reproduced Sfcnn	Original Sfcnn
Pearson R	0.7678	0.7928
RMSE	1.4647	1.3263
MAE	1.1633	1.0277
SD	1.3928	1.3252

The 7-fold cross-validation yielded a mean test Pearson correlation of 0.7375 ± 0.0145 and the highest Pearson correlation of 0.7678 on the CASF-2016 core set, as shown in Table 2, demonstrating consistent model performance across different data partitions and confirming the reliability of our reproduced implementation.

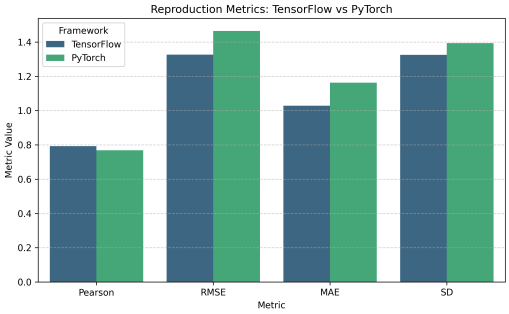

Fig. 3. Highest reproduced metrics.

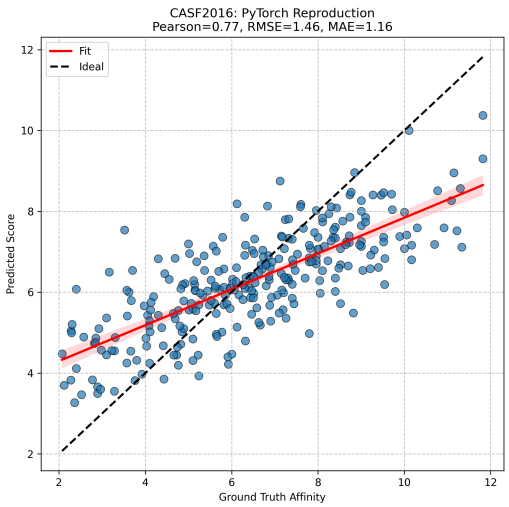

Fig. 4. Prediction scatter plot of the best reproduced result.

Although the original Sfcnn reports superior metrics, its training process **did not converge** during reproduction; our cross-validation results also raise concerns regarding the reliability of the original Sfcnn result. Due to the lack of access to the original training data and the absence of author responses to data requests on GitHub[2], the original training process is deemed irreproducible.

The training and validation curves for the four metrics are presented in Figure 5 and Figure 6.

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

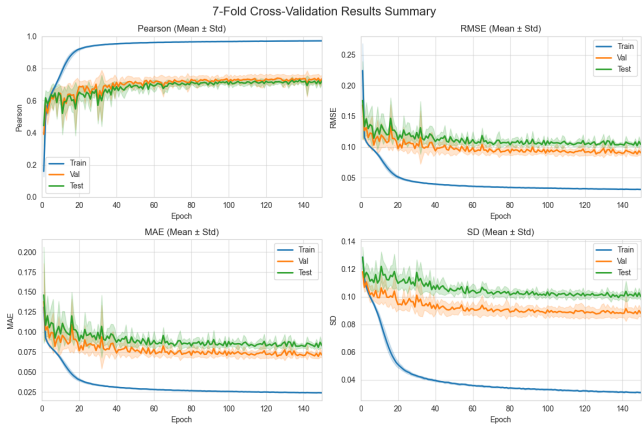

Fig. 5. Training curve for reproduced hyperparameters.

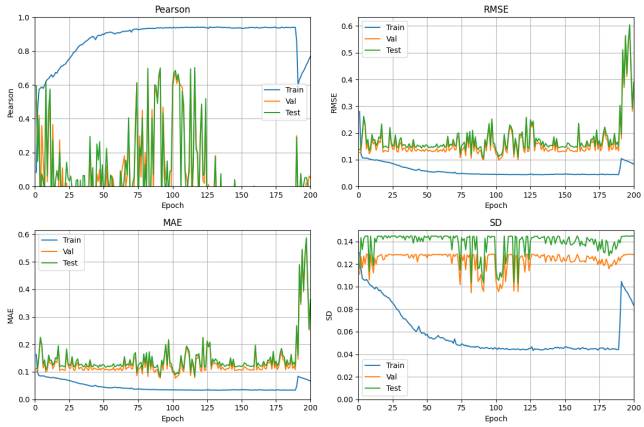

Fig. 6. Training curve for original hyperparameters.

