Machine learning approaches to predicting protein-ligand binding

Dr Pedro J Ballester
MRC Methodology Research Fellow
EMBL-EBI, Cambridge, United Kingdom

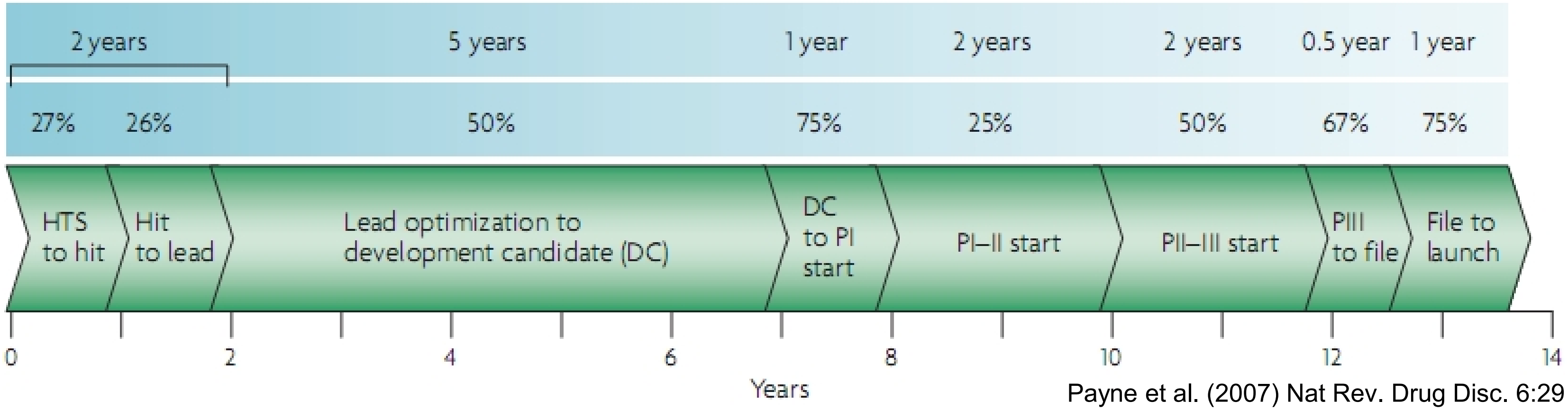
EMBL-EBI

EBI is an Outstation of the European Molecular Biology Laboratory.

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

The Drug Discovery Process



Payne et al. (2007) Nat Rev. Drug Disc. 6:29

- Developing new drug = average US$4 billion and 15 years

  http://www.forbes.com/sites/matthewherper/2012/02/10/the-truly-staggering-cost-of-inventing-new-drugs/

- While clinical trials are the most expensive stages, the research influencing approval the most at early stages:
  - Finding a target linked to the disease and a molecule modulating the function of target without trigering harmful side effects.

- Goal: finding drug leads for new targets (challenging)

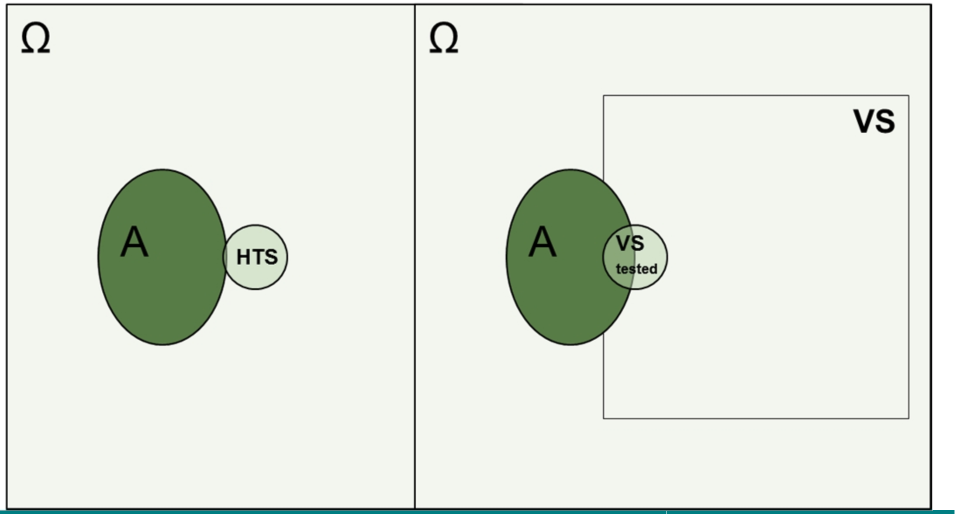
Virtual Screening: Why?

- HTS: Main strategy for identifying active molecules (hits) by wet-lab testing a library of molecules against a target.
- Computational methods (Virtual Screening) are needed:
  - HTS is slow: HTS of corporate collections → many months
  - HTS is expensive: Average cost US$1M per screen. Payne et al. 2007
  - Growing # of research targets → no HTS until target validation
- Limited diversity in HTS:

  HTS 10^6 cpds...

  but 10^{60} small molecules!

  (Dobson 2004 Nature)



- Target really undruggable?

Drug Design: goals

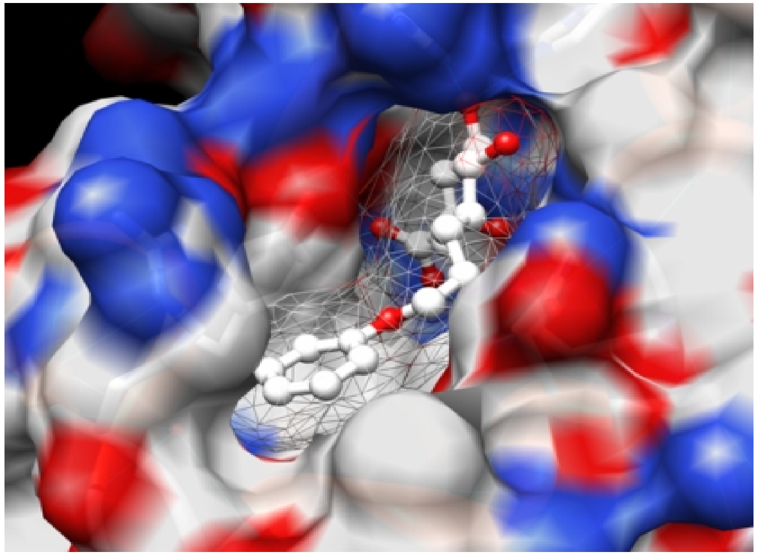
- Identifying active molecules among a large number of inactive molecules (i.e. extremely weak binders).
- Drugs must selectively bind to their intended target, as binding to other proteins may cause harmful side-effects
- Optimising selectivity: e.g. identify hits that occupy a subpocket that is not in related proteins w/≠ functions
- Increasing potency of the drug lead: predicting which analogues are more potent.
- How well these goals are met depend on the accuracy of structure-based tools for the considered target.

