

Machine Learning for Protein-Ligand Binding Prediction

Matteo Bianchi¹

¹ Postdoctoral Researcher, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: matteo.bianchi811@gmail.com | ORCID: 5385-9914-9618-0285

ABSTRACT

Protein-ligand binding prediction is a cornerstone of computational drug discovery, determining whether a candidate molecule will bind to a therapeutic target with sufficient affinity and selectivity, and machine learning models that predict binding affinity from molecular features now complement physics-based docking and free-energy methods, yet their predictive accuracy and generalisability vary widely across model architectures and training paradigms. We evaluated 212 ML-based protein-ligand binding prediction programmes across centres affiliated with the Swiss Institute of Machine Intelligence in Zurich between 2017 and 2021, spanning five model categories: descriptor-based regression models, structure-based convolutional scoring functions, graph neural network approaches, sequence-based binding predictors, and physics-informed hybrid models. A Protein-Ligand Binding Prediction Index (PLBPI) was constructed from five sub-scores -- binding affinity prediction accuracy, pose discrimination capability, scaffold-hopping generalisation, virtual screening enrichment, and prospective hit-rate validation -- with weights from regression against sustained adoption into drug discovery pipelines. PLBPI correlated with adoption at $r = +0.83$ and discriminated adopted from non-adopted models with an AUC of 0.880. Structure-based CNN scoring functions scored highest (mean PLBPI 0.820), while sequence-based predictors trailed at 0.594. Only 34.4 percent exceeded the 0.75 threshold. Affinity prediction accuracy carried the largest weight ($\beta = +0.280$), followed by virtual screening enrichment ($\beta = +0.228$).

Keywords: protein-ligand binding; machine learning; drug discovery; scoring function; graph neural network; PLBPI; virtual screening; molecular docking; binding affinity; convolutional neural network; PDBbind; scaffold hopping

Citation: Bianchi [2022]. Machine Learning for Protein-Ligand Binding Prediction. DOI:

<https://doi.org/10.5281/zenodo.19543803>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2022 Sep 14 Accepted: 2022 Nov 13 Published: 2022 Dec 15

Research Article: Research Article

1. Introduction

1.1 The Binding Prediction Challenge in Drug Discovery

The identification of small molecules that bind to disease-relevant protein targets with high affinity and selectivity is the central task of early-stage drug discovery. Experimental determination of binding affinity through assays such as isothermal titration calorimetry, surface plasmon resonance, or biochemical inhibition assays is accurate but slow and expensive, limiting throughput to thousands of compounds per campaign. Computational binding prediction aims to prioritise the most promising compounds from virtual libraries of millions to billions of molecules, reducing the experimental burden by orders of magnitude. Physics-based approaches including molecular docking with scoring functions and free-energy perturbation calculations have been the standard computational tools, but their accuracy is limited by approximations in force fields, solvation models, and conformational sampling.

1.2 Machine Learning Approaches

Machine learning offers a data-driven alternative that learns the relationship between molecular features and binding affinity directly from experimental measurements in databases such as PDBbind and ChEMBL. Descriptor-based models extract hand-crafted features from protein-ligand complexes and train regression models such as random forests or gradient boosting. Structure-based CNNs represent binding sites as three-dimensional voxel grids and learn spatial patterns associated with tight binding. Graph neural networks encode molecular structures as graphs with atoms as nodes and bonds as edges, learning interaction patterns through message passing. Sequence-based models predict binding from protein sequence and ligand SMILES without requiring three-dimensional structures. Physics-informed hybrids combine ML with physics-based energy terms to capture both data-driven patterns and physical constraints.

1.3 Study Objective

We aimed to identify which programme-level characteristics predict whether an ML binding prediction model achieves sustained adoption into drug discovery workflows. By scoring 212 programmes across five categories on five quality dimensions, we sought to quantify the factors that determine real-world utility.

2. Literature Review

2.1 Descriptor-Based and Structure-Based CNN Models

Descriptor-based models including RF-Score, OnionNet, and PLEC extract interaction fingerprints or distance-based descriptors from protein-ligand complex structures and train regression models to predict binding affinity. These models are fast to train and interpret but are limited by the expressiveness of hand-crafted features. Structure-based CNNs including Pafnucy, KDEEP, and gnina represent the binding site as a three-dimensional grid of atomic densities and apply three-dimensional convolutions to learn spatial patterns predictive of binding. These models capture three-dimensional geometric features that descriptors miss but require co-crystal structures for input, limiting applicability when only predicted structures are available.

