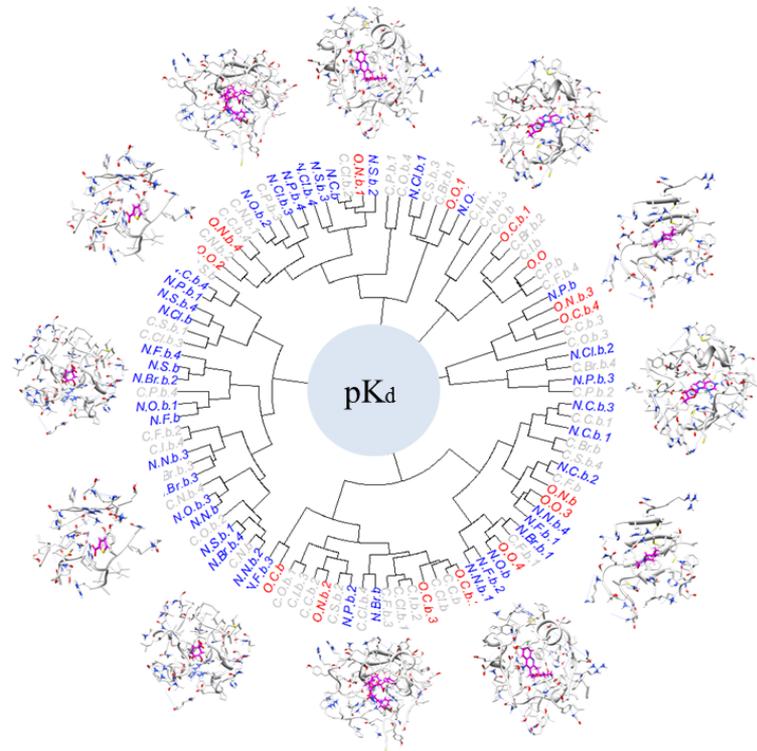


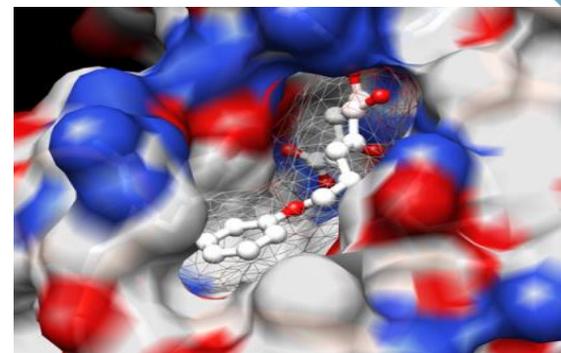
## ML scoring functions to improve structure-based binding affinity prediction and virtual screening

Dr Pedro Ballester

Group Leader at CRCM (France)



# Related docking applications



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

Predict whether a **molecule docked** to a given **target** is a **true binder** from the **docked structure**

## Virtual Screening (VS)

- \* Usually benchmarking as binary classification
- \* Goal: identifying new binders of a given target
- \* +ve and -ve instances
- \* Disadvantage: more confounding factors s.t. docking pose error, decoys assumed inactive, etc.

Predict the **affinity** of a **molecule bound** to a given **target** from the **crystal structure**

## Binding Affinity Prediction

- \* Usually benchmarking as regression problem
- \* Goal: determinations of  $K_d/K_i$  by ITC reduced
- \* only +ve instances (necessary condition VS)
- \* Advantage: any decrease in performance is due to worse modelling

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

*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<sup>1,\*</sup>† and John B. O. Mitchell<sup>2,\*</sup>

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

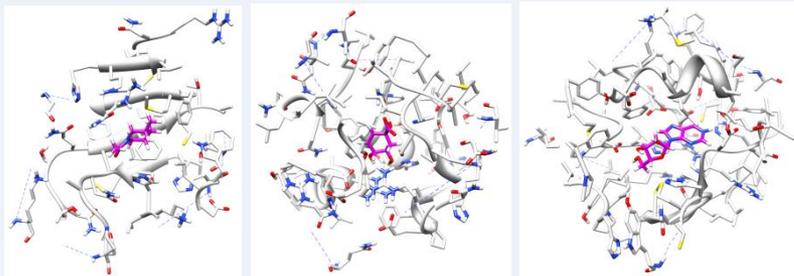
Associate Editor: Burkhard Rost

1. Generic: structures of proteins from other families may improve prediction (complement structures of target)
2. Providing that a sufficiently flexible regression model is used → Random Forest (Breiman, 2001)
3. Advantage: circumventing a priori assumptions about the SF's functional form may reduce modelling error



# Training and testing RF-Score

Training set (1105 complexes)

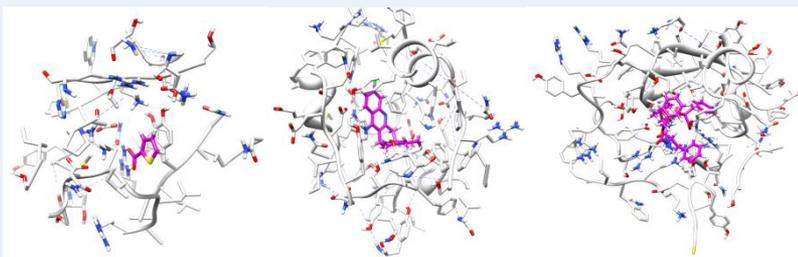


1w8l  
pK<sub>i</sub>=0.49

1gu1  
pK<sub>i</sub>=4.52

2ada  
pK<sub>i</sub>=13

Test set (195 complexes)



2hdq  
pK<sub>i</sub>=1.4

1e66  
pK<sub>i</sub>=9.89

7cpa  
pK<sub>i</sub>=13.96

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

pK <sub>d/i</sub>	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 <sub>d/i</sub>	C.C	-	C.I	N.C	-	I.I	PDB
1.40	858	-	0	0	-	0	2hdq
-	-	-	-	-	-	-	-
13.96	4476	-	0	283	-	0	7cpa