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

Docking

- If X-ray structure of the target is available → Docking:
  - predicting whether and how a molecule binds to the target.



- Docking = Pose generation + Scoring
  - Pose generation: estimating the conformation and orientation of the ligand as bound to the target.
  - Scoring: predicting how strongly the ligand binds to the target.
- Many relatively accurate algorithms for pose generation, but imperfections of scoring functions continue to be the major limiting factor for the reliability of docking.

Scoring Functions for Docking: functional forms

- 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)

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

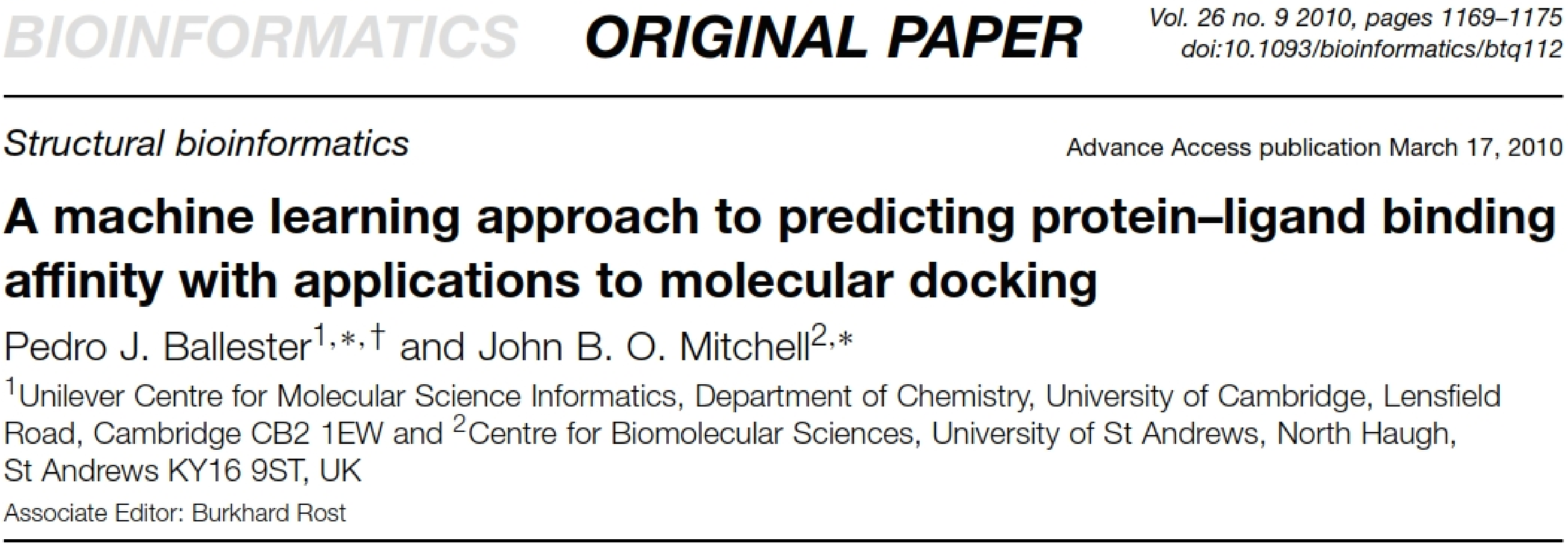
- SFs are trained on pK data usually through MLR:
  - FF (A_{ij}, B_{ij}), Emp(w_0,…,w_4) and sometimes KB ($\rho^{ij}_{\text{ref state.}}$)

Scoring Functions for Docking: limitations

- Two major sources of error affecting all SFs:
  1. Limited description of protein flexibility.
  2. Implicit treatment of solvent.
- This is necessary to make SFs sufficiently fast.
- 3rd source of error has received little attention so far:
  - Conventional scoring functions assume a theory-inspired predetermined functional form for the relationship between:
    - the structure-based description of the p-l complex
    - and its measured/predicted binding affinity
  - Problem: difficulty of explicitly modelling the various contributions of intermolecular interactions to binding affinity.
  - Also, SFs use an additive functional form, but this has been specificly shown to be suboptimal (Kinnings et al. 2011 JCIM).

