IMPERIAL

Machine-Learning Scoring Functions for Structure-Based Binding Affinity Prediction: a 15-Year Perspective

Pedro J. Ballester

Department of Bioengineering, Imperial College London, UK

p.ballester@imperial.ac.uk

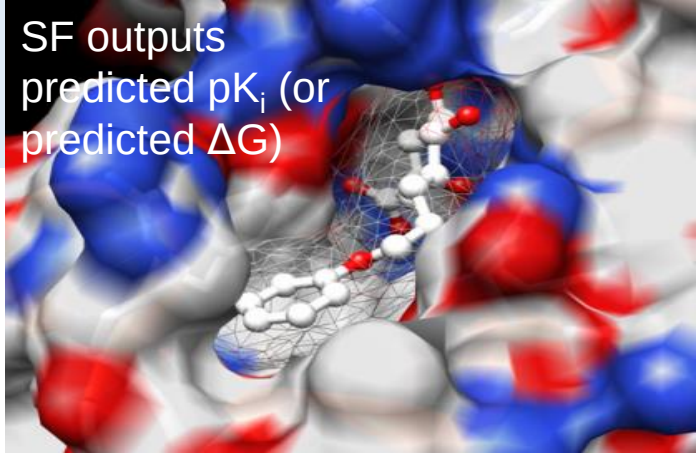
x.com/pjballester

linkedin.com/in/pedroballester/

ballestergroup.github.io

XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery (BCADD-2026)

Molecular docking applications



Predict whether or not a **molecule docked** to a given **target** is a **true binder** from the **docked structure**

Predict the **affinity** (~activity) of a **molecule bound** to a given **target** from the **crystal structure**

Virtual Screening (VS)

* *Benchmarking posed as binary classification*
* *Goal: identifying new binders of a given target*
* *+ve and –ve instances*
* *Disadvantage: more confounding factors s.t. docking pose error, strong class imbalance, bias, etc.*

Binding Affinity Prediction (BAP)

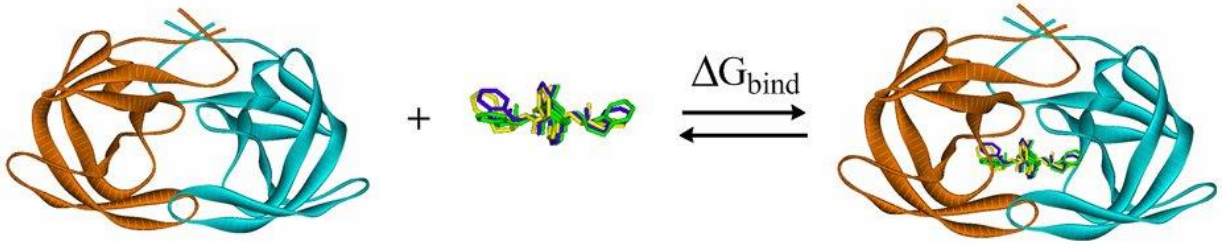
* Benchmarking posed as regression problem
* Goal: Kd/Ki/Activity measurements reduced
* mostly +ve instances (eg. target-docked actives)
* Advantage: fewer confounding factors, esp. when using crystal structs.

Each application rely on the prediction provided by a **scoring function (SF)** → important to develop SFs that are optimal for the intended application

$$[EI]_{aq} \rightleftharpoons [E]_{aq} + [I]_{aq}$$

$$K_A = K_i^{-1} = \frac{[EI]}{[E]\,[I]}$$

$$\Delta G = -RT\ln K_A$$



Classical scoring functions (SFs) – fixed linear form

- Force Field-based SFs (e.g. DOCK score)

$$E_{binding} = \sum_{protein} \sum_{ligand} \left(\frac{A_{ij}}{d_{ij}^{12}} - \frac{B_{ij}}{d_{ij}^{6}} + 332.0 \times \frac{q_i q_j}{\varepsilon(d_{ij}) \times d_{ij}} \right)$$

- Empirical SFs (e.g. X-Score)

$$\Delta G_{bind} = w_0 + w_1 \Delta G_{vdW} + w_2 \Delta G_{h-bond} + w_3 \Delta G_{rotor} + w_4 \Delta G_{hydrophobic}$$

- Knowledge-based SFs (e.g. PMF)

$$PMF = \sum_{prot} \sum_{lig} A_{ij}(d_{ij}) \qquad A_{ij}(d_{ij}) = -k_B T \ln \left[f^{j}_{Vol_corr}(r) \frac{\rho^{ij}_{seg}(r)}{\rho^{ij}_{bulk}} \right]$$

- Mathematically, all other SFs are a **linear** combination of nonlinear terms! → as with SFC, RF better than ΣΣ

Machine-learning Scoring Functions (RF-Score, 2010)

BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 9 2010, pages 1169–1175
doi:10.1093/bioinformatics/btq112

Structural bioinformatics Advance Access publication March 17, 2010

A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking

Pedro J. Ballester[1,*,†] and John B. O. Mitchell[2,*]

[1]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW and [2]Centre for Biomolecular Sciences, University of St Andrews, North Haugh, St Andrews KY16 9ST, UK

Associate Editor: Burkhard Rost

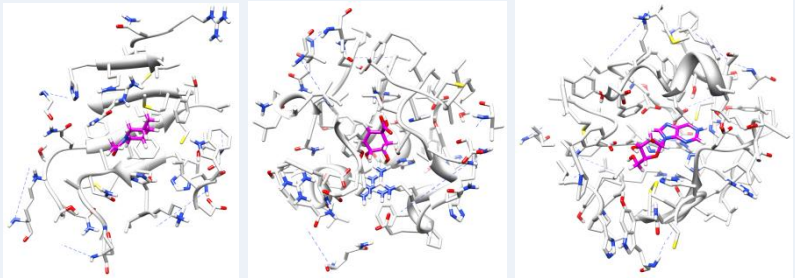
1. Generic: structures of proteins from other families may improve prediction (complement structures of target)

2. This requires an algorithm that builds sufficiently flexible regression models → Random Forest (Breiman, 2001)

3. Advantage: circumventing a priori assumptions about the SF's functional form (linearity), which may reduce modelling error

Training and testing RF-Score on CASF-2007

Training set (1105 complexes)

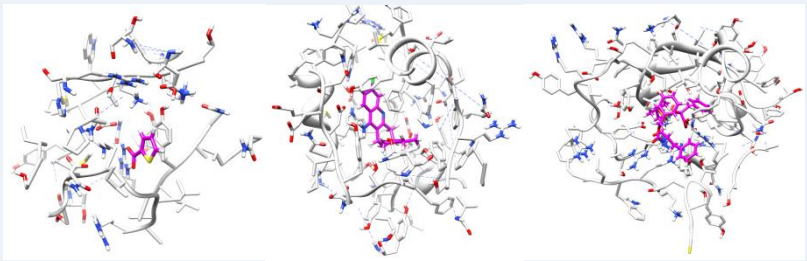


1w8l
pKi=0.49

1gu1
pKi=4.52

2ada
pKi=13

Test set (195 complexes)



2hdq
pKi=1.4

1e66
pKi=9.89

7cpa
pKi=13.96

Generation of descriptors (d_{cutoff}, binning, interatomic types)

$pK_{d/i}$	C.C	–	C.I	N.C	–	I.I	PDB
0.49	1254	–	0	166	–	0	1w8l
–	–	–	–	–	–	–	–
13.00	2324	–	0	919	–	0	2ada

1105

$pK_{d/i}$	C.C	–	C.I	N.C	–	I.I	PDB
1.40	858	–	0	0	–	0	2hdq
–	–	–	–	–	–	–	–
13.96	4476	–	0	283	–	0	7cpa

195

Random Forest (RF) training
(descriptor selection, model selection)

RF-Score
(description and training choices)