The residuals of the test set (gaps between predicted and true values) are visualized in Figure 7. The histogram approximately follows a normal distribution, indicating that the model's predictions are generally unbiased, with most errors concentrated around zero.

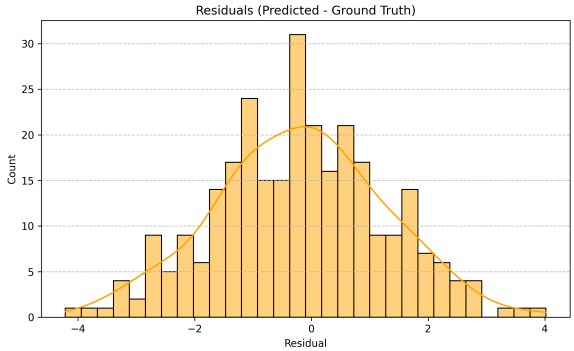

Fig. 7. Test residual histogram

We conjectured that the divergent results are caused by unusually **high learning rate** and **dropout rate**; therefore, the reproduced results are considered the baseline for subsequent AF3 result assessment.

AlphaFold3 Result Assessment

Dataset Selection

The assessment utilizes the CASF-2016 core set, excluding 6 proteins with structural complexity beyond AlphaFold3's predictive capacity, resulting in a subset of 279 proteins for this assessment. 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

Structure Generation and Processing

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.

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 the `Bio.PDB` Python package, followed by featurization and grid mapping.

Atoms or isotopes not included in the 14 predefined atom types are categorized as `other`.

Scoring Protocol

To quantitatively assess the reliability of AlphaFold3 (AF3)-predicted structures for protein-ligand affinity (PLA) prediction, we evaluated the predicted complexes using the reproduced Sfcnn network, initialized with the best-performing weights from 7-fold cross-validation (Pearson $R = 0.7678$). The experimentally determined PLA values from the CASF-2016 core set served as the ground truth. For benchmarking, the predicted scores for AF3-generated structures are compared against both the ground truth and the Sfcnn scores obtained from experimentally resolved structures, employing the same evaluation metrics as in Section 4.

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

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

Table 4. Performance metrics for Sfcnn AF3-predicted structures compared to Sfcnn CASF-2016 and ground truth.

Metric	vs. Groundtruth	vs. CASF2016 Sfcnn
Pearson R	0.2930	0.3620
RMSE	2.0836	1.2543
MAE	1.6933	0.9962
SD	2.0825	1.1761

As summarized in Table 4, the use of AF3-generated structures impacts the Sfcnn model's performance against the CASF2016 Sfcnn. The Pearson correlation decreases to 0.3620, compared to 0.7678 achieved with experimentally determined structures (Table 2).

The RMSE and MAE values also reflect a decline in predictive accuracy against the ground truth when using AF3 structures. Furthermore, Sfcnn predictions derived from AF3 structures show a Pearson correlation of only 0.2930 with experimental structures (i.e., "vs. Groundtruth" in Table 4), with an RMSE of 2.0836 and MAE of 1.6933 in this comparison, indicating considerable deviation.

Residual Analysis and Visualization

To further dissect the sources and distribution of prediction errors, a series of visual analyses are presented below:

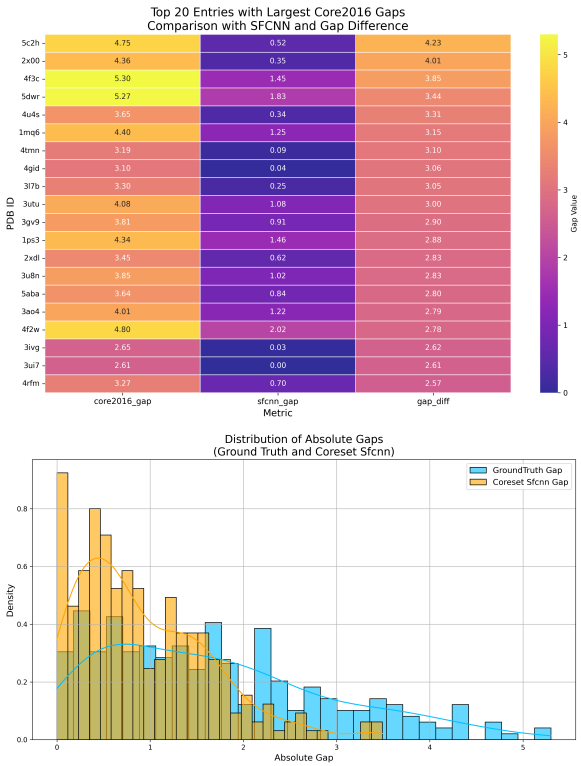

Fig. 8. Heatmap (left) of the top 20 complexes with the largest prediction gaps and histogram (right) of absolute prediction gaps for AF3 structures.

