Development and Application of Virtual Screening Methods for G Protein-Coupled Receptors

by

Wallace K. Chan

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biological Chemistry)
in The University of Michigan
2018

Doctoral Committee:

Professor Yang Zhang, Chair
Professor Philip C. Andrews
Professor Heather Carlson
Associate Professor Mark Saper
Professor John R. Traynor

Wallace Kin Bon Chan

wallakin@umich.edu

ORCID ID: 0000-0001-5104-4159

To my wife

Acknowledgements

I would first like to thank my advisor, Dr. Yang Zhang, for his support throughout my training in graduate school. I initially entered his lab with no experience in programming, and to be quite frank, it was an incredibly daunting transition from wet-lab biochemistry to computational chemistry and bioinformatics. However, thanks to his patience, I eventually developed the proper skillsets in programming and became able to develop databases and algorithms to address chemical and biological problems. Given the scientific freedom that he gave, I was allowed to explore areas of research that I found infinitely fascinating and thus made the experience more enjoyable in the end. Moreover, I would like to thank him for giving me the opportunity to participate in grant writing. It is not in every lab that the principal investigators allow students into this process, but I believe that his doing so has helped me with a skill that will benefit me greatly in the future.

Next, I would like to thank my committee members for their unwavering support and guidance. I appreciate the efforts from Dr. Mark Saper, who proved invaluable as a mentor. Additionally, it was an honor working with Dr. Traynor, who kindly extended his assistance on numerous occasions with the opioid receptors. Given their late arrival to my committee, I very much appreciate and thank Dr. Heather Carlson and Dr. Philip Andrews for their willingness to come on board and provide their feedback. Lastly, even though they are no longer currently at the university, I would like to thank Dr. John Tesmer, Dr. Matthew Young, and Dr. Barry Grant for their help and suggestions at my committee meetings for the first few years of my training.

Many of my fellow lab mates in the Zhang lab have helped me greatly along the way, and without them, I would not have gotten nearly as far. Dr. Brandon Govindarajoo, who was still a PhD student when I first joined, helped me greatly with the transition, offering me a multitude of life advice, programming tips, and general camaraderie. Also, Hongjiu Zhang taught me invaluable lessons in programming, which greatly accelerated my ability to program. Without him, I do not believe that my programming would have been sufficient enough to make it through the program.

Dr. Jeffrey Brender was also there for helpful discussions about research, and despite his short time in the lab, I believe he made a lasting impact. I would also like to acknowledge Dr. Golam Mortuza and Dr. Jarrett Johnson for their friendship in the lab.

Apart from the PhD program, I have to acknowledge all of my friends that have been so important to me through these years. Dr. John Hyatt, my best friend since we were both 5 years old, has been there for me from almost the start. We've always been there for each other through all the rough patches in life and have managed to cultivate a long-term friendship that many people have never had. For that, I am greatly thankful to him (hoo ah!), as well as his mom (Aunt Elayne) and step-dad (Gordon Niessen). Also, I would like to acknowledge Justin and Alyssa McNally and Matthew DeMars, who have been some of the greatest friends, not just in Ann Arbor, but in my life. Dr. Alex Ninfa has been a great mentor and friend, and it was both a pleasure and honor to have been your GSI and music jam buddy through these years. Dr. Stephen Ragsdale was also an integral part of my graduate school life, providing guidance prior to my preliminary exam. Of course, music was involved, as well. Moreover, I will always enjoy the philosophical discussions I had with Alex Terzian about the human condition. Lastly, I greatly cherish my friendship with Alice Sano. There have been countless others out there that have made a mark on my life during my graduate school years, and I apologize if I left anyone out.

Last, but definitely not least, I would like to thank my parents, Louis and Rose Chan, for raising me with a high moral standard and providing me access to a good education. Without them, I do not think that I would be the man I am today. Additionally, I would like to acknowledge my sister, Jennifer, and her husband Aurelien. Above all, I would like to sincerely thank my wife, Mia Peng, for her steadfast support through the whole PhD program. Words cannot express my gratitude.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Tables

Abstract

G protein-coupled receptors (GPCR) constitute one of the largest family of transmembrane proteins that have been implicated in a multitude of diseases, including cancer and diabetes, and have been an important target in drug deve lopment. While experiment-based high-throughput screening for the unearthing of novel chemical compounds remains the *de facto* standard for drug discovery, virtual screening has been gaining acceptance as an important complementary method due to its high speed and low cost, which instead employs computers.

This dissertation is aimed at the development of virtual screening algorithms as applied to GPCR's, in addition to the construction of GPCR-related databases (GPCR-EXP, GLASS). MAGELLAN is a ligand-based virtual screening algorithm that makes inferences about what a GPCR would potentially bind based on sequence- and structure-based alignments. Building on top of this work, a sequential virtual screening pipeline combining MAGELLAN with AutoDock Vina was constructed for the discovery of novel, bifunctional opioids with mu opioid receptor (MOR) agonist and delta opioid receptor (DOR) antagonist activity.

In the process of developing the virtual screening algorithms, two GPCR-related databases were constructed to provide necessary data for the study. GPCR-EXP is a database of experimentally-validated and predicted GPCR structures. Important features include semi-manual curation of data, weekly updates, a user-friendly web interface, and high-resolution structure models with GPCR-I-TASSER, which many of the other GPCR-related databases lack. Additionally, GLASS database was developed in response to the absence of databases dedicated to GPCR experimental data. As a result, pharmacological data was pooled and integrated into a single source, resulting in over 500,000 unique GPCR-ligand associations; this made it the most comprehensive database of its kind thus far, providing the community with an accessible web interface, freely-available data, and ligands ready for docking.

MAGELLAN utilized pharmacological data from GLASS to infer from the ligands of sequence- and structure-based homologues what a target GPCR would bind. It was tested on two public virtual screening databases (DUD-E and GPCR-Bench) and achieved an average EF of 9.75 and 13.70, respectively, which compared favorably with AutoDock Vina (1.48/3.16), DOCK 6 (2.12/3.47), and PoLi (2.2). Lastly, case studies with the mu opioid and motilin receptors demonstrated its applicability to virtual screening in general, as well as GPCR de-orphanization. Subsequently, MAGELLAN was combined with AutoDock Vina into a novel, sequential virtual screen pipeline against both MOR and DOR to compensate for the weaknesses of each algorithm. Retrospective virtual screens against both MAGELLAN and AutoDock Vina were established for both receptors, and both methods were reported to have over-random discrimination between actives and decoys using the GPCR-Bench dataset.

In conclusion, structure (GPCR-EXP) and pharmacological data (GLASS) databases were constructed to provide users with a comprehensive source of GPCR data. Moreover, GLASS made it possible for MAGELLAN to be developed, providing it a rich source of experimental data. In return, this resulted in greater performance than competing algorithms. Lastly, a prospective sequential virtual screening pipeline was established for the discovery of novel bifunctional opioids, in which the models for both methods were validated to perform well. In future studies, cAMP and β-arrestin assays will be run on a subset of compounds from a prospective virtual screen in the hopes of discovering a novel opioid with reduced tolerance and withdrawal.

CHAPTER 1.
Introduction to Chemoinformatics and Bioinformatics

1. Brief Overview

In research today, the modern scientist will likely at some point of their career encounter chemical or biological problems when manual analysis of their data is impractical due to its large size. Situations such as these require the usage of computational methods in order to make it possible to sift through data. However, this has typically required some knowledge of programming in the past, which can unfortunately be a steep learning curve. Fortunately, many web servers are available nowadays to provide an accessible interface to an algorithm of interest for the scientist lacking experience in computational methods. As a result, computational methods for life science have seen an explosion in usage in the scientific community

Chemoinformatics is the application of computers in solving chemical problems, where a major application is in the *in silico* design of drugs. Within this field, cheminformatics is a specialized discipline that aims for efficiently working with enormous amounts of chemical data. For example, chemical similarity can be calculated between a reference compound and a virtual compound library in order to very quickly find chemically-similar compounds, as opposed to manually checking by eye. Whereas cheminformatics works with chemical data, bioinformatics was developed to address biological problems on a large scale. For example, the alignment of DNA or protein sequences have historically been done manually, but with the introduction of sequence alignment algorithms decades ago, a reference sequence can be aligned with a database of sequences in a blink of an eye.

Throughout this chapter, I will cover various aspects of chemoinformatics and bioinformatics, including computer representations of chemical compounds and proteins, chemical and biological databases, general bioinformatics, and virtual screening. Focus will be spent on concepts used in

the following chapters. Additionally, background information for G protein-coupled receptors (GPCR) will be presented. Finally, I will end with a segue into their role in my dissertation.

2. Computer Representation of Chemical Compounds and Proteins

Chemical compounds and proteins can be represented by a computer in a multitude of ways. Some formats are highly descriptive (i.e. PDB file format), while others are abstracted for fast calculations (i.e. molecular fingerprints). There are a dizzyingly large number of formats available, many frustratingly developed for proprietary purposes by companies, adding to the increasing lack of a standard. In this section, I will describe all of those used within this dissertation.

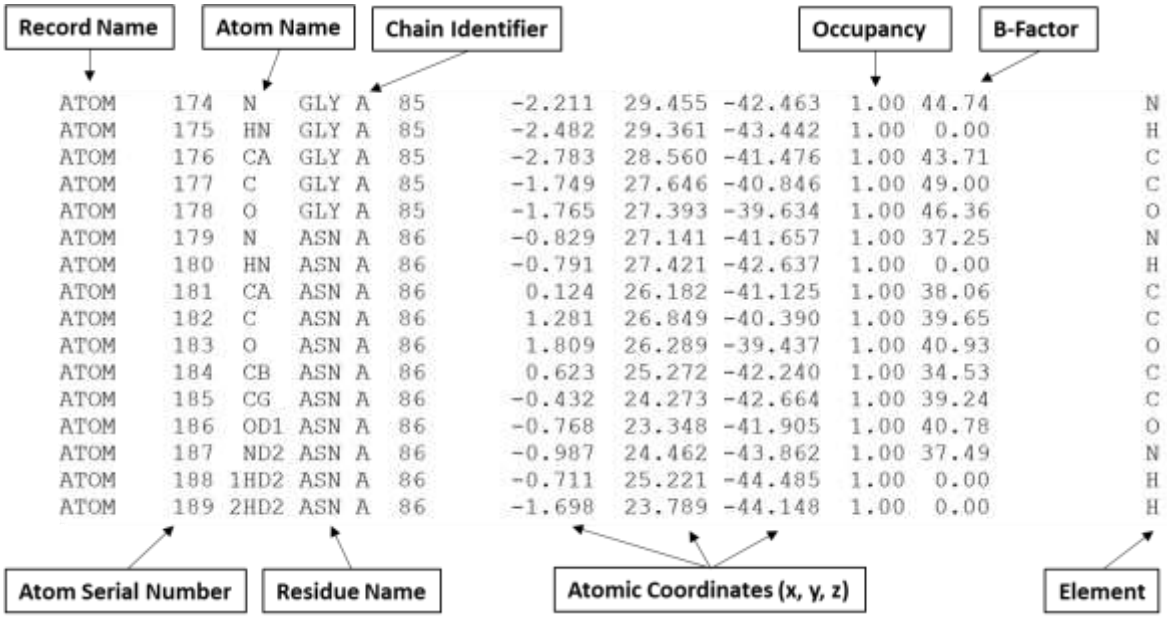


Figure 1.1 – Representative Portion of PDB File. The portions with the 'Record Name' of ATOM helps software understand the identity and location of atoms and therefore help correctly process relevant information from the file. The amino acids, Glycine (GLY) in position 85 and asparagine (ASN) in position 86, from this structure are shown.

2.1 PDB File Format

Most researchers in biochemistry will be fairly acquainted with the PDB file format, since it has been primarily used to describe the three-dimensional structure of proteins, DNA, and RNA. This file format was first conceived in 1976 as a means to help researchers exchange protein coordinates through a database.[1] Not surprisingly, its format has been revised and updated numerous times over the years. Essentially, a PDB file is a text file that contains various information about the

structure provided in specified ranges of columns. The file contains a variety of data, ranging from resolution and method used to solved structure to atomic coordinate specifications.

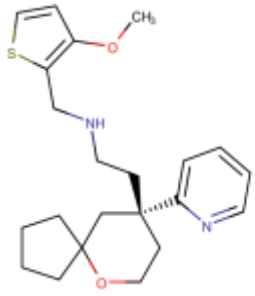
One of the most important pieces of information within the PDB file is the 'ATOM' record name. An example is shown in Figure 1.1 that depicts the coordinates for two representative amino acids from a PDB structure. Each line depicts a single atom in the structure. For example, the first line corresponds to the backbone nitrogen of Gly-85. Furthermore, the atomic coordinates of this atom (-2.211, 29.344, -42.463) are given so that whichever algorithm or molecular visualization software is used can correctly process this representation.

2.2 Line Notation

Line notation allows for the representation of a chemical compound using a string of ASCII characters. Despite looking rather odd to the untrained eye, they are completely human readable, and someone familiar with the format would be able to interconvert between it and the corresponding 2D chemical structure. Nowadays, these are primarily used for chemical database searching. The Simplified Molecular-Input Line-Entry System (SMILES) and International Chemical Identifier (InChI) formats are currently the most widely used.

SMILES strings were initially conceived in the 1980s as a means to make chemical compounds machine readable. Figure 1.2 shows the typical format of a SMILES string. Each letter represents an organic atom (B, C, N, O, P, S, F, Cl, Br, or I), and aromaticity is denoted with alternating equal signs (i.e. pyridine moiety: C4=CC=CC=N4). Additionally, rings are classified by including an opening and closing number (i.e. thiophene moiety: C1=C(SC=C1)). Single bonds are usually implicit, while double and triple bonds are represented as '=' and '#', respectively. The use of parentheses indicates branching, and stereochemistry is specified at chiral centers with '@'. SMiles ARbitrary Target Specification (SMARTS) strings were developed by the Daylight Chemical Information Systems as a robust extension of the SMILES string that provided expanded functionality, such as the ability to filter a compound database by substructure. However, one of the biggest drawbacks of this format is that there is no standard way in which to generate the SMILES string.[2] This can complicate database searching, especially when a compound of interest cannot be found due to this problem.

InChI strings were developed in 2005 by the International Union of Pure and Applied Chemistry (IUPAC) in response to the inconsistencies produced by SMILES strings.[3] Additionally, they were able to express more information than SMILES strings. An example is shown for oliceridine in Figure 1.2. All InChI strings start with 'InChI=', followed by the version number and an 'S', which corresponds to its standardization. Subsequently, there are six layers of information; the first layer is the most important and gives the chemical formula, atomic connections, and hydrogen atoms, while the others focus on other chemical aspects such as charge, stereochemistry, and isotopes. Also, it should be noted that the InChI format is conspicuously more difficult to read than SMILES. As seen from Figure 1.2, InChI strings can be long and unwieldy, so a shorter version was also developed as a companion to the original. Known as InChI keys, they are 27-character hashed versions of InChI strings that allow for extremely fast chemical database searches due to their reduced length. Nevertheless, a previous study has demonstrated that a single duplicate for the first 14 characters could theoretically occur 0.014% of the time in a database of 100 million compounds.[4] Given that most chemical databases have well below this number of chemical compounds, it can be assumed that a duplication will likely not occur. A drawback of using the InChI key is that it cannot be converted back to its respective InChI string, thus these two descriptors always need to be paired.



SMILES: COC1=C(SC=C1)CNCC[C@]2(CCOC3(C2)CCCC3)C4=CC=CC=N4

InChI: InChI=1S/C22H30N2O2S/c1-25-18-7-15-27-19(18)16-23-13-10-21(20-6-2-5-12-24-20)11-14-26-22(17-21)8-3-4-9-22/h2,5-7,12,15,23H,3-4,8-11,13-14,16-17H2,1H3/t21-/m1/s1

InChI Key: DMNOVGJWPASQDL-OAQYLSRUSA-N

Figure 1.2 – SMILES and InChI Strings for Oliceridine. MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

2.3 Molecular Fingerprints

Molecular fingerprints provide an abstraction of the chemical features of compounds into binary vectors. All have a fixed length for purposes of comparison and can be used to calculate chemical similarity mind-bogglingly fast. Though efficient, they likely have the least specific information

packed into its form. Over the years, various developments have aimed to squeeze as much information into a small space as possible.

Substructure key-based fingerprints consist of a predefined set of substructures, and the number of possible bits is defined by their number. One of the most commonly-used fingerprints of this type is Molecular ACCess System (MACCS), first developed by MDL Information Systems (formerly Molecular Design Limited) in 1979. Interestingly, they were initially intended for use in database searching as opposed to virtual screening,[5] which is the common method it is used for today. They come in two different flavors: one with 960 substructures, and the other with 166 of the most interesting substructures with corresponding SMARTS strings for drug discovery.[5] Not surprisingly, the latter is far more popular in usage. The principle in which these work is that each position in the fingerprint corresponds to a substructure. If the compound has the substructure in its chemical structure, then the bit will be set to '1'. Else, it would be set to '0'. A drawback to using these types of fingerprints is that they are usually relatively sparse in content, in that they will have mostly zeros, as typical molecules will have very few of the substructures.

Path-based fingerprints are constructed by analyzing every possible fragment in a molecule of a given linear path length, then hashing them all to produce the fingerprint. An example is given in Figure 1.3 for oliceridine, using a path length of 3. Occasionally, bit collisions occur when the same bit is assigned to two different fragments. However, this is not a common occurrence and can be reduced by increasing the fingerprint length. The Daylight fingerprint, developed by Daylight Chemical Information Systems (hence the namesake), is the most used out of all of these and typically consists of 1,028 bits.

Circular fingerprints are very similar to path-based fingerprints in that they are hashed from a collection of molecular fragments. However, their method of fragment analysis is not based on fragments generated in a linear path, but rather, the chemical environment centered around each atom within a certain radius. An example for oliceridine is given in Figure 1.4, where a radius of 2 was used. Here, fragments are generated by moving a certain radius away from a starting atom up until a diameter of 4, resulting in 3 fragments for the specified starting atom. The ECFP4

fingerprint is the industry standard, and not surprisingly, it has been shown to be among the best performing fingerprints in a recent benchmark that ranked diverse structures by similarity.[6]

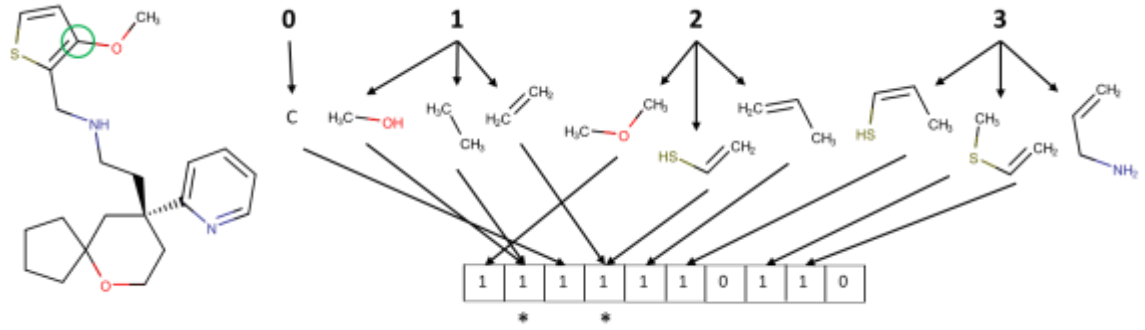


Figure 1.3 – Hypothetical 10-bit Path-Based Fingerprint for Oliceridine. A fingerprint with a path length of 3 was used in this example. Only fragments found from a single starting atom (green circle) are shown. The path lengths of the fragments (0, 1, 2, 3) are numbered in bold. The asterisks (*) denote where there are bit collisions. MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

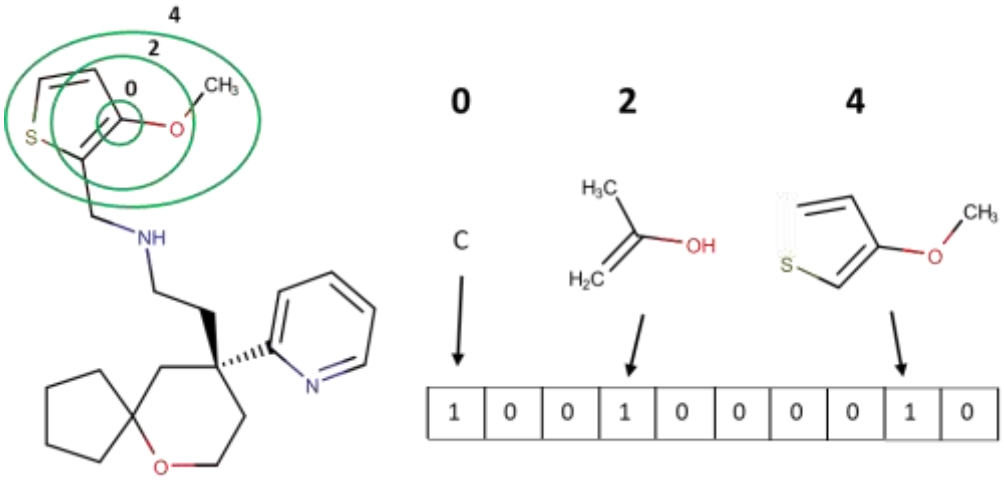


Figure 1.4 – Hypothetical 10-bit Circular Fingerprint for Oliceridine. A fingerprint with a radius of 2 was used in this example. Only fragments found from a single starting atom (innermost green circle) and onwards are shown. The diameters of the fragments (0, 2, 4) are numbered in bold. MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

2.4 Chemical Table Files

Another strategy for storing chemical information in a text file is chemical table file family of file formats. Originally developed by MDL Information Systems starting in the late 1970's,[7] they have become one of the most widely-used file formats, having been adopted by a vast majority of computational chemistry software. Of those in the family, I will focus upon the Structure-Data File (SDF) format. An example for monosodium glutamate is shown in Figure 1.5. In brief, the file

starts with three lines of a header block, which is mandatory but can be left empty if desired. This is followed by a counts line, which consists of specifications such as number of atoms, number of bonds, and so forth. The atoms block provides information about the coordinates and identity of the atom, while the bond block describes the connectivity between atoms. 2D coordinates are shown in the example, but 3D coordinates are also commonly used. The properties block denotes any existing charges or isotopes, as well as ending the molecular description. Up until this point, the file is essentially in Molfile format. Thus, SDF is different in that the subsequent associated data allow the inclusion of miscellaneous information not allowed in the main form, such as

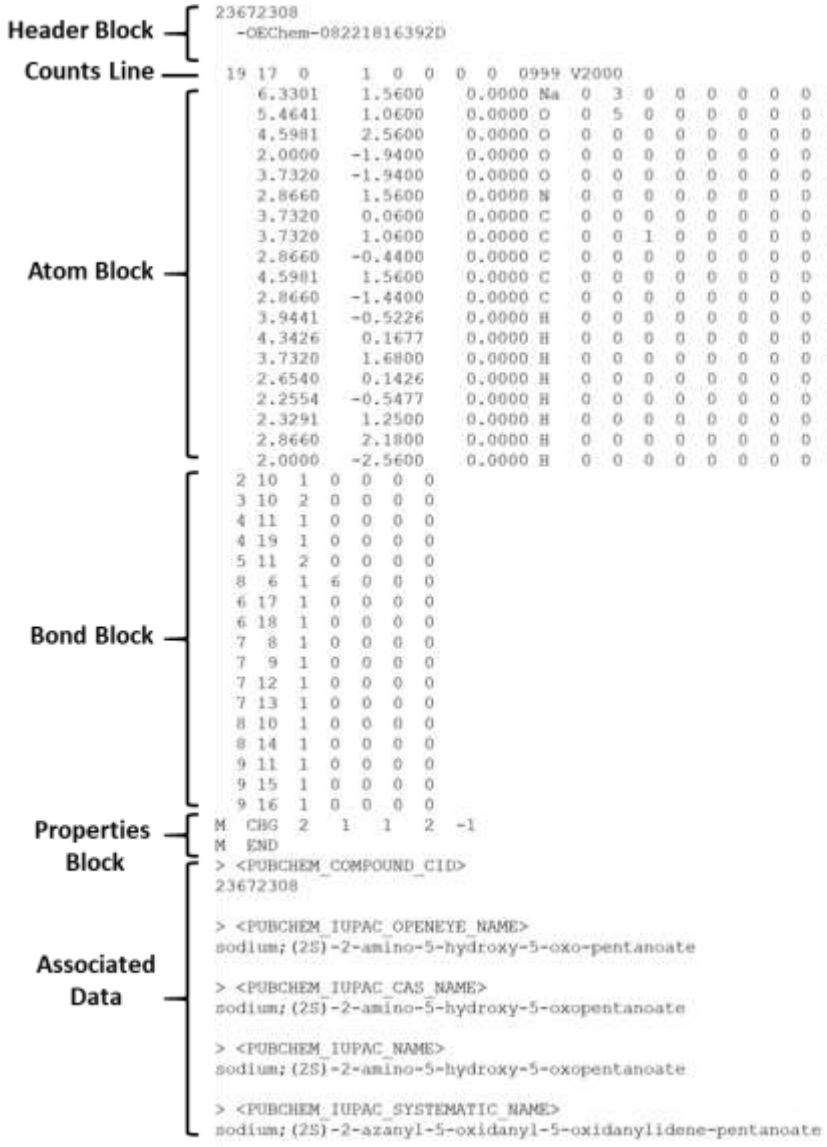


Figure 1.5 – Example SDF Format for Monosodium Glutamate.

IUPAC name and database identifiers. The tag for the data type is included inside angle brackets (‘<’, ‘>’), and the relevant data is placed on the line immediately following it.

Originating from the now-defunct Tripos, the Mol2 format has achieved a similar level of popularity and usage as the SDF format. An example of this format for benzene is given in Figure 1.6. Various aspects are almost identical to the SDF format, where various blocks are designated for counts, atom, and bond information, though with different column formatting. Moreover, each block is recognized starting with a record type indicator (i.e. @<TRIPOS>ATOM), followed by the corresponding data. Apart from these main record type indicators, there exists many others not available in SDF format, such as substructures and rotatable bonds.

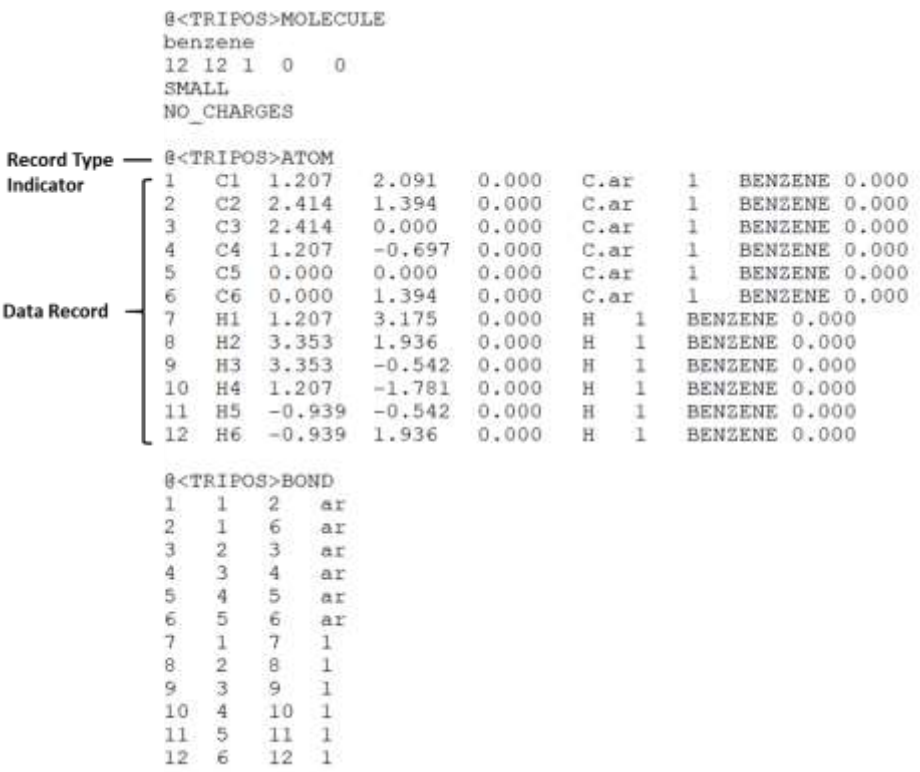


Figure 1.6 – Example Mol2 File Format for Benzene.