2010: 1st version → Rp=0.776

Scoring Function Corsortium using RF-Score code

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

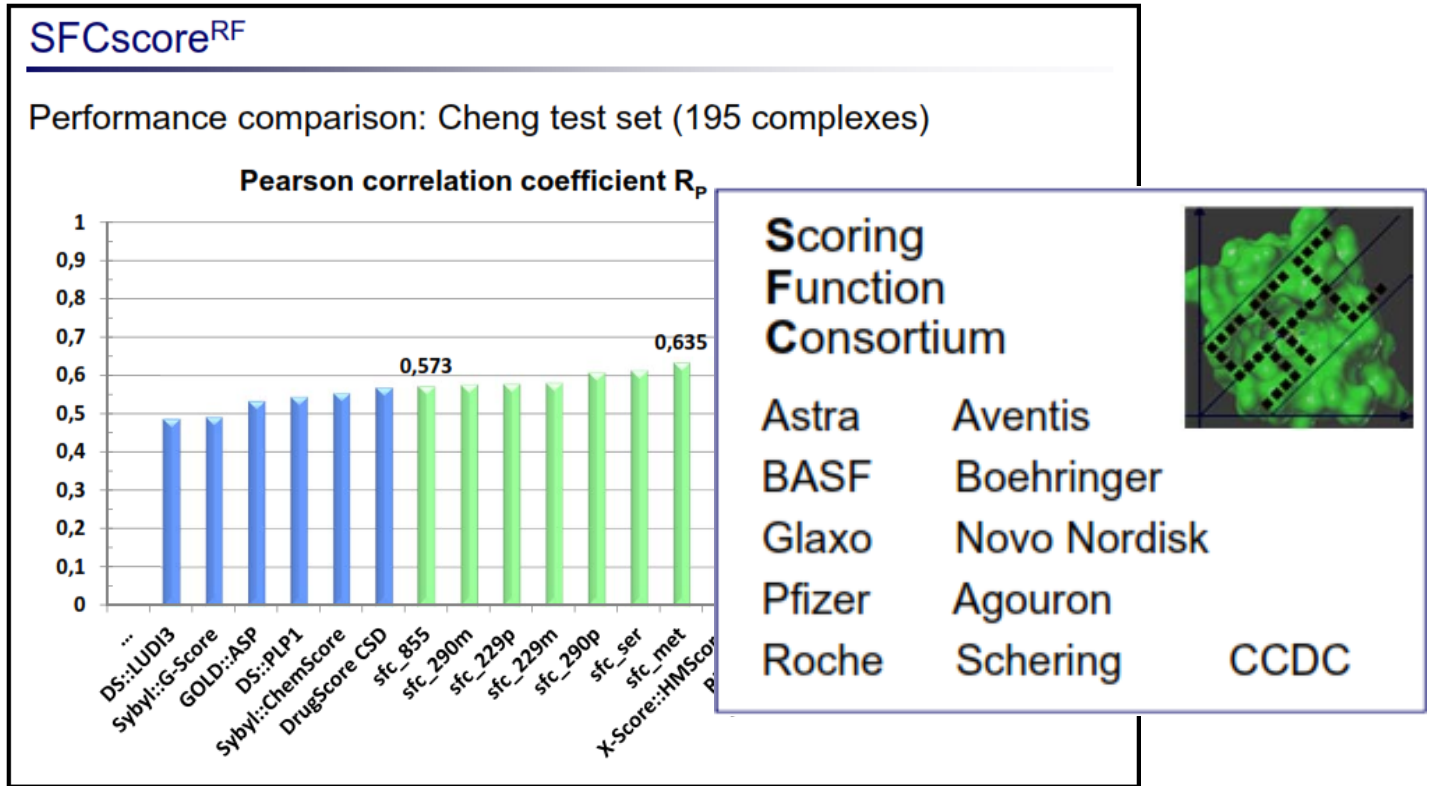
2008
2013

SFCscore^{RF}: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes

David Zilian and Christoph A. Sotriffer*

Institute of Pharmacy and Food Chemistry, University of Wuerzburg, Am Hubland, D-97074 Wuerzburg, Germany

Department of Pharmaceutical Chemistry, Philipps-Universität Marburg, D-35032 Marburg, Germany

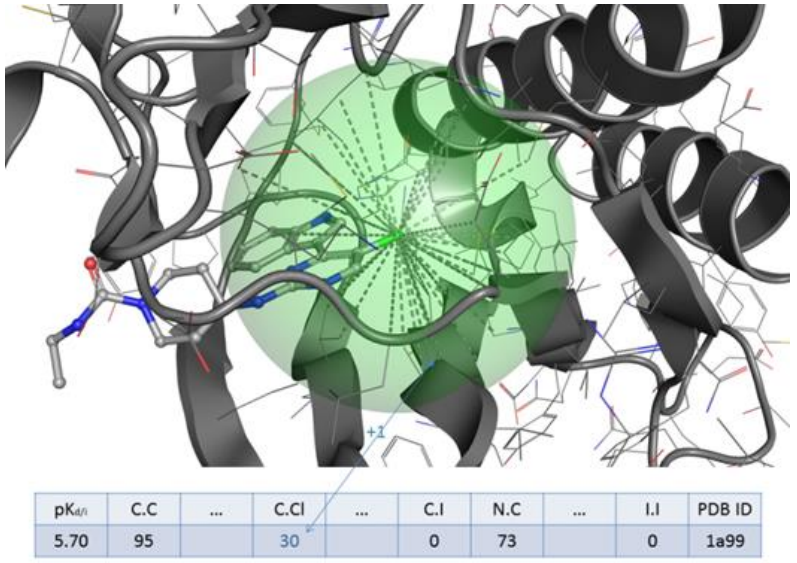


http://infochim.u-strasbg.fr/CS3_2012/Lectures/Sotriffer.pdf

Best features? RF-Score v2 in 2014

Systematic numerical study:

- Interatomic distance cutoffs
- Interatomic distance bin sizes
- Atom hybridisation and protonation state
- Angle between HBD, HBA and H atoms.
- Covalent and van der Waals radius of atoms.
- Basic feature selection.
- Model selection by OOB



$pK_{d/i}$	C.C	...	C.Cl	...	C.I	N.C	...	I.I	PDB ID
5.70	95		30		0	73		0	1a99

Elem(c12,b2)_spr1_oob

- Rp: 0.787 → 0.803

(= training set, = test set)

→ a data-driven feature selection procedure was more effective than a knowledge-based one

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Terms of Use

Does a More Precise Chemical Description of Protein–Ligand Complexes Lead to More Accurate Prediction of Binding Affinity?

Pedro J. Ballester,[†,*] Adrian Schreyer,[‡] and Tom L. Blundell[‡]

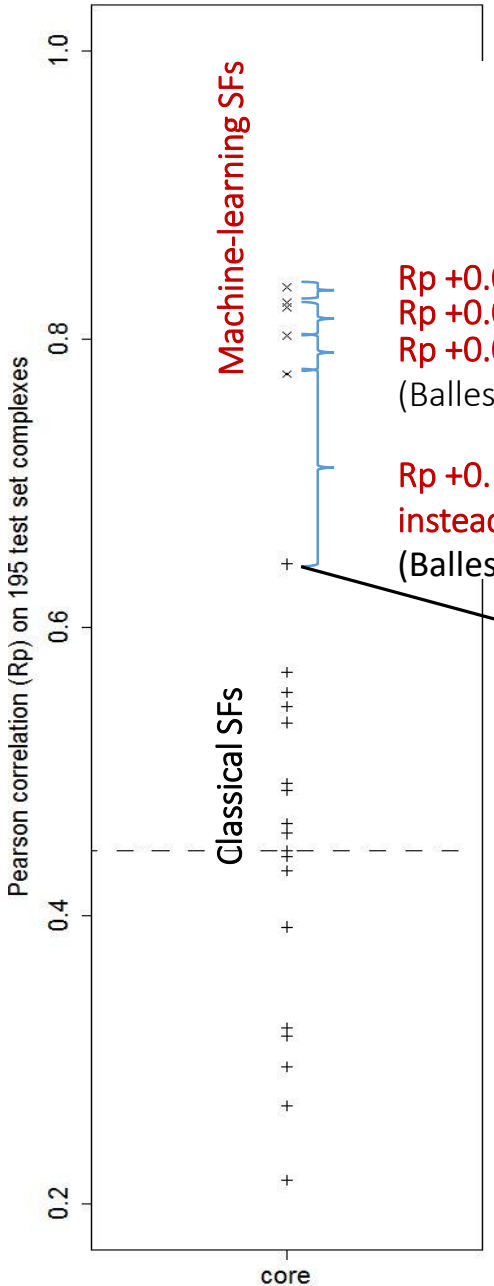
[†]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton - CB10 1SD, United Kingdom