2010

A Machine Learning Approach



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

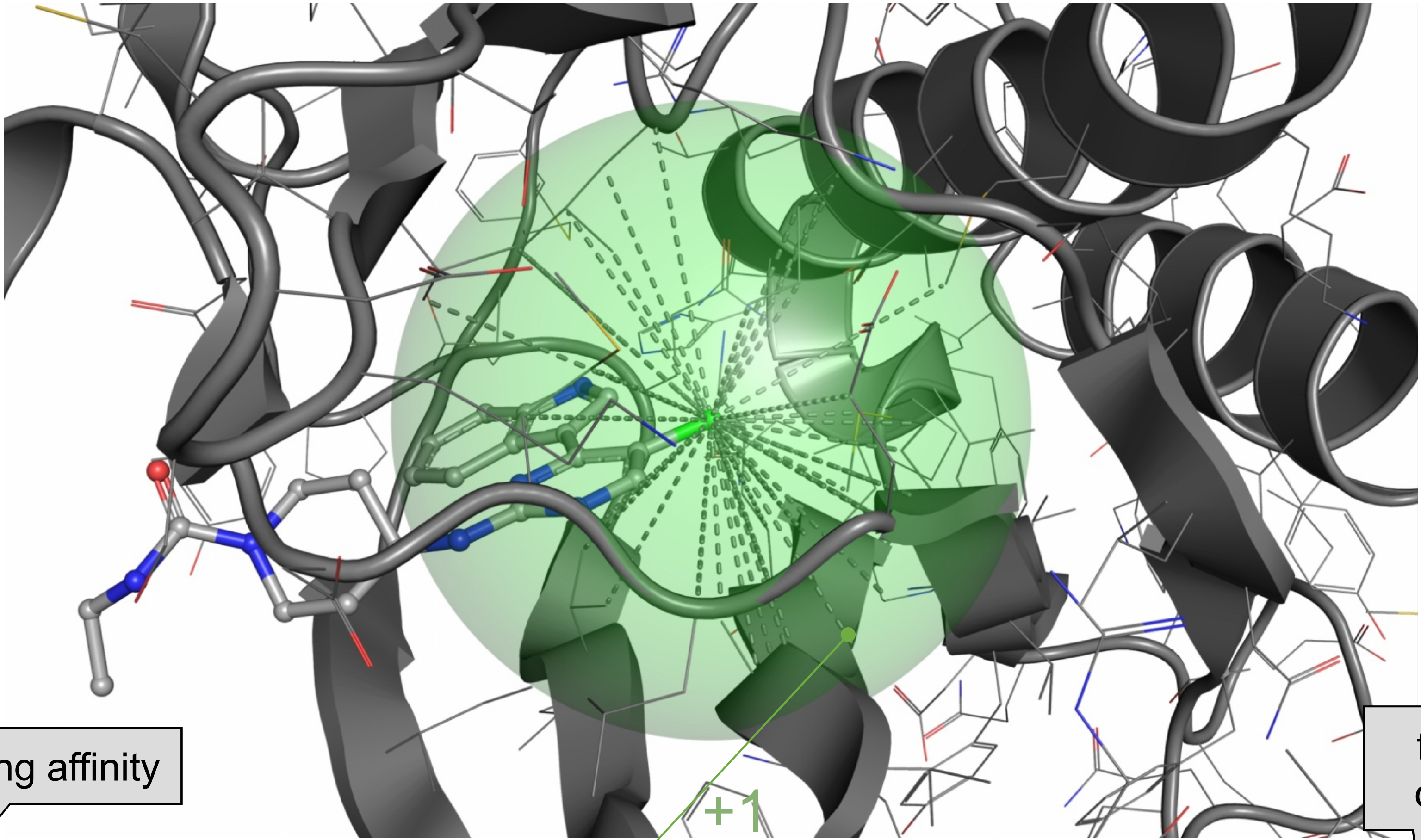
Associate Editor: Burkhard Rost

non-parametric machine learning can be used to implicitly capture the functional form (data-driven, not knowledge-based)

A machine learning approach

- Main idea: a priori assumptions about the functional form introduces modelling error → no asumptions!
- reconstruct the physics of the problem implicitly in an entirely data-driven manner using non-parametric ML.
- Random Forest (Breiman, 2001) to learn how the atomic-level description of the complex relates to pK:
  - Random Forest (RF): a large ensemble of diverse DTs.
  - Decision Tree (DT): recursive partition of descriptor space s.t. training error is minimal within each terminal node.
- But how do we characterise a protein-ligand complex as set of numerical descriptors (features)?

Characterising the protein-ligand complex



binding affinity

features or descriptors

$pK_{d/i}$	C.C	…	C.Cl	…	C.I	N.C	…	I.I	PDB ID
5.70	95		30		0	73		0	2p33

PDBbind benchmark

- *De facto* standard for SFs benchmarking:
  Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. (2009) *JCIM* **49**, 1079-1093

- Refined set → 1300 manually curated protein-ligand complexes with measured binding affinity (↑ diverse):

Training: $D_{\text{train}} = \{(y_j, \vec{x}_j)\}_{j=1}^{1105}$ $\quad y_j = -\log K_j$ $\quad \rightarrow$ $\quad f = f(\vec{x}_j)$

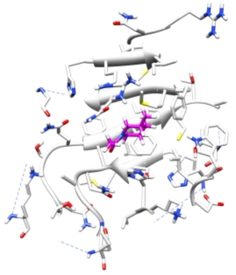
Testing: $D_{\text{test}} = \{(y_j, \vec{x}_j)\}_{j=1106}^{1300}$ $\quad \leftrightarrow$ $\quad \tilde{D}_{\text{test}} = \{(f(\vec{x}_j), \vec{x}_j)\}_{j=1106}^{1300}$

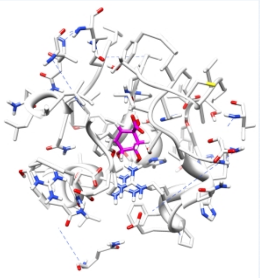
- Benchmark: 16 state-of-the-art SFs → test set error

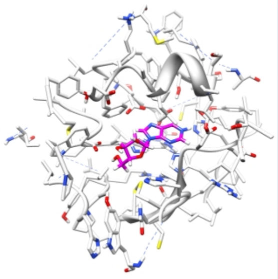
- RF-Score vs 16 SFs on test set error, but:
  - Other SFs have an undisclosed number of cmpxes in common!
  - RF-Score & X-Score (best) non-overlapping training-test sets.

Training and testing machine learning SFs

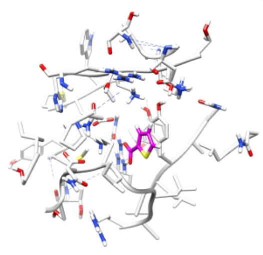
Training set (1105 complexes)

