On fine-tuning Boltz-2 for protein-protein affinity prediction

James King, Lewis Cornwall, Andrei Cristian Nica, James Day, Aaron Sim, Neil Dalchau, Lilly Wollman, Joshua Meyers
Synteny
London, UK

Abstract

Accurate prediction of protein–protein binding affinity is vital for understanding molecular interactions and designing therapeutics. We adapt Boltz-2, a state-of-the-art structure-based protein–ligand affinity predictor, for protein–protein affinity regression and evaluate it on two datasets, TCR3d and PPB-affinity. Despite high structural accuracy, Boltz-2-PPI underperforms relative to sequence-based alternatives in both small- and larger-scale data regimes. Combining embeddings from Boltz-2-PPI with sequence-based embeddings yields complementary improvements, particularly for weaker sequence models, suggesting different signals are learned by sequence- and structure-based models. Our results echo known biases associated with training with structural data and suggest that current structure-based representations are not primed for performant affinity prediction.

1 Introduction

Protein–protein interactions (PPIs) underpin nearly all cellular processes, and accurate prediction of their binding affinities is essential for understanding molecular mechanisms and guiding therapeutic design [1]. Computational methods for affinity prediction are increasingly critical, offering scalable alternatives to experimental approaches [2]. This is especially true in the development of immunomodulatory bispecifics, precise modelling of T cell receptor (TCR)–peptide–MHC and antibody–antigen interactions is particularly important, as small changes in affinity can determine clinical efficacy and safety [3].

Recent advances in machine learning have transformed multimeric protein structure prediction, exemplified by AlphaFold-Multimer [4] and its successor AlphaFold 3 (AF3) [5]. Open-source implementations such as Boltz [6] have broadened accessibility, enabling large-scale exploration of protein–protein complex structures. Although work remains to predict challenging interfaces such as those mediating antibody or TCR target recognition, these tools represent substantial progress and lay the foundation for affinity prediction and downstream therapeutic applications.

2 Background and Related Work

Previous methods for predicting protein-protein affinity have taken advantage of both sequence-based features and structure-based representations [2]. Standardised benchmark datasets such as PPB-affinity [7] and the more recent PPB-affinity (filtered) [8] provide an invaluable resource for systematically training and evaluating such methods. Related developments in protein–ligand affinity prediction, such as Boltz-2 [9], demonstrate the potential of modern machine learning architectures, motivating their adaptation and extension to the protein–protein affinity setting.

Structure-based affinity prediction models, however, have been shown to exploit dataset biases by memorising entities or pocket geometry [10] while sequence-based models risk learning global se-

quence similarity rather than binding determinants [11]. In both cases, careful mitigation strategies must be employed to ensure models capture interactions, not artifacts.

In this work, we make minimal modifications to Boltz-2 to permit training of a protein-protein affinity model. We show using two datasets that training an structure-based affinity model is subject to pitfalls introduced by aspects of the data and contrast this with sequence-based alternatives. We discuss the hurdles that must be leapt in order to train a useful structure-based affinity predictor.

3 Datasets

3.1 TCR3d

TCR3d 2.0 [12] comprises TCR-pMHC complex structures curated from the Protein Data Bank (PDB). A subset of 251 of these are both complexed with class-I or class-II MHC molecules, and associated with affinity measurements (dissociation constant, K_d) obtained via surface plasmon resonance (SPR) or other biophysical techniques. We transform K_d values to pK_d and exclude three entries for which the TCR binding mode is peptide-agnostic (9eji, 9ejg and 9ejh).[1]

We design hard splits between our train, validation, and test splits using the algorithm presented in Appendix A. In this way, similar sequences are discouraged from appearing in different splits.

3.2 PPB-affinity

The PPB-affinity (filtered) dataset [8] describes 8,207 unique protein-protein interaction entries with binding affinity measurements (K_d). This dataset is a filtered version of the original PPB-affinity dataset [7] which collates varied PPI data from multiple sources. Each entry is associated with a PubMed ID (PMID) indicating the source of the affinity experiment. The filtered dataset resolves annotation inconsistencies by removing duplicate entries from multi-chain protein interactions. The *PPB-affinity* dataset contains many different types of proteins and therefore allows a more general model to be trained, considering protein-protein affinities beyond TCR-pMHC complexes.

