HMG-CoA reductase

Maarten Koffijberg, Youry van Uum | Leiden

Universiteit
Leiden
The Netherlands

Target

3-hydroxy-3-methyl-glutaryl-coenzyme A Reductase (HMG-CoA reductase)

High levels of cholesterol in blood plasma is a leading cause of atherosclerosis, which in turn might lead to an increased risk of heart attack, stroke and peripheral artery disease. According to the World Health Organization, a third of ischaemic heart disease is attributable to high cholesterol levels[1] and incidence of high cholesterol is of worrisome proportions. Previous studies have shown that 86 million U.S. adults have total cholesterol levels above 200mg/dL[2], which is considered high. These numbers illustrate the size and impact high cholesterol levels have.

In the pathway that leads to cholesterol synthesis, 3-hydroxy-3-methyl-glutaryl-coenzyme A (HMG-CoA) gets converted to mevalonic acid by HMG-CoA reductase. Downstream this conversion eventually leads to an increase in the synthesis of cholesterol[3]. This conversion is a rate limiting step, therefore inhibiting HMG-CoA is instrumental for lowering cholesterol levels downstream.

Drugs that are prescribed for high cholesterol levels are such inhibitors and are collectively known as statins. Examples include atorvastatin, fluvastatin, lovastatin and others (see table 1) which are all classified as plasma lipid modifying agents[4]. These competitively inhibit HMG-CoA reductase, bind to the enzyme's active site and therefore prevent HMG-CoA from being converted into mevalonic acid which is a precursor to cholesterol[5]. Table 1 shows an overview of statins, their ATC codes, administrative route and defined daily dose.

Table 1 An overview of HMG-CoA reductase inhibitors

ATC code	Name	DDD	U	Adm.R
C10AA01	simvastatin	30	mg	O
C10AA02	lovastatin	45	mg	O
C10AA03	pravastatin	30	mg	O
C10AA04	fluvastatin	60	mg	O
C10AA05	atorvastatin	20	mg	O
C10AA06	cerivastatin	0.2	mg	O
C10AA07	rosuvastatin	10	mg	O
C10AA08	pitavastatin	2	mg	O

The HMG-CoA reductase protein is part of the HMG-CoA reductase family (Pfam: PF00368), which can be found in multiple species such as humans, mice and rice [6]. It is expressed in cells all over the body as can be seen in figure 1, adapted from Fagerberg et al.

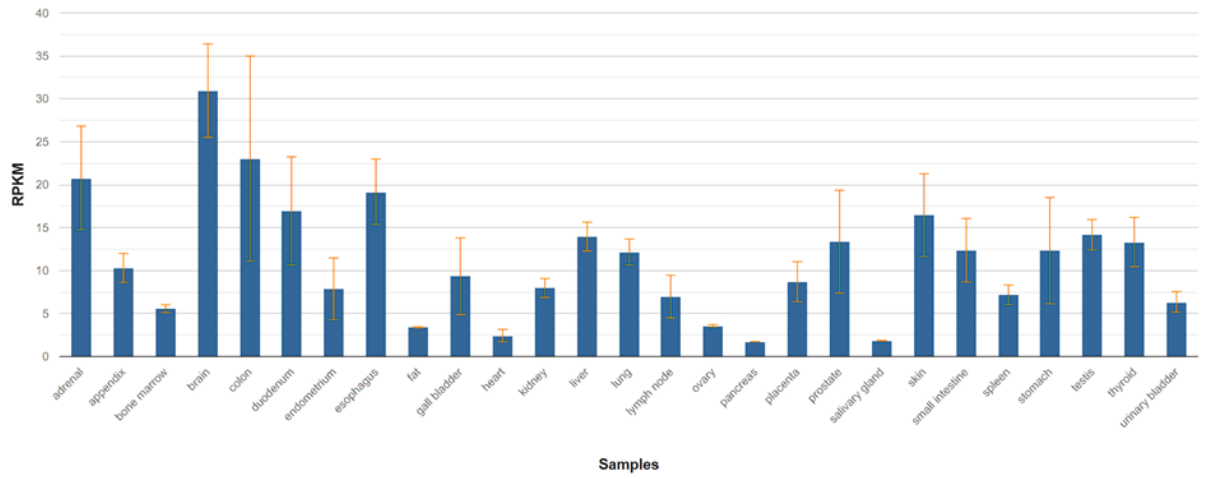

Figure 1 Expression of HMGCR in human tissue

Related proteins (off-targets), based on sequence

Human HMG-CoA Reductase
Entry Name: HMDH_HUMAN
Uniprot primary accession: P04035
Sequence length: 888
Status: UniProtKB reviewed
Protein existence: Evidence at protein level
The value 'Experimental evidence at protein level' indicates that there is clear experimental evidence for the existence of the protein. The criteria include partial or complete Edman sequencing, clear identification by mass spectrometry, X-ray or NMR structure, good quality protein-protein interaction or detection of the protein by antibodies.
Mass: 97,476 kDa

Pan Paniscus CoA Reductase
Entry Name: A0A2R9B572_PANPA
Uniprot primary accession: A0A2R9B572
Sequence length: 888
Species: Pan paniscus (Pygmy chimpanzee) (Bonobo)
Status: UniProtKB unreviewed
Protein existence: Inferred from homology
The value 'Protein inferred by homology' indicates that the existence of a protein is probable because clear orthologs exist in closely related species.
Mass: 97,420 kDa

Based on the BLAST data as summarized above, we can conclude that there is a high level of gene conservation between humans and pygmy chimpanzees. The mass of the proteins is a close macht and the length of the protein sequences are exactly the same. Evolutionary speaking, the two species are closely related and it is therefore not unexpected that the two proteins show a high level of similarity.

Similar Human target:
Entry Name: SCAP_HUMAN
Uniprot primary accession: Q12770
Sequence length: 1279
Status: UniProtKB reviewed
Protein existence: Evidence at protein level
Mass: 139,719 kDa

Scap is a cholesterol-regulated transporter of SREBPs from the endoplasmic reticulum to the Golgi. Both Scap and HMG-CoA are involved in the cholesterol synthesis pathway and both contain a sterol-sensing domain (SSD)[7]. This SSD triggers the ubiquitination and degradation of HMG CoA in the presence of cholesterol. Scap is also regulated by the presence of cholesterol by the same SSD. Any similarity between the protein sequence of Scap and HMG-CoA stem from this SSD.