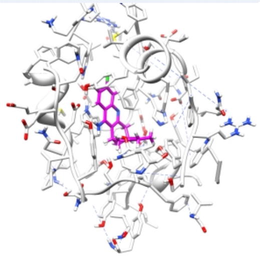

1w8l
pKi=0.49

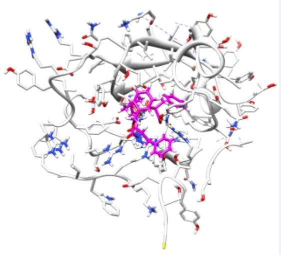

1gu1
pKi=4.52


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

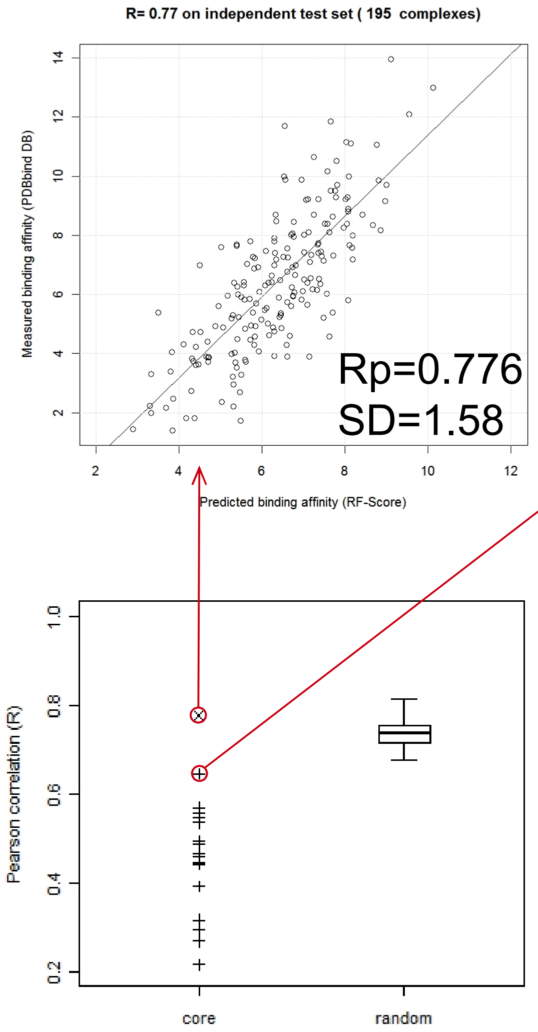
195

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

Random Forest training
(descriptor selection, model selection)

RF-Score
(description and training choices)

RF-Score‘s performance



Rp=0.776
SD=1.58

COMPARATIVE ASSESSMENT OF SCORING FUNCTIONS *J. Chem. Inf. Model., Vol. 49, No. 4, 2009* **1087**

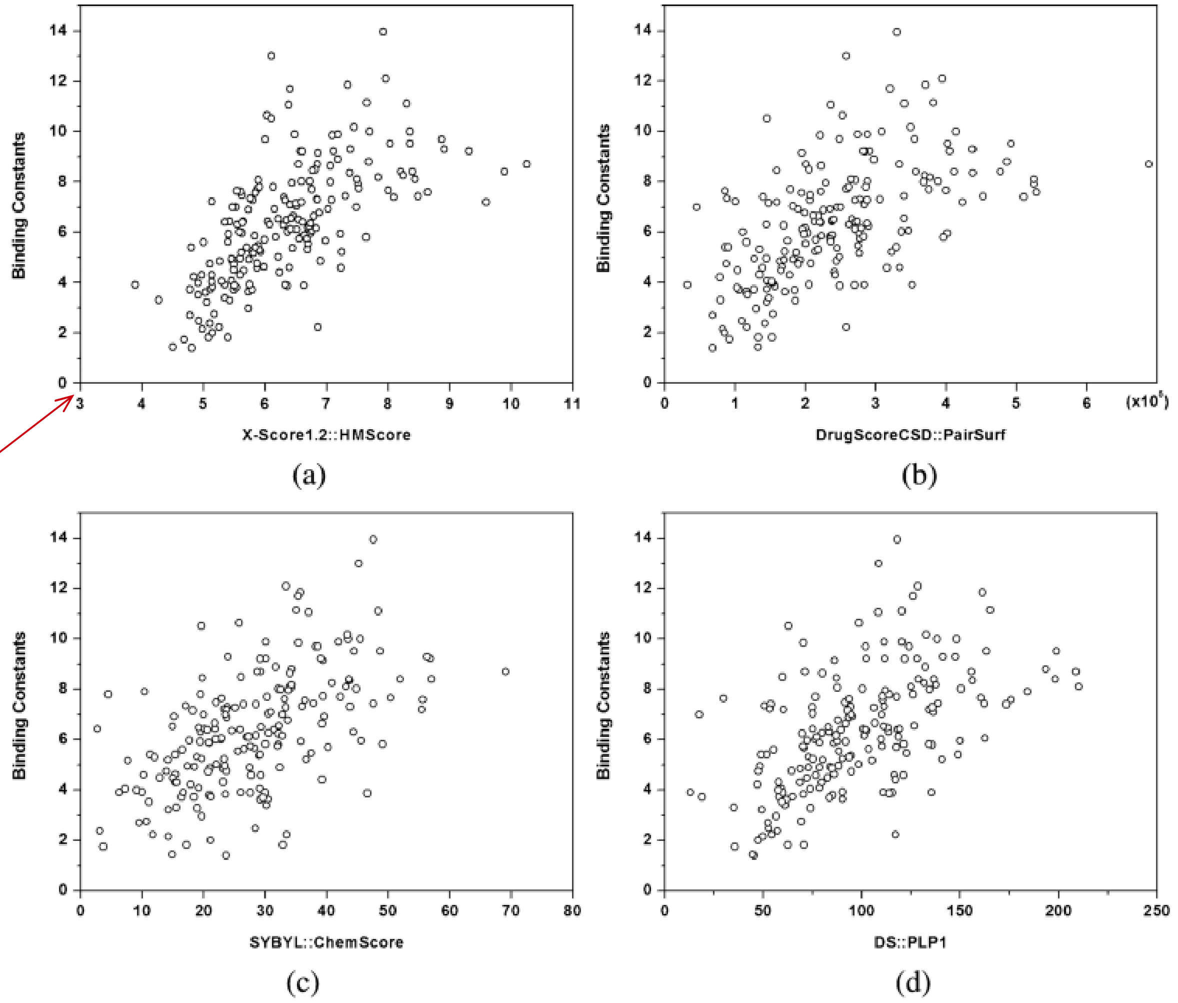
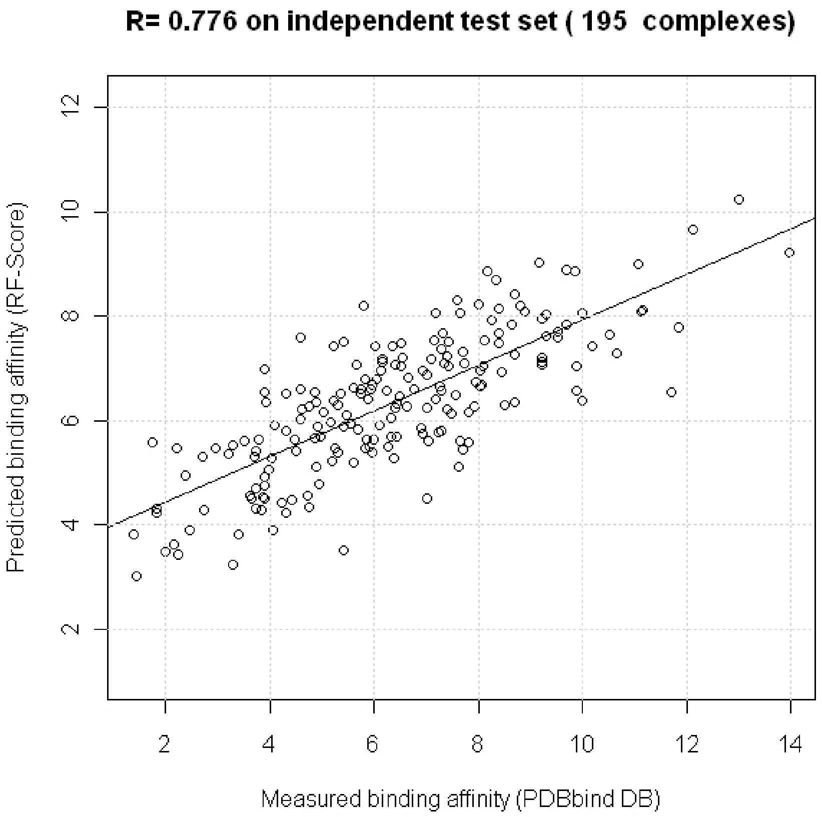
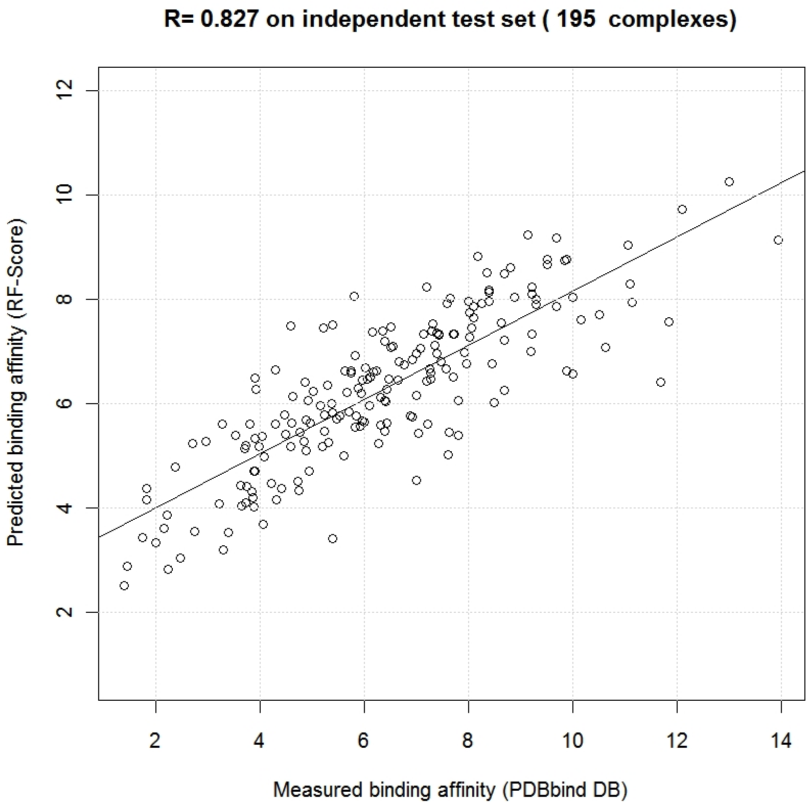