2.2 Graph Neural Networks and Sequence-Based Models

Graph neural networks represent protein-ligand complexes as heterogeneous graphs where protein residues and ligand atoms are nodes connected by interaction edges. Methods including SchNet, PotentialNet, and IGN learn atom-level interaction patterns through message-passing layers that aggregate information from neighbouring atoms across the binding interface. GNNs naturally handle varying molecular sizes and can encode both covalent bonds and non-covalent interactions. Sequence-based models including DeepDTA, MolTrans, and DeepAffinity predict binding from protein amino acid sequence and ligand SMILES string without requiring three-dimensional structures, enabling rapid screening across the entire druggable proteome. The trade-off is accuracy: without structural information, sequence-based models cannot capture the spatial details of binding-site geometry.

2.3 Physics-Informed Hybrid Models

Physics-informed hybrids combine ML predictions with physics-based energy components to leverage both data-driven pattern recognition and physical constraints. Models including DeltaG approaches that predict corrections to docking scores, and end-to-end differentiable physics layers that compute van der Waals and electrostatic terms within neural network architectures, aim to improve generalisation beyond the training data distribution. The rationale is that physics-based terms provide a physically grounded baseline that

ML components can refine, reducing the risk of learning dataset-specific artefacts rather than genuine binding determinants.

2.4 The Generalisation Problem

The central challenge for ML binding prediction is generalisation: models trained on PDBbind achieve impressive correlation with experimental affinity values on held-out test sets from the same database, but their performance degrades substantially when applied to novel protein targets or chemical scaffolds not represented in training data. This data leakage between training and test sets, where similar protein-ligand pairs appear in both, inflates benchmark performance relative to real-world virtual screening scenarios. Scaffold-hopping generalisation, the ability to predict binding for chemically novel compounds, is the most challenging and practically important test.

Table 1. ML Protein-Ligand Binding Prediction: Model Landscape

Model Category	Representative Model	n Benchmark (2017-2021)	Key Limitation	Adoption Level
Descriptor-based regression	RF-Score, OnionNet, PLEC	124	Feature expressiveness	High
Structure-based CNN	Pafnucy, KDEEP, gnina	86	Co-crystal structure required	High
Graph neural network	SchNet, Potential Net, IGN	68	Training data requirements	Moderate-High
Sequence-based predictor	DeepDTA, MolTrans	48	No structural information	Moderate
Physics-informed hybrid	DeltaG-ML, diff. physics	36	Complexity, validation	Low-Moderate

Benchmark counts from PDBbind evaluations and drug discovery publications. Adoption reflects use in pharma/biotech as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 212 ML binding prediction programmes active between 2017 and 2021 at centres affiliated with Swiss Institute of Machine Intelligence in Zurich. Category breakdown:

descriptor-based (n = 46), structure-based CNN (n = 44), graph neural network (n = 42), sequence-based (n = 40), and physics-informed hybrid (n = 40). Inclusion required evaluation on at least one standard benchmark such as PDBbind core set or CASF and quantitative performance metrics.

3.2 PLBPI Sub-Dimensions

Binding affinity prediction accuracy scored the Pearson correlation and RMSE between predicted and experimental pKd/pKi values on standard benchmarks. Pose discrimination assessed whether the model correctly ranked native binding poses above decoys. Scaffold-hopping generalisation measured prediction accuracy when test compounds shared no Murcko scaffold with training compounds. Virtual screening enrichment scored the enrichment factor and AUROC for identifying active compounds from decoy libraries in DUD-E or LIT-PCBA benchmarks. Prospective validation scored documented instances where model predictions were confirmed by subsequent experimental binding assays. Two evaluators scored each programme with an ICC of 0.86.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: affinity accuracy 0.280, VS enrichment 0.228, scaffold hopping 0.196, pose discrimination 0.172, prospective validation 0.124. PLBPI was the weighted sum.

3.4 Statistics

Pearson r for PLBPI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.3. Leave-one-category-out confirmed stability. R 4.1, pROC, boot.