2.5 FASTA File Format

As a ubiquitous format for proteins, the FASTA file format is a text file containing a head followed by the primary structure. On an interesting note, its namesake stems from a legacy sequence alignment software of the same name.[8] An example for the mouse mu opioid receptor is given in Figure 1.7. The first line always begins with an angle bracket, ‘>’, followed with a personalized

description of the protein of interest. In the following lines, the primary structure is shown, starting from the N-terminus. Though not required, there are typically 80 characters per line; this is because old terminals back in the 1980's were only able to display this much text per line.

```
>sp|P42866|OPRM_MOUSE Mu-type opioid receptor
MDSSAGPGNISDCSDPLAPASCSPAPGSWLNLSHVDGNQSDPCGPNRTGLGGSHSLCPQT
GSPSMVTAITIMALYSIVCVVGLFGNFLVMYVIVRYTKMKTATNIYIFNLALADALATST
LPFQSVNYLMGTWPFGNILCKIVISIDYYNMFTSIFTLCTMSVDRYIAVCHPVKALDFRT
PRNAKIVNVCNWILSSAIGLPVMFMATTKYRQGSIDCTLTFSHPTWYWENLLKICVFIFA
FIMPVLIITVCYGLMILRLKSVRMLSGSKEKDRNLRRITRMVLVVVAVFIVCWTPIHIYV
IIKALITIPETTFQTVSWHFCIALGYTNSCLNPVLYAFLDENFKRCFREFCIPTSSTIEQ
QNSARIRQNTREHPSTANTVDRTNHQLENLEAETAPLP
```

Figure 1.7 – Example FASTA File Format for Mouse Mu Opioid Receptor.

3. Chemical and Biological Databases

As the amount of data available to the scientific community increased over time, there became a distinct need to catalogue and organize it so that it could be easily accessible. Truly, gone are the days of hours-long expeditions to the library in search of publications that may or may not have been helpful to the question at hand. Amazingly, there now exist public databases that index data anywhere from the primary structures of proteins to various experimental values of ligands for a given receptor. To list and survey them would be out of the scope of this section, so I will briefly mention some of the more important databases used for the current dissertation.

UniProt is the *de facto* standard source of information for proteins.[9] This database originated from the merging of data from European Bioinformatics Institute (EBI), Swiss Institute of Bioinformatics (SIB), and Protein Information Resource (PIR) into an entity known as the UniProt consortium. The most commonly-used portion of the database is referred to as UniProt Knowledgebase (UniProtKB), which is subdivided into Swiss-Prot and TrEMBL. The former collection of data is manually annotated and reviewed by scientists for their respective proteins, while the latter refers to those that are computationally annotated from genomic data. Not surprisingly, TrEMBL contains a far larger quantity of data than Swiss-Prot. Within Swiss-Prot, a multitude of information about a protein of interest is available, such as primary structure, post-translational modifications, function, subcellular localization, and known protein-protein interactions.

The Protein Data Bank (PDB) is the single largest repository for protein, DNA, and RNA structures solved by structural biologists.[10] It began as a united effort in the 1970's to provide the scientific community with protein structures coded into punch cards.[1] As the Internet came into fruition, it became possible to move the data onto an online platform for a higher throughput distribution. Thus, the first web-server for browsing the PDB was developed at Brookhaven National Laboratory in 1996.[11] With the explosion of solved structures starting in the 1980's, this resource became increasingly invaluable to life science researchers around the world.

Chemical databases vary greatly in their content, providing anywhere from pure chemical data to experimental data for ligands and proteins. However, the base content of these databases is always chemical in nature. First released in 2009, ChEMBL is arguably the most massive database for molecules with drug-like properties and biological activity.[12] As of the latest release (ChEMBL 24.1), the database contain 1,828,820 unique compounds corresponding to 12,091 targets and 15,207,914 activities from 69,861 publications, all manually annotated. A similar database founded over a decade earlier at University of California at San Diego is BindingDB,[13] which also contain a large number of manually-curated affinity data. However, it has less of a focus on membrane receptors than ChEMBL and more strongly emphasizes enzymes targets.[14] DrugBank is a chemical database whose topic of interest is information on drugs and their corresponding targets.[15] Another interesting database of note is Psychoactive Drug Screening Program's (PDSP) K_i database,[16] which houses a sizeable number of experimental affinities. A large portion of their data is dedicated to G protein-coupled receptors (GPCR). Also, the International Union of Basic and Clinical PHARmacology's (IUPHAR) Guide to Pharmacology is a chemical database that deals primarily with popular pharmacological targets, such as GPCR's and ion channels.[17] It is manually curated by experts, and only ligands that have been well characterized are included. In contrast, ChEMBL, BindingDB, and PDSP Ki are looser in their criteria for inclusion, where the binding mode or mechanism are largely unknown for most ligands. Lastly, PubChem is a pure chemical database maintained by the National Center for Biotechnology Information (NCBI),[18] containing approximately 93.9 million chemical compounds. Additionally, they have a gargantuan collection of bioactivity data from about 1.25 million high-throughput screening campaigns, each with several million values.

One of the most interesting aspects of chemical and biological databases is their interconnectivity to one another. In each chemical database, chemical compounds have a unique ID for identification purposes. For example, the ID for morphine in DrugBank is DB00295, while that for the same compound in ChEMBL is CHEMBL70. In many of the large databases, cross-references are provided so that other databases can be accessed for the same compound. Alternatively, ID mapping files are sometimes provided to facilitate the mapping between databases. Another important feature of most chemical databases is the utilization of chemical line notation, such as SMILES and InChI, for substructure or chemical similarity searching. Alternatively, the user can typically draw the molecule into a web applet, which would get translated into chemical line notation, as well. Finally, the data from these databases are all downloadable, which form the basis of the data used in many areas of chemoinformatics and bioinformatics research.

4. Concepts in Bioinformatics

The field of bioinformatics is concerned mainly with working on biological problems with computers when infeasible with manual human ability. The necessity for computers in biological problems first began when the first protein and nucleic acid sequences were acquired. One can only imagine the difficulties and tedium of having to compare multiple sequences by eye, thus having an algorithm compute the alignments proved to be the best tool for the task. From there, the field blossomed beyond sequences into the prediction of macromolecular structure of proteins, analysis of the regulation of gene and protein expression, and understanding of networks concerning protein interactions. Certainly, bioinformatics has grown into an important discipline in its own right and to cover it in its entirety would warrant a textbook. However, basic concepts used in this dissertation will be covered in this section as follows.

4.1 Sequence Alignment

Before the advent of sequence alignment algorithms, pioneering work in the analysis of the substitution of amino acids in the primary structure of proteins was performed by Margaret Dayhoff in 1978, which led to the development of the Point Accepted Mutation (PAM) matrix.[19] The premise of this was to check for the frequency of amino acid substitutions observed in nature among closely-related homologues. Phylogenetic trees were manually constructed for each of 71 families of proteins with at least 85% sequence identity. For each branch in the phylogenetic trees,

the number of mismatched amino acids and their identity were recorded. For all 20 amino acids, the propensity for one residue to mutate to another was evaluated, which led to a 20 x 20 frequency table. The values from this table were ultimately transformed into a log odds values and became what is known as a substitution matrix. Another attempt was made in redefining the substitution matrix by Jorja and Steven Henikoff in 1992, which resulted in the BLOcks SUbstitution Matrix (BLOSUM) matrix.[20] While Dayhoff's PAM matrix was based on sequences with high global similarity, the BLOSUM matrix was derived on blocks that consists of un-gapped regions of aligned primary structures. Similar to PAM, the propensities for the substitution of one residue for another were calculated and converted into log odds ratio. Because of how they were designed, each matrix has their own set of strengths. The PAM matrix is better for tracking the evolutionary origin of proteins, while the BLOSUM matrix is great for finding conserved domains.[21]

A seminal publication from Needleman and Wunsch in 1970 presented an algorithm for the alignment of amino acid sequences using dynamic programming. What this algorithm aims to find is an optimal alignment based on the global similarity of two given sequences. A matrix m x n is constructed, where m and n are the lengths of the first and secondary sequences, respectively. Each position in the matrix is scored by either insertion of a gap (up or left) or an alignment (diagonal). The former typically consists of a gap opening or gap extension penalty, depending on the current direction, while the alignment score will be taken from a scoring matrix, such as PAM or BLOSUM. The highest of the three sub-scores will be taken, and the direction will be recorded. After all possible paths have been computed, the algorithm backtracks starting from the last aligned residues for both sequences. The path with the highest scores is taken and stops when the first aligned residues are reached. The Smith-Waterman algorithm is a variation of the Needleman-Wunsch algorithm in that it performs a local sequence alignment as opposed to global. Here, negative scores are set to zero, while the traceback starts instead on the cell with the highest score and ends when zero is reached. As a result, this is how local alignment is made possible. Additionally, the Gotoh extension to the Needleman-Wunsch algorithm allowed for the favoring of long consecutive gaps, as opposed to a collection of short gaps, with the introduction of the affine gap penalty.[22] Though sequence alignment with dynamic programming is thorough and can yield optimal sequence alignments, its application to large databases with millions of sequences results in long computation times due to its need to compute every possibility in the alignment.

Thus, heuristic methods (i.e. something not optimal but sufficient for reaching the end goal) were developed to speed things up by orders of magnitude.

Basic Local Alignment Search Tool (BLAST) is a landmark heuristic sequence alignment algorithm developed in 1990 by various scientists at NCBI.[23] In brief, the reference sequence is broken down into *k*-letter words. Words matching a sequence in the database are then scored based on how well they match, then the high-scoring words are retained. If there is an exact match between a high-scoring word and the sequence, then this serves as a seed for an un-gapped alignment. If the alignment scores above a threshold value, then it is considered a match. Using this method, a large database of sequences can be aligned with a reference sequence in a short amount of time. A variant of BLAST called Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) was developed in 1997 again at NCBI, which included some of the same scientists from the original version.[24] This algorithm made it possible to discover distant relationships between proteins. It operated by first running BLAST on a reference sequence against a sequence database. Any proteins sequences found that are above a threshold are retained and used to construct a Position-Specific Scoring Matrix (PSSM); this is a L x 20 scoring matrix based on the amino acid conservation information for each residue of the reference sequence for each of the 20 amino acid residues among the aligned sequences. From the second iteration onwards, the PSSM is used for scoring, and any new protein sequences found above a threshold are retained for the re-generation of the PSSM. This is typically repeated 3 times or until convergence. Overall, both BLAST and PSI-BLAST have revolutionized bioinformatics, providing an efficient tool for the analysis of a large amount of sequence data.

In contrast with pairwise sequence alignment methods such as BLAST, multiple sequence alignment methods operate by comparing a set of homologous protein sequences, such as GPCR's, and can thus reveal conserved motifs or residues. Importantly, it should be noted that an all-against-all comparison is made, where every sequence is aligned to each other for the optimal alignment. Motifs have been found using multiple sequence alignment, such as GCM motif for DNA-binding activity.[25] Clustal Omega is one of the best algorithms freely available today for this purpose, being able to efficiently, accurately align hundreds of thousands of sequences in only a few hours.[26]

4.2 Structure Alignment

There exist numerous metrics for the evaluation of the superposition of two distinct structures. Root-Mean-Square Deviation (RMSD) is frequently-used to measure the average distance between the atoms of two superposed protein structures, given as follows:

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{i,x} - w_{i,x})^2 + (v_{i,y} - w_{i,y})^2 + (v_{i,z} - w_{i,z})^2)}$$

where there are *n* points corresponding to proteins *v* and *w* in three dimensions, *x*, *y*, and *z*. During the structure alignment, this value is minimized to represent the best superposition in three-dimensional space. For many applications, RMSD is usually sufficient for general structural comparison. However, deficiencies in this metric arise when local errors in structure affect the overall score. For example, two proteins with a similar fold may visually superpose well, but the RMSD may be very high due to equal weight on all pairwise alignments. Furthermore, comparisons between large proteins will result in higher RMSD values, which can be misleading. As a result, the Template Modeling (TM)-score was developed by Yang Zhang in 2004, originally for the assessment of protein structure template quality in threading.[27] The equation is shown as follows:

$$TM - score = max\left[\frac{1}{L_{target}}\sum_{i}^{L_{aligned}}\frac{1}{1 + (\frac{d_i}{d_o(L_{target})})^2}\right]$$

where L_{target} and $L_{aligned}$ are respectively the lengths of the target protein and aligned region, d_i is the distance between the *i*th pair of residues, and $d_o(L_{target}) = 1.24\sqrt[3]{L_{target} - 15} - 1.8$ is a distance scale for normalization. During a structure alignment, this value is maximized; the range of the score is between 0 and 1, with 1 being a perfect match. A TM-score above 0.5 denotes that the pair of proteins likely belong to the same fold.[28]

Structure has been previously shown to be more conserved than sequence.[29] Though it is well-known that homologous proteins can be found by sequence alignment, this is not always the case for proteins with distant homologues with the same fold, due to low sequence identity. Thus, TM-align was written as a means to generate accurate, fast structural alignments for different proteins.[30] In brief, initial alignments are first made based on features such as secondary structure, where they are used to generate the first superposition. Subsequently, the first rotation is made based on these initial alignments. This process is iterated until the TM-score is maximized.

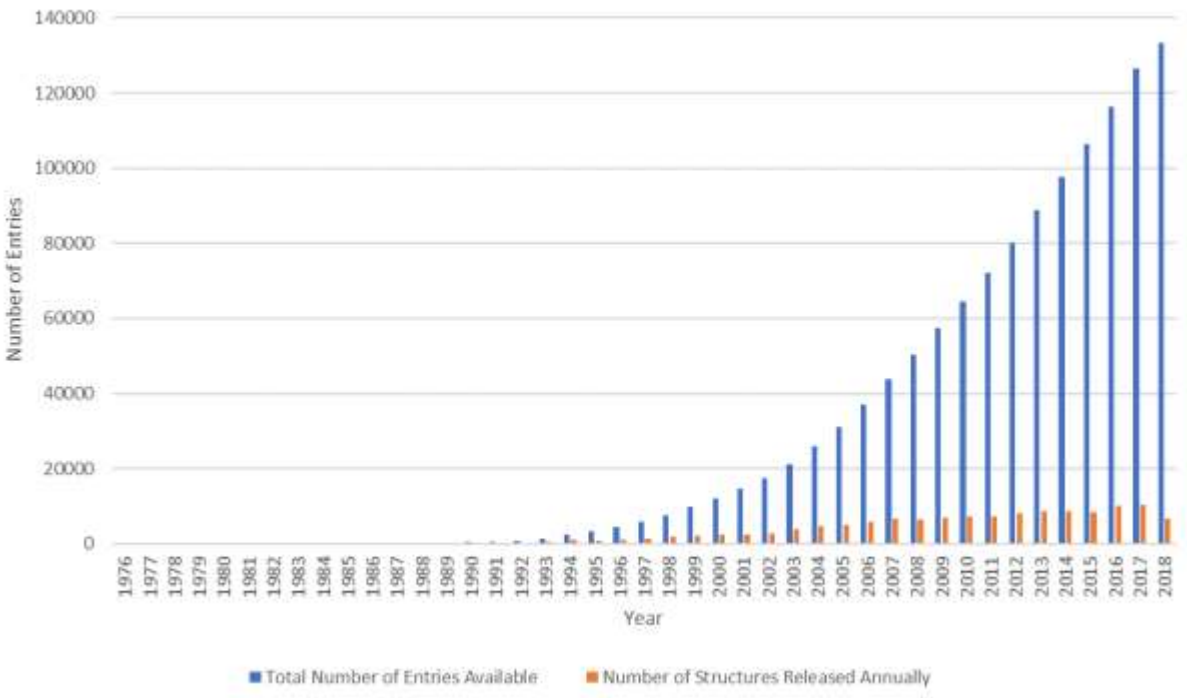


Figure 1.8 – PDB Structure Statistics for Proteins. Data was acquired from the PDB web server: https://www.rcsb.org/stats/growth/protein in August 2018.

4.3 Structure Prediction

Structural bioinformatics is a branch of bioinformatics concerned with the analysis and prediction of macromolecular structure. Despite the explosion in availability of protein structures (Figure 1.8), the majority of proteins within the human proteome (~20,000) have yet to be solved. Thus, computational models are typically generated for when the experimental structure does not exist. Three families of structure prediction exist for proteins: 1.) *ab initio* modeling, 2.) comparative or homology modeling, and 3.) threading or fold recognition. *Ab initio* methods seek to build protein

structure models from scratch. Some groups have approached this problem with physics-based approaches, such as molecular dynamics simulations. For instance, one study utilized replica-exchange molecular dynamics simulations on a set of 9 small proteins from the PDB;[31] 8 of the 9 structures folded correctly, though the experiment took about 6 months to run. Other strategies include using reduced models with only the peptide backbone and side chain centers of mass in folding simulations. Algorithms such as QUARK[32] and ROSETTA[33] are able to produce reasonable models, but unfortunately, the resolution of the structures are usually not high enough for in-depth analysis. Additionally, typically only small proteins can be folded, due to the high computational costs. Comparative modeling methods, such as MODELLER,[34] operate under the notion that sequence similarity implies structural similarity.[35] Homologous proteins with solved structures are as used as input for modeling, and when they are below 30% sequence identity, the accuracy of the models takes a hit and may end up with an entirely different fold.[36] Therefore, the usage of this methodology is limited in many cases, when a good homologue with a solved structure is not available. In response to this, fold recognition methods aim to overcome this drawback by selecting templates for modeling through fold-level homology. This is made possible by the prediction that that there is a limited number of folds found in nature.[37] Furthermore, it is likely that a correct fold will be selected a majority of the time, as there are already approximately 1,300 folds currently known. Numerous algorithms have been developed in the same vein, such as HHpred,[38] Phyre,[39] and MUSTER.[40]

Iterative Threading ASSEmbly Refinement (I-TASSER)[41-43] is a composite protein structure prediction algorithm that begins with secondary structure prediction, fold recognition, and replica-exchange Monte Carlo simulations. All that is required as input is the primary structure of the protein of interest. Following the first simulation, centroids are selected from the clustering of structural decoys, then subjected to a second round of folding simulations. After the decoys are clustered again, the lowest-energy structures are selected for refinement through the optimization of hydrogen bond networks. I-TASSER has been consistently ranked as the top method in Critical Assessment of Structure Prediction (CASP), an international benchmark for structure prediction algorithms, and has also generated structure models for hundreds of thousands of proteins submitted by researchers worldwide. A variant of this algorithm is GPCR-I-TASSER,[44] which was specifically designed to predict the structure of GPCR's. When no suitable template is found, the

transmembrane domain undergoes *ab initio* folding. Experimental restraints from GPCR-RD[45] guide the folding simulations, while a membrane repulsive potential keeps non-TM domains from being inserted into the virtual cell membrane.

4.3 Function Prediction

Another problem in bioinformatics is the assignment of biological function to an uncharacterized protein with a known sequence. This can be accomplished in a variety of ways, including inference from sequence and structure. One algorithm, COFACTOR[46-47], combines sequence, structure, and protein-protein interaction into a consensus prediction for Gene Ontology (GO) terms[48], which is standardized vocabulary for the annotation of proteins. Another algorithm, COnsensus ApproaCH (COACH)[49], utilizes experimental protein-ligand structure data from BioLiP[50] to predict the binding site of the query protein. In essence, it is a meta-algorithm that employs other state-of-the-art binding site prediction software, such as COFACTOR[46], FINDSITE,[51] and ConCavity.[52] Additionally, TM-STE and S-SITE, respectively employing structure- and sequence-based methods for the detection of binding pockets, were developed specifically for COACH. Altogether, the top-scoring predictions from each algorithm are fed into a previously-trained support vector machine model, producing a final consensus prediction. Continuous Automated Model EvaluatiOn (CAMEO), similar to CASP, is a community-wide effort to benchmark the accuracy of protein structure prediction servers. It also contains a section dedicated to ligand binding site predictions in proteins, in which COACH consistently performs better than its competitors.

5. Concepts in Chemoinformatics

Similar to bioinformatics, chemoinformatics is concerned with the analysis and processing of large amounts of data, though affiliated with chemistry. In this section, various concepts in chemoinformatics will be introduced to provide a glimpse into the processing of chemical data with computers.

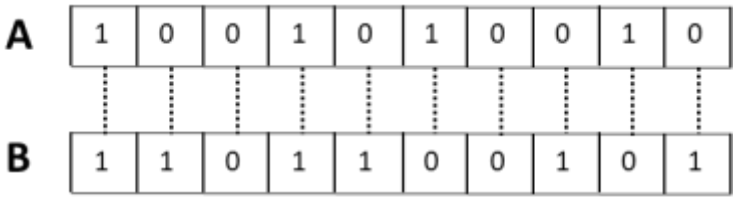
5.1 Chemical Similarity

Molecular fingerprints, as described previously in the chapter, are most often used in the calculation of chemical similarity. As a brief refresher, they are composed of a fixed-length string of bits; the presence of ‘1’ in a position denotes the presence of a chemical fragment, while ‘0’

denotes its absence. The simplicity of this form allows for the possibility of dreadfully fast calculations. A multitude of similarity metrics are available, but the Tanimoto coefficient has proven to be among the best and therefore has been most in use.[53] The equation is shown as follows:

$$Tanimoto\ Coefficient = \frac{c}{a + b - c}$$

where a is the number of bits in the first molecule, b is the number of bits in the second molecule, and c is the number of shared bits between the two molecules. A visual representation for calculation of the Tanimoto coefficient is given in Figure 1.9 for hypothetical 10-bit fingerprints. Only the same type of molecular fingerprint can be compared among molecules and mixing different types will lead to erroneous results.



$$TC = \frac{c}{a+b-c} = \frac{2}{4+6-2} = 0.25$$

Figure 1.9 – Sample Tanimoto Coefficient Calculation. Hypothetical 10-bit fingerprints are given for molecules **A** and **B**. 2 bits are shared between the two molecules, while molecules **A** and **B** respectively have 4 and 6 bits in their form. The Tanimoto coefficient (TC) is calculated to be 0.25, which means that there is 25% similarity between the two molecules.

5.2 Cluster Analysis

In many chemoinformatic applications, it is often necessary to group the data by some quantitative means for classification purposes. This is known as cluster analysis or clustering. Frequently-occurring chemotypes can be revealed this way and thus provide insight into a collection of chemical compounds. In many of the applications throughout this work, various clustering algorithms were employed to isolate groups of chemically-similar compounds.

The Taylor-Butina algorithm[54-55] is a non-hierarchical, unsupervised clustering method. For every chemical compound in a set, the chemical similarity is calculated using the corresponding

molecular fingerprints. Neighbors are acquired based on a similarity threshold, which is user defined, and become members of the cluster if they meet the threshold. Subsequently, the compounds (centroids) are sorted in descending order based on the number of members they have. Starting from the centroid with the largest number of neighbors, an exclusion sphere is set that flags the members of the cluster, disallowing them from becoming a centroid or a member of another cluster. Chemical compounds with no neighbors are referred to as singletons, as they bear no chemical similarity with any other in the set. Moreover, every cluster is guaranteed to have a neighbor at least the similarity cutoff of the centroid.

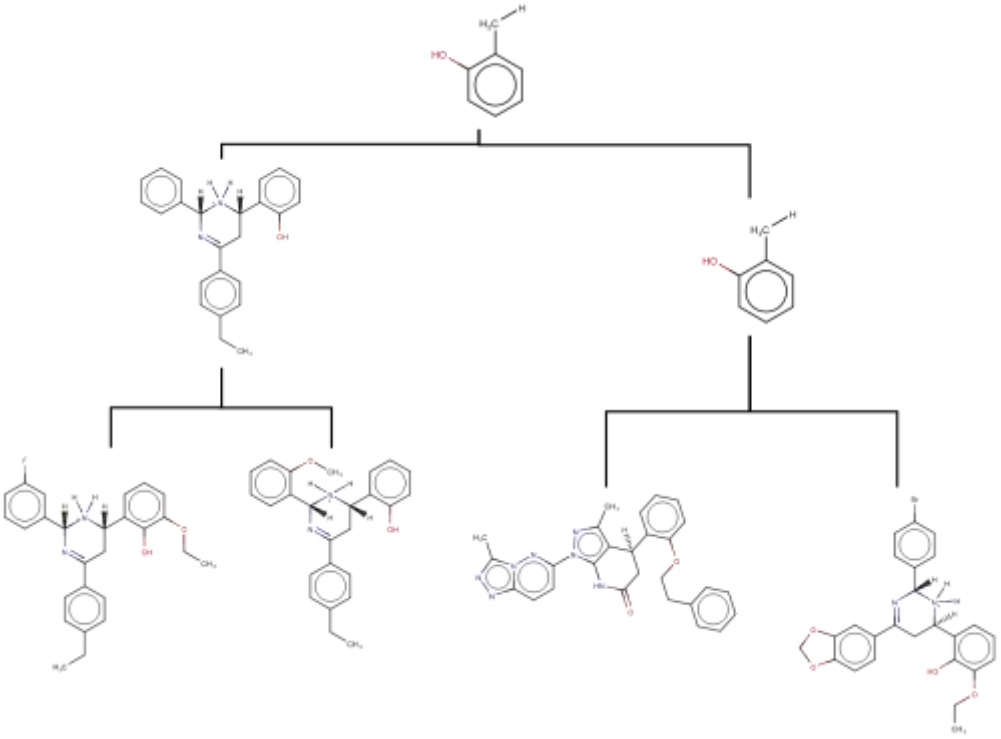


Figure 1.10 – Clustering with LibMCS. Three levels of clustering are shown, with the starting chemical compounds being at the first level (bottom). MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

LibMCS from Chemaxon is a hierarchical clustering algorithm rooted in chemical substructure. Though the algorithm is proprietary, the method is straightforward, and a description is provided by the company website.[56] First, chemical compounds are decomposed to chemical graphs, and clusters are formed where the compounds share the same substructure. This process is then repeated for further levels of clustering for each cluster until a single common substructure is acquired or until a size threshold for a substructure is reached. This ultimately results in a dendrogram, such as that shown in Figure 1.10. Indeed, this is a powerful method that can be used

to group sets of chemical compounds with the same substructure and is useful because they likely have similar bioactivities. Furthermore, there is a greater level of user specification as to which clusters at which level to analyze, as opposed to a single clustering output generated from methods like the Taylor-Butina algorithm. Overall, there is no best clustering method, and it all boils down to which one is suitable for the question at hand.

