

Virtual screening for non-carbohydrate inhibitors for cholera toxin B

Searching for non-carbohydrate Cholera toxin B inhibitors using virtual screening techniques and data analysis

Abstract

Carbohydrate ligands are present in various biological processes which makes them interesting targets in the pharmacological world. Inhibitors can be made to correct biological processes or target carbohydrate-binding pathogens. However, carbohydrates do have multiple drawbacks which makes them hard targets for inhibitor-synthesis. The cholera toxin, especially subunit B, is a carbohydrate binding pathogen. It binds the GM1 pentasaccharide present in epithelial cell membranes. Inhibitors have been made to target this binding and stop the internalisation of the cholera toxin. However, to date these inhibitors mostly consist of complex carbohydrate molecules. This study describes a virtual screening of multiple drug-like molecule databases to find non-carbohydrate inhibitors, which are less complex and have better pharmacological properties. The virtual screening consists of structural screening using HYBRID and similarity screening using ROCS. The results have been combined using reciprocal data-fusion, resulting in a hitlist which contains information of both screening methods and is ranked based on both methods. Multiple interesting molecules have been found, showing similar interactions when compared to the carbohydrate GM1 ligand. From the final hit list molecules can be selected for further in-vitro research.

Layman's summary (Dutch)

Suikers zijn onderdeel van veel biologische processen. Meerdere ziektemakers, waaronder de ziektemaker van de cholera bacterie, binden aan suikers van cellen om zo cellen binnen te kunnen dringen. Meerdere inhibitie-medicijnen proberen deze interacties tegen te gaan, door te binden aan de cholera-toxine voordat deze aan de cel kan binden. De meeste van deze inhibitors zijn echter complexe suikers en lastig te maken. Hierdoor is er in een grote database met niet-suiker moleculen gezocht naar alternatieven met gebruik van meerdere computer programma's die deze moleculen (relatief) snel kunnen screen op basis van structuur en overeenkomsten met bestaande inhibitie-medicijnen. Meerdere moleculen zijn gevonden die gunstige interacties laten zien met de cholera toxine, die vergelijkbaar zijn met de cel-suikermoleculen. Moleculen die goed uit deze computer-screening komen kunnen in de toekomst in het lab worden bekeken op daadwerkelijke binding aan de cholera toxine, met het doel om minder complexe niet-suiker inhibitors te kunnen vinden die makkelijker gemaakt kunnen worden.

CONTENTS

1. Introduction	- 5 -
1.1 Carbohydrate ligands	- 5 -
1.2 Cholera toxin	- 7 -
1.3 Carbohydrate cholera ligands.....	- 10 -
1.4 Cholera ligands interactions	- 15 -
1.5 Drawbacks of non-carbohydrate ligands	- 16 -
1.6 Non-carbohydrate cholera ligands.....	- 16 -
1.7 Virtual screening for non-carbohydrates and research setup	- 17 -
2. Materials & methods	- 19 -
2.1 Utrecht HPC	- 19 -
2.2 Target crystal structures	- 19 -
2.3 OpenEye SPRUCE	- 23 -
2.4 Superposition - Protein RMSD study.....	- 24 -
2.5 Databases	- 25 -
2.6 Database preparation - OpenEye FILTER.....	- 26 -
2.7 Database preparation - OpenEye Flipper	- 27 -
2.8 Database preparation - OpenEye Tautomers.....	- 27 -
2.9 Database preparation - OpenEye OMEGA.....	- 28 -
2.10 Active compounds	- 29 -
2.11 OpenEye ROCS – similarity search	- 30 -
2.12 OpenEye HYBRID – ligand guided docking	- 33 -
2.13 Pre-docking actives study.....	- 34 -
2.14 Datafusion - Reciprocal Rank Fusion	- 35 -
2.15 Clustering results - Butina clustering	- 36 -
2.16 Visualization of hit molecules	- 38 -
3. Results.....	- 40 -
3.1 Receptor Studies.....	- 40 -
3.1 HYBRID.....	- 40 -
3.2 ROCS.....	- 43 -
3.3 Reciprocal Sum Rank calculations	- 46 -
3.4 Butina clustering results	- 50 -
3.5 Hit molecule visualization	- 53 -
4. Conclusion	- 59 -
5. Discussion.....	- 60 -
6. Future research recommendations	- 62 -

7. References.....	- 63 -
8. Supplementary.....	- 68 -

1. INTRODUCTION

1.1 Carbohydrate ligands

Carbohydrate ligand molecules form a large group of biological ligands in the human body. One important group of proteins that is connected to carbohydrate ligand binding is the lectin protein family (Sethi et al., 2021). The lectin protein family is important for many physiological and cellular processes. Selectins for example are important adhesion lectins, which are crucial in the immune system. Galectins are associated with many cell regulation processes and also cancer and collectins are linked to pattern recognition of respiratory viruses and pathogenic bacteria (Ghazarian et al., 2011).

Many diseases and pathogens are linked to carbohydrates. *Escherichia coli* binds to host mannoses and Influenza A Virus binds to sialic acids. Other pathogens that interact with carbohydrates for example are *Neisseria gonorrhoeae*, *Streptococcus pneumoniae* and *Pseudomonas aeruginosa* (Ghazarian et al., 2011).

Due to the many interactions of carbohydrates with many processes, proposed carbohydrate-based therapeutics and drugs are very common. Using glycosylation-modification these drugs can target different carbohydrate-binding processes. For example, anti-adhesion therapy is used for blocking pathogen interactions with the cell by binding the pathogen lectins and therefore blocking interactions with human glycoproteins (Fig 1) (Ghazarian et al., 2011).

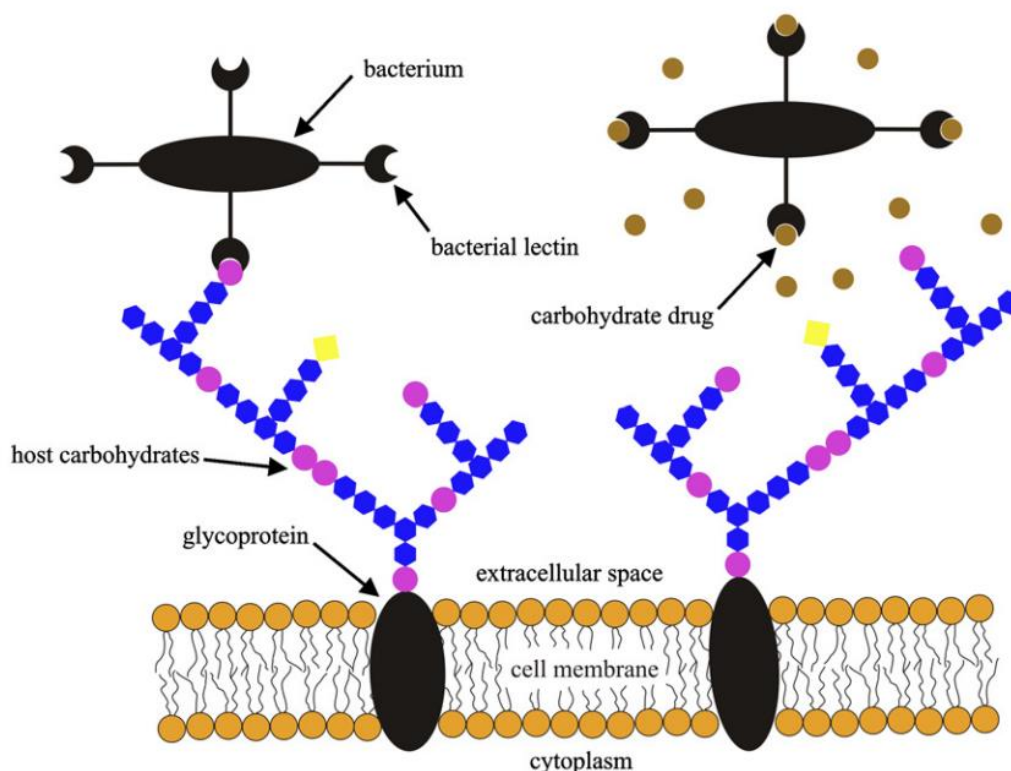


Figure 1 - Schematic illustration of bacterium binding a surface glycan (left). Binding can be stopped by carbohydrate drugs, called anti-adhesive therapy (right) (Ghazarian et al., 2011).

Carbohydrates are also used for carbohydrate-based vaccines, lectin-based drug targeting for anti-cancer therapy and overall drug targeting (Ghazarian et al., 2011).

1.2 Cholera toxin

Another pathogen which is connected to carbohydrate binding is the cholera toxin. The cholera toxin is made by the *Vibrio cholerae* bacterium. *Vibrio cholerae* is a Gram-negative bacterium, which produces the cholera toxin, the pathogen of the cholera disease. Cholera is an acute secretory diarrheal disease, which without correct therapy has a mortality of 50%. The cholera toxin causes these symptoms by increasing water excretion in the small intestine. (Haan & Hirst, 2004; Harris et al., 2012)

The cholera toxin is a so-called AB₅-toxin, which means it has an enzymatically active A subunit and a receptor-binding B subunit. Cholera toxin consists of an A subunit and five B subunits, which makes it a hexamer protein (Haan & Hirst, 2004) (Fig 2).

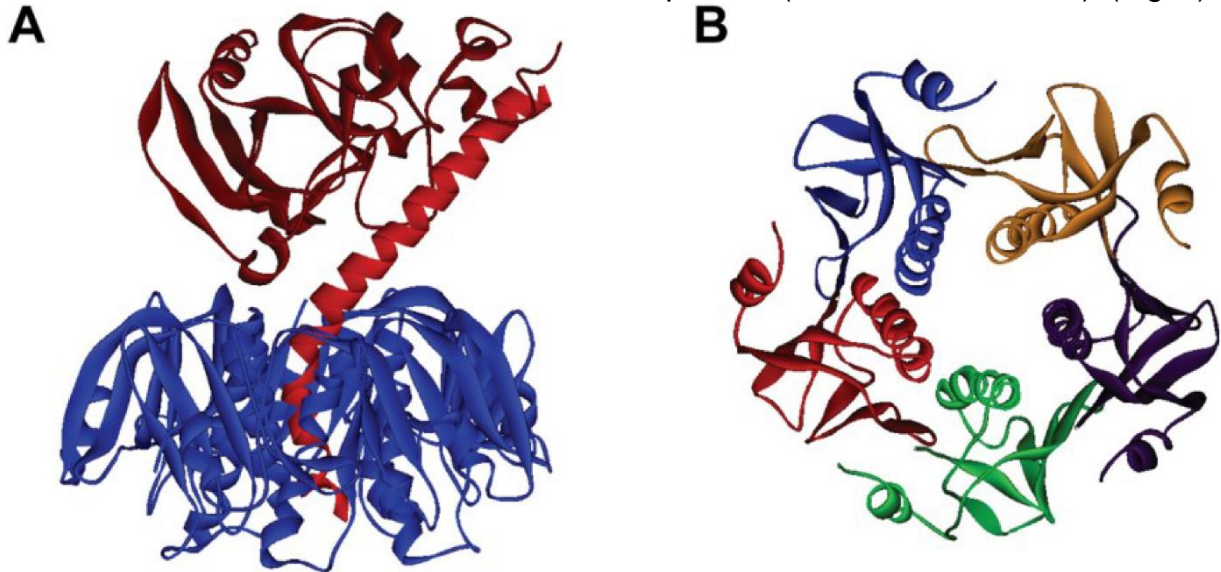


Figure 2 - Schematic representation of the cholera toxin. A shows the A subunit in red and the B subunit in blue. The dark red part is the A1 domain, which is the enzymatic domain. The light red A2 domain connects the A subunit to the B subunit. B shows the B subunit from the top, where 5 distinct monomers are visible that together form the B subunit pentamer (Baldauf et al., 2015).

The B subunits are lectins that bind to apical membrane receptors. The target it binds to is the GM1 ganglioside (Fig 3), which is a carbohydrate pentasaccharide connected to a ceramide unit. It is present in almost all plasma membranes, however the cholera toxin mostly binds to epithelial cells in the small intestines. All monomers of the pentamer have a GM1 binding pocket and interact mostly with the galactose of the GM1 ganglioside, although interaction is also found with the sialic acid and N-acetyl galactosamine. It binds to GM1 with a K_D of 7.3×10^{-10} M. This is the K_D when taking a single-site and a single receptor into account. When looking at a multi-step binding model, that takes receptor cross-linking into account and therefore multiple sites and receptors, the K_D was determined to be 56 nM (Haan & Hirst, 2004; Lauer et al., 2002).

Although lower K_D 's (40 nM) have been found since then (Kumar & Turnbull, 2018; Turnbull et al., 2004). (Haan & Hirst, 2004)

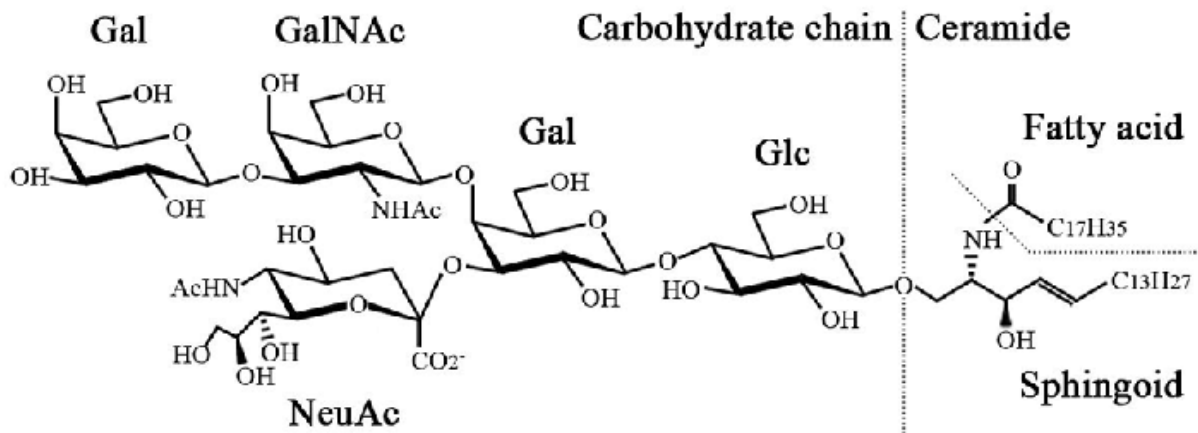


Figure 3 - GM1 ganglioside (Yu et al., 2011).

After binding to GM1 the complex gets stabilized due to hydrogen-bond interactions and interactions with water molecules. After this the A subunit will get translocated intracellularly (Fig 4). The A subunit consists of two domains, the A1 domain and the A2 domain. The A1 domain is responsible for enzymatic activities, while the A2 domain links the A subunit to the B subunit. When the A subunit is in the cytosol, the A1 domain acts as a ADP-ribosyltransferase and NAD-glycohydrolase (Gill & Coburn, 1987; Haan & Hirst, 2004; Moss et al., 1977, 1981; Moss & Richardson, 1978). This leads to ADP-ribosylation of G proteins, which impacts multiple metabolic pathways. The reason for the diarrheal symptoms is due to the involvement of G proteins in activating the adenylate cyclase complex which leads to higher cAMP levels (Chen et al., 2002). The higher cAMP levels lead to a activation of protein kinase A, which over-activates CFTR chloride channels by phosphorylation (Burch et al., 1988; Cheng et al., 1991; Denning et al., 1992; Field et al., 1989a, 1989b; Picciotto et al., 1992). This means that more chloride will be secreted than usual, which will be followed by excretion of water leading to diarrhoea. (Haan & Hirst, 2004)

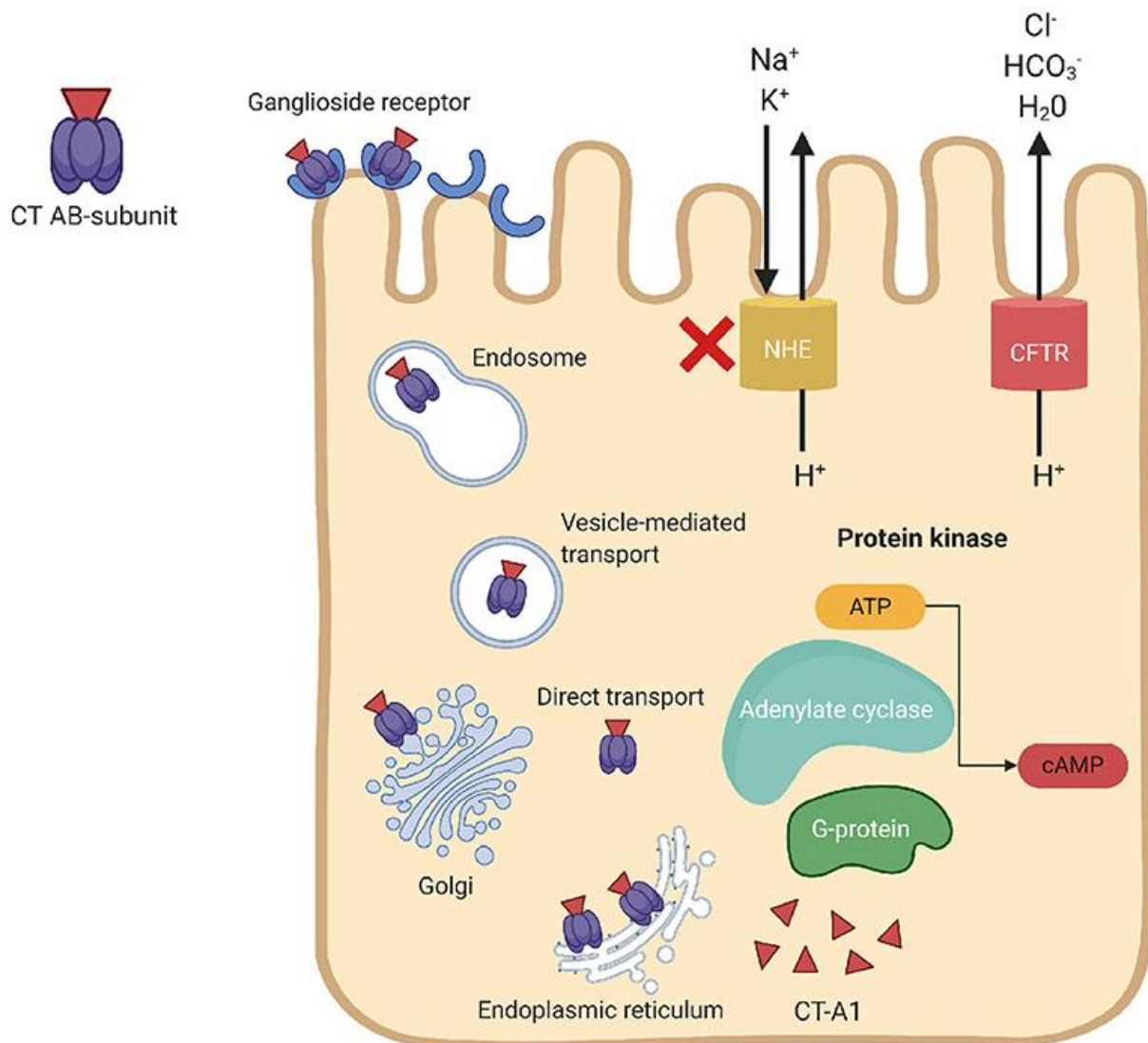


Figure 4 - Schematic representation of the Cholera Toxin Pathway (Ramamurthy et al., 2020).

To target the cholera toxin, inhibitors have been made. These inhibitors are anti-adhesion inhibitors and target the binding of the cholera toxin B subunit to the GM1 gangliosides, thereby neutralizing the toxin in its whole (Kumar & Turnbull, 2018).

1.3 Carbohydrate cholera ligands

As the natural ligand of the B subunit was well-studied and known, the first iteration of subunit B inhibitors were based on GM1 ganglioside, so-called GM1-mimics. Most of these target the galactose and sialic acid of the inhibitors and bring them into a similar orientation as they are in GM1. Bernardi *et al.* created a GM1-mimic with binding affinity equal to the GM1-oligosaccharide (Fig 5) (Bernardi *et al.*, 1999). However, this molecule still had the same issues as the GM1 pentasaccharide itself, it is fairly complex and therefore exhaustive and expensive to synthesize for actual inhibitor use.

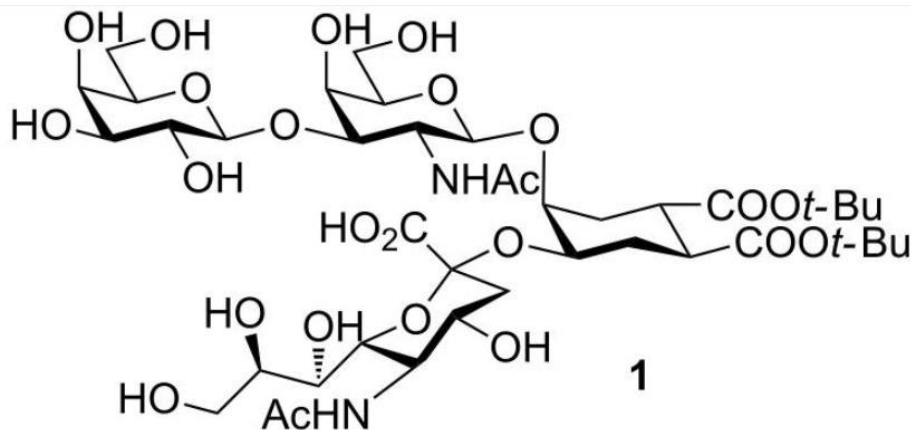


Figure 5 - GM1-mimic designed by Bernardi *et al.* This mimic showed equal affinity to CTB as GM1 (Bernardi *et al.*, 1999).

Bernardi *et al.* tried to tackle this problem by replacing the α -Neu5Ac with alpha-hydroxy acids (Bernardi *et al.*, 2000, 2002). These second generation GM1-mimics (Fig 6) by Bernardi *et al.* showed a K_D of 190 μ M, which is a clear drop compared to the 40 nM of GM1-os and the first generation GM-1 mimic (Bernardi *et al.*, 2000, 2002).