Figure 8 details the distribution and magnitude of prediction errors. The heatmap identifies complexes exhibiting the most significant discrepancies when comparing Sfcnn scores from AF3-predicted structures against ground truth PLA

values and against scores from experimentally-derived coreset structures. These discrepancies highlight instances where AF3-based predictions deviate substantially within the Sfcnn PLA framework, with pronounced differences predominantly falling within the 2.5-4.2 range. The histogram of residuals from AF3 Sfcnn predictions, when compared against both ground truth and coreset Sfcnn scores, displays a right-skewed distribution. This skewness signifies that a considerable number of complexes are associated with large prediction errors.

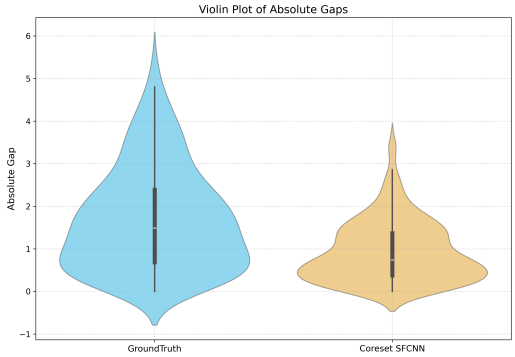

Fig. 9. Violin plot of prediction gaps for AF3 structures.

Figure 9 offers a complementary view through a violin plot, which depicts the complete error distribution of AF3 Sfcnn predictions relative to both ground truth and coreset Sfcnn scores. This visualization highlights that the majority of absolute prediction gaps are concentrated within the 0-2 range, while also corroborating the existence of larger deviations previously identified.

Interpretation and Implications

Collectively, these findings indicate that although the reproduced Sfcnn model achieves moderate concordance with experimental affinities on the CASF-2016 core set, its predictive efficacy diminishes when utilizing AF3-predicted structures. The observed alterations in performance metrics, notably the diminished Pearson correlation against ground truth, suggest that contemporary AF3 structural models introduce a greater degree of uncertainty into affinity prediction workflows.

The top 10 error complexes and their corresponding protein structures and Chai-1 overall confidence scores are presented in Table 5.

Table 5. Top 10 complexes with the largest prediction gaps. 'heter' denotes heterogeneous, 'iso' denotes isomorphic.

PDB ID	Aggregate Score	Prediction Gap	chain structure
5c2h	0.95	4.23	1 chain
2x00	0.93	4.01	5 heter
4f3c	0.97	3.85	2 iso
5dwr	0.90	3.44	1 chain
4u4s	0.89	3.31	2 heter
1mq6	0.97	3.15	2 heter
4tmn	**0.63**	**3.10**	**2 iso**
4gid	0.96	3.06	1 chain
3l7b	0.93	3.05	2 iso
3utu	0.97	3.00	2 heter

Statistics of these complexes show no significant correlation with protein structures, suggesting that AF3's performance degradation is not attributable to difficulties in handling specific types of complex structures. However, the overall high aggregate scores raise concerns about the reliability of AF3's confidence scores in the context of PLA prediction.

Discussion

Hypotheses for AF3's Underperformed Results

The observed decrease in PLA prediction accuracy for AF3-generated structures can be attributed to several methodological limitations and training dataset constraints.

Architectural and Scoring Limitations

AF3's Diffusion Transformer architecture lacks explicit physics-based energy calculations, resulting in minimal correlation between ranking metrics and experimental binding affinities. Unlike physics-based methods that explicitly model electrostatic interactions, van der Waals forces, and solvation effects, AF3's learned representations may fail to capture subtle energetic differences that distinguish strong from weak binders. AF3's confidence scores, which correlate with structural accuracy rather than binding strength, further compound this limitation by providing misleading assessments of prediction quality for affinity-related tasks. Additionally, the model's single-shot prediction approach cannot account for the ensemble of conformational states that contribute to experimental binding affinities, particularly for flexible protein-ligand systems where induced-fit mechanisms play crucial roles.