195

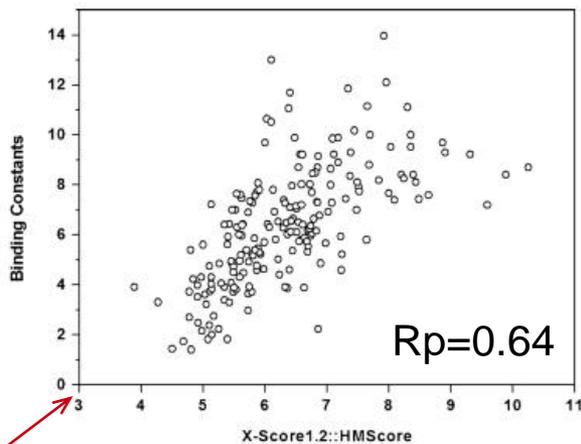
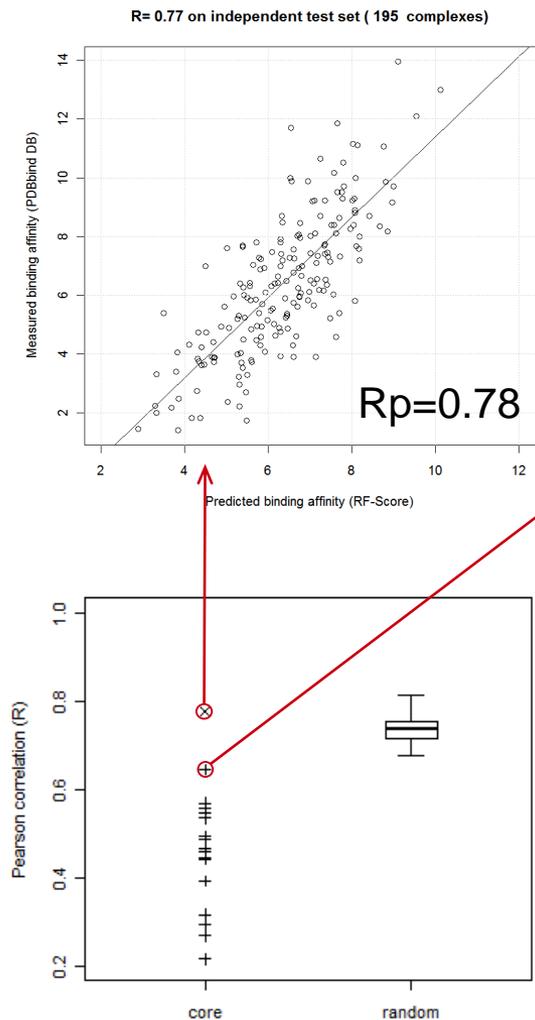
Random Forest (RF) training  
(descriptor selection, model selection)

RF-Score  
(description and training choices)

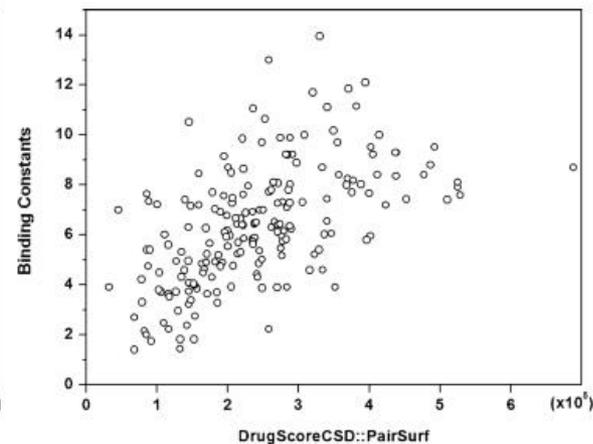
# RF-Score-v1 performance on diverse test set

COMPARATIVE ASSESSMENT OF SCORING FUNCTIONS

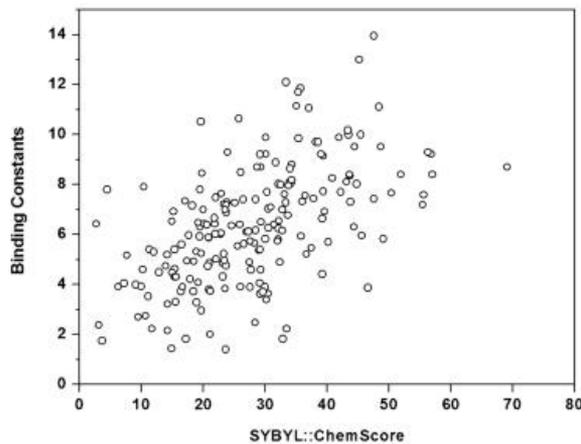
*J. Chem. Inf. Model.*, Vol. 49, No. 4, 2009 1087



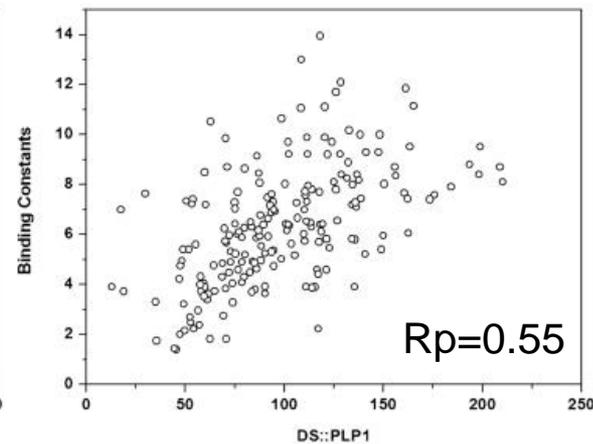
(a)



(b)



(c)



(d)



# Scoring Function Consortium using RF-Score code

JOURNAL OF  
CHEMICAL INFORMATION  
AND MODELING

Article

pubs.acs.org/jcim

2008

2013

## SFCscore<sup>RF</sup>: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes

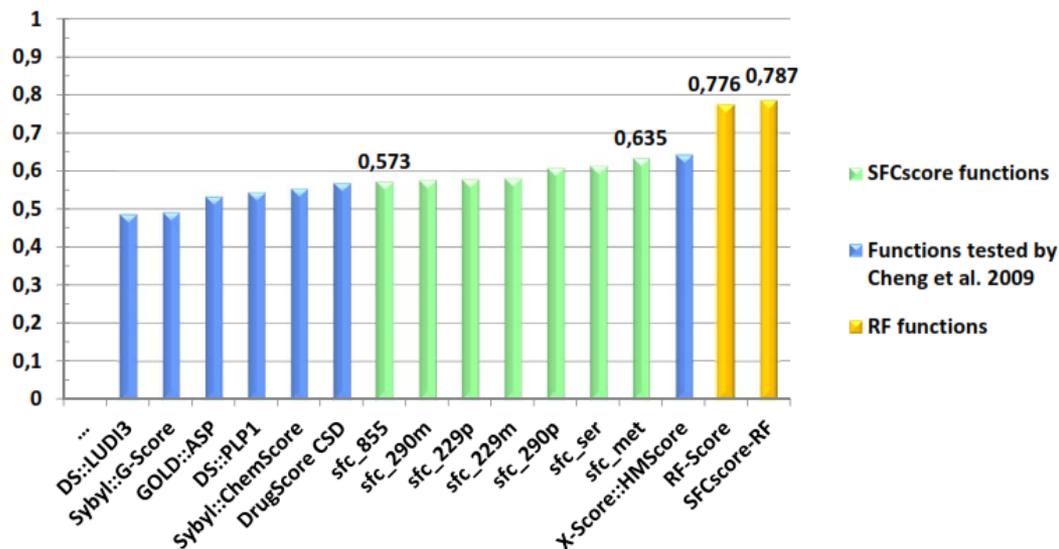
David Zilian and Christoph A. Sotriffer\*