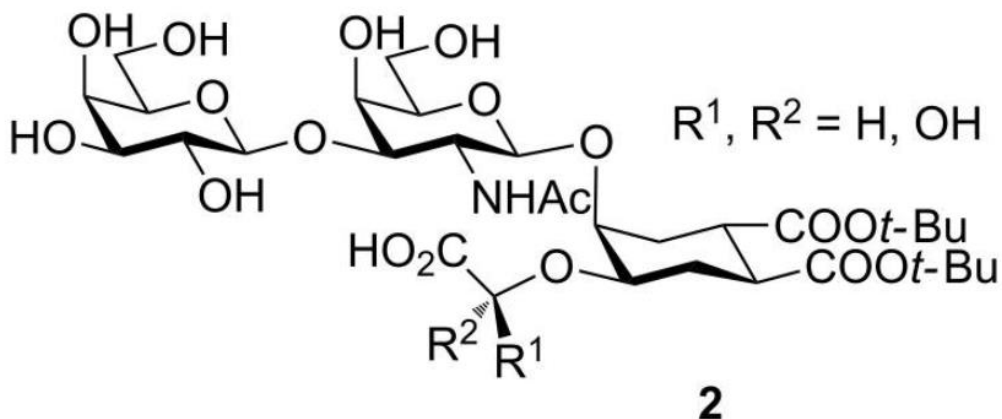


Figure 6 - Second generation GM1 mimic from Bernardi *et al.* (Bernardi *et al.*, 2000, 2002)

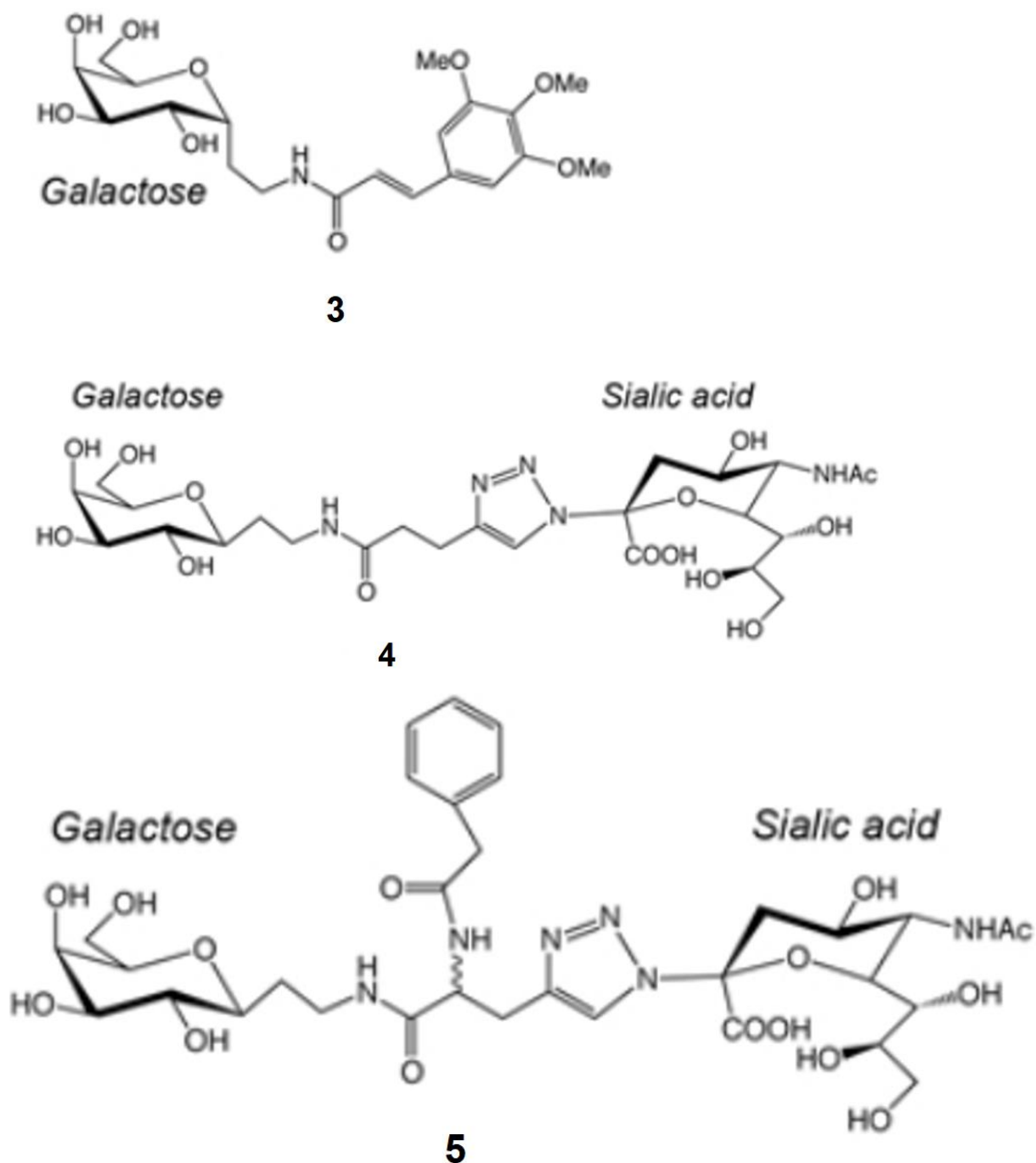


Figure 7 - More GM1-mimics by Bernardi et al. (Heggelund et al., 2017)

Another generation GM1 mimics has been made by Bernardi *et al.* (Cheshev et al., 2010; Heggelund et al., 2017; Podlipnik et al., 2007) (Fig 7). These were aimed to be less complex and focus on the galactose binding site, together with the sialic acid. All three mimics had a binding strength in the upper μM range.

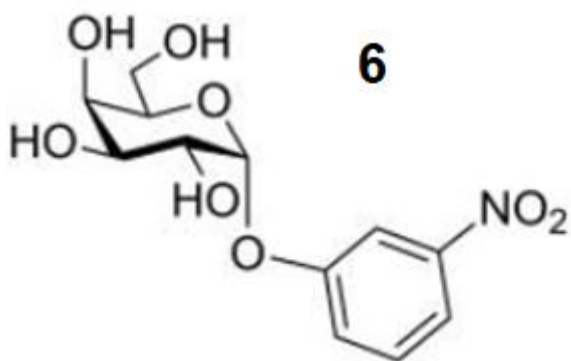


Figure 8 - MNPG ligand (Minke et al., 1999)

Another class of cholera toxin B inhibitors was found by screening for galactose derivatives by Hol & Verlinde *et al.*. They found *m*-nitrophenyl α -D-galactoside (MNPG) as the best binder in this screening, which bound 100 times better than galactose (Kumar & Turnbull, 2018; Merritt et al., 1997; Minke et al., 1999) (Fig 8). MNPG showed an IC_{50} 0.6 mM and 0.72 mM in LT ELISA and CT assays respectively, while galactose showed an IC_{50} of 58mM and 45mM. However, this is still relatively low when comparing to the GM1-oligosachharide which showed IC_{50} 's of 1×10^{-5} mM and 14×10^{-6} . Still MNPG was used as stepping stones to other ligands, due to its simplicity and better binding affinity than galactose.

MNPG derivatives were made by Mitchell *et al.*. Here the MNPG molecule found by Hol & Verlinde *et al.* was elongated, resulting in better affinity than the original MNPG (Fig 10) (Mitchell et al., 2004).

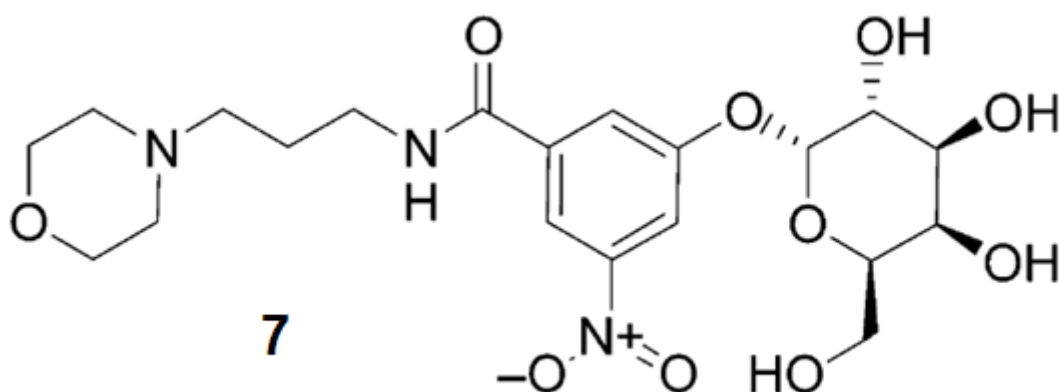


Figure 9 - Elongated version of MNPG (Mitchell et al., 2004).

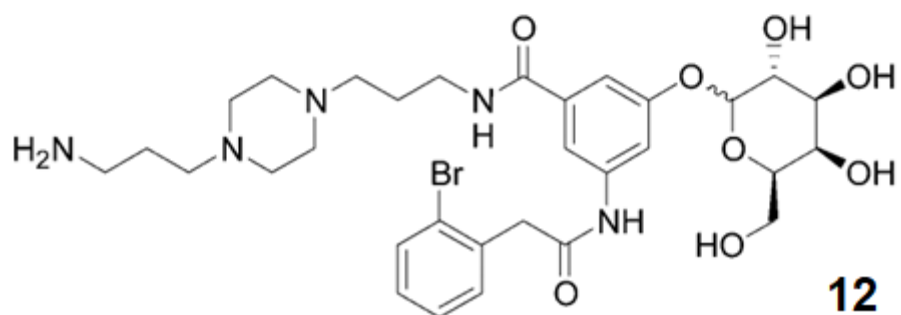
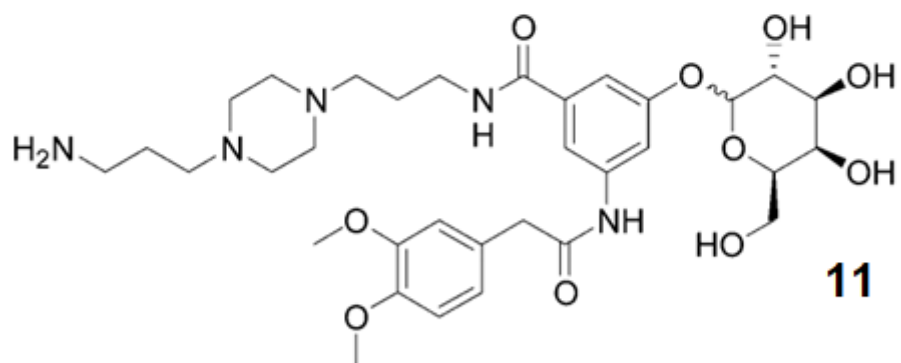
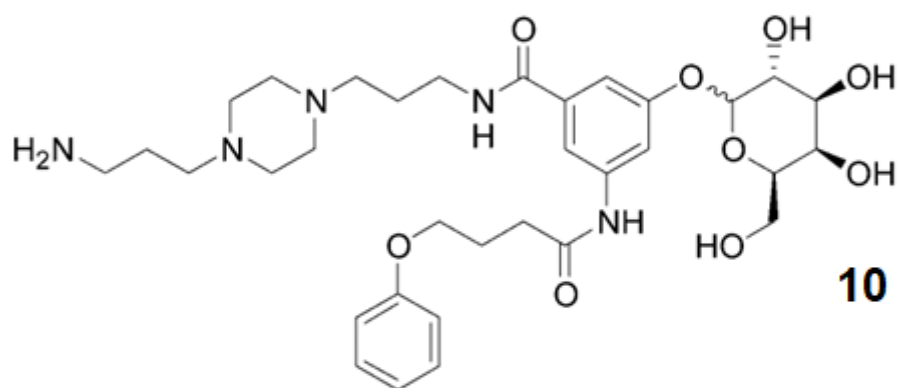
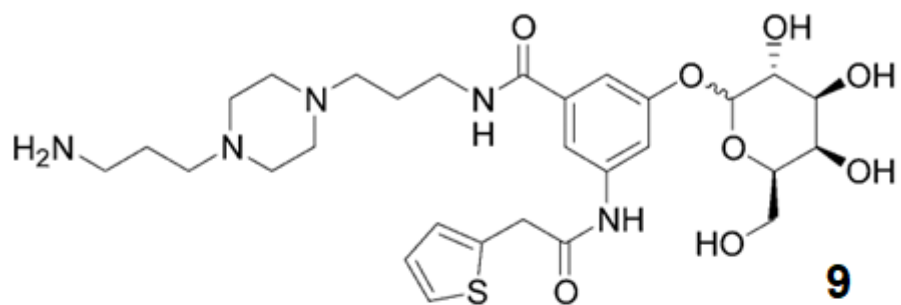
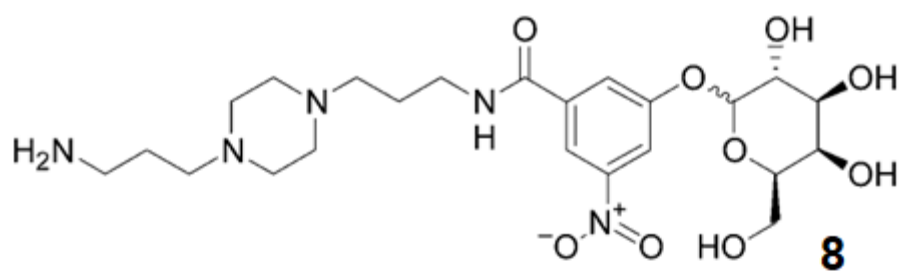


Figure 10 - Multiple MNPG variants made by Mitchell et al (Mitchell et al., 2004).

In this study GD_{1b} Direct Enzyme Linked Assay (DELA) was used. This assay resulted in an IC₅₀ of 1-1.4mM for MNPG. The base elongated MNPG was altered by adding other subgroups (Fig 10). All molecules tested (Fig 10) showed an increase in affinity. Molecule **8** (Fig 10), which still has the original nitro group, had an IC₅₀ of 0.32 mM. Only one molecule was made where replacing the nitro group resulted in a better affinity. Compound **9**, where the nitro group is substituted/elongated with a 2-thiophene acetyl group. This increased the affinity from 0.32mM (IC₅₀) to 0.20 mM (IC₅₀). Compounds **10** and **11** showed a small decrease of affinity compared to **5**, with IC₅₀'s of 0.43 mM and 0.35 mM respectively. **12** showed even a bigger decrease in affinity with an IC₅₀ of 0.79 mM. It is clear that replacing the nitro group in most instances will lead to a decrease, although some improvements are possible. (Kumar & Turnbull, 2018; Mitchell et al., 2004)

Lactose also has affinity for GM1 with a comparable IC₅₀ to galactose. Vrasidas *et al.* created a lactose-2-aminothiazoline that bound cholera fairly well with a K_D value of 23 μM and thereby is one of the better monovalent and relatively simple carbohydrate inhibitors of the cholera toxin B subunit. (Kumar & Turnbull, 2018; Vrasidas et al., 2002)

1.4 Cholera ligands interactions

But how do these inhibitors and in particular GM1 bind to cholera toxin B? GM1-os has been crystalized and studied in different studies and many interactions actually lead to the good binding of GM1-os to the cholera toxin B binding pocket (Fig 11) (Turnbull et al., 2004). It is a bivalent binding, meaning it goes into the protein with a so-called “two-fingered grip”. This bivalent binding is led by the galactose-GalNac and sialic acid groups present in the GM1-oligosaccharide. In terms of binding area, Galactose, GalNac and the sialic acid contribute 39%, 17% and 43% respectively of the total intermolecular contacts. (Turnbull et al., 2004)

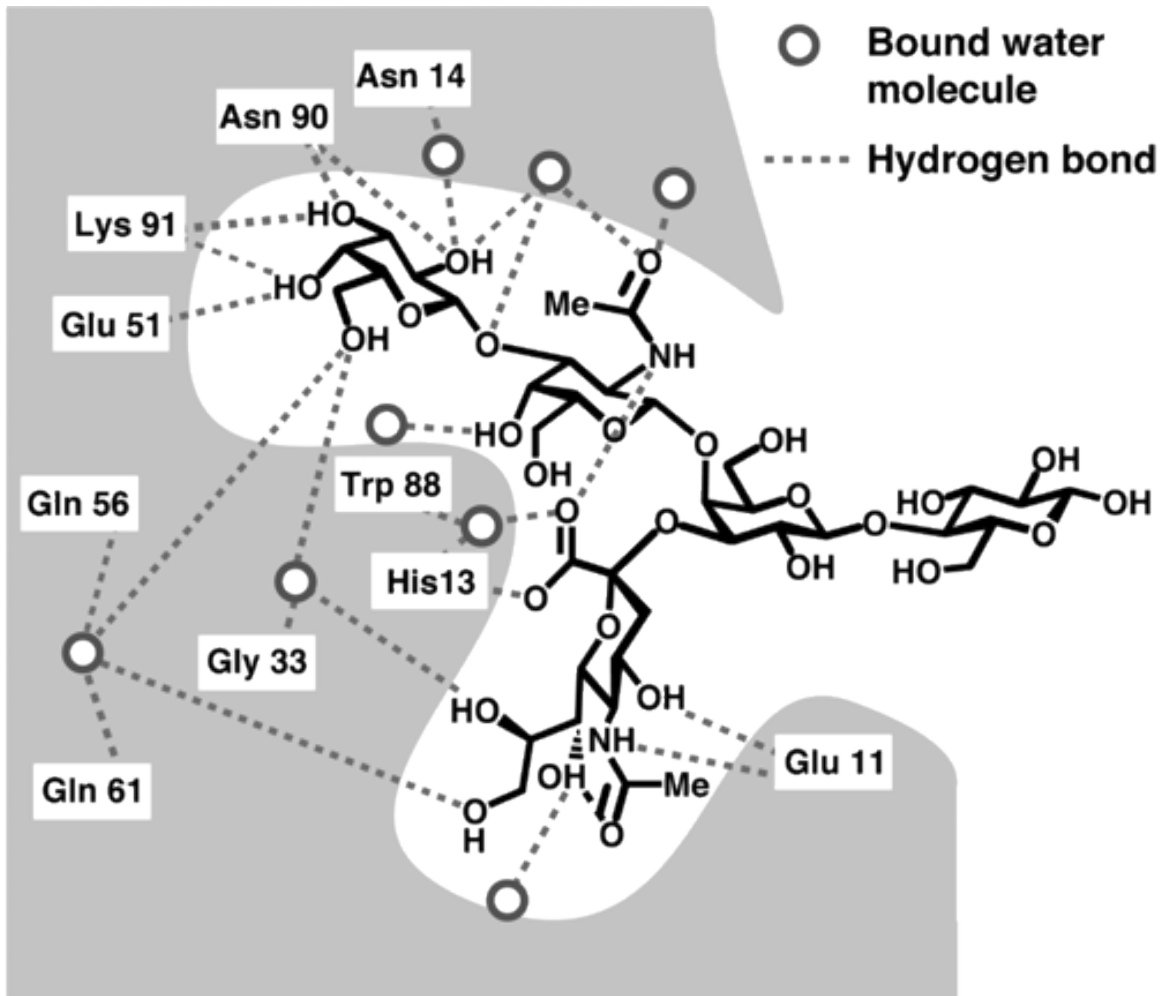


Figure 11 - Schematic representation of GM1 interactions in the Cholera Toxin Subunit B pocket (Turnbull et al., 2004).

Due to the nature of carbohydrates and the many OH groups, a lot of hydrogen bonds are present between cholera toxin amino acids and the GM1-os ligand. Also the binding is stabilized due to multiple interactions with bound water molecules. As most ligands focus on the galactose pocket, the direct hydrogen bonds with glutamic acid-51, asparagine 90 and lysine 91 can be seen as the galactose pocket defining bonds. It was found that indeed galactose is the most important fragment of the GM1-os binding when looking at Free Energies, where galactose had the biggest part of the total Free Energy created by GM1-os. And although sialic acid also has a big part in this, it is mostly a stabilizing binding fragment as it has not shown high affinity on its own when tested. This confirms that galactose forms the more important pocket and

therefore binding fragment of most interest and the target to find replacements for. (Turnbull et al., 2004)

1.5 Drawbacks of non-carbohydrate ligands

However, the use of carbohydrates has several drawbacks which makes them hard to use in several circumstances. First of all they are very complex to make, which makes it expensive and laborious to construct them. Besides that they have multiple unwanted pharmacokinetic properties, like short half-lives and bad bioavailability due to metabolic degradation and bad permeability. This has led to multiple studies to replace carbohydrate-based mimic ligands with non-carbohydrate ligands, which are more pharmacokinetic appropriate and easier to synthesize. An example are glycomimetics (Fig 12). These so called glycomimetic ligands mimic the structure and function of carbohydrates, while improving properties compared to their carbohydrate counterparts and enhancing selectivity and affinity. (Hevey, 2019)

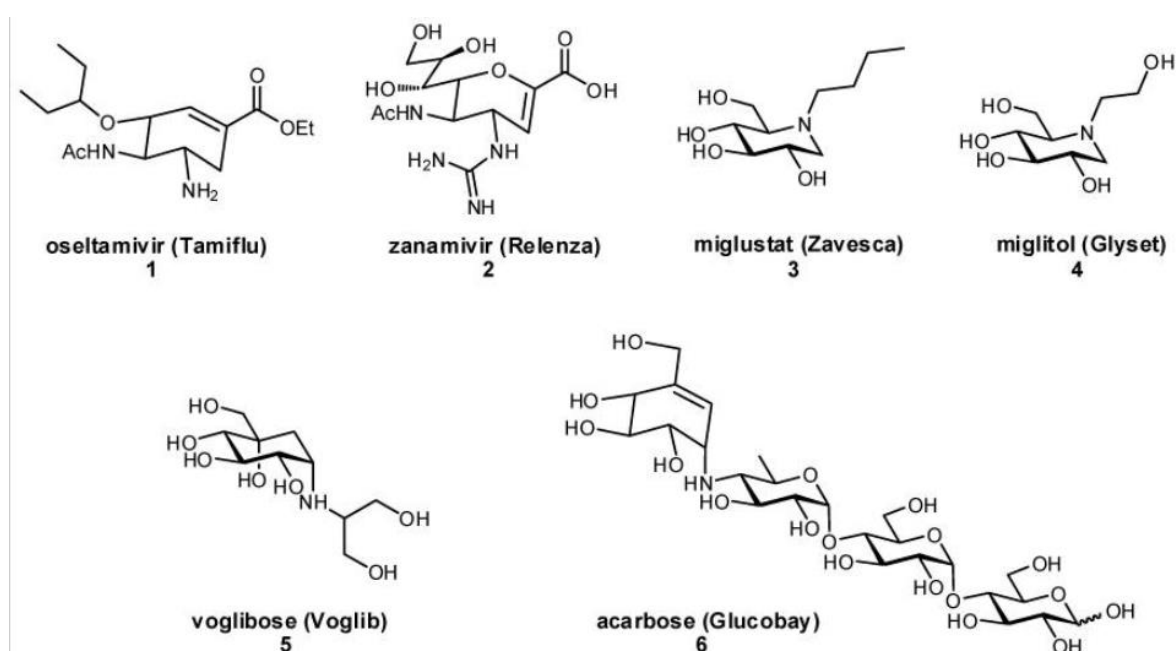


Figure 12 - Glycomimetic inhibitors that are actually used (Hevey, 2019).