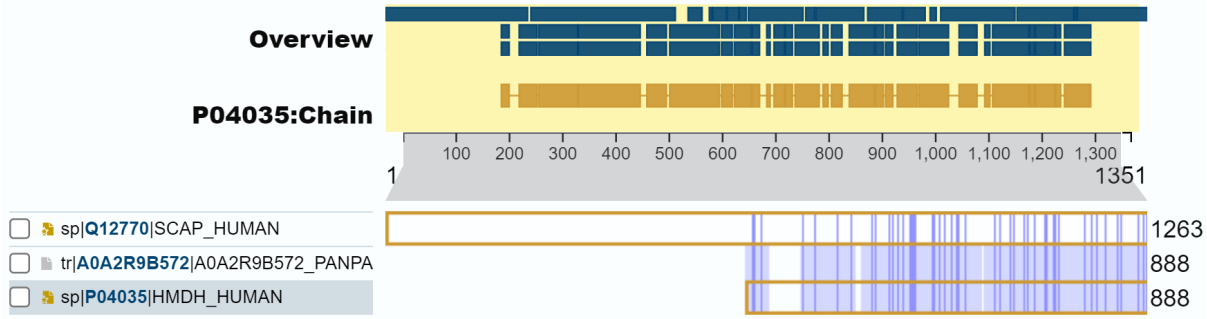

Figure 2 Sequences aligned, showing "Similarity"

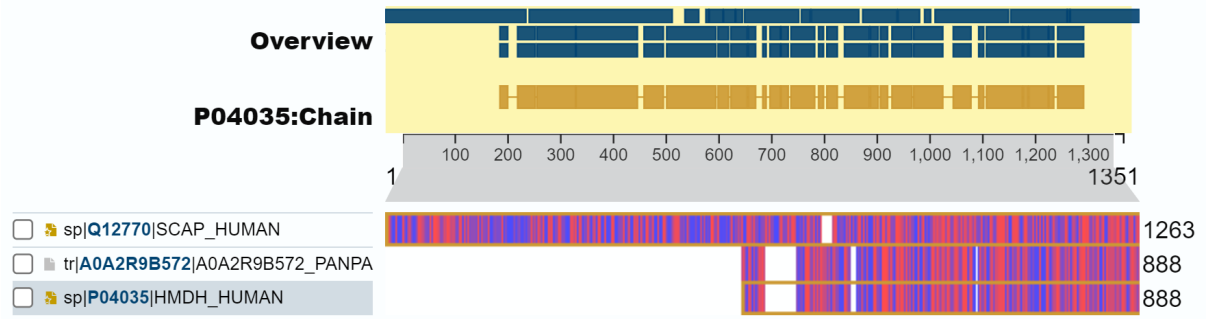

Figure 3 Sequences aligned, showing "Hydrophobic" commonality

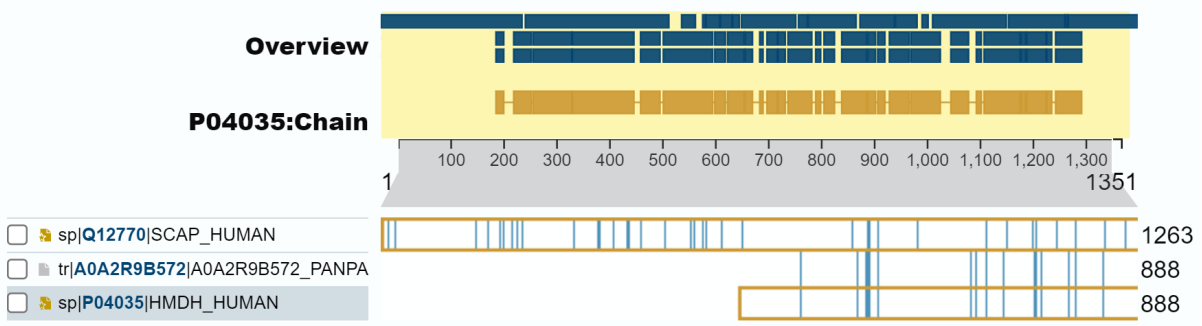

Figure 4 Sequences aligned, showing "Negative" commonality

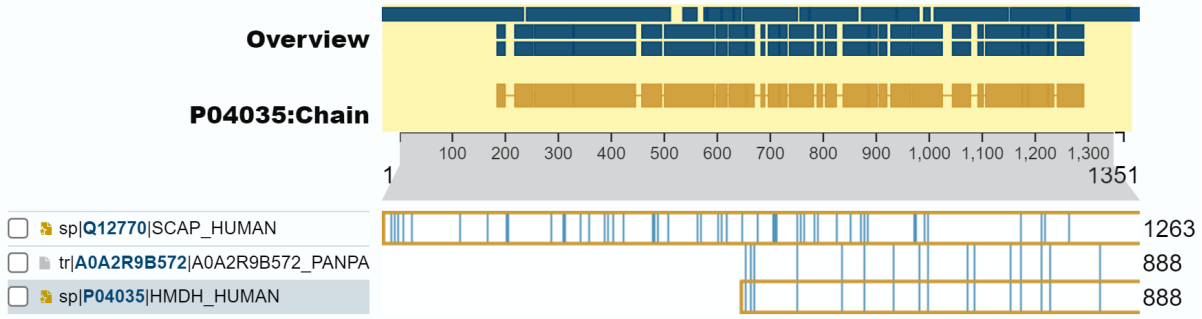

Figure 5 Sequences aligned, showing "Positive" commonality

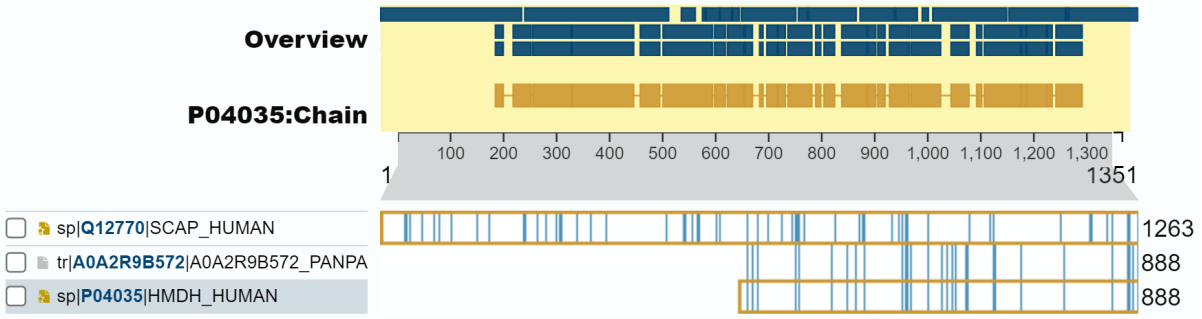

Figure 6 Sequences aligned, showing "Aromatic" commonality

As can be seen in figures 2-6, the protein sequences show a large similarity in all aspects between humans and pygmy chimpanzees and far less between our target and the closest related protein in humans, which is Scap.

Data management

In order to gather the compound data needed for training the various models, we used two different methods which will be explained in this section.

Data gathering

Our first method was based on the provided talktorials which can be found in the "teachopencadd.git" folder, included in this report.

1. Talktorial 1 - We fetched data from ChEMBL based on uniprot id P04035 and ChEMBL id "ChEMBL402". After filtering on IC50 values (>0) and removing duplicates, only 208 compounds with accompanying smiles and bioactivity data remained.
2. Talktorial 2 - Filtering our dataset (HMG-CoA_compounds.csv) on Lipinski's Rule of 5 (Ro5) created in step 1, resulted in the following numbers:
   # compounds in unfiltered data set: 208
   # compounds in filtered data set: 159
   # compounds not compliant with the Ro5: 49
   Based on a pIC50-cutoff value of 6.3, the dataset consisted of:
   Number of active compounds: 113
   Number of inactive compounds: 46
   Figures 7 and 8 show a radar plot of Lipinski's rule of 5 compliant and non-compliant compounds respectively.

