Machine learning approaches to predicting protein-ligand binding

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

- Developing new drug = average US$4 billion and 15 years
  [Link](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 triggering 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.\(^\text{Payne et al. 2007}\)
  - 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 sub-pocket 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)

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

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

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

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

\[ PMF = \sum_{\text{prot-lig}} A_{ij}(d_{ij}) \ln \left[ \frac{f_{\text{vol-corr}}^j(r) \rho_{\text{seg}}^{ij}(r)}{\rho_{\text{bulk}}^{ij}(r)} \right] \]

- SFs are trained on pK data usually through MLR:
  - FF \((A_{ij}, B_{ij})\), Emp\((w_0, \ldots, w_4)\) and sometimes KB \((\rho_{\text{ref state}}^{ij})\)
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 specifically shown to be suboptimal (Kinnings et al. 2011 JCIM).
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 assumptions!

• 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

Machine learning approaches to predicting protein-ligand binding

<table>
<thead>
<tr>
<th>pK\textsubscript{d/i}</th>
<th>C.C</th>
<th>...</th>
<th>C.Cl</th>
<th>...</th>
<th>C.I</th>
<th>N.C</th>
<th>...</th>
<th>I.I</th>
<th>PDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.70</td>
<td>95</td>
<td></td>
<td>30</td>
<td></td>
<td>0</td>
<td>73</td>
<td></td>
<td>0</td>
<td>2p33</td>
</tr>
</tbody>
</table>

binding affinity

features or descriptors
PDBbind benchmark

- *De facto* standard for SFs benchmarking:

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

\[
\begin{align*}
\text{Training: } & \quad D_{\text{train}} = \{(y_j, \bar{x}_j)\}_{j=1}^{1105}, \quad y_j = -\log K_j \\
\text{Testing: } & \quad D_{\text{test}} = \{(y_j, \bar{x}_j)\}_{j=1106}^{1300}, \quad \tilde{D}_{\text{test}} = \{f(\bar{x}_j), \bar{x}_j\}_{j=1106}^{1300}
\end{align*}
\]

- Benchmark: 16 state-of-the-art SFs \( \rightarrow \) 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)**

<table>
<thead>
<tr>
<th>PDB</th>
<th>pKᵢ</th>
<th>C.C</th>
<th>C.I</th>
<th>N.C</th>
<th>I.I</th>
<th>PDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1w8l</td>
<td>0.49</td>
<td>1254</td>
<td>0</td>
<td>166</td>
<td>0</td>
<td>1w8l</td>
</tr>
<tr>
<td>1gu1</td>
<td>4.52</td>
<td>1254</td>
<td>0</td>
<td>166</td>
<td>0</td>
<td>1gu1</td>
</tr>
<tr>
<td>2ada</td>
<td>13.00</td>
<td>2324</td>
<td>0</td>
<td>919</td>
<td>0</td>
<td>2ada</td>
</tr>
</tbody>
</table>

**Test set (195 complexes)**

<table>
<thead>
<tr>
<th>PDB</th>
<th>pKᵢ</th>
<th>C.C</th>
<th>C.I</th>
<th>N.C</th>
<th>I.I</th>
<th>PDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>2hdq</td>
<td>1.40</td>
<td>858</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2hdq</td>
</tr>
<tr>
<td>1e66</td>
<td>9.89</td>
<td>4476</td>
<td>0</td>
<td>283</td>
<td>0</td>
<td>1e66</td>
</tr>
<tr>
<td>7cpa</td>
<td>13.96</td>
<td>4476</td>
<td>0</td>
<td>283</td>
<td>0</td>
<td>7cpa</td>
</tr>
</tbody>
</table>

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

Random Forest training (descriptor selection, model selection)

RF-Score (description and training choices)
RF-Score’s performance

Comparative Assessment of Scoring Functions


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\(\text{GSD}\):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
• In predicting pK\textsubscript{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\textsubscript{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
• 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)

http://istar.cse.cuhk.edu.hk/idock/
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
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. Sternberg¹*

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

Pedro J. Ballester¹,*;†, Martina Mangold²;†, Nigel I. Howard², Richard L. Marchese Robinson², Chris Abell², Jochen Blumberger³ and John B. O. Mitchell⁴

• 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).

<table>
<thead>
<tr>
<th>Overall Performance</th>
<th>K_i ≤ 100µM</th>
<th>K_i ≤ 250µM</th>
<th>(L¹, L², L³)[µM]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Against Mtb DHQase</td>
<td>35 (23.6%)</td>
<td>89 (60.1%)</td>
<td>(23, 24, 40)</td>
</tr>
<tr>
<td>Against Scl DHQase</td>
<td>40 (27.0%)</td>
<td>91 (61.5%)</td>
<td>(4, 21, 29)</td>
</tr>
</tbody>
</table>
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
• 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