Predicting protein-ligand binding affinity with machine learning

Emanuel Ieremita

Proteins and drugs

- Drugs are useful: they allow us to fix things that go wrong in our body (and more), by binding to proteins and influencing their function
- But making drugs is hard, expensive, and time consuming
- Traditionally laboratory methods were mainly used, but recently more computational approaches are becoming popular

Fig. 1. *Excitatory synapse, D.S. Goodsell, 2018*

Binding affinity

- A successful drug is able to bind reasonably well to its target protein (among other things), so we need ways to assess this binding affinity as a first step in developing a drug
- Normally it can be done by experimentally measuring the affinity, or through molecular docking simulations, which compute all the physical and chemical interactions between the protein's components (amino acids) and the drug
- There are also predictive models that can be trained on available experimental data that could estimate the binding affinity much quicker (and potentially reasonably accurately too)
- This project is an attempt at training such a model

Predictive models

- There are different predictive models out there, but they all use different approaches (CNNs, regression models), datasets (oftentimes small), and features
- In this project I made an attempt of my own, trying to make use of appropriate data, think of the most descriptive features that I could extract from it, and identify a machine learning model that performs the best

Research

- The first phase in the project was the research phase, where I learned as much as I could about protein-ligand interactions
- After familiarising myself with the field, I found data sources, and I had to see what features this data could provide, to be able to perform feature engineering to the level that will best encapsulate the binding process

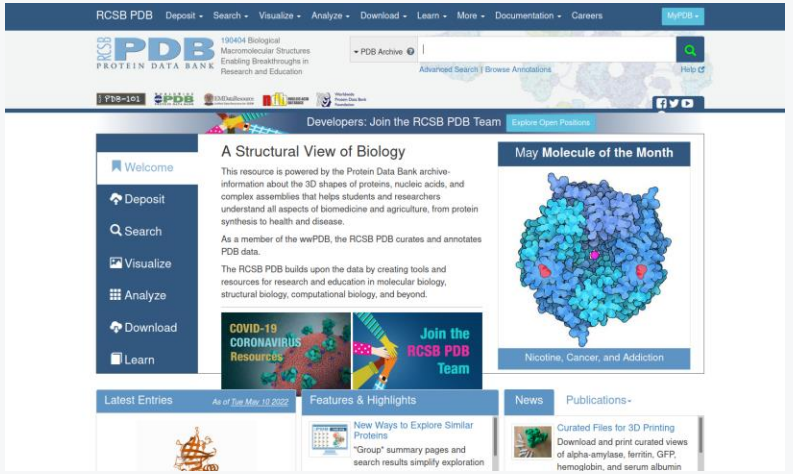
Fig. 2. *Inhibitory synapse, D.S. Goodsell, 2018*

Data sources

- The **Protein Data Bank** (PDB) (https://www.rcsb.org/) is an archive of curated protein structures deposited by researchers (since 1971!)
- A variety of data is available, including structural files, sequence data, ligand structural files and more

- **PDBbind** (http://pdbbind.org.cn/) is a comprehensive database of curated, experimentally measured, binding affinity data of protein-ligand complexes from the PDB.
- This is the main dataset that was used for the labels and training
- The latest version contains more than 19,000 protein-ligand complexes
- **DSSP (**https://swift.cmbi.umcn.nl/gv/dssp/) is a database of secondary structure assignments for entries in the PDB