[‡]Dept. of Biochemistry, University of Cambridge, 80 Tennis Court Rd, Cambridge - CB2 1GA, United Kingdom

https://pubs.acs.org/doi/10.1021/ci500091r

Deep learning < 2020: incremental impact on CASF-2007

scoring function	R_p
RF-Score::Elem-v2	0.803
RF-Score::Elem-v1	0.776
X-Score::HMScore	0.644
DrugScore^{CSD}	0.569
SYBYL::ChemScore	0.555
DS::PLP1	0.545
GOLD::ASP	0.534
SYBYL::G-Score	0.492
DS::LUDI3	0.487
DS::LigScore2	0.464
GlideScore-XP	0.457
DS::PMF	0.445
GOLD::ChemScore	0.441
by NHA	0.431
SYBYL::D-Score	0.392
IMP::RankScore	0.322
DS::Jain	0.316
GOLD::GoldScore	0.295
SYBYL::PMF-Score	0.268
SYBYL::F-Score	0.216



Rp +0.010: RF-Score v3 with RDKit ligand features in 2019 → (**)
Rp +0.023: Deep Neural Networks instead of RF (2017-18) → (*)
Rp +0.027: RF with data-selected features → RF-Score-v2
(Ballester, Schreyer & Blundell 2014; doi:10.1021/ci500091r)

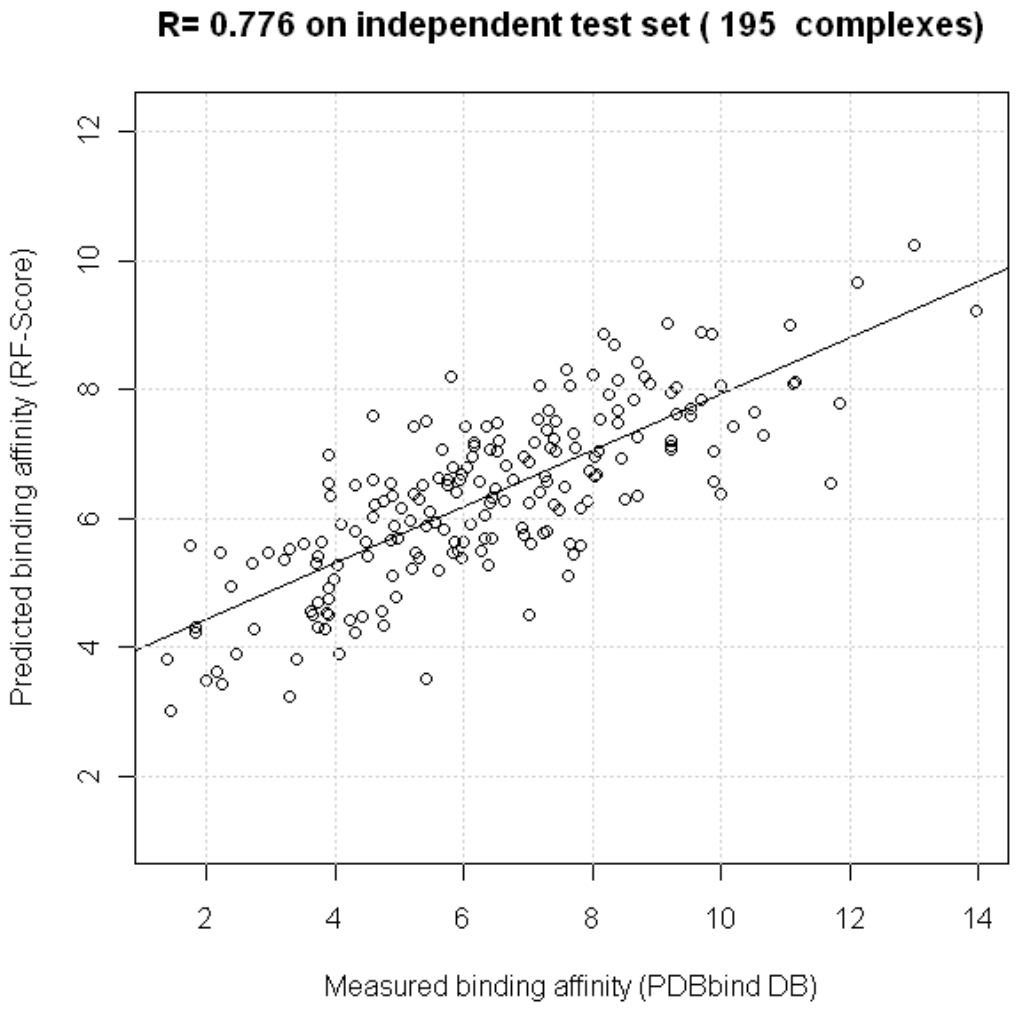
Rp +0.132: Random Forest (RF) with simple intermolecular features instead of linear regression and expert-selected features → RF-Score-v1
(Ballester & Mitchell **2010**; doi:10.1093/bioinformatics/btq112r)

Performance of the **best classical SF for this problem** → X-Score (Cheng et al. **2009**; doi:10.1021/ci9000053)

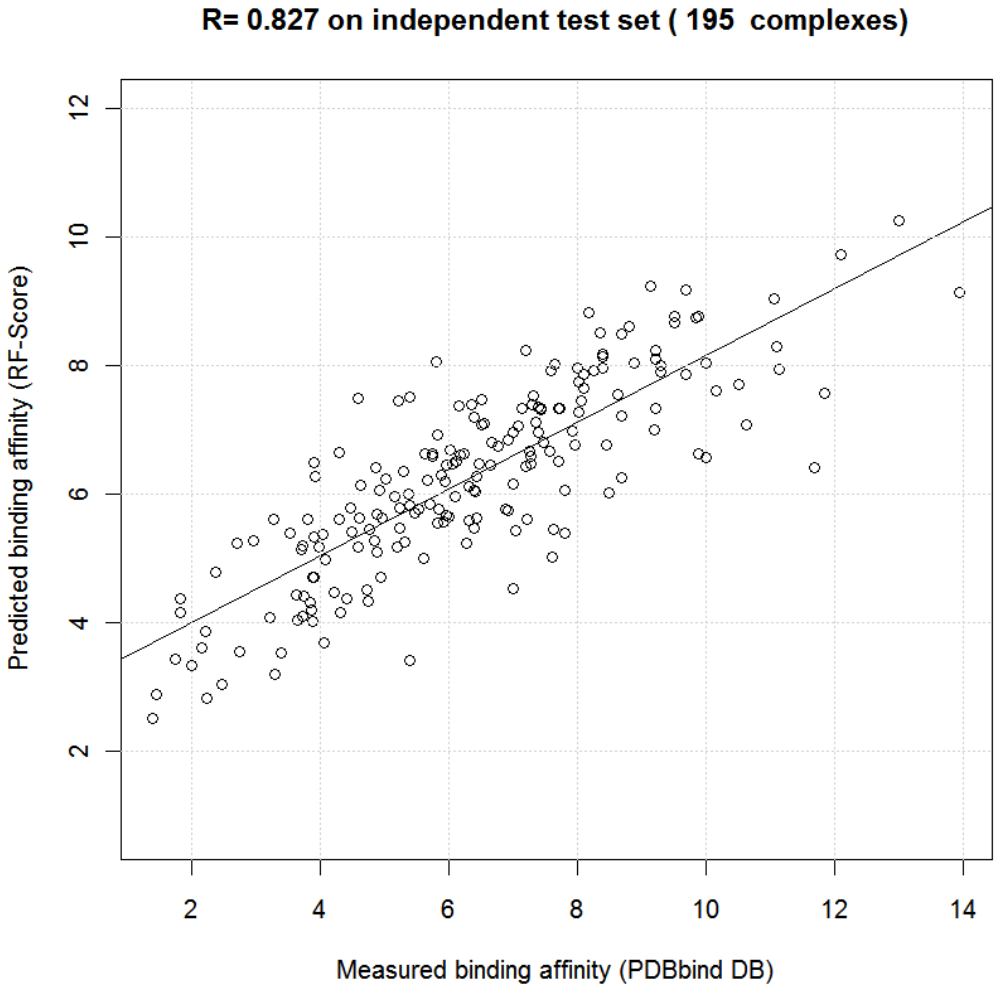
(*) Deep Learning applied to this problem
→ TNet-BP (Cang & Wei 2017; doi:10.1371/journal.pcbi.1005690) &
→ Spatial MPNN (Feinberg et al. 2018; arXiv:1803.04465v1)