Figure 6. Correlations between the experimentally measured binding constants (in −log K_d units) of the 195 protein−ligand complexes in the primary test set and the binding scores computed by (a) X-Score::HMScore (R = 0.644), (b) DrugScore^{CSD}::PairSurf (R = 0.569), (c) SYBYL::ChemScore (R = 0.555), and (d) DS::PLP1 (R = 0.545).

Careful with biases when comparing SFs!





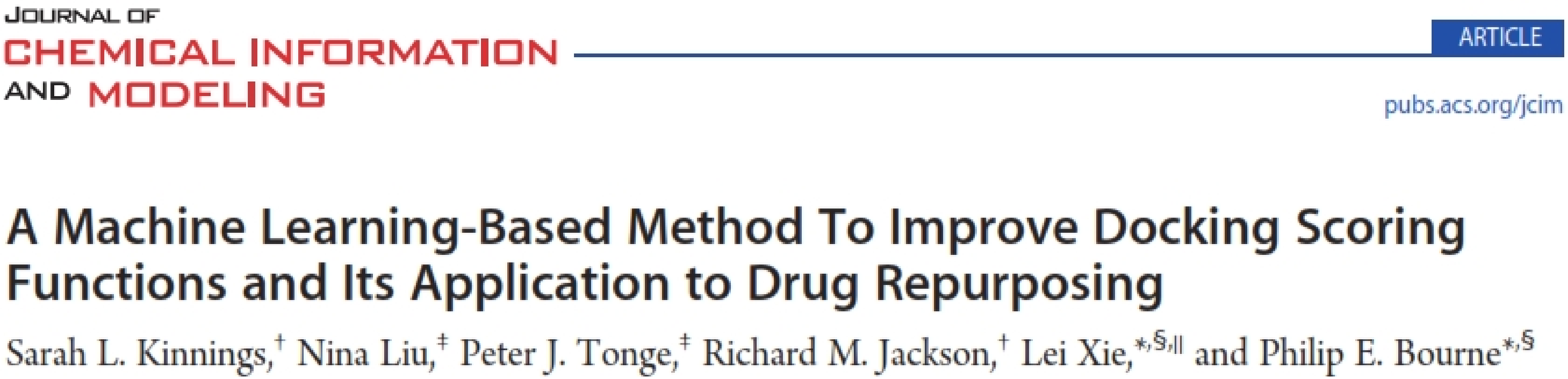
No overlap (unlike other SFs but X-Score) → R_p=0.776