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

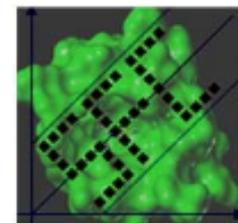
### SFCscore<sup>RF</sup>

Performance comparison: Cheng test set (195 complexes)

Pearson correlation coefficient  $R_p$



Scoring  
Function  
Consortium



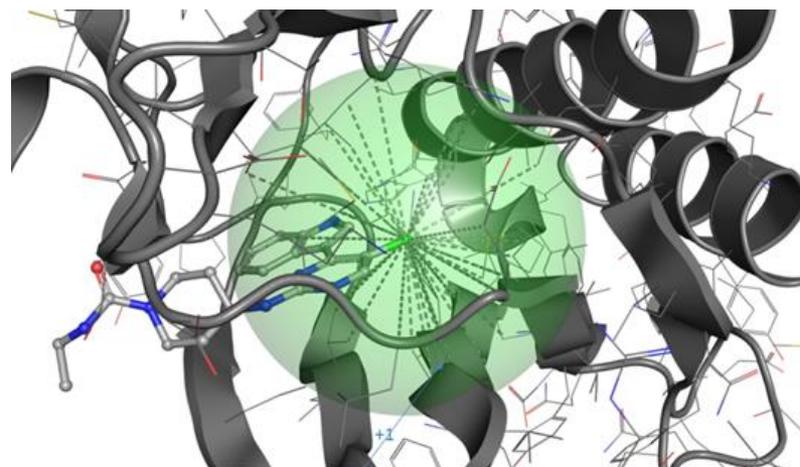
Astra	Aventis
BASF	Boehringer
Glaxo	Novo Nordisk
Pfizer	Agouron
Roche	Schering

CCDC

# Best structural descriptors? 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 <sub>in</sub>	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

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

Pedro J. Ballester,<sup>†,\*</sup> Adrian Schreyer,<sup>‡</sup> and Tom L. Blundell<sup>‡</sup>

<sup>†</sup>European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton - CB10 1SD, United Kingdom

<sup>‡</sup>Dept. of Biochemistry, University of Cambridge, 80 Tennis Court Rd, Cambridge - CB2 1GA, United Kingdom

# Analysing the improvement over classical SFs

Full Paper

www.molinf.com

molecular  
informatics

DOI: 10.1002/minf.201400132

## Improving AutoDock Vina Using Random Forest: The Growing Accuracy of Binding Affinity Prediction by the Effective Exploitation of Larger Data Sets

Hongjian Li,<sup>[a]</sup> Kwong-Sak Leung,<sup>[a]</sup> Man-Hon Wong,<sup>[a]</sup> and Pedro J. Ballester<sup>\*[b, c]</sup>

Classical SFs

additive functional form:

$$p = \sum_{m=1}^M w_m x_m$$

DOCK (force-field SF):

$$E_{bind} \equiv \sum_{k=1}^K \sum_{l=1}^L \left( \frac{A_{kl}}{d_{kl}^{12}} - \frac{B_{kl}}{d_{kl}^6} \right) + \sum_{k=1}^K \sum_{l=1}^L \left( 332.0 \frac{q_k q_l}{\epsilon(d_{kl}) d_{kl}} \right)$$

PMF (knowledge-based SF):

$$PMF \equiv \sum_i \sum_j \sum_{k=1}^{K_j} \sum_{l=1}^{L_i} H_{kl}(d_{kl})$$

X-Score (empirical SF):

$$\Delta G_{bind} \equiv w_0 + w_1 \Delta G_{vdW} + w_2 \Delta G_{hBonds} + w_3 \Delta G_{rotor} + w_4 \Delta G_{hydrophob}$$

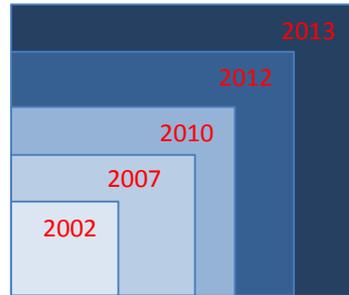
Machine-learning SFs

$$p = f_{RF}(x_m) \quad x_{ij} \equiv \sum_{k=1}^{K_j} \sum_{l=1}^{L_i} \Theta(d_{cutoff} - d_{kl})$$

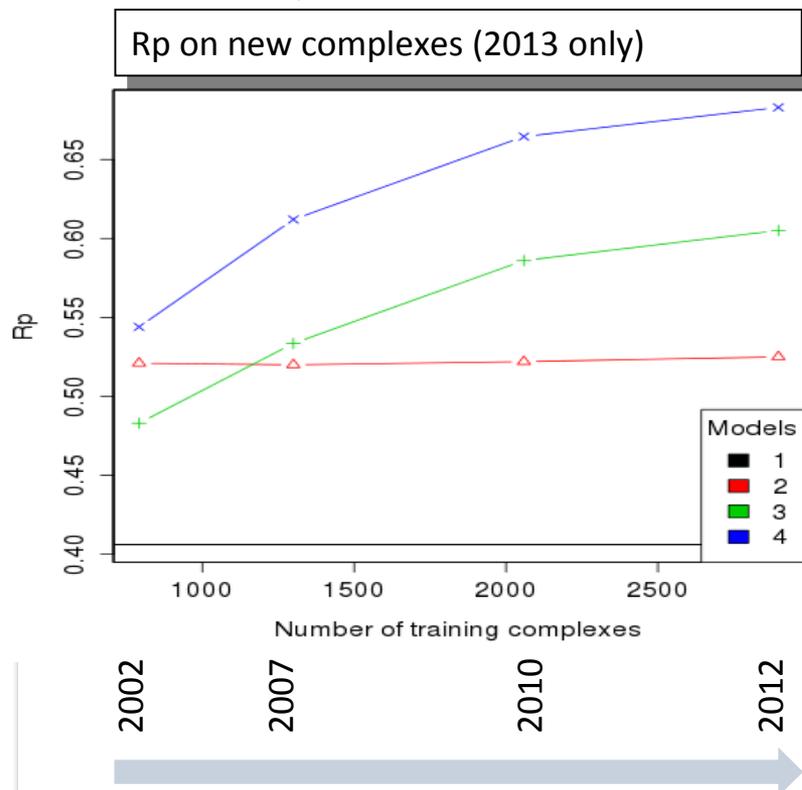
1. Autodock Vina off-the-shelf as a baseline (model 1)
2. Retrained Vina using its 6 features and MLR (model 2)
3. Retrained Vina using its 6 features and RF (model 3)
4. RF training on a merged vector with the 42 RF-Score v1 + Vina features (model 4)

# The effect of training with more data in all models

From 5 releases of the PDBbind database



Generated 4 time-stamped data partitions with same test set (new structures released in 2013)



**model 2** vs **model 3**: substituting MLR by RF using Vina features → MLR does not improve with more data, but RF does!