5.3 Molecular Docking

Though not technically a chemoinformatic technique, the abundant results from docking-based virtual screens are regularly subjected to processing with chemoinformatics. Molecular docking is a method used in computational chemistry to predict how a ligand binds with a receptor through scoring functions. Prior knowledge of the binding site is typically required in order to specify the area to be examined. Protocols for most docking programs start with adding hydrogens and partial charges to both the receptor and the compounds. This is then followed by the docking algorithm doing a conformational search of the most favorable ligand pose, which is evaluated with a scoring function at each step. Subsequently, the top poses are generated for the user, who can then visually check with a molecular viewer. Additionally, the final scores for each model are also given.

There have been dozens of docking software developed over the years, and each has approached docking in a different way. Some of the major differences between these are: 1.) the search algorithm, 2.) scoring function, and 3.) flexibility of ligand or receptor. Among the top methods employed for conformational searches are the Lamarckian genetic algorithm (AutoDock[57]), genetic algorithm (GOLD[58]), local search global optimizer (AutoDock Vina[59]), ant colony optimization, (PLANTS[60]), anchor-and-grow (DOCK 6[61]), and exhaustive search (Glide[62-63]). Though these strategies differ greatly in their search algorithms, their basic premise remains the same; that is, they aim to achieve the most favorable ligand pose. To do so, a scoring function must be calculated at each step of conformational sampling to evaluate the pose. Many of the ones used currently are physics-based force fields that approximate the binding energy of the ligand pose in the binding site. For example, the scoring function from DOCK 6 simply uses van der Waals and electrostatics terms for computational efficiency.[64] Various others take other physical terms into account, such as hydrogen bonding, desolvation, and hydrophobic contributions.[58] Additionally, there exist empirical scoring functions, which estimate the binding energy using a set of weighted

energy terms, and knowledge-based scoring functions, which utilize energy potentials derived from experimentally-solved structures.[65-66] Finally, there is the option of how to treat the receptor and ligand during docking. Most software packages make the receptor rigid because of the computational rigor involved in sampling the receptor conformation. However, some given an option to make certain side chains of the receptor flexible, such as AutoDock Vina[59] and GOLD.[58] Schrödinger has an induced-fit docking protocol that allows for both ligand flexibility and conformational changes in the binding site, though its application to virtual screening of a large number of compounds is limited due to its computational intensity. Most of the earliest docking methods, such as the original DOCK,[67] treated the ligand as a rigid body in order to find molecules with shape complementarity to the binding site. Nevertheless, this methodology is limited by the conformation of the molecule being docked, whereas vastly different conformations could be observed in reality. Therefore, multiple conformers would have to be generated for docking, making shape matching methods, such as ROCS from Open Eye Software, a far more attractive option. Modern docking algorithms all treat the ligand as flexible, allowing for it to find its most optimal pose in the binding pocket of a receptor.

5.4 Virtual Library and Benchmark Design

It is important that the contents of a virtual library meet the criteria for what a virtual screen is aiming. To screen a library of completely random compounds would not be sense in a drug discovery setting, as a clear majority would likely not fit the mold of what a typical drug would be like. Thus, a virtual library must be fashioned around a certain set of properties corresponding to a research question. As a prime example for this concept, ZINC database[68] is a database of commercially-available compounds available for virtual screening. Multiple subsets of data have been pre-compiled, where the chemical compounds have been selected using various property filters. For example, the drug-like subset is filtered by Lipinski's rule of five, which evaluates a compound based on a set of chemical and physical properties;[69] in order to be druglike, it must have: 1.) no more than 5 hydrogen bond donors, 2.) no more than 10 hydrogen bond acceptors, 3.) molecular weight of between 150 and 500 Daltons, and 4.) an octanol-water partition coefficient of not greater than 5. In another subset, the compounds are filtered based on rules defined by Teague *et al* for lead-like compounds,[70] where they pass if they are: 1.) between 250 and 350 Daltons in molecular weight, 2.) an octanol-water partition coefficient of no greater than 3.5, and

3.) no more than 7 rotatable bonds. In another study involving the serotonin receptor, ZINC database was filtered for compounds containing one or more aliphatic nitrogens, a characteristic feature found in most serotonin receptor ligands.[71] The design of these types of focused virtual libraries is required for the success of a virtual screen in an experimental setting.

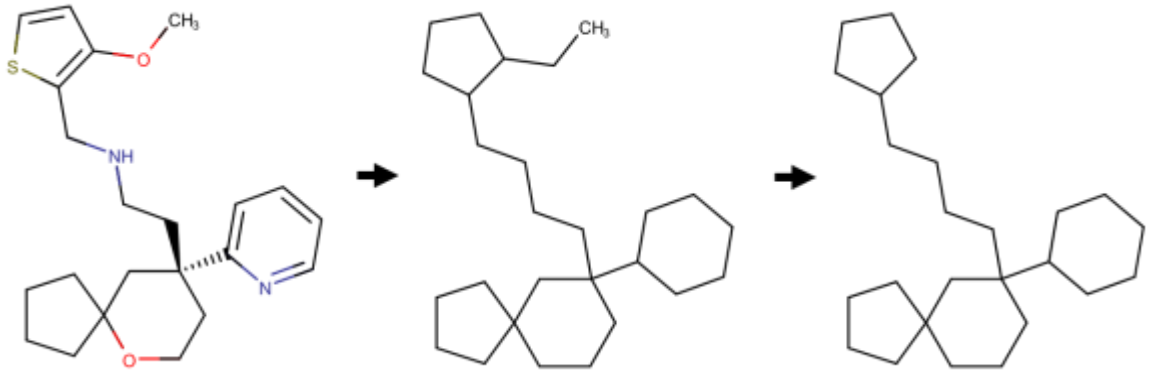


Figure 1.11 – Decomposition of Oliceridine into a Bemis-Murcko Framework. First, all non-carbon atoms are converted to carbon, and aromaticity and stereochemistry are removed. Then, all side chains not linking to another ring are removed. MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

In chemoinformatics, benchmarks datasets are used to provide a gold standard for the comparison of the performance of algorithms. The premise involves testing the ability an algorithm in distinguishing known active compounds from among a sea of decoy molecules. The Directory of Useful Decoys (DUD)[72] and its successor, DUD-Enhanced (DUD-E),[73] will be used as examples for successful design with docking benchmarks. For every receptor, the original study that developed DUD generated 33 decoys per active compound that were chemically similar but topologically different. However, a few drawbacks were found for this method. First, several chemotypes were overrepresented in many data sets.[74] Second, some of the sets had a very low number of active ligands, indicating the need for more.[73] Last, it was observed that the net formal charge of was imbalanced between the actives and decoys.[75] Thus, DUD-E was designed to address all of these issues from the first variant. Actives were now drawn through ChEMBL database in bulk and filtered by bioactivity, solving the deficiency in active compounds for some sets. Furthermore, more attention was drawn to balancing out charged compounds between actives and decoys. Most importantly, overrepresentations of certain chemotypes was eiminated by the clustering of Bemis-Murcko frameworks.[76] An example of a decomposition into a Bemis-Murcko

framework is given in Figure 1.11. Actives having the same framework were clustered together, and based on a set of rules, a certain number were chosen from each cluster so that there were between 100 and 600 for each set.

5.5 Virtual Screening

High throughput screening is an essential methodology in the pharmaceutical industry for the screening of chemical libraries for hits against a drug target. Due to its being costly, time-intensive, and laborious, virtual screening (or *in silico* screening) has emerged as a complementary strategy to reduce the chemical search space and to prioritize hits for experimental validation. Virtual screens can typically be categorized as ligand based and structure based, depending on what algorithm is used; the former utilizes pure chemical information in its search process, whereas the latter uses structural information to determine how well a compound would bind. As with any method, there are advantages and disadvantages to each. Ligand-based methods, such as chemical similarity, are computationally inexpensive and can screen millions of compounds within a short time but have the drawback of being biased towards the known ligands used to build the model. Conversely, structure-based methods, such as molecular docking, inherently have no bias, but they are extremely computationally expensive. A trend in recent years has culminated in the combination of these methods to address their respective shortcomings.[77]

Validation of virtual screening methods can be categorized as retrospective or prospective. In retrospective methods, the screening is usually performed where a certain amount of known active compounds and inactive compounds or decoys are all scored and ranked. The goal here is to try and get as many active compounds into the top-ranking portion of the list as possible. A typical metric for evaluation is the enrichment factor of the top 1%, given as follows:

$$EF_{1\%} = \frac{N_{act}^{1\%} / N_{select}^{1\%}}{N_{act} / N_{tot}}$$

where N_{act} and N_{tot} are the total numbers of the active and all compounds, respectively. $N_{act}^{1\%}$ and $N_{select}^{1\%}$ are, respectively, the numbers of active ligands and the number of all candidates in

the top 1% of the ranked database. An illustration of different results achievable in a retrospective virtual screen is given in Figure 1.12.

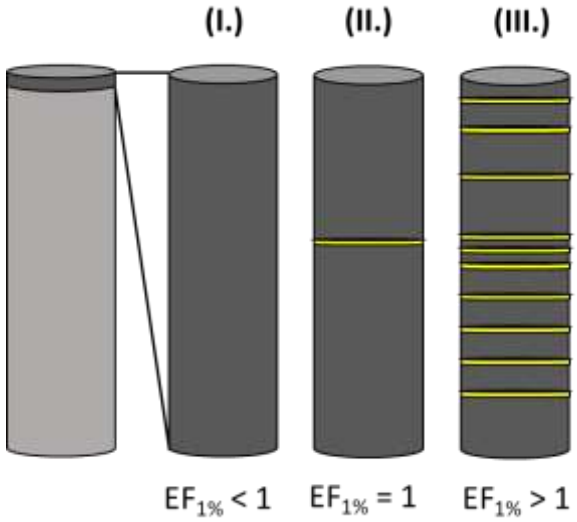


Figure 1.12 – Illustration of Enrichment Factors from Retrospective Virtual Screens. Assume that there 10 total compounds at the top 1% of the ranked database (dark gray), 1,000 compounds overall (dark gray + light gray), and 100 total active compounds. Each active compound is represented as a yellow bar in the top 1%. (I.) In this example, no actives were retrieved, thus the algorithm would be considered performing worse than random. (II.) Only 1 out of the 100 actives were acquired, so this would be considered random. (III.) In this scenario, 10 actives were found in the top 1%, performing greater than random.

A drawback of retrospective virtual screens is that no real results will have been produced that officially translate to biology. They are usually run to benchmark algorithms or to validate the virtual screening model before choosing compounds for experimental validation, and hence, they are useful in that regard. However, a model will always just be a model, and the only practical way to verify whether a virtual screen resulted in a prediction of any merit would be to test the compounds in the lab. They are usually chosen based on various criteria, such as manual inspection of the ligand-receptor complex for key interactions or chemical diversity. The usual metric for success is the hit rate, which is the percent of selected compounds that were considered active. This is not the greatest measure used, as the selection can be very subjective. Moreover, there is a risk of underreporting inactive compounds to boost the hit rate. Nevertheless, it remains an unwritten standard in the field to list hit rates as a measure of how successful the virtual screen was.

6. G Protein-Coupled Receptors

G protein-coupled receptors (GPCR) have been an important target in drug discovery, as they represent the target for almost a third of all drugs on the market. Marked by their distinctive seven-pass transmembrane domain, the GPCR superfamily astoundingly accounts for almost 5% of the human proteome. Consequently, the function of each member varies widely from pain modulation to vision. In this section, I will provide a brief overview of the biology of GPCR's, as well as applications of computer-aided drug design.

6.1 Family Organization

Originally, GPCR's were observed to contain members sharing very little sequence similarity with others, hinting at the existence of distinct families (a.k.a. classes) within the superfamily.[78] Further on down, phylogenetic analysis of GPCR primary structures with PSI-BLAST proposed a division into the following families:[79]

- Class A – Rhodopsin-like
- Class B – Secretin-receptor Family
- Class C – Metabotropic Glutamate / Pheromone
- Class D – Fungal Mating Pheromone Receptors
- Class E – Cyclic AMP Receptors
- Class F – Frizzled / Smoothened

Alternative classifications include the GRAFS system (Glutamate, Rhodopsin, Adhesion, Frizzled, Secretin), which separated out the adhesion receptors from Class B.[80] Of these, the Class A GPCR's by far outnumber the other families. Additionally, it was later stratified into 19 subgroups following phylogenetic analysis.[81] One of these subgroups, the olfactory receptors, comprise a whopping 390 members out of 719 Class A receptors. Class D and E GPCR's are usually excluded from vertebrate-based studies, as they are of fungal and slime mold origin, respectively. Type 2 taste receptors, which have been linked to detecting bitterness, have been distantly linked with Class A GPCR's; they are also sometimes separated out into their own group.[82] Clearly, the classification of GPCR's has been an arduous task that has taken decades to come to fruition, and thankfully there now stands a relatively clear-cut organizational system.

With such a large number of GPCR's, there will undoubtedly be members that have had not been study as rigorously. These are referred to as orphan receptors, having no known endogenous ligand or function. According to IUPHAR,[17] there are currently 87 Class A GPCR's, 8 Class C GPCR's, and 26 adhesion GPCR's that are still not well understood. The process of de-orphanisation is aimed at elucidating the pharmacology of these receptors in order to shed some new light on potentially medically-relevant conditions. Since the 1980's, there have been dozens of success stories in the de-orphanisation of GPCR's,[83] and it is only a matter of time before the function and endogenous ligands of all members are revealed.

6.2 Structure and Physiological Roles

All members of the GPCR superfamily share the characteristic seven-transmembrane domain that winds through the plasma membrane in a serpentine fashion. Additionally, they contain an N-terminal domain, C-terminal tail, 3 extracellular loops, and 3 intracellular loops. Not surprisingly, the most conserved regions lie in the transmembrane domains, where the original classification has been shown to be approximately replicated using only the transmembrane domains.[84] Moreover, this is the classical site where ligand binding occurs for many Class A receptors and was also replicated for this family using only the binding site residues in a previous study.[85] Interestingly, the barrel-like fold of the TM domain is highly conserved and observed in all solved structures (Figure 1.13B). The N-terminal domain is oriented to face the extracellular matrix and participates in ligand binding, while the C-terminal faces the cytosol and is involved with the association with downstream effector proteins (Figure 1.13A).

Since 2000, the number of solved GPCR structures in all families has skyrocketed, revealing a wealth of information about how they function through their structure. Both active- and inactive-state structures have been produced, with the former kind bound with either heterotrimeric G protein complexes or β-arrestin. Many of the structures have ligands bound, though a smaller set contain allosteric modulators. Truly, the history of GPCR structural biology is a fascinating topic, and thus, a comprehensive review of this will be given in Chapter 2. Lastly, details of receptor activation with respect to structure will not be discussed, as it is beyond the scope of the dissertation, though a good review of this topic exists by Katrich *et al*.[86]

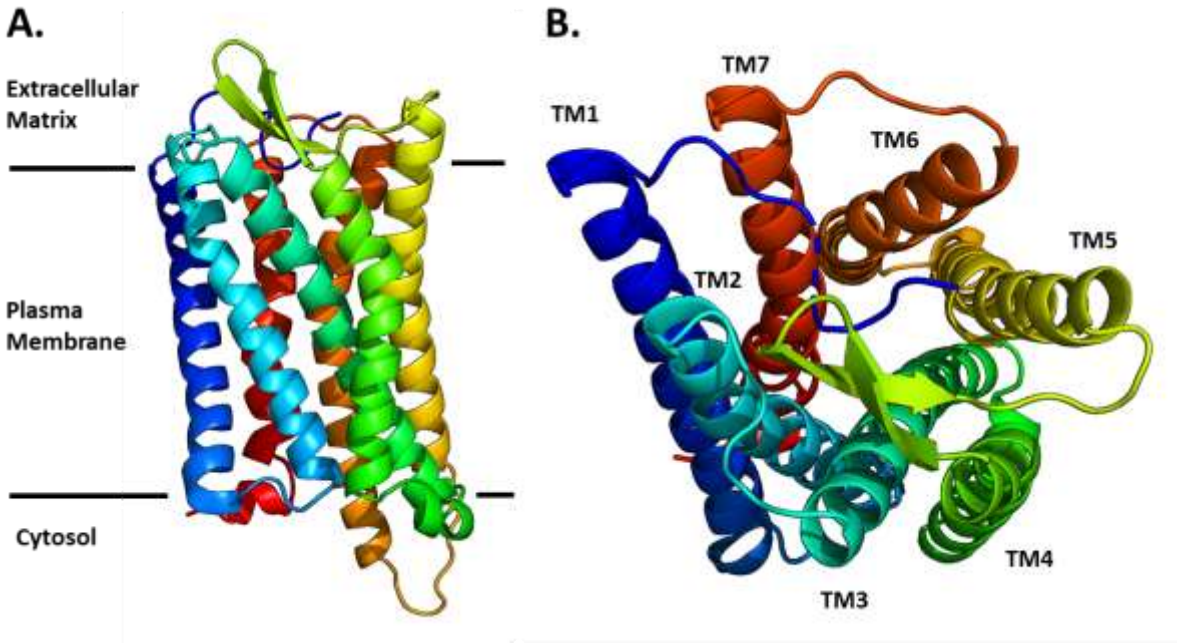


Figure 1.13 – Structure of Mouse Mu Opioid Receptor. The structure was taken and modified from the structure with PDB ID: 5C1M. (A.) A side view is shown to emphasize the cellular localization of the GPCR. (B.) A top-down view from the perspective of the extracellular matrix is shown to emphasize the conserved, barrel-like topology.

GPCR's are involved in a wide range of physiological roles in higher vertebrates and likely arose through multiple gene duplication events.[87] The duplicates would have had little evolutionary pressure, allowing them the freedom to mutate and develop new functions. This has resulted in roles as far flung as olfaction to light perception. Besides being involved in various senses, they have also played roles in fight-or-flight response and pain modulation. However, many GPCR's have important signaling pathways that could lead to diseases when dysregulated by mutations.[88] For example, activating mutations in the thyroid-stimulating hormone receptor have been shown to cause thyroid carcinoma,[89] while inactivating mutations in the same receptor were associated with hypothyroidism.[90] Other diseases, such as addiction, have been shown to be caused by the kind of ligand be used. In the case of the mu opioid receptor, an unwanted signaling pathway is activated enough to the point where the development of addiction can develop.[91] No matter the cause of the disease, the development of therapeutic compounds targeting GPCR's has remained an important part of drug development in order to combat suffering.

6.3 Application of Computer-Aided Drug Design

To date, there exist a plethora of prospective virtual screening campaigns applied to GPCR's. Some interesting studies have included the discovery of: 1.) a biased agonist for the mu opioid receptor,[92] 2.) selective agonists for the serotonin 1B receptor over the serotonin 2B receptor,[71] and 3.) antagonists for the C-X-C chemokine receptor 4.[93] Though there have been numerous computational studies claiming to have found the next greatest potential drug, the question bugging countless experimentalist scientists remains: Does computer-aided drug design actually produce compounds that make it to clinical trials? The answer would be a resounding 'yes'.

In 2006, a group from Predix Pharmaceuticals (now known as Epix Pharmaceuticals) ran a docking-based virtual screen on a homology model of serotonin 1A receptor that resulted in a potent, selective agonist.[94] The reporting came as the drug candidate, Naluzotan, was in a phase III clinical trial, though ultimately, it failed to perform better than the placebo and was discontinued.[95] In another study from Heptares Therapeutics, a novel adenosine A_{2A} receptor antagonist was discovered through a docking-based virtual screen on homology models, called AZD4635.[96] It is currently in phase I clinical trials. These are but a fraction of success stories stemming from computer-aided drug design, but this is proof enough that the field has undoubtedly been advancing to produce viable results.

7. Goal of Dissertation

In the following two chapters, I introduce two GPCR-related databases in which I developed. The first to be introduced is GPCR-EXP, a semi-manually curated database for experimentally-solved and predicted GPCR structures, which was created in response to other resources being slow to update and not user friendly. Next, GLASS database was constructed, which processes and unifies GPCR experimental data from various other pharmacological and biological databases. To date, it is the largest database of its kind and was made to fill in the gaps of other similar databases that have ceased updates.

The latter two work-related chapters delve deeper into algorithm development using bioinformatics and chemoinformatics, as well as its application. In the 4th chapter, I discuss the development of MAGELLAN, which is a ligand-based virtual screening algorithm that

incorporates sequence and structure information, and data from GLASS database was heavily involved in its usage. Its purpose was primarily to aid in the de-orphanisation of GPCR's, but it could also be used as for general ligand-based virtual screening. Finally, the 5th chapter builds on the work from MAGELLAN by coupling it with a docking-based virtual screen. The goal of this work was to discover novel bifunctional compounds that act as agonists towards the mu opioid receptor and as an antagonist towards the delta opioid receptor, which would potentially lead to a safer opioid with reduced tolerance and withdrawal.

8. References

1. Berman, H. M., The protein data bank: a historical perspective. *Acta Crystallographica Section A: Foundations of Crystallography* **2008,** *64* (1), 88-95.
2. O'Boyle, N. M., Towards a Universal SMILES representation-A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics* **2012,** *4* (1), 22.
3. Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I., InChI-the worldwide chemical structure identifier standard. *Journal of cheminformatics* **2013,** *5* (1), 7.
4. Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S., InChIKey collision resistance: an experimental testing. *Journal of cheminformatics* **2012,** *4* (1), 39.
5. Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G., Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002,** *42* (6), 1273-1280.
6. O'Boyle, N. M.; Sayle, R. A., Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of cheminformatics* **2016,** *8* (1), 36.
7. Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L.; Leland, B. A.; Laufer, J., Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of chemical information and computer sciences* **1992,** *32* (3), 244-255.
8. Lipman, D. J.; Pearson, W. R., Rapid and sensitive protein similarity searches. *Science* **1985,** *227* (4693), 1435-1441.
9. Consortium, U., UniProt: the universal protein knowledgebase. *Nucleic acids research* **2016,** *45* (D1), D158-D169.
10. Rose, P. W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K., The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic acids research* **2017,** *45* (D1), D271-D281.
11. brPrilusky, J., OCA, a browser-database for protein structure/function. *URL http://oca. weizmann. ac. il and mirrors worldwide* **1996**.
12. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012,** *40* (Database issue), D1100-7.

13. Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J., BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **2015,** *44* (D1), D1045-D1053.
14. Wassermann, A. M.; Bajorath, J., BindingDB and ChEMBL: online compound databases for drug discovery. *Expert opinion on drug discovery* **2011,** *6* (7), 683-687.
15. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A. C.; Liu, Y.; Maciejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; Tang, A.; Gabriel, G.; Ly, C.; Adamjee, S.; Dame, Z. T.; Han, B.; Zhou, Y.; Wishart, D. S., DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **2014,** *42* (Database issue), D1091-7.
16. Roth, B. L.; Lopez, E.; Patel, S.; Kroeze, W. K., The multiplicity of serotonin receptors: uselessly diverse molecules or an embarrassment of riches? *The Neuroscientist* **2000,** *6* (4), 252-262.
17. Alexander, S. P.; Davenport, A. P.; Kelly, E.; Marrion, N.; Peters, J. A.; Benson, H. E.; Faccenda, E.; Pawson, A. J.; Sharman, J. L.; Southan, C.; Davies, J. A.; Collaborators, C., The Concise Guide to PHARMACOLOGY 2015/16: G protein-coupled receptors. *Br J Pharmacol* **2015,** *172* (24), 5744-869.
18. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research* **2009,** *37* (Web Server issue), W623-33.
19. Dayhoff, M.; Schwartz, R.; Orcutt, B., 22 a model of evolutionary change in proteins. *Atlas of protein sequence and structure* **1978**, 345-352.
20. Henikoff, S.; Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **1992,** *89* (22), 10915-10919.
21. Mount, D. W., Comparison of the PAM and BLOSUM amino acid substitution matrices. *Cold Spring Harbor Protocols* **2008,** *2008* (6), pdb. ip59.
22. Gotoh, O., An improved algorithm for matching biological sequences. *Journal of molecular biology* **1982,** *162* (3), 705-708.
23. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403-10.
24. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997,** *25* (17), 3389-402.
25. Akiyama, Y.; Hosoya, T.; Poole, A. M.; Hotta, Y., The gcm-motif: a novel DNA-binding motif conserved in Drosophila and mammals. *Proceedings of the National Academy of Sciences* **1996,** *93* (25), 14912-14916.
26. Sievers, F.; Higgins, D. G., Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. *Multiple Sequence Alignment Methods* **2014,** *1079*, 105-116.
27. Zhang, Y.; Skolnick, J., Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics* **2004,** *57* (4), 702-710.
28. Xu, J.; Zhang, Y., How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics* **2010,** *26* (7), 889-895.
29. Illergård, K.; Ardell, D. H.; Elofsson, A., Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics* **2009,** *77* (3), 499-508.
30. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **2005,** *33* (7), 2302-9.

31. Ozkan, S. B.; Wu, G. A.; Chodera, J. D.; Dill, K. A., Protein folding by zipping and assembly. *Proceedings of the National Academy of Sciences* **2007,** *104* (29), 11987-11992.
32. Xu, D.; Zhang, Y., Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins: Structure, Function, and Bioinformatics* **2012,** *80* (7), 1715-1735.
33. Simons, K. T.; Kooperberg, C.; Huang, E.; Baker, D., Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions1. *Journal of molecular biology* **1997,** *268* (1), 209-225.
34. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M.; Eramian, D.; Shen, M. y.; Pieper, U.; Sali, A., Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* **2006,** *15* (1), 5.6. 1-5.6. 30.
35. Martí-Renom, M. A.; Stuart, A. C.; Fiser, A.; Sánchez, R.; Melo, F.; Šali, A., Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* **2000,** *29* (1), 291-325.
36. Baker, D.; Sali, A., Protein structure prediction and structural genomics. *Science* **2001,** *294* (5540), 93-96.
37. Magner, A.; Szpankowski, W.; Kihara, D., On the origin of protein superfamilies and superfolds. *Scientific reports* **2015,** *5*, 8166.
38. Söding, J.; Biegert, A.; Lupas, A. N., The HHpred interactive server for protein homology detection and structure prediction. *Nucleic acids research* **2005,** *33* (suppl_2), W244-W248.
39. Kelley, L. A.; Sternberg, M. J., Protein structure prediction on the Web: a case study using the Phyre server. *Nature protocols* **2009,** *4* (3), 363.
40. Wu, S.; Zhang, Y., MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. *Proteins: Structure, Function, and Bioinformatics* **2008,** *72* (2), 547-556.
41. Roy, A.; Kucukural, A.; Zhang, Y., I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols* **2010,** *5* (4), 725.
42. Zhang, Y., I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **2008,** *9* (1), 40.
43. Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y., The I-TASSER Suite: protein structure and function prediction. *Nature methods* **2015,** *12* (1), 7.
44. Zhang, J.; Yang, J.; Jang, R.; Zhang, Y., GPCR-I-TASSER: a hybrid approach to G protein-coupled receptor structure modeling and the application to the human genome. *Structure* **2015,** *23* (8), 1538-1549.
45. Zhang, J.; Zhang, Y., GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation. *Bioinformatics* **2010,** *26* (23), 3004-3005.
46. Roy, A.; Yang, J.; Zhang, Y., COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research* **2012,** *40* (W1), W471-W477.
47. Zhang, C.; Freddolino, P. L.; Zhang, Y., COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic acids research* **2017,** *45* (W1), W291-W299.
48. Ashburner, M.; Ball, C. A.; Blake, J. A.; Botstein, D.; Butler, H.; Cherry, J. M.; Davis, A. P.; Dolinski, K.; Dwight, S. S.; Eppig, J. T., Gene Ontology: tool for the unification of biology. *Nature genetics* **2000,** *25* (1), 25.