(**) RF with protein-ligand and ligand features
By Boyles, Deane & Morris 2019; doi:10.1093/bioinformatics/btz665)

Training set released to prevent training-test overlaps



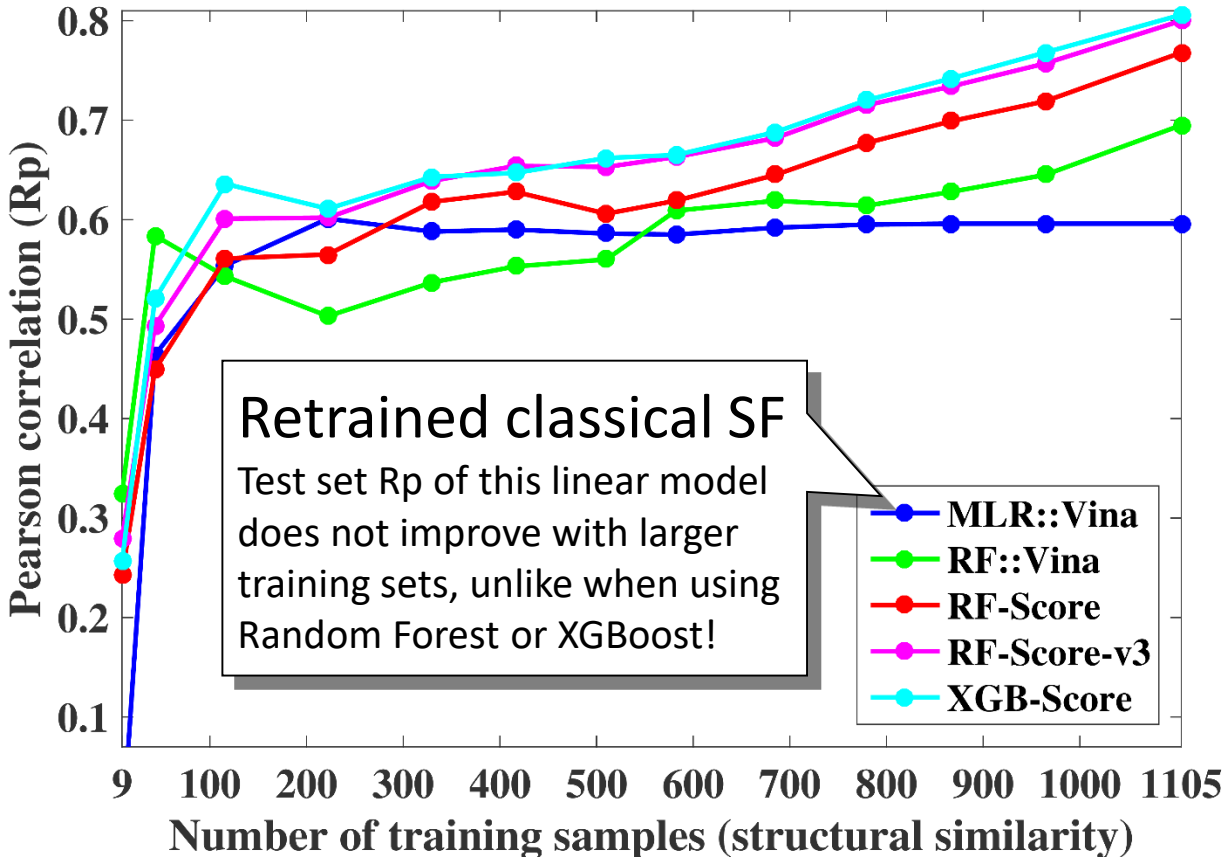
No overlap (unlike other SFs but X-Score) → R_p=0.776

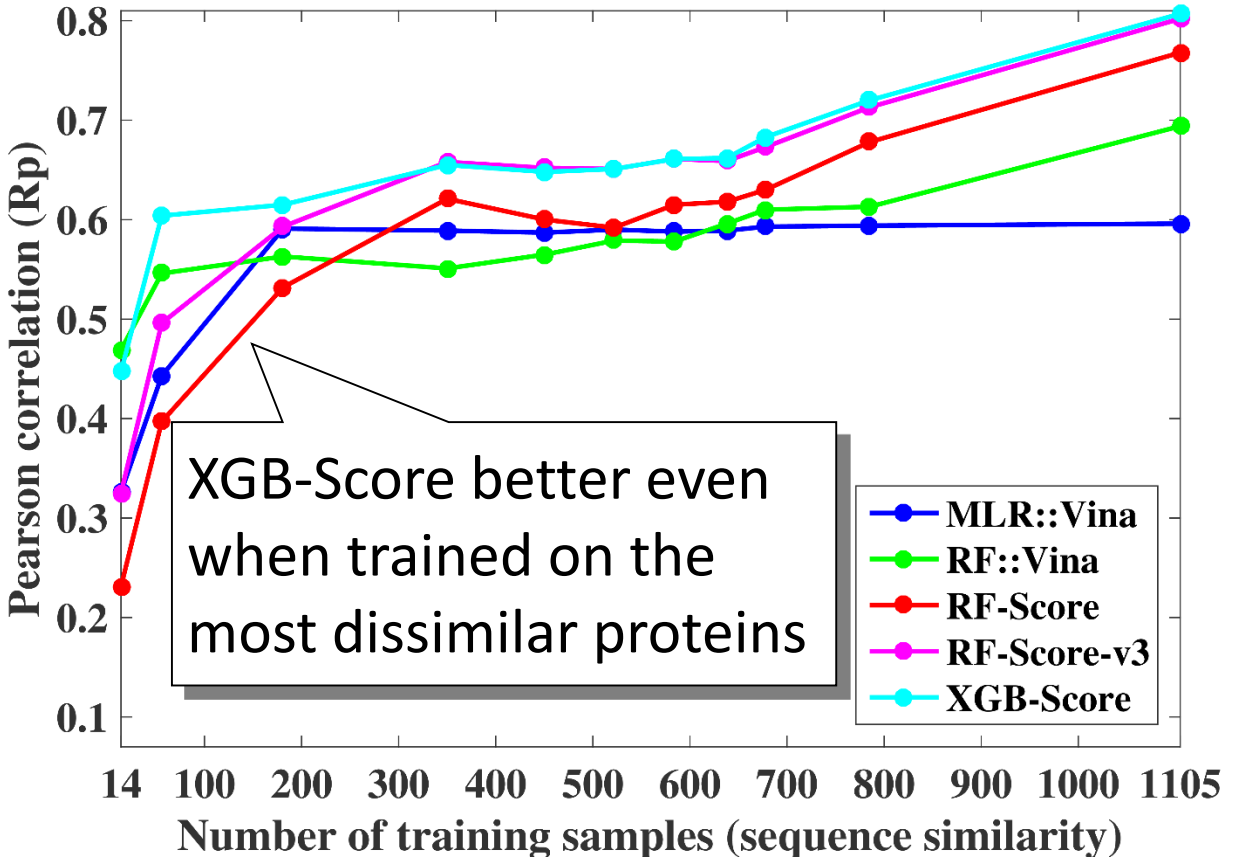


If we allow 65 complexes overlap → R_p=0.827

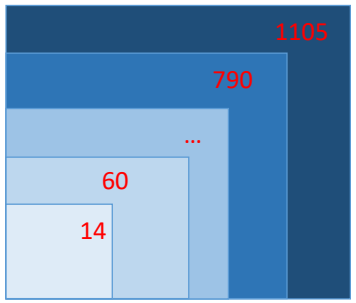
ΔR_p=+0.051

How well SFs predict complexes w/dissimilar proteins





CASF-2007 core (195 complexes): the same test set for all points
Each point ≠ nested training set or learning algorithm. Left: training complexes with the most dissimilar subsets of proteins