For lectins, research has been done resulting in many possible candidates mainly containing heterocyclic inhibitors and peptide-based inhibitors (Sethi et al., 2021).

It is clear that the search for non-carbohydrate ligands is an important, albeit difficult subject. In this research we targeted the cholera toxin, where no real non-carbohydrate ligands have been found yet.

1.6 Non-carbohydrate cholera ligands

There have been some studies to find non-carbohydrate inhibitors for cholera toxin. Robina et al made non-hydrolyzable S-galactosides and non-carbohydrate ligands based on polyhydroxyalkylfluorate. Using weak affinity chromatography the structures were analysed. The molecule with the highest affinity in this study showed a K_D of 0.2 mM, which was 5.5 times higher than MNPG. Although, all other molecules in this study did not represent very good binders. All in all, research to non-carbohydrate

monovalent cholera inhibitors is still scarce and mostly still GM1-mimics or carbohydrate containing MNPG derivatives. (Kumar & Turnbull, 2018; Ramos-Soriano et al., 2013)

1.7 Virtual screening for non-carbohydrates and research setup

For finding non-carbohydrates however, our target was to find a small drug-like (and therefore good pharmacokinetic properties) molecule, which binds in this important galactose pocket. This was done using a big drug-containing screening database, two different virtual screening tools and data fusion to get to a combined data-driven conclusion. Virtual screening is using computational methods to analyse large databases containing molecules of interest and identify hits. In this context we were trying to find non-carbohydrate inhibitors in a large self-combined database of drug-like and bioactive compounds. (Lavecchia & Di Giovanni, 2013; Oliveira et al., 2023)

Virtual screening can be categorized in two approaches: Structure-based virtual screening methods and ligand-based methods. With structure-based methods we are talking about methods that use the structure of interest, the protein, and interactions with this structure to find suitable hits. Ligand-based methods however use only the ligand and no structure, hits are purely found by ligand information. (Lavecchia & Di Giovanni, 2013; Oliveira et al., 2023)

One common method for structure-based virtual screening is docking, where molecules get docked into a target protein receptor. The goal is to score database molecules based on how well they bind to the target receptor, searching for possible ligand or inhibitor candidates. The docking tool used in this study is the OpenEye HYBRID docking tool, which uses ligand information to boost performance. It does this by using the Chemical Gaussian Overlay (CGO). Instead of just docking the molecule, it first matches the 3D arrangement of shape and chemical features of the bound ligand and database ligand before docking. This means it will dock the database molecules in a similar pose to the already bound ligand. It was found that this enhances performance when compared to their “old-school” docking tool FRED, which does not take bound-ligand information into account. (McGann, 2012)

For finding a suitable structure for this docking multiple Protein Data Bank (PDB) files containing protein structures and ligands have been considered. After comparing seven PDB files using multiple superpose methods of the OpenEye Superposition tool, the differences between the protein and binding site structures in the PDB's were considered insignificant. This means that the protein can be considered rigid.

Also a self-docking study was done, where the original ligand was docked into its corresponding receptor. The idea was to look if the receptor would dock the ligand in the same pose it was crystalized in. However, this was later omitted from the study as HYBRID is not a pose prediction tool and this is not a correct way of grading receptor performance.

After this a pre-docking study was done using known active compounds, where two PDB's clearly scored better. Both PDB's, 1EEI and 1PZJ, have been used for this research as the ligands were clearly different size and therefore most likely will lead to different hits and hit-poses when using ligand-guided docking.

Besides a structure-based method, also a ligand-based method was used. Rapid Overlay of Chemical Structures (ROCS) is another tool from OpenEye. ROCS is a similarity search method, where database molecules are compared to the input molecule based on 3D shape and 3D arrangement of chemical features. ROCS is a very fast screening tool and was also found to be more consistent and often better than structure-based docking. However ligand-based methods only take the ligand into account and can therefore not predict things like affinity, binding modes, interactions and poses. (Hawkins et al., 2007)

All in all both structure-based and ligand-based methods do have positives and negatives. Therefore after applying both methods, the results of both methods can be combined using data fusion. Data fusion is the act of combining multiple data-sources, here HYBRID and ROCS results, to one explicit and usable data-output. This data-output contains a score which includes information of all input methods.

The data fusion method used is the Reciprocal Rank Sum method. This method uses the rank of hits in both score-sorted HYBRID and ROCS outputs and combines this information resulting in a Reciprocal Rank Sum. This Reciprocal Rank Sum contains information about the rank in both ROCS and HYBRID scoring outputs of the given molecule and therefore combines the information of both methods. Using this output we were able to select molecules that scored very well in one of either methods or scored well in both. The Reciprocal Rank Sum has been used in Virtual Screenings before and is deemed as a quick and good fusion scoring method. (Bajusz et al., 2019; Cormack et al., 2009; Willett, 2013)

After fusing the data, the top 1000 of the fused list was clustered. Molecule clustering is done to dissect the data in different groups of similar molecules. This is usable as in the databases many similar looking molecules were present. Clustering the data lead gave the option to look at scores of different classes of potential hits and also made it easy to omit carbohydrates and unwanted molecules that are for example too complex or had other drawbacks. The clustering method of choice was Butina Clustering. This algorithm is based on Tanimoto similarity using chemical fingerprints. The method is a very quick and unsupervised method to cluster large sets of molecules using similarity exclusion spheres, where each molecule in a cluster has a minimal set similarity to the centroid of the cluster. The main reasons for using Butina are the speed and ease -of use as well as the discrete scoring system. (Butina, 1999; T005 · *Compound Clustering*, z.d.)

At last, using the scores and Butina clustering, molecules could be selected for further visual inspection using VIDA visualization tools. A select number of molecules will be further researched *in vitro* and hopefully lead to usable inhibitors in the future.

2. MATERIALS & METHODS

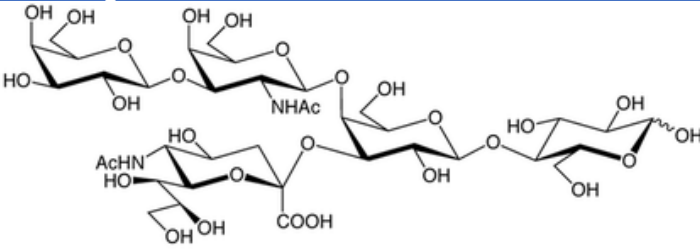
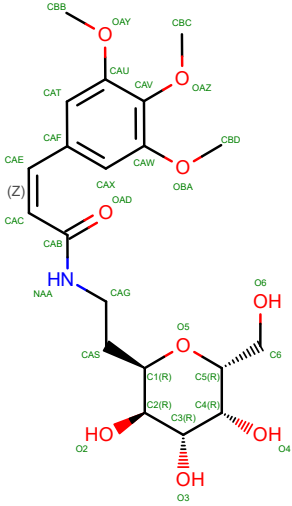
2.1 Utrecht HPC

All work has been done on the Utrecht High Performance Cluster, running either CentOS7 or Rocky8. All calculations have been done using 40-80 cores per run. Runs have been done in chunks to fasten the process and run multiple runs in parallel. Memory was always set at 3GB per CPU. GPU was not used for any of the runs.

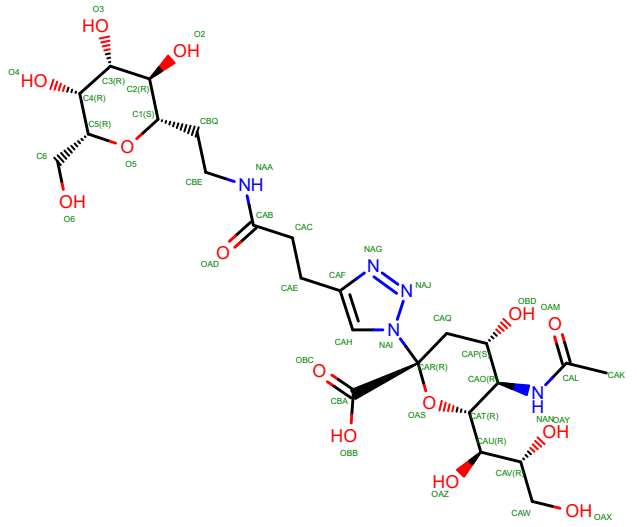
2.2 Target crystal structures

Seven PDB crystal structures containing a monovalent ligand were found in the RCSB PDB (Table 1). The seven PDB's contained the Cholera Toxin B-pentamer from either classic *Vibrio cholerae* or *Vibrio cholerae* O1 biovar El Tor. The classic and El Tor toxins do not have any structural, biological or binding differences at the binding site and therefore could be used interchangeably in this virtual screening as they show comparable binding affinity to GM1 (Heggelund et al., 2017).

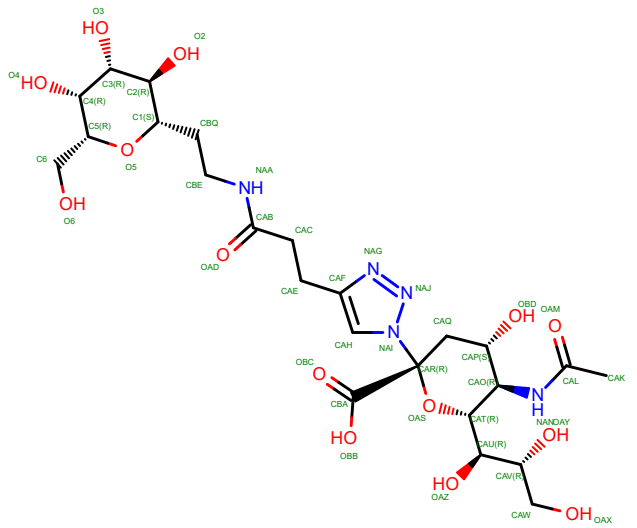
Table 1 - PDB's chosen for the virtual screening

PDB title	Authors	Year	Organism	Bound ligand
3CHB	Merritt, E.A. Hol, W.G.J.	1998	<i>Vibrio cholerae</i>	
5LZJ	Heggelund, J.E. Krengel, U.	2017	<i>Vibrio cholerae</i> O1 biovar El Tor str. N16961	

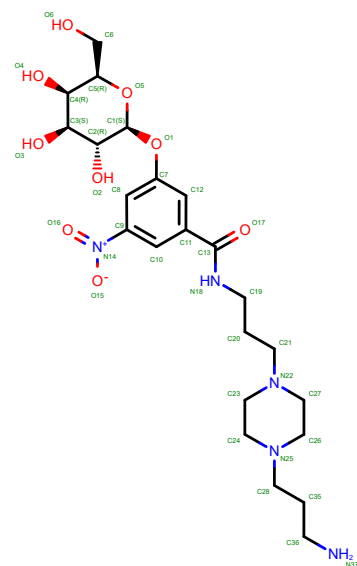
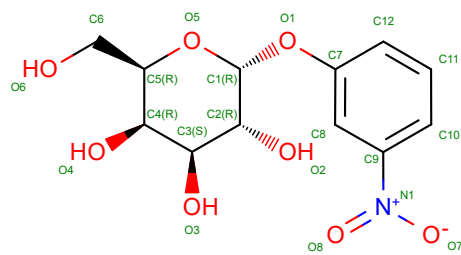
5LZG	Heggelund, J.E. Martinsen, T. Krengel, U.	2017	<i>Vibrio cholerae</i> O1 biovar El Tor str. N16961
------	---	------	--



5LZH	Heggelund, J.E. Martinsen, T. Krengel, U.	2017	<i>Vibrio cholerae</i>
------	---	------	------------------------



1EEI	Merrit, E.A. Hol, W.G.J.	2000	<i>Vibrio cholerae</i>
1PZJ	Mitchell, D.D. Pickens J.C. Korotkov, K. Fan, E. Hol, W.G.J.	2004	<i>Vibrio cholerae</i>

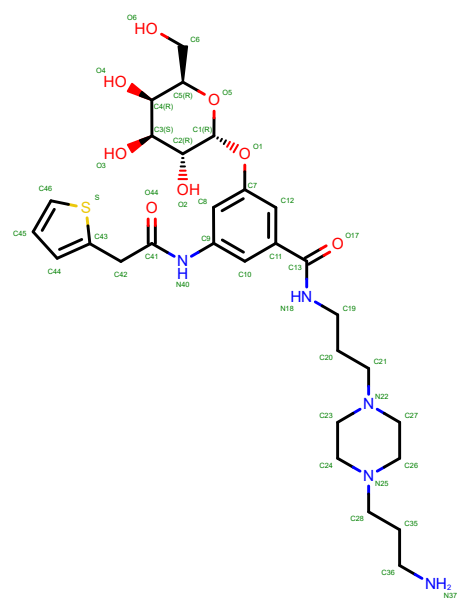


1PZK

Mitchell, D.D.
Pickens J.C.
Korotkov, K.
Fan, E.
Hol, W.G.J.

2004

Vibrio cholerae



2.3 OpenEye SPRUCE

All PDB's were converted to usable receptors for OpenEye docking tools using the OpenEye SPRUCE program. SPRUCE determines if the quality fulfils the need for docking. It also expands the asymmetric unit to the biological unit if necessary, enumerates alternate locations, builds missing pieces of the structure, optimizes hydrogen atoms and enumerates tautomers of cofactors and ligands.

For the building step, a loop database can be added for loopbuilding. The database added was *rscsb_spruce_1_0_20200422.loop_db*.

Spruce creates multiple design units per PDB and scores them. Per PDB, one design unit has been chosen based on the Iridium score. The Iridium score estimates the model quality of a structure, based on different global and local quality metrics. It scores the design units in four categories:

- HT Highly Trustworthy
- MT Mildly Trustworthy
- NT Not Trustworthy
- NA Not Applicable

The highest scoring design unit was chosen. If design units were placed in the same category, further distinction was made by underlying sub-scores. SPRUCE ranks design units from best to worst, so the best was chosen. (Spruce 1.5.2.0. OpenEye, cadence Molecular Sciences, Santa Fe NM. <http://eyesopen.com>)

2.4 Superposition - Protein RMSD study

Due to runtime we wanted to choose the best PDB to dock into and therefore did a similarity study between the PDB's if there possibly was a most representative PDB among the chosen PDB's. This was done with both PDB's and design unit files using a code written in Python (ver. 3.10) and the use the OpenEye Superposition application.

With different methods the RMSD was calculated between all possible pairs of the PDB's discussed in table 1. The Root Mean Square Deviation is average distance between atoms in the two given structures. These RMSD calculations have been done using multiple OpenEye Superposition methods.

- Global: Calculates the RMSD between all matched alpha carbon atoms. Alpha carbons are the first carbon atoms that attach to a functional group.
- Site: Calculates the RMSD between alpha carbons, focusing on the subset in the active site. For this calculation Spruce was done beforehand, marking the active site of the crystal structure.
- DifferenceDistanceMatrix (DDM) – Calculates the pairwise distance matrix of alpha carbon atoms for both structures given. After this, it calculates the difference from these two matrices. The resulting values either show a contraction or expansion of one structure when compared to the other.
- Weighted DDM – This method calculates Gaussian weights for all matching alpha carbons using the DDM, after this it uses the weights in the RMSD calculation.
- SecondaryStructureElements (SSE): Calculates shape overlap of the secondary structure elements of the two given structures, it does not report a RMSD score but a shape Tanimoto (between 0 and 1; 0 no overlap, 1 total overlap) instead.

(Spruce 1.5.2.0. OpenEye, cadence Molecular Sciences, Santa Fe NM. <http://eyesopen.com>)

2.5 Databases

The compound screening set is composed from multiple databases of Enamine and MCULE.

Table 2 - Databases and contents

Database	Number of molecules	Content type	Format	Database version
Enamine CLOUD	293	Diverse structures resembling FDA-approved chemical entities	2D SDF	20181010
Enamine Bioactive	2.320	Broad set of active compounds in diverse therapeutic areas	2D SDF	20220627
Enamine FDA approved	1.040	Relevant selection of drugs	2D SDF	20220708
Enamine HTS collection	1.371.031	High throughput screening set, diverse screening set	2D SDF	202303
Enamine Hit Locator	460.160	Diverse library with high MedChem tractability	2D SDF	20220221
MCULE full database	40.291.174	All of MCULE screening database	2D SMILES	230323
Total	42.126.028			

The databases were selected based on content, focussing on drug-like molecules and bioactive molecules. Besides that, MCULE and Enamine have been chosen due to availability and good customer support.

2.6 Database preparation - OpenEye FILTER

The OpenEye FILTER program (ver. 4.2.1.1 built 20221128) was used to eliminate unwanted compounds from the MCULE database. This was done only for the MCULE database as this database is the most diverse and is not pre-filtered on pharmacological properties while all Enamine databases were already drug/active focused. FILTER also deletes molecules too complex for the screenings, like molecules that are too big, flexible or contain filter-specified elements.

The default filter Blockbuster was used. Blockbuster is based on a set of drug molecules and therefore filters on drug-like pharmacokinetic properties. The three filtering categories are physical properties, atomic and functional group content and molecular graph topology. Some of the most important rules can be seen in Table 3. These rules were used to keep or omit molecules from the MCULE database.

Table 3 - Blockbuster filter rules

Rule	Minimum - Maximum
Molecular Weight	130 – 781
Heavy Atom Count	9 – 55
Carbon Count	3 – 41
Non-C and non-H atoms	1 – 14
Chiral Atoms	0 – 21
H-bond acceptors	0 – 13
H-bond donors	0 – 9
Total functional group count	0 – 7

Besides these rules a specified elemental filter is applied, which is based on metal removal and salt removal.

Metals eliminated by Blockbuster are: Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd.

Allowed elements are: H, C, N, O, F, P, S, Cl, Br, I.

Salts get removed by deleting atoms that are not part of the largest connected component of a compound.

(OMEGA 4.2.1.1. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.)

2.7 Database preparation - OpenEye Flipper

The OpenEye Flipper program (ver. 4.2.1.1 built 20221128) is part of the OMEGA OpenEye distribution. Flipper was used for enumerating **only** unspecified stereocenters (Fig 13).

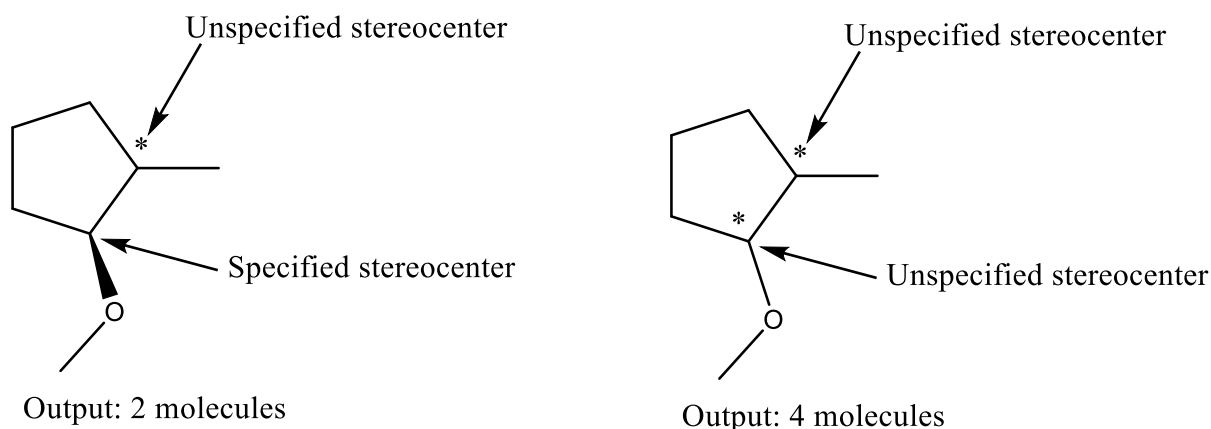


Figure 12 - Only unspecified stereocenters get enumerated by Flipper.

The filtered MCULE database and Enamine databases were put in Flipper which resulted in a total increase of database molecules. Flipper ran at default settings, which meant a maximal of 12 stereocenters were enumerated per molecule, if unspecified.

(OMEGA 4.2.1.1. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.)

2.8 Database preparation - OpenEye Tautomers

The OpenEye Tautomers program (ver. 2.2.1.1 built 20221128) is part of the QUACPAC OpenEye distribution. Tautomers was used for enumerating tautomeric forms of our database molecules (Fig 14).

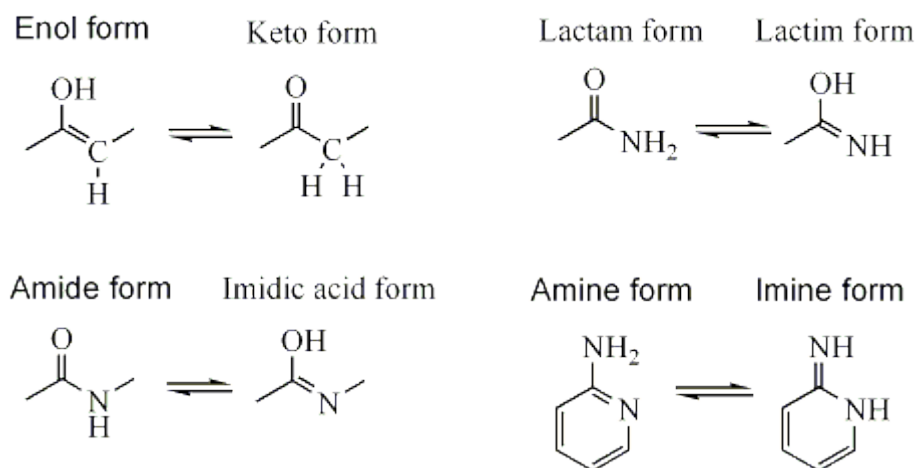


Figure 14 - Tautomer generation examples (*Tautomer* - wikidoc, z.d.).