49. Yang, J.; Roy, A.; Zhang, Y., Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013,** *29* (20), 2588-95.
50. Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research* **2013,** *41* (Database issue), D1096-103.
51. Brylinski, M.; Skolnick, J., A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **2008,** *105* (1), 129-34.
52. Capra, J. A.; Laskowski, R. A.; Thornton, J. M.; Singh, M.; Funkhouser, T. A., Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS computational biology* **2009,** *5* (12), e1000585.
53. Bajusz, D.; Rácz, A.; Héberger, K., Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics* **2015,** *7* (1), 20.
54. Taylor, R., Simulation Analysis of Experimental-Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *Journal of Chemical Information and Computer Sciences* **1995,** *35* (1), 59-67.
55. Butina, D., Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences* **1999,** *39* (4), 747-750.
56. Library MCS (LibMCS) clustering.
57. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **2009,** *30* (16), 2785-2791.
58. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology* **1997,** *267* (3), 727-748.
59. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010,** *31* (2), 455-61.
60. Korb, O.; Stutzle, T.; Exner, T. E., Empirical scoring functions for advanced protein− ligand docking with PLANTS. *Journal of chemical information and modeling* **2009,** *49* (1), 84-96.
61. Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C., DOCK 6: Impact of new features and current docking performance. *J Comput Chem* **2015,** *36* (15), 1132-56.
62. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K., Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of medicinal chemistry* **2004,** *47* (7), 1739-1749.
63. Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L., Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *Journal of medicinal chemistry* **2004,** *47* (7), 1750-1759.
64. Meng, E. C.; Shoichet, B. K.; Kuntz, I. D., Automated docking with grid-based energy evaluation. *Journal of computational chemistry* **1992,** *13* (4), 505-524.
65. Huang, S.-Y.; Grinter, S. Z.; Zou, X., Scoring functions and their evaluation methods for protein–ligand docking: recent advances and future directions. *Physical Chemistry Chemical Physics* **2010,** *12* (40), 12899-12908.
66. Liu, J.; Wang, R., Classification of current scoring functions. *Journal of chemical information and modeling* **2015,** *55* (3), 475-482.

67. Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E., A geometric approach to macromolecule-ligand interactions. *Journal of molecular biology* **1982,** *161* (2), 269-288.
68. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012,** *52* (7), 1757-68.
69. Lipinski, C. A., Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004,** *1* (4), 337-341.
70. Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T., The design of leadlike combinatorial libraries. *Angewandte Chemie International Edition* **1999,** *38* (24), 3743-3748.
71. Rodriguez, D.; Brea, J.; Loza, M. I.; Carlsson, J., Structure-based discovery of selective serotonin 5-HT(1B) receptor ligands. *Structure* **2014,** *22* (8), 1140-1151.
72. Irwin, J. J., Community benchmarks for virtual screening. *Journal of computer-aided molecular design* **2008,** *22* (3-4), 193-199.
73. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** *55* (14), 6582-94.
74. Good, A. C.; Oprea, T. I., Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of computer-aided molecular design* **2008,** *22* (3-4), 169-178.
75. Mysinger, M. M.; Shoichet, B. K., Rapid context-dependent ligand desolvation in molecular docking. *Journal of chemical information and modeling* **2010,** *50* (9), 1561-1573.
76. Bemis, G. W.; Murcko, M. A., The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996,** *39* (15), 2887-93.
77. Drwal, M. N.; Griffith, R., Combination of ligand- and structure-based methods in virtual screening. *Drug Discov Today Technol* **2013,** *10* (3), e395-401.
78. Attwood, T.; Findlay, J., Fingerprinting G-protein-coupled receptors. *Protein Engineering, Design and Selection* **1994,** *7* (2), 195-203.
79. Josefsson, L.-G., Evidence for kinship between diverse G-protein coupled receptors. *Gene* **1999,** *239* (2), 333-340.
80. Bjarnadóttir, T. K.; Gloriam, D. E.; Hellstrand, S. H.; Kristiansson, H.; Fredriksson, R.; Schiöth, H. B., Comprehensive repertoire and phylogenetic analysis of the G protein-coupled receptors in human and mouse. *Genomics* **2006,** *88* (3), 263-273.
81. Joost, P.; Methner, A., Phylogenetic analysis of 277 human G-protein-coupled receptors as a tool for the prediction of orphan receptor ligands. *Genome biology* **2002,** *3* (11), research0063. 1.
82. Adler, E.; Hoon, M. A.; Mueller, K. L.; Chandrashekar, J.; Ryba, N. J.; Zuker, C. S., A novel family of mammalian taste receptors. *Cell* **2000,** *100* (6), 693-702.
83. Civelli, O.; Reinscheid, R. K.; Zhang, Y.; Wang, Z.; Fredriksson, R.; Schioth, H. B., G protein-coupled receptor deorphanizations. *Annu Rev Pharmacol Toxicol* **2013,** *53*, 127-46.
84. Cvicek, V.; Goddard III, W. A.; Abrol, R., Structure-based sequence alignment of the transmembrane domains of all human GPCRs: Phylogenetic, structural and functional implications. *PLoS computational biology* **2016,** *12* (3), e1004805.
85. Lin, H.; Sassano, M. F.; Roth, B. L.; Shoichet, B. K., A pharmacological organization of G protein-coupled receptors. *Nat Methods* **2013,** *10* (2), 140-6.
86. Katritch, V.; Cherezov, V.; Stevens, R. C., Structure-function of the G protein–coupled receptor superfamily. *Annual review of pharmacology and toxicology* **2013,** *53*, 531-556.

87. Schöneberg, T.; Hofreiter, M.; Schulz, A.; Römpler, H., Learning from the past: evolution of GPCR functions. *Trends in pharmacological sciences* **2007,** *28* (3), 117-121.
88. Schöneberg, T.; Schulz, A.; Biebermann, H.; Hermsdorf, T.; Römpler, H.; Sangkuhl, K., Mutant G-protein-coupled receptors as a cause of human diseases. *Pharmacology & therapeutics* **2004,** *104* (3), 173-206.
89. Russo, D.; Arturi, F.; Schlumberger, M.; Caillou, B.; Monier, R.; Filetti, S.; Suarez, H., Activating mutations of the TSH receptor in differentiated thyroid carcinomas. *Oncogene* **1995,** *11* (9), 1907-1911.
90. Biebermann, H.; Grüters, A.; Schöneberg, T.; Gudermann, T., Congenital hypothyroidism caused by mutations in the thyrotropin-receptor gene. *New England Journal of Medicine* **1997,** *336* (19), 1390-1391.
91. Porter-Stransky, K. A.; Weinshenker, D., Arresting the development of addiction: the role of β-arrestin 2 in drug abuse. *Journal of Pharmacology and Experimental Therapeutics* **2017,** *361* (3), 341-348.
92. Manglik, A.; Lin, H.; Aryal, D. K.; McCorvy, J. D.; Dengler, D.; Corder, G.; Levit, A.; Kling, R. C.; Bernat, V.; Hübner, H., Structure-based discovery of opioid analgesics with reduced side effects. *Nature* **2016,** *537* (7619), 185.
93. Mysinger, M. M.; Weiss, D. R.; Ziarek, J. J.; Gravel, S.; Doak, A. K.; Karpiak, J.; Heveker, N.; Shoichet, B. K.; Volkman, B. F., Structure-based ligand discovery for the protein-protein interface of chemokine receptor CXCR4. *Proc Natl Acad Sci U S A* **2012,** *109* (14), 5517-22.
94. Becker, O. M.; Dhanoa, D. S.; Marantz, Y.; Chen, D.; Shacham, S.; Cheruku, S.; Heifetz, A.; Mohanty, P.; Fichman, M.; Sharadendu, A.; Nudelman, R.; Kauffman, M.; Noiman, S., An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. *J Med Chem* **2006,** *49* (11), 3116-35.
95. Kirchhoff, V. D.; Nguyen, H. T.; Soczynska, J. K.; Woldeyohannes, H.; McIntyre, R. S., Discontinued psychiatric drugs in 2008. *Expert opinion on investigational drugs* **2009,** *18* (10), 1431-1443.
96. Langmead, C. J.; Andrews, S. P.; Congreve, M.; Errey, J. C.; Hurrell, E.; Marshall, F. H.; Mason, J. S.; Richardson, C. M.; Robertson, N.; Zhukov, A., Identification of novel adenosine A2A receptor antagonists by virtual screening. *Journal of medicinal chemistry* **2012,** *55* (5), 1904-1909.

CHAPTER 2.
GPCR-EXP: A Semi-Manually Curated Database for Experimentally-Solved and Predicted GPCR Structures

1. Introduction

G protein-coupled receptors (GPCR) constitute one of the largest family of transmembrane proteins and have been implicated in a multitude of human diseases, such as cancer and diabetes.[1] With the rapid rise and availability of GPCR structures within the past decade, structure-based drug design has become a prominent area in drug development utilizing structure to facilitate the optimization of known compounds or to provide a means for virtual screening.

The centralized resource that houses all three-dimensional structural data for biological macromolecules is the Protein Data Bank (PDB).[2] As such, all GPCR structures are submitted and stored in this database, so that the scientific community can effortlessly access this data. However, there is no easy way to survey the breadth of GPCR structures within the confines of PDB, and thus, researchers have to rely upon manually sifting through the scientific literature to identify which structures have been solved. This process can be cumbersome and ineffective, and hence a database consolidating all aspects of GPCR structural data would be extremely useful in this regard.

To our knowledge, there is only one resource, GPCRdb,[3] which focuses on experimental GPCR structures. Some of the difficulties in establishing such databases are: 1) precise curation of the data, 2) timely updates, and 3) usability of the web interface. As GPCR structures are increasingly being released at a higher rate, the time in between database updates can greatly affect the availability of new data to the scientific community. Furthermore, purely manual methods of data acquisition can lead to inaccuracies and take a long time to curate, while an inefficient database

interface with endless options can lead to utter frustration in perusing the information in need. As a result, the need for a GPCR structure database addressing such needs is of utmost importance.

Despite the recent renaissance in GPCR structural biology, 1,029 out of 1.076 human GPCR genes do not yet have an experimental structure. Therefore, apart from experimental GPCR structures, the prediction of GPCR structures holds equal importance. In particular, as many efforts in drug discovery utilize structure for drug design, computational methods have been developed to predict the protein structure for those targets without a solved structure. Many web servers, such as GPCRM,[4] GoMoDo,[5] and GPCR-ModSim,[6] and databases, including GPCR-SSFE[7] and GPCRdb[3] rely upon homology modeling with MODELLER[8] for generating their structure models. While this method works well for targets that have close relatives with known structures, it can be less effective when applied to targets with no good homologous structure templates. As a majority of human GPCR proteins share less than 30% sequence identity to any experimentally determined GPCR structures, a robust structure prediction program capable of modeling distantly or non-homologous GPCR structures is required. GPCR-I-TASSER is our in-house method that utilizes the LOMETS[9] meta-threading approach to find templates, from which structure fragments and distance restraints are extracted to guide structure assembly simulation. Our method performed well in the GPCRDock2010 competition, in which our models were among the most accurate for the transmembrane (TM) domains for both the C-X-C chemokine 4 and dopamine D3 receptors; this was significant because we achieved the most accurate structure for the former target, which was considered very difficult.[10] This indicates the ability of GPCR-I-TASSER to produce accurate structure models for use in various avenues of drug discovery.

In the present study, we have developed GPCR-EXP, a semi-manually curated database for experimentally-solved and predicted GPCR structures. Experimental data is scraped and processed from the PDB, producing a comprehensive, accurate collection of protein and ligand data. An important feature of the update pipeline for keeping the GPCR data current is the manual curation of ligand data, which is crosschecked with PubMed. Additionally, structure models are generated for over 1,000 GPCRs from the human genome, followed by binding site prediction. All data are freely downloadable and browsable in a convenient database interface. Weekly updates to GPCR-EXP's experimentally-validated GPCR structures will ensure that the GPCR community gets the

latest structures on a consistent basis. Lastly, we present a brief analysis of trends in GPCR structural biology using the experimental data in GPCR-EXP.

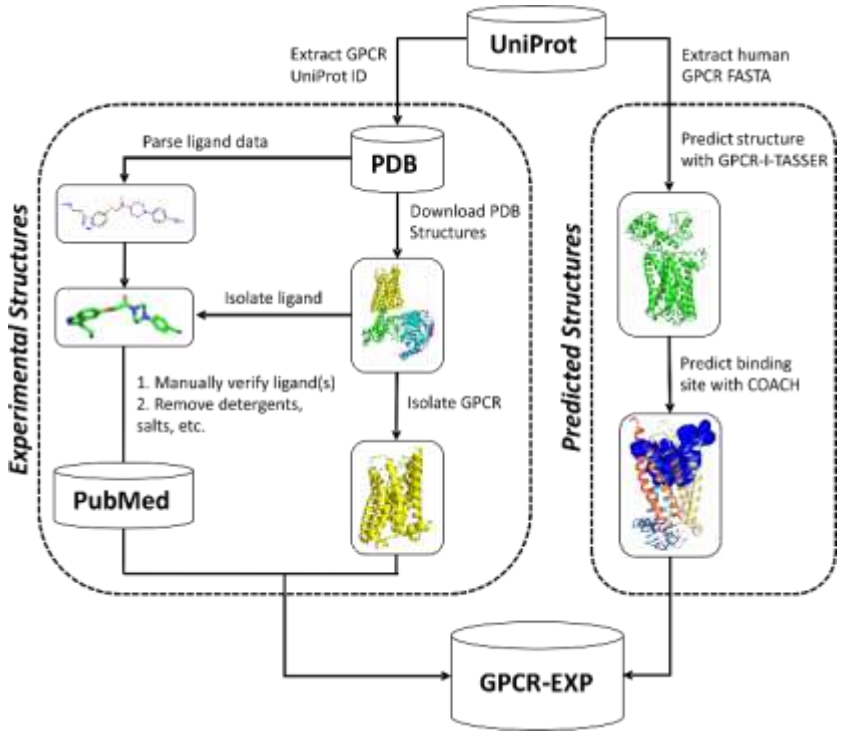


Figure 2.1 - GPCR-EXP Pipeline for Data Processing.

2. Methods

GPCR-EXP contains data for both experimentally-solved and predicted structures of GPCRs. Data for the former are primarily extracted from PDB, while high-resolution structure models and binding site predictions are generated for the latter with GPCR-I-TASSER[11] and COACH[12-13], respectively. The pipeline for the acquisition and processing of data for GPCR-EXP is illustrated in Figure 2.1.

2.1 Processing Data for Experimental Structures

The entire data acquisition and processing pipeline was written in Perl, while additional custom Python scripts were used for 2D chemical image generation.

2.1.1 General Experiment Data

UniProt IDs are programmatically acquired from '7tmrlist.txt' in UniProt (https://www.uniprot.org/docs/7tmrlist.txt), which are then queried against the RESTful API from PDB.[2, 14] Subsequently, a list of PDB IDs mapped from the UniProt IDs are returned and used to scrape data and download PDB structures.

Using the list of PDB IDs, comma-separated value (CSV) report files are scraped from the RESTful API from PDB using their 'Custom Report Web Services', corresponding to 'Structure Summary Report', 'Sequence Report', 'Ligand Report', and 'Citation Report'. All relevant data, such as chain IDs and ligands, from these files are used in the following data processing steps, in addition to forming the basis for the compilation of a summary tab-separated values (TSV) file of all processed data in one of the final steps.

2.1.2 PDB Structures

The original PDB structures are downloaded from PDB using their PDB ID. GPCR structures with full TM domains are filtered for using the following criteria: 1) they are greater than 150 residues, and 2) the full TM domain has at least 80% sequence identity with the corresponding reference TM domain from UniProt. With respect to the latter condition, the reference TM domain sequences are combined and treated as a single sequence, which is then aligned with the GPCR sequence parsed from the PDB file. GPCR structures that have been solved with only the extracellular domain or portions of the transmembrane domain are filtered out and included as a separate download; for the purposes of this study, these will be referred to as 'GPCR fragments'.

All GPCR structures with full length TM domains are then processed to isolate a single chain representing only the GPCR, removing any fusion (e.g. lysozyme) or associated (e.g. G protein) proteins. If the PDB structure has multiple GPCR chains, the chain ID of the GPCR with the alphabetically-lowest letter is designated and used for the rest of the pipeline. For purposes of viewability, the isolated GPCR structures are structurally aligned with a reference structure using TM-align in order to position the GPCR with a side view of the TM domain where the N-terminal domain faces up. Moreover, PNG files of the aligned GPCRs are generated using MolScript [15] for use as thumbnail images on the web page.

Data about fusion and associated proteins are extracted from the 'Sequence Report' CSV files on the basis of being a non-GPCR protein within the PDB structure. Numerous associated proteins are observed to be peptide ligands and typically appeared as a distinct chain in the PDB structure. An associated protein is classified as a peptide ligand if there are fewer than 50 residues. Interestingly, the largest peptide that we observe is Exendin-P5 (PDB: 6B3J), which contains 40 residues. All peptides are kept as ligands alongside small molecules, which will be discussed in the following section.

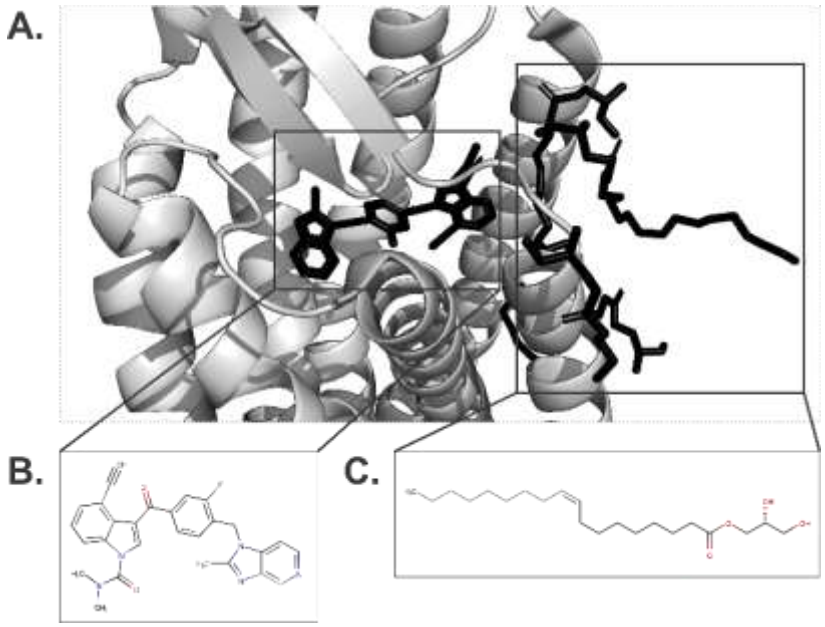


Figure 2.2 - Ligands for Platelet-Activating Receptor (PDB: 5ZKQ). (A) The structure includes four ligands (ABT-491, (2R)-2,3-dihydroxypropyl (9Z)-octadec-9-enoate, zinc, and sulfate). The former two are shown in black. (B) ABT-491 is the ligand, which acts as an inverse-agonist, while (C) (2R)-2,3-dihydroxypropyl (9Z)-octadec-9-enoate is a lipid that assists in membrane protein crystallography. MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

2.1.3 Ligand Data

Small molecules are parsed from the 'Ligand Report' CSV files. As commonly seen, there is an abundance of non-ligand molecules alongside the actual ligand or ligands. Consequently, this makes programmatic acquisition of the ligand difficult. To address this problem, we designate ligands through manual inspection of the corresponding publication through PubMed. Figure 2.2A shows an example case with the platelet-activating receptor (PDB: 5ZKQ). Here, the ligand was determined to be the inverse agonist ABT-491 (Figure 2B). Molecules ignored included a

detergent used in membrane protein crystallization, (2R)-2,3-dihydroxypropyl (9Z)-octadec-9-enoate (Figure 2.2C), as well as the ions, zinc and sulfate.

We initially compiled a list of non-ligand molecules following a comprehensive survey of PDB structures of GPCRs and incorporated them into a filter. This allowed us to streamline the process of verifying ligands while updating the database. If a new molecule appears in an update that the database has yet encountered, then it is manually checked and designated as a ligand or added to the filter if it is a non-ligand. Over the course of developing the update pipeline, we observed that the current filter is sufficient for most of the new GPCR structures coming out recently, indicating reoccurrence of common, non-ligand molecules.

PNG images of 2D chemical structures are generated for each ligand through RDKit using their respective InChI strings.[16] Additionally, both small-molecule and peptide ligands were extracted from the original PDB structures as PDB files. Also, it should be noted that allosteric modulators were included alongside ligands that bound in the orthosteric site.

2.1.4 GPCR Structure Superposition

Many GPCRs have multiple PDB structures bound with different ligands or in different activation states. For example, the 5-hydroxytryptamine receptor 1B has four PDB structures (PDB: 6G79, 5V54, 4IAQ, 4IAR) as of the time of preparation of the manuscript. For each GPCR, one structure is chosen as the reference, while the others are aligned to it using TM-align.[17] All small molecule and peptide ligands are included in this step, as well.

2.2 Generating Predicted Structures and Binding Sites

UniProt IDs for 825 human GPCRs were programmatically acquired from '7tmrlist.txt' in UniProt, while 251 additional entries were obtained from TrEMBL. In total, we modelled 1,076 human GPCRs in this study, with 703, 49, 22, 11, 44, and 247 from class A, B, C, F, other, and TrEMBL, respectively.

The full-length sequence of the human GPCRs were modelled using GPCR-I-TASSER, which operates by reassembling structural fragments from threading through replica-exchange Monte

Carlo simulations.[11] There were a number of GPCRs that had very long sequences. In fact, there were 22 GPCRs that had over 1,000 residues, which consisted of a mixture of Class B and C GPCRs with gigantic extracellular domains. Though GPCR-I-TASER can model proteins over 1,000 residues, the models with more than 650 residues are poorly modeled with incompact transmembrane helix packing, due to poor template coverage and insufficient conformation sampling for long proteins. Thus, for those GPCRs with over 650 residues, we truncated the N- and C-termini so that there would be at most 30 residues extending from either side of the TM domain. The top 5 models were selected for use in the database.

COACH[12-13] is an algorithm developed to detect ligand binding residues through composite sequence-profile and structure comparisons and was thus used to predict the binding site for the top model from GPCR-I-TASSER.

2.3 Web Server Construction

The web server was constructed on top of a MySQL database. The server side was coded with a combination of Python CGI scripting and PHP, while the client side was controlled with JavaScript. NGL Viewer was used to display all protein structures,[18] and Plotly facilitated the visual display of graphs for database statistics. Lastly, the jQuery plugin, tablesorter (https://plugins.jquery.com/tablesorter/), was used to generate the data tables.

3. Results

3.1 Brief Analysis of GPCR-EXP

As of the time of writing, GPCR-EXP contains 271 unique, experimental GPCR structures. A total of eight species is represented in this set, ranging from common species such as *Homo sapiens* and *Mus musculus* to more uncommon ones, such as human cytomegalovirus (strain AD169). Additionally, there are 52 unique types of GPCRs that have been solved independent of species, such as the mu opioid receptor. We also modelled the structures of 1,076 human GPCRs, as well as predicted their binding sites.

Analysis of the experimental data of the GPCR structures reveals fascinating trends in structural biology. As has been frequently observed in recent years, the rate of release of GPCR structures

has been steadily increasing (Figure 2.3). In particular, the number of structures solved in the current year (as of July 27, 2018) is 42, as compared to 46 in 2017. Being only halfway through the year, it is very likely that the amount of releases in this year will dwarf the previous. One interesting observation is that no GPCR structures were solved in 2009. Up until then, the only available GPCRs were rhodopsin, β1 and β2 adrenergic receptors, and A2A adenosine receptor; the chemokine CXCR4 receptor was the first new GPCR to be released in 2010 after the drought of 2009. With the current overrepresentation of rhodopsin (50 structures), β1 and β2 adrenergic receptors (39 structures), and A2A adenosine receptor (45 structures) among all the GPCR structures, it is likely that they are easier to crystallize or have a better-established protocol than the others.

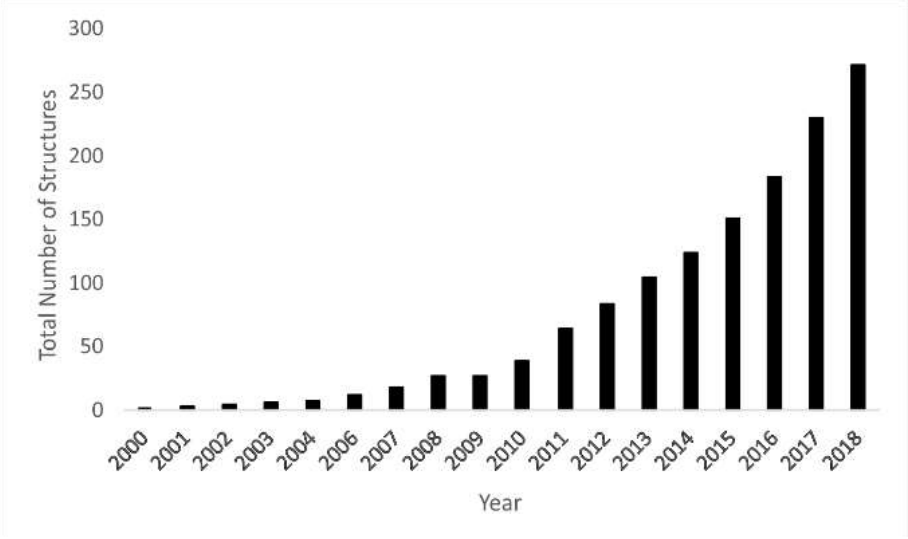


Figure 2.3 - Cumulative Number of Experimentally-Solved GPCR Structures Over Time. The first GPCR structure ever solved was in 2000. Since then, structures have been released at higher rates in recent years, likely due to advances in the field. Note that no structures were solved in 2009.

Starting from 2000, all of the GPCRs initially solved were Class A GPCRs. This is not a surprising, as they constitute the majority of members in the superfamily. However, it took until 2013 for the first Class B and F GPCRs to finally be solved,[19-21] while it took until the following year to finally get a solved structure for Class C GPCRs (Figure 2.4A).[22-23] For each year up until 2017, non-Class A GPCRs have constituted at most only ~10% of all structures per year. Regardless, their proportion has grown in recent years, and their prevalence is expected to increase due to their medical relevance. Another interesting facet of GPCR structural biology was the inclusion of

fusion proteins to facilitate the formation of crystal contacts in crystallography. Starting in 2007, Cherezov *et al* was the first study to engineer a chimeric GPCR construct, where the third intracellular loop was replaced with lysozyme.[24] In 2012, researchers started using soluble apocytochome b562 (i.e. bRIL), and this has steadily grown in popularity, having largely replaced lysozyme in recent years (Figure 2.4B). Meanwhile, other fusion proteins (rubredoxin, flavodoxin, and GlgA glycogen synthase) have also been used to lesser extents.