We reuse splits described by the authors [8], which were generated by implementing a $\leq 30\%$ sequence identity threshold to split proteins into training, validation, and test sets. We reuse baseline results from this paper directly, these are sequence-based approaches which operate directly on protein sequences of the ligand and receptor to learn an association with affinity.

4 Methods

4.1 Adapting Boltz-2 for protein-protein affinity prediction

Boltz-2 extends Boltz-1 [13] by improving structure prediction and, crucially for this work, adding an affinity module trained to predict protein-ligand affinity. The single and pairwise Pairformer representations are passed through an *affinity module*, which consists of a further distance-conditioned Pairformer stack, followed by cross-pair pooling and MLP readouts for affinity prediction.

The Boltz-2 affinity module is trained on protein-ligand interactions. In order to adapt the affinity module for protein-protein interactions, we make the following modifications.

1. **Loss function.** We define the optimisation objective to be

$$L(y, \hat{y}) \coloneqq \lambda L_{\text{Huber}}(y, \hat{y}) + (1 - \lambda) L_{\text{rank}}^{\text{PMID}}(y, \hat{y}),$$

where y and $\hat{y}$ denote a batch of true and predicted affinities, respectively. L_{Huber} is the Huber loss,

$$L_{\text{Huber}}(y, \hat{y}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{1}{2}(y_i - \hat{y}_i)^2 & \text{if } |y_i - \hat{y}_i| \leq \delta, \\ \delta |y_i - \hat{y}_i| - \frac{1}{2}\delta^2 & \text{otherwise,} \end{cases}$$

[1] as of October 1, 2025

and $L_{\text{rank}}^{\text{PMID}}$ is a ranking loss between pairs within a batch

$$L_{\text{rank}}^{\text{PMID}}(y, \hat{y}) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}_{\hat{y}_i > \hat{y}_j} \log\left(1 + e^{y_i - y_j}\right).$$

2. **Batch construction.** The ranking loss function involves comparison of predicted affinities within a batch. In order to mitigate study-wise batch effects, we construct each batch from a single PMID only.
3. **Affinity module input.** The Boltz-2 affinity module uses representations from both the inter-chain peptide-ligand interactions, and the intra-chain ligand-ligand interactions, but not intra-chain protein-protein interactions. In order to respect the symmetry of protein-protein interactions, we included *all* inter- and intra-chain representations.

We freeze the weights of the Boltz-2 trunk and trained the affinity module from scratch using He initialisation [14].

4.2 Combining sequence and structure representations

We perform a simple concatenation of the MLP representations extracted from our Boltz-2-PPI affinity module with the final linear embeddings from the sequence models in [8] and construct a simple linear projection as our combined sequence-structure model. We selected the ESM2-650M [15] and ProtT5 [16] models as the two different architectures for our sequence models in our experiments.

5 Results

5.1 Performance on the TCR3d Dataset

We report the results of fine-tuning Boltz-2 on the TCR3d dataset in Table 1. Our fine-tuned model *Boltz-2-PPI*, is shown alongside our sequence-based ESM2 650M baseline [15], which was fine-tuned with an MSE loss on the same data. Unfortunately the sequence-based models achieve stronger predictive performance than Boltz-2-PPI.

A clear limitation is the small size of the dataset, which hinders effective training of a structure-based model. To mitigate this, we experimented with reducing the size of the affinity module to a simple 2-layer MLP, but this did not yield improvements.

We also investigated whether the quality of predicted structures was influencing the training of affinity models. We observed that Boltz-2 achieves high structure prediction accuracy (DockQ ≈ 0.91) on TCR-pMHC structures that were present in its training set, but reduced accuracy (DockQ ≈ 0.70) for unseen complexes [6]. Since our affinity dataset spans both included and more recent complexes, we trained an additional model using experimentally determined structures in place of Boltz-predicted coordinates to assess the impact of poor structure predictions on performance. This model did not outperform Boltz-2-PPI trained using predicted structures, suggesting that structural quality is not the primary performance bottleneck.