All databases were put through tautomers, after Flipper. Tautomers ran at default settings, which means a maximum of 4096 tautomers per input compound were generated and a maximum of the 256 lowest energy tautomers were saved.

(QUACPAC 2.2.1.1. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.)

2.9 Database preparation - OpenEye OMEGA

The OpenEye OMEGA (ver. 2.2.1.1 built 20221128) is a program for generating 3D conformations of molecules. The original databases were in 2D SDF and 2D SMILES formats and therefore needed to be put through OMEGA for virtual screening use. Two different OMEGA methods were used.

OMEGA ROCS was used for preparing 3D conformers for ROCS similarity search. This method only creates a maximum of 50 conformers per input molecule as more conformers resulted in diminishing returns when comparing accuracy to running speed.

OMEGA POSE was used for creating 3D conformers for docking use. OpenEye docking requires more conformers than similarity search. It also takes flexibility into account, where more flexible molecules get more conformers. The standard is a maximum of 200 conformers for an input molecule. However, if a molecule contains 8 or more rotatable bonds and therefore is quite flexible, the maximum conformers created will be set to 800.

An enhanced fragment database was used to enhance the 3D conformer generation. It was obtained from OpenEye: *omega_fragment_lib_2020.oeb.gz*.

(Hawkins et al., 2010)

(OMEGA 4.2.1.1. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.)

2.10 Active compounds

The active compounds which will be used as positive control have been collected by literature study. They all showed binding to cholera with decent affinity and most important represent different classes of carbohydrate cholera toxin B ligands. They all have been discussed in the introduction section. The actives will be prepared with the same methods used for the database molecules and will be added to the database for the virtual screening runs. All molecules (1-12) that are described in the introduction are used.

2.11 OpenEye ROCS – similarity search

OpenEye Rapid Overlay Of Chemical Structures (ROCS ver. 3.5.1.1 built 20221128) is used to find molecules based on similarity to already known ligands. It performs shape based overlays of database molecule conformers on top of chosen query molecules (Fig 15). Besides shape it also takes the overlap in chemical features into account. These chemical features include:

- Donors – H-bond donors
- Acceptors – H-bond acceptors
- Hydrophobe groups – Groups with localized or delocalized negative charge
- Cations – Groups with localized or delocalized positive charge
- Anions – Aliphatic groups
- Rings – Rings of size between 4 and 7 atoms

For each molecule it will maximize heavy atom volume overlap and color overlap and save the best scoring conformer. The amount of similarity is expressed in the Tanimoto score, which is based on these overlaps.

$$Tanimoto_{A,B} = \frac{overlapAB}{selfA+selfB-overlapAB}$$

The formula divides the total rigid overlap between molecule A and B and divides this by the area of A and B minus the overlapped area. This results in a score between 0 and 1, where 1 is total overlap and 0 is no overlap at all. This formula is used for both the shape and color.

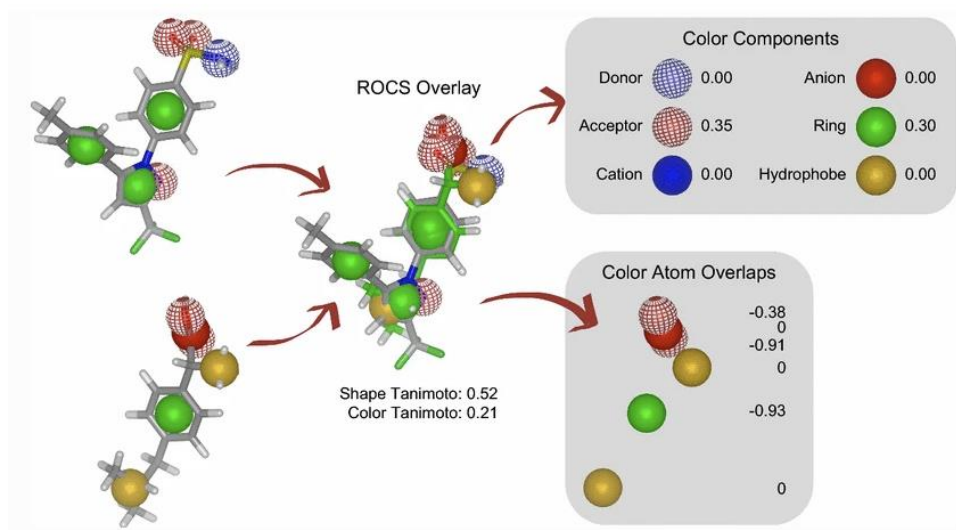


Figure 15 - Schematic overview of ROCS overlap calculations (Kearnes & Pande, 2016).

This scores can be easily combined, which leads to the TanimotoCombo score. This TanimotoCombo score therefore is scored between 0 and 2 and contains both shape and color information.

The query molecules used for ROCS are the ligands present in the receptors used for HYBRID docking (Figure 16). The ligands were extracted from their respective design units using the OpenEye VIDA visualization suite.

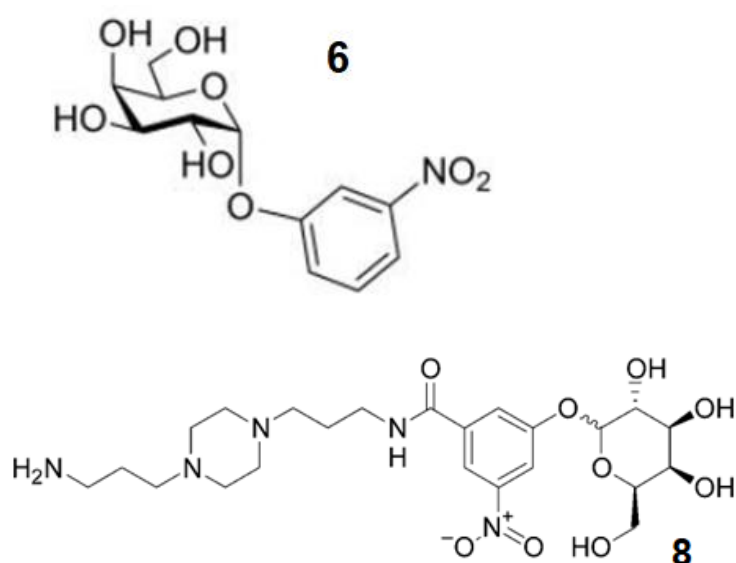


Figure 16 - 1PZJ ROCS query (8) and 1EEI ROCS query (6)

All database molecules were scored against these two queries and per molecule the best conformer score was saved.

(ROCS 3.5.1.1. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.)

2.12 OpenEye HYBRID – ligand guided docking

OpenEye HYBRID (ver. 4.2.0.1 built 20221128) is a docking program that uses ligand guided docking. This means that it also takes information of the ligand into account when docking your database molecules in the receptor. Therefore, to use HYBRID, a structure with a bound ligand is required.

As we used two separate runs with each one receptor, the HYBRID run consists of two steps.

Exhaustive search

1. Enumerates all rotations and translations of each conformer within the active site box.
2. Discards clashing poses.
3. Discard poses that do not contain chosen constraints. Constraints are user specified interactions that are required to dock database molecules. If these interactions are not present, the molecule will not be docked.
4. Score the remaining poses with Chemical Gaussian Overlay, which scores similarities based on the overall shape and the position of hydrogen bonding and metal chelating groups between original bound ligand and database molecule. This makes it unique when compared to non-ligand guided docking alternative FRED.
5. Sort poses by score and only save the best poses

Optimization

1. Enumerate nearby positions of each pose
2. Each pose is scored with the Chemgauss4 score. The Chemgauss4 score measures the complementarity to the active site of the pose by looking at shape, hydrogen bonding between ligand and protein, hydrogen bonding with the solvent and metal-chelator interactions.
3. The poses are sorted on Chemgauss4 score and the best pose is saved to represent the molecule

The docking was done with two receptor design units. 1EEI contained the small MNPG ligand and 1PZJ contained a MNPG derivative (See ROCS). Both these were chosen for docking due to the results in the PDB similarity study and pre-docking study and their vast differences in ligand-size.

1PZJ was run with constraints on Asn90, Lys91 and Glu51. This means that for a molecule to dock, it does to have these interactions or it will not be docked at all. This was done for time-saving reasons and also to add more information to the docking run, which leads to more accurate docking of the database molecules in the 1PZJ receptor.

(OEDOCKING 4.2.0.1. OpenEye, Cadence Molecular Sciences, Inc., Santa Fe, NM.)

2.13 Pre-docking actives study

All actives have been docked into the 1PZJ, 1EEI, 5LZG, 5LZJ and 1PZK design units. This was done to make a decision which design unit(s) to use further into the study. Docking was done using the HYBRID docking tool, without constraints and everything on default settings. The receptor(s) showing the best median and/or mean docking score would be deemed most fit for use for the virtual screening.

2.14 Datafusion - Reciprocal Rank Fusion

HYBRID scoring lists and ROCS scoring lists were sorted on score from best to worst. After this the molecules were ranked according to their spot in these sorted lists. The reciprocal was calculated per molecule per method using the following formula. The reciprocal is the inverse of the rank. When using multiple methods, for us HYBRID and ROCS, the reciprocals per molecule can be combined resulting in a reciprocal sum. This reciprocal sum therefore contains information about the rank in the HYBRID scoring output and ROCS scoring output.

$$\text{Reciprocal Rank Sum} = \left(\frac{1}{\text{rank}}\right)_{\text{HYBRID}} + \left(\frac{1}{\text{rank}}\right)_{\text{ROCS}}$$

The resulting reciprocal lists contains information about the molecules that were present in both HYBRID and ROCS results. Molecules that were omitted in one of the two methods due to errors were not kept for reciprocal ranking as both HYBRID and ROCS ranks are necessary for reciprocal calculations.

(Cormack et al., 2009)

2.15 Clustering results - Butina clustering

Butina clustering is an unsupervised molecule database clustering method created by Darko Butina in 1998. It was made to tackle the problems of the industry standard Jarvis-Patrick's clustering method. This method required lots of manual tinkering and finetuning to get correct results. To tackle this problem Butina created a method based on Tanimoto similarity by the use of Daylight's fingerprints. Daylight's fingerprints are bit-patterns that contain atom and molecule feature data, where each bit marks the presence or absence of a certain feature. This method only requires user chosen Tanimoto level, which can be quickly chosen. This makes the method easy to use and very fast. For this research the Volkamer Lab Python Butina Clustering method was used, which uses the original principle while using modern Python coding (*T005 - Compound Clustering*, z.d.).

Fingerprints need to be created from the SMILES input molecules. This needs to be done, as SMILES are not easily comparable by algorithms while fingerprints are. The fingerprint type used is the RDKit fingerprint, which is similar to the original Daylight fingerprints used by Butina.

Fingerprints are bit-patterns that describe chemical substructures of the input molecule. Each bit represents a feature being present with "1" or not present with "0"

(Fig 17).

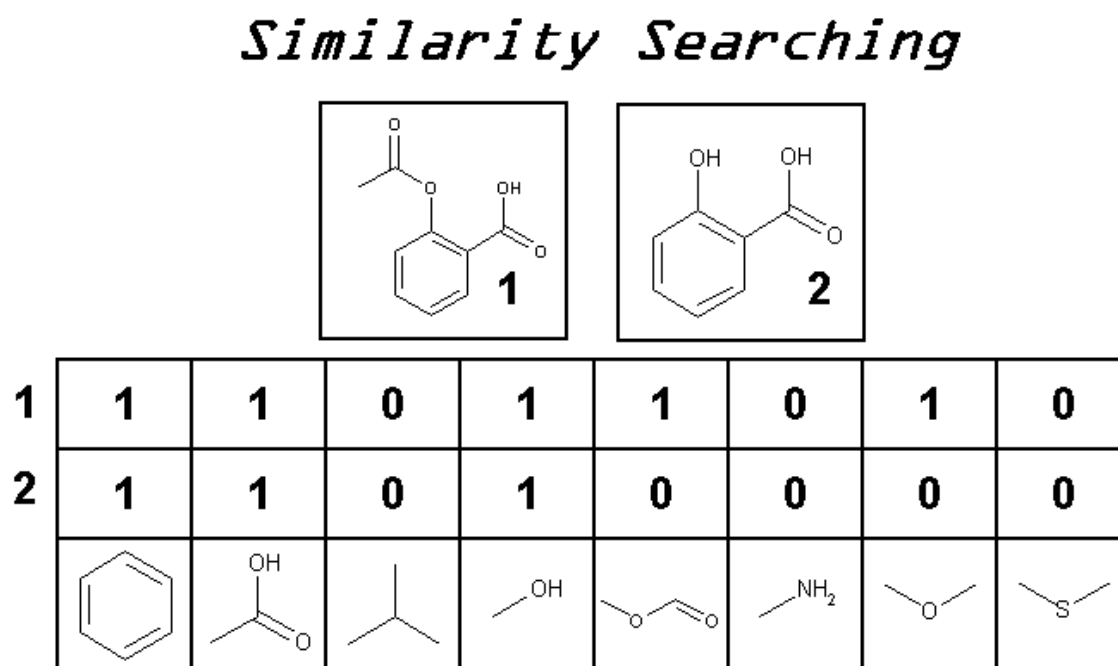


Figure 17 - Schematic representation of fingerprint generation. Each bit contains a certain chemical feature (Lasserre, 2018).

The RDKit fingerprints have a length of 2048 bits and a max path size of 5. The max path size of 5 means a bit-feature can maximal be 5 bonds long (C-C-C-C-C for example). As these fingerprints are made for the whole input database, the molecules you want to cluster, they can be compared by comparing the fingerprint bit-pattern.

For Butina clustering this comparison between fingerprints is done using Tanimoto similarity. The Tanimoto similarity can be expressed as follows:

$$\text{Tanimoto} = BC / (B1 + B2 - BC)$$

BC are the number of on-bits (1's) in both fingerprint 1 and 2. B1 are the on-bits in fingerprint 1 and B2 are the on-bits in fingerprint 2. For the figure above this means the following (Fig 16).

3 on-bits present in both.

Molecule 1 has a total of 5 on-bits.

Molecule 2 has a total of 3 on-bits.

$$\text{Tanimoto} = 3 / (5 + 3 - 3) = 0.6$$

The Tanimoto similarity is 0.6. The Tanimoto similarity again can be scored from 0 to 1, where 0 are two totally different molecules and 1 means it is the exact same molecule.

After translating the SMILES to these easily comparable fingerprints, all possible comparisons are made and stored in a matrix. This matrix therefore contains Tanimoto similarity scores between all possible molecule pairs.

To create clusters, a similarity threshold needs to be set also known as the Tanimoto level. This Tanimoto level is the maximal difference between molecule A and B allowed to be considered neighbours. Using the set Tanimoto level, all possible neighbours per molecule are calculated. After that the list is sorted on amount of neighbours. After that the dataset is clustered using this list. From the top (molecule with most neighbours) the data will be clustered, omitting molecules already clustered for further clustering. This will result in X amount of clusters, with some molecules left out that are not clustered at all. These molecules are called singletons and are not similar enough to any of the centroid molecules and therefore form their own cluster.

The Tanimoto level is chosen manually, by looking at the clustering data and 3 metrics. Those metrics are:

- Amount of clusters & cluster-size: It is required not to create too many clusters, as you actually want to divide your data into similar groups, but not divide them in only singletons. At the other end, too less clusters will lead to big clusters and not enough distribution.
- Inter-cluster similarity: The similarity between two random molecules from two different clusters. A lower value is better, as a clear distinction between clusters is wanted.
- Intra-cluster similarity: The similarity between two random molecules from the same cluster. A higher value is better, as you want similar molecules in a cluster.

Looking at the clustered data and these metrics will help choosing a suitable Tanimoto level for the dataset. The most important is to find a clear distinction between the intra- and inter-cluster similarity and have them as far away from each other as possible. For both 1PZJ and 1EEI runs the Tanimoto level was set at 0.6 after inspection.

The Butina clustering was done with the top 1000 set molecules of the reciprocal sum lists of both 1PZJ and 1EEI. After that lists containing the most representative molecules per cluster were extracted, as are the best HYBRID hits. This was all done with the Volkamer Lab Python Code. (Butina, 1999; *T005 · Compound Clustering*, z.d.)

2.16 Visualization of hit molecules

Visualization of hit molecules has been done using the VIDA OpenEye visualization tool. Using the FRED view function molecules could be visualized in bound state in the binding site. In FRED view, possible interactions have been highlighted by distance measuring. The two original ligands have also been visualized in FRED view for comparison (Fig 18 & 19).

The 1PZJ molecule shows interactions similar to GM1. For 1PZJ constraints have been chosen for the obvious interactions Asn90, Lys91 and Glu51. These interactions were found in more of the 7 PDB's and therefore chosen as constraints. Trp88 does also show possible stacking with the galactose ring.

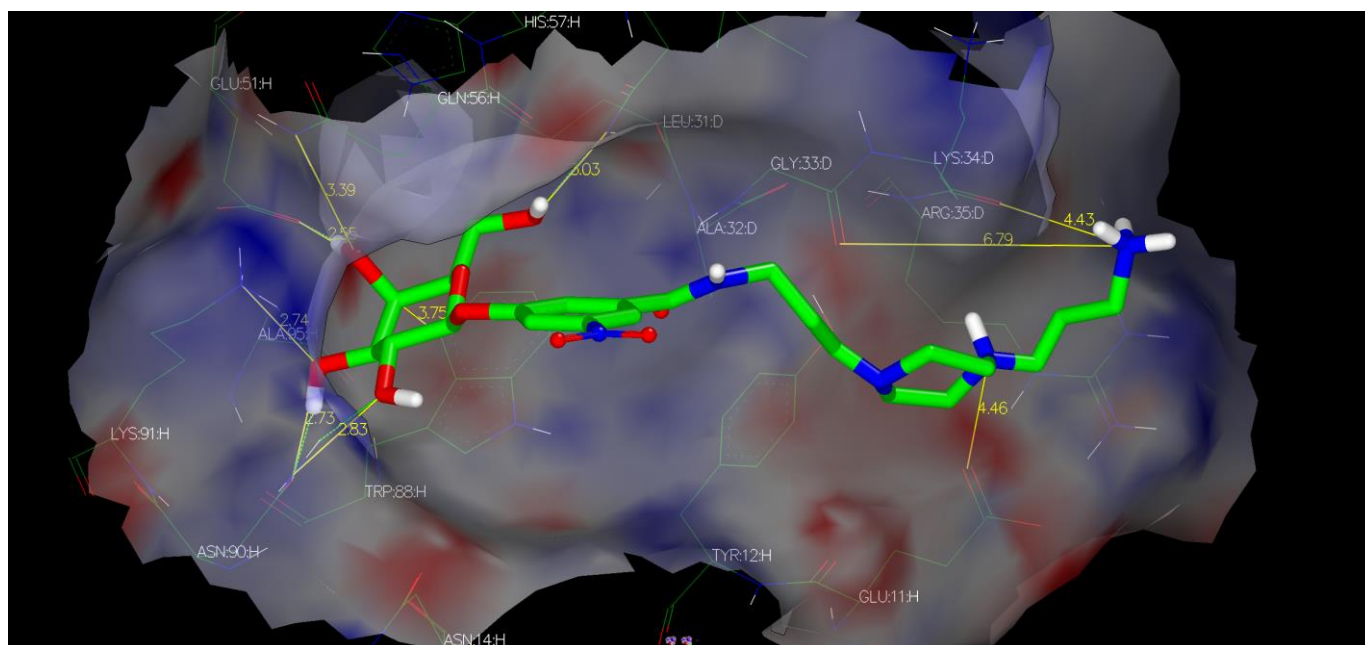


Figure 18 - 1PZJ receptor and ligand in FRED view.

The 1EEI molecule also shows interactions with Asn90, Lys91, Glu51. Just as 1PZJ stacking is observable between the galactose ring and Trp88. 1EEI, in contrast to 1PZJ has been run without constraints.

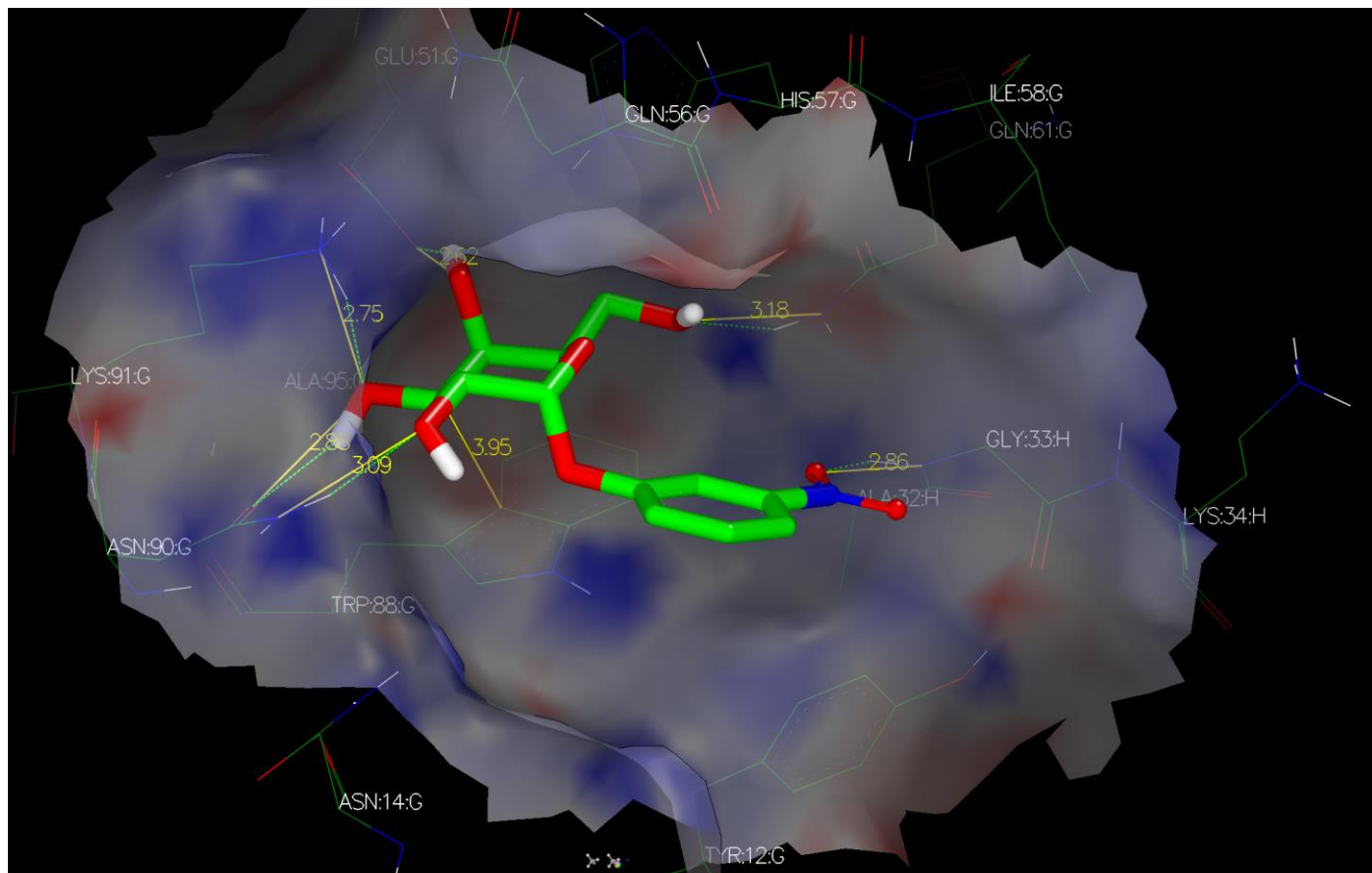


Figure 19 - 1EEI receptor and ligand in FRED view.