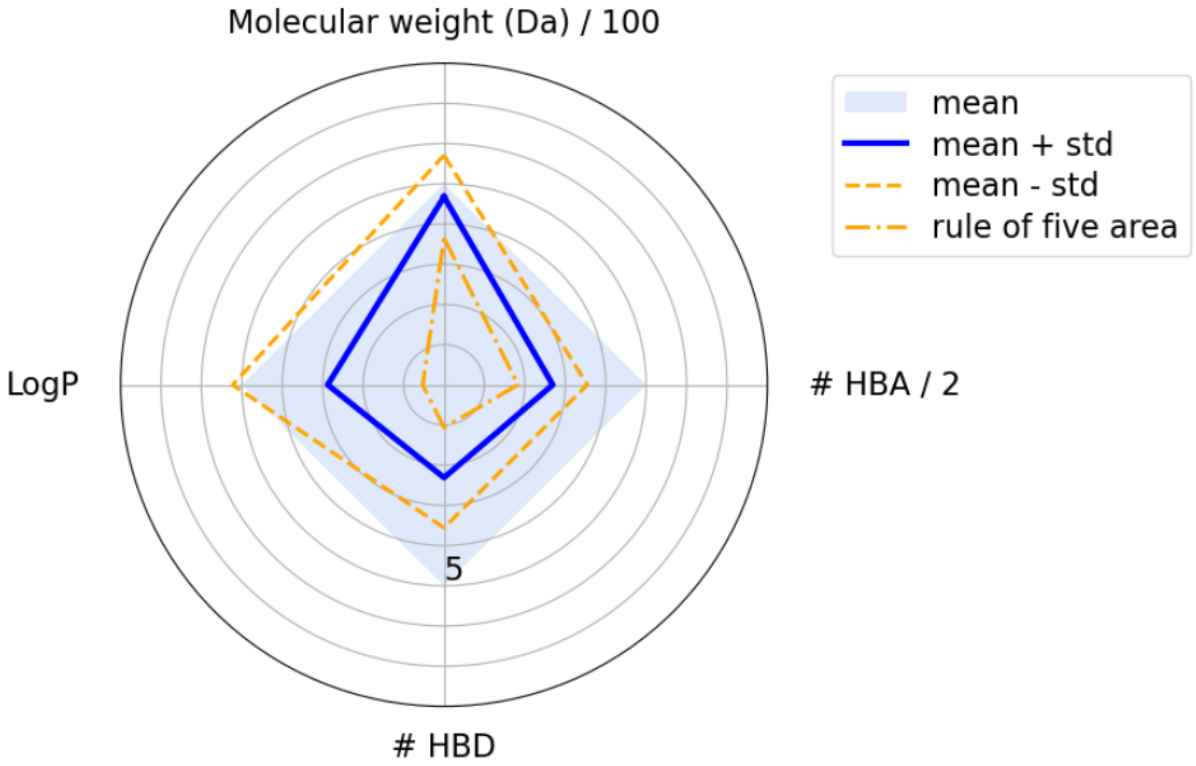

Figure 7 Radar plot showing mean values and standard deviation of all compounds that do not violate Lipinski's Ro5.

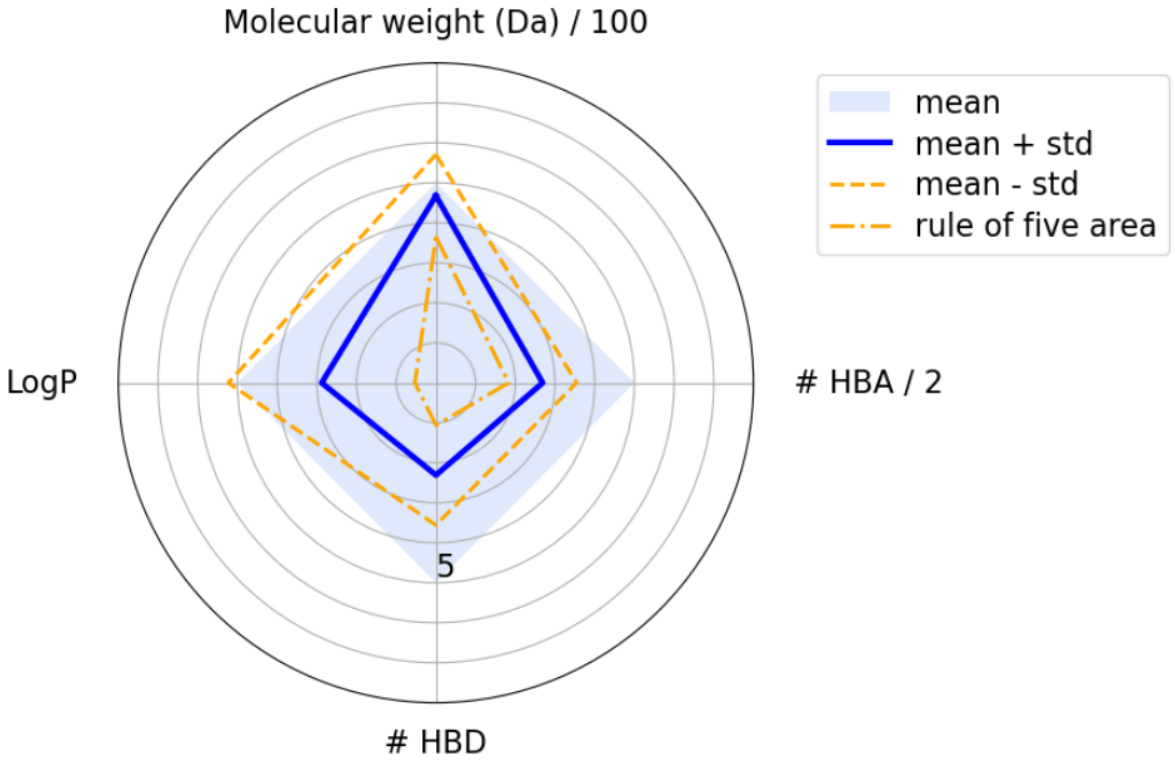

Figure 8 Radar plot showing mean values and standard deviation of all compounds that do violate Lipinski's Ro5.

A dataset of this size (159) turned out to be insufficient for the intended use of training models.

3. Talktorial 7 - Training on this dataset yielded Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN) models that all performed poorly on specificity, as can be seen in table 2. Note the low scores and/or high standard deviation (std) scores, especially those of the specificity of the models.

Table 2 Performance of models trained on the first (small) dataset after cross validation (n=5).

Model	Accuracy (std)	Sensitivity (std)	Specificity (std)	Area Under Curve (std)
RF	0.91 (0.06)	0.96 (0.05)	0.83 (0.18)	0.98 (0.02)
SVM	0.74 (0.07)	0.98 (0.04)	0.18 (0.09)	0.90 (0.07)
ANN	0.86 (0.08)	0.97 (0.04)	0.64 (0.23)	0.96 (0.04)

The Reciever Operating Characteristics (ROC) plot for this dataset returned sub-optimal results (see figure 13).