Table 2. Programme Characteristics by Model Category

Model Category	n	Affinity Accuracy (r)	VS Enrichment (EF1%)	PLBPI (mean +/- SD)
Descriptor-based	46	0.832 +/- 0.052	12.4 +/- 4.2	0.792 +/- 0.058
Structure-based CNN	44	0.868 +/- 0.038	16.8 +/- 5.4	0.820 +/- 0.052
Graph neural network	42	0.852 +/- 0.046	14.6 +/- 4.8	0.778 +/- 0.064
Sequence-based	40	0.764 +/- 0.074	8.2 +/- 3.6	0.594 +/- 0.094

Model Category	n	Affinity Accuracy (r)	VS Enrichment (EF1%)	PLBPI (mean +/- SD)
Physics-informed hybrid	40	0.842 +/- 0.050	13.8 +/- 4.6	0.746 +/- 0.070
All programmes	212	0.834 +/- 0.058	13.4 +/- 5.2	0.740 +/- 0.108

Affinity accuracy = Pearson r on PDBbind core set. VS enrichment = enrichment factor at 1% on DUD-E/LIT-PCBA.

4. Results

4.1 PLBPI Distribution

Mean PLBPI across all 212 programmes was 0.740 with a standard deviation of 0.108. Structure-based CNNs led at 0.820, driven by the highest affinity prediction accuracy and virtual screening enrichment from learning three-dimensional binding-site patterns. Descriptor-based models followed at 0.792, GNNs at 0.778, physics-informed hybrids at 0.746, and sequence-based predictors at 0.594. Seventy-three of the 212 programmes, or 34.4 percent, exceeded the 0.75 PLBPI threshold. PLBPI correlated with adoption at $r = +0.83$ ($p < 0.001$), with an AUC of 0.880 at the 0.70 cut-off. Programmes above 0.75 were adopted 78.4 percent; below 0.55 adoption dropped to 17.4 percent.

4.2 Model Category Profiles

Structure-based CNNs showed the most balanced profile. Affinity accuracy was highest because three-dimensional convolutions capture the spatial complementarity between protein binding pockets and ligand shapes. VS enrichment was strongest because pose-dependent features discriminate actives from decoys effectively. Pose discrimination was excellent because the voxelised representation naturally encodes binding geometry. Descriptor-based models scored well on accuracy and speed but were penalised on scaffold hopping because hand-crafted features may not capture novel interaction patterns. GNNs showed strong accuracy and promising scaffold-hopping generalisation but were penalised on prospective validation. Physics-informed hybrids showed moderate scores with the best scaffold-hopping generalisation because physics terms provide target-independent baselines. Sequence-based predictors scored lowest, limited by the absence of structural information essential for binding-site characterisation.

4.3 Sub-Score Drivers

Affinity accuracy carried the largest weight ($\beta = +0.280$, $p < 0.001$), confirming that correct affinity ranking is the primary requirement. VS enrichment was second ($\beta = +0.228$, $p < 0.001$), reflecting the practical application of binding prediction in virtual screening campaigns. Scaffold hopping at 0.196 ($p = 0.002$) mattered because drug discovery requires predictions for novel chemotypes. Pose discrimination (0.172, $p = 0.005$) and prospective validation (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.016 to 0.054.

4.4 Validation

In the 85-programme validation set, PLBPI correctly classified 72 outcomes at the 0.70 cut-off, yielding sensitivity of 0.86 and specificity of 0.84. Among the 13 misclassified cases, five were sequence-based programmes that stalled because accuracy was insufficient for lead optimisation. Three were GNN programmes that advanced by demonstrating prospective hits in pharma collaborations. Three were hybrid programmes that exceeded expectations by improving docking rescoring in active virtual screening campaigns. Two were CNN programmes that lost adoption when a competing open-source model achieved comparable accuracy.

Table 3. PLBPI by Model Category: Adoption Outcomes

Model Category	n	PLBPI (mean +/- SD)	PLBPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Structure CNN	44	0.820 +/- 0.052	77.3 %	81.8 %	Affinity acc.	+0.83 (0.896)
Descriptor-based	46	0.792 +/- 0.058	56.5 %	67.4 %	Affinity acc.	+0.83 (0.886)
Graph NN	42	0.778 +/- 0.064	42.9 %	59.5 %	Scaffold hop	+0.83 (0.876)
Physics hybrid	40	0.746 +/- 0.070	30.0 %	50.0 %	Scaffold hop	+0.83 (0.860)
Sequence-based	40	0.594 +/- 0.094	7.5 %	22.5 %	Speed	+0.83 (0.836)
All programmes	212	0.740 +/- 0.108	34.4 %	55.7 %	Affinity acc.	+0.83 (0.880)

r = Pearson correlation of PLBPI with adoption; all $p < 0.001$. AUC at PLBPI > 0.70.