(VIDA 5.03. OpenEye, Cadence Molecular Sciences, Santa Fe, NM.)

3. RESULTS

3.1 Receptor Studies

Prior to running HYBRID and ROCS the prepared design units were compared using an RMSD study and pre-docking actives study described in the method section. The 3CHB and 5LZH design units were omitted after SPRUCE, as 3CHB could not be ranked (NA) and 5LZH resulted in unknown and unsolvable problems. The design units used were all scored HT or MT by SPRUCE.

The RMSD study showed very low RMSD values for all methods, which means there is no significant difference between PDB's. Based on these results we could not chose a most representative PDB.

It was decided to dock all actives into the 5LZJ, 5LZG, 1PZJ, 1PZK and 1EEI receptors. 1EEI and 1PZJ showed the best mean and median Chemgauss4 scores and therefore were the chosen targets for HYBRID and ROCS runs in this study.

Exact results of these studies can be found in the supplementary.

3.1 HYBRID

HYBRID runs were performed for both 1PZJ and 1EEI with a total of 17.202.162.458 conformers from 54.798.546. This resulted in a scoring list per receptor. 1PZJ was run with the constraints mentioned in the method section, while 1EEI was run without constraints.

The HYBRID 1PZJ run resulted in 32.211.095 molecules docked into the 1PZJ receptor. This is however still with Flipper and Tautomeric forms, which means molecules are present multiple times. After deleting the duplicates 18.764.548 unique docked duplicates were found. The first ranking molecule MCULE-2839448221 showed a score of Chemgauss4 -13.951094 (Fig 20).

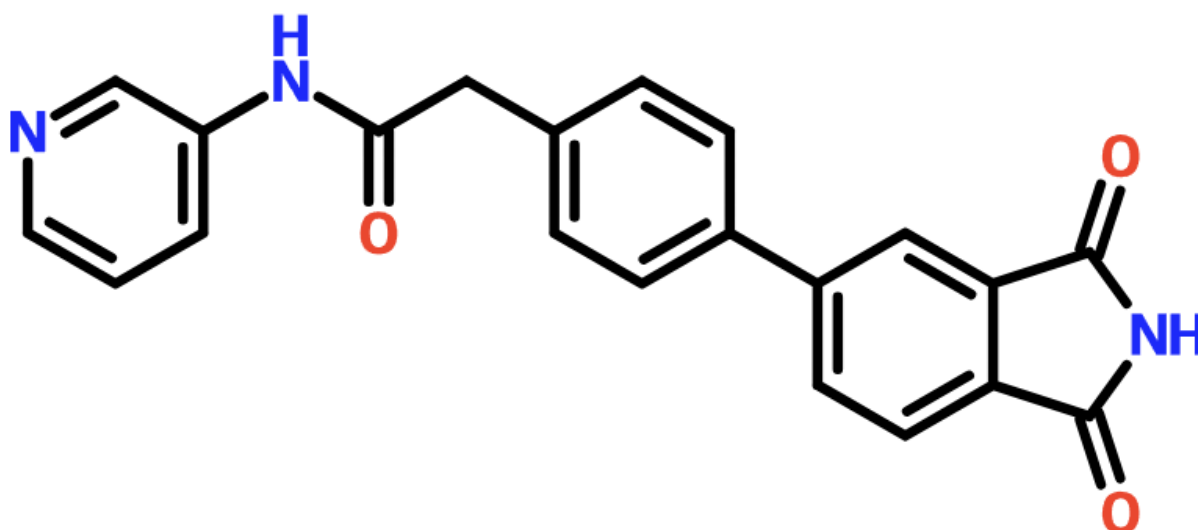


Figure 20 - 2D structure of MCULE-2839448221

All scores were between 30.396610 and -13.951094. Most actives were present in the negative score range.(Fig 21)

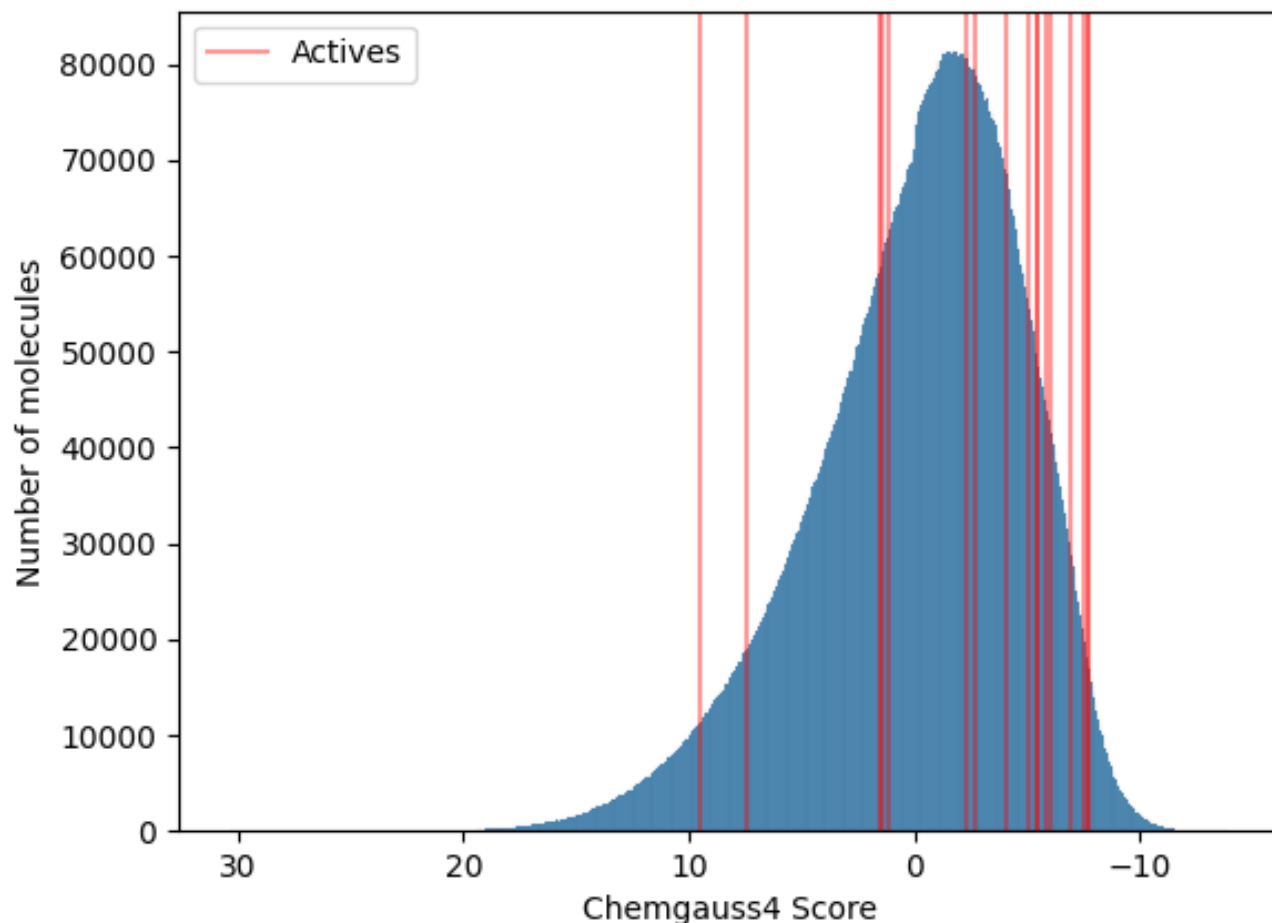


Figure 21 - Score distribution HYBRID 1PZJ

The HYBRID 1EEI run resulted in 53,015.957 molecules docked into the 1EEI receptor. However, without Flipper and Tautomeric forms it ended up being 34,336.463 unique molecules docked. The molecules were docked between scores 22.572247 and -13.132102. The first ranking molecule is MCULE-7853952702 with the score of -13.132102 (Fig 22).

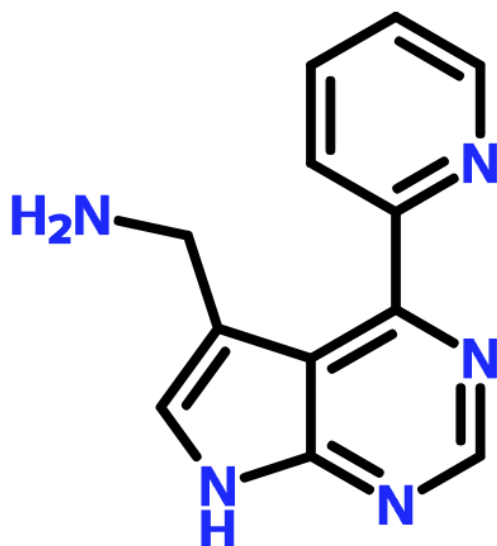


Figure 22 - 2D structure of MCULE-7853952702

Most scores are in the negative range. 1EEI docked less actives, although most actives are still present in the negative range and the best docking active is the original MNPG ligand. (Fig 23)

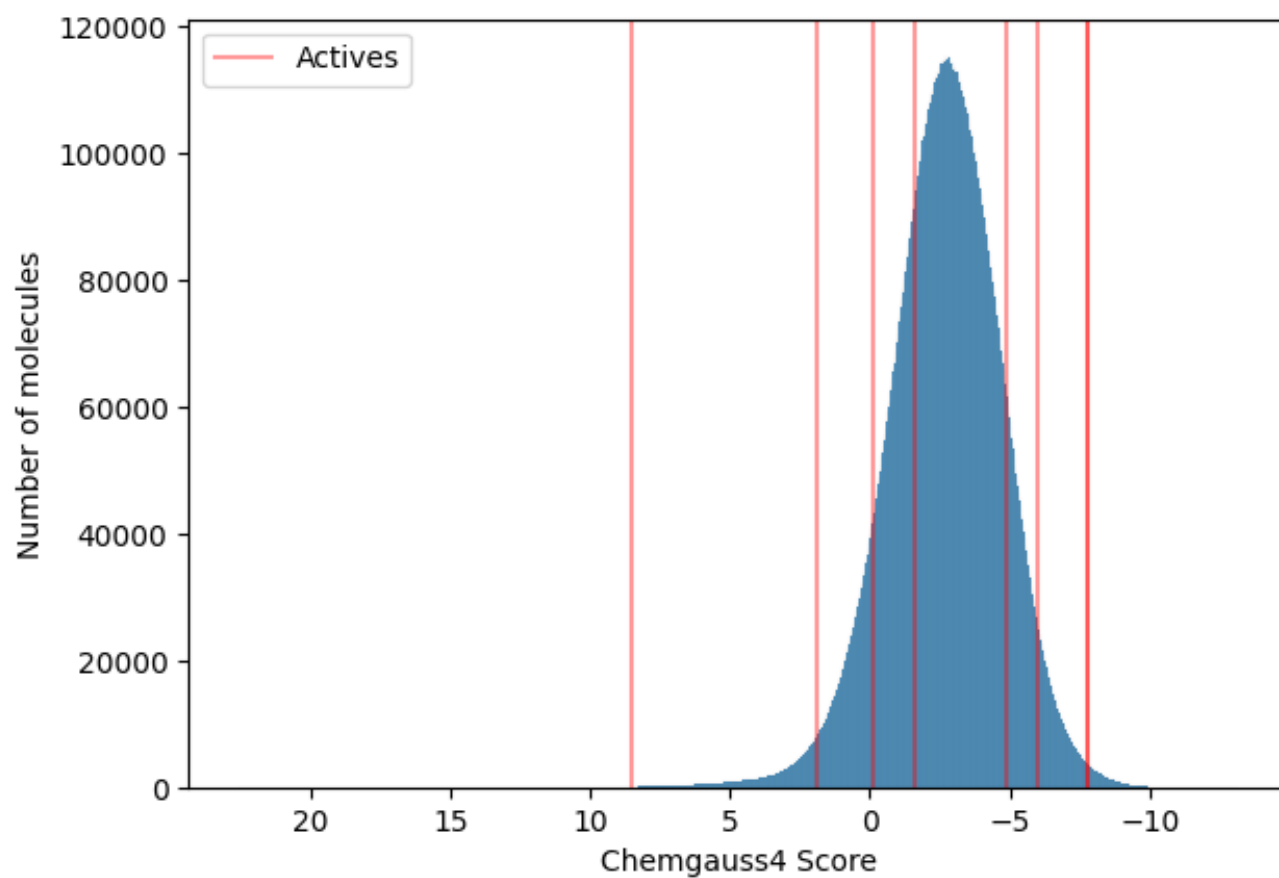


Figure 23 - Score distribution of HYBRID 1EEI

3.2 ROCS

ROCS was a combined run, where the database molecules were compared to both 1EEI and 1PZJ ligand. The database contained 54.798.546 molecules and 2.454.999.185 conformers.

The 1PZJ ROCS run resulted in a scoring list of 54.798.543 molecules, with scores between 1.111 and 0.122. This list was still with all duplicates due to Tautomers and Flipper. When selecting the best scores per molecules and deleting duplicates, the final result was a list of 35.357.425 unique molecules which have been compared to the 1PZJ ligand.

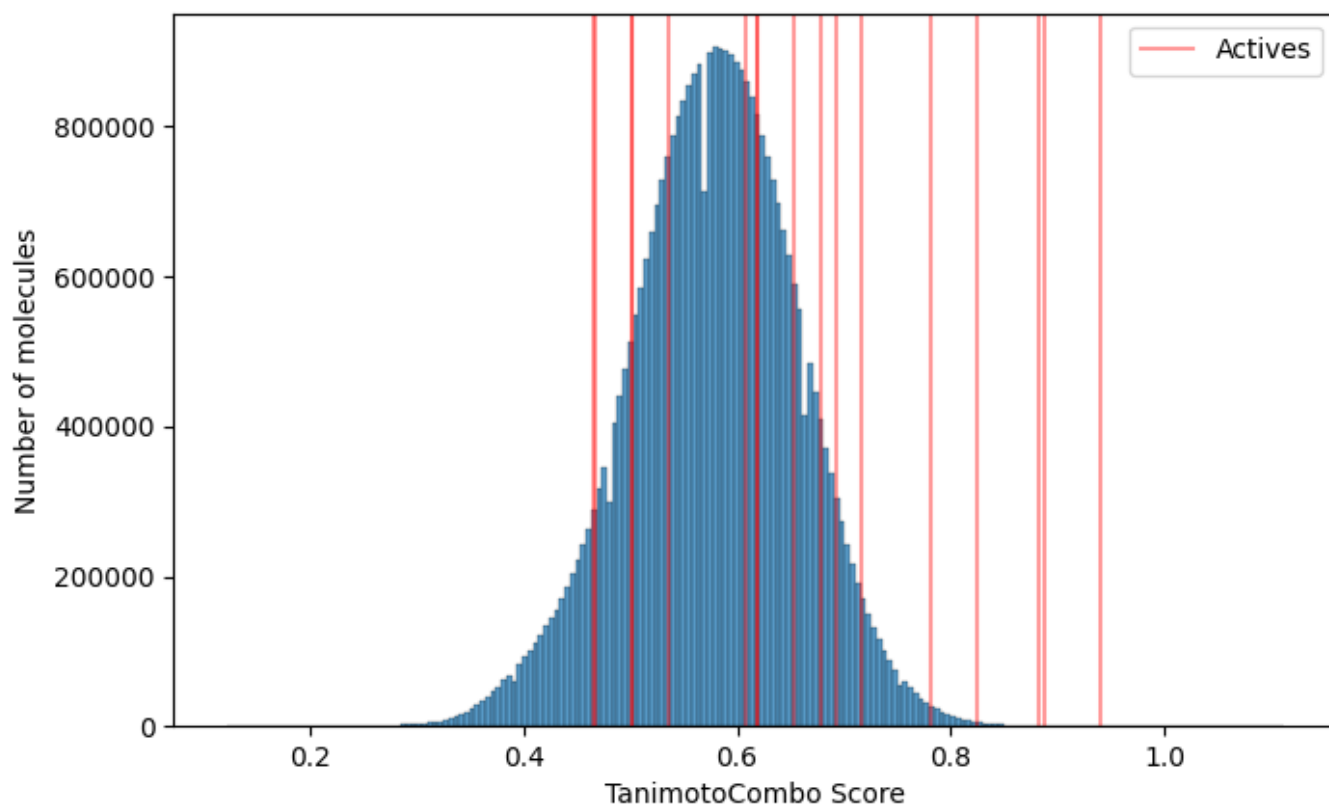


Figure 24 - Score distribution ROCS 1PZJ

No high TanimotoCombo scores were found, the best scoring active has a score of 0.940. Which was the original 1PZJ ligand (Fig 24). The best scoring molecule was MCULE-8937237797 with 1.111 (Fig 25).

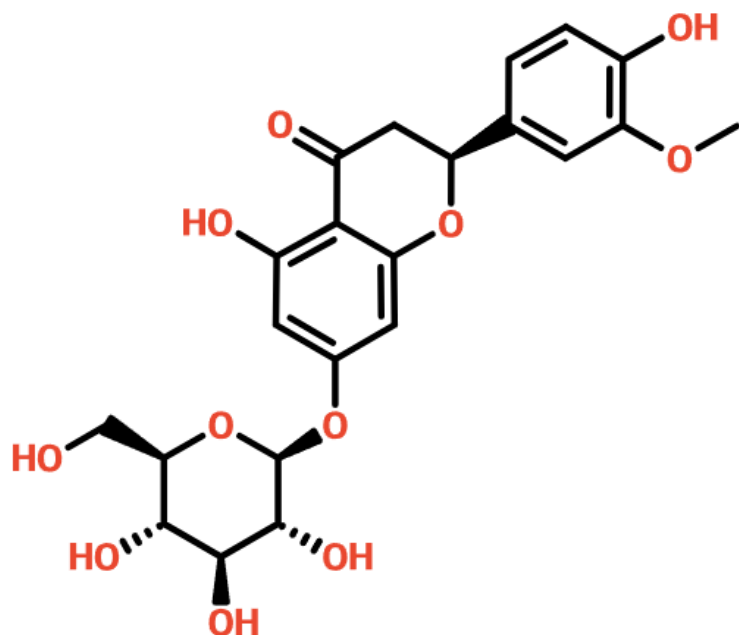


Figure 25 - 2D structure of MCULE-8937237797.

The 1EEI ROCS run had the same original size of 54.798.543 molecules. It had scores between 0.266 and 1.682. After removing duplicates it again resulted in a total list of 35.357.425 unique molecules. The highest scoring active was the original 1EEI ligand with a score of 1.656 (Fig 26).

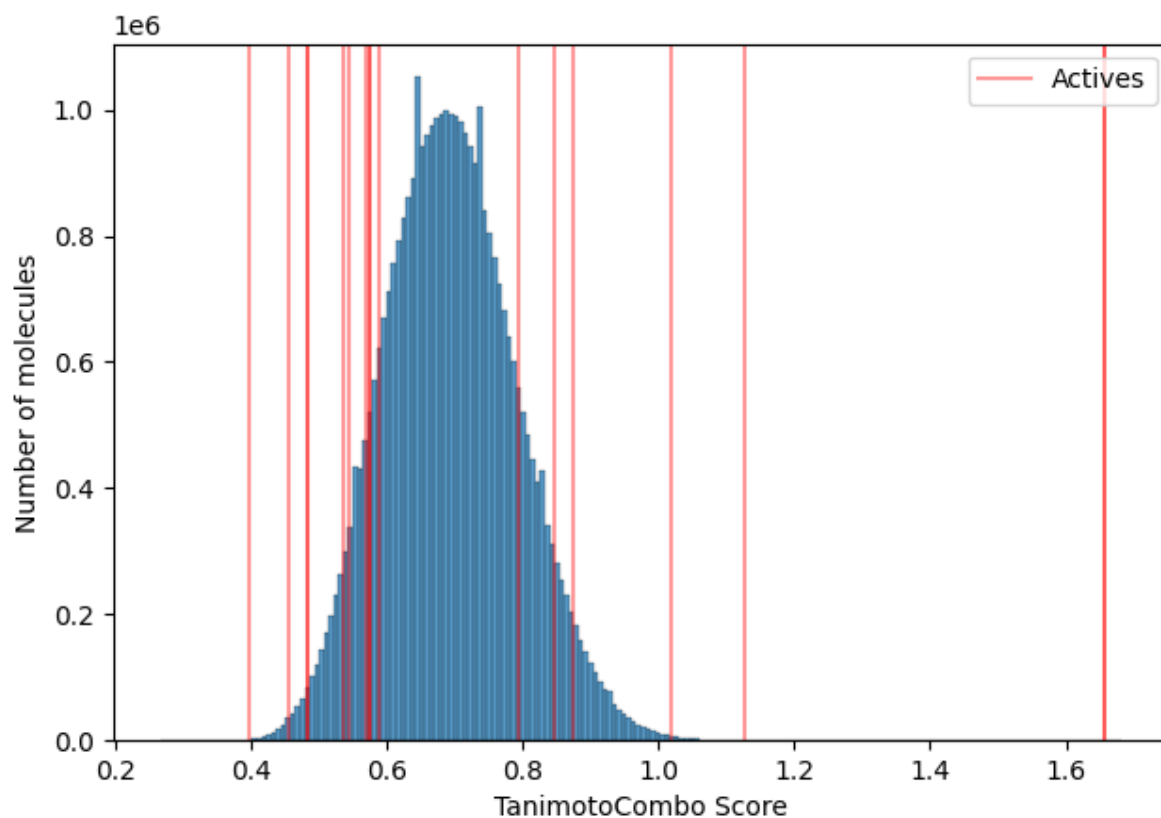


Figure 26 - Score distribution ROCS 1EEI

The best scoring molecule was MCULE-1513474432 with a TanimotoCombo score of 1.682 (Fig 27).