Since models trained on this dataset are not sufficiently reliable, a larger dataset was needed. In order to gather a larger set, we changed our approach and used different selection parameters. Instructions from “02-MachineLearning” were used.

1. We fetched data from ChEMBL based on uniprot id P04035 and ChEMBL id “ChEMBL402“. After filtering for pChEMBL values and calculating the mean values of any duplicates, our dataset consisted of 1291 compounds.
   Based on a pChEMBL-cutoff value of 6.5, the dataset consisted of:
   Number of active compounds: 181
   Number of inactive compounds: 1110
   Note: all compounds without a pChEMBL value are classified as inactive.
2. In order to keep a large enough dataset to train our classification and regression models on, the decision was made to not filter on Lipinski’s Ro5. This gives our models the chance to learn what structures are useful, while filtering on compounds for the Ro5 is still possible during virtual screening.
3. For training of the regression model, only data from compounds with an actual pChEMBL value could be used. After removing all compounds without the required data, our dataset consisted of 264 compounds, a significant improvement over our smaller dataset. Performance of classification models trained on this dataset can be reviewed in the section: Machine Learning (see table 3).

Compound clustering

Our goal in this project is to design a few compounds by hand and evaluate them using machine learning. The design of these ligands is based on the Most Common Substructures (MCS) from our dataset. In order to get the MCS, our dataset had to be clustered based on the Tanimoto similarity index. Talktorial 5, which can be found in the included files, was used for this purpose. A distance cutoff of 0.4 was chosen as this value resulted in a smooth distribution and only a few singletons. Data on the distribution of the clusters can be found in figure 9.

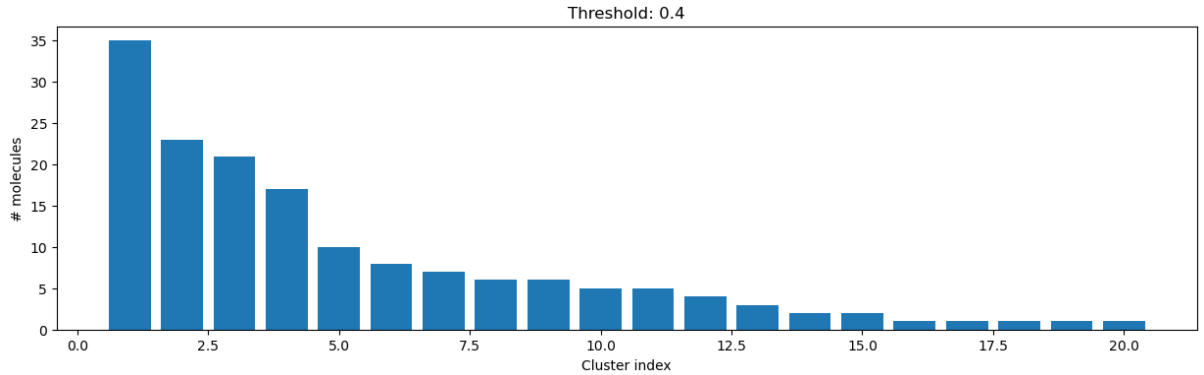

Figure 9 Cluster distribution. Number of clusters: 20 from 159 molecules at distance cut-off 0.40. Number of molecules in largest cluster: 35. Similarity between two random points in same cluster: 0.64. Similarity between two random points in different cluster: 0.40.

The largest cluster of compounds consists of 35 molecules (figure 10). Visual inspection shows a few substructures that all molecules share.

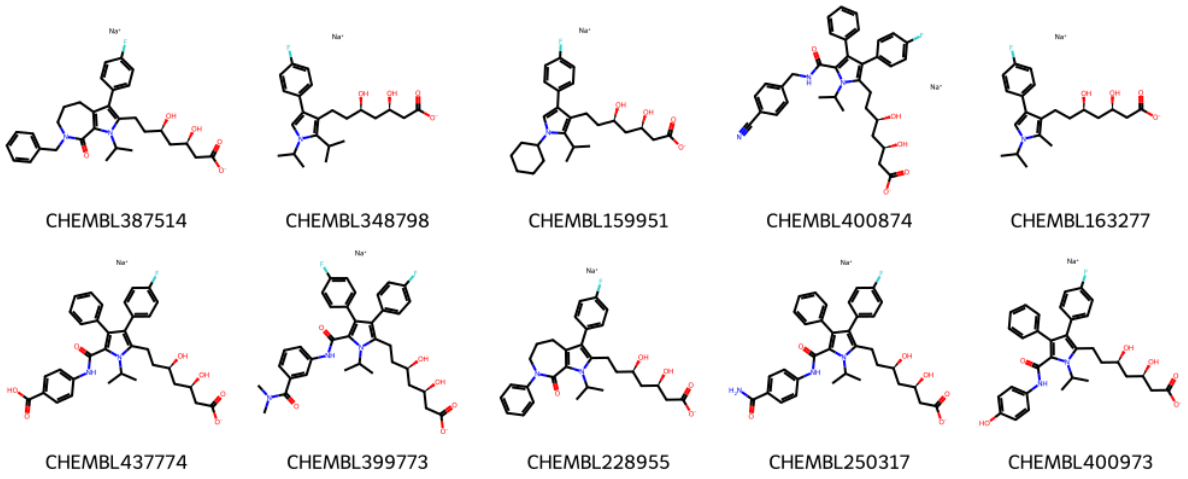

Figure 10 The first 10 molecules from the largest cluster.

Design of the novel ligands was based on the MCS of both the largest cluster and unclustered data of all compounds with pIC50 values > 9, see figures 11 and 12. Creation of the MCS overview was done according to the instructions in talktorial 6.

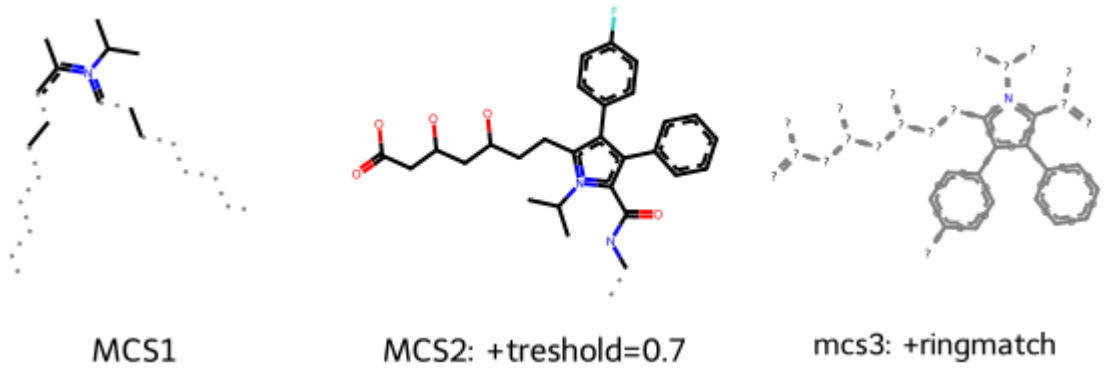

Figure 11 Overview of MCS of compounds in the largest cluster.

