Deep Learning Scoring Function for Flexible Molecular Docking

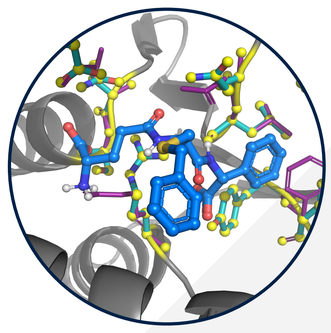
Rocco Meli,[1] Jocelyn Sunseri,[2] Philip C. Biggin,[1] David R. Koes[2]

[1] Department of Biochemistry, University of Oxford
[2] Department of Computational and Systems Biology, University of Pittsburgh

Introduction

Problems with **standard docking** studies:

- Semi-empirical scoring functions (SFs)
- Knowledge-based scoring functions (SFs)
- **Rigid receptor**



Goals of this Google Summer of Code (GSoC) Project:

- **Train a CNN SF on docking with flexible side chains**
- **Implement CNN optimisation of flexible side chains**

Implemented in `gnina`*, a deep learning framework for molecular docking*

Flexible Docking

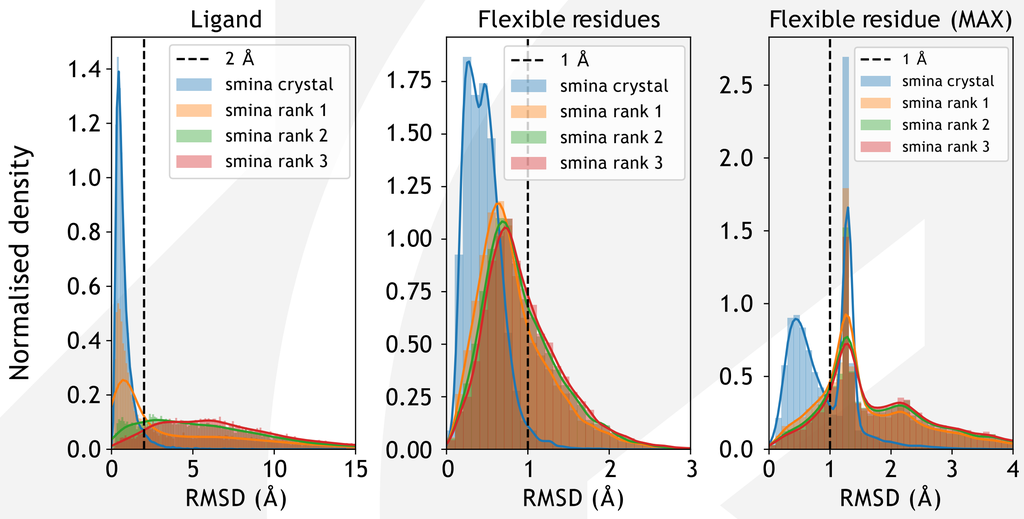
Flexible docking principles (with `smina`):

- Ligand rotations and translations
- Ligand rotatable bonds
- **Rotatable bonds of protein side chains**
- Fixed (rigid) backbone

PDBbind 2018 [4]:

- 16151 P-L complexes
- PDB and MOL2 files
- Binding affinity: K_i, K_d, IC_{50}

Re-docking: 203786 (+15840) protein-ligand poses



Ligand and flexible side chains poses can be **annotated based on RMSD**

Good	Bad
• Ligand RMSD < 2 Å • Flex RMSD (MAX) < 1 Å	• Ligand RMSD > 4 Å • Flex RMSD (MAX) > 1.5 Å