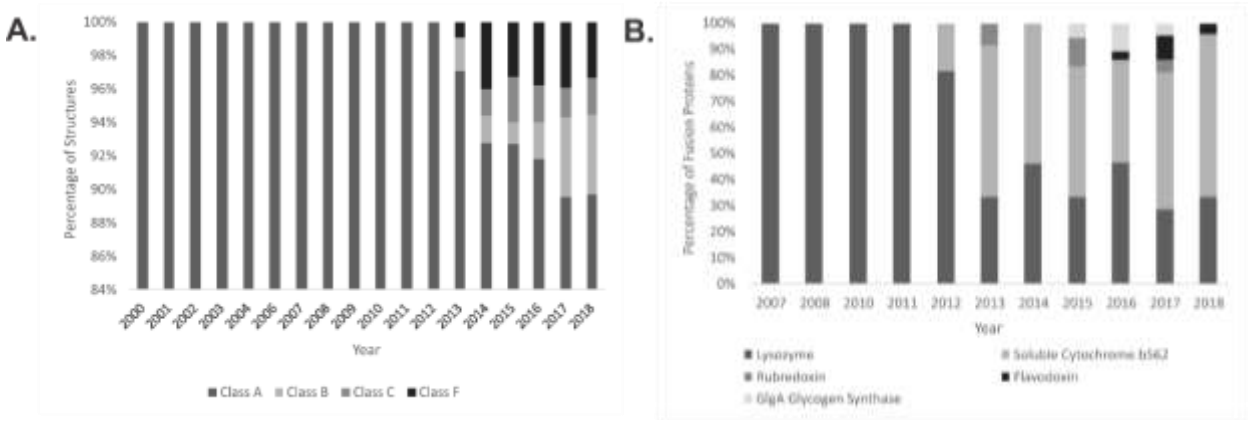


Figure 2.4 - Percentage of Structures per GPCR Classes or Fusion Proteins by Year. (A) Starting in 2000, only structures for Class A GPCRs were solved until 2013, when the first structures of Class B and F GPCRs were unveiled. (B) In 2007, fusion proteins were added to GPCR constructs in order to increase crystallographic contacts. A popular choice was lysozyme initially, but soluble cytochrome b562 quickly became a recent favorite. Note that no structures were solved in 2009.

Another aspect we examined was structural resolution and number of point mutations used with GPCR constructs. The resolution of the structure of bovine rhodopsin, the first GPCR to have its structure solved by X-ray crystallography, was 2.8 Å.[25] This was quite the milestone at the time, though its relative abundance and stability probably greatly facilitated crystallization. With all the advances of GPCR structural biology, one would think that the resolution would improve overall, even slightly, through time. Though this is true only for some cases, the overall trend has remained approximately the same on average up until now (Figure 2.5A). It is noted though, while Cryo-EM structures tend to have lower resolution in general, the 9 structures currently solved by this method have not contributed greatly to diminishing the average. In fact, the resolution of the structures range from 3.3 – 4.5 Å, which is quite an achievement given the technique's earlier limitations in resolution.[26] Starting in 2007, the number of GPCR structures solved with stabilizing point mutations has on average grown over the years (Figure 2.5B). This is likely due to its necessity because of challenges in the crystallography of certain GPCRs, as other unexplored

targets are being addressed. An extreme example of this would be the neurotensin receptor 1 (PDB: 4BV0/4BWB), of which two structures had greater than 20 point mutations resulting from directed evolution.[27] Regardless, the trend appears to be target dependent, as numerous GPCR structures have also recently been solved without any mutations.

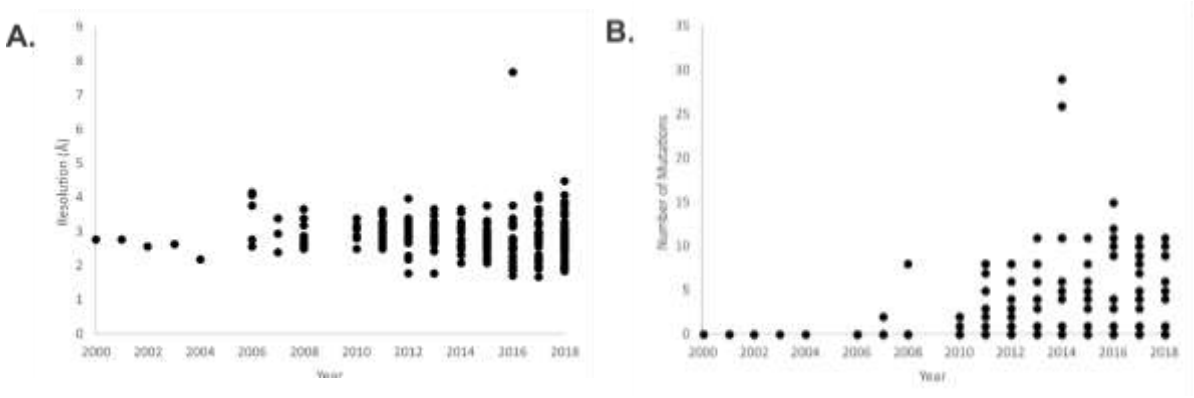


Figure 2.5 - Structure Resolution and Number of Mutations by Year. (A) The overall resolution of GPCR structures has remained approximately the same, while (B) the overall number of mutations used per structure has increased.

3.2 Browsing GPCR-EXP

GPCR-EXP was developed to provide the user with an intuitive web interface for browsing GPCR structural data, which is divided into experimentally-solved structures and predicted structures. We designed the database to allow the access of GPCR structure-related data as fluid as possible. As opposed to going through numerous menus, pages, and options for one PDB structure, an integrated information set can be accessed in GPCR-EXP with minimal hassle.

3.2.1 Experimentally-Solved Structures

All experimentally-solved GPCR structures are arranged in a sortable table, which includes the following data: 1) PDB ID, 2) UniProt ID with species, 3) method used to solve the structure, 4) resolution (if applicable), 5) release date, and 6) reference with PubMed ID. The user can browse all structures in a single table or by class. The former method is useful for checking the latest GPCRs or sorting structures by resolution, while the latter method allows for a more controlled browsing experience.

If browsing by class, the structures of each GPCR class are stratified into tables grouped by GPCR name. For example, the 5-hydroxytryptamine receptor 2C has two solved structures (Figure 2.6A).

By clicking on the GPCR structure thumbnail, users will be directed to a popup page that includes further data about the structure, such as mutations, fusion/associated proteins, and ligands. A link to our in-house GPCR-ligand database, GLASS,[28] provides a more comprehensive set of pharmacological data corresponding to the user's GPCR of interest. Moreover, an embedded structure viewer displays the processed structure, along with any applicable crystallographic ligands. For convenience, PDB files of the original structure, GPCR-only structure, or ligands can also be downloaded with a single click from the table. Additionally, there are GPCR-specific services that can be accessed through various popup pages. First, the user can view pre-superposed structures of a GPCR of interest by clicking 'Overlay Structures' (Figure 2.6B). Second, any small molecule or peptide ligands associated with a structure can be seen by clicking on 'Display Ligands' (Figure 2.6C). Third, users can click 'Download Structures', which will lead to a popup page allowing the users to customize the structures they wish to download. This feature will be

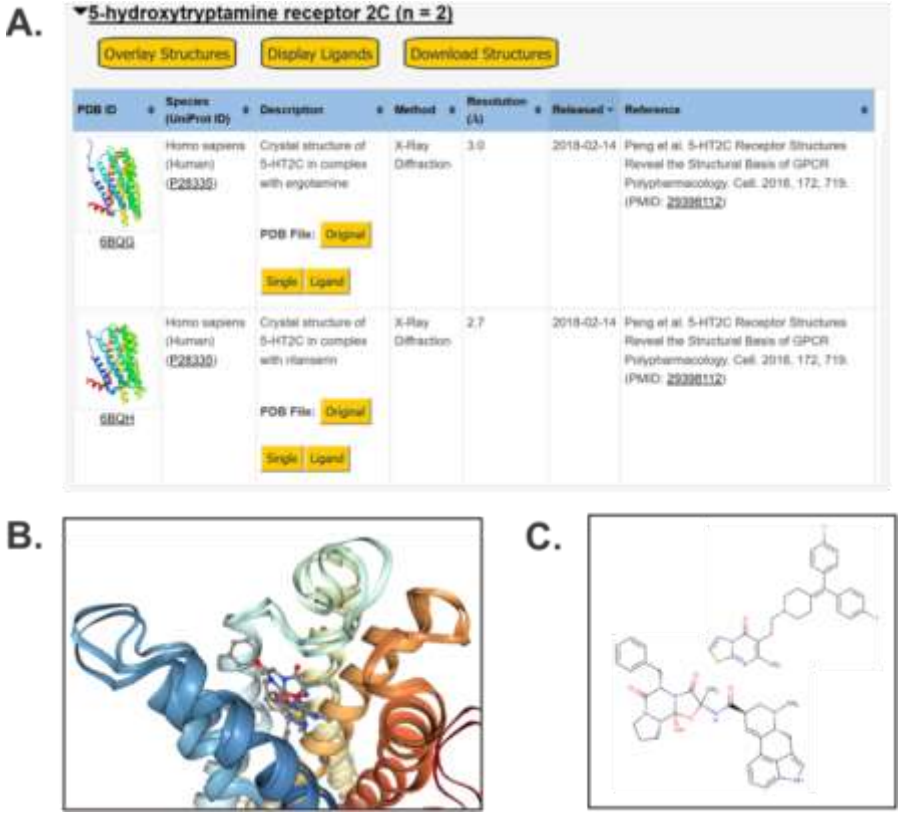


Figure 2.6 - Experimentally-Solved GPCR Structures on GPCR-EXP. The 5-hydroxytryptamine receptor 2C is used here as an example. (A) Structural data is shown in a sortable table. (B) Clicking on 'Overlay Structures' will allow the user to view superposed structures, while (C) clicking on 'Display Ligands' will pull up and display any associated crystallographic ligands.

important for users interested in examining structures of interest in more details with advanced structure visualization software.

3.2.2 Predicted Structures

GPCR-I-TASSER was run to generate structure models for 1,076 GPCRs from the human genome. Like the experimentally-solved structures, tables for these structures are also sortable and contain the following data: 1) UniProt ID, 2) GPCR name, 3) C-score of the best structure model, 4) estimated TM-score, and 5) estimated RMSD. The user can browse all structures in one table or by class. If browsing by all structures, two additional columns (family / subfamily) are provided for reference. Class A GPCRs are stratified by subfamily, and an example with the acetylcholine receptors is given in Figure 2.7A. On the contrary, each of the non-Class A GPCRs are grouped into their respective tables because they represent a minority of GPCRs.

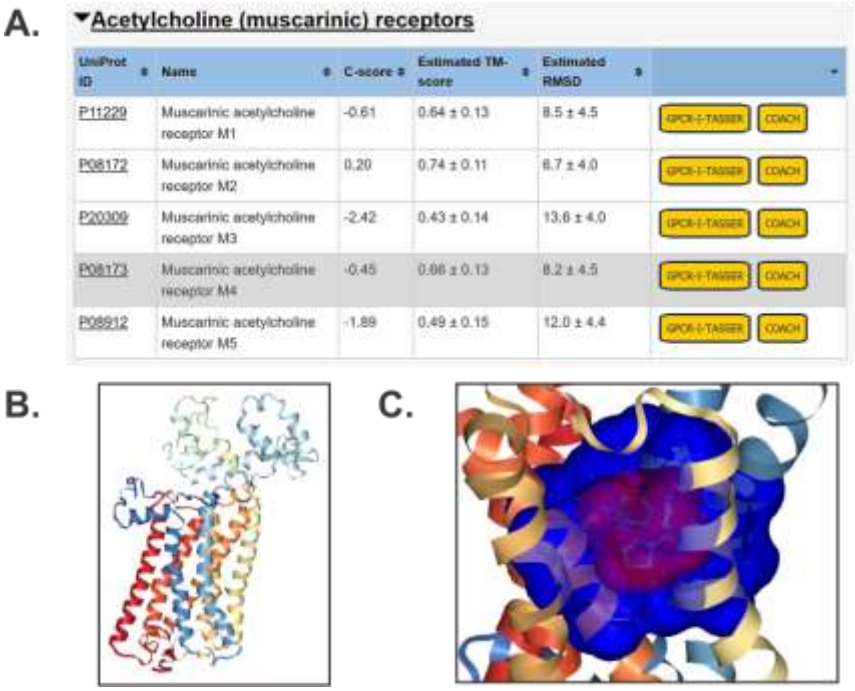


Figure 2.7 - Predicted GPCR Structures on GPCR-EXP. The acetylcholine receptors are used here as an example. (A) Structural data is shown in a sortable table. Clicking on 'GPCR-I-TASSER' or 'COACH' will provide the user with detailed information on the predicted structure and predicting binding site, respectively. (A) In the former, the user can view models (top-ranked model shown) . (B) In the latter, the user can select and view a predicted binding site of interest. The blue surface represents the predicted binding site, while the red surface is a possible binding ligand.

Detailed information on a structure model of interest can be accessed by clicking on the 'GPCR-I-TASSER' button. This provides various predictions alongside the structure, such as secondary structure, solvent accessibility, normalized B-factor, and local accuracy. The top 10 templates used by GPCR-I-TASSER and associated statistics are also given, while the top 5 models are available for viewing and download. The top-ranked model is shown for muscarinic acetylcholine receptor M1 as an example in Figure 2.7B. Additionally, a visual assessment of the local structure quality of the model is given by ResQ,[29] an algorithm which makes predictions about residue-level model quality and B-factor estimations through the combination of sequence- and structure-based profiling. Alongside the structure models, binding site predictions from COACH[12-13] based on the top-ranked models are also shown, where detailed information can be retrieved by clicking on the 'COACH' button. Here, the predicted binding sites from COACH and each of its component methods are displayed along with all related statistics from the prediction. Furthermore, the binding pocket can be visually examined from the GPCR on the web interface (Figure 2.7C).

3.3 Downloading GPCR-EXP

All data compiled and processed from PDB are freely available for download on the web server. Bulk downloads include a TSV file of all data pertaining to experimentally-solved GPCR structures, as well as a text file containing detailed statistics. Moreover, the following PDB file compilations are also available for download: 1) all original PDB files, 2) GPCR structures modified to contain single chain, 3) superposed GPCRs along with respective ligands, 4) GPCR fragments, and 5) predicted GPCR structures from human genome.

3.4 Maintenance of GPCR-EXP

The update pipeline for the experimentally-solved GPCRs is almost fully automated, apart from the manual inspection of ligands. As such, updating the database is streamlined and requires minimal oversight for implementation. As of the time of writing, we are running weekly updates for this data. On the other hand, GPCR-I-TASSER and COACH predictions on all GPCRs from the human genome are only updated annually due to its high computational cost.

4. Summary

We have developed a database, GPCR-EXP, which combines data for experimental and predicted structures related to GPCRs. Overall, GPCR-EXP contains the following unique features:

(1) Semi-manual curation of the data allows for a quick and high-quality update of the database content. Most of the updating process is implemented through a custom script, while the ligand data is manually cross-examined with the literature using PubMed, in order to avoid the inclusion of ions, detergents, and other non-pharmacological ligands.

(2) Computational models from cutting-edge methods are provided for the GPCRs that do not have experimental structure available. In particular, GPCR-I-TASSER creates structural models using iterative fragment assembly, which allows reliably modeling on some of the distant-homologous protein targets, while COACH generates high-quality binding site predictions.

(3) Given the critical importance of the quality information to the biomedical users, a confidence scoring system is given to estimate the global quality of all the predicted GPCR-I-TASSER models. Meanwhile, a reliable B-factor modeling method,[29] which was validated in the last CASP experiment,[30] is used to estimate the accuracy of the local structures. These model quality annotations are important in better assisting the use of the predicted models for users working in the biomedical sciences.

(4) All experimental PDB structures with a full transmembrane domain have been preprocessed to have a single chain. Fusion proteins (such as lysozyme) have been programmatically removed as well.

(5) Experimental structures of the same GPCR have been pre-superposed and are made viewable on the web page, as well as downloadable. This feature will be useful for researchers interested in studying differences in the binding pocket of different ligand-bound complexes, as well as between active and inactive state structures.

(6) All ligand data have been curated and are viewable on the web page. Additionally, PDB files of the crystallographic poses are provided.

(7) All data are freely available for download. Batch PDB files of all the original structures, single chain structures, and superposed structures are provided with the detailed statistics and general data.

(8) GPCR-EXP is updated weekly to account for the increasing frequency of release of new GPCR structures.

In conclusion, all of these features address the aforementioned issues brought up earlier. With semi-manual curation of data, weekly updates are possible because of the programmatic nature of process, while human error is significantly cut down. Additionally, the inclusion of a user-friendly interface allows researchers to easily access whatever data is desired in an expedient fashion. For these reasons, we believe that GPCR-EXP will be an invaluable resource to researchers in drug discovery. Interestingly, GPCR-EXP was originally conceived in 2015 and has managed to garner mentions in several scientific publications[31-36] from biomedical users, despite the manuscript describing the database yet to be published. With the latest, more-comprehensive version of our database, we are confident to have a lasting impact in GPCR research, and hence in drug discovery on the whole.

2.6 References

1. Dorsam, R. T.; Gutkind, J. S., G-protein-coupled receptors and cancer. *Nature reviews cancer* **2007,** *7* (2), 79.
2. Rose, P. W.; Prlić, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z., The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic acids research* **2016**, gkw1000.
3. Pándy-Szekeres, G.; Munk, C.; Tsonkov, T. M.; Mordalski, S.; Harpsøe, K.; Hauser, A. S.; Bojarski, A. J.; Gloriam, D. E., GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic acids research* **2017,** *46* (D1), D440-D446.
4. Miszta, P.; Pasznik, P.; Jakowiecki, J.; Sztyler, A.; Latek, D.; Filipek, S., GPCRM: a homology modeling web service with triple membrane-fitted quality assessment of GPCR models. *Nucleic acids research* **2018**.
5. Sandal, M.; Duy, T. P.; Cona, M.; Zung, H.; Carloni, P.; Musiani, F.; Giorgetti, A., GOMoDo: a GPCRs online modeling and docking webserver. *PLoS One* **2013,** *8* (9), e74092.
6. Esguerra, M.; Siretskiy, A.; Bello, X.; Sallander, J.; Gutiérrez-de-Terán, H., GPCR-ModSim: A comprehensive web based solution for modeling G-protein coupled receptors. *Nucleic acids research* **2016,** *44* (W1), W455-W462.
7. Worth, C. L.; Kreuchwig, F.; Tiemann, J. K.; Kreuchwig, A.; Ritschel, M.; Kleinau, G.; Hildebrand, P. W.; Krause, G., GPCR-SSFE 2.0—a fragment-based molecular modeling web tool for Class A G-protein coupled receptors. *Nucleic acids research* **2017,** *45* (W1), W408-W415.
8. Eswar, N.; Webb, B.; Marti-Renom, M. A.; Madhusudhan, M.; Eramian, D.; Shen, M. y.; Pieper, U.; Sali, A., Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics* **2006,** *15* (1), 5.6. 1-5.6. 30.
9. Wu, S.; Zhang, Y., LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic acids research* **2007,** *35* (10), 3375-3382.
10. Kufareva, I.; Rueda, M.; Katritch, V.; Stevens, R. C.; Abagyan, R.; Dock, G., Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. *Structure* **2011,** *19* (8), 1108-1126.

11. Zhang, J.; Yang, J.; Jang, R.; Zhang, Y., GPCR-I-TASSER: a hybrid approach to G protein-coupled receptor structure modeling and the application to the human genome. *Structure* **2015,** *23* (8), 1538-1549.
12. Yang, J.; Roy, A.; Zhang, Y., Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013,** *29* (20), 2588-95.
13. Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research* **2013,** *41* (Database issue), D1096-103.
14. Consortium, U., UniProt: the universal protein knowledgebase. *Nucleic acids research* **2016,** *45* (D1), D158-D169.
15. Kraulis, P. J., MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *Journal of applied crystallography* **1991,** *24* (5), 946-950.
16. Landrum, G., RDKit: Open-source cheminformatics. *Online). http://www. rdkit. org. Accessed* **2006,** *3* (04), 2012.
17. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **2005,** *33* (7), 2302-9.
18. Rose, A. S.; Hildebrand, P. W., NGL Viewer: a web application for molecular visualization. *Nucleic acids research* **2015,** *43* (W1), W576-W579.
19. Hollenstein, K.; Kean, J.; Bortolato, A.; Cheng, R. K.; Doré, A. S.; Jazayeri, A.; Cooke, R. M.; Weir, M.; Marshall, F. H., Structure of class B GPCR corticotropin-releasing factor receptor 1. *Nature* **2013,** *499* (7459), 438.
20. Siu, F. Y.; He, M.; De Graaf, C.; Han, G. W.; Yang, D.; Zhang, Z.; Zhou, C.; Xu, Q.; Wacker, D.; Joseph, J. S., Structure of the human glucagon class B G-protein-coupled receptor. *Nature* **2013,** *499* (7459), 444.
21. Wang, C.; Wu, H.; Katritch, V.; Han, G. W.; Huang, X.-P.; Liu, W.; Siu, F. Y.; Roth, B. L.; Cherezov, V.; Stevens, R. C., Structure of the human smoothened receptor bound to an antitumour agent. *Nature* **2013,** *497* (7449), 338.
22. Doré, A. S.; Okrasa, K.; Patel, J. C.; Serrano-Vega, M.; Bennett, K.; Cooke, R. M.; Errey, J. C.; Jazayeri, A.; Khan, S.; Tehan, B., Structure of class C GPCR metabotropic glutamate receptor 5 transmembrane domain. *Nature* **2014,** *511* (7511), 557.
23. Wu, H.; Wang, C.; Gregory, K. J.; Han, G. W.; Cho, H. P.; Xia, Y.; Niswender, C. M.; Katritch, V.; Meiler, J.; Cherezov, V., Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. *Science* **2014,** *344* (6179), 58-64.
24. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H.-J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K., High-resolution crystal structure of an engineered human β2-adrenergic G protein–coupled receptor. *science* **2007,** *318* (5854), 1258-1265.
25. Palczewski, K.; Kumasaka, T.; Hori, T.; Behnke, C. A.; Motoshima, H.; Fox, B. A.; Le Trong, I.; Teller, D. C.; Okada, T.; Stenkamp, R. E., Crystal structure of rhodopsin: AG protein-coupled receptor. *science* **2000,** *289* (5480), 739-745.
26. Cheng, Y., Single-particle cryo-EM at crystallographic resolution. *Cell* **2015,** *161* (3), 450-457.
27. Egloff, P.; Hillenbrand, M.; Klenk, C.; Batyuk, A.; Heine, P.; Balada, S.; Schlinkmann, K. M.; Scott, D. J.; Schütz, M.; Plückthun, A., Structure of signaling-competent neurotensin receptor 1 obtained by directed evolution in Escherichia coli. *Proceedings of the National Academy of Sciences* **2014,** *111* (6), E655-E662.

28. Chan, W. K.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Ozgur, A.; Zhang, Y., GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* **2015,** *31* (18), 3035-42.
29. Yang, J.; Wang, Y.; Zhang, Y., ResQ: an approach to unified estimation of B-factor and residue-specific error in protein structure prediction. *Journal of molecular biology* **2016,** *428* (4), 693-701.
30. Zhang, C.; Mortuza, S.; He, B.; Wang, Y.; Zhang, Y., Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. *Proteins: Structure, Function, and Bioinformatics* **2018,** *86*, 136-151.
31. Schneider, S.; Provasi, D.; Filizola, M., The dynamic process of drug–GPCR binding at either orthosteric or allosteric sites evaluated by metadynamics. In *G Protein-Coupled Receptors in Drug Discovery*, Springer: 2015; pp 277-294.
32. D Cromie, K.; Van Heeke, G.; Boutton, C., Nanobodies and their use in GPCR drug discovery. *Current topics in medicinal chemistry* **2015,** *15* (24), 2543-2557.
33. Cooke, R. M.; Brown, A. J.; Marshall, F. H.; Mason, J. S., Structures of G protein-coupled receptors reveal new opportunities for drug discovery. *Drug discovery today* **2015,** *20* (11), 1355-1364.
34. Stockert, J. A.; Devi, L. A., Advancements in therapeutically targeting orphan GPCRs. *Frontiers in pharmacology* **2015,** *6*, 100.
35. Jakubík, J.; El-Fakahany, E. E.; Doležal, V., Towards predictive docking at aminergic G-protein coupled receptors. *Journal of molecular modeling* **2015,** *21* (11), 284.
36. Thomas, T.; Chalmers, D. K.; Yuriev, E., Homology Modeling and Docking Evaluation of Human Muscarinic Acetylcholine Receptors. In *Muscarinic Receptor: From Structure to Animal Models*, Springer: 2016; pp 15-35.

CHAPTER 3.
GLASS: A Comprehensive Database for Experimentally-Validated GPCR-Ligand Associations[1]

1. Introduction

G protein-coupled receptors (GPCR) represent one of the largest families of transmembrane proteins that bind extracellular molecules and activate intracellular signal transduction pathways, which mediate many physiological functions through their interaction with heterotrimeric G proteins. Many human diseases, including cancer and diabetes, have been found to be associated with the malfunction of the biological roles of GPCRs.[2] Currently, approximately 30-50% of drugs on the market target GPCRs, making them one of the most attractive membrane receptors for drug development.[3-4] While experiment-based assays for novel chemical compounds remain the standard procedure for drug discovery, *in silico* screening is gaining increasing acceptance as an important complementary method to narrow down the drug searching scope and to guide experimental design. Another advantage of the computational approach is due to its high speed and low cost, which enables high-throughput and large-scale database screening.[5]

Both the experimental and computational drug discovery approaches rely on existing GPCR-ligand experimental data to provide insight for screening and selecting new drugs. A variety of GPCR-orientated databases, such as GPCRDB,[6] TinyGRAP,[7] GPCR-OKB,[8] GDD,[9] and GPCR-RD,[10] have been developed, which generated important impacts on various molecule-level studies on the elucidation of GPCR structure and function.

There are however very few databases that can provide comprehensive resources for GPCR-ligand interactions that are essential in assisting GPCR virtual screening studies.[11-13] One difficulty in

[1] This chapter was adapted from a previously-published work in Bioinformatics, entitled "GLASS: a comprehensive database for experimentally validated GPCR-ligand associations" by WKB Chan, H Zhang, J Yang, JR Brender, J Hur, A Özgür, and Yang Zhang. WKB Chan and H Zhang shared co-authorship in the study.

developing such databases is that the GPCRs can be associated with a large number of ligands in various binding affinities, and the GPCR-ligand association data in many chemical libraries are often mixed with various false-positives. A collection of GPCR-ligand associations with stringent experimental validations and careful human curation is essential to ensure the quality of the datasets. Second, with the success of the sequencing and structural genomics projects, the number of available GPCR and ligand interactions increase rapidly. But most of the new studies are scattered in a wide spread of publications and archives, which makes it difficult to keep the databases up to date. For example, GLIDA[14] was a useful GPCR-ligand binding database designed for chemical genomic drug discovery; but it has ceased updates to its server since October 2010. The current GLIDA library contains around 39,000 GPCR-ligand entries, whereas the amount of unique GPCR-ligand interactions available in the literature in our estimation is above 500,000. The missing of such a substantial amount of new data significantly degrades the usefulness of the databases to the experimental and computational drug discovery studies.