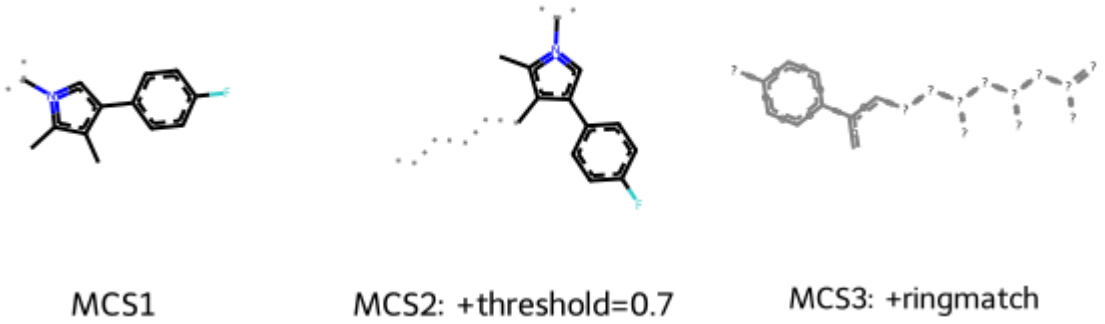

Figure 12 Overview of MCS of all compounds in the dataset with an pIC50 value >9.

Machine learning

We decided to run the machine learning models based on the material that is described in the module that was supplied with the course, which can be found at https://github.com/jesperswillem/CBR_teaching/blob/main/02-MachineLearning.ipynb. Our own version of this file is available at (https://github.com/Allmetalhotend/ACMDD/blob/main/CBR_teaching/02-MachineLearning.ipynb).
We decided to stick to the random forest regression model as the random forest classifier yielded the best performance. For this we decided to use the morgan3 fingerprint, which was the best performing overall fingerprint except for the support vector machine (SVM) learning method. It is possibly the case that the SVM has a hard time using high bitsize fingerprints, as it performed relatively better using the MACCS fingerprints.

The following ROC curves and metrics were determined from our classification models and displayed in figure 13 and table 3, respectively. All models were trained using morgan3 fingerprints. The resulting ROC curves show a marked improvement of the large dataset over the small dataset.

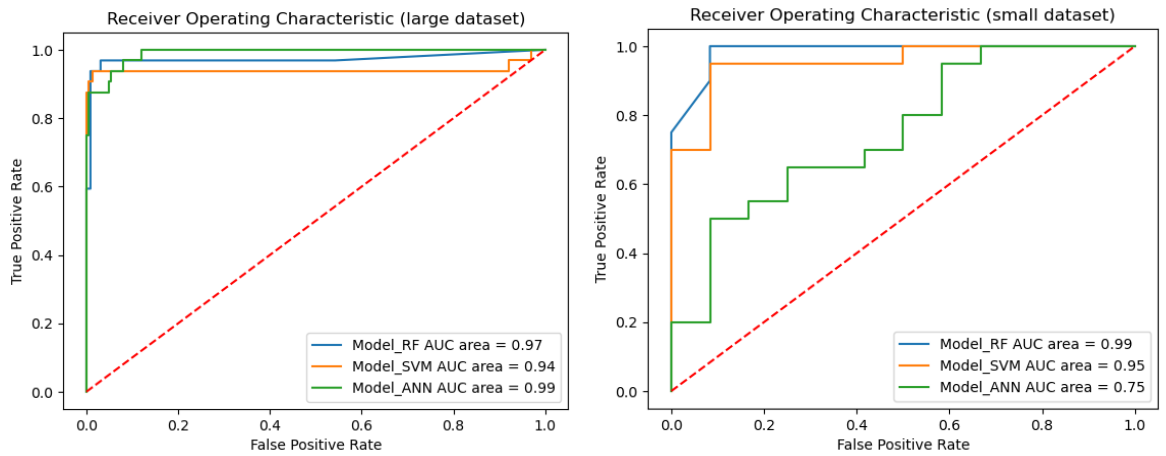

Figure 13 ROC curves of all three classifier models trained on the large and small dataset. Models trained on the large dataset return more favorable results.

From table 3 we can see that the SVM models underperform with a sensitivity of only 0.15 versus 0.80+ scores for both the RF model and the ANN model. between the latter two models the RF model seemed to perform best with an AUC of 0.99.

Table 3 Performance of models trained on the second (large) dataset after cross validation (n=5).

Model	Accuracy (std)	Sensitivity (std)	Specificity (std)	Area Under Curve (std)
RF	0.97 (0.01)	0.87 (0.06)	0.98 (0.01)	0.99 (0.01)
SVM	0.88 (0.02)	0.15 (0.08)	1.00 (0.00)	0.97 (0.03)
ANN	0.97 (0.01)	0.88 (0.05)	0.99 (0.00)	0.98 (0.01)

While the RF and ANN perform almost the same, we preferred the former because of the much reduced runtimes. The RF method and large dataset (as described in Data Management) were used for training our regression model. This yielded the following performance metrics, as can be seen in table 4.

Table 4 Performance of the RF regression model

	Value	Standard deviation (std)
MAE	0.52	0.01
RMSE	0.71	0.06

A MAE value below 0.6 and/or RMSE below 0.6 is considered quite decent. Since this model returns an MAE value of 0.52 and an RMSE value of 0.71, we can conclude that our model performs quite adequately. However, in general for a reliable regression model more data should be fed into the mode, as the dataset that was used is quite small.

We used the regression model to predict the affinities of a series of molecules (see table 5). from this we can get a quick glance on model results. However these results cannot be used to assess model performance, as the molecules here are incorporated in the training sets.

Table 5 Overview of predicted pChEMBL values for some compounds of interest

Name	Function	Predicted pChEMBL value (QSAR)	Predichted pChEMBL value (Docking)	Actual pChEMBL value
3HI / CHEMBL1565	HMG-CoA Inhibitor	7.92	5.08	7.56
LIGAND001	Self-designed ligand	8.40	5.30	-
LIGAND002	Self-designed ligand	5.87	5.28	-
HMG-CoA	Substrate	5.62	4.84	-
Atorvastatin	Approved drug (HMG-CoA inhibitor)	8.63	5.42	8.21

We decided to generate 2 different new ligands (figure 14) and predict their affinity using the random forest regression model. The first ligand (ligand_01/ligand1) was based on the largest cluster that we isolated from the Pchembl data. We decided to replace one position with a -COOH group attached to the ‘2 position in the pyrrole scaffold. The resulting structure performed rather well in the regression, however this is to be expected as the similarity of the cluster in the pChEMBL data is large. A future prospect is to investigate what type of group is preferred on what position of the pyrrole. Therefore more ligands must be generated and eventually tested experimentally. We also tested other groups on the ‘2 position such as an amine group or C=O group, but the -COOH group performed best. as some drugs in the cluster also have a lipophilic group attached to the ‘2 position of the pyrrole which could also be a future subject of investigation.