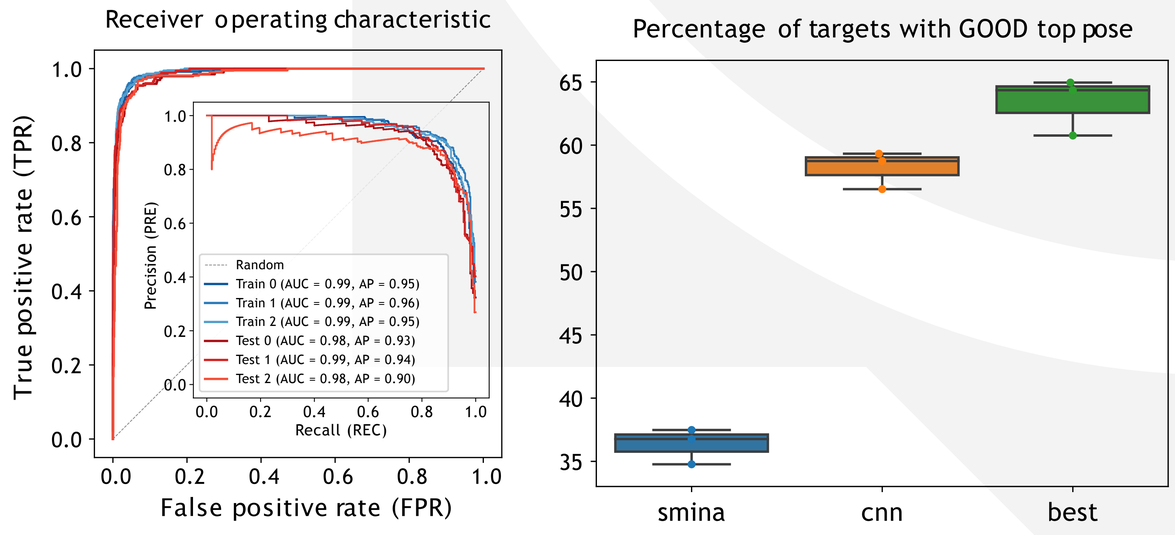
Training and Validation

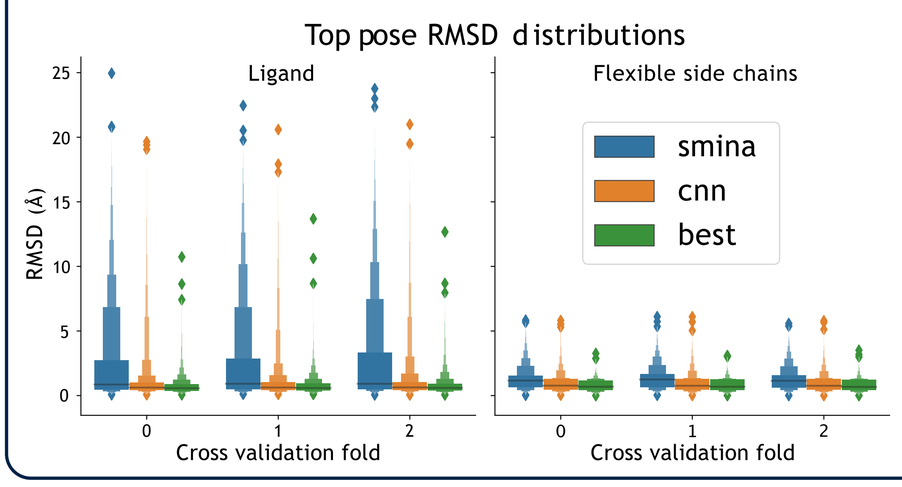
RMSD-based annotation:

- **3232 (+ 8284) positive** examples
- **75404 (+ 25) negative** examples
- **Class imbalance** problem

3-fold **cross-validation**:

- Protein sequence distance
- Ligand similarity
- Targets per fold: [5598, 5639, 4889]





$$\mathrm{TPR} = \frac{t_\mathrm{P}}{t_\mathrm{P} + f_\mathrm{N}}$$

$$\mathrm{FPR} = \frac{f_\mathrm{P}}{f_\mathrm{P} + t_\mathrm{N}}$$

$$\mathrm{PRE} = \frac{t_\mathrm{P}}{t_\mathrm{P} + f_\mathrm{P}}$$

$$\mathrm{REC} = \frac{t_\mathrm{P}}{t_\mathrm{P} + f_\mathrm{N}}$$

gnina

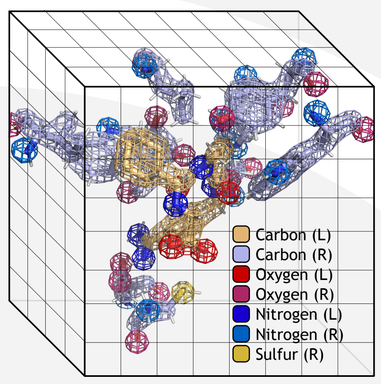
`gnina` [1]:

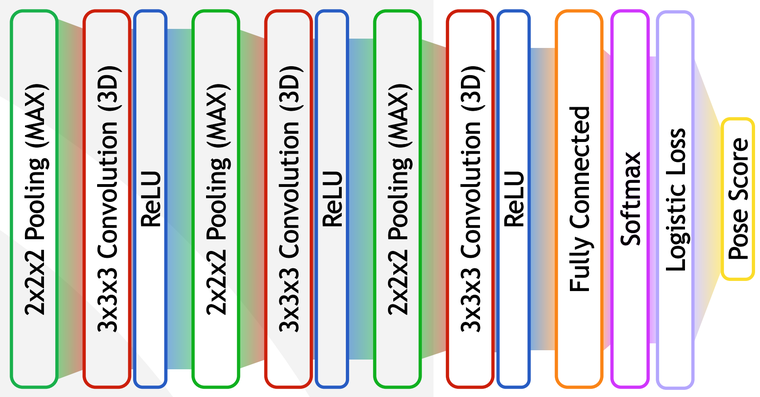
- **Deep learning framework for molecular docking**
- `caffe` [2] + `smina` [3] + `libmolgrid`
- Developed in David Koes group (Pittsburgh)