**model 3** vs **model 4**: increasing the number of features also beneficial with RF

**model 2** vs **model 4**: large gain over classical SFs due to RF being able to assimilate larger feature and data sets

**model 4**

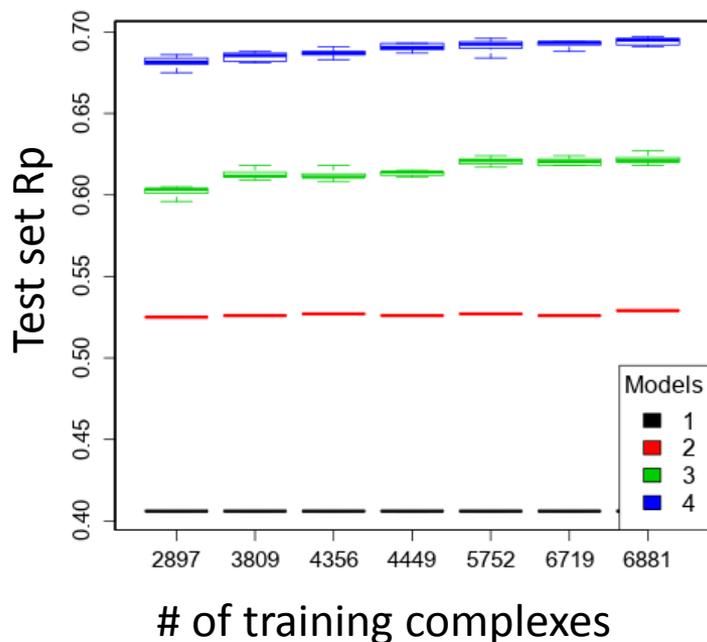
→ released as RF-Score v3

# Trade-off quality vs quantity of training data

## Low-Quality Structural and Interaction Data Improves Binding Affinity Prediction via Random Forest

Hongjian Li <sup>1</sup>, Kwong-Sak Leung <sup>1</sup>, Man-Hon Wong <sup>1</sup> and Pedro J. Ballester <sup>2,\*</sup>

Molecules 2015, 20, 10947-10962; doi:10.3390/molecules200610947

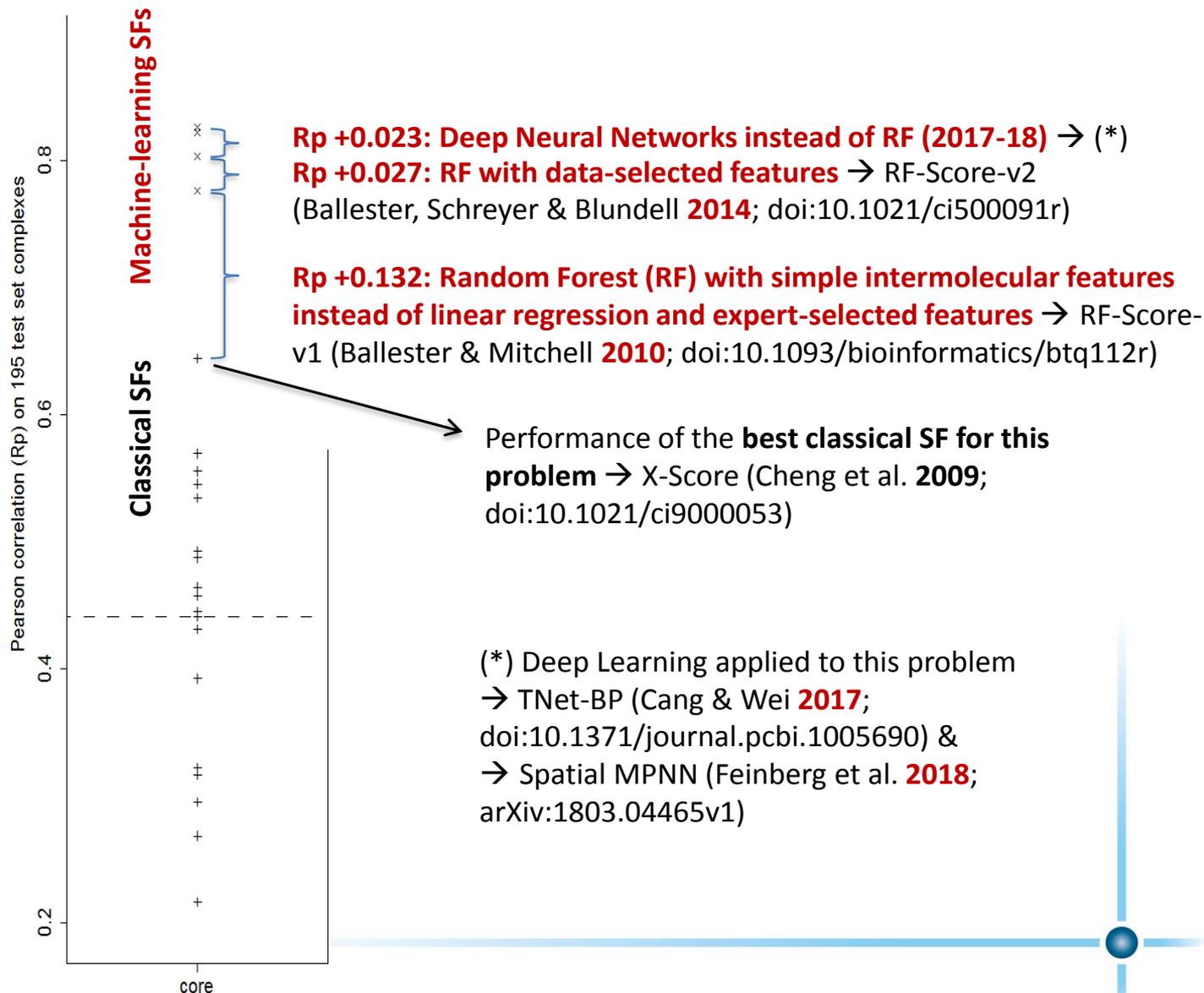


Test Set	Training Sets
	refined12 (2897)
	general12_Kd/KiOnly $\leq 2.5$ Å (3809)
	general12_Kd/KiOnly $\leq 3.0$ Å (4356)
refined13\refined12 (382)	general12_Kd/KiOnly (4449)
	general12 $\leq 2.5$ Å (5752)
	general12 $\leq 3.0$ Å (6719)
	general12 (6881)

