Predicting the activity of protein-ligand complexes

Lukas Fallmann

BACHELORARBEIT

eingereicht am

Fachhochschul-Bachelorstudiengang

Medizin- und Bioinformatik

in Hagenberg

im Juni 2023

Advisor:

Micha Johannes Birklbauer, M.Sc.

Declaration

I hereby declare and confirm that this thesis is entirely the result of my own original work. Where other sources of information have been used, they have been indicated as such and properly acknowledged. I further declare that this or similar work has not been submitted for credit elsewhere. This printed copy is identical to the submitted electronic version.

Hagenberg, June 27, 2023

Lukas Fallmann

Contents

Preface

Abstract

This should be a 1-page (maximum) summary of your work in English.

Kurzfassung

An dieser Stelle steht eine Zusammenfassung der Arbeit, Umfang max. 1 Seite. ...

Chapter 1

Introduction

The discovery of new drugs or any chemically active compounds for that matter is an expensive and time-consuming process. It has been estimated, that it takes about 14 Years from the initial discovery of a promising new compound to the release of a marketable drug[1]. In addition to that the price of this drug-discovery circle ranges up to 800 Million Dollars[2]. All techniques which aim to improve the efficiency of drug discovery can be generalized as one of two methods. These two are called High-throughput-screening (HTS) and virtual screening (VS)[1].
When using an HTS-approach there are many compounds which are tested against some type of target protein. During testing, it is measured whether a certain compound biochemically interacts with a protein. Those interacting combinations are considered active and are marked by researchers as hits. To improve the performance of HTS there are a number of factors to consider. Through miniaturization, it is possible to investigate more compounds at the same time. With a higher throughput quality-control is more time-consuming and leads to an overall more expensive process. For this reason HTS is most efficient, when analyzing a small set of compounds as the technology is not suitable for large datasets[3].
In contrast to the in vitro approach of HTS, VS is a theoretical in silico approach. To save resources in the laboratory the activity of certain compounds is predicted using a preexisting library of small molecules. The activity can be predicted using the ligands of a compound and their respective binding sites or the 3D structure of a compound. The Key idea behind the ligand based approach (LBVS) is that similar compounds have similar chemical properties. Therefore, the goal of LBVS is to find molecules which have similar or identical chemical properties as the sample compound[4]. Structure-based VS uses the 3D structure of a compound to predict which molecules from the dataset will bind to the provided sample. Each molecule of a certain database subset is fitted (docked) to the sample. Hereby it is important to differentiate between rigid and flexible docking.
In rigid docking the dataset sample is rotated and translated in a six-dimensional space in order to fit the sample protein. For each fitted molecule a score is calculated based on how well the molecule fits to the sample[5]. Although this algorithm often predicts actual possible binding sites and bound proteins, there is no guarantee that this compound will actually bind in vitro. Therefore, predicted interactions should be seen as a hypothesis. Still rigid docking provides a great baseline at a comparatively low cost[4]. The low

accuracy of rigid docking is due to the nature of biochemical substances as samples in a database can only provide a snapshot of a sample. With flexible docking it is possible to simulate moving binding sites. The flexibility can be introduced at different stages. Implicit flexibility is achieved by smoothing protein surfaces and therefore allowing room for interpretation when docking. Cross- or Ensemble docking can be done by repeating the docking process with different conformations. Explicit flexibility is reached through allowing side-chain flexibility. Most commonly utilized is the approach where the ligand is flexible, and the receptor is rigid. Even though this approach does provide better more accurate results it takes considerably longer to compute[6].
Regardless of the docking type the score decides which pose between a protein and a ligand is most likely to exist. In addition to that, the score also determines whether a protein-ligand complex is considered active. There are a lot of different scoring functions which can be grouped into four categories: physics-based, empirical, knowledge-based, and machine learning-based[7].
The focus of this work is on implementing a machine-learning based scoring approach. Machine-learning based scoring functions work by training on pre-classified data and finding the best model for predicting future data. To accurately and efficiently train a model crucial binding sites need to be identified beforehand. The basis of this thesis is the master thesis of Birklbauer Micha[8]. In his thesis a selection of eleven proteins from the directory of useful decoys have been selected to be analyzed. For the selected proteins all possible interactions have been analyzed. Based on the interaction-data a few basic scoring functions have been implemented. The direct result of this thesis are proteins and the frequency of their interactions.
Since this work aims to implement different machine learning algorithms for use in scoring functions the state of the art is described in the following.