Fig. 3. *Protein Data Bank homepage*



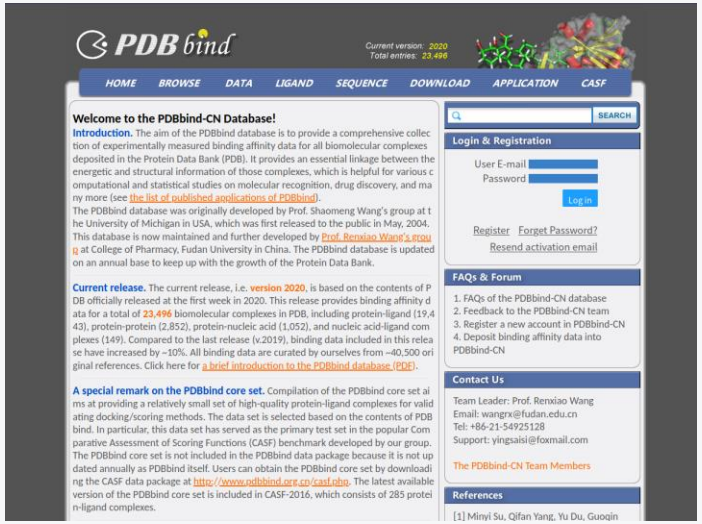
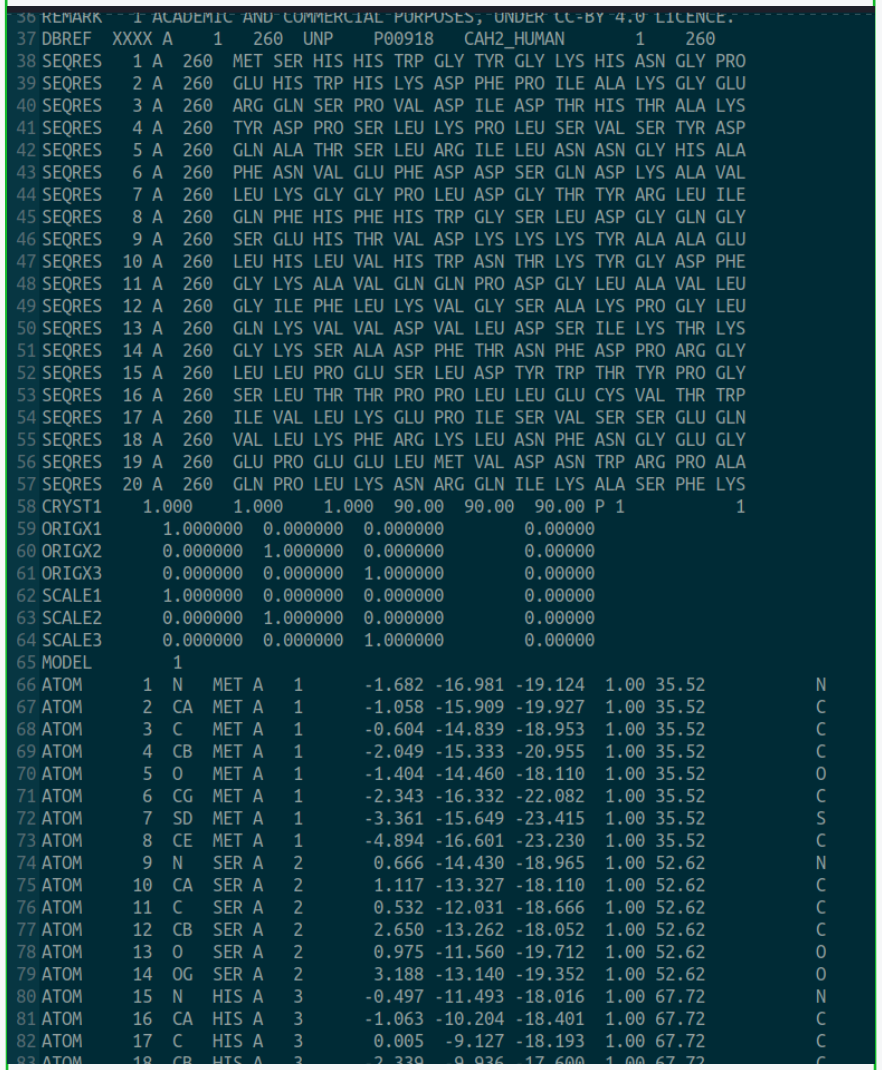