Exploiting a larger data volume is more important for the performance of RF-Score than restricting to a smaller set of higher data quality

# SFs tested on PDBbind benchmark (now CASF2007)

scoring function	$R_p$
RF-Score::Elem-v2	0.803
RF-Score::Elem-v1	0.776
X-Score::HMScore	0.644
DrugScore <sup>CSD</sup>	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



# RF-Score codes for binding affinity prediction

---

- RF-Score v4 (trained on 3441 complexes with 47 features):  
<http://ballester.marseille.inserm.fr/rf-score-4.tgz>
- RF-Score v3 (trained on 2959 complexes with 42 features):  
<http://ballester.marseille.inserm.fr/rf-score-3.tgz>
- RF-Score v2 (python code to generate v2 descriptors):  
<https://bitbucket.org/aschreyer/rfscore>
- RF-Score v1 (C code to generate v1 descriptors and R scripts):  
<http://ballester.marseille.inserm.fr/RF-Score-v1.zip>

# Machine-learning SFs for structure-based VS

**REVIEW including both classes of ML SFs up to 2015:** Ain et al. Wiley Interdiscip Rev Comput Mol Sci. 2015 Nov-Dec; 5(6): 405–424. doi: 10.1002/wcms.1225

**REST OF TALK:** RF-Score-VS

**Machine-learning SFs for VS** are not trained on crystal structures, but trained on much larger numbers of +ves (docking poses of actives) and -ves (docking poses of inactives)

## Virtual Screening (VS)

- \* Usually benchmarking as binary classification
- \* Goal: identifying new binders of a given target
- \* +ve and -ve instances
- \* Disadvantage: more confounding factors s.t. docking pose error, decoys assumed inactive, etc.

## Binding Affinity Prediction

- \* Usually benchmarking as regression problem
- \* Goal: determinations of Kd/Ki by ITC reduced
- \* only +ve instances (necessary condition VS)
- \* Advantage: any decrease in performance is due to worse modelling

**Machine-learning SFs (trained only on actives) are not optimal to apply to VS** (due to their high proportion of inactives)

# DUD-E: Generating data to build ML SFs for VS

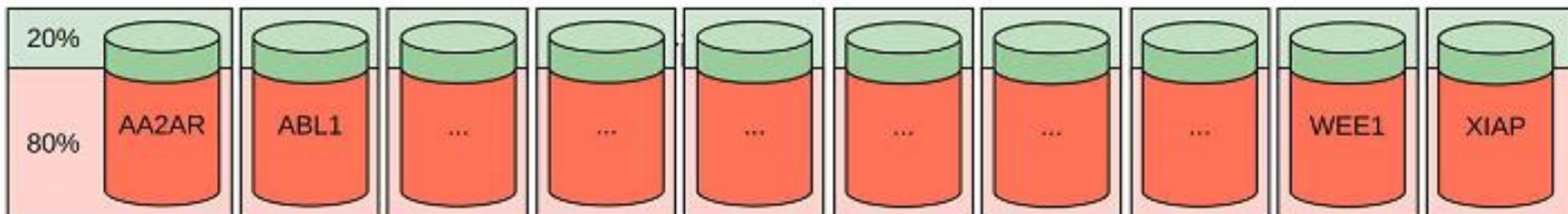
---

- DUD-E: Mysinger et al. 2012 ([dx.doi.org/10.1021/jm300687e](https://doi.org/10.1021/jm300687e))
- 102 protein targets: on average 224 actives per target with their reported activities + 50 decoys per active ( $\leq 1\mu\text{M}$ )
- Decoys: assumed inactive ( $\sim$ physicochemical to actives, but dissimilar chemical structure to reduce likelihood of being active)
- After docking with Smina implementation of Vina, 50 docking poses x (15 426 actives and 893 897 decoys) across targets
- 50 poses per molecule, but only kept best for training.
- Three RF-Score features: v1 (2010), v2 (2014) and v3 (2015).
- i.e. 909,323 data instances to train and validate each SF.
- Available at <https://wojcikowski.pl/rfscorevs/data/>

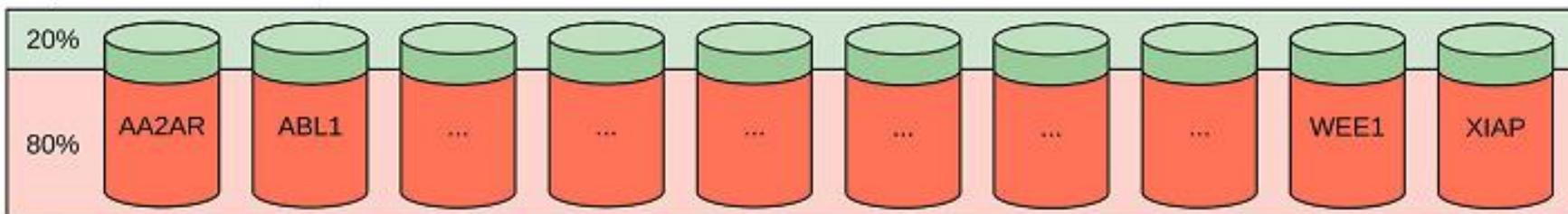
# Cross-validating machine-learning SFs for VS



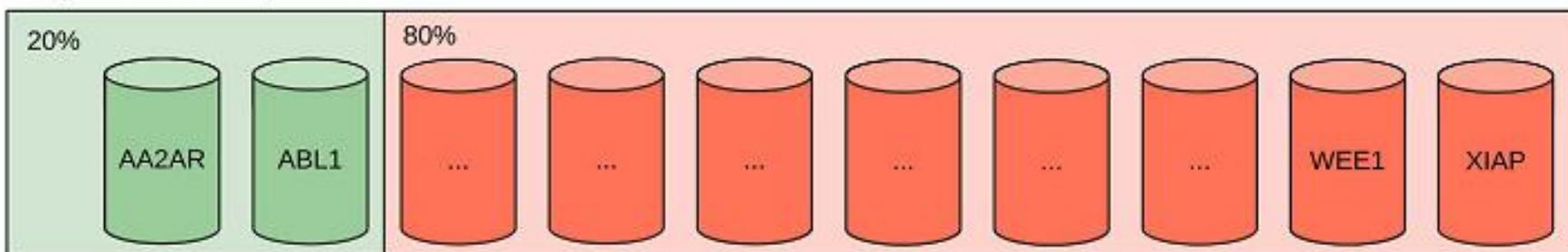
A) Per-Target Split: i.e. some actives known for the target



