DeepSite: protein-binding site predictor using 3D-convolutional neural networks FREE

J Jiménez, S Doerr, G Martínez-Rosell, A S Rose, G De Fabritiis ✉

Bioinformatics, Volume 33, Issue 19, 01 October 2017, Pages 3036–3042,
https://doi.org/10.1093/bioinformatics/btx350
Published: 31 May 2017 **Article history** ▾

https://academic.oup.com/bioinformatics/article/33/19/3036/3859178

Survey by Eric Wang

2019 Spring @ https://qdata.github.io/deep2Read/

- Introduction
- Materials
- Methods
- Experiments
- Future work

Introduction

- One of the first steps in the structure-based drug design (SBDD) pipeline is identifying viable **druggable binding sites** on the target protein. This task is defined as identifying and delimiting **protein cavities**, potentially at the surface that are likely to bind to a small compound.
- Four different and complementary approaches: geometric, evolutionary, energetic or statistical
- These approaches typically exploit known binding site properties.
- In this work, we propose a machine learning algorithm based on DCNNs for predicting **ligand-binding sites in proteins**, and show that given enough training data, they are able to capture binding site characteristics and can outperform other two competitive algorithms by providing an extensive test set based on more than 7000 proteins of the **scPDB** database.

Introduction - Non-NN approaches

1. fpocket (Le Guilloux *et al.*, 2009) uses the concept of alpha-spheres (spheres that contact four atoms but contain none) introduced by the MOE Site Finder to find cavities.
2. POCKET (Levitt and Banaszak, 1992) and LigSite (Hendlich *et al.*, 1997) search for protein-solvent-protein events on a determined enclosing of the protein.
3. Pocket-Picker (Weisel *et al.*, 2007) uses a uniform grid of points which get assigned a buriedness index.
4. PASS (Brady and Stouten, 2000) computes a coating of probe spheres with protein as substrate, with additional layers of probes accreted onto previously found probe spheres, keeping in the end low solvent exposure spheres at the point when an accretion layer no longer produces new probe spheres.
5. Concavity (Capra *et al.*, 2009) additionally makes use of evolutionary sequence conservation information in combination with other structure-based methods.

- Introduction
- Materials
- Methods
- Experiments
- Future work

Materials

- scPDB v.2013 database (Desaphy *et al.*, 2015)
- Up to date selection of high-quality, non-redundant druggable binding sites extracted from the Protein Data Bank (PDB), focusing mostly on small synthetic or natural ligands.
- It contains information for the pocket, the corresponding ligand and its binding mode.
- Out of a total of 9190 structures found in the file, 12 were discarded as they contained multiple erroneous entries. Furthermore, the scPDB database also provides a clustering of binding sites for all PDB structures by Uniprot entry. To exclude identical binding sites in the training and test sets for an unbiased evaluation of the performance of our method, 1556 structures without clustering information were removed resulting in a **final set of 7622 structures** for analysis.

Materials

- *k* = 10-fold cross-validation
- The **separation method** ensures that the same protein pocket (possibly existing in different PDB structures) does not occur on both the training and test set in a split, therefore limiting the possibility of over-fitting.
- In particular, we checked whether training and test sets were dissimilar enough in each of the splits by using directly the Shaper similarity metric matrix (Desaphy *et al.*, 2012) provided by the scPDB database which reports a **structural distance** between binding pockets.
- No identical pair of binding sites was found on both training and test set in any split.

- Introduction
- Materials
- Methods
- Experiments
- Future work

Methods - Descriptor Computation and labeling

- Treat protein structures as 3D images. Coordinates of this 3D image are defined to span the bounding box of the protein plus a buffer of 8 Å to account for pockets located close to its edges.
- The 3D image is then discretized into a grid of 1 × 1 × 1 Å^3 sized voxels.
- Voxel occupancies are defined with respect to the atoms in the protein depending on their **excluded volume** and other seven atom properties: hydrophobic, aromatic, hydrogen bond acceptor or donor, positive or negative ionizable and metallic. These are called ***channels***, to draw a comparison to computer vision, where an image can be represented with three different color arrays: red, green and blue.
- The AutoDock 4 (Morris *et al.*, 2009) atom types found in Table 1 were used with the rules of Table 2 to assign each atom to a specific channel.
- Atom occupancies were calculated by taking the simplest approximation for the pair correlation function defined by

$$n(r) = 1 - \exp\left(-(r_{\mathrm{vdw}}/r)^{12}\right).$$

Table 1.