Fig. 4. *PDBbind homepage*

Data processing

PDB data processing

- Initially I thought I would make use of the information contained in the .pdb coordinate files. This required understanding the file format (quite complex) and finding ways to parse them to extract the relevant data.
- After some consideration, I decided that I will not use the 3D information, so the only useful information left in the files was the protein sequence, which can be very informative in itself.
- The protein sequence could have been extracted from the .pdb file, but instead I was able to download the protein sequences of all entries from the PDB
- All the proteins that I was working with from the PDBbind database had a PDB ID, so I wrote a script that extracted from the sequences file all the sequences associated with the PDBIDs that I had
- I then wrote some functions to extract some potentially relevant features from the protein sequences (different quantities and metrics determined from the amino acid composition)



```
REMARK   1 ACADEMIC AND COMMERCIAL PURPOSES, UNDER CC-BY 4.0 LICENCE.
DBREF  XXXX A    1   260  UNP    P00918   CAH2_HUMAN       1    260
SEQRES   1 A  260  MET SER HIS HIS TRP GLY TYR GLY LYS HIS ASN GLY PRO
SEQRES   2 A  260  GLU HIS TRP HIS LYS ASP PHE PRO ILE ALA LYS GLY GLU
SEQRES   3 A  260  ARG GLN SER PRO VAL ASP ILE ASP THR HIS THR ALA LYS
SEQRES   4 A  260  TYR ASP PRO SER LEU LYS PRO LEU SER VAL SER TYR ASP
SEQRES   5 A  260  GLN ALA THR SER LEU ARG ILE LEU ASN ASN GLY HIS ALA
SEQRES   6 A  260  PHE ASN VAL GLU PHE ASP ASP SER GLN ASP LYS ALA VAL
SEQRES   7 A  260  LEU LYS GLY GLY PRO LEU ASP GLY THR TYR ARG LEU ILE
SEQRES   8 A  260  GLN PHE HIS PHE HIS TRP GLY SER LEU ASP GLY GLN GLY
SEQRES   9 A  260  SER GLU HIS THR VAL ASP LYS LYS LYS TYR ALA ALA GLU
SEQRES  10 A  260  LEU HIS LEU VAL HIS TRP ASN THR LYS TYR GLY ASP PHE
SEQRES  11 A  260  GLY LYS ALA VAL GLN GLN PRO ASP GLY LEU ALA VAL LEU
SEQRES  12 A  260  GLY ILE PHE LEU LYS VAL GLY SER ALA LYS PRO GLY LEU
SEQRES  13 A  260  GLN LYS VAL VAL ASP VAL LEU ASP SER ILE LYS THR LYS
SEQRES  14 A  260  GLY LYS SER ALA ASP PHE THR ASN PHE ASP PRO ARG GLY
SEQRES  15 A  260  LEU LEU PRO GLU SER LEU ASP TYR TRP THR TYR PRO GLY
SEQRES  16 A  260  SER LEU THR THR PRO PRO LEU LEU GLU CYS VAL THR TRP
SEQRES  17 A  260  ILE VAL LEU LYS GLU PRO ILE SER VAL SER SER GLU GLN
SEQRES  18 A  260  VAL LEU LYS PHE ARG LYS LEU ASN PHE ASN GLY GLU GLY
SEQRES  19 A  260  GLU PRO GLU GLU LEU MET VAL ASP ASN TRP ARG PRO ALA
SEQRES  20 A  260  GLN PRO LEU LYS ASN ARG GLN ILE LYS ALA SER PHE LYS
CRYST1    1.000    1.000    1.000  90.00  90.00  90.00 P 1           1
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      1.000000  0.000000  0.000000        0.00000
SCALE2      0.000000  1.000000  0.000000        0.00000
SCALE3      0.000000  0.000000  1.000000        0.00000
MODEL        1
ATOM      1  N   MET A   1      -1.682 -16.981 -19.124  1.00 35.52           N
ATOM      2  CA  MET A   1      -1.058 -15.909 -19.927  1.00 35.52           C
ATOM      3  C   MET A   1      -0.604 -14.839 -18.953  1.00 35.52           C
ATOM      4  CB  MET A   1      -2.049 -15.333 -20.955  1.00 35.52           C
ATOM      5  O   MET A   1      -1.404 -14.460 -18.110  1.00 35.52           O
ATOM      6  CG  MET A   1      -2.343 -16.332 -22.082  1.00 35.52           C
ATOM      7  SD  MET A   1      -3.361 -15.649 -23.415  1.00 35.52           S
ATOM      8  CE  MET A   1      -4.894 -16.601 -23.230  1.00 35.52           C
ATOM      9  N   SER A   2       0.666 -14.430 -18.965  1.00 52.62           N
ATOM     10  CA  SER A   2       1.117 -13.327 -18.110  1.00 52.62           C
ATOM     11  C   SER A   2       0.532 -12.031 -18.666  1.00 52.62           C
ATOM     12  CB  SER A   2       2.650 -13.262 -18.052  1.00 52.62           C
ATOM     13  O   SER A   2       0.975 -11.560 -19.712  1.00 52.62           O
ATOM     14  OG  SER A   2       3.188 -13.140 -19.352  1.00 52.62           O
ATOM     15  N   HIS A   3      -0.497 -11.493 -18.016  1.00 67.72           N
ATOM     16  CA  HIS A   3      -1.063 -10.204 -18.401  1.00 67.72           C
ATOM     17  C   HIS A   3       0.005  -9.127 -18.193  1.00 67.72           C
```