B) Horizontal Split : i.e. some actives known for all 102 targets



C) Vertical Split: i.e. absolutely no actives known for the target



# DUD-E: Performance of classical and ML SFs for VS

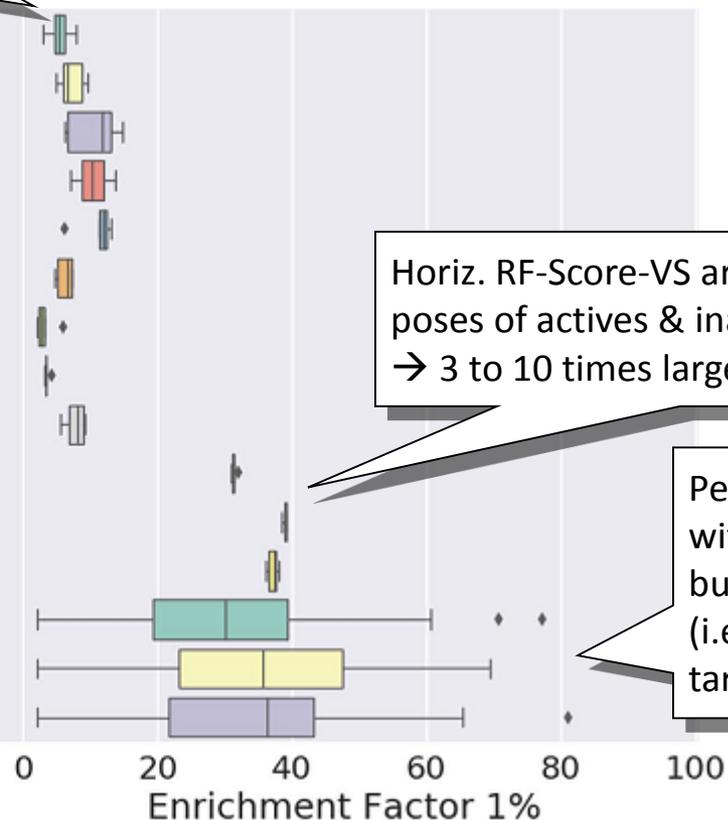
RF-Score v3 is trained on PDBbind crystal structures  
(not on docked poses of actives and inactives)

Docking engine  
(similar results  
with DOCK3.6  
& DOCK6.6)

Averages in 5  
folds:  $EF_{1\%}=38.96$   
ROC AUC=0.84  
across all targets

Autodock Vina

RF-Score v3  
Vertical RF-Score-VS v1  
Vertical RF-Score-VS v2  
Vertical RF-Score-VS v3  
Autodock Vina Score  
Chemscore  
D score  
G score  
PMF score  
Horizontal RF-Score-VS v1  
Horizontal RF-Score-VS v2  
Horizontal RF-Score-VS v3  
Per Target RF-Score-VS v1  
Per Target RF-Score-VS v2  
Per Target RF-Score-VS v3