AutoDock 4 atom types

Element	Description
C	Non H-bonding aliphatic carbon
A	Non H-bonding aromatic carbon
NA	Acceptor 1 H-bond nitrogen
NS	Acceptor S Spherical nitrogen
OA	Acceptor 2 H-bonds oxygen
OS	Acceptor S Spherical oxygen
SA	Acceptor 2 H-bonds sulfur
HD	Donor 1 H-bond hydrogen
HS	Donor S Spherical hydrogen
MG	Non H-bonding magnesium
ZN	Non H-bonding zinc
MN	Non H-bonding manganese
CA	Non H-bonding calcium
FE	Non H-bonding iron

Table 2.

Property-atom type (AutoDock 4) correspondence used for Deepsite's 3D descriptor computation

Property	Rule
Hydrophobic	atom type C or A
Aromatic	atom type A
Hydrogen bond acceptor	atom type NA or NS or OA or OS or SA
Hydrogen bond donor	atom type HD or HS with O or N partner
Positive ionizable	atom with positive charge
Negative ionizable	atom with negative charge
Metal	atom type MG or ZN or MN or CA or FE
Excluded volume	all atom types

1: **function** Occupancy(atomCoords, centerCoords, radii, channels)

2: **for** each atom A in protein **do**

3: $\boldsymbol{a} \leftarrow \text{atomCoords}_A$

4: $\boldsymbol{h} \leftarrow \text{channels}_A$

5: $\boldsymbol{r}_{\text{vdw}} \leftarrow \text{radii}_A$

6: **for** each center $\boldsymbol{c}$ in centerCoords **do**

7: $r \leftarrow L_2\text{Dist}(\boldsymbol{c}, \boldsymbol{a})$

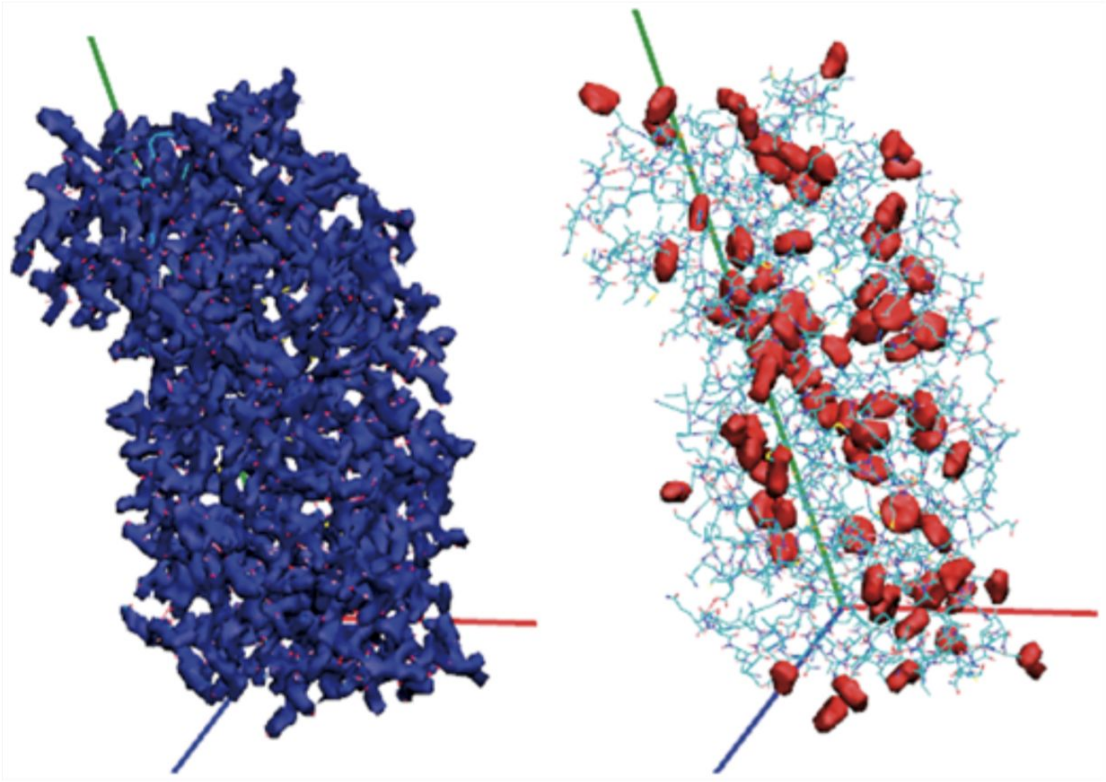
8: $x \leftarrow \frac{r_{\text{vdw}}}{r}$

9: $n \leftarrow 1 - \exp\left(-x^{12}\right)$

10: **for** each channel p in $\boldsymbol{h}$ **do**

11: $O_{\boldsymbol{c},p} \leftarrow \max\{n, O_{\boldsymbol{c},p}\}$

Fig. 1.