Fig. 5. *Example .pdb file*

Data processing

DSSP processing

- The secondary structure of proteins is also a likely relevant aspect in the binding of ligands
- To get the DSSP files associated with the proteins in my dataset, I made use of the API that is available.

```
==== Secondary Structure Definition by the program DSSP, NKI version 4.0                    ==== DATE=2021-11-26
REFERENCE W. KABSCH AND C.SANDER, BIOPOLYMERS 22 (1983) 577-2637
HEADER    LYASE (OXO-ACID)                        21-JUL-95   1CNX
COMPND    MOL_ID: 1; MOLECULE: CARBONIC ANHYDRASE II; CHAIN: A; SYNONYM: CARBONATE DEHYDRATASE, HCA II; EC: 4.2.1.1
SOURCE    MOL_ID: 1; ORGANISM_COMMON: HUMAN; ORGANISM_SCIENTIFIC: HOMO SAPIENS; ORGANISM_TAXID: 9606
AUTHOR    P.A.Boriack; D.W.Christianson
  256  1  0  0  0 TOTAL NUMBER OF RESIDUES, NUMBER OF CHAINS, NUMBER OF SS-BRIDGES(TOTAL,INTRACHAIN,INTERCHAIN)
 11760.0   ACCESSIBLE SURFACE OF PROTEIN (ANGSTROM**2)
  168 65.6   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(J)  , SAME NUMBER PER 100 RESIDUES
   18  7.0   TOTAL NUMBER OF HYDROGEN BONDS IN     PARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
   57 22.3   TOTAL NUMBER OF HYDROGEN BONDS IN ANTIPARALLEL BRIDGES, SAME NUMBER PER 100 RESIDUES
    2  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-5), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-4), SAME NUMBER PER 100 RESIDUES
    2  0.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-3), SAME NUMBER PER 100 RESIDUES
    1  0.4   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-2), SAME NUMBER PER 100 RESIDUES
    1  0.4   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I-1), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+0), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+1), SAME NUMBER PER 100 RESIDUES
   25  9.8   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+2), SAME NUMBER PER 100 RESIDUES
   37 14.5   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+3), SAME NUMBER PER 100 RESIDUES
   16  6.2   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+4), SAME NUMBER PER 100 RESIDUES
    0  0.0   TOTAL NUMBER OF HYDROGEN BONDS OF TYPE O(I)-->H-N(I+5), SAME NUMBER PER 100 RESIDUES
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30     *** HISTOGRAMS OF ***
  0  0  0  1  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    RESIDUES PER ALPHA HELIX
  1  2  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    PARALLEL BRIDGES PER LADDER
  2  2  2  1  1  1  0  1  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    ANTIPARALLEL BRIDGES PER LADDER
  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    LADDERS PER SHEET
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    4 A H              0   0  201      0, 0.0     0, 0.0     0, 0.0     0, 0.0   0.000 360.0 360.0 360.0-173.5    9.8    0.2    9.0