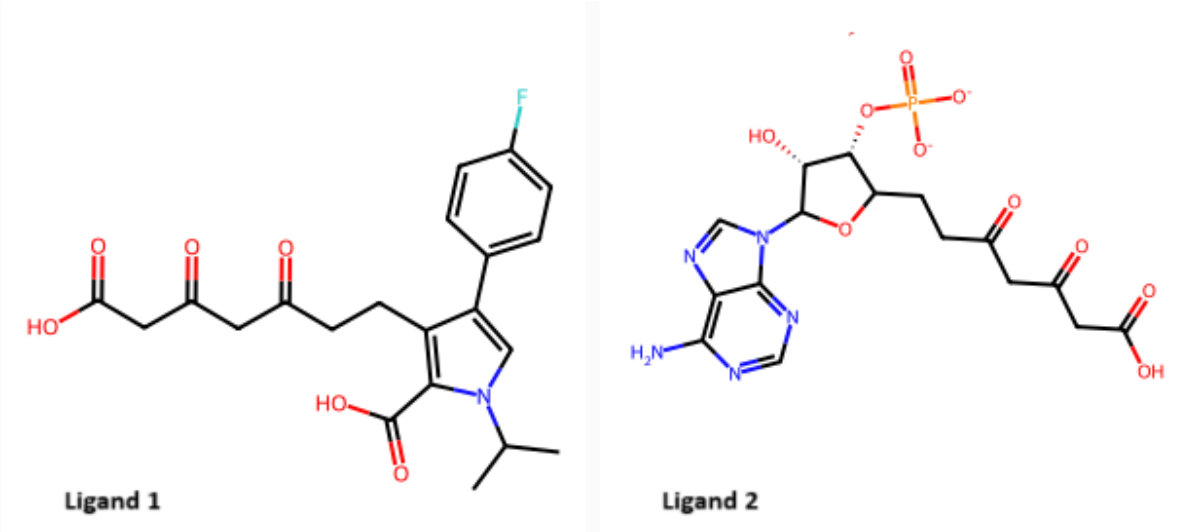

Figure 14: structures of the 2 ligands that were generated. ligand_01 (left) and ligand_02 (right)

The second ligand was based on the substrate of the HMG-Coa reductase enzyme. Here we decided to replace the diphosphate group of HMG-CoA with the ketone containing chain that is found on all ligands found in the largest cluster. This group seems to be important for activity on the receptor, and we believe this group is inserted into the binding pocket where normally the diphosphate group of HMG-Coa binds. As this diphosphate is a large linear group, a similar linear group may be preferred on the ligand to ensure blocking of the catalytic domain of the protein. We also tried to investigate this using docking, but the results of our docking experiments were insufficient (*see section: Docking).*

Docking

For docking we used the code that was made available on https://github.com/jesperswillem/CBR_teaching. Our own version of this file is available on: https://github.com/Allmetalhotend/ACMDD/blob/main/CBR_teaching/03-Docking.ipynb. We started with grabbing the PDB structure 1DQ9 (https://www.rcsb.org/structure/1dq9) and the ligand HMG. We extracted the smiles of the ligand HMG and docked this using autodock vina. We used the same molecule as was in the crystal structures for the initial dock to make a first evaluation of docking performance. this yielded the following structure:

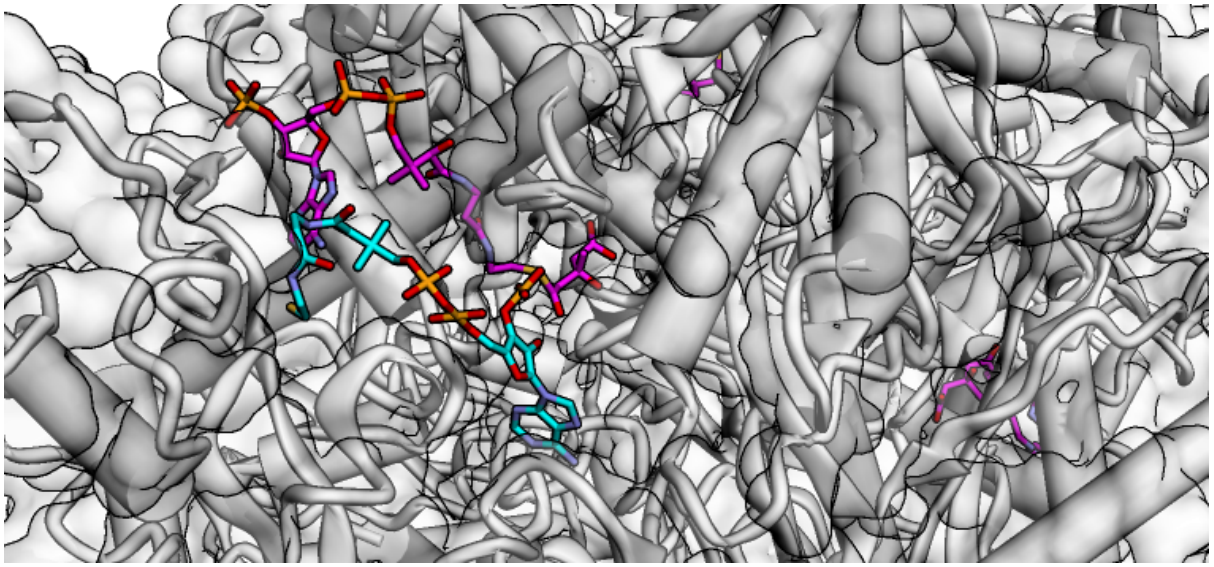

Figure 15: Dock of substrate HMG-COA (cyan) on the PDB structure 1DQ9 with the same molecule of HMG-COA substrate (purple).

Here (figure 15) we can see the HMG molecule that was in the crystal structure in magenta and the docked version of HMG in cyan. Ideally the two molecules would overlap completely suggesting a perfect dock. However this was not the case and the HMG molecule was flipped and bound differently in the binding pocket. However the docking program was able to find the right binding spot so we decided to dock another structure which included a drug structure that was extracted from the pdb (https://www.rcsb.org/ligand/3HI):