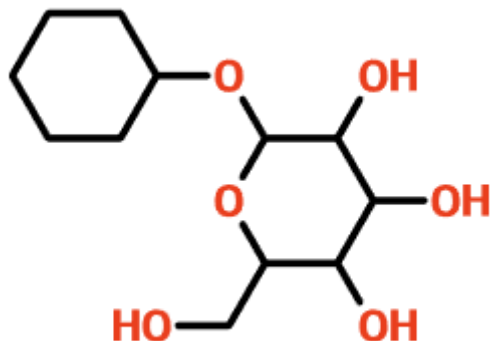


Figure 27 - 2D structure of MCULE-1513474432

3.3 Reciprocal Sum Rank calculations

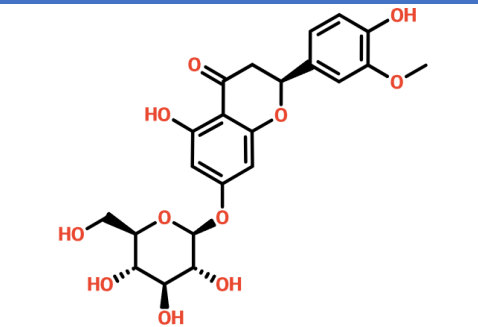
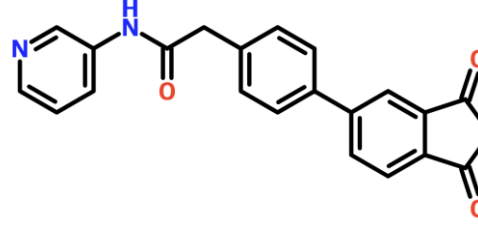
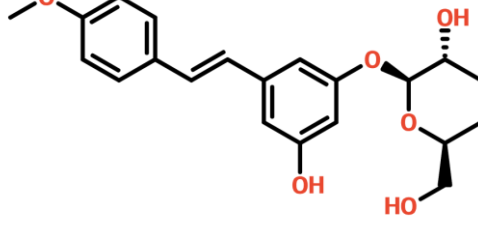
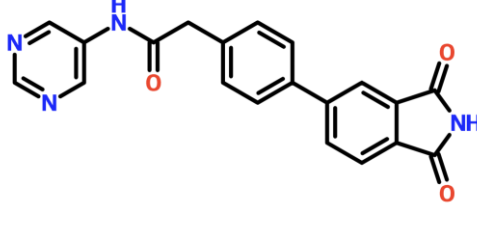
The total length of the reciprocal sum lists represented the length of the HYBRID scoring lists, as molecules needed to be present in both lists.

This resulted in a 1PZJ reciprocal list size of 18.764.548 unique molecules

For 1EEI 34.336.463 unique molecules made up the reciprocal scoring list of this run.

From these lists the top 1000 was used for further examination using the Butina Clustering. The reciprocal sum rank calculations resulted in the following top 10 molecules for both runs, which includes the top 5 of ROCS and HYBRID for both receptors.

Table 4 - 1PZJ reciprocal top 10

Molecule	HYBRID rank	ROCS rank	Reciprocal sum	2D structure
MCULE-8937237797	179115	1	1.00000558301	
MCULE-2839448221	1	4228376	1.000000236497	
MCULE-4211909652	23562	2	0.5000424412	
MCULE-5030435107	2	4702056	0.500000212673	

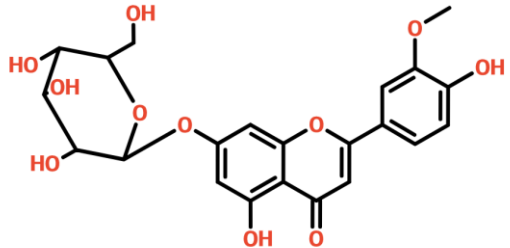
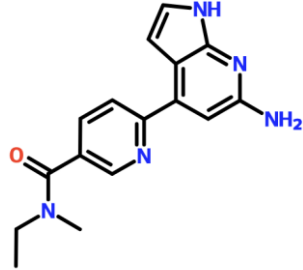
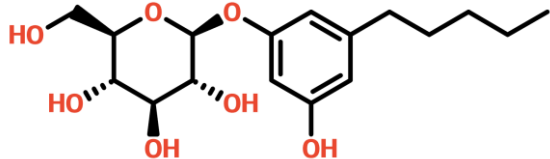
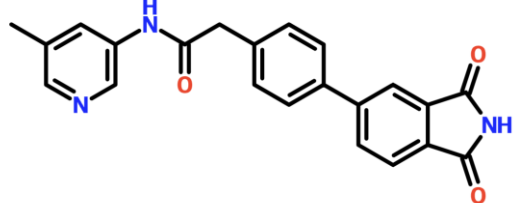
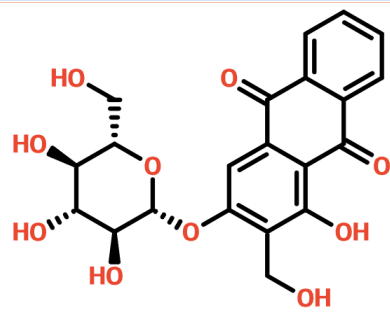
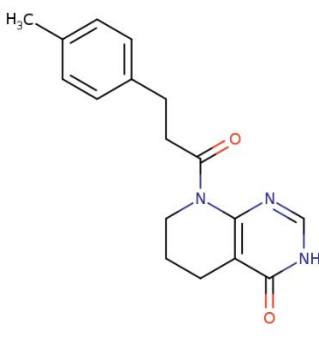
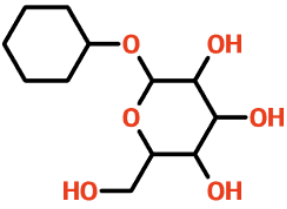
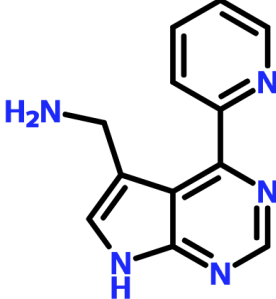
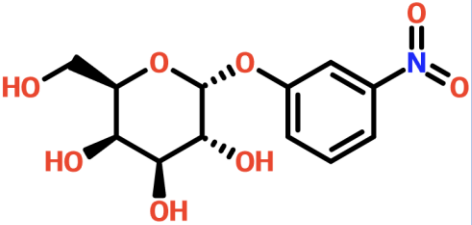
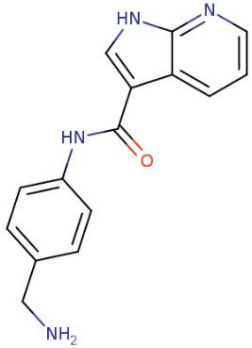
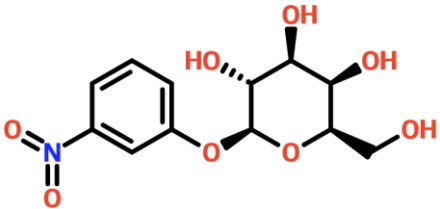
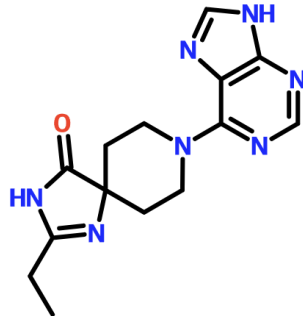
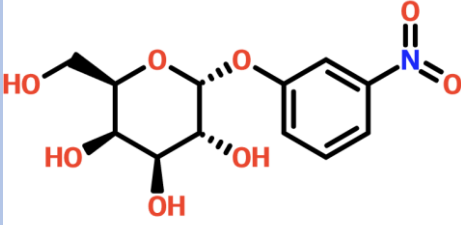
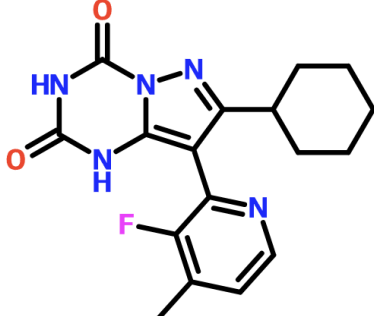
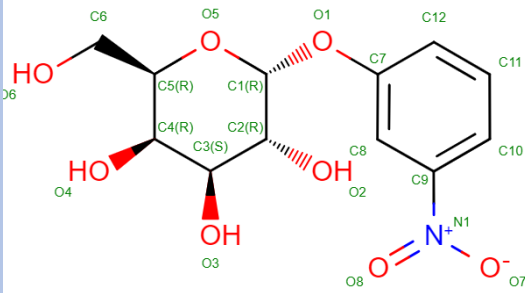
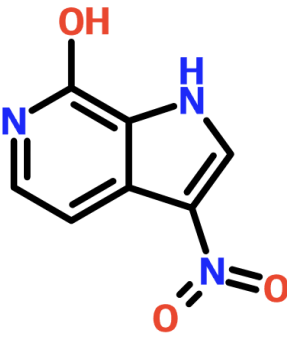
MCULE-1885456555	390281	3	0.33333556226	
MCULE-8133751560	3	24803220	0.3333330403173	
MCULE-1057184189	33665	4	0.2500297044	
MCULE-9436696942	4	3209110	0.250000311613	
MCULE-9482069112	97288	5	0.20001027880000002	
Z2874994383	5	10696566	0.20000009348790002	

Table 5- 1EEI reciprocal top 10

Molecule	HYBRID rank	ROCS rank	Reciprocal sum	2D structure
MCULE-1513474432	24296	1	1.000041159	
MCULE-7853952702	1	1916955	1.000000521661	
MCULE-5557798235	189178	2	0.50000528603	
Z3711495684	2	7829797	0.500000127717	
MCULE-4783577005	30310	3	0.3333659924	

MCULE-2430515474	3	2050198	0.3333330403173	
MCULE-4686674347	188178	4	0.25000528605	
MCULE-9486643068	4	2760633	0.250000362236	
MNPG Active	189176	5	0.20000528608	
MCULE-1407049348	5	10369943	0.20000009643250002	

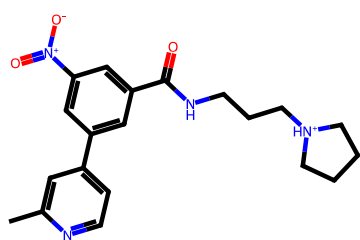
3.4 Butina clustering results

Clustering of the top 1000 reciprocal molecules of 1PZJ resulted in 102 clusters with the Tanimoto level set at 0.6 and showed clusters of different sizes (Table 6). The inter-cluster similarity, where two random molecules of two different clusters are taken, showed a Tanimoto similarity of 0.13. The intra-cluster similarity, where two random molecules in the same cluster are taken, was 0.42.

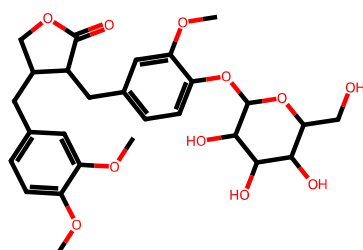
Table 6 - Butina cluster sizes 1PZJ

Cluster size	Number of clusters
Singletons	44
>5 molecules	24
>25 molecules	8
>100 molecules	2

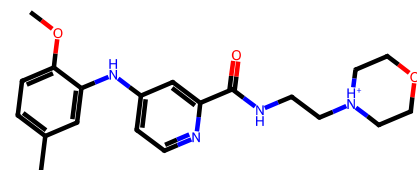
The 8 bigger clusters, which have 25 or more molecules, had the following most representative molecules. Many ringlike-structures are present. Cluster 2 and 7 do contain carbohydrate structures. Other clusters contain nitrogen-heavy ring structures and side-groups.



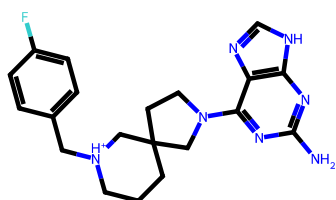
Cluster 1



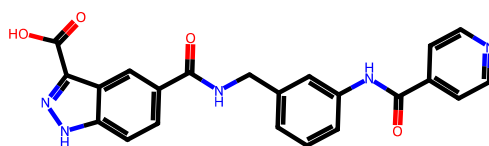
Cluster 2



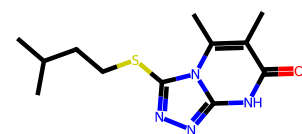
Cluster 3



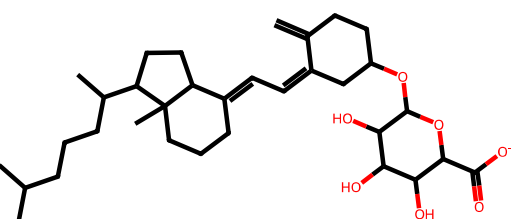
Cluster 4



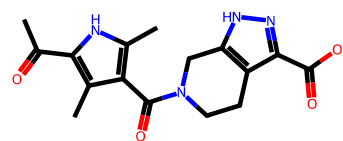
Cluster 5



Cluster 6



Cluster 7



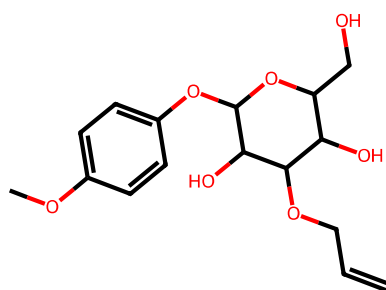
Cluster 8

The clustering of 1EEI resulted in a total of 66 clusters with a Tanimoto level set at 0.6 (Table 7). The inter-cluster similarity was Tanimoto 0.10 and the intra-cluster similarity is 0.73.

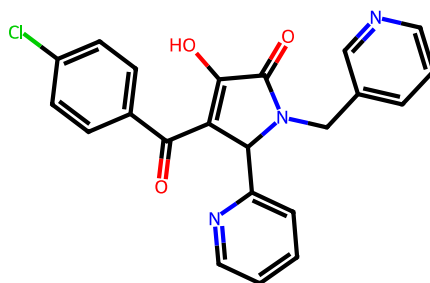
Table 7 - Butina clusters sizes 1EEI

Cluster size	Number of clusters
Singletons	24
>5 molcules	21
>25 molecules	6
>100 molecules	2

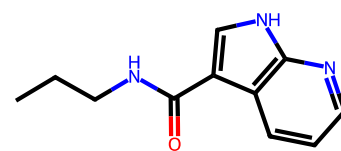
The 6 bigger clusters, which had 25 or more molecules, had the following molecules.



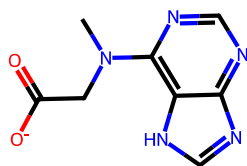
Cluster 1



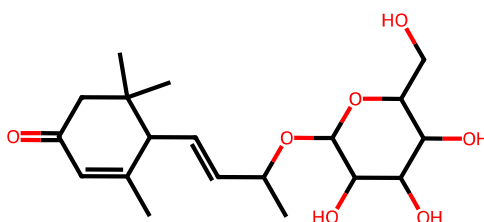
Cluster 2



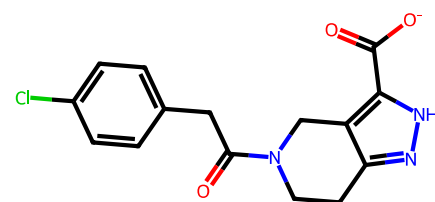
Cluster 3



Cluster 4



Cluster 5



Cluster 6

A clear separation can be made between carbohydrate-like molecules, which are in cluster 1 and cluster 5. The other clusters show again nitrogen rich cyclic groups.

3.5 Hit molecule visualization

Using VIDA the best docking molecules from different clusters have been visualized and interactions have been marked. This was done for both runs.

MCULE-2839448221 was the best docking molecule in 1PZJ with a Chemgauss4 score of -13.95 (Fig28). All interactions with the constraints are clearly present as there are interactions with Asn90, Lys91 and Glu51 (~ Distance between 2.6 and 3.3Å). The distance to Asn90 is somewhat longer, but still worth to mention. Another possible interaction is interactions with Gln56, which is also observed with GM1. Hydrophobic aromatic stacking interaction is possible between the Trp88 and the molecule, with a distance of 3.87Å. Interactions are also possible at the other end of the molecule with Gly33 and due to flexibility maybe with other groups as well.

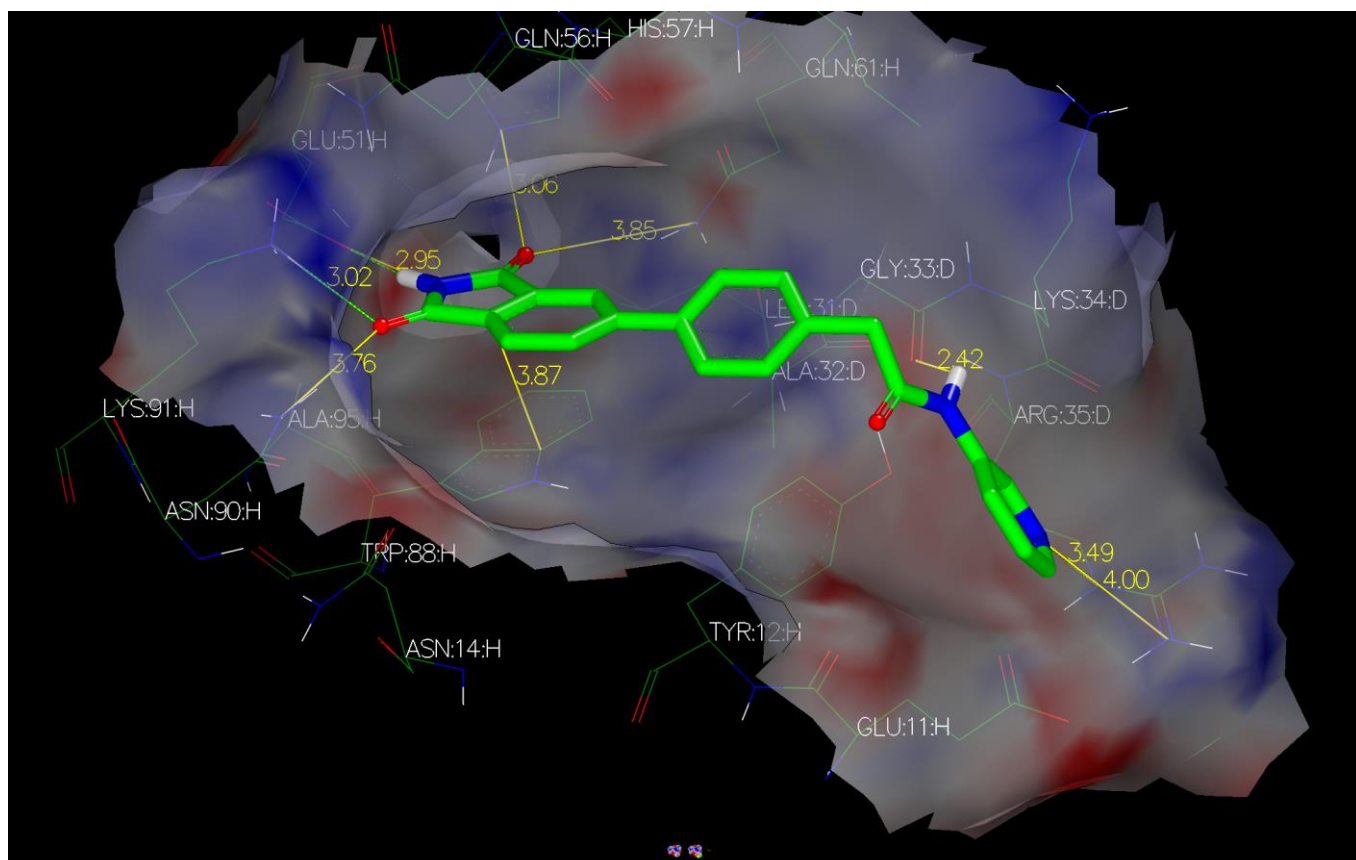
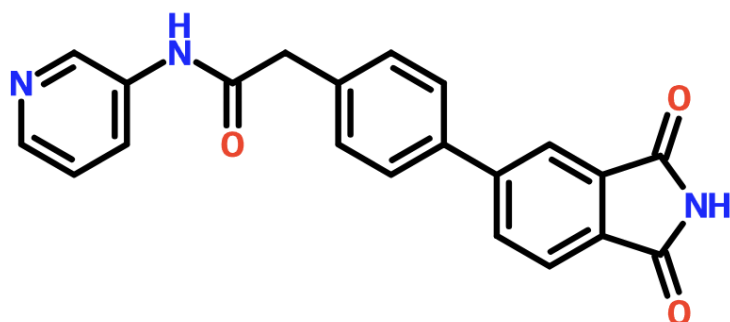


Figure 28 - 2D structure MCULE-2839448221 of with below the 3D pose visualized in VIDA FRED view. The yellow lines highlight the interactions with corresponding distances.

The second best docking molecule from 1PZJ was MCULE-8133751560 (Fig 29). Again all constraints are present. The distance is still short enough for stacking. However, other interactions are absent, only an extra possible interaction can be seen with Trp88.