For completeness, we also tested the pre-trained Boltz-2 model by representing the peptide as a small molecule (SMILES) – despite the peptide being an out-of-distribution input for Boltz-2, and not considering the MHC as part of the ligand – which showed poor performance.

Taken together, these results suggest that Boltz-2 embeddings do not easily yield signal for affinity prediction in this low-data regime, especially when compared to sequence-based alternatives.

5.2 Performance on the PPB-affinity Dataset

We fine-tuned Boltz-2-PPI on the larger PPB-affinity (filtered) dataset [8]. A summary of results is presented in Table 2. The structure-based Boltz-2-PPI model underperforms when compared with sequence-based baselines reported in Alsamkary *et al.* [8]. This reinforces the conclusions from experiments with the TCR3d dataset that the Boltz embeddings provide weaker signals for affinity prediction relative to direct sequence representations.

Table 1: Performance of Boltz-2-PPI and baselines on the TCR3d test set. Reported values are Pearson correlation (r) and Spearman correlation (ρ).

Model	**Pearson** r (↑)	**Spearman** ρ (↑)
Boltz-2-PPI	0.153	0.091
Boltz-2-PPI (small affinity module)	0.144	0.111
Boltz-2-PPI (true structures)	0.159	0.111
Sequence baseline (ESM2-650M)	**0.239**	**0.193**

To investigate further, we combine Boltz-2-PPI embeddings with sequence embeddings from literature-baselines. We re-train the Prot-T5 and ESM2 sequence models using using published code with default parameters [8] (achieving comparable results to those reported on the test set - Appendix B), and combine the final embeddings with representations extracted from the affinity module of our best Boltz-2-PPI model. For the weaker sequence-based model (ESM2-650M-SC), incorporating Boltz-2-PPI features yields modest improvements. The effect is less pronounced for the stronger sequence model (ProtT5-PAD). These results suggest that structural representations learned from Boltz embeddings contain complementary information, albeit significantly less for high-capacity sequence transformers.

Table 2: Performance of Boltz-2-PPI and sequence-based models on the PPB-affinity (filtered) test set. Reported values are Pearson correlation (r), Spearman correlation (ρ), and the root mean squared error (RMSE). Combined models are computed by us, baseline results are taken from [8].

Model	**Pearson** r (↑)	**Spearman** ρ (↑)	**RMSE** (↓)
Boltz-2-PPI (fine-tuned, structure only)	0.338	0.357	1.362
Sequence baseline (ProtT5-PAD)	0.48	0.51	1.42
Sequence baseline (ESM2-650M-SC)	0.47	0.48	1.74
Combined (ESM2-650M-SC + Boltz-2-PPI)	0.487	0.483	1.367
Combined (ProtT5-PAD + Boltz-2-PPI)	**0.496**	**0.515**	**1.326**

6 Discussion

Across both the TCR3d and PPB-affinity (filtered) datasets, sequence-based models consistently outperform Boltz-2-PPI. Adapting the Boltz-2 affinity module for protein–protein interactions yielded limited gains, even when training on true structures, suggesting that Boltz representations, while strong for structure prediction, lack the expressiveness needed for affinity regression. In contrast, pre-trained protein language models capture these signals more easily.

Our combined models show that Boltz embeddings can add complementary value to weaker sequence models, pointing to the promise of integrating structural and sequence-derived representations. Progress will likely depend on more sophisticated fusion strategies and larger, more homogeneous affinity datasets.

Finally, while Boltz-2 leverages diverse, multi-fidelity supervision (e.g. docking decoys, quality scores, affinity proxies), we restricted training to curated affinity datasets for controlled benchmarking. Incorporating broader, lower-fidelity binding data could provide richer supervision, paralleling Boltz-2's strategy and potentially improving generalisation.

References

[1] Haiying Lu, Qiaodan Zhou, Jun He, Zhongliang Jiang, Cheng Peng, Rongsheng Tong, and Jianyou Shi. Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials. *Signal transduction and targeted therapy*, 5(1):213, 2020.

[2] Tao Tang, Xiaocai Zhang, Yuansheng Liu, Hui Peng, Binshuang Zheng, Yanlin Yin, and Xiangxiang Zeng. Machine learning on protein–protein interaction prediction: models, challenges and trends. *Briefings in Bioinformatics*, 24(2):bbad076, 03 2023.