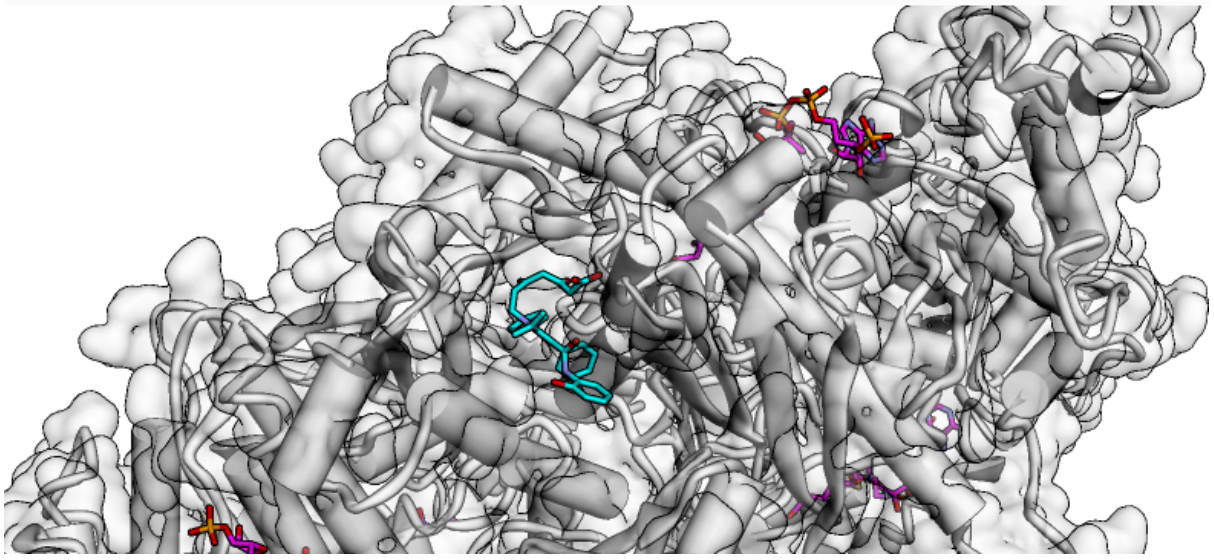

Figure 16: Dock of ligand 3HI (cyan) on the PDB structure 1DQ9 with HMG-COA substrate (purple).

Here Autodock Vina was not able to find the right binding spot suggesting allosteric behaviour or bad docking performance (see figure 16). To investigate this we take a look at another crystal structure with the bound state of the 3HI ligand (https://www.rcsb.org/ligand/3HI). We docked 3HI itself on its own crystal structure. (for the crystal structure with ligand see: https://www.rcsb.org/structure/3cct).

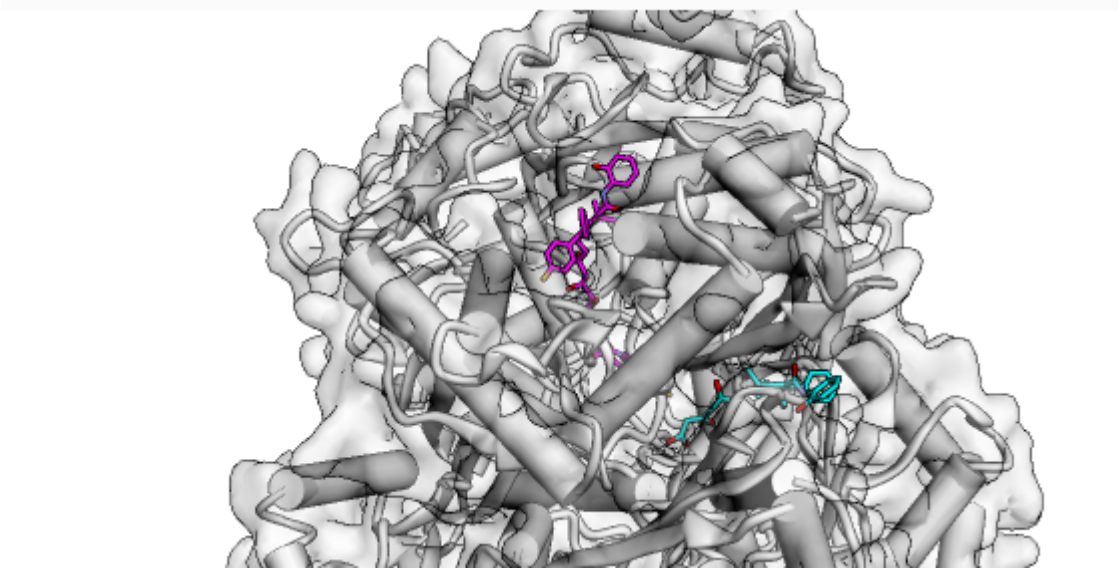

Figure 17: Dock of ligand 3HI (cyan) on the PDB structure 3CCT with 3HI ligand (purple).

Here we see the same trend of the ligand not finding the binding spot demonstrating bad docking performance (figure 17). A future prospect to resolve this issue is to use another docking program to replace Autodock Vina for this subject. However, as no alternatives were accessible we kept on docking using Autodock Vina. Next we looked at another structure, this time of a known drug called Atorvastatin which yielded the same trend as using 3HI. Here (figure 18) we tested atorvastatin on the crystal structure with COA; we see that the atorvastatin still is not able to find the right binding spot:

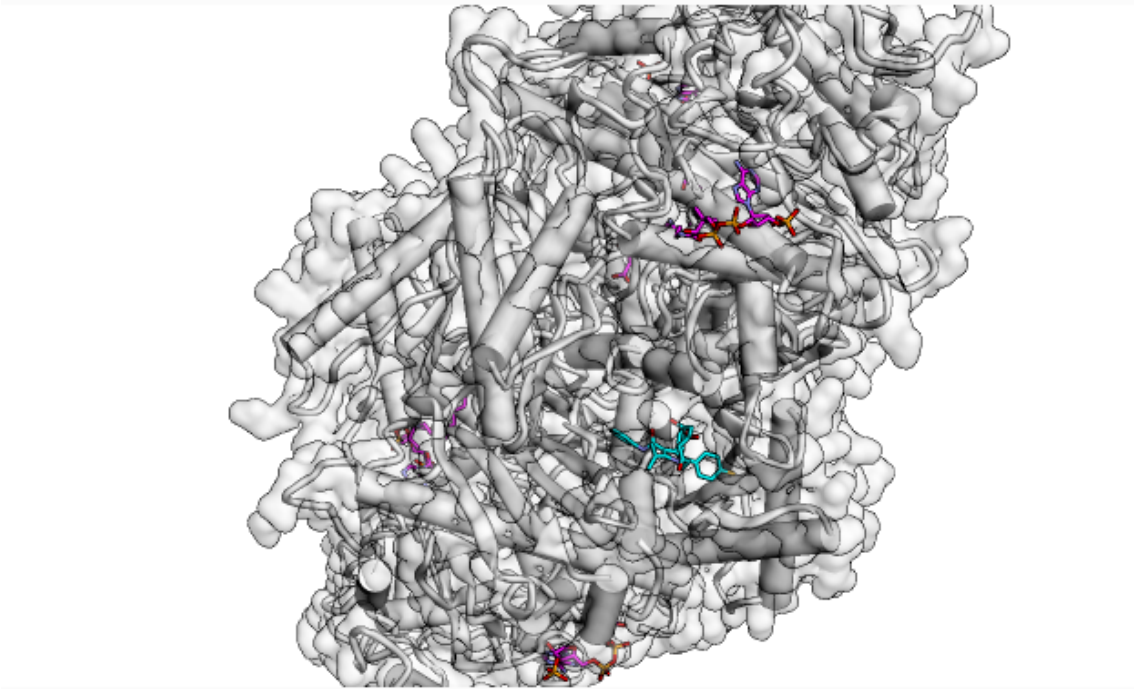

Figure 18: Dock of ligand atorvastatin (cyan) on the PDB structure 1DQ9 with HMG-COA substrate (purple).