1.1 Machine Learning in drug design and activity prediction

The following chapter summarizes the recent developments in drug design using various machine learning techniques.

Today there exist a multitude of machine learning approaches in the field of drug design and activity prediction. As a result of various AI breakthroughs in recent years there have been numerous research projects regarding the usability of artificial intelligence in various bioinformatic domains. MILCDock uses the Output of five traditional Scoring Functions as input for a neural Network. This technique has a slight performance benefit when compared to traditional scoring functions [9].

1.2 Goals

The goals of this thesis are twofold:

1. Evaluate common machine learning approaches for activity prediction and compare results with current literature.
2. Fine tune the provided datasets using a multitude of feature-engineering practices, to increase performance of the ML approaches.

The second goal can be viewed as an extension of the first one since its primary aim is to improve the results achieved while pursuing the first goal.

Chapter 2

Methods

2.1 Data description

The following chapter is dedicated to explaining the data used for this thesis. This includes detailed descriptions of the protein-ligand complexes as well as their interaction types.

2.1.1 Proteins

The following five proteins have been used to conclude this thesis:

Acetylcholinesterase

Cyclooxygenase 1

Dipeptidyl peptidase IV

Monoamine oxidase B

Soluble epoxide hydrolase

2.1.2 Interactions

Overview of the interactions between protein and ligands used in this thesis.

2.1.3 Data structure

One or two sentences describing the data csv structure

2.2 Data partitioning

To validate the results of training various machine learning methods the provided data concerning the five targets was split into a training-set as well as a test-set. To achieve a 70/30 train/test split ratio each sample was randomly assigned to one of the two datasets[10].

In order to validate the machine learning approaches during training 10-fold cross-validation has been applied. For the process of cross-validation the training dataset is

split into n equally large subsets. The type of cross-validation implemented in this thesis uses all but one of these partitions to train the classification model and validates the results with the remaining partition. This process is repeated for all possible validation partitions[11].

2.3 Machine-Learning approaches

Introduction to the ML Approaches used for the thesis.

2.4 Quality metrics

In order to make the results from this thesis comparable to the results from the scoring function introduced in [8] by Micha Birklbauer the same quality metrics have been implemented for this thesis. The following will provide an overview for the used metrics.

2.4.1 Terminology

To calculate the metrics that are mentioned within this chapter a few base numbers are necessary:

TP – **T**rue **P**ositives are active samples, which are classified as such

TN – **T**rue **N**egatives are inactive samples, which are classified as such

FP – **F**alse **P**ositives are inactive samples, which are classified as active

FN – **F**alse **N**egatives are active samples, which are classified as inactive

2.4.2 Visual metrics

For better visualization of the four base metrics mentioned in 2.4.1 this thesis displays the resulting data in a confusion matrix. This metric displays distribution of the results over the four base metrics.

In addition to that, the ROC(receiver operating characteristic) curve will also be displayed for the results. The ROC curve is a collection of points in a two-dimensional space, where their location is defined by the FPR 2.4.4 on the x-axis and the TPR($\frac{\#TP}{\#TP+\#FN}$) on the y-axis. Each point on this line depicts the ratio of FPR to TPR at a certain cutoff. [12]

2.4.3 Accuracy

Accuracy (ACC) describes which portion of the predicted samples was accurate and is defined as follows:

$$\text{ACC} = \frac{\#TP + \#TN}{\#TP + \#TN + \#FP + \#FN}$$

[13]

2.4.4 False positive Rate

False positive rate (FPR) describes the incorrect as active identified compounds in relation to all inactive compounds and is defined as follows:

$$\text{FPR} = \frac{\#FP}{\#TN + \#FP}$$

[12]

2.4.5 Area under the curve

Area under the curve (AUC) is a metric which stems from the ROC curve. The area under the ROC curve is calculated using the scikit-learn package.