    2    5 A W        +     0   0   67      1,-0.1     2,-0.2    11,-0.1     7,-0.2  -0.146 360.0 136.3 -56.1 150.4    7.5   -1.5   11.5
    3    6 A G        -     0   0   12      5,-2.9    10,-0.1     2,-0.2    -1,-0.1  -0.770  63.5 -89.6-164.9-153.5    5.4   -4.5   10.5
    4    7 A Y  S    S+     0   0   33     -2,-0.2     5,-0.2     5,-0.2   237,-0.1   0.143  82.2 112.6-127.2  17.1    4.4   -7.9   11.7
    5    8 A G  S  > S-     0   0   26      4,-0.1     4,-0.9     1,-0.1    -2,-0.2  -0.172  84.6 -98.0 -79.6 176.9    7.3   -9.9   10.2
    6    9 A K  T  4 S+     0   0  170      1,-0.2    -1,-0.1     2,-0.2     5,-0.0   0.745 122.6  41.3 -68.6 -29.0   10.1  -11.6   12.2
    7   10 A H  T  4 S+     0   0  167      1,-0.1    -1,-0.2     4,-0.0    -5,-0.0   0.569 129.4  24.0 -96.8 -15.7   12.5   -8.7   11.7
    8   11 A N  T  4 S+     0   0   55     -3,-0.1    -5,-2.9    -6,-0.0    -2,-0.2   0.100  94.6 128.1-136.6  19.5   10.2   -5.7   12.2
    9   12 A G  S >X S-     0   0    0     -4,-0.9     3,-2.6    -5,-0.2     4,-0.8   0.018  79.0 -77.0 -73.0 175.7    7.4   -7.2   14.3
   10   13 A P  T 34 S+     0   0   26      0, 0.0     3,-0.5     0, 0.0     4,-0.4   0.740 126.1  64.0 -41.4 -40.1    5.9   -6.1   17.7
   11   14 A E  T 34 S+     0   0  108      1,-0.2     3,-0.1     2,-0.1    -5,-0.1   0.661 108.1  42.1 -63.9 -21.9    8.9   -7.4   19.6
   12   15 A H  T X4 S+     0   0   45     -3,-2.6     3,-2.0     1,-0.1     4,-0.4   0.572  86.5  91.5 -98.3 -21.4   11.0   -4.8   17.8
   13   16 A W  G >X S+     0   0    5     -4,-0.8     4,-2.7    -3,-0.5     3,-0.7   0.706  72.9  71.3 -51.9 -29.9    8.7   -1.8   17.9
   14   17 A H  G 34 S+     0   0   68     -4,-0.4    -1,-0.3     1,-0.2     6,-0.1   0.747  88.0  66.1 -56.1 -31.5   10.2   -0.5   21.3
   15   18 A K  G <4 S+     0   0  134     -3,-2.0    -1,-0.2     1,-0.2    -2,-0.2   0.872 116.5  23.0 -61.3 -43.4   13.4    0.4   19.4
   16   19 A D  T <4 S+     0   0  126     -3,-0.7    -2,-0.2    -4,-0.4    -1,-0.2   0.709 135.3  37.1 -94.3 -25.4   11.7    3.1   17.4
   17   20 A F    ><  -     0   0   59     -4,-2.7     3,-2.2     1,-0.1    -1,-0.2  -0.809  65.0-177.2-131.4  84.0    8.8    3.8   19.7
   18   21 A P  G >  S+     0   0  104      0, 0.0     3,-1.7     0, 0.0     -4,-0.1   0.610  76.8  76.1 -56.7 -21.2   10.0    3.5   23.3
   19   22 A I  G >  S+     0   0   61     -1,-0.3     3,-3.1   179,-0.2     -5,-0.1   0.641  70.3  93.1 -65.8 -12.4    6.4    4.0   24.7
```

Fig. 6. *Example .dssp file*

- I wrote a script that queried the PDBIDs of the proteins in my dataset, and downloaded their respective DSSP file
- After getting all the DSSP files, they had to be parsed by another python program that I wrote which extracted the relevant features
- All these, combined with the features obtained from the protein sequence, formed the descriptors for the protein