If we allow 65 cpxes overlap → R_p=0.827

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

2011



JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

A Machine Learning-Based Method To Improve Docking Scoring Functions and Its Application to Drug Repurposing

Sarah L. Kinnings,[†] Nina Liu,[‡] Peter J. Tonge,[‡] Richard M. Jackson,[†] Lei Xie,[*,§,∥] and Philip E. Bourne[*,§]

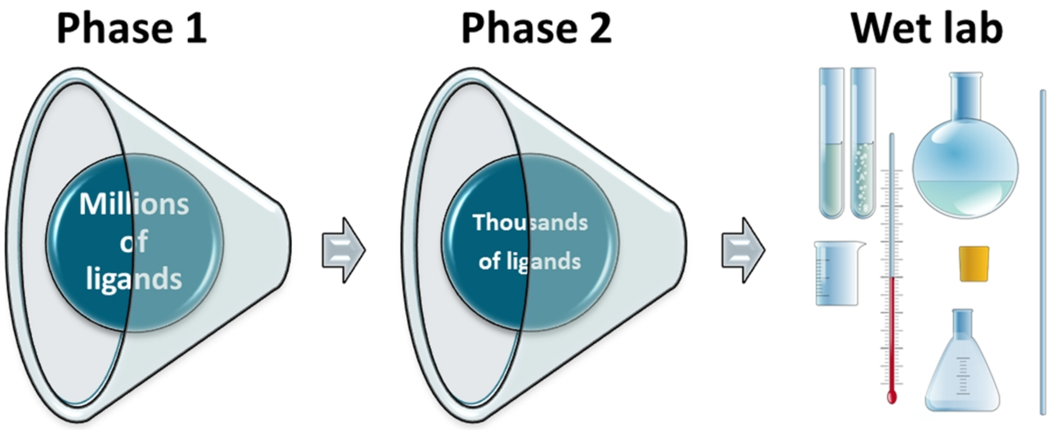
- In predicting $pK_{d/i}$, nonlinear combination of energy terms performs better than the linear regression of energy terms
- Target-specific SF by only considering complexes of anti-TB enzyme InhA (SVR on 80 structures with IC_{50} values)
- SVM classifier better than SVR at retrospective Virtual Screening, partly because negative data in training set.

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

2013

idock



http://istar.cse.cuhk.edu.hk/idock/

- RF-Score is now integrated in istar, a web platform for large-scale online protein-ligand docking

- Multi-threaded Idock on >12M commercially-available compounds → docking poses re-scored with RF-Score.

- Together with Hongjian Li, Kwong-Sak Leung, Man-Hon Wong (Chinese University of Hong Kong)

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

2007

WILEY InterScience® DISCOVER SOMETHING GREAT

PROTEINS STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

A general approach for developing system-specific functions to score protein–ligand docked complexes using support vector inductive logic programming

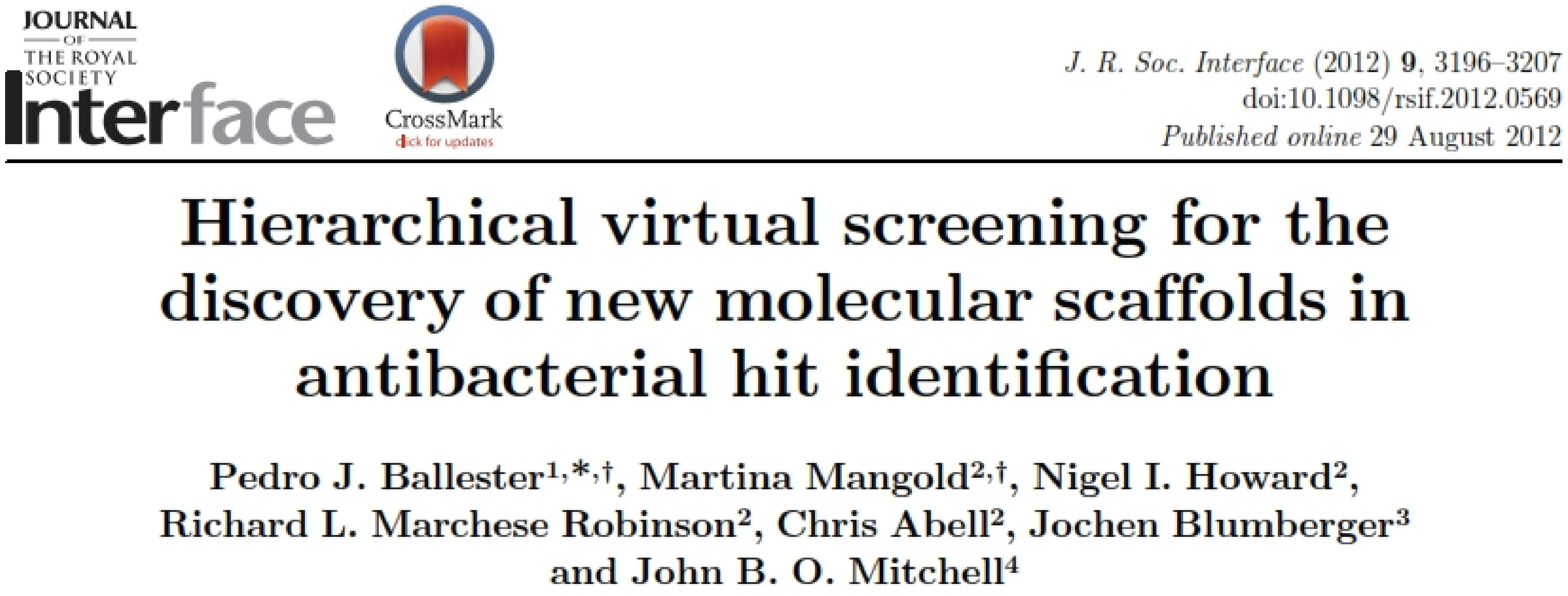
Ata Amini,[1] Paul J. Shrimpton,[1] Stephen H. Muggleton,[2] and Michael J. E. Sternberg[1]*