In this study, we have developed a new GPCR-ligand association (GLASS) database for use as a general platform in assisting GPCR-related drug screening studies. Drawing from multiple primary data sources, GLASS focuses on a comprehensive and yet precise collection of the experimentally-validated GPCR-ligand interactions with strong affinities. All the GPCR-ligand association data are manually-curated and made freely-available to the community.

2. Data and Methods

The GPCR-ligand association data in GLASS consist of two major resources. The first resource consists of five primary pharmacological datasets from ChEMBL,[15] BindingDB,[16] IUPHAR,[17] DrugBank,[18] and PDSP,[19] which contain various bioactive ligand and protein interaction data. A flowchart of the construction of GLASS is depicted in Figure 3.1.

2.1. Database Recombination Pipeline

A list of all reviewed UniProt IDs pertaining to GPCRs was first collected from UniProtKB.[20] Data relevant to each GPCR, such as species, gene name, and primary sequence, were simultaneously extracted. We used a combination of synonymous GPCR names from IUPHAR and UniProtKB.

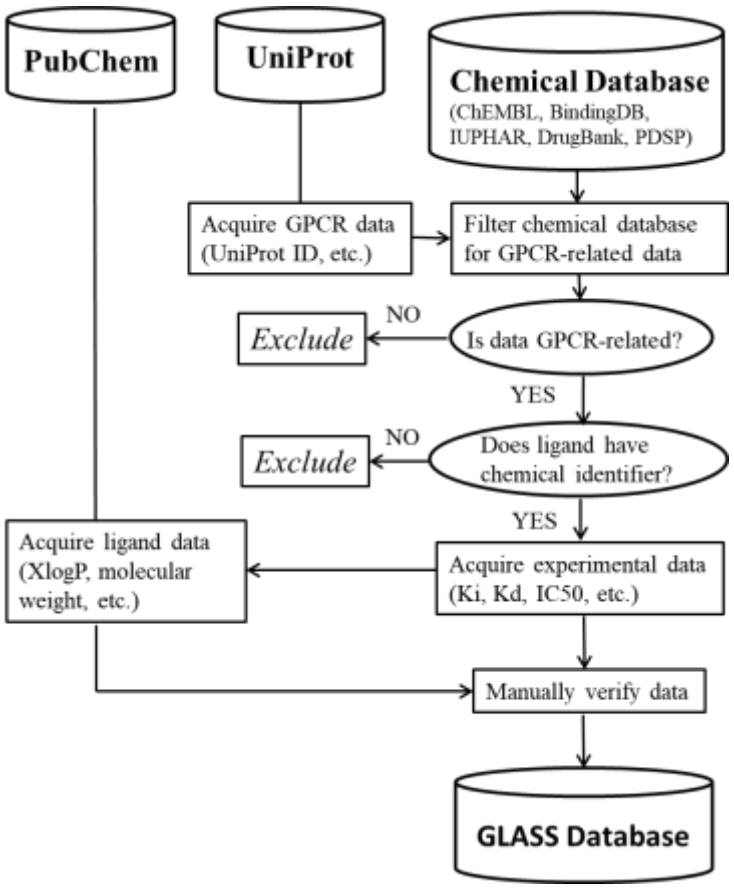


Figure 3.1 - Flowchart for the Construction of GLASS Database.

In the second step, flat line databases were downloaded from the pharmacological databases of ChEMBL, BindingDB, IUPHAR, DrugBank, and PDSP. Data entries were filtered only for GPCR-related ones using UniProt ID and compiled together. The ligands without chemical identifiers were eliminated. Meanwhile, the statistical analysis of the distributions among the Ki, Kd, IC_{50}, and EC_{50} values revealed that the majority (>95%) of the experimental ligand-GPCR associations have the activity values below 10 μM. Thus, an activity filter was implemented, i.e. the entries with a Ki, Kd, IC_{50}, and EC_{50} higher than 10 μM were designated as inactive compounds, in order to sieve out weak and suspicious GPCR-ligand associations. Once an entry passes all criteria, records on the pharmacological data (e.g. ligand activities), the references to the original literature of study, and the chemical identifiers such as SMILES or InChI, are collected from the original pharmacological databases. Overall, the compiled compound files for both active and inactive compounds were generated and included as download for users.

2.2. Architecture of the GLASS Library

The GLASS database was built using MySQL, while the Internet webpage was augmented with a combination of Perl and Python CGI scripts to facilitate the communication of the interfaces with the MySQL database.

For each GPCR-ligand association, relevant chemical information, such as XlogP, molecular weight, hydrogen bond acceptor and donor, 2D structure image, synonyms, and IUPAC name, were extracted from PubChem using the compound identifier (CID) of each ligand via their Chemical Identifier Exchange service. The 3D SDF files were generated from respective canonical SMILES strings using Open Babel.[21]

For the GPCRs from the human genome, the associated conditions and diseases from experiments were compiled from TTD[22] when available. The 3D structure information is provided for each GPCR by cross-linking to the PDB when the experimental structures are available; otherwise, a link is provided to the GPCR-HGmod,[23] a comprehensive human GPCR structure database with all models constructed by the GPCR-I-TASSER algorithm assisted with the mutagenesis experimental restraints. A confidence score is provided for each of the GPCR structure models to calibrate the quality. An NGL Viewer image is created for each GPCR to allow users to view the 3D structure of the receptor.

To facilitate comparative interaction studies, GLASS provides an interactive search engine to collect homology ligand/compounds through either substructure or chemical similarity. Using the JSME molecular editor,[24] users are allowed to draw a chemical structure of the compounds, which is then converted into a SMILES string. Subsequently, it is transferred to Open Babel for either a substructure or similarity search against the indexed ligands. An SDF file is pre-created containing all ligand indexes in order to expedite the searching process. For the chemical similarity search, users are able to select the Tanimoto coefficient cutoffs. The resultant ligands are returned as SMILES strings. Finally, the SMILES strings are used as probes to search against the database in order to collect homologous ligands, which are returned as images of the chemical structure and their names. Tanimoto coefficients are returned, as well, if the similarity search was selected.

3. Results

3.1 GLASS in Numbers

As of the time of writing, GLASS contains 994,751 GPCR-ligand entries, collected from multiple sources of experiments. Some associations appear more than once in different experiments. After removing the redundant entries, there are 549,792 unique associations each containing a species-specific GPCR paired with an interacting ligand (444,959 unique associations remain if removing the redundancy across species and accounting for orthologues).

A total of 3,056 GPCR entries in GLASS were extracted from UniProt,[20] where 733 GPCRs have at least one ligand associations. The other 2,323 GPCR entries have no ligand associated data in the experiment literature as of the present time. Among the GPCR's with ligand associations, there are approximately 750 different types of ligand/compound associations per receptor on average; but the median value is only 77 due to the fact that several receptor families have a dominantly high number of ligand associations (see below). The total number of unique ligands in GLASS is 335,040. A summary of the current GLASS database is presented in Table 3.1.

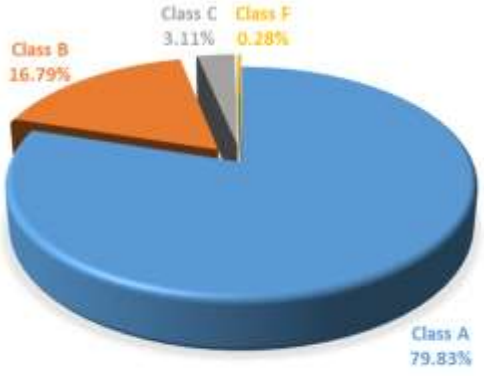


Figure 3.2 - Distribution of GPCR-Ligand Data in GLASS by Family. All values presented as percentage of total. Fungal, cyclic AMP, slime mold, OA, and T2R receptors, which have insufficient (<10 entries) or no data, were excluded from the plot.

Table 3.1 - Summary of GLASS Database

Type of entry	Number of entries
All GPCRs	3,056
With ligand association	733
Without ligand association	2,323
Unique ligands	335,040
Drug-like ligands	238,027
All GPCR-ligand associations	994,751
Unique associations	549,792

Most of the ligand asso ciations in GLASS are skewed towards the Class-A rhodopsin-like family of GPCRs, which makes up approximately 80% of the association data (Figure 3.2). The top four receptors in the rhodopsin-like family, all of which have more than 65,000 ligand associations, are from serotonin, adenine and adenosine nucleotide, opioid, and dopamine receptors. These

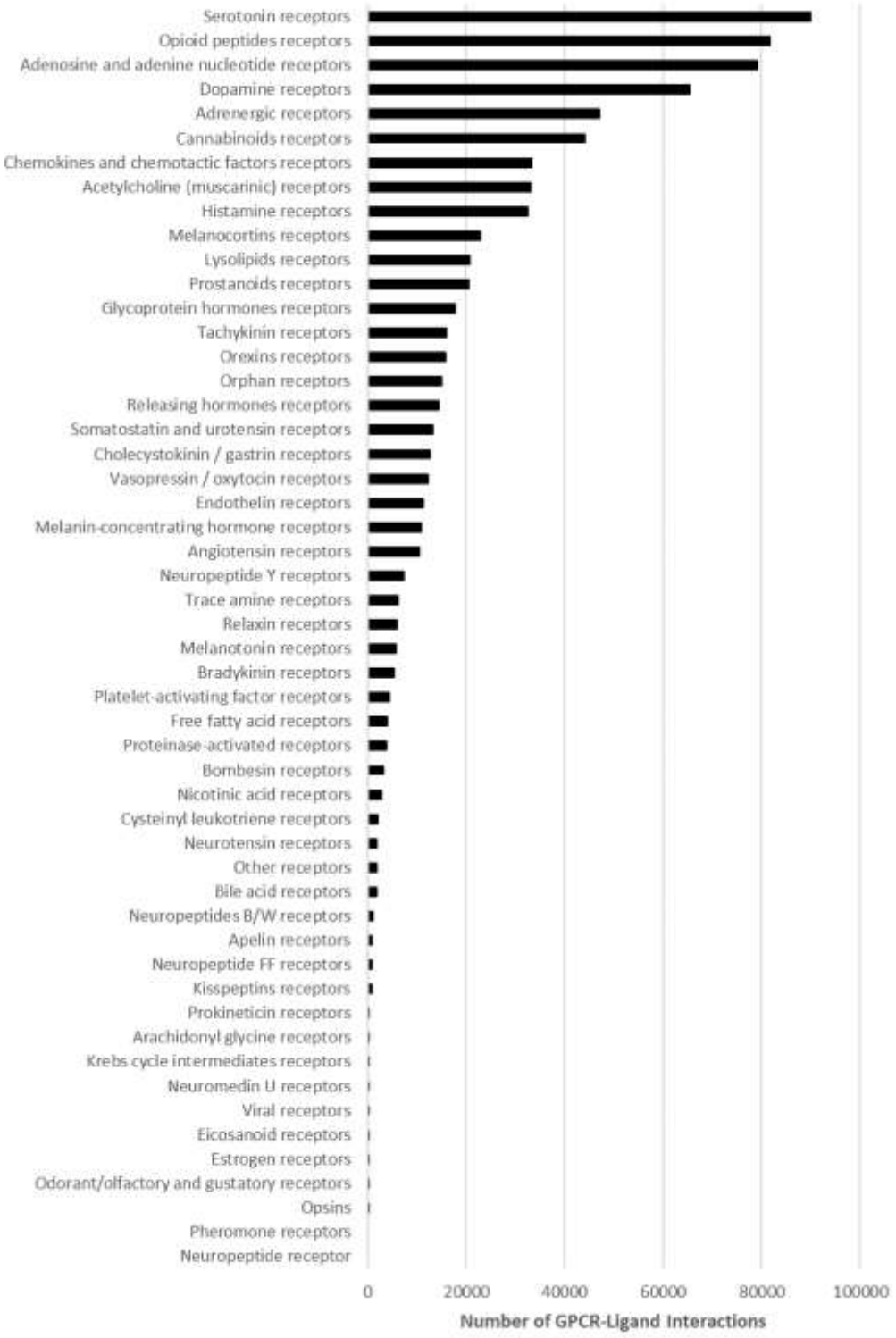


Figure 3.3 - Histogram of Ligand Associations with GPCRs in the Class A (Rhodopsin-like) Family. The figure only displays the GPCRs with at least one ligand associations.

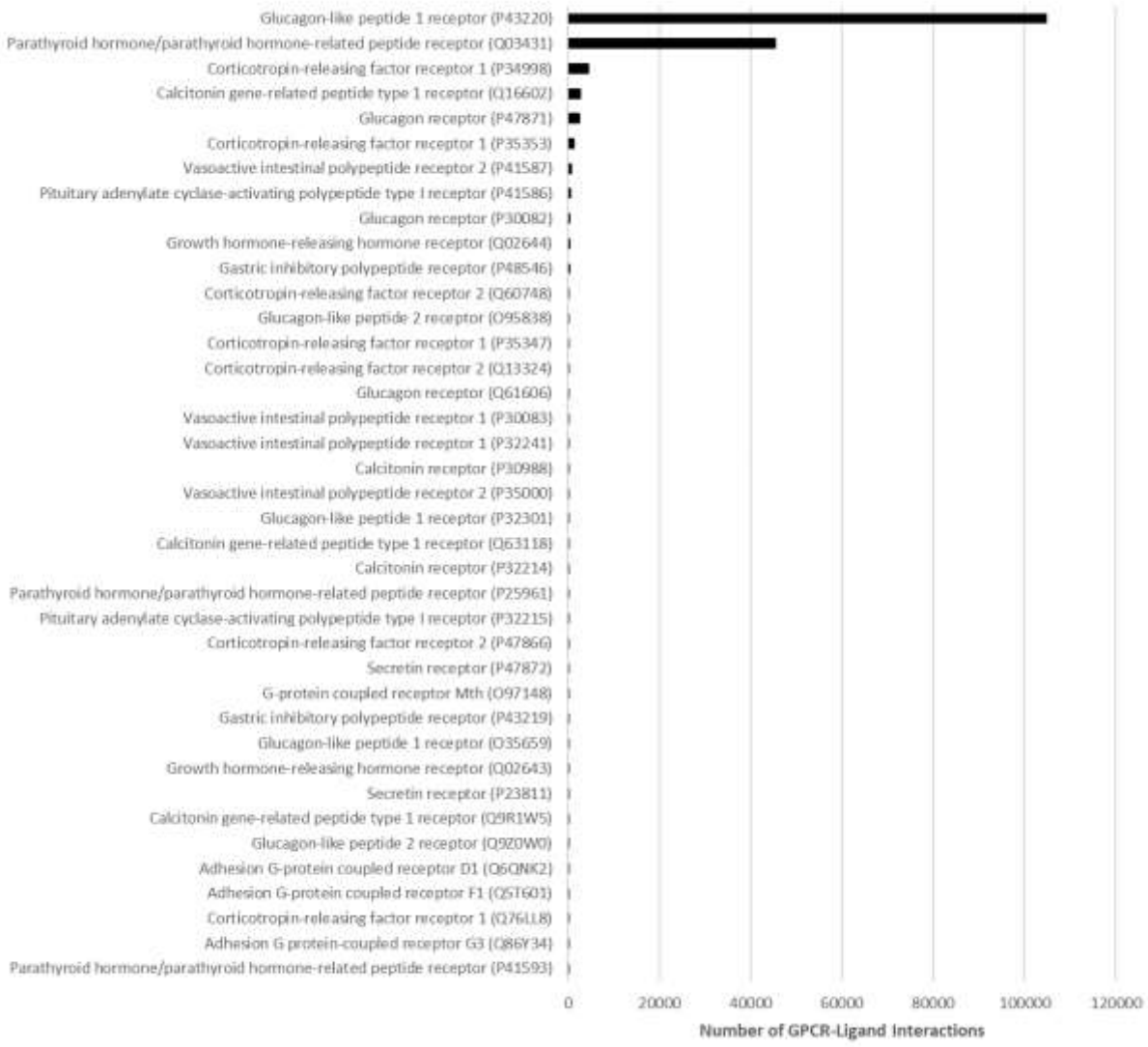


Figure 3.4 - Histogram of Ligand Associations with GPCRs in the Class B (Secretin) Family. The figure only displays the GPCRs with at least one ligand associations.

receptors also represent the set of the most popularly studied GPCRs in literature due to their importance in pharmaceutical applications and research. A histogram of the ligand associations for the entire Class-A family is shown in Figure 3.3.

The non-rhodopsin-like families of GPCRs constitute a far lesser proportion of ligand associations. Nevertheless, the human glucagon-like peptide 1 receptor from the Class-B secretin family contains the most abundant GPCR-ligand associations among all the human GPCRs, containing over 100,000 entries. The other non-rhodopsin-like GPCRs with more than 2,000 GPCR-ligand

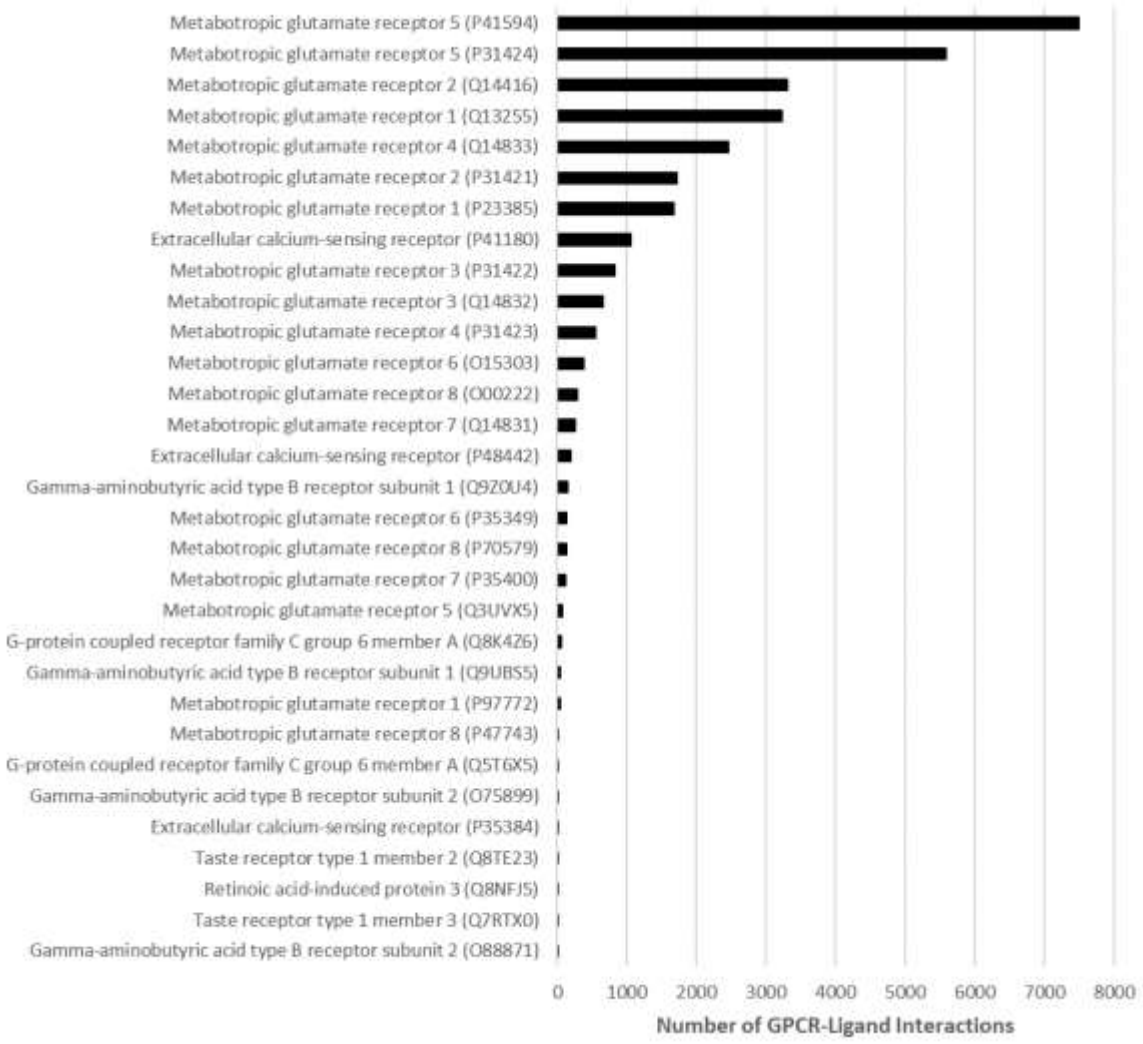


Figure 3.5 - Histogram of Ligand Associations with GPCRs in the Class C (Metabotropic Glutamate/Pheromone) Family. The figure only displays the GPCRs with at least one ligand associations.

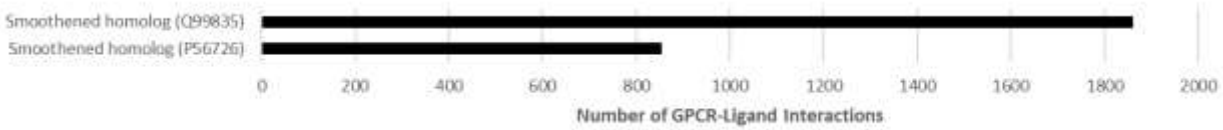


Figure 3.6 - Histogram of Ligand Associations with GPCRs in the Class F (Frizzled /Smoothened) Family. The figure only displays the GPCRs with at least one ligand associations.

associations are the metabotropic glutamate group of receptors from the Class-C metabotropic glutamate/pheromone family. There are only two members (UniProt ID: Q88935 and P56726) from the Class-F family that have associated experimental data, while little to no GPCR-ligand associations are found for the GPCRs from the fungal mating pheromone (Class-D), cyclic AMP

(Class-E), slime mold, ocular albinism (OA), and taste receptor (T2R) families. Figures 3.4-3.6 list the detailed data distributions of ligand associations for Class-B, C, and F families. This highly uneven ligand association distribution explains the reason that the median number of ligands per receptor is much lower than the average.

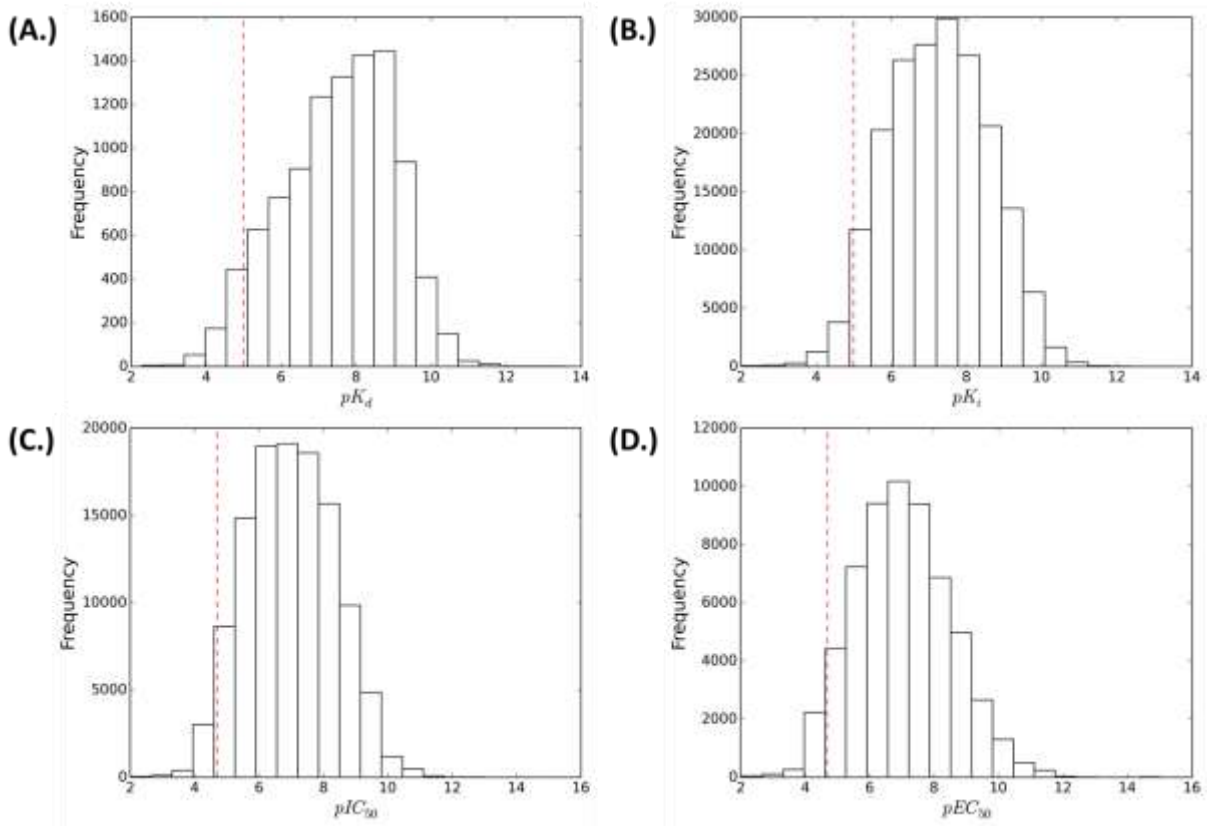


Figure 3.7 - Activity Distributions of Ligands from GLASS Database. Pharmacological data covered include (A.) K_d, (B.) K_i, (C.) IC_{50}, and (D.) EC_{50}. All experimental values are converted to the negative log form. Thus, $pK_d = -\log K_d$. The dashed red line indicates the activity cutoff used for the creation of ligand sets in MAGELLAN, which was 10 µM for K_i and K_d and 20 µM for IC_{50} and EC_{50}.

3.2 Survey of Experimental Data

All experimental data related to K_d, K_i, IC_{50}, and EC_{50}, some common experimental measures in pharmacology, were collected for analysis and summarized in Figure 3.7. All four of these followed an approximately normal distribution when transformed to their negative log form and conformed to expectations. Typical cutoffs of 10 µM were used as activity filters for GLASS database in creating a filtered chemical subset. However, a cutoff of 20 µM for IC_{50} and EC_{50} was

used in the following chapter with MAGELLAN and will be explained in there. Overall, there was only a small reduction in the amount of data after filtration. The most shocking revelation while surveying the publicly-available experimental data was the inconsistencies between studies. Many experimental values for the same ligand and receptor were found to be over 2 orders of magnitude different from one another (Figure 3.8). This is consistent with a previous study on IC_{50} data from ChEMBL.[25] The variation is understandable for measures such as EC_{50} and IC_{50} because they are assay specific and comparable only under certain conditions. However, for constants like K_d and K_i, the disagreements are deeply troubling and potentially reflective of the problem of irreproducibility in science.