Molecular descriptors

- The **Mordred** python library (Moriwaki et. al 2018) was used to calculate the descriptors for the ligands
- Doing this required some extensive research into molecular descriptor types, since the library provided more than 1800 options, and they would have been excessive for my model
- Unfortunately, information was very hard to find regarding molecular descriptors, and the library was not particularly well documented. I had to choose the descriptors that I could find information about that also seemed relevant (107 in total)

Issues

- Mordred encountered issues when trying to compute values from most of the ligand files (because of the coordinate system they were in), so I had to replace them by downloading them from the PDB (in a different coordinate system)
- some occasional issues still occurred, like missing data points, but they were fixed by just imputing them with an average value of that field

Predictive model

- 4 machine learning models from the python Scikitlearn library were trained (linear regression, decision tree regressor, random forest regressor, multilayer perceptron)
- The best performing was a random forest regressor, which achieved a Pearson's correlation coefficient of 0.789 and a root mean square error of 1.25
- A brief exploration of some different parameter combinations (such as the number of estimator trees and max features) was done to get to this result

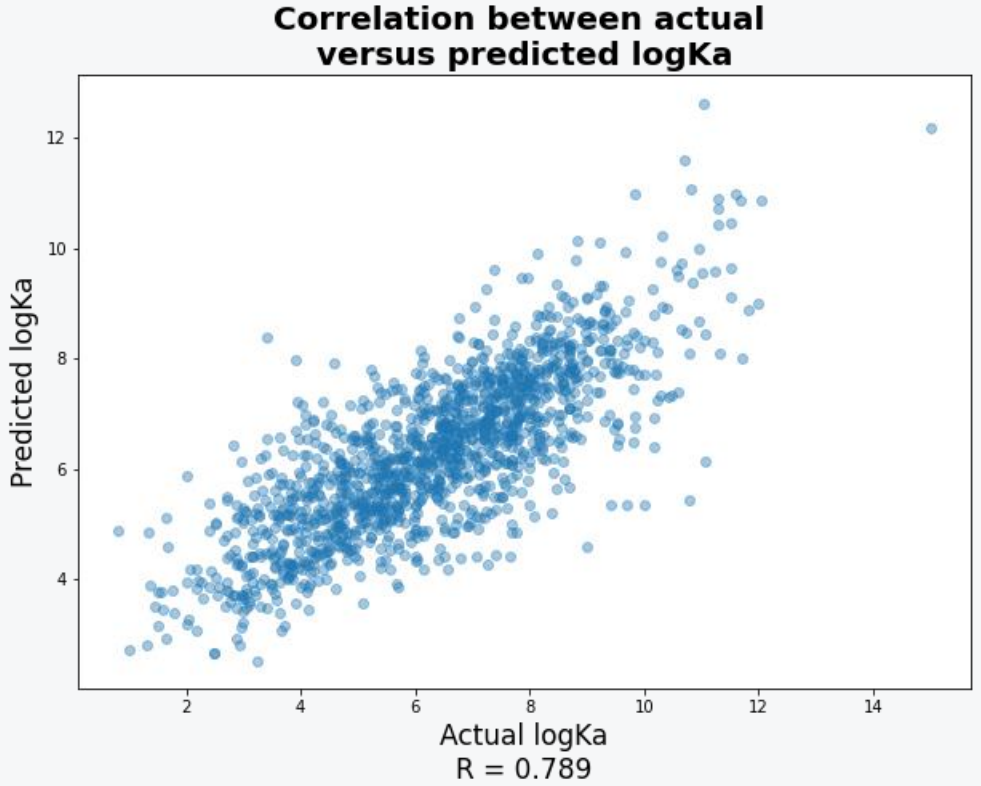


Fig. 7. *Random forest regressor predicted versus actual logKa*

Lack of more rigorous testing

- The project plan mentioned that the model will be tested on the CASF-2016, which is a refined set of over 200 high quality protein structures from a wide range of protein classes
- Unfortunately, due to the many unexpected operations that were needed to prepare the data, I did not have enough time to perform extensive testing
- The CASF-2016 test would have required some more data processing and sourcing, and it was also in a slightly different format, so new methods would have had to be developed to process it