View large Download slide

Example of descriptor computation output for the hydrophobic and aromatic channels, respectively for PDB ID 4NIE

Methods - Descriptor Computation and labeling

- Subgrids of 16 × 16 × 16 voxels out of these arrays are then sampled, defining smaller protein areas with local properties.
- Label each of the subgrids as positive, if its geometric center is closer than 4 Å to the pocket geometric center ; negative otherwise.
- We then design a DCNN which uses as input the various features (channels) of the subgrids and outputs the binding site label probability, in the hope of capturing local patterns in the structure of the grid that may help characterize binding pockets.

- Introduction
- Materials
- Methods
- Experiments
- Future work

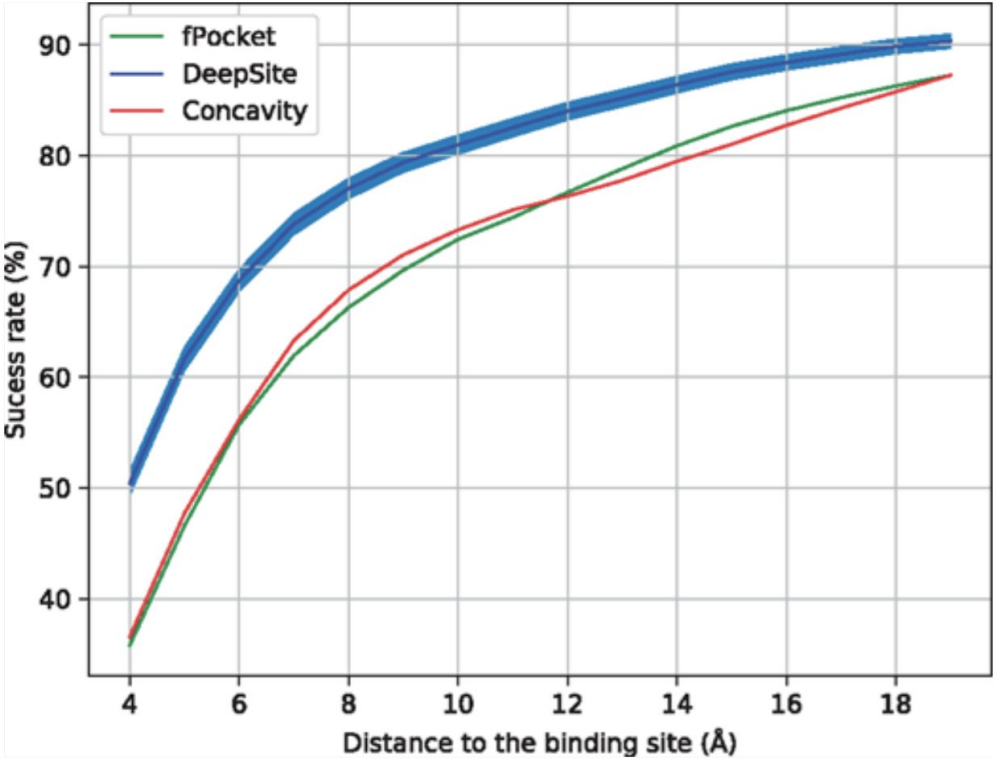
Experiments - Evaluating pocket prediction

- **Distance to the center of the binding site (DCC)**. This metric considers a prediction successful if a point prediction of the pocket is closer than a given distance threshold to the geometric center of the real-binding site. Values ranging between 4 and 20 Å are typically used for success rate plots. This evaluation ignores altogether the shape of the predicted pocket.
- **Discretized volumetric overlap (DVO)**. For this metric, we discretize protein space into 1 × 1 × 1 Å3 voxels, and consider the convex hulls determined by both the real and predicted-binding site volume. We then compute a Jaccard Index, defined by

$$J = \frac{\#|V_r \cap V_p|}{\#|V_r \cup V_p|},$$

where V_r and V_p are the set of 1 × 1 × 1 Å3 voxels that fall inside the convex hull of the real and predicted binding pocket respectively. This measure takes into account both pocket shape and size of the pocket prediction. The average of this measured is considered across all different test splits.

Fig. 3.



View large Download slide

DCC metric averaged over all 10 splits for DeepSite, fPocket and Concavity. Metric shown for distances ranging between 4 and 20 Å

Table 3.

Evaluation of DeepSite volumetric performance using the DVO metric

	Average DVO	SD	*P*-value
DeepSite	**0.652**	0.129	—
fPocket	0.619	0.169	0
Concavity	0.489	0.172	0

Note: DVO is defined as a discretized Jaccard Index of the predicted and real pocket volumes. Mann–Whitney's U *P*-value shown for the unilateral test using DeepSite as baseline ($n = 7622$).