2.4.6 Yield of Actives

Yield of actives (Ya) describes the correct as active identified compounds in relation to all as active labeled compounds and is defined as follows:

$$\text{Ya} = \frac{\#TP}{\#TP + \#FP}$$

[14]

2.4.7 Enrichment Factor

The enrichment factor (EF) describes the relation of the truly active compounds among all as active predicted complexes and the relative share of active compounds in the dataset. This metric is defined as follows:

$$\text{EF} = \frac{\frac{\#TP}{\#TP+\#FP}}{\frac{\#TP+\#FN}{\#TP+\#TN+\#FP+\#FN}}$$

[12]

2.4.8 Relative Enrichment Factor

The relative enrichment factor (REF) describes the relation of the EF to the maximum achievable EF. The REF is defined as follows:

$$\text{REF} = \frac{100 * \#TP}{\min(\#TP + \#FP, \#TP + \#FN)}$$

[12]

2.5 Feature engineering

Explain concept of feature engineering and possible implications for thesis.

Chapter 3

Results

3.1 Performance per Protein-Complex

Evaluate models tuned for selected proteins.

3.2 Performance Overview – Comparing ML-approaches

Compare performance of overall ml approaches.

Chapter 4

Discussion

4.1 Conclusion

Recap findings of thesis.

4.2 Improvements and outlook

Explain possible improvements to used technique. Provide general outlook on topic.

Appendix A

Technical Details

Appendix B

Supplementary Materials

List of supplementary data submitted to the degree-granting institution for archival storage (in ZIP format).

B.1 PDF Files

Path: /

thesis.pdf . . . . . Master/Bachelor thesis (complete document)

B.2 Media Files

Path: /media

*.ai, *.pdf . . . . . Adobe Illustrator files
*.jpg, *.png . . . . . raster images
*.mp3 . . . . . audio files
*.mp4 . . . . . video files

B.3 Online Sources (PDF Captures)

Path: /online-sources

Reliquienschrein-Wikipedia.pdf [15]

Appendix C

Questionnaire

Appendix D

LaTeX Source Code

References

Literature

[1] S. Myers and A. Baker. "Drug discovery–an operating model for a new era". eng. *Nature Biotechnology* 19.8 (Aug. 2001), pp. 727–730. DOI: 10.1038/90765 (cit. on p. 1).

[2] Joseph A. DiMasi, Ronald W. Hansen, and Henry G. Grabowski. "The price of innovation: new estimates of drug development costs". eng. *Journal of Health Economics* 22.2 (Mar. 2003), pp. 151–185. DOI: 10.1016/S0167-6296(02)00126-1 (cit. on p. 1).

[3] Lorenz M. Mayr and Peter Fuerst. "The Future of High-Throughput Screening". *SLAS Discovery* 13.6 (July 2008), pp. 443–448. DOI: 10.1177/1087057108319644. (Visited on 02/29/2024) (cit. on p. 1).

[4] Aleix Gimeno et al. "The Light and Dark Sides of Virtual Screening: What Is There to Know?" *International Journal of Molecular Sciences* 20.6 (Mar. 2019), p. 1375. DOI: 10.3390/ijms20061375. (Visited on 02/29/2024) (cit. on p. 1).

[5] A. Lavecchia and C. Di Giovanni. "Virtual screening strategies in drug discovery: a critical review". eng. *Current Medicinal Chemistry* 20.23 (2013), pp. 2839–2860. DOI: 10.2174/09298673113209990001 (cit. on p. 1).