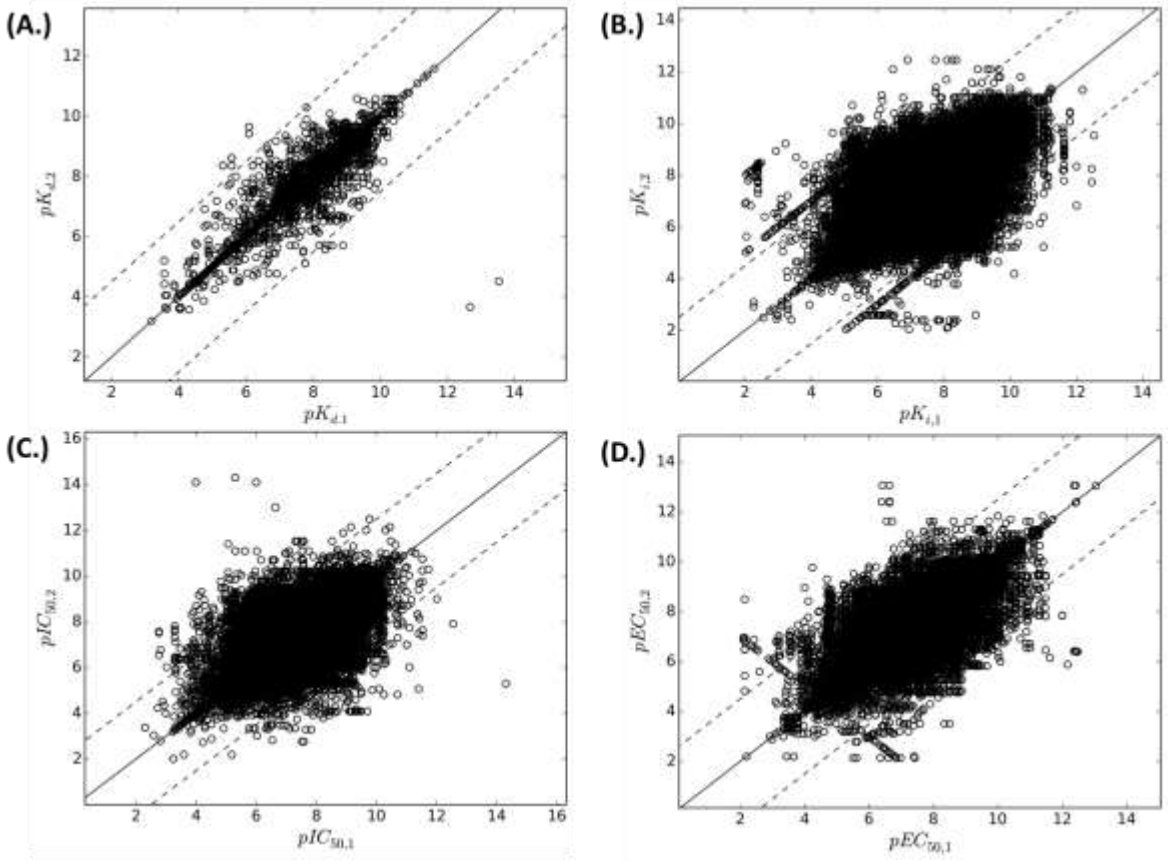


Figure 3.8 - Pairs of Activity Data from GLASS Database. Pharmacological data covered include (A.) K_d, (B.) K_i, (C.) IC_{50}, and (D.) EC_{50}. Non-redundant pairs of experimental values are shown between all GPCRs and their ligands. All experimental values are converted to the negative log form. Thus, $pK_d = -\log K_d$. The outer dashed lines represent a 2.5 log unit threshold.

3.3. Database Features

The GLASS database is updated every month, and all data are made freely available at: http://zhanglab.ccmb.med.umich.edu/ GLASS/. Three features have been developed for searching, browsing, or downloading of the GPCR-ligand association data in GLASS, as shown in Figure 3.8, which are outlined in the following.

3.3.1. Searching GLASS

An efficient search function is essential to the development of biomedical databases. GLASS provides three options on the home page for searching the database based on three types of queries: (1) GPCR-based, (2) ligand-based, and (3) GPCR-ligand-based. Users can choose these options by selecting the radio button of interest before or after typing the desired input (Figure 3.8).

Figure 3.8 - A screen shot of the GLASS homepage showing options for searching, browsing, and downloading of database-related data.

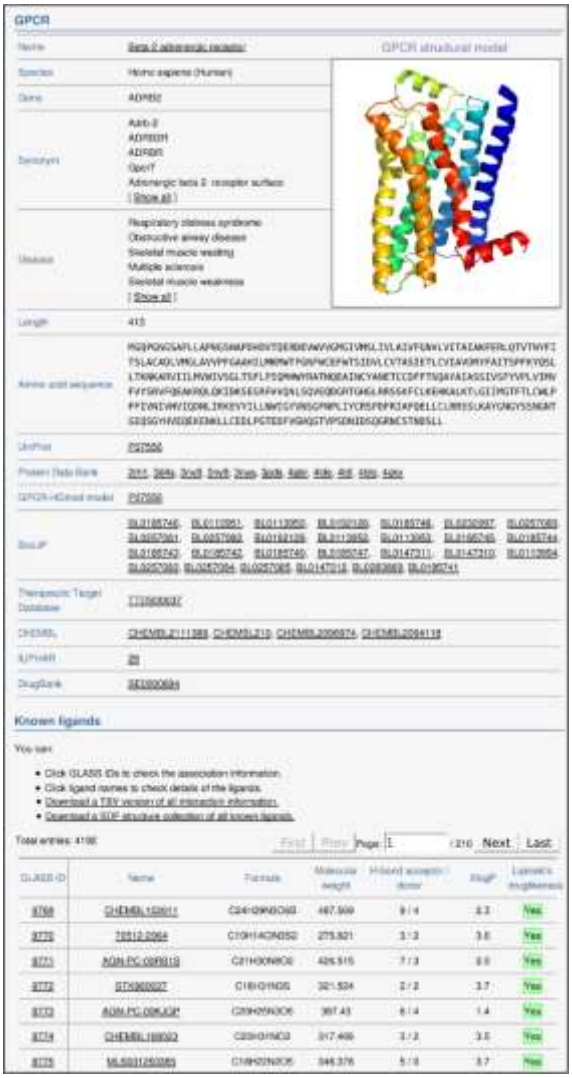


Figure 3.9 - Illustration of the output of GPCR-based search from GLASS. This example is from the human β2 adrenergic receptor, where ligands associated with the receptor are listed at the bottom of the page. The 3D structure shown was from the PDB entry solved by Cherezov et al.[1]

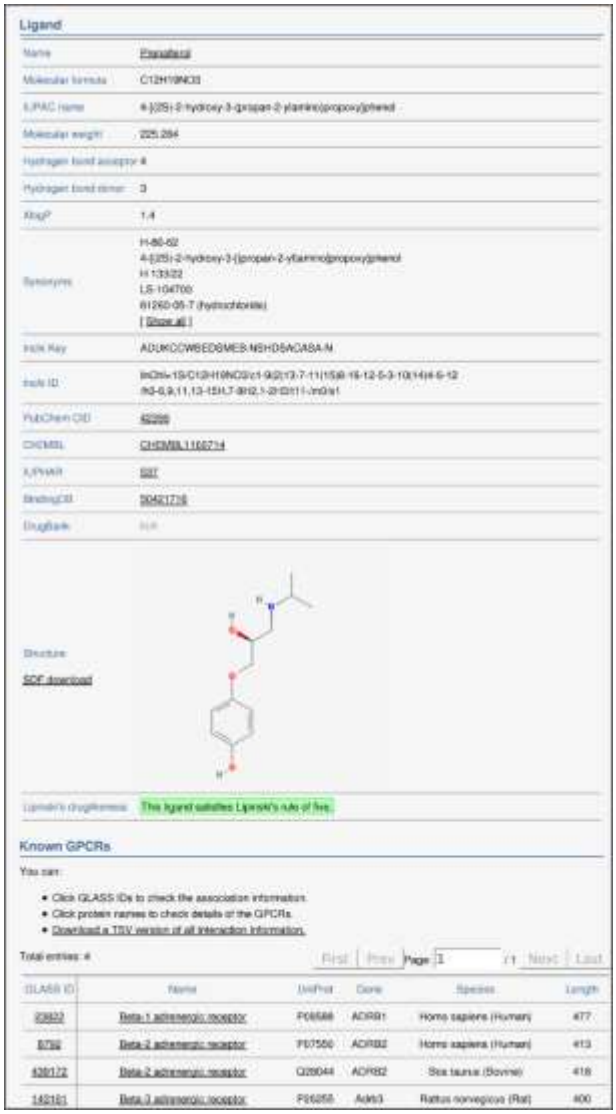


Figure 3.10 - Illustration of the output page for the ligand-based search on GLASS. The ligand shown is prenalterol, one of the associationed ligands for the human β2 adrenergic receptor in Figure 5. GPCRs bound with prenalterol are listed at the bottom of the page.

Using the GPCR-based search, users can search for a GPCR of interest using a variety of inputs, including UniProt ID, gene name, or associated medical conditions. Clicking on the 'Search' button takes the user to a page listing all GPCRs that match the query; clicking and following the link of the GPCR of interest will bring the user to a detailed page with GPCR-related information, including GPCR name, species, gene name, synonyms, associated diseases, primary sequence and its length, atomic structural model, and database identifiers. All ligands that are associated with

the GPCR are listed at the bottom of the page. Figure 3.9 presents an example of output of the GPCR-based search from the human β2 adrenergic receptor.

The ligand-based search requires knowledge of the name, chemical identifier, or PubChem ID of the ligand of interest. Clicking on the 'Search' button will bring the user to a page of results of all ligands matching the query. Clicking and following the link of the ligand of interest will bring up a detailed page with the ligand name, molecular formula, IUPAC name, synonyms, physico-chemical properties, chemical identifiers, database identifiers, 2D chemical structure, and a list of GPCR targets with experimental data. An example output involving the ligand, prenalterol, is shown in Figure 3.10, where all GPCRs that bind with the ligand are listed at the bottom of the page.

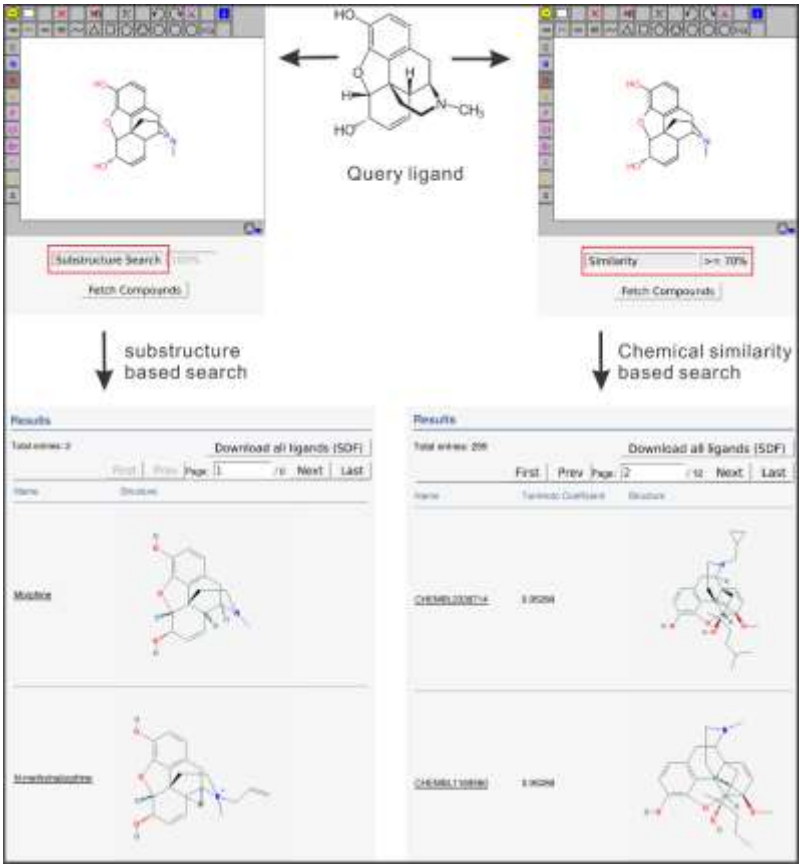


Figure 3.11 - Searching GLASS database for ligands using either the substructure similarity (Left Panel) or chemical similarity (Right Panel). The users first specify the ligand by importing a MOL or SDF file or draw the molecule into the JSME molecular editor. In this example, morphine is the query molecule. By clicking on the 'Fetch Compounds' button, the substructure search pipeline will look for ligands that have the molecule that the users specified as part of its chemical structure; the chemical similarity search pipeline will return all ligands that are at least 70% chemically identical to the query (the cutoff is adjustable). While the ligand name and chemical structure are provided for the users in both searches, Tanimoto coefficients are seen only with the chemical similarity search.

Although the GPCR-ligand association information can be retrieved from the GPCR- and ligand-based searches, GLASS provides a third GPCR-ligand-based search option if the respective GLASS ID of the interaction is known. In the above example, the GLASS ID of the human β2 adrenergic receptor and prenalterol association is '8792'. By searching on '8792', the users will be brought to a page containing GPCR and ligand information, as well as experimental binding affinity data. In this example, the free energy of binding was reported to be 9.76 kcal/mol from the reference with the PubMed ID 24063433.

In addition to the ligand-, GPCR- and ligand-GPCR-based searching options, GLASS provides a target-based search for users who wish to locate a particular ligand by either chemical similarity or match of substructure (Figure 3.11). Using the JSME chemical editor, the user can manually draw a ligand of interest or import a MOL or SDF file. Substructure search queries should be for the ligands of sufficient chemical complexity, as it would otherwise match too many ligands and result in an unreasonably long search. Searching by chemical similarity, there are options to select for a percentage cutoff. Results are returned with respective ligands and 2D chemical structure images; Tanimoto coefficients are also provided for similarity searches. All ligands found can be downloaded in SDF file format. An example to search homologies of morphine is illustrated in Figure 3.10.

3.3.2. Browsing GLASS

A comprehensive list of GPCRs and ligands from GLASS is provided on the home page to enable browsing of all entries in bulk. Additionally, the user can also browse all GPCRs as sorted by their respective families as designated by UniProt.[20] According to this schema, the rhodopsin-like family GPCR entries are further divided into the level of sub-families due to the high volume of entries, while the rest of the families remain in one level.

3.3.3. Downloading GLASS

Tables of GPCR, GPCR-ligand, and ligand data are all made available for download in TSV file format. A zipped SDF file of all GLASS ligands in 3D format is available and ready for use in molecular docking experiments; physicochemical properties and molecular descriptors are included within the property tags for the user's convenience.

4. Summary

We have developed a new database, GLASS, which encompasses a wide breadth of GPCR-related pharmacological data, gathered from a multitude of data sources. GLASS contains over ten times more ligand and GPCR-ligand interaction data than the leading databases, which makes GLASS the most comprehensive and up-to-date GPCR-ligand association repository in the field. It is however the novel sets of data collection and feature setting, rather than the sheer amount of data, which makes GLASS database unique.

The current structure of GLASS database has been made to retain the majority of GPCR-ligand pharmacological data after some definitive filters to rule out false positives; this gives users options to choose proper cutoff values for certain experimental parameters, such as binding constants. This will avoid any subjective pre-cutoffs that limit user's flexibility. Certain GPCR-ligand databases, such as GLIDA,[14] only give a list of ligands with biological activities as opposed to experimental parameters. For example, a ligand could be designated as an agonist for a GPCR, but we are left unaware of how it came to be as such. The pre-cutoff setting makes it difficult to customize ligand datasets by experimental values for analysis. GLASS database was designed to ensure all of its extracted data available for user manipulation. The presence of this option means that analyses can be performed on individual GPCRs to elucidate their ligand preferences based on various cutoff values.

One of the major difficulties in studying GPCR-ligand association stems from the lack of 3D structure of the receptors due to the notorious recalcitrance to crystallization.[26] Currently, GLASS has integrated X-ray crystal structures for numerous GPCRs from the PDB library and predicted models by GPCR-I-TASSER for all 1,073 human GPCRs from GPCR-EXP database (http://zhanglab.ccmb.med.umich.edu/GPCR-EXP/). In the next step, we are extending GPCR-I-TASSER to generate atomic structure models for the rest of 2,020 GPCRs from other species. Meanwhile, we will extend the cutting-edge ligand binding prediction approaches, including COFACTOR[27] and COACH[28], to deduce the ligand-binding sites of all GPCRs based on the GPCR-I-TASSER models. The high-resolution 3D structure and ligand-binding prediction data will provide useful insights to the physical landscape of the GPCR-ligand associations.

One of the focuses of GLASS is to provide references to various experimental and computational virtual screening studies. For instance, an important approach to GPCR virtual screening is to collect ligand profiles from homologous ligand-GPCR interactions,[13] where the completeness of the ligand-GPCR associations in GLASS will be essential to increase the sensitivity and recognition power of the ligand profiles. With its comprehensive coverage of datasets and consistent updates of data, we expect that GLASS become an important primary GPCR resource and impart its usefulness in many other biomedical studies, including *in silico* GPCR drug discovery, GPCR de-orphanization, and functional annotation.

5. References

1. Cherezov, V.; Rosenbaum, D. M.; Hanson, M. A.; Rasmussen, S. G.; Thian, F. S.; Kobilka, T. S.; Choi, H. J.; Kuhn, P.; Weis, W. I.; Kobilka, B. K.; Stevens, R. C., High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **2007,** *318* (5854), 1258-65.
2. Dorsam, R. T.; Gutkind, J. S., G-protein-coupled receptors and cancer. *Nat Rev Cancer* **2007,** *7* (2), 79-94.
3. Klabunde, T.; Hessler, G., Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *Chembiochem* **2002,** *3* (10), 928-944.
4. Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L., How many drug targets are there? *Nature reviews Drug discovery* **2006,** *5* (12), 993-996.
5. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* **2001,** *46* (1-3), 3-26.
6. Horn, F.; Weare, J.; Beukers, M. W.; Horsch, S.; Bairoch, A.; Chen, W.; Edvardsen, O.; Campagne, F.; Vriend, G., GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res* **1998,** *26* (1), 275-9.
7. Beukers, M. W.; Kristiansen, I.; AP, I. J.; Edvardsen, I., TinyGRAP database: a bioinformatics tool to mine G-protein-coupled receptor mutant data. *Trends in pharmacological sciences* **1999,** *20* (12), 475-7.
8. Khelashvili, G.; Dorff, K.; Shan, J.; Camacho-Artacho, M.; Skrabanek, L.; Vroling, B.; Bouvier, M.; Devi, L. A.; George, S. R.; Javitch, J. A.; Lohse, M. J.; Milligan, G.; Neubig, R. R.; Palczewski, K.; Parmentier, M.; Pin, J. P.; Vriend, G.; Campagne, F.; Filizola, M., GPCR-OKB: the G Protein Coupled Receptor Oligomer Knowledge Base. *Bioinformatics* **2010,** *26* (14), 1804-5.
9. Gatica, E. A.; Cavasotto, C. N., Ligand and decoy sets for docking to G protein-coupled receptors. *Journal of chemical information and modeling* **2012,** *52* (1), 1-6.
10. Zhang, J.; Zhang, Y., GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation. *Bioinformatics* **2010,** *26* (23), 3004-5.
11. van Laarhoven, T.; Nabuurs, S. B.; Marchiori, E., Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* **2011,** *27* (21), 3036-43.

12. Weill, N.; Rognan, D., Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *Journal of chemical information and modeling* **2009,** *49* (4), 1049-62.
13. Zhou, H.; Skolnick, J., FINDSITE(X): A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. *Mol Pharm* **2012,** *9* (6), 1775-84.
14. Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C., GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update. *Nucleic acids research* **2008,** *36* (Database issue), D907-12.
15. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012,** *40* (Database issue), D1100-7.
16. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic acids research* **2007,** *35* (Database issue), D198-201.
17. Sharman, J. L.; Mpamhanga, C. P.; Spedding, M.; Germain, P.; Staels, B.; Dacquet, C.; Laudet, V.; Harmar, A. J.; Nc, I., IUPHAR-DB: new receptors and tools for easy searching and visualization of pharmacological data. *Nucleic acids research* **2011,** *39* (Database issue), D534-8.
18. Knox, C.; Law, V.; Jewison, T.; Liu, P.; Ly, S.; Frolkis, A.; Pon, A.; Banco, K.; Mak, C.; Neveu, V.; Djoumbou, Y.; Eisner, R.; Guo, A. C.; Wishart, D. S., DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic acids research* **2011,** *39* (Database issue), D1035-41.
19. Roth, B. L., http://pdsp.med.unc.edu/pdsp.php. **2015**.
20. Magrane, M.; Consortium, U., UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011,** *2011*, bar009.
21. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *Journal of cheminformatics* **2011,** *3*, 33.
22. Qin, C.; Zhang, C.; Zhu, F.; Xu, F.; Chen, S. Y.; Zhang, P.; Li, Y. H.; Yang, S. Y.; Wei, Y. Q.; Tao, L.; Chen, Y. Z., Therapeutic target database update 2014: a resource for targeted therapeutics. *Nucleic Acids Res* **2014,** *42* (Database issue), D1118-23.
23. Zhang, J.; Yang, J.; Jang, R.; Zhang, Y., Hybrid structure modeling of G protein-coupled receptors in the human genome. **2015**, submitted.
24. Bienfait, B.; Ertl, P., JSME: a free molecule editor in JavaScript. *Journal of cheminformatics* **2013,** *5*, 24.
25. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P., Comparability of Mixed IC50 Data - A Statistical Analysis. *Plos One* **2013,** *8* (4), e61007.
26. Michino, M.; Abola, E.; Brooks, C. L.; Dixon, J. S.; Moult, J.; Stevens, R. C.; Participants, G. D., Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. *Nature Reviews Drug Discovery* **2009,** *8* (6), 455-463.
27. Roy, A.; Yang, J.; Zhang, Y., COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic Acids Res* **2012,** *40* (Web Server issue), W471-7.
28. Yang, J.; Roy, A.; Zhang, Y., Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013,** *29* (20), 2588-95.

CHAPTER 4.

MAGELLAN: Incorporation of Sequence and Structure Information in a Ligand-Profile Based Virtual Screen for Human Class-A G Protein-Coupled Receptors

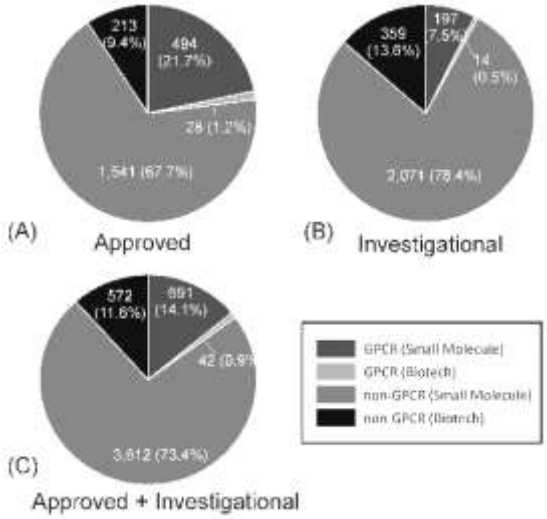


Figure 4.1 - DrugBank Statistics for GPCRs. Percentage of small molecule and biotech drugs shown for GPCR and non-GPCR targets under the groups of (A) approved, (B) investigational, and (C) total drugs.

1. Introduction

G protein-coupled receptors (GPCR) are a large superfamily of transmembrane receptors responsible for cellular signal transductions. The malfunction of the receptors is the cause of a wide array of pathologies, such as cancer and diabetes.[1-2] Consequently, GPCRs are among the most clinically-studied targets in drug discovery. A detailed analysis of the DrugBank shows that 23% of the 2,276 FDA-approved drugs on the market, including both small molecules and biologics, target GPCRs, while out of other 2,641 drugs that are under some form of clinical trial, 8% target GPCRs. Overall, GPCRs represent targets of 15% of all drugs that are either approved and investigational (Figure 1).

A mainstay in drug development is high-throughput screening (HTS), a technique biochemically assaying a pharmacological target against the candidate compounds. However, HTS is usually costly and laborious, where various *in silico* approaches have found useful to assist and complement HTS.[3] There are two general approaches that are commonly employed in the computer-aided drug screening: ligand-based and receptor-based virtual screening. In the former, knowledge of what ligands the receptor targets tend to bind is used to develop a model for drug screening,[4-5] while the latter utilizes structural information of the receptors to predict ligand-binding affinity, normally through docking.[6-7] Although structure-based approaches are typically very computationally expensive and can take a long time to run, they can be very useful when the structure of the receptor is known, producing results that may be biochemically relevant; on the other hand, ligand-based approaches are usually very fast, but they tend to be biased towards ligands that are currently known.[8]

However, both structure and ligand-based approaches require some sort of information, either known active ligands or a structure, and this may not be available for a drug target of interest. The orphan GPCRs are one such example, many of which lack known endogenous ligands.[9] In this regard, chemical genomics approaches are often applied to infer ligand binding information, based on the assumption that similar receptors bind similar ligands.[10] One of the earliest applications of the idea was with the algorithm, FINDSITE, which uses ligand information from structurally-homologous receptors found through fold-recognition in a ligand-based virtual screen.[11-12] Another more-recent algorithm is PoLi, developed by the same lab, which looks for similar protein receptors by performing binding pocket structure comparison between the query and targets, followed by a ligand-based screening search.[5]

Though structure is generally considered to be more conserved than sequence in evolution, relying solely on structural similarities can result in high false positives in ligand information deduction, as receptors of similar structures often bind with different ligands. In particular, experimental structures are not always available for many medically-relevant target proteins, where low-resolution models would have to be generated for the target receptors; this would further impact the accuracy and specificity of the structure-based ligand inferences. This is true especially for the case of GPCR families, which all have similar global folds but different local structure at the

binding sites.[13] Moreover, the majority of the structure-based approaches rely on selecting homologous proteins and their respective ligand sets from the Protein Data Bank (PDB);[14] however, pharmacological data are often found in low quantities within the PDB. Currently, there are many more proteins with known pharmacological data than those with known structures. The largest sources of publicly-available ligand data reside in various manually curated databases, such as ChEMBL,[15] BindingDB,[16] and GLASS (for GPCRs).[17] Using the wealth of information from such resources should help enhance the accuracy of the ligand-based approaches.

In this study, we present a novel ligand-profile based virtual screening approach, MAGELLAN (standing for Michigan G protein-coupled receptor ligand-based virtual screen), specifically designed for G protein-coupled receptors. To enhance the reliability and robustness of ligand-based screening approach, multiple methods, utilizing both structure- and sequence-based alignments, are employed for detecting heterogeneous receptor homologies, from which consensus ligand profiles are created for the next step of virtual screening. To examine the strength and weakness of the pipeline, large-scale tests are performed on 224 representative Class A GPCRs, which are carefully controlled with various component and state-of-the-art methods. Here, Class A GPCRs were selected as the focus mainly because of their high diversities in structure and function and clinical importance in drug discovery. Moreover, the conserved transmembrane domains of these receptors make it an ideal case for examining the sequence- and structure-based alignment pipelines. An on-line MAGELLAN webserver, together with the virtual screening results for all human GPCRs and the filtered ligand sets, are available and downloadable at *https://zhanglab.ccmb.med.umich.edu/MAGELLAN*.

2. Methods

The virtual screening process of MAGELLAN consists of three distinct stages: 1) GPCR alignment and selection, 2) ligand profile construction, and 3) virtual screening. The flowchart of the MAGELLAN pipeline is depicted in Figure 2, which starts with a single primary sequence of the target (or query) GPCR in FASTA format, where the output consists of a list of predicted ligands bound with the target.