- One of the two previous non-parametric ML to build SFs. ≠ from RF-Score: target-specific & modelling assumptions
- Very useful for lead optimisation: Support Vector Inductive Logic Programming (SVILP) predicts binding + rules
- Which protein-ligand interatomic features are associated to potent binding? e.g. O.2_C.2, N.am, 51, 2.8, 0.5

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

2012


JOURNAL OF THE ROYAL SOCIETY Interface

CrossMark click for updates

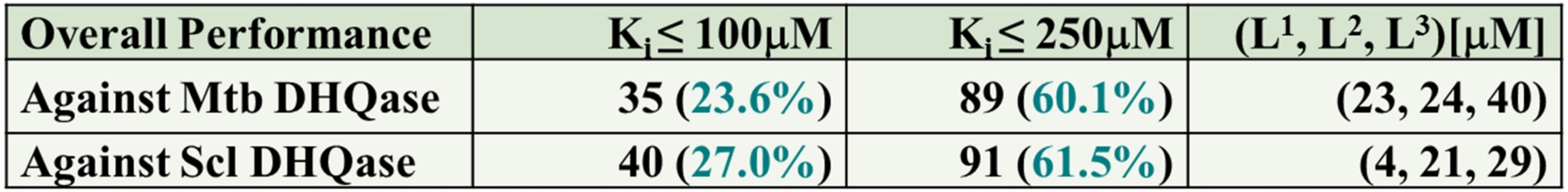
J. R. Soc. Interface (2012) **9**, 3196–3207
doi:10.1098/rsif.2012.0569
Published online 29 August 2012

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

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

- First prospective VS application of RF-Score to two antibacterial targets. Hierarchical, screening 9M cpds.

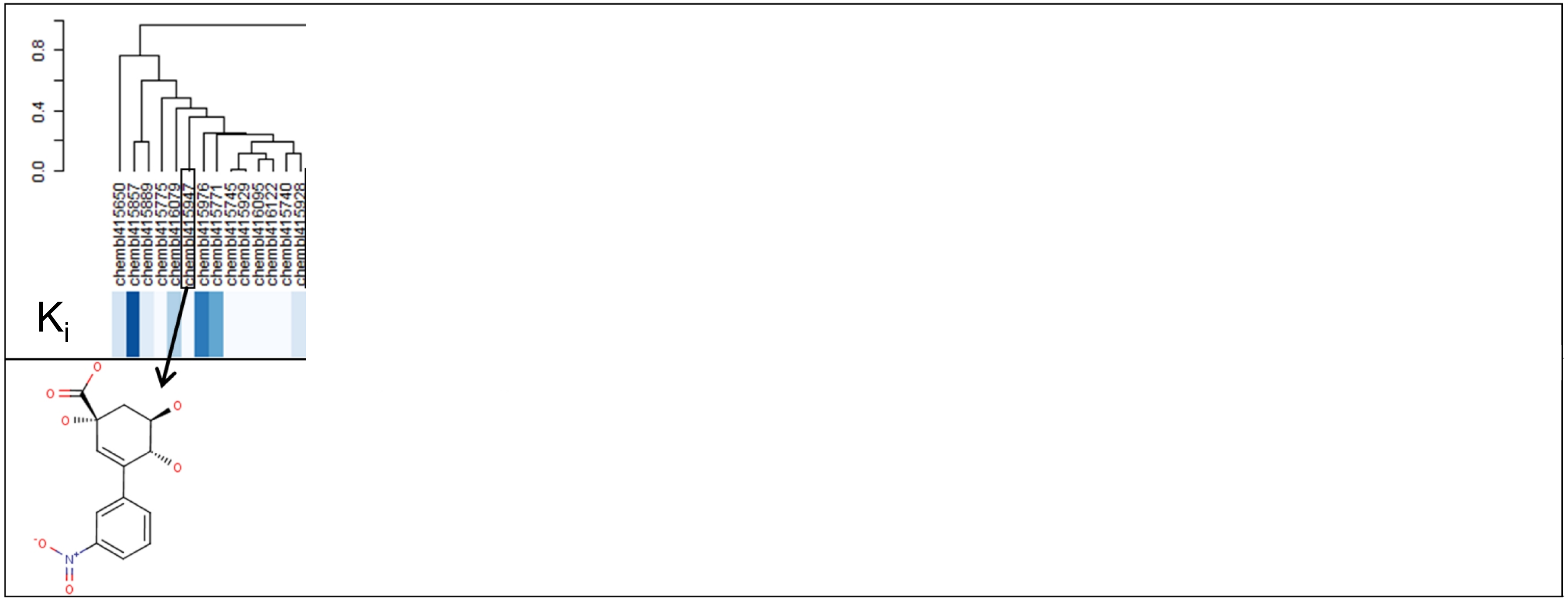
- Outstanding hit rates of ~ 60% with Ki ≤ 250 μM → 100 new and structurally diverse actives (£5,000 cost).