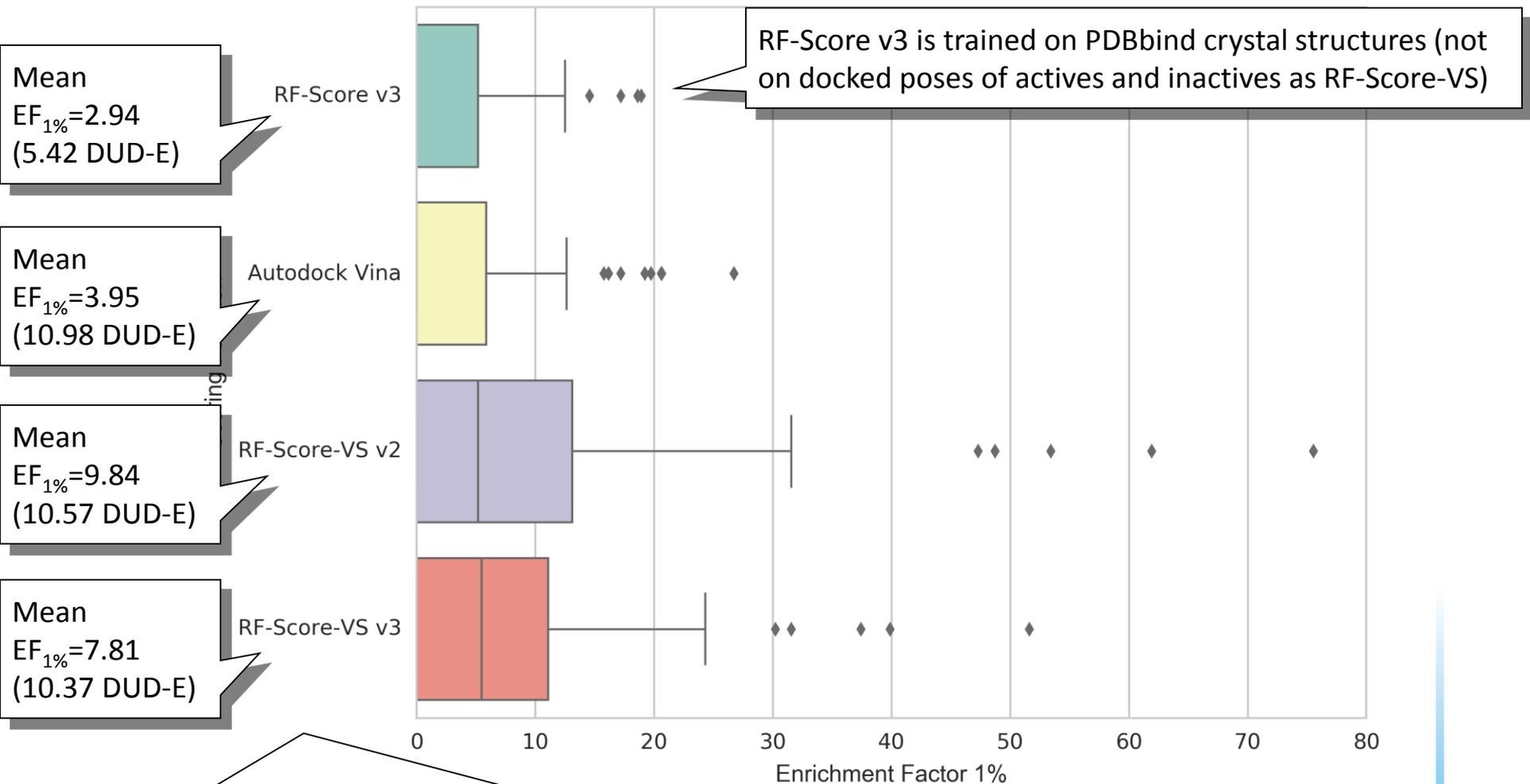


# DEKOIS: validating RF-Score-VS on unseen targets

---

- DEKOIS 2.0: Bauer et al. 2013 (dx.doi.org/10.1021/ci400115b)
- 81 targets: we used 76 targets (4 in DUD-E, 1 w/out crystal struct.)
- filtered out any nearly identical ligand or decoy to any ligand/decoy present in DUD-E (tanimoto score of at least 0.99; OpenBabel FP2 fingerprints).
- Each DEKOIS target has 40 ligands and 1200 decoys, our pruning removed on average 18.6 (46.5%) ligands and 188 (15.7%) decoys.
- retained DEKOIS 2.0 ligands & decoys were docked using Autodock Vina with default settings, as previously done with DUD-E.
- Re-score with RF-Score-VS trained on entire DUD-E data > available at [https://github.com/oddt/rfscorevs\\_binary](https://github.com/oddt/rfscorevs_binary)
- NEXT: reporting performance of RF-Score-VS on DEKOIS data

# DEKOIS: validating RF-Score-VS on unseen targets



Even less favourable than the 'vertical split' scenario (RF-Score-VS is applied to a not included in the training set): here targets, actives & inactives in this test set were not in the training set

# Summary

---

- Machine-learning SFs shown to be more accurate than classical SFs at predicting pKd of diverse protein-ligand complexes
- The performance of RF-Score improves with training set size, but not that of classical SFs (MLR-based) → gap will broaden
- Regarding structural descriptors, data-driven Feature Selection (FS) leads to more predictive SFs than knowledge-based FS
- A target-specific RF-Score-VS can be built for any target with at least one ligand-bound crystal structure and some actives
- Much room for improvement: other ML algorithms, more targets and training data, identifying the best SF for each target...

# Acknowledgements

## Collaborators

- Hongjian Li, Kwong-Sak Leung, Man-Hon Wong (Chinese University of Hong Kong)
- Adrian Schreyer, Tom Blundell (University of Cambridge)
- John Mitchell (University of Cambridge)
- Maciek Wójcikowski (University of Warsaw)
- Pawel Siedlecki (University of Warsaw)

## Funding

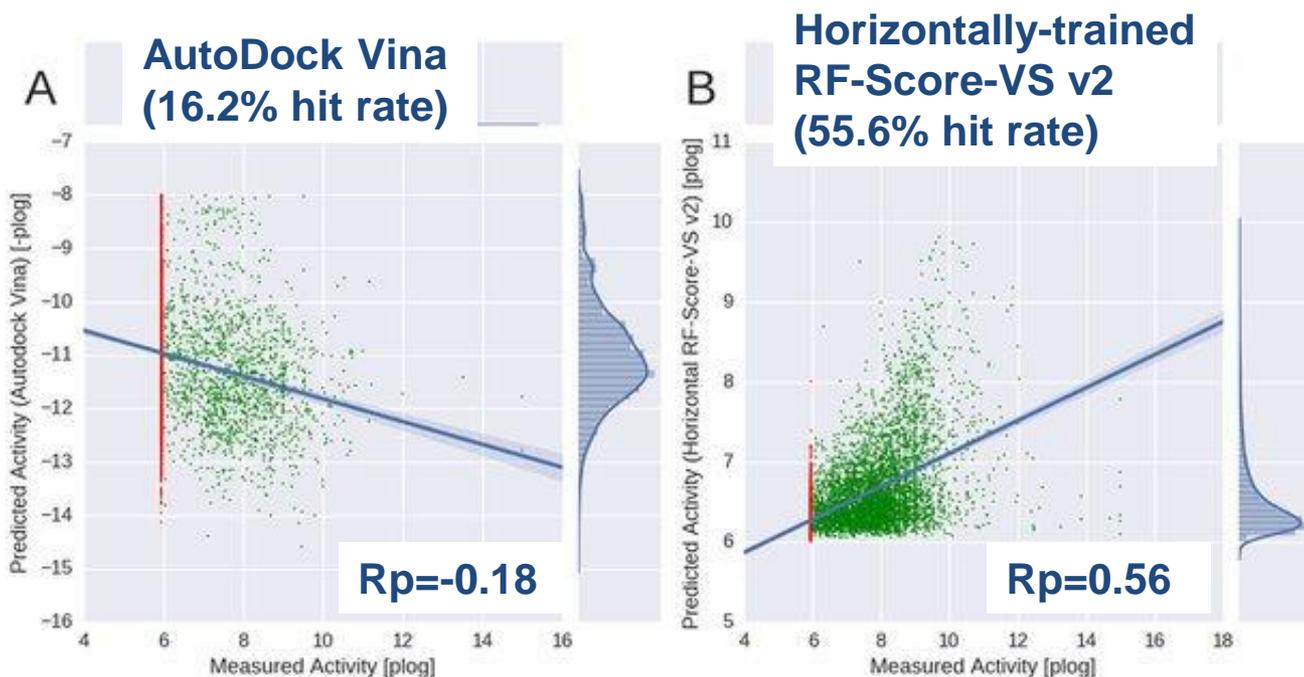


Institut national  
de la santé et de la recherche médicale



# DUD-E: scoring in the presence of inactives

- **RF-Score-VS: need to exploit inactive data too**
  - top 1% of all docked molecules ranked by predicted binding affinity (from 5CV of RF-Score-VS v2, directly with Vina)
  - **Red** points indicate **inactive** compounds (**false positives**), **green** points are **actives** (**true positives**) within top 1% of each SF



# Prospective application of RF-Score v1 (2012)

## Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification

Pedro J. Ballester<sup>1,\*†</sup>, Martina Mangold<sup>2,†</sup>, Nigel I. Howard<sup>2</sup>, Richard L. Marchese Robinson<sup>2</sup>, Chris Abell<sup>2</sup>, Jochen Blumberger<sup>3</sup> and John B. O. Mitchell<sup>4</sup>