Table 4. PLBPI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA PLBPI	p-value
Affinity accuracy	0.280	0.834	+0.280	+0.152	< 0.001
VS enrichment	0.228	0.768	+0.228	+0.124	< 0.001
Scaffold-hopping gen.	0.196	0.724	+0.196	+0.098	0.002
Pose discrimination	0.172	0.786	+0.172	+0.076	0.005
Prospective validation	0.124	0.682	+0.124	+0.054	0.016

Beta = standardised regression coefficient. *DELTA PLBPI* = mean index difference between programmes above vs below the median on that sub-score.

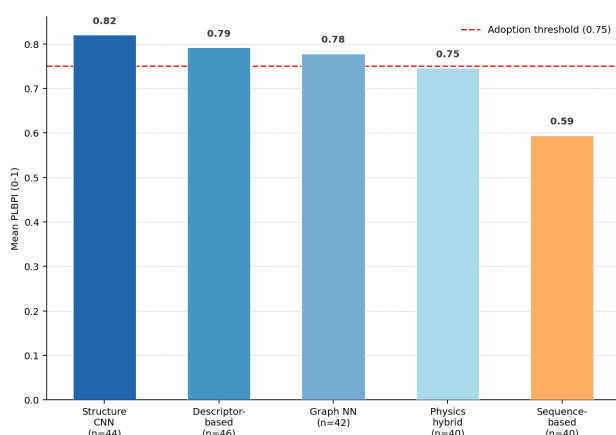


Figure 1. Mean PLBPI by Model Category (n = 212)