`github.com/gnina`

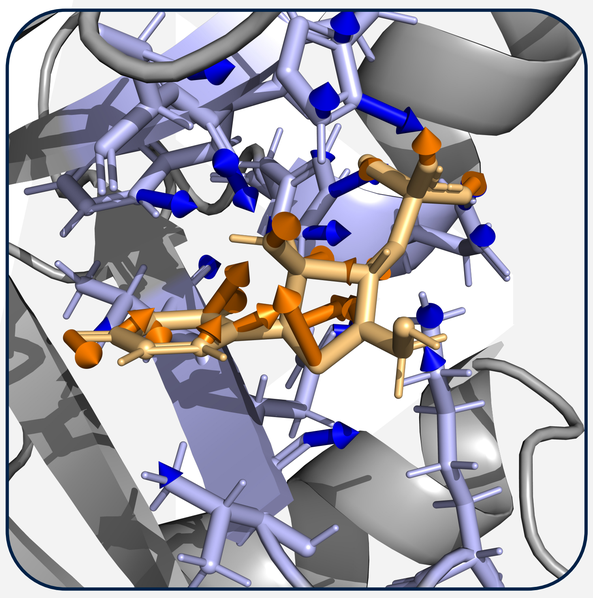
Workflow:

- **Discretisation** (gridding) of protein-ligand binding site (3D)
- Computation of **atomic densities** for different `smina` atom types
- **Standard CNN** machinery for computer vision (with data augmentation)





CNN Atomic Gradients



Backpropagation of gradients to ligand and receptor atoms [5]:

$$\frac{\partial \mathcal{L}}{\partial \vec{a}} = \sum_{g \in G_{\vec{a}}} \frac{\partial \mathcal{L}}{\partial g} \frac{\partial g}{\partial d} \frac{\partial d}{\partial \vec{a}}$$

$$g(d;R) = \begin{cases} e^{-\frac{2d^2}{R^2}} & 0 \leq d < R \\ \frac{4}{e^2 R^2} d^2 - \frac{12}{e^2 R} d + \frac{9}{e^2} & R \leq d < 1.5R \\ 0 & d \geq 1.5R \end{cases}$$

Ligand
Ligand Gradients
Flexible Side Chains
Flexible Side Chains Gradients

Ligand pose and side chains optimisation, with respect to the CNN loss function, using standard optimisation techniques (BFGS).

CNN Optimisation

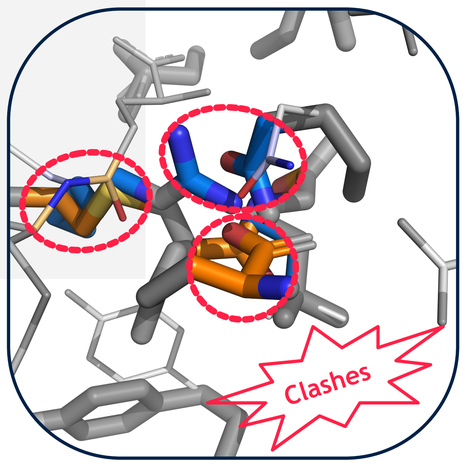
The CNN is trained on `smina` poses, which are **physically sensible poses**.

Problems:

- Steric clashes (L-L, R-L, R-R)
- RMSD distributions (L and R) worsen

CNN-optimised (L)
CNN-optimised (R)
`smina` (L)
`smina` (R)

Including CNN-optimised poses in the training set **broadens the conformational space** with unrealistic poses and should **improve the CNN performance** [5].



Conclusions

- `smina` SF is **not parametrised** well to score flexible side chains
- CNN SF is learned and therefore **performs better** and is **more robust**
- CNN SF trained on docked poses does not learn **steric interactions**

Next steps:

- **Enrich training dataset** with CNN-optimised poses
- **Re-train** CNN SF on the enriched dataset to **improve performance**

References

[1] M. Ragoza *et al.*, J. Chem. Inf. Model. **57**, 942-957 (2017)
[2] Y. Jia *et al.*, arXiv, arXiv:1408.5093 (2014)
[3] D. R. Koes *et al.*, J. Chem. Inf. Model. **53**, 1893-1904 (2013)
[4] Z. Liu *et al.*, Acc. Chem. Res. **50**, 302-309 (2017)
[5] M. Ragoza *et al.*, arXiv, arXiv:1710.07400 (2017)

evotec
BBSRC bioscience for the future
Open Chemistry
OXFORD Interdisciplinary Bioscience DTP
Department of Biochemistry

@RMeli
@RoccoMeli