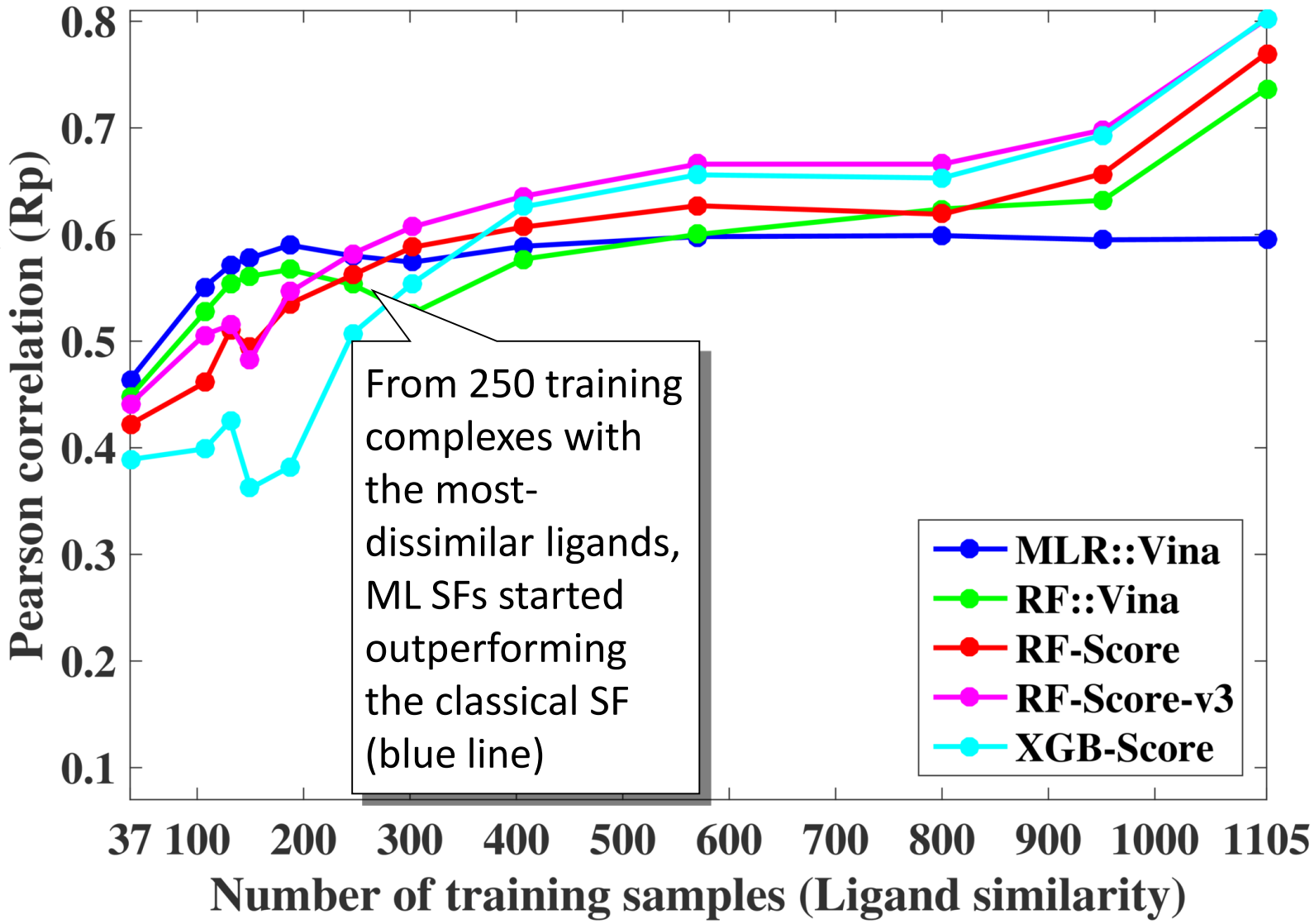


Source: https://academic.oup.com/bioinformatics/article/35/20/3989/5380764

How well SFs predict complexes w/dissimilar ligands?

These ligand molecules are of low diversity. Still much harder split than any of the protein splits (remember, XGB-Score was better despite training with dissimilar proteins)

Also known as "Cold Start", "Blind ligand", "Blind drug" setting



Source: https://academic.oup.com/bioinformatics/article/35/20/3989/5380764

Deep learning in 2025: SOTA features + graph NNs

AEV-PLIG

https://doi.org/10.1038/s42004-025-01428-y

Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data

Check for updates

Ísak Valsson [1,4], Matthew T. Warren [2,4], Charlotte M. Deane [1], Aniket Magarkar [3] ✉, Garrett M. Morris [1] ✉ & Philip C. Biggin [2] ✉

Deep learning in 2025: still rather incremental...

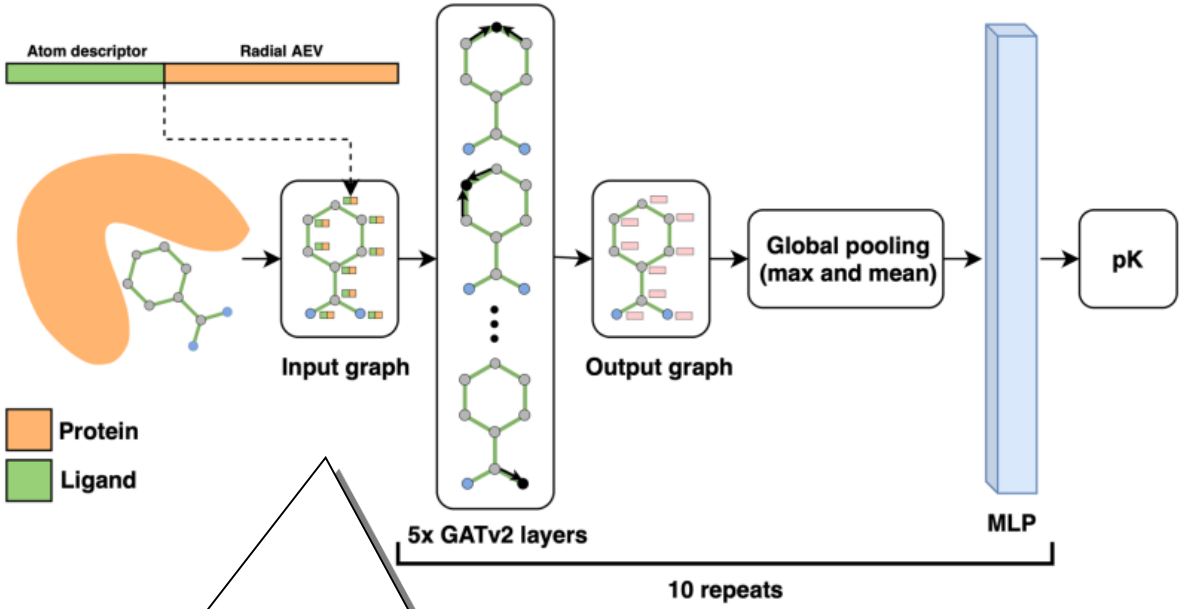


Table 1 | Performance metrics for AEV-PLIG and other state-of-the-art binding affinity predictors on five different benchmarks