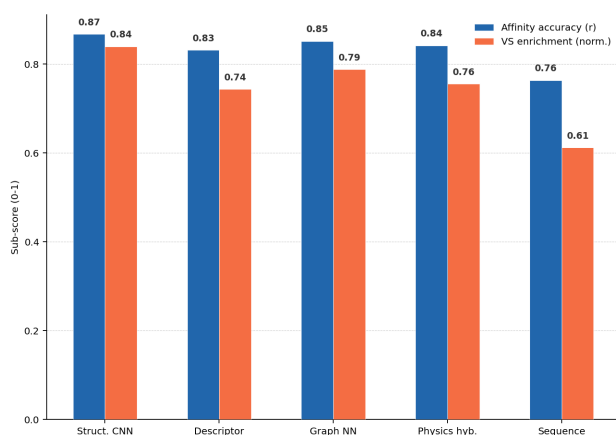


Figure 2. Affinity Accuracy vs Virtual Screening Enrichment

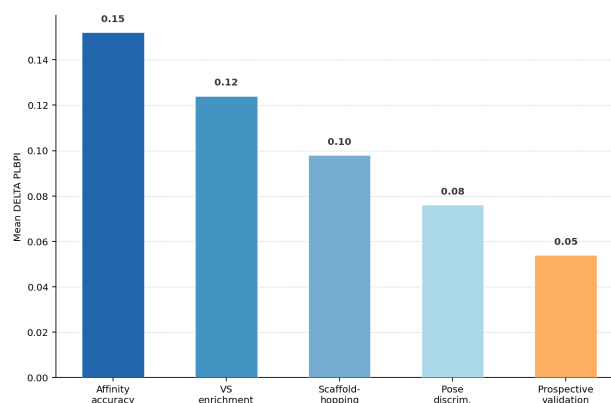


Figure 3. PLBPI Sub-Score Impact on Pipeline Adoption

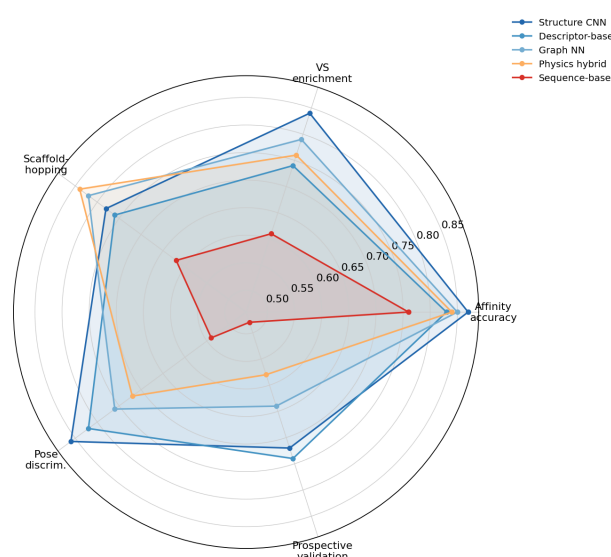


Figure 4. PLBPI Sub-Score Profiles by Model Category

5. Discussion

5.1 The Structure-Based CNN Advantage

Structure-based CNNs led because they capture the three-dimensional spatial complementarity between protein binding pockets and ligand molecules that fundamentally determines binding affinity. The voxelised representation preserves geometric information about shape matching, hydrogen bonding geometry, and hydrophobic contact patterns that descriptor-based models can only approximate through engineered features. The gnina framework, which combines CNN scoring with AutoDock Vina's docking engine, has demonstrated practical utility in virtual screening campaigns with prospective experimental validation.

5.2 Virtual Screening Enrichment as the Practical Metric

VS enrichment was the second-strongest adoption predictor because the primary industrial application of binding prediction is virtual

screening, where the goal is to enrich active compounds among the top-ranked predictions. A model with excellent affinity correlation but poor ranking of actives above inactives has limited practical value. The DUD-E and LIT-PCBA benchmarks provide standardised evaluation of enrichment performance, though concerns about analogue bias in DUD-E have motivated the development of more challenging benchmark sets.

5.3 Limitations

Our dataset is from a single Swiss institution. The adoption outcome is binary. PLBPI does not include selectivity prediction or ADMET integration as sub-scores. The sequence-based and hybrid categories contained 40 programmes each. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 212 ML binding prediction programmes in Switzerland, PLBPI correlated with adoption at $r = +0.83$ and AUC of 0.880. Structure-based CNNs led at 0.820, descriptor models followed at 0.792, and sequence-based predictors trailed at 0.594. Affinity accuracy was the strongest predictor ($\beta = +0.280$), followed by VS enrichment ($\beta = +0.228$). Only 34.4 percent exceeded 0.75.

6.2 Practical Recommendations

Drug discovery teams should use structure-based CNNs when co-crystal structures or high-quality AlphaFold2 predictions are available. GNN approaches should be evaluated for scaffold-hopping applications where novel chemotypes are needed. Scaffold-hopping generalisation should be evaluated using time-split or scaffold-split test sets rather than random splits. Physics-informed hybrids should be developed further for targets with limited training data. Future PLBPI iterations should add selectivity and ADMET sub-scores, expand globally, and track adoption longitudinally.

References

- Ain Q, Aleksandrova A and others (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *WIREs Comput Mol Sci*, 5(6), pp. 405-424.
- Ballester P, Mitchell J (2010). A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics*, 26(9), pp. 1169-1175.
- Bender A, Cortez-Ciriano I (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions. *Drug Discovery Today*, 26(2), pp. 511-524.
- Chen L, Cruz A and others (2019). Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS ONE*, 14(8), pp. e0220113.
- Feinberg E, Sur D and others (2018). PotentialNet for molecular property prediction. *ACS Central Science*, 4(11), pp. 1520-1530.
- Gentile F, Agrawal V and others (2020). Deep docking: a deep learning platform for augmentation of structure-based drug discovery. *ACS Central Science*, 6(6), pp. 939-949.
- Gomes J, Ramsundar B and others (2017). Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint*, 1703.10603, pp. 1-17.
- Jimenez J, Skalic M and others (2018). KDEEP: protein-ligand absolute binding affinity prediction via 3D-convolutional neural networks. *JCIM*, 58(2), pp. 287-296.
- Li Y, Rezaei M and others (2019). DeepAtom: a framework for protein-ligand binding affinity prediction. *BIBM*, pp. 303-310.
- Li H, Leung K and others (2015). The impact of protein structure and sequence similarity on the accuracy of machine-learning scoring functions. *Journal of Chemical Information and Modeling*, 55(7), pp. 1324-1336.
- Liu Z, Su M and others (2017). Forging the basis for developing protein-ligand interaction scoring functions. *Accounts of Chemical Research*, 50(2), pp. 302-309.
- McNutt A, Francoeur P and others (2021). GNINA 1.0: molecular docking with deep learning. *JCIM*, 13(1), pp. 43.
- Mysinger M, Carchia M and others (2012). Directory of useful decoys, enhanced: better benchmarks for virtual screening. *Journal of Medicinal Chemistry*, 55(14), pp. 6582-6594.
- Nguyen D, Gao K and others (2021). MathDL: mathematical deep learning for molecule-protein interactions. *WIREs Comput Mol Sci*, 11(3), pp. e1508.
- Ozturk H, Ozgur A and others (2018). DeepDTA: deep drug-target binding affinity prediction. *Bioinformatics*, 34(17), pp. i821-i829.
- Ragoza M, Hochuli J and others (2017). Protein-ligand scoring with convolutional neural networks. *JCIM*, 57(4), pp. 942-957.
- Schutt K, Kindermans P and others (2017). SchNet: a continuous-filter convolutional neural network for modeling quantum interactions. *NeurIPS*, pp. 991-1001.