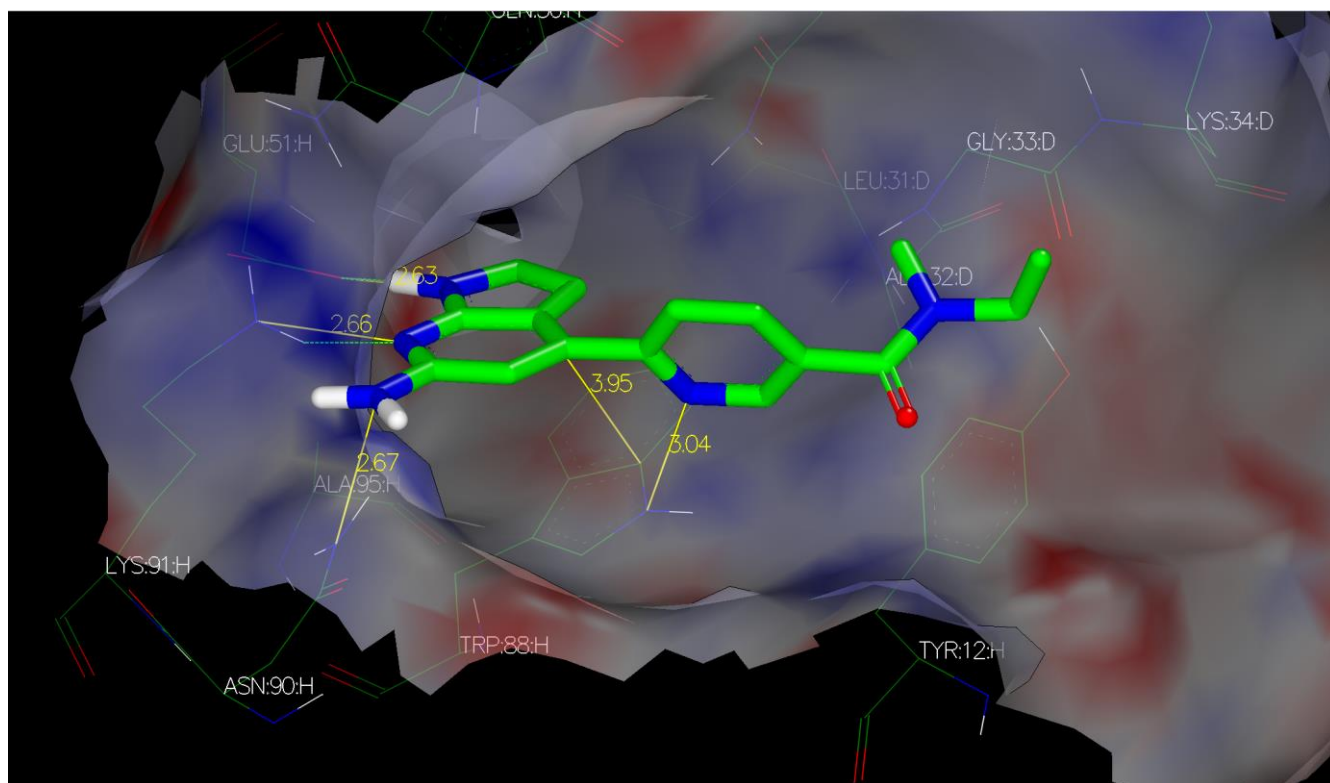
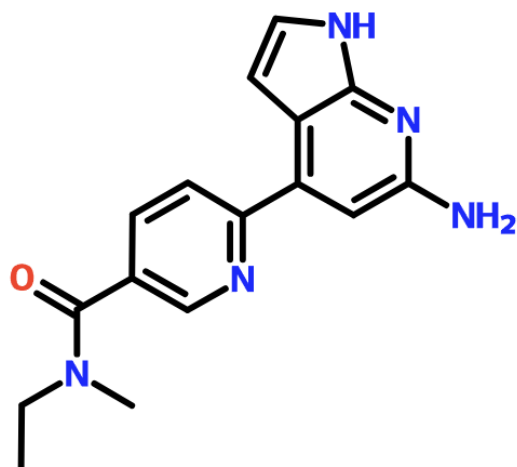


Figure 29 - 2D structure MCULE-8133751560 of with below the 3D pose visualized in VIDA FRED view. The yellow lines highlight the interactions with corresponding distances.

The third best molecule in the 1PZJ visualization was Z2874994383 (Fig 30). Again all constraints are present, although the distance to the Asn90 can be considered too long. Hydrophobic stacking with Trp88 is also present. Elsewhere in the molecule, the carboxyl (=O) engages in some possible interactions with Gln61 and Glu33.

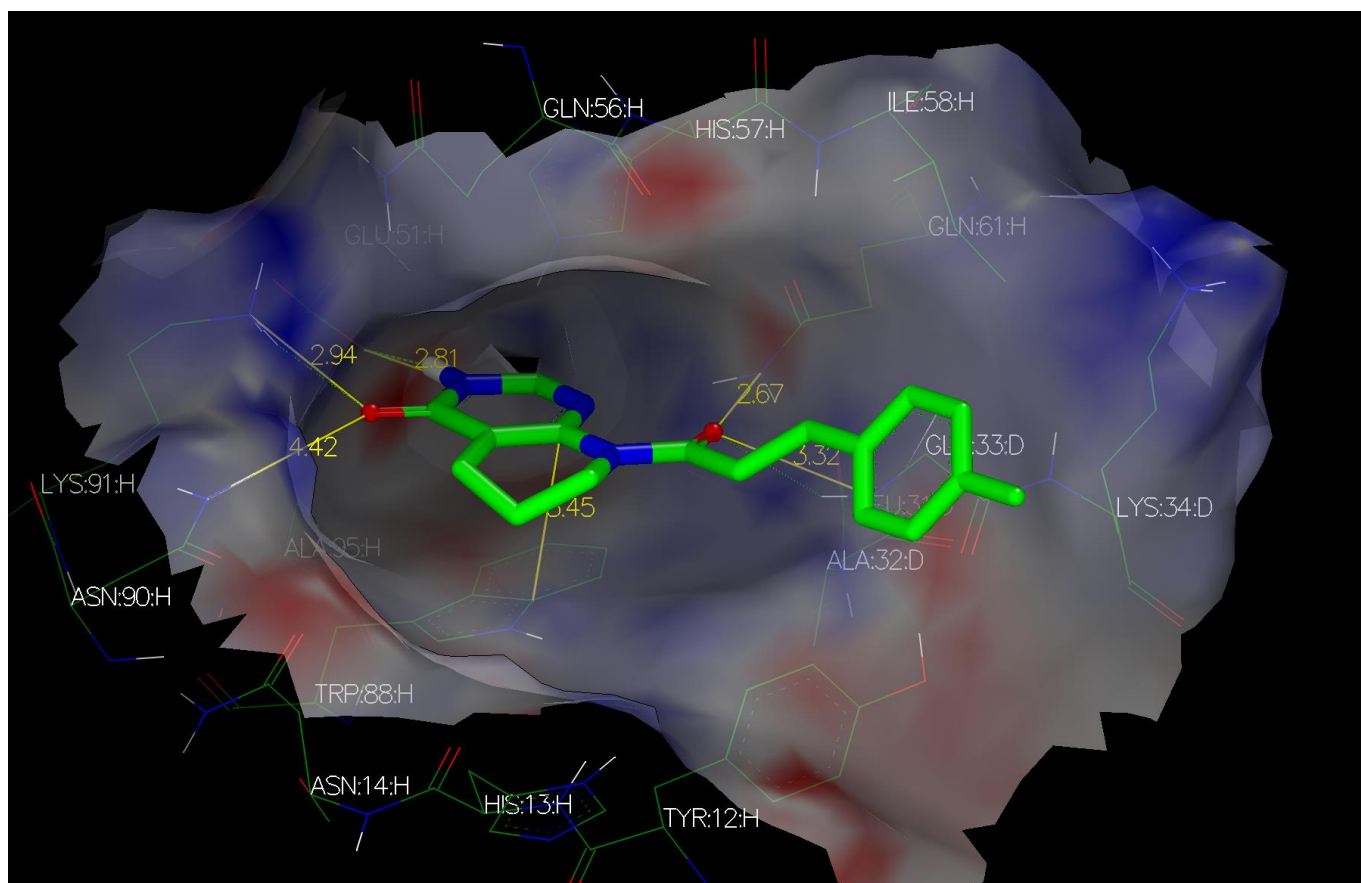
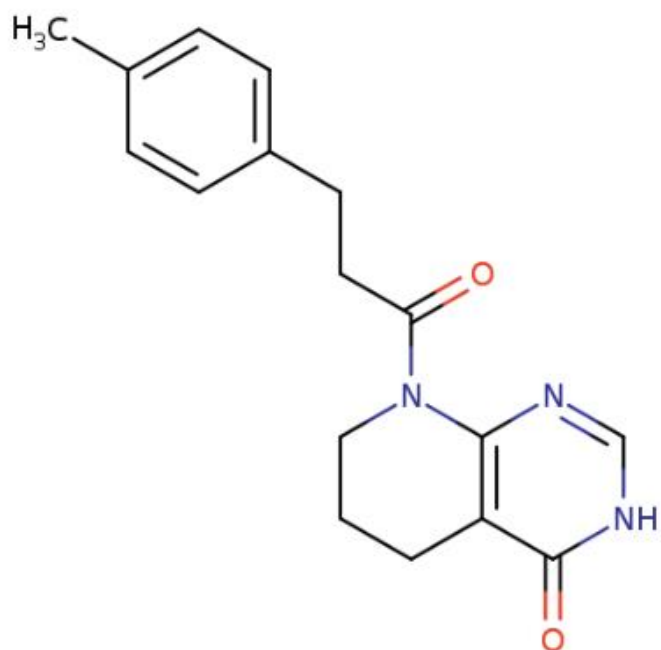


Figure 30 - 2D structure Z2874994383 of with below the 3D pose visualized in VIDA FRED view. The yellow lines highlight the interactions with corresponding distances.

For 1EEI the best scoring molecule was MCULE-7853952702 with a score of -13.31 (Fig 31). Although this run was not with constraints, possible interactions with Asn90,

Glu51 and Lys91 were all present. Stacking again was possible with the Trp88 and even another interaction with Trp88 is possible by the nitrogen in the most right ring structure. The NH₃⁺ shows possible interactions with Gln56 and Gln61.

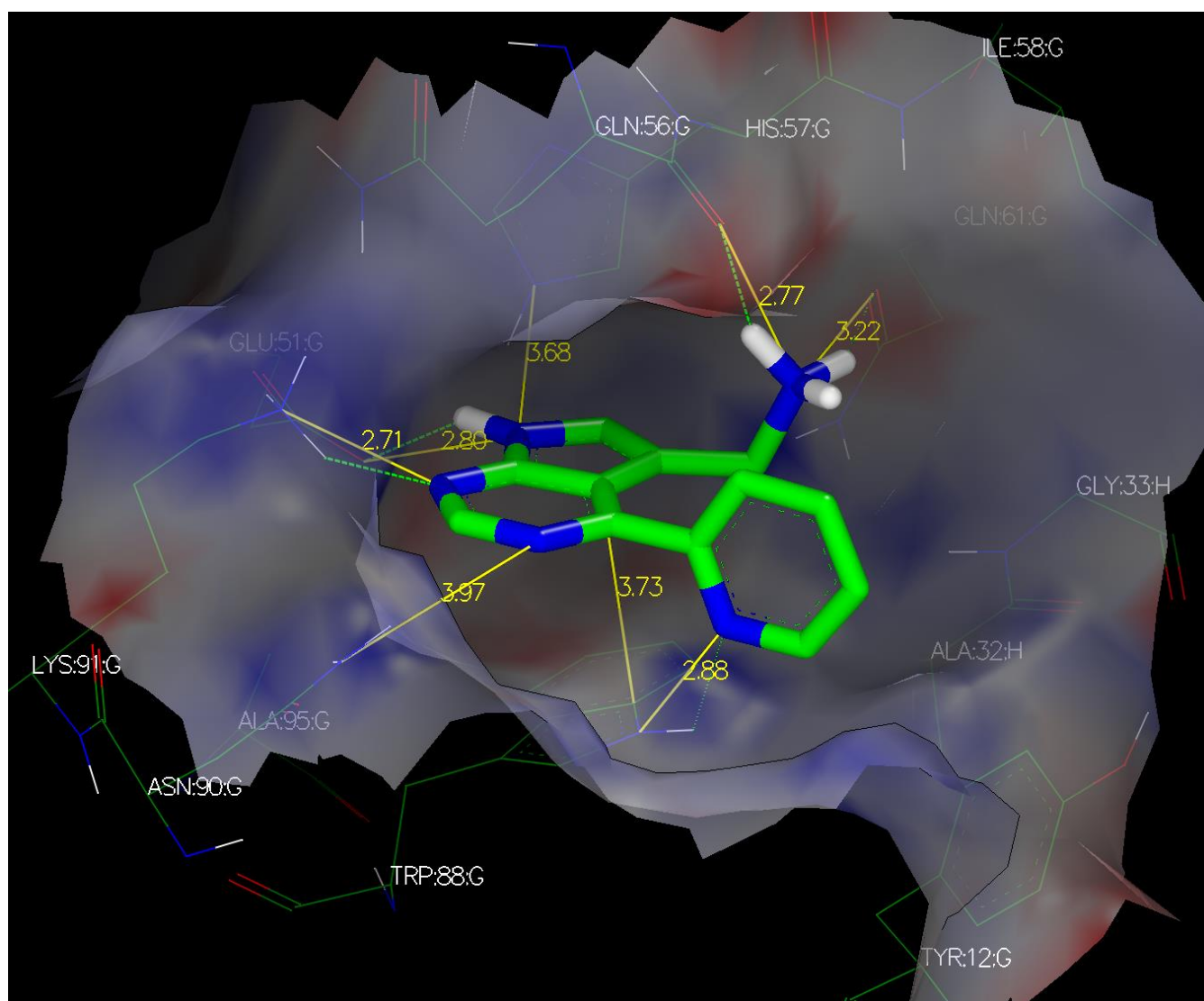
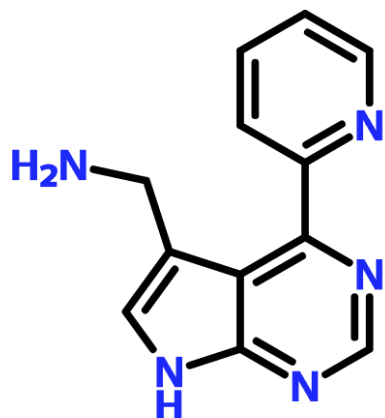


Figure 31 - 2D structure MCULE-7853952702 of with below the 3D pose visualized in VIDA FRED view. The yellow lines highlight the interactions with corresponding distances.

The second molecule after clustering for 1EEI was MCULE-2430515474 with a score of -12.59 (Fig 32). Possible interactions are observable with the Lys91 and Glu51. The possibility for stacking has decreased with, although still possible. Further to the right we see possible interactions with Gly33 by a carboxyl and NH.

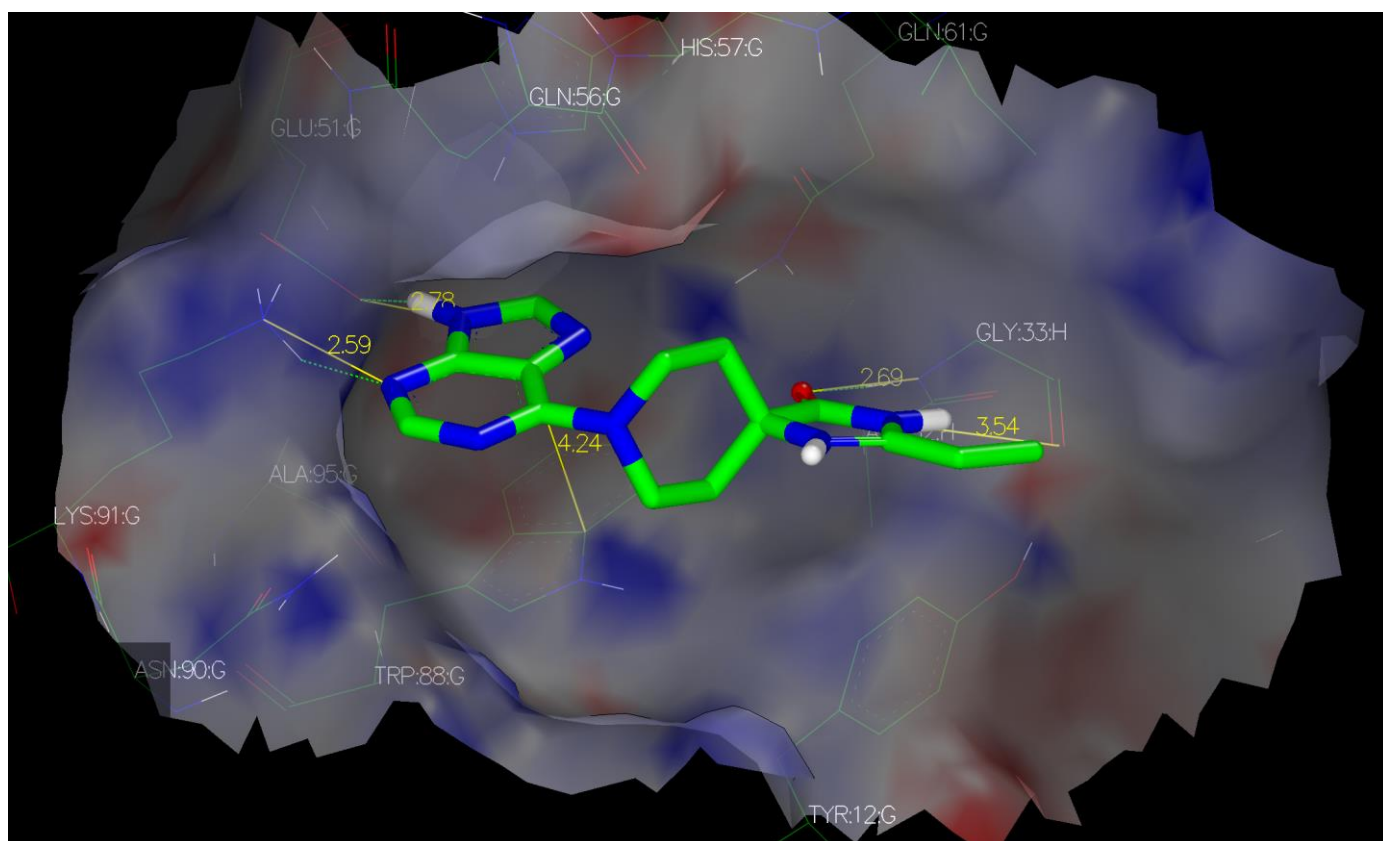
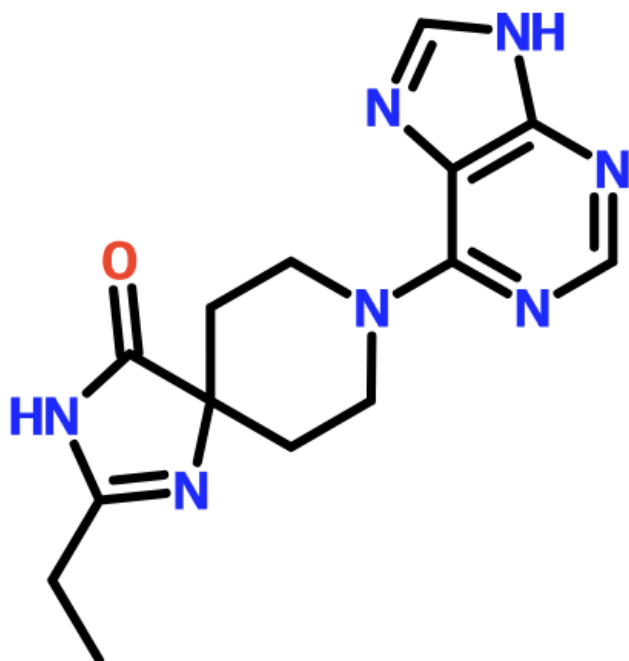


Figure 32 - 2D structure of MCULE-2430515474 with below the 3D pose visualized in VIDA FRED view. The yellow lines highlight the interactions with corresponding distances.

The third best molecule is MCULE-9486643068 and has a score of -12.44 (Fig 33). It shows possible interactions with Glu51, Lys91, Gln56 and Gln61. Again there is stacking possible with Trp88 and there is also another interaction present with the Trp88.