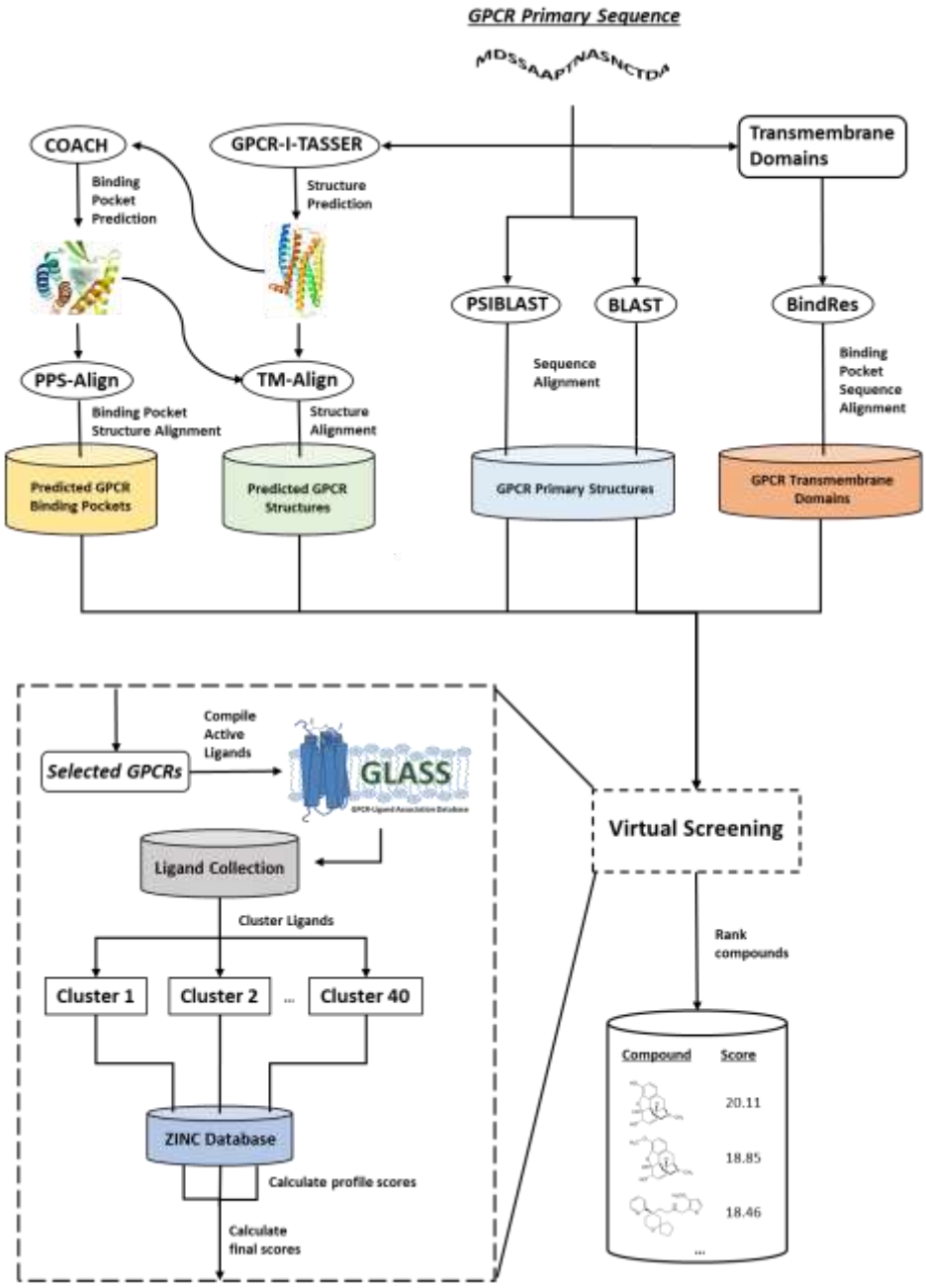


Figure 4.2 - MAGELLAN pipeline

2.1 Construction of Ligand-GPCR Association Library

The central assumption of MAGELLAN is that similar receptors bind similar ligands.[10] For this purpose, a comprehensive library of GPCR-ligand associations is constructed from the GLASS database [17]. Here, only experimental values of K_i, K_d, IC_{50}, and EC_{50} were used. In case that multiple experimental values exist for the same GPCR-ligand pair from different studies, the

median was taken as the representative value to avoid outliers. To filter out inactive ligands, a common threshold of 10 μM is used for both K_i and K_d values. However, a threshold of 20 μM was set for IC_{50} and EC_{50}, justified by a previous study that found a K_i-IC_{50} conversion factor of 2 to be suitable.[18] This relatively loose criterion could account for variability in assay conditions, inherent in the types of experiments used to determine these values. After filtration, the library contains 238,108 GPCR-ligand associations attached with 644 GPCRs.

2.2 Detection of Homologous GPCRs

The first stage of MAGELLAN is to select homologous GPCRs, in order to construct a predictive model for what the target GPCR would potentially bind. Five complementary algorithms, including TM-align,[19] PPS-Align,[20] BLAST,[21] PSI-BLAST[22] and BindRes, are extended to detect analogous GPCRs, where the first two are structure based and the other three are built on sequence and sequence-profile comparisons.

In the first GPCR detection pipeline, TM-align[19] is used to align the global structures of query to template GPCRs. To obtain a structure model of the query GPCR, its query sequence is submitted to GPCR-I-TASSER, which was designed to create full-length GPCR structures by reassembling the structural fragments from threading through replica-exchange Monte Carlo simulations.[23] The resulting structure models are then compared against the GPCRs in the pre-compiled GPCR-ligand library, where the structures are also generated with GPCR-I-TASSER. The resultant GPCRs detected by TM-align are scored by:

$$S_{TMalign} = \frac{2}{1 + e^{-(0.2\mathrm{T}+\mathrm{f}(0.4\mathrm{S} + 0.3\mathrm{E} + 0.2J) + R)^2}} - 1 \qquad (1)$$

Here, $T = \frac{1}{L}\sum_i^{L_{ali}} \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}$ is the TM-score to measure the global structure similarity of the query and template models, where L is the length of the query sequence, L_{ali} is the number of aligned residues by TM-align, d_i is the distance of ith pair of aligned residues between query and template and $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ is the scale factor; $f = m/n$ is the fraction of the aligned residues in the binding pocket (m) normalized by the total number of binding residues (n) on the template; $S = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}$ accounts for the local structural similarity of the binding pockets between query

and template; $E = \frac{1}{n}\sum_{i=1}^{n} B(A_i^q, A_i^t)$ measures the evolutionary relation between the aligned binding residues, where $B(A_i^q, A_i^t)$ is the BLOSUM mutation score; $J = \frac{1}{n}\sum_{i=1}^{n}\left(\sum_a^{20} p_i^a \log\frac{p_i^a}{p_i^a+q^a} + \sum_a^{20} q^a \log\frac{p_i^a}{p_i^a+q^a}\right)$ is the average Jensen-Shannon divergence over the binding pocket, where p_i^a is the frequency of amino acid a at ith column of multiple sequence alignment (MSA) identified by PSI-Blast for the query GPCR and q^a is the background frequency; and R is the residue chemical similarity of the binding site residues, where Figure 4.3 provides an illustrative example for how the residue chemical similarity was calculated.

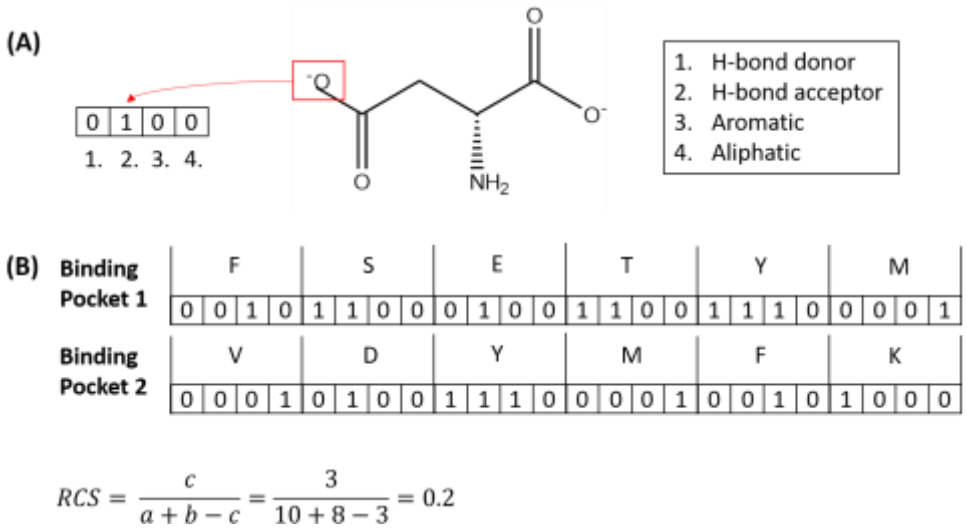


Figure 4.3 - Explanation of the Residue Chemical Similarity (RCS) Term. (A) Shown is the binary representation of the chemical features of aspartate. Each bit is position dependent and corresponds to one of four chemical features (H-bond donor, H-bond acceptor, aromatic, and aliphatic). The anionic oxygen from the carboxyl group (red box) is a H-bond acceptor, thus a bit is set in the second position. (B) In this simplified example, the two binding pockets are a binary representation of the chemical features of the aligned binding pockets produced from PPS-Align. The RCS calculation is essentially a Tanimoto coefficient calculation between the two bit-string representations of the binding pockets, where a is the number of bits in Binding Pocket 1, b is the number of bits in Binding Pocket 2, and c is the number of bits shared between the two. Here, the two binding pockets have a RCS of 0.2, as shown in the figure.

Second, PPS-Align is an algorithm recently designed for sequence-order independent structure alignments of binding pockets.[20] In this pipeline, the GPCR-I-TASSER models of query and template GPCRs are submitted to COACH,[24] which was designed to detect ligand binding residues through composite sequence-profile and structure comparisons. The ligand binding pockets are then constructed by clustering the COACH binding-site predictions with the highest confidence score, which are finally aligned by PPS-Align for pocket comparisons. The GPCR templates from the library are scored by

$$S_{PPSalign} = \frac{2}{1 + e^{-(PPS + 0.25S + 0.25J + I_{bs})}} - 1 \qquad (2)$$

where PPS in [0,1] is the pocket similarity score returned by PPS-Align, S and J are the same as defined in Eq. (1), and I_{bs} is the sequence identity of the binding-site residues in the PPS-align aligned region between query and template GPCRs.

In the third BindRes pipeline, we first parse the transmembrane (TM) domains of the query GPCR according to the UniProtKB/SwissProt annotation, which are then aligned with the TM domains of all template GPCRs in the library using Clustal Omega.[25] The template GPCRs are ranked by

$$S_{BindRes} = \frac{2}{1 + e^{-(I_{bs} + R + 0.2J)}} - 1 \qquad (3)$$

where I_{bs}, R and J are defined similarly as in Eqs. (1-2). The calculations focus solely on the 44 orthosteric binding site residues on the TM-domains, as specified by Gloriam et al.[26] Since these orthosteric residues have been labelled in Ballesteros-Weinstein numbering system,[27] the identities can be conveniently referred through the most conserved residue of each TM domain according to the Clustal Omega alignments.

Finally, the BLAST and PSI-BLAST pipelines use the programs from the NCBI BLAST+ software suite (V2.2.29). For BLAST, the query GPCR sequence is matched against the GPCR templates, which are sorted by descending sequence identity to the query. The same is done for PSI-BLAST but with sequence-profile alignment, where the profiles were collected with 4 iterations from the non-redundant (NR) sequence database from NCBI under an E-value cutoff of 0.001. The results are also ranked by descending sequence identity.

2.3 Ligand Profile Construction and Profile-Based Virtual Screening

Associated active ligands from the ten top-ranked GPCRs are compiled for each of the five GPCR alignment methods. It should be noted that all ligands are originally represented as InChI identifiers and keys, and converted into 1,024-bit Morgan fingerprints with a radius of 2 using RDKit.[28]

To capture the most common chemotypes, the resulting ligand collections are clustered with the Taylor-Butina algorithm[29-30] using the Chemfp Python library,[31] where a Tanimoto coefficient (TC)

cutoff of 0.8 was used. The 40 largest GPCR clusters are selected for use in the next step of virtual screening. If there are fewer than 40 clusters, all of them are used.

For a given cluster (*k*), a ligand profile is constructed for the query GPCR, which is represented by a $1024 \times N_k$ matrix, where N_k is the number of non-redundant ligands in the cluster and each ligand has 1024-bit fingerprints taken from the ZINC12 database (Figure 3). A profile-compound calculation is performed through the compound library using a profile score of

$$PrS_k(z) = \frac{1}{\mathrm{N_k}} \sum_{i=1}^{N_k} w_i T_{i,z} \qquad (4)$$

where $T_{i,z} = \frac{\sum_{j=1}^{1024} b_i^j b_z^j}{\sum_{j=1}^{1024} b_i^j b_i^j + \sum_{j=1}^{1024} b_z^j b_z^j - \sum_{j=1}^{1024} b_i^j b_z^j}$ is the Tanimoto coefficient between the *i*th ligand and the *z*th compound in the database. Here, b_i^j and b_z^j are the bits in the *i*th and the *z*th compounds, respectively. $w_i = \frac{1}{M_i} \sum_{m=1}^{M_i} S_m(i)$ is the weighting factor for ligand *i*, where $S_m(i)$ is the scoring function of *m*th alignment method as defined in Eqs. (1-3) and M_i is the total number of methods that identifies the *i*th ligand (Figure 4.4).

For each cluster (k), each $PrS_k(\mathrm{z})$ is converted into a Z-score, where $Z_k(z) = \frac{PrS_k(\mathrm{z}) - \mu_k}{\sigma_k}$, where μ_k and σ_k are, respectively, the mean and standard deviation of all PrS in the kth cluster. A final score for ZINC compound, *z*, is calculated by taking the maximum PrS among all clusters,

$$S(z) = \max_k \{Z_k(z)\} \qquad (5)$$

Here, we note that there are overall 3 free parameters in the MAGELLAN pipeline, including number of GPCRs used from the alignments, TC cutoff for clustering, and number of clusters used. These parameters have been optimized using an independent dataset of 56 GPCRs that are non-redundant from the test proteins reported in this study. During the training process, the parameters were determined by maximizing the average enrichment factor of virtual screening as defined in Eq. (7) below.

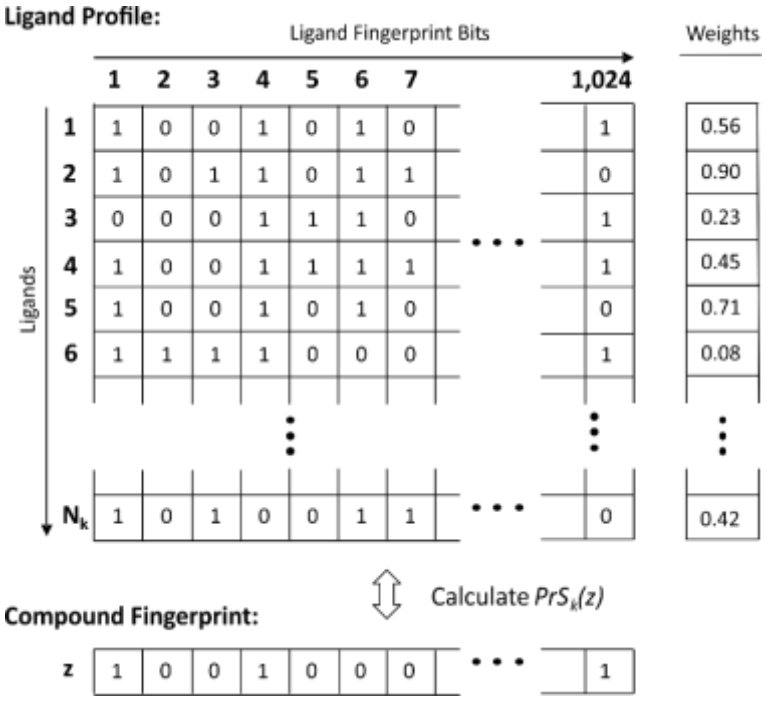


Figure 4.4 - Illustration of the ligand profile that summarizes feature information of all ligands from the *k*th cluster. In the ligand profile, the horizontal axes represent the ligand fingerprint bits, while the vertical axes indicate the ligands; each ligand contains a corresponding weight (w_i) based on GPCR alignment scores. The bottom shows a compound fingerprint from ZINC database which is scored. For each ligand in the ligand profile, a Tanimoto coefficient ($T_{i,z}$) is calculated against the ZINC compound and multiplied with its corresponding weight. The average of these values among the ligand profile is the $PrS_k(\mathrm{z})$.

2.4 Construction of Minimum Spanning Tree by Similarity Ensemble Approach

To evaluate similarity of GPCR proteins based on their ligand similarity, we construct a minimum spanning tree for the targets using the similarity ensemble approach (SEA).[32-33] To assess the ligand set similarities in a statistically stringent base, we first collect multiple random ligand sets with sizes between 10 and 1,000 ligands from the GLASS database, and calculate the TC score between the randomly collected ligand pairs. The relation of the TC-score distribution and the size of ligand sets follows well with

$$\begin{cases} \mu = ks \\ \sigma = ms^r \end{cases} \qquad (6)$$

where μ and σ are mean and standard deviation of the TC-score distribution, s is the product of the size of two ligand sets compared, and k, m and r are parameters to fit (see Figure 4.5). Thus,

the fitting parameters are used to convert any raw TC-score of two ligand sets to a size-independent Z-score by $Z = (TC - \mu)/\sigma$.

Here, only the ligand pairs with TC score above a threshold are used in the statistical calculation, where a TC threshold of 0.84 is found to be optimal, which has the Z-score distribution follow the Gumbel distribution (Figure 4.6). Using the extreme value distribution data, a BLAST-like E-value can be calculated for each GPCR pairs. Finally, a minimum spanning tree based on the significance of E-value can be calculated using Kruskal's algorithm,[34] with the image generated with Cytoscape.[35]

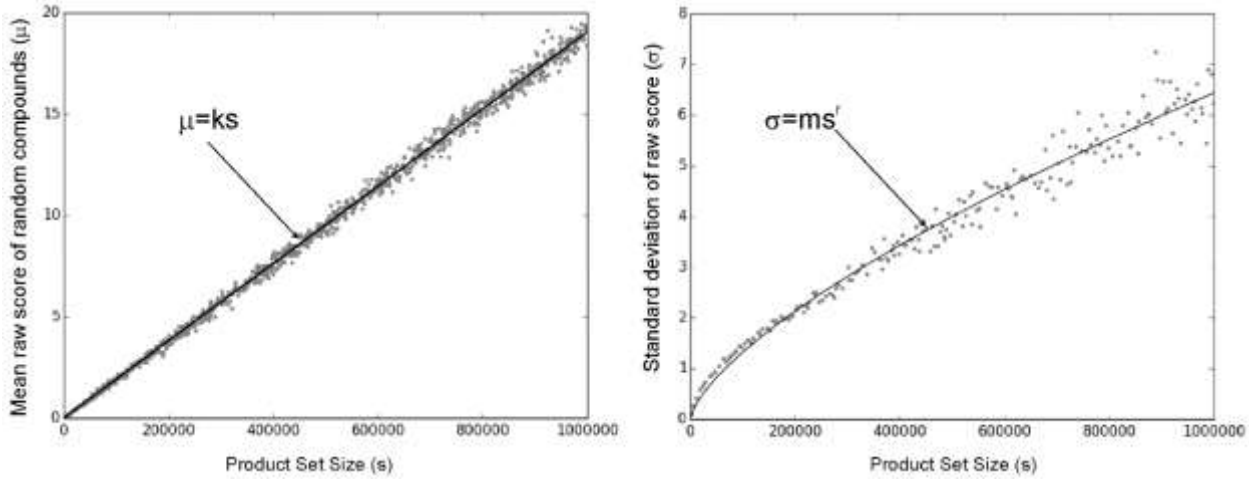


Figure 4.5 - Mean and standard deviation of ligand similarities between random ligand sets from the GLASS database versus the product of set sizes. Only the ligand pairs with a Tanimoto coefficient above 0.84 are calculated and the data follow well the linear and power-law equation shown in Eq. (6).

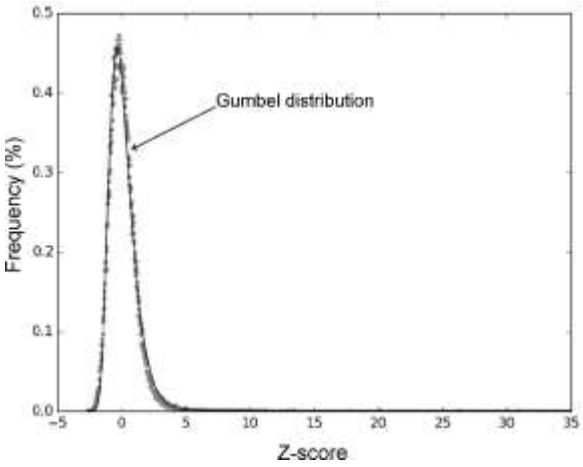


Figure 4.6 - Z-score distribution of the random background data from GLASS database. The Tanimoto coefficient of 0.84 was found to be the best fit with the Gumbel distribution.

2.5 On-line Webserver Construction

An online web server of MAGELLAN is constructed at https://zhanglab.ccmb.med.umich.edu/MAGELLAN/, using Python CGI scripting, complemented with MySQL, Javascript, and PHP. The MAGELLAN algorithm implemented on the web server is the same as that described in the present study, with the exception that it uses TM-HMM to determine the transmembrane domains of the unknown query GPCR.[36] The user is able to supply optional inputs, such as annotated transmembrane domains, a GPCR structure, or the binding site residues, in order to bypass GPCR-I-TASSER and COACH that can reduce the job runtime. All ligand and test sets filtered from GLASS database are provided for download for the user. Additionally, the top 1% of results from screening the full ZINC database for all human GPCRs are pre-generated and made available publicly.

3. Results

3.1 Comparison of MAGELLAN with Component Methods

Five different alignment methods have been used in MAGELLAN for GPCR model collection. To justify the profile approach, we first examine the performance of MAGELLAN in comparison with the individual alignment methods, in which the same procedure is implemented as shown in Figure 4.2, but only with the GPCR models detected by an individual method.

3.1.1 Testing Dataset Construction and Performance Evaluation

The test datasets are constructed from a comprehensive list of all 224 Class A GPCRs. For each GPCR, the active ligands are collected from GLASS database,[17] which are filtered with a stringent activity threshold of 1 μM for K_i, K_d, IC_{50}, and EC_{50} values. In order to increase chemical diversity and to even out set sizes, ligands were clustered for each GPCR by their Bemis-Murcko frameworks.[37] If there were between 100 and 600 frameworks, the highest activity ligand was selected from each cluster. If there were fewer than 100 frameworks, the highest activity ligands were chosen regardless of their framework until a total of 100 ligands was achieved. If there were greater than 600 frameworks, the activity threshold was decreased by a factor of 2 until there were fewer than 600 frameworks, wherein the highest activity ligands were selected from each framework. As a result, the test set consists of 224 Class A GPCRs, which are associated with in total 54,438 active ligands, or on average 258 per GPCRs.

To test the methods, the active ligands from each GPCR are mixed with a set of 500,000 randomly-selected compounds as decoys from the "Clean Drug-Like" subset of the ZINC database. The downloaded compounds were in SMILES string format and subsequently converted into Morgan fingerprints with RDKit,[28] consisting of 1,024-bit fingerprints with a radius of 2. A retrospective virtual screen (RVS) experiment is implemented by different methods, where the goal of RVS is to prioritize the active ligands using the proposed scoring functions. The performance of RVS can be qualitatively measured by the enrichment factor (*EF*):

$$EF_{x\%} = \frac{N_{act}^{x\%} / N_{select}^{x\%}}{N_{act} / N_{tot}} \qquad (7)$$

where N_{act} and N_{tot} are the total numbers of the active and all compounds in the ligand pool, respectively. $N_{act}^{x\%}$ and $N_{select}^{x\%}$ are, respectively, the numbers of true positive ligands and the number of all candidates in the top $x\%$ of the compounds selected by the RVS methods. A higher EF_x indicates a better RVS performance, where $EF_x = 1$ means a random selection without enrichment. While $x\%$ can be taken as different cutoff (1%, 2%, 5% etc), we focus mainly on 1% for the briefness of data presentation.

To rule out the effect from using close homologous targets, a handicap was applied to the selection of GPCRs, where any homologous GPCR templates with greater than 30% sequence identity to the query, based on the BLAST alignment, were excluded. Without this handicap, only the query GPCR is excluded. Meanwhile, to challenge the pipeline, a homologous cutoff has been applied in the GPCR-I-TASSER structure modeling and COACH binding prediction, i.e., all structures with a sequence identity >30% to the query GPCR sequence are excluded from the threading template library no matter if the handicap of binding GPCR is applied.

3.1.2 MAGELLAN Significantly Outperforms Component Pipelines in RVS Experiment

In Figure 4.7A, we present a scatter plot of the enrichment factors ($EF_{1\%}$) acquired from MAGELLAN, in comparison with that from the five individual pipelines, where a cutoff of 30% sequence identity was used for filtering out the close homologous GPCR templates when inferring the ligand profiles. It was shown that MAGELLAN achieves a higher enrichment factor than the individual pipelines for most of the GPCRs. For example, MAGELLAN outperforms the BindRes

pipeline in 130 cases, while BindRes does so in 72 cases. These numbers are 146/52, 115/77, 121/67, and 129/70 for BLAST, PSI-BLAST, PPS-align, and TM-align pipelines respectively.

Table 4.1 - Summary of RVS results by MAGELLAN and component methods. Data shown are median and average (in parentheses) $EF_{1\%}$ values on 224 test Class A GPCRs. P-value is calculated in the Wilcoxon signed-rank test between MAGELLAN and the control methods.

Methods	With handicap		Without handicap	
	$EF_{1\%}$	p-value	$EF_{1\%}$	p-value
MAGELLAN	14.38 (23.03)	--	62.03 (59.13)	--
BindRes	10.19 (17.90)	5×10^{-7}	56.39 (56.35)	4×10^{-6}
BLAST	7.04 (16.49)	5×10^{-14}	53.02 (53.26)	1×10^{-14}
PSI-BLAST	11.83 (20.79)	5×10^{-3}	54.31 (54.18)	2×10^{-14}
TM-align	13.76 (20.15)	8×10^{-9}	56.92 (56.39)	6×10^{-4}
PPS-Align	10.99 (20.25)	2×10^{-6}	56.90 (55.55)	1×10^{-6}

In Table 4.1 (Columns 2 and 3), we also list the average and median $EF_{1\%}$ values for each method, which again shows that MAGELLAN achieved a higher enrichment factor than all the individual methods. To examine the significance of the difference, a Wilcoxon signed-rank test is calculated for each pair of the comparison, where the two-tailed p-value is equal to or below $5 \, x \, 10^{-3}$ in all the cases, which indicates that the differences between MAGELLAN and the individual methods are statistically significant.

When comparing the individual methods, TM-Align performed better than the other component methods, as evidenced by its higher median $EF_{1\%}$ of 13.76. As structure is more conserved than sequence, it is of no surprise that it was able to achieve such a result. However, no single method by itself contributed dominantly to MAGELLAN because the p-values between MAGELLAN and each of them were significant, signifying the synergistic effect of