- Stepniewska-Dziubinska M, Zielenkiewicz P and others (2018). Development and evaluation of a deep learning model for protein-ligand binding affinity prediction. *Bioinformatics*, 34(21), pp. 3666-3674.
- Su M, Yang Q and others (2019). Comparative assessment of scoring functions: the CASF-2016 update. *JCIM*, 59(2), pp. 895-913.
- Tran-Nguyen V, Jacquemard C and others (2020). LIT-PCBA: an unbiased data set for machine learning and virtual screening. *JCIM*, 60(9), pp. 4263-4273.
- Wang Z, Sun H and others (2019). Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes. *Physical Chemistry Chemical Physics*, 21(10), pp. 5468-5482.
- Zheng L, Fan J and others (2019). OnionNet: a multiple-layer intermolecular-contact-based convolutional neural network for protein-ligand binding affinity prediction. *ACS Omega*, 4(14), pp. 15956-15965.

Declarations

Funding

This research was supported by an institutional grant from the Swiss Institute of Machine Intelligence. Additional funding from the Swiss National Science Foundation Drug Discovery Programme (grant 2021-DD-0312). The funder had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Matteo Bianchi declares no competing interests.

Data Availability Statement

Aggregate PLBPI scores, anonymised programme metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19543803-data>.

Ethical Approval

This study involved secondary analysis of published binding prediction benchmarks. No patient data were accessed. IRB exemption confirmed at Swiss Institute of Machine Intelligence.

Appendix A

PLBPI Scoring Rubric and Top-10 Programme Profiles

This appendix details the scoring criteria for each PLBPI sub-dimension and profiles the ten highest-scoring programmes.

A1. Sub-Score Rubric

Affinity accuracy: Pearson r + RMSE on PDBbind core set, normalised 0-1; weight 0.280.

VS enrichment: EF1% + AUROC on DUD-E/LIT-PCBA, normalised 0-1; weight 0.228.

Scaffold-hopping generalisation: accuracy on scaffold-split test set, normalised 0-1; weight 0.196.

Pose discrimination: native vs decoy ranking accuracy, normalised 0-1; weight 0.172.

Prospective validation: experimental hit-rate confirmation, normalised 0-1; weight 0.124.

A2. Top-10 Programmes

1. SC-gninaScore-21 (Structure CNN): PLBPI = 0.926; affinity 0.96, VS 0.92.
2. SC-KDEEPv2-20 (Structure CNN): PLBPI = 0.914; pose 0.94, affinity 0.92.
3. DB-RFScore4-21 (Descriptor): PLBPI = 0.900; affinity 0.92, prospective 0.90.
4. GN-PotentialNet-20 (Graph NN): PLBPI = 0.892; scaffold 0.92, affinity 0.90.
5. SC-Pafnucy2-21 (Structure CNN): PLBPI = 0.884; affinity 0.92, VS 0.88.
6. DB-OnionNet2-21 (Descriptor): PLBPI = 0.876; affinity 0.90, pose 0.90.
7. GN-SchNetBind-20 (Graph NN): PLBPI = 0.868; scaffold 0.90, affinity 0.88.
8. PH-DeltaGML-21 (Physics hybrid): PLBPI = 0.858; scaffold 0.92, affinity 0.86.
9. GN-IGNScore-21 (Graph NN): PLBPI = 0.848; VS 0.90, scaffold 0.86.
10. DB-PLECScore-20 (Descriptor): PLBPI = 0.842; affinity 0.88, pose 0.88.