[3] Jim Middelburg, Kristel Kemper, Patrick Engelberts, Aran F Labrijn, Janine Schuurman, and Thorbald van Hall. Overcoming challenges for cd3-bispecific antibody therapy in solid tumors. *Cancers*, 13(2):287, 2021.

[4] Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, et al. Protein complex prediction with alphafold-multimer. *biorxiv*, pages 2021–10, 2021.

[5] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

[6] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, pages 2024–11, 2025.

[7] Huaqing Liu, Peiyi Chen, Xiaochen Zhai, Ku-Geng Huo, Shuxian Zhou, Lanqing Han, and Guoxin Fan. Ppb-affinity: Protein-protein binding affinity dataset for ai-based protein drug discovery. *Scientific data*, 11(1):1316, 2024.

[8] Hazem Alsamkary, Mohamed Elshaffei, Mohamed Soudy, Sara Ossman, Abdallah Amr, Nehal Adel Abdelsalam, Mohamed Elkerdawy, and Ahmed Elnaggar. Beyond simple concatenation: Fairly assessing plm architectures for multi-chain protein-protein interactions prediction. *arXiv preprint arXiv:2505.20036*, 2025.

[9] Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pages 2025–06, 2025.

[10] Jie Li, Xingyi Guan, Oufan Zhang, Kunyang Sun, Yingze Wang, Dorian Bagni, and Teresa Head-Gordon. Leak proof pdbbind: A reorganized dataset of protein-ligand complexes for more generalizable binding affinity prediction. *ArXiv*, pages arXiv–2308, 2024.

[11] Matsvei Tsishyn, Fabrizio Pucci, and Marianne Rooman. Quantification of biases in predictions of protein–protein binding affinity changes upon mutations. *Briefings in bioinformatics*, 25(1):bbad491, 2024.

[12] Valerie Lin, Melyssa Cheung, Ragul Gowthaman, Maya Eisenberg, Brian M Baker, and Brian G Pierce. Tcr3d 2.0: expanding the t cell receptor structure database with new structures, tools and interactions. *Nucleic Acids Research*, 53(D1):D604–D608, 09 2024.

[13] Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, et al. Boltz-1 democratizing biomolecular interaction modeling. *BioRxiv*, 2024.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[15] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[16] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7112–7127, 2021.

A Dataset splitting

Algorithm 1

Input: Complexes $\mathcal{C} = \{A_1, A_2, \ldots, A_{251}\}$, threshold $\tau = 20.0$, test ratio $r = 0.40$
Output: Train set $\mathcal{T}$, test set $\mathcal{V}$
for each pair $(A_i, A_j) \in \mathcal{C} \times \mathcal{C}$ where $i \neq j$ **do**
    Compute $D(A_i, A_j) = \frac{1}{N_{A_i}} \sum_{k=1}^{N_{A_i}} \min_\ell \text{Levenshtein}(S_{A_i,k}, S_{A_j,\ell})$
    if $D(A_i, A_j) \leq \tau$ **then**
        Add edge (A_i, A_j) to graph G
    end if
end for
Find connected components $\{C_1, C_2, \ldots, C_m\}$ in G
for each remaining component C_k **do**
    if $|\mathcal{V}| < r|\mathcal{C}|$ **and** $|\mathcal{V}| + |C_k| \leq 1.2r|\mathcal{C}|$ **then**
        $\mathcal{V} \leftarrow \mathcal{V} \cup C_k$
    else
        $\mathcal{T} \leftarrow \mathcal{T} \cup C_k$
    end if
end for

B Re-trained sequence-based model PPB-affinity performance

Table 3: Performance re-trained sequence-based models on the PPB-affinity (filtered) test set versus reported values.

Model	**Pearson** r **(↑)**	**Spearman** ρ **(↑)**	**RMSE (↓)**
Re-trained (ProtT5-PAD)	0.484	0.506	6.055
Re-trained (ESM2-650M-SC)	0.462	0.456	1.360
Reported (ProtT5-PAD)	0.48	0.51	1.42
Reported (ESM2-650M-SC)	0.47	0.48	1.74