Model	CASF-2016		OOD Test		0 Ligand Bias		FEP benchmark	
	PCC (↑)	K_{τ} (↑)	PCC (↑)	K_{τ} (↑)	PCC (↑)	K_{τ} (↑)	wmPCC (↑)	wmK_{τ} (↑)
RF-score	0.82	0.63	0.64	0.48	0.24	0.16	0.29	0.18
	[0.77, 0.85]	[0.58, 0.67]	[0.58, 0.70]	[0.42, 0.53]	[0.14, 0.34]	[0.09, 0.22]	[0.17, 0.40]	[0.10, 0.27]
Pafnucy	0.76	0.55	0.55	0.39	0.17	0.10	0.26	0.18
	[0.70, 0.80]	[0.50, 0.60]	[0.47, 0.62]	[0.32, 0.45]	[0.07, 0.28]	[0.03, 0.18]	[0.17, 0.36]	[0.12, 0.23]
OnionNet-2	0.83	0.64	0.57	0.39	0.35	**0.23**	0.27	0.18
	[0.79, 0.86]	[0.59, 0.68]	[0.49, 0.63]	[0.32, 0.45]	[0.23, 0.44]	**[0.16, 0.30]**	[0.18, 0.35]	[0.11, 0.25]
PointVS	0.80	0.60	0.69	0.51	0.28	0.17	0.33	0.22
	[0.76, 0.84]	[0.55, 0.64]	[0.63, 0.74]	[0.46, 0.56]	[0.19, 0.38]	[0.10, 0.24]	[0.23, 0.43]	[0.15, 0.29]
SIGN	0.82	0.63	0.66	0.49	0.27	0.18	0.37	**0.26**
	[0.78, 0.86]	[0.58, 0.67]	[0.59, 0.71]	[0.43, 0.54]	[0.16, 0.37]	[0.11, 0.25]	[0.29, 0.45]	**[0.21, 0.32]**
AEScore	0.85	0.64	0.63	0.44	0.26	0.17	0.28	0.21
	[0.81, 0.88]	[0.60, 0.68]	[0.57, 0.69]	[0.38, 0.49]	[0.15, 0.36]	[0.09, 0.23]	[0.19, 0.37]	[0.14, 0.28]
AEV-PLIG	**0.86**	**0.67**	**0.73**	**0.55**	**0.37**	0.21	**0.41**	**0.26**
	[0.82, 0.89]	**[0.62, 0.71]**	**[0.66, 0.77]**	**[0.50, 0.60]**	**[0.25, 0.49]**	[0.14, 0.29]	**[0.32, 0.49]**	**[0.20, 0.32]**

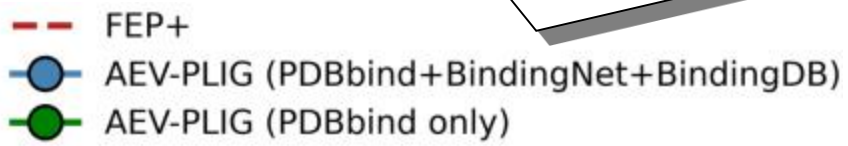
For each heavy atom in the ligand:
- PLIG (Protein-Ligand Interaction Graph) calculates 5 atom features (e.g. ring membership) and 22 protein interaction features (e.g. # of contacts with protein atoms of a given type with 6 Å).
- AEV (Atomic Environment Vectors) spatial distribution of protein atoms for 22 ECIF atom types at 16 radial distances.

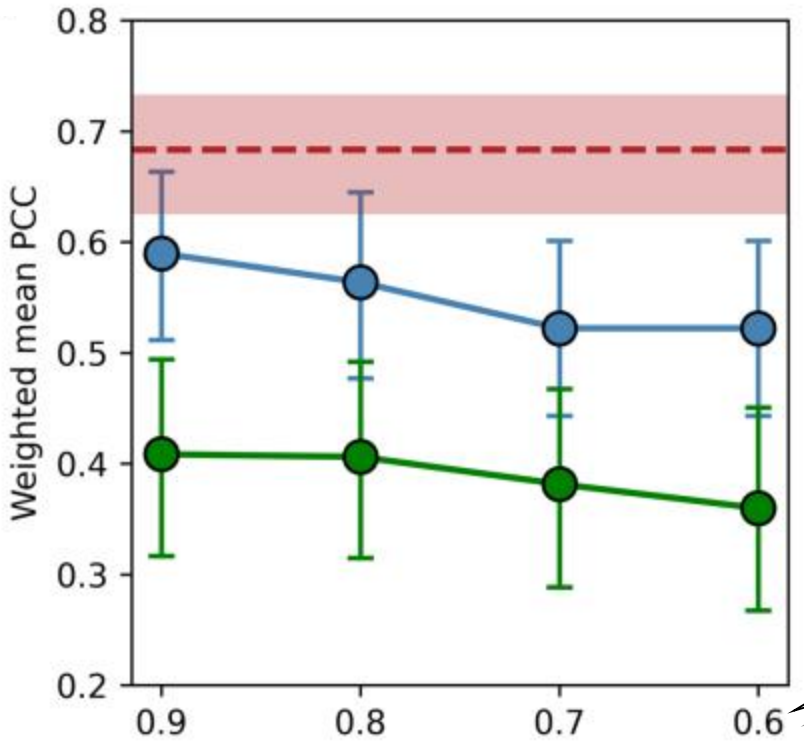
Interesting: when trained and rigorously evaluated on the same datasets, 2010's RF-score v1 vs AEV-PLIG PCCs 0.82 vs 0.86 (CASF-2016 core set)

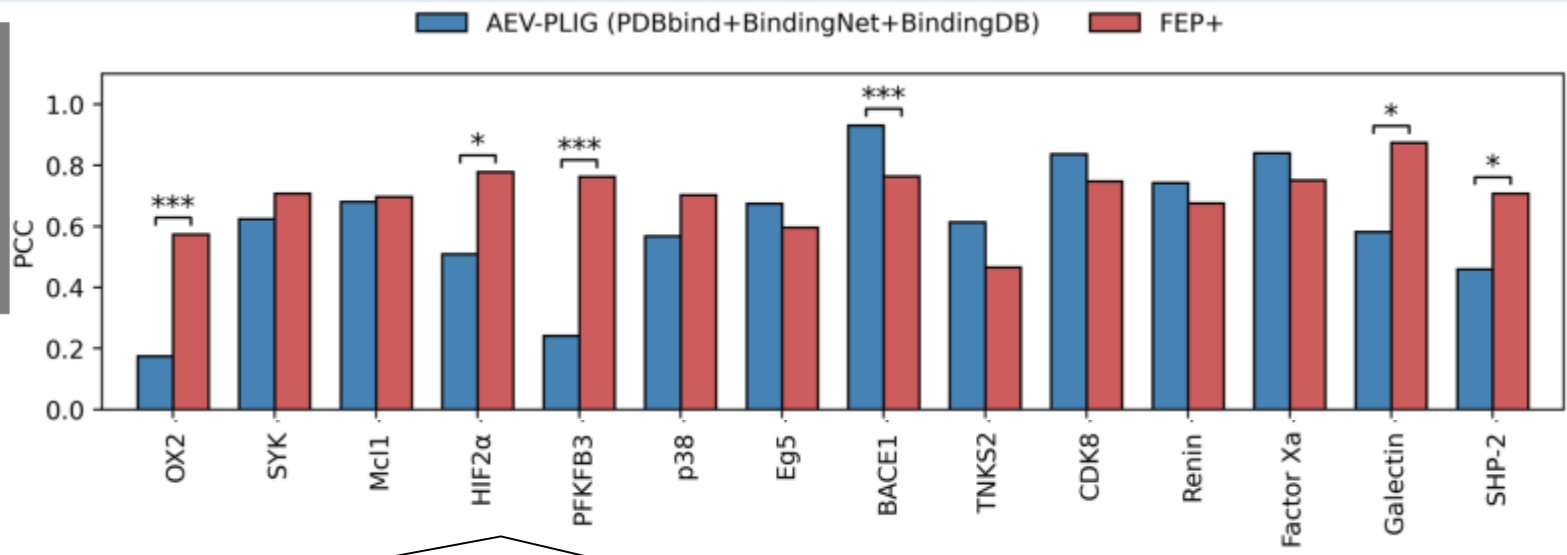
Same comparison on the most realistic benchmark: 2010's RF-score v1 vs AEV-PLIG PCCs 0.29 vs 0.41 (FEP test set).