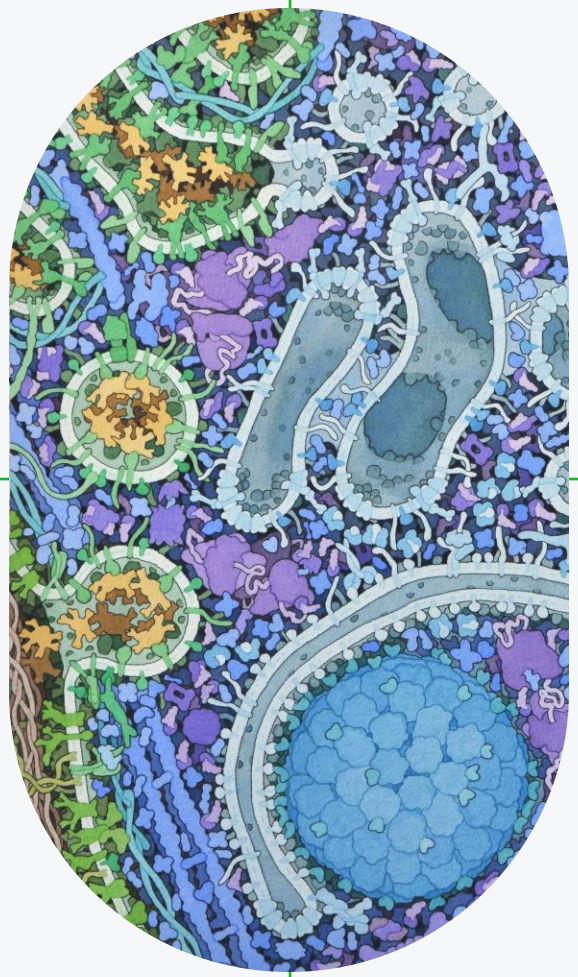


Fig. 8. *Autophagy, D.S. Goodsell & D. Klionsky, 2011*

Command line functionality

- Originally, the project was intended to be usable, but because of poor time management I only succeeded to train the model
- A command line program was developed nonetheless, after the submission deadline
- The script can take .pdb and .sdf (molecule structure file) files as inputs and will output a pandas dataframe with the predicted scores

Issues

- The program is quite slow because it uses the DSSP API to retrieve the DSSP files live, and it can take up to a few seconds per .pdb file
- Error handling is not implemented yet, so if unexpected issues arise during usage, it might be hard for inexperienced users to diagnose what went wrong

Future improvements

The model is not ideal in many aspects, so here are some ways I plan to update it

- Remove features that have a low correlation with the label data (binding strength)
- Many of the DSSP features had low correlation, and the retrieval of the files is slow, so not using them at all might be better
- More protein related descriptors could be added, possibly calculated with a library for this (at the time of writing the report none was found, but now I found some)
- Try to add more or different molecular descriptors, since the model did not take long to train with the current number
- Create a web interface for better usability

References

- Figure 1: Illustration by David S. Goodsell. doi: 10.2210/rcsb_pdb/goodsell-gallery-016
- Figure 2: Illustration by David S. Goodsell. doi: 10.2210/rcsb_pdb/goodsell-gallery-016
- Figure 3: Protein Data Bank website - https://www.rcsb.org/
- Figure 4: PDBbind website - http://pdbbind.org.cn/
- Figure 7: Plot generated with matplotlib python library - https://matplotlib.org/
- Figure 8: Illustration by David S. Goodsell and Daniel Klionsky. doi: 10.2210/rcsb_pdb/goodsell-gallery-012
- Moriwaki, H., Tian, YS., Kawashita, N. *et al.* Mordred: a molecular descriptor calculator. *J Cheminform* **10,** 4 (2018). https://doi.org/10.1186/s13321-018-0258-y