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

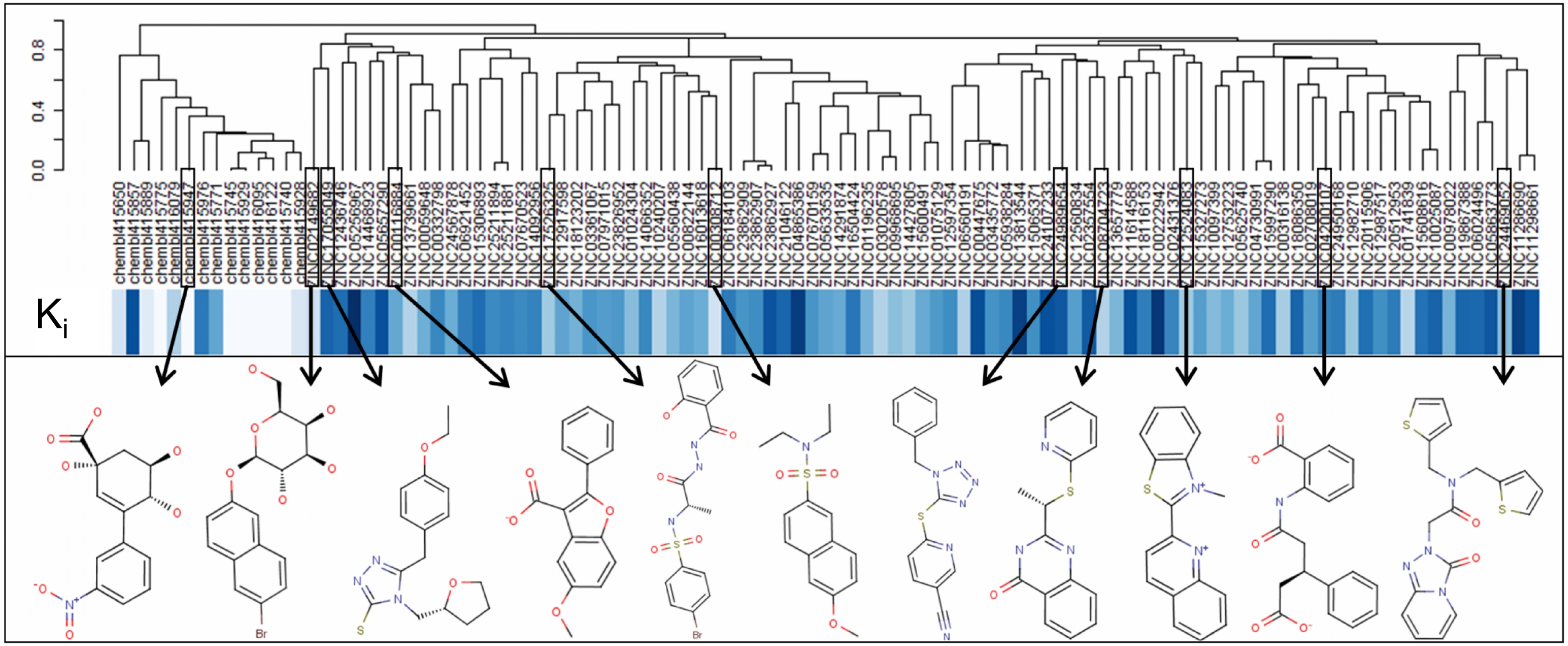
One known scaffolds for Type II DHQase

M. Tuberculosis



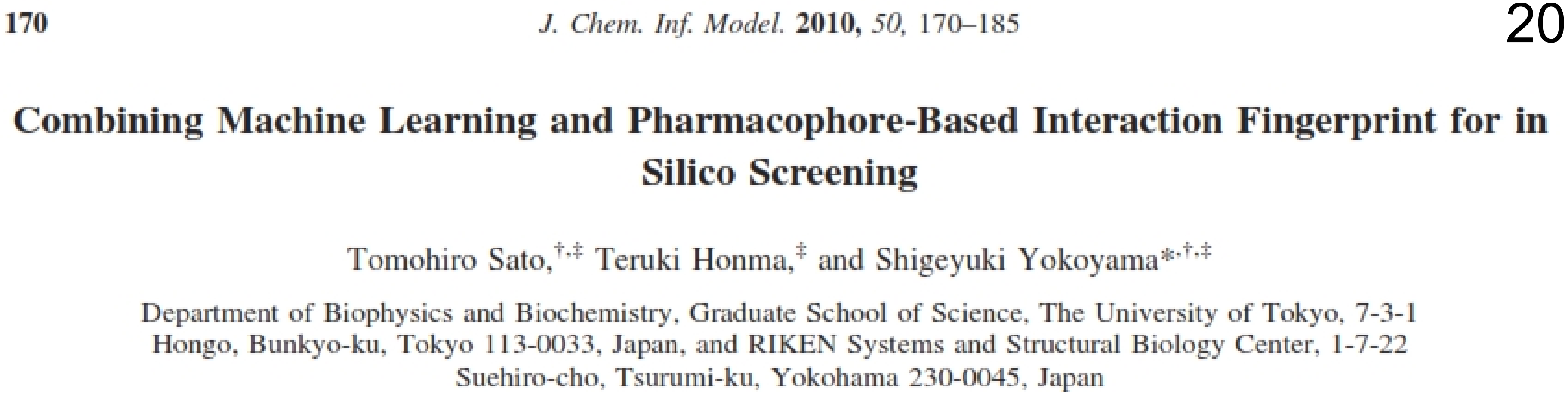
New active scaffolds for Type II DHQase

M. Tuberculosis



Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects



170 *J. Chem. Inf. Model.* **2010,** *50,* 170–185

Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening

Tomohiro Sato,[†,‡] Teruki Honma,[‡] and Shigeyuki Yokoyama*[,†,‡]

Department of Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, 7-3-1
Hongo, Bunkyo-ku, Tokyo 113-0033, Japan, and RIKEN Systems and Structural Biology Center, 1-7-22
Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

2010

- Not a MLSF predicting binding affinity, ML classifier to discriminate between actives and inactives of a target.
- Interesting: uses docking poses of active and inactives to supplement ligand-bound crystal structures of the target.
- SVM, RF and NNs. Five target-specific classifiers. Implementations generally outperform GlideScore::SP

Talk outline

1. Motivation
2. Predicting $K_{d/i}$ of diverse protein-ligand structures
3. Ranking protein-ligand structures of a target
4. Ranking protein-ligand docking poses of a target
5. Analysing binding: feature importance and selection
6. Virtual Screening based on ML regression
7. Virtual Screening based on ML classifiers
8. Future prospects

Future prospects – reviews highlighting MLSFs

- 2010 Xiaoqin Zou & co-workers (U. of Missouri, USA):
  - MLSFs shown to be able to exploit very large training sets
- 2012 Stephen Bryant & co-workers (NCBI, USA):
  - RF-Score strikingly outperforms all 16 state-of-the-art traditional SFs.
  - MLSFs avoid explicit error-prone modelling of solvation & entropy.
- 2012 Christoph Sotriffer (U. of Würzburg, Germany):
  - MLSFs are becoming increasingly popular.
- 2012 Russ Altman & co-workers (Stanford U., USA):
  - MLSFs improve rank-ordering of series of related molecules.
  - As structural dbs grow, MLSFs are expected to further improve.
- 2013 Chung-Hang Leung & co-workers (U. of Macau, China):
  - MLSFs are attracting increasing attention in estimation of binding affinity