...but data augmentation takes it to the next level

of training complexes: 19,443 (PBDbind), 69,816 (BindingNet), 8822 (BindingDB).
AEV-PLIG's weighted mean PCC from 0.41 to 0.59 vs FEP+'s weighted mean PCC of 0.68





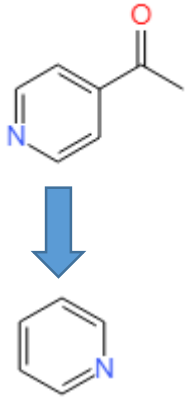


FEP benchmark has 14 targets with >= 25 congeneric ligands. Data-augmented AEG-PLIG approaches the accuracy of molecular dynamics (MD) simulation-based FEP+ method and is 400,000 times faster

Predictive even when only molecules with ECFP6-based Tanimoto similarity <= 0.6 are allowed in the training set

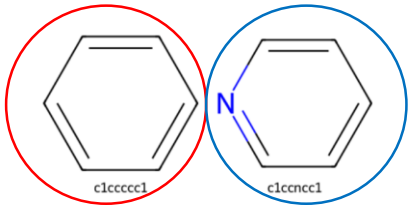
Scaffold split: unrealistically high train-test similarities!

Bemis-Murcko scaffold: core structure of a molecule by removing its side chain atoms and focusing on its central ring systems and linkers.

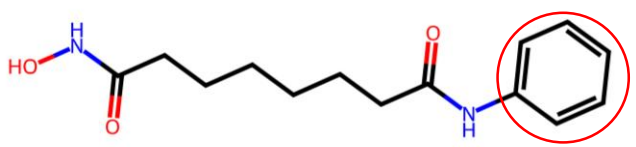


Scaffold split will often permit high similarities between training and test molecules (scaffolds different in a single atom, one scaffold containing the other) that rarely occur prospectively (massive diversity of screening libraries used as real-world test set)

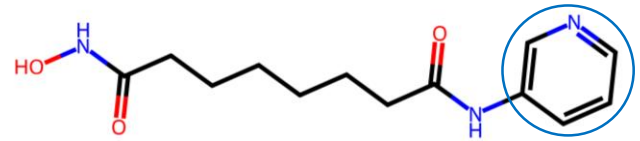
Frequent scaffolds among drugs



Scaffold split can place the molecule on the left in the training set and that on the right in the test set!



C(CCCCCC(=O)NO)C(=O)Nc1ccccc1



c1cc(cnc1)NC(=O)CCCCCCC(=O)NO

Guo et al. (2024) “Scaffold Splits Overestimate Virtual Screening Performance”. https://link.springer.com/chapter/10.1007/978-3-031-72359-9_5

Many retrospective papers, but which improve SOTA?

nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-026-01331-8

Guidelines for evaluating structure-based methods for protein–ligand binding prediction

Received: 15 January 2026

Accepted: 28 September 2026

Viet-Khoa Tran-Nguyen [1] & Pedro J. Ballester [2]

Evaluation criterion	Guidelines
Careful scrutiny of claims in preprints and social media	• Caution with strong claims that have not been peer reviewed • Superlative assertions demand greatest scrutiny • Marginal improvements must be statistically validated to distinguish stochastic variability from true performance gains
Training data full availability	• Explicit disclosure of all training and test data instances: protein-ligand structures and their binding affinities • Full source code released too to enable independent scrutiny • Rejecting "available upon request" data-sharing practices
Prevention of training-test data leakage and sampling biases	• Prevent data leakage by verifying that the test set was not used in any way to train or select the model • Avoid sampling bias by using realistic splits for the considered task • Use degradation curves to reveal generalization limits by mapping performance against training-test dissimilarity
Equal data partitioning	• Not only the test set, but also the training set and validation set, must be common to all the compared scoring functions • Alternative training and pre-training sets can also be used, but these should be fully disclosed and made available for others
Run-dependent variability assessment	• Report multiple independent runs, instead of just the best run, to enable reliable and robust comparisons of scoring functions • Report cross-validation (CV) metrics over multiple seeds to comprehensively evaluate all complexes and identify the hard-to-predict ones
Dataset growth vs AI algorithmic innovation	• New methods must be evaluated against the best non-deep learning ML scoring functions on the same data partition • Include simple baselines (e.g. ligand's heavy atom count or protein's cluster-membership affinity) to assess added value • How much of the performance gain is due to algorithmic innovation and how much is due to using more training data?
Benchmark-task alignment	• Align benchmark to task, e.g. leave-target-out validation to identify models that generalize to novel targets • A realistic benchmark can have high training-test similarity • Task-aligned benchmarking identifies best model for the task

To sum up: machine-learning SFs for BAP

These fast SFs are now highly predictive for the structure-based potency optimisation of drug leads across a range of targets

How to improve them further beyond the obvious action*?

1. User-friendly tools to facilitate the broadest adoption.
2. Training data augmentation without *, e.g. target-docked actives
3. Target-specific SFs which closely tailor the model to the target

*generating more data

nature protocols https://doi.org/10.1038/s41596-023-00885-w

A practical guide to machine-learning scoring for structure-based virtual screening

Viet-Khoa Tran-Nguyen [1], Muhammad Junaid[1], Saw Simeon [1] & Pedro J. Ballester [2]

Acknowledgements

IMPERIAL

Collaborators

University of Oxford (UK): Edith Sim, Graham Richards.
University of Cambridge (UK): Taufiq Rahman, Adrian Schreyer, Tom Blundell, John Mitchell, Chris Abell.
EMBL-EBI (UK): Janet Thornton.
The Francis Crick Institute (UK): Ghita Ghislat.
Chinese University of Hong Kong (China): Hongjian Li
University of Warsaw (Poland): Maciej Wojcikowski, Pawel Siedlecki
University of Paris Cité (France): Viet-Khoa Tran-Nguyen

Funding

MRC Medical Research Council
BBSRC bioscience for the future
THE ROYAL SOCIETY
EPSRC Engineering and Physical Sciences Research Council
FONDATION ARC POUR LA RECHERCHE SUR LE CANCER
ANR

Join our group for a postdoc

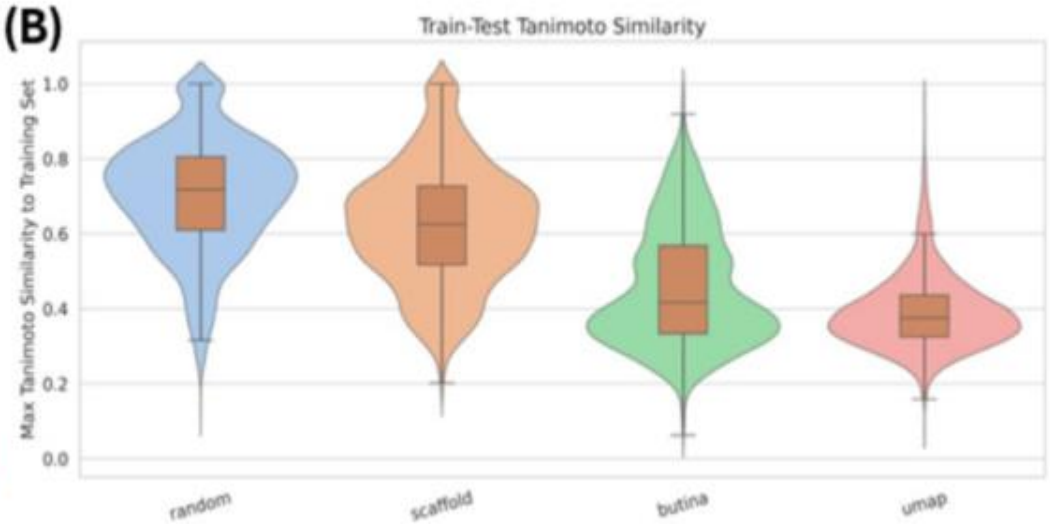
If interested, please quickly email me p.ballester@imperial.ac.uk with a CV with publications and a motivation letter. Deadline: 26 October 2026