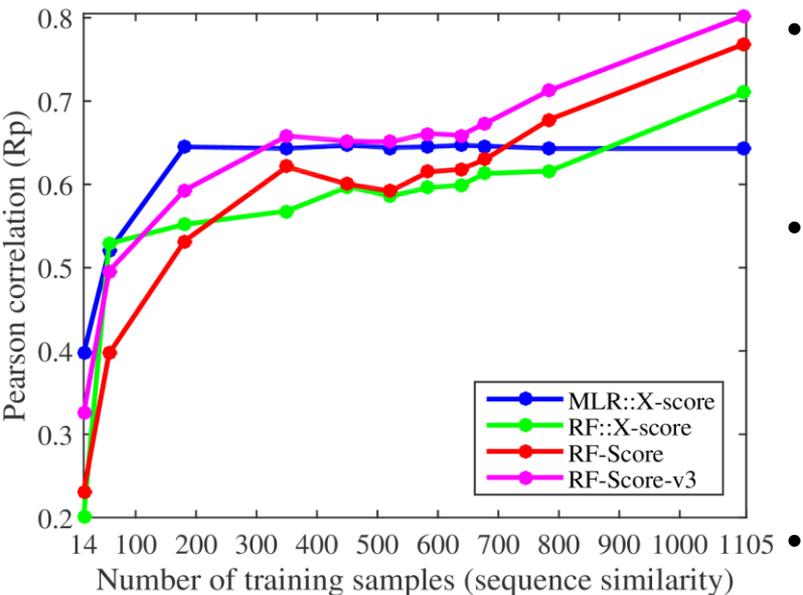
- DHQase-2: only three known active scaffolds (1 from HTS)
- Hierarchical VS: USR (3) on 9M cpds > GOLD on 4K USR hits → RF-Score → test top 148 cpds (N. Howard, Univ of Cambridge)
- Very high hit rates of ~ 25% with  $K_i \leq 100 \mu\text{M}$  → 100 new and structurally diverse actives (£5,000 cost)

Overall Performance	$K_i \leq 100 \mu\text{M}$	$K_i \leq 250 \mu\text{M}$	$(L^1, L^2, L^3)[\mu\text{M}]$
Against Mtb DHQase	35 (23.6%)	89 (60.1%)	(23, 24, 40)
Against Scl DHQase	40 (27.0%)	91 (61.5%)	(4, 21, 29)

# Impact of similarities between training/test proteins

## The Impact of Protein Structure and Sequence Similarity on the Accuracy of Machine-Learning Scoring Functions for Binding Affinity Prediction

Hongjian Li <sup>1,2,3</sup> , Jiangjun Peng <sup>2,4</sup>, Yee Leung <sup>2</sup>, Kwong-Sak Leung <sup>2,3</sup>, Man-Hon Wong <sup>3</sup>, Gang Lu <sup>5</sup> and Pedro J. Ballester <sup>6,7,8,9,\*</sup> *Biomolecules* 2018, 8(1), 12; doi:10.3390/biom8010012



- Test set from PDBbind benchmark (2007 core set with 195 complexes). Each point is the Rp of a SF on this common test set
- Each SF trained on nested training sets (e.g. 1<sup>st</sup> training set with 14 complexes comes from removing all training complexes with protein sequence similarity > 0.2 to the protein of at least one test complex. Larger sets obtained with higher thresholds
- Unlike X-Score, RF-Score improves when the complexes with the most similar proteins are included in the training set, but not only from them!

# Train/test on crystal structures vs docking poses

## Correcting the impact of docking pose generation error on binding affinity prediction



Hongjian Li<sup>1</sup>, Kwong-Sak Leung<sup>1</sup>, Man-Hon Wong<sup>1</sup> and Pedro J. Ballester<sup>2,3,4,5\*</sup>

BMC Bioinformatics 2016, 17(Suppl 11):308; doi:10.1186/s12859-016-1169-4

**Table 1** Performance of the four models trained on crystal and docked poses and tested also on crystal and docked poses (schemes 1 and 2) on the PDBbind v2007 benchmark. Comparing the same models from the two first blocks (crystal:crystal and crystal:docked) shows that the pose generation error also introduces a small degradation in the test set performance. Making the same comparisons between the second and fourth blocks shows that a substantial part of this error has been corrected

Model	Training	Test	RMSE	SD	Rp	Rs
1 (Vina)	Crystal	Crystal	2.41	1.99	0.554	0.608
2 (MLR:Vina)	Crystal	Crystal	1.88	1.85	0.630	0.680
3 (RF:Vina)	Crystal	Crystal	1.66	1.59	0.744	0.752
4 (RF:VinaElem)	Crystal	Crystal	1.52	1.42	0.803	0.799
1 (Vina)	Crystal	Docked	2.02	1.98	0.557	0.597
2 (MLR:Vina)	Crystal	Docked	1.90	1.87	0.622	0.670
3 (RF:Vina)	Crystal	Docked	1.76	1.72	0.693	0.710
4 (RF:VinaElem)	Crystal	Docked	1.60	1.52	0.772	0.771
2 (MLR:Vina)	Docked	Crystal	1.91	1.88	0.618	0.648
3 (RF:Vina)	Docked	Crystal	1.74	1.69	0.705	0.716
4 (RF:VinaElem)	Docked	Crystal	1.58	1.45	0.794	0.790
2 (MLR:Vina)	Docked	Docked	1.86	1.83	0.640	0.667
3 (RF:Vina)	Docked	Docked	1.69	1.63	0.730	0.730
4 (RF:VinaElem)	Docked	Docked	1.55	1.45	0.795	0.789

- Binding affinity prediction is often carried out on the docked pose of a known binder rather than its co-crystallised pose.
- Our results suggest that pose generation error is in general far less damaging for binding affinity prediction than it is currently believed.
- Another contribution of our study is the proposal of a procedure that largely corrects for this error.
- The resulting machine-learning scoring function, RF-Score v4 is freely available at <http://ballester.marseille.inserm.fr/rf-score-4.tgz>

# Support Vector Regression (SVR)-Score

The SVR RBF kernel implementation in the caret package [34] of the statistical software suite R was used. As with previous studies [16], grid search was conducted on the gamma parameter in the RBF kernel ( $\gamma$ ) and the cost of constraint violation parameter ( $C$ ) to give the best performance in a five-fold cross-validation of the training set. In each cross-validation, SVR was trained using the 36 combinations of parameter values arising from  $\gamma \in \{0.01, 0.1, 1, 10, 100, 1000\}$  and  $C \in \{0.25, 0.5, 1, 2, 4, 8\}$ . Thereafter, the average root mean square error between predicted and measured binding affinity across the five cross-validation sets (i.e. those not used to train the SVR) was calculated for each ( $\gamma, C$ ) combination and that with the lowest value was selected to train on the entire training set to give  $\text{SVR-Score} \equiv \text{SVR}(\gamma=0.1, C=1)$ . This model selection procedure is intended to find the model that is most likely to generalize to independent test data sets. When ran on the independent test set, SVR-Score achieved a Pearson's correlation of  $R=0.726$ , Spearman's correlation  $R_s=0.739$  and standard deviation  $SD=1.70$  as illustrated in Figure 2 (left).

[https://doi.org/10.1007/978-3-642-34123-6\\_2](https://doi.org/10.1007/978-3-642-34123-6_2)