As could be concluded the use of autodock Vina for this target/ligands is questionable and is characterized by bad docking performance. interestingly using our own ligand (3-(6-carboxy-3,5-dioxohexyl)-4-(4-fluorophenyl)-1-isopropyl-1H-pyrrole-2-carboxylic acid) (figure 19, cyan) we see a resemblance in binding localisation:

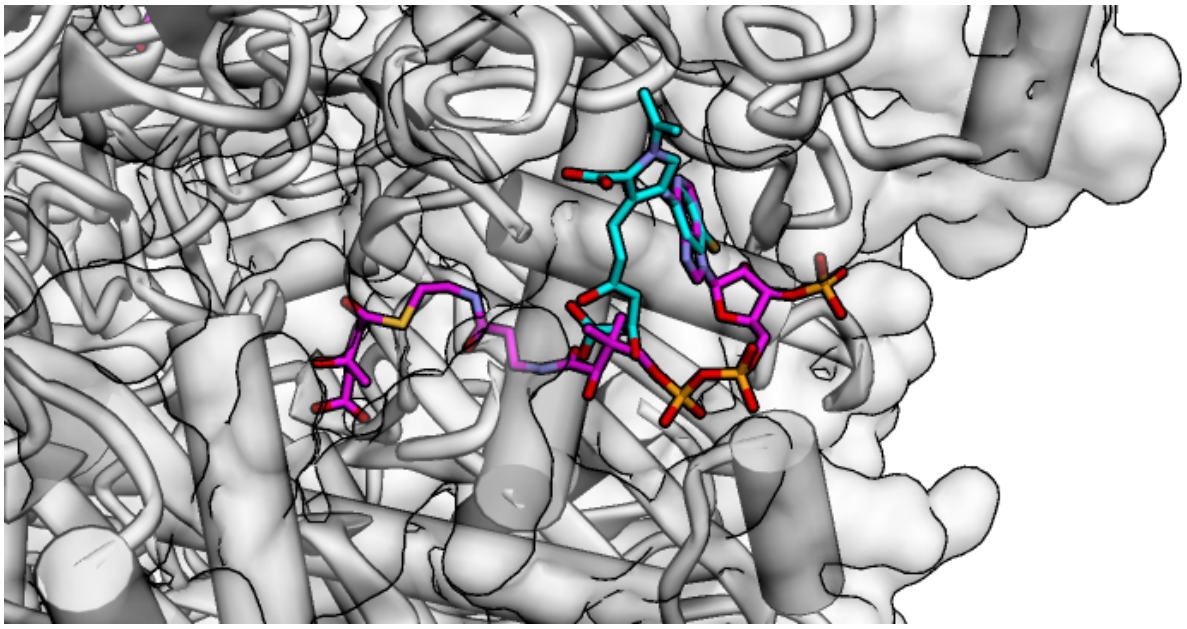

Figure 19: Dock of ligand Ligand_01 (cyan) on the PDB structure 1DQ9 with HMG-COA substrate (purple).

We see here overlap between the substrate COA and our own ligand (1). We observed the same with our other ligand (figure 20, cyan):

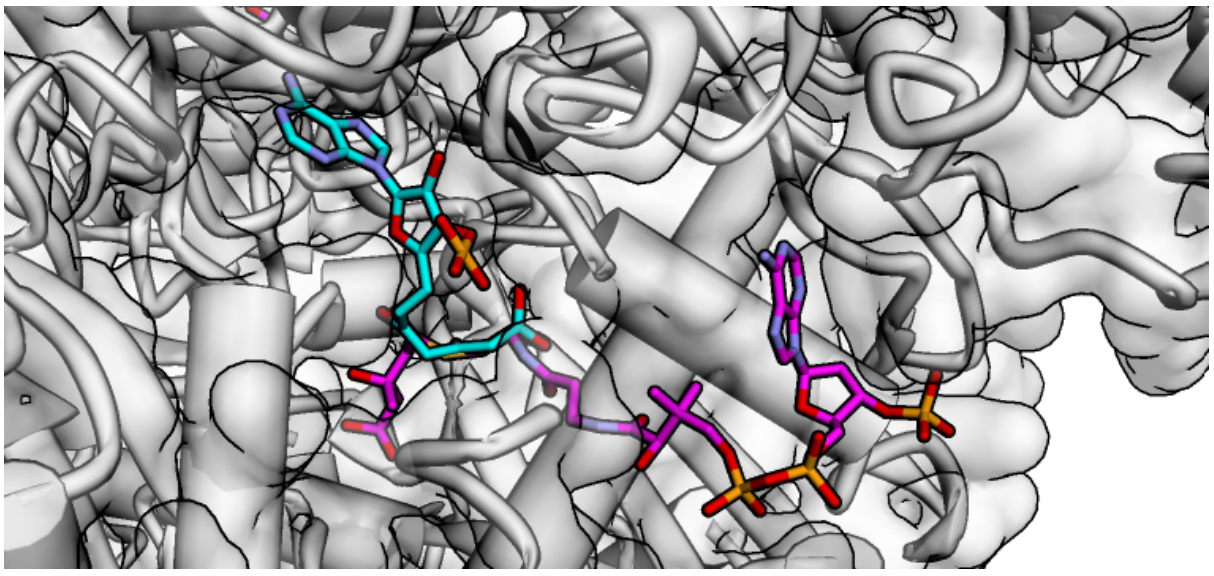

Figure 20: Dock of ligand ligand_02 (cyan) on the PDB structure 1DQ9 with HMG-COA substrate (purple).

If we add up all results from the docking we can conclude that the results are disappointing and not suitable for drawing conclusions. To resolve this issue other docking software could be tried. Another option to try first could be to also visualize other docking poses using autodock Vina, as we only visualized one docking pose.

References

1. Noncommunicable disease and Health promotion (NHP) unit. Indicator Metatdata Registery List. Global Health Observatory.

2. Tsao CW, Aday AW, Almarzooq ZI, et al. Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association. *Circulation*. 2023;147(8):e93-e621. doi:10.1161/CIR.0000000000001123

3. Goldstein JL, Brown MS. *Regulation of the Mevalonate Pathway*.; 1990.

4. WHO Collaborating Centre for Drug Statistics Methodology. ATC classification index with DDDs.

5. Istvan ES, Deisenhofer J. Structural Mechanism for Statin Inhibition of HMG-CoA Reductase. *Science (1979)*. 2001;292(5519):1160-1164. doi:10.1126/science.1059344

6. Paysan-Lafosse T, Blum M, Chuguransky S, et al. InterPro in 2022. *Nucleic Acids Res*. 2023;51(D1):D418-D427. doi:10.1093/nar/gkac993

7. Brown MS, Radhakrishnan A, Goldstein JL. Retrospective on Cholesterol Homeostasis: The Central Role of Scap. *Annu Rev Biochem*. 2018;87:783-807. doi:10.1146/annurev-biochem-062917-011852