Training Data Constraints and Generalization Deficiencies

AF3's performance is constrained by training data limitations and poor generalization capabilities. A significant performance decline on structures released after the training cutoff date suggests memorization rather than true physical understanding of molecular interactions. Additionally, AF3 exhibits a persistent bias toward active GPCR conformations regardless of ligand type and performs poorly on ternary complex prediction. These limitations indicate that AF3 requires complementary approaches to address deficiencies in conformational sampling, affinity ranking, and complex system modeling. Recent work by Zheng et al. [5] suggests that optimal strategies involve integration into hybrid computational pipelines combining AI-based prediction with physics-based refinement and experimental validation, with enhanced sampling techniques showing promise for overcoming current limitations.

Conclusion

This study presents a systematic evaluation of AlphaFold3-predicted protein structures for protein-ligand affinity prediction using a reproduced Sfcnn model. Our results reveal significant performance degradation when using AF3-generated structures compared to experimentally determined structures, with Pearson correlation against ground truth decreasing from 0.7678 to 0.3850. This was accompanied by substantial increases in RMSE and MAE, indicating reduced predictive accuracy.

The findings highlight critical limitations of current AF3 models for affinity prediction tasks, primarily attributed to the absence of physics-based energy calculations and training dataset constraints biased toward static crystallographic

structures. While AF3 demonstrates remarkable capabilities in structural prediction, its direct application to binding affinity estimation requires careful consideration of these methodological limitations.

The importance of reproducibility in deep learning models for structural biology is underscored by our inability to reproduce the original Sfcnn results, emphasizing the need for transparent reporting and accessible training protocols. Future developments should focus on integrating physics-based scoring functions with structure prediction models and incorporating dynamic conformational information to enhance reliability for drug discovery applications.

External Libraries

- **PyTorch**: Custom neural network implementation and model training.
- **NumPy**: Numerical operations and data manipulation, especially for array handling.
- **Pandas**: Data analysis and manipulation, potentially for initial data handling (kept as per existing document).
- **scikit-learn**: Machine learning utilities, including K-fold cross-validation, linear regression models, and evaluation metrics.
- **H5py**: Reading and writing HDF5 files for efficient storage of large datasets (e.g., featurized grids).
- **OpenBabel (pybel)**: Parsing and processing molecular file formats (PDB, MOL2) and feature extraction.
- **Matplotlib**: Generating static plots, such as training/validation curves and metric visualizations.
- **Seaborn**: Creating informative statistical graphics, enhancing Matplotlib plots.
- **tqdm**: Displaying progress bars for iterative processes like training epochs and data loading.

Personal Reflections: Guo Yu

Contributions

- Designed and implemented the data pipeline, including dataset curation, preprocessing, featurization, data augmentation, and storage.
- Managed exclusion of overlapping complexes and ensured data compatibility with the reproduced network.
- Designed and maintained the AF3 generation-evaluation workflow, implemented K-fold cross-validation.

Problems Encountered

- Extremely large memory consumption issues arose with the original data storage and retrieval methods, leading to the adoption of HDF5 format for efficient handling of large datasets.
- K-fold cross-validation required extensive training time, necessitating experimentation with optimization techniques such as learning rate scheduling and mixed precision training to balance computational efficiency with model performance.

Acknowledgments

This work was completed as the final project for CS177: Bioinformatics—Software Development and Applications at ShanghaiTech University. The authors express their gratitude to the course instructors and teaching assistants for their valuable guidance and support throughout this project.

We acknowledge the assistance of GitHub Copilot, an AI-powered programming assistant, which provided support in several key areas of this research:

- Hyperparameter optimization recommendations for the Sfcnn training process
- Technical insights into AlphaFold3's diffusion model architecture
- Insights in visualization techniques for model performance metrics

References

1. Dunger J. et al. Abramson J., Adler J. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.
2. 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.
3. 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.
4. 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.
5. Haiyang Zheng, Hanfeng Lin, Adebowale A. Alade, Jingjing Chen, Erika Y. Monroy, Min Zhang, and Jin Wang. Alphafold3 in drug discovery: A comprehensive assessment of capabilities, limitations, and applications. *bioRxiv*, 2025.