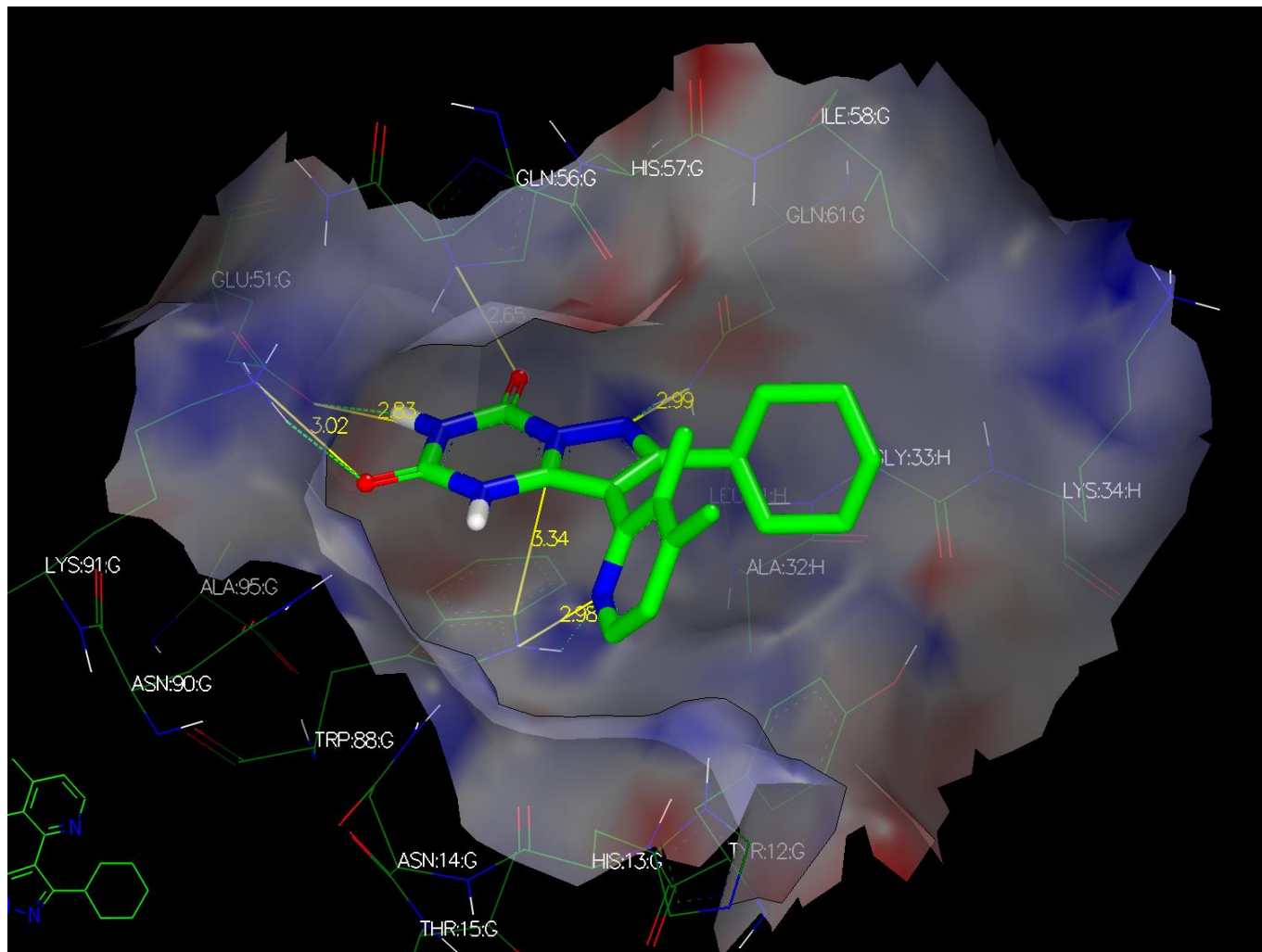


Figure 33 - 2D structure of MCULE-9486643068 with below the 3D pose visualized in VIDA FRED view. The yellow lines highlight the interactions with corresponding distances.

4. CONCLUSION

Non-carbohydrate molecules are clearly visible in the clusters and using the clustering they can be clearly separated from the carbohydrate molecules found. In the clusters it is clear that most molecules in the top 1000 reciprocal for both molecules consist of molecules containing cyclic groups which are rich of nitrogen atoms. These nitrogen groups can replace the interactions made by hydroxyl groups, which is shown in the VIDA visualization.

HYBRID 3D poses did show interactions we expected, as they represent interactions present in the original GM1-CTB binding. The non-carbohydrate ligands mostly did show logical poses, where the same interactions are visible that were present for the original ligands or GM1 galactose group. Instead of the OH rich carbohydrate structures, the interactions are made by the nitrogen rich cyclic groups in the non-carbohydrate molecules

Cyclic stacking interactions with the Trp88 seems to be possible with the rings we see present in the new-found ligands as well. From these clusters, molecules can be selected on either HYBRID, ROCS or the reciprocal score. It is obvious though that high-scoring ROCS molecules mostly contain carbohydrate-groups. Therefore, the HYBRID score could be seen as more leading for finding non-carbohydrate molecules from this set of results.

Actives did not all score well, but this can be due to the complexity of these molecules. Especially ROCS does not handle big molecules well, which will result in poor results for most of the actives. For ROCS actives scored between rank 141 and 32554168 for the 1PZJ run and between 5 and 35352752 for the 1EEI run. In HYBRID, not all actives scored well. For 1PZJ actives did score between rank 388211 and 18049007. 1EEI had the actives scoring between rank 189175 and 34314015. Both which can be explained by the non-validation of sugar-structures in OpenEye. Although, the original ligands 1PZJ and 1EEI ligands did score best in their respective runs of all actives, which is expected.

All in all after two virtual screening methods on two receptors, candidate molecules were found. These molecules do show explainable features which could mimic the carbohydrates present in the original GM1-mimics and MNPG(-derivative) ligands. Now the next step is to test ligands in in vitro for affinity.

5. DISCUSSION

First of all the targets we chose for HYBRID docking and ROCS are both MNPG-like ligands. This was done based on how well 1PZJ and 1EEI docked the actives, however the 1EEI and 1PZJ do not chemically and structurally represent all actives. Therefore you could want to dock with a bigger range of PDB's with more variety in bound ligands. Most actives however are carbohydrate-heavy molecules or are quite big and complex. OpenEye tools are not validated for these type of molecules and this can also explain that most GM1-mimics did not show good docking performance. However, if possible, it would be nice to do docking with a bigger variety of molecules as of now a big chunk of cholera ligands could not be used due to software limitations.

The option of multi-receptor docking in HYBRID could tackle this problem. With this HYBRID-mode, first similarity scores will be calculated between all receptor-bound ligands and database molecules before docking. This will lead to database molecules being docked in receptors containing ligands most similar to it, which means this receptor will most likely fit them better. This mode of HYBRID indeed showed better performance than single-receptor docking HYBRID. However, we found that the increase in running time was too much for the time available and we opted for two single-receptor runs using 1PZJ and 1EEI instead.

For clustering, Butina clustering was used for its speed, ease of use and the fact that it suited our use case. Butina does however return many singletons, which can be problematic when looking at bigger sets of molecules or other use cases. For this study, the separation of carbohydrates and non-carbohydrates and a quick view of molecule types was the goal, which it did correctly. For better clustering performance and looking at bigger molecule sets, other clustering methods can be considered. Examples are hierarchical clustering and UMAP clustering. In a study comparing Butina, hierarchical and UMAP clustering, UMAP came out as the best clustering method for that specific database. (Hernández-Hernández & Ballester, 2023)

The Reciprocal Rank Sum (RRS) method used in this study is also only usable when looking at smaller molecule sets. With the relatively small set cut-off used in this study RRS is usable, however when looking at very large number of molecules RRS becomes almost unusable due to the scaling of the reciprocal scores. When looking at huge datasets the reciprocal becomes too small to actually mark significant differences. The ease of use of the RRS however and also the option to expand easily by just adding another reciprocal score makes it very adaptable and usable when using set and relatively small cut-off values. Another remark is that the current cut-off is self-chosen. The top 1000 is an arbitrary value which can be or not be enough molecules to look at. Cut-offs can be chosen using algorithms and data-driven decisions, however earlier used algorithms did not work with our data and therefore we chose a cut-off by hand. However, more research could and should be done to other data-driven cut-off algorithms for future virtual screenings.

ROCS did show worse performance for the 1PZJ ligand, which could be explained by the complexity, length and flexibility of the molecule. The set of 40 conformers for OMEGA-ROCS runs could be seen as not enough to describe the flexibility of this molecule. Increasing the amount of conformers for ROCS however did show insignificant performance gains when compared to an increase of runtime. Still an

increase of conformer numbers could be considered for these complex molecules in the future.

Finally of lastly, the cholera toxin B receptor is a hard target to actually dock into, due to the shallowness of the binding site. The poor performance of actives in both ROCS and HYBRID runs do confirm this. However, the original ligands of both 1PZJ and 1EEI scored well in HYBRID.

6. FUTURE RESEARCH RECOMMENDATIONS

From the results obtained molecules can be selected for further lab and *in vitro* studies. Using different kind of cholera toxin essays the affinity and effectiveness of the candidate inhibitors can be tested. These can be an ELISA-type assay or an intestinal organoid swelling inhibition assay. (Haksar et al., 2019; Zomer-van Ommen et al., 2016)

Another recommendation is to possibly expand the virtual screening by multi-receptor docking like described in the discussion. Also, other screening methods can be applied like 2D screening methods, like 2D fingerprint screening (Duan et al., 2010). The nice part about the reciprocal sum is that, when using the same molecules, different virtual screening methods can be easily added to the already existing reciprocal sum. This makes this study easily expandable by adding different and new methods where ranking is possible.

7. REFERENCES

- Bajusz, D., Rácz, A., & Héberger, K. (2019). Comparison of Data Fusion Methods as Consensus Scores for Ensemble Docking. *Molecules*, 24(15), 2690. <https://doi.org/10.3390/molecules24152690>
- Baldauf, K. J., Royal, J. M., Hamorsky, K. T., & Matoba, N. (2015). Cholera Toxin B: One Subunit with Many Pharmaceutical Applications. *Toxins*, 7(3), Article 3. <https://doi.org/10.3390/toxins7030974>
- Bernardi, A., Arosio, D., & Sonnino, S. (2002). Mimicking Gangliosides by Design: Mimics of GM1 Headgroup. *Neurochemical Research*, 27(7), 539-545. <https://doi.org/10.1023/A:1020251428217>
- Bernardi, A., Carrettoni, L., Ciponte, A. G., Monti, D., & Sonnino, S. (2000). Second generation mimics of ganglioside GM1 as artificial receptors for cholera toxin: Replacement of the sialic acid moiety. *Bioorganic & Medicinal Chemistry Letters*, 10(19), 2197-2200. [https://doi.org/10.1016/S0960-894X\(00\)00428-5](https://doi.org/10.1016/S0960-894X(00)00428-5)
- Bernardi, A., Checchia, A., Brocca, P., Sonnino, S., & Zuccotto, F. (1999). Sugar Mimics: An Artificial Receptor for Cholera Toxin. *Journal of the American Chemical Society*, 121(10), 2032-2036. <https://doi.org/10.1021/ja983567c>
- Burch, R. M., Jelsema, C., & Axelrod, J. (1988). Cholera toxin and pertussis toxin stimulate prostaglandin E2 synthesis in a murine macrophage cell line. *The Journal of Pharmacology and Experimental Therapeutics*, 244(2), 765-773.
- Butina, D. (1999, juni 29). *Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets* (world) [Research-article]. ACS Publications; American Chemical Society. <https://doi.org/10.1021/ci9803381>
- Chen, A., Hu, T., Mikoryak, C., & Draper, R. K. (2002). Retrograde transport of protein toxins under conditions of COPI dysfunction. *Biochimica Et Biophysica Acta*, 1589(2), 124-139. [https://doi.org/10.1016/s0167-4889\(02\)00163-5](https://doi.org/10.1016/s0167-4889(02)00163-5)
- Cheng, S. H., Rich, D. P., Marshall, J., Gregory, R. J., Welsh, M. J., & Smith, A. E. (1991). Phosphorylation of the R domain by cAMP-dependent protein kinase regulates the CFTR chloride channel. *Cell*, 66(5), 1027-1036. [https://doi.org/10.1016/0092-8674\(91\)90446-6](https://doi.org/10.1016/0092-8674(91)90446-6)
- Cheshev, P., Morelli, L., Marchesi, M., Podlipnik, Č., Bergström, M., & Bernardi, A. (2010). Synthesis and Affinity Evaluation of a Small Library of Bidentate Cholera Toxin Ligands: Towards Nonhydrolyzable Ganglioside Mimics. *Chemistry – A European Journal*, 16(6), 1951-1967. <https://doi.org/10.1002/chem.200902469>
- Cormack, G. V., Clarke, C. L. A., & Buettcher, S. (2009). Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 758-759. <https://doi.org/10.1145/1571941.1572114>

Denning, G. M., Ostedgaard, L. S., Cheng, S. H., Smith, A. E., & Welsh, M. J. (1992). Localization of cystic fibrosis transmembrane conductance regulator in chloride secretory epithelia. *Journal of Clinical Investigation*, *89*(1), 339-349.

Duan, J., Dixon, S. L., Lowrie, J. F., & Sherman, W. (2010). Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *Journal of Molecular Graphics and Modelling*, *29*(2), 157-170. <https://doi.org/10.1016/j.jmgm.2010.05.008>

Field, M., Rao, M. C., & Chang, E. B. (1989a). Intestinal electrolyte transport and diarrheal disease (1). *The New England Journal of Medicine*, *321*(12), 800-806. <https://doi.org/10.1056/NEJM198909213211206>

Field, M., Rao, M. C., & Chang, E. B. (1989b). Intestinal electrolyte transport and diarrheal disease (2). *The New England Journal of Medicine*, *321*(13), 879-883. <https://doi.org/10.1056/NEJM198909283211307>

Ghazarian, H., Idoni, B., & Oppenheimer, S. B. (2011). A glycochemistry review: Carbohydrates, lectins and implications in cancer therapeutics. *Acta Histochemica*, *113*(3), 236-247. <https://doi.org/10.1016/j.acthis.2010.02.004>

Gill, D. M., & Coburn, J. (1987). ADP-ribosylation by cholera toxin: Functional analysis of a cellular system that stimulates the enzymic activity of cholera toxin fragment A1. *Biochemistry*, *26*(20), 6364-6371. <https://doi.org/10.1021/bi00394a009>

Haan, L. de, & Hirst, T. R. (2004). Cholera toxin: A paradigm for multi-functional engagement of cellular mechanisms (Review). *Molecular Membrane Biology*, *21*(2), 77-92. <https://doi.org/10.1080/09687680410001663267>

Haksar, D., de Poel, E., van Ufford, L. Q., Bhatia, S., Haag, R., Beekman, J., & Pieters, R. J. (2019). Strong Inhibition of Cholera Toxin B Subunit by Affordable, Polymer-Based Multivalent Inhibitors. *Bioconjugate Chemistry*, *30*(3), 785-792. <https://doi.org/10.1021/acs.bioconjchem.8b00902>

Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T., & Calderwood, S. B. (2012). Cholera. *Lancet*, *379*(9835), 2466-2476. [https://doi.org/10.1016/S0140-6736\(12\)60436-X](https://doi.org/10.1016/S0140-6736(12)60436-X)

Hawkins, P. C. D., Skillman, A. G., & Nicholls, A. (2007). Comparison of Shape-Matching and Docking as Virtual Screening Tools. *Journal of Medicinal Chemistry*, *50*(1), 74-82. <https://doi.org/10.1021/jm0603365>

Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., & Stahl, M. T. (2010). Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *Journal of Chemical Information and Modeling*, *50*(4), 572-584. <https://doi.org/10.1021/ci100031x>

Heggelund, J. E., Mackenzie, A., Martinsen, T., Heim, J. B., Cheshev, P., Bernardi, A., & Krengel, U. (2017). Towards new cholera prophylactics and treatment: Crystal structures of bacterial enterotoxins in complex with GM1 mimics. *Scientific Reports*, *7*, 2326. <https://doi.org/10.1038/s41598-017-02179-0>

- Hernández-Hernández, S., & Ballester, P. J. (2023). On the Best Way to Cluster NCI-60 Molecules. *Biomolecules*, 13(3), 498. <https://doi.org/10.3390/biom13030498>
- Hevey, R. (2019). Strategies for the Development of Glycomimetic Drug Candidates. *Pharmaceuticals*, 12(2), 55. <https://doi.org/10.3390/ph12020055>
- Kearnes, S., & Pande, V. (2016). ROCS-derived features for virtual screening. *Journal of Computer-Aided Molecular Design*, 30(8), 609-617. <https://doi.org/10.1007/s10822-016-9959-3>
- Kumar, V., & Turnbull, W. B. (2018). Carbohydrate inhibitors of cholera toxin. *Beilstein Journal of Organic Chemistry*, 14, 484-498. <https://doi.org/10.3762/bjoc.14.34>
- Lasserre, P. (2018, december 5). Molecular Similarity Search: A Simple but Powerful Drug Discovery Tool. *GSI Technology*. <https://medium.com/gsi-technology/molecular-similarity-search-a-simple-but-powerful-drug-discovery-tool-2b991d78c191>
- Lauer, S., Goldstein, B., Nolan, R. L., & Nolan, J. P. (2002). Analysis of cholera toxin-ganglioside interactions by flow cytometry. *Biochemistry*, 41(6), 1742-1751. <https://doi.org/10.1021/bi0112816>
- Lavecchia, A., & Di Giovanni, C. (2013). Virtual screening strategies in drug discovery: A critical review. *Current Medicinal Chemistry*, 20(23), 2839-2860. <https://doi.org/10.2174/09298673113209990001>
- McGann, M. (2012). FRED and HYBRID docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design*, 26(8), 897-906. <https://doi.org/10.1007/s10822-012-9584-8>
- Merritt, E. A., Sarfaty, S., Feil, I. K., & Hol, W. G. (1997). Structural foundation for the design of receptor antagonists targeting Escherichia coli heat-labile enterotoxin. *Structure*, 5(11), 1485-1499. [https://doi.org/10.1016/S0969-2126\(97\)00298-0](https://doi.org/10.1016/S0969-2126(97)00298-0)
- Minke, W. E., Roach, C., Hol, W. G. J., & Verlinde, C. L. M. J. (1999). Structure-Based Exploration of the Ganglioside GM1 Binding Sites of Escherichia coli Heat-Labile Enterotoxin and Cholera Toxin for the Discovery of Receptor Antagonists. *Biochemistry*, 38(18), 5684-5692. <https://doi.org/10.1021/bi982649a>
- Mitchell, D. D., Pickens, J. C., Korotkov, K., Fan, E., & Hol, W. G. J. (2004). 3,5-Substituted phenyl galactosides as leads in designing effective cholera toxin antagonists: Synthesis and crystallographic studies. *Bioorganic & Medicinal Chemistry*, 12(5), 907-920. <https://doi.org/10.1016/j.bmc.2003.12.019>
- Moss, J., Osborne, J. C., Fishman, P. H., Brewer, H. B., Vaughan, M., & Brady, R. O. (1977). Effect of gangliosides and substrate analogues on the hydrolysis of nicotinamide adenine dinucleotide by cholera toxin. *Proceedings of the National Academy of Sciences of the United States of America*, 74(1), 74-78. <https://doi.org/10.1073/pnas.74.1.74>
- Moss, J., Osborne, J. C., Fishman, P. H., Nakaya, S., & Robertson, D. C. (1981). Escherichia coli heat-labile enterotoxin. Ganglioside specificity and ADP-ribosyltransferase activity. *The Journal of Biological Chemistry*, 256(24), 12861-12865.

Moss, J., & Richardson, S. H. (1978). Activation of adenylate cyclase by heat-labile Escherichia coli enterotoxin. Evidence for ADP-ribosyltransferase activity similar to that of cholera toxin. *The Journal of Clinical Investigation*, 62(2), 281-285. <https://doi.org/10.1172/JCI109127>

Oliveira, T. A. de, Silva, M. P. da, Maia, E. H. B., Silva, A. M. da, & Taranto, A. G. (2023). Virtual Screening Algorithms in Drug Discovery: A Review Focused on Machine and Deep Learning Methods. *Drugs and Drug Candidates*, 2(2), Article 2. <https://doi.org/10.3390/ddc2020017>

Piccioletto, M. R., Cohn, J. A., Bertuzzi, G., Greengard, P., & Nairn, A. C. (1992). Phosphorylation of the cystic fibrosis transmembrane conductance regulator. *The Journal of Biological Chemistry*, 267(18), 12742-12752.

Podlipnik, Č., Velter, I., Ferla, B. L., Marcou, G., Belvisi, L., Nicotra, F., & Bernardi, A. (2007). First round of a focused library of cholera toxin inhibitors. *Carbohydrate Research*, 342(12), 1651-1660. <https://doi.org/10.1016/j.carres.2007.06.006>

Ramamurthy, T., Nandy, R. K., Mukhopadhyay, A. K., Dutta, S., Mutreja, A., Okamoto, K., Miyoshi, S.-I., Nair, G. B., & Ghosh, A. (2020). Virulence Regulation and Innate Host Response in the Pathogenicity of Vibrio cholerae. *Frontiers in Cellular and Infection Microbiology*, 10. <https://www.frontiersin.org/articles/10.3389/fcimb.2020.572096>

Ramos-Soriano, J., Niss, U., Angulo, J., Angulo, M., Moreno-Vargas, A. J., Carmona, A. T., Ohlson, S., & Robina, I. (2013). Synthesis, Biological Evaluation, WAC and NMR Studies of S-Galactosides and Non-Carbohydrate Ligands of Cholera Toxin Based on Polyhydroxyalkylfuroate Moieties. *Chemistry – A European Journal*, 19(52), 17989-18003. <https://doi.org/10.1002/chem.201302786>

Sethi, A., Sanam, S., & Alvala, M. (2021). Non-carbohydrate strategies to inhibit lectin proteins with special emphasis on galectins. *European Journal of Medicinal Chemistry*, 222, 113561. <https://doi.org/10.1016/j.ejmech.2021.113561>

T005 · Compound clustering. (z.d.). TeachOpenCADD. Geraadpleegd 24 juli 2023, van https://projects.volkamerlab.org/teachopencadd/talktorials/T005_compound_clustering.html

Tautomer—Wikidoc. (z.d.). Geraadpleegd 24 juli 2023, van <https://www.wikidoc.org/index.php/Tautomer>

Turnbull, W. B., Precious, B. L., & Homans, S. W. (2004). Dissecting the cholera toxin-ganglioside GM1 interaction by isothermal titration calorimetry. *Journal of the American Chemical Society*, 126(4), 1047-1054. <https://doi.org/10.1021/ja0378207>

Vrasidas, I., Kemmink, J., Liskamp, R. M. J., & Pieters, R. J. (2002). Synthesis and Cholera Toxin Binding Properties of a Lactose-2-aminothiazoline Conjugate. *Organic Letters*, 4(10), 1807-1808. <https://doi.org/10.1021/ol025909w>

Willett, P. (2013). Fusing similarity rankings in ligand-based virtual screening. *Computational and Structural Biotechnology Journal*, 5, e201302002. <https://doi.org/10.5936/csbj.201302002>

Yu, R. K., Tsai, Y.-T., Ariga, T., & Yanagisawa, M. (2011). Structures, biosynthesis, and functions of gangliosides—An overview. *Journal of Oleo Science*, 60(10), 537-544. <https://doi.org/10.5650/jos.60.537>

Zomer-van Ommen, D. D., Pukin, A. V., Fu, O., Quarles van Ufford, L. H. C., Janssens, H. M., Beekman, J. M., & Pieters, R. J. (2016). Functional Characterization of Cholera Toxin Inhibitors Using Human Intestinal Organoids. *Journal of Medicinal Chemistry*, 59(14), 6968-6972. <https://doi.org/10.1021/acs.jmedchem.6b00770>

8. SUPPLEMENTARY

All used codes and extra results can be found at

https://github.com/damianv98/choleratoxinb_screening/

This contains codes, property graphs of databases and receptor study results.