Q&A

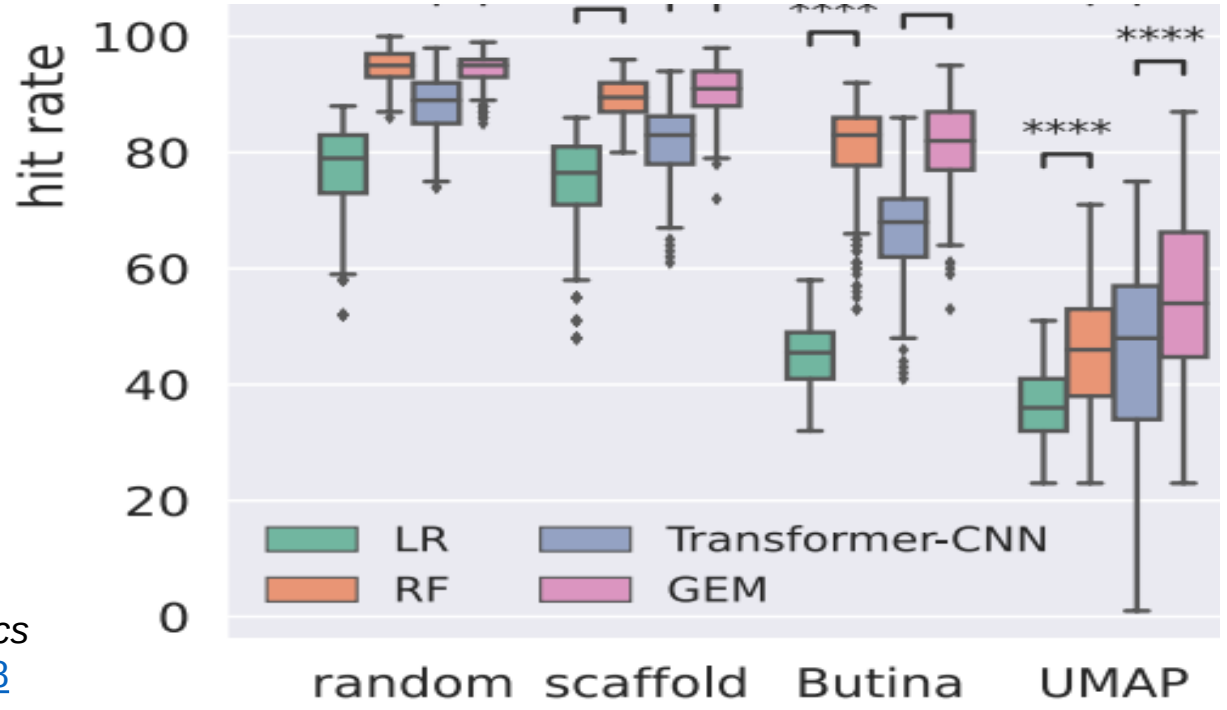


UMAP-based clustering split: the most challenging



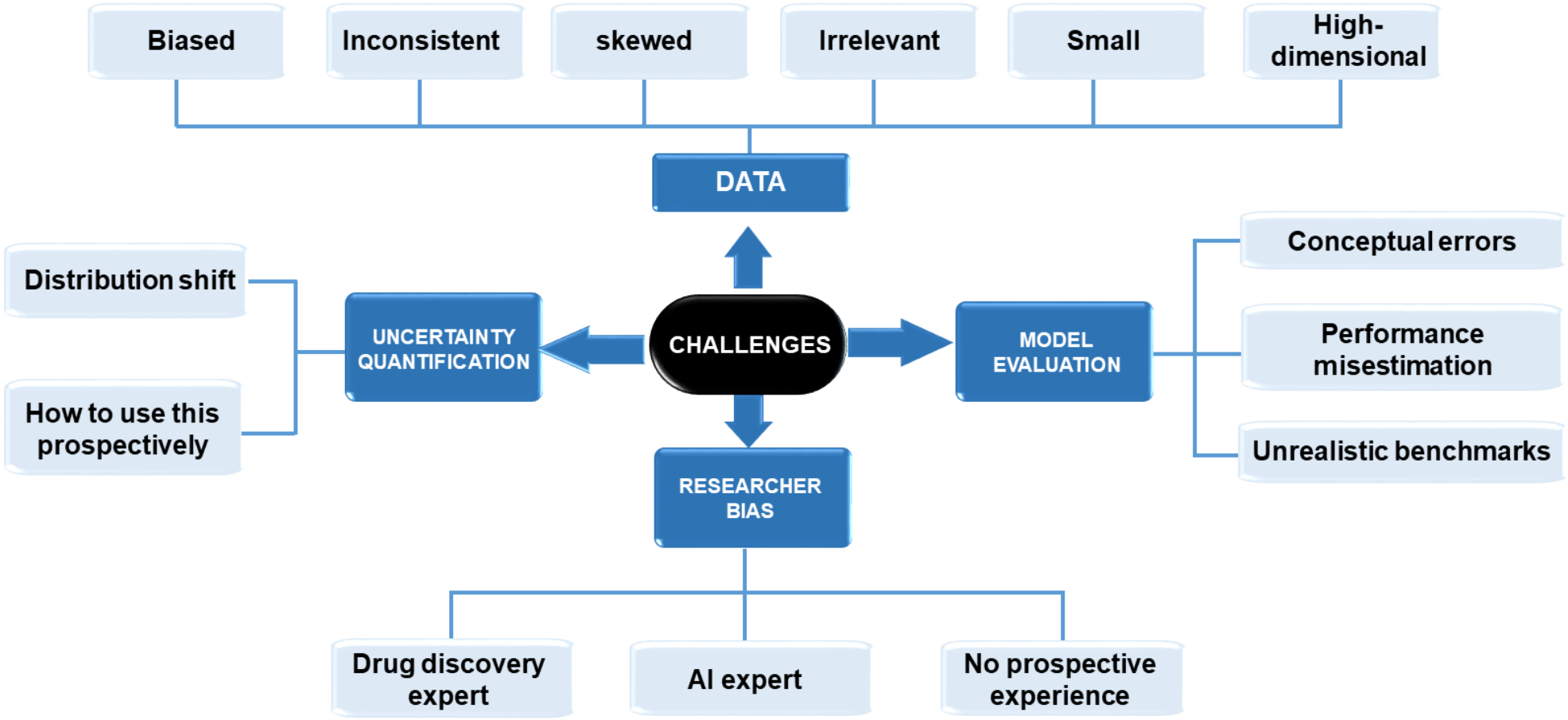


UMAP clustering: UMAP captures the manifold structure of the data in a topology-preserving manner assuming k clusters. Here outputs a two-dimensional embedding. K= 7 fods was optimal in this cancer application. fold = UMAP cluster



Guo et al. (2025) "UMAP-based clustering split for rigorous evaluation of AI models for virtual screening on cancer cell lines". *J. of Cheminformatics* https://jcheminf.biomedcentral.com/articles/10.1186/s13321-025-01039-8

Future progress by tackling data-centric challenges



Ghislat et al. (2024) "Data-centric challenges with the application and adoption of artificial intelligence for drug discovery" *Expert Opinion on Drug Discovery*. https://arxiv.org/abs/2407.05150