[6] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. "Software for molecular docking: a review". *Biophysical Reviews* 9.2 (Jan. 2017), pp. 91–102. DOI: 10.1007/s12551-016-0247-1. (Visited on 02/29/2024) (cit. on p. 2).

[7] Jin Li, Ailing Fu, and Le Zhang. "An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking". en. *Interdisciplinary Sciences: Computational Life Sciences* 11.2 (June 2019), pp. 320–328. DOI: 10.1007/s12539-019-00327-w. (Visited on 02/29/2024) (cit. on p. 2).

[8] Micha Johannes Birklbauer. "Automatic identification of important interactionsand interaction-frequency-based scoring inprotein-ligand complexes". MA thesis. FH Hagenberg, Aug. 31, 2021 (cit. on pp. 2, 5).

[9] Connor Morris et al. "MILCDock: Machine Learning Enhanced Consensus Docking for Virtual Screening in Drug Discovery". *Journal of chemical information and modeling* 62 (Nov. 2022). DOI: 10.1021/acs.jcim.2c00705 (cit. on p. 3).

[10] Yun Xu and Royston Goodacre. "On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning". en. *Journal of Analysis and Testing* 2.3 (July 2018), pp. 249–262. DOI: 10.1007/s41664-018-0068-2. (Visited on 04/21/2024) (cit. on p. 4).

[11] Annette M. Molinaro, Richard Simon, and Ruth M. Pfeiffer. "Prediction error estimation: a comparison of resampling methods". *Bioinformatics* 21.15 (May 2005), pp. 3301–3307. DOI: 10.1093/bioinformatics/bti499. eprint: https://academic.oup.com/bioinformatics/article-pdf/21/15/3301/50340684/bioinformatics_21_15_3301.pdf (cit. on p. 5).

[12] Julio Cesar Dias Lopes et al. "The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability". *Journal of Cheminformatics* 9.1 (Feb. 2017), p. 7. DOI: 10.1186/s13321-016-0189-4. (Visited on 04/20/2024) (cit. on pp. 5, 6).

[13] Mohammad Hossin and Sulaiman M.N. "A Review on Evaluation Metrics for Data Classification Evaluations". *International Journal of Data Mining & Knowledge Management Process* 5 (Mar. 2015), pp. 01–11. DOI: 10.5121/ijdkp.2015.5201 (cit. on p. 5).

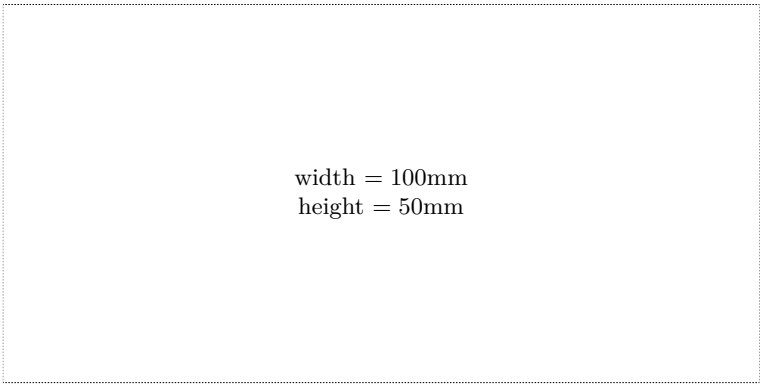
[14] Deborah Giordano et al. "Drug Design by Pharmacophore and Virtual Screening Approach". *Pharmaceuticals* 15.5 (May 2022), p. 646. DOI: 10.3390/ph15050646. (Visited on 04/20/2024) (cit. on p. 6).

Online sources

[15] *Reliquienschrein*. Aug. 29, 2022. URL: https://de.wikipedia.org/wiki/Reliquienschrein (visited on 02/11/2023).

Check Final Print Size

— Check final print size! —



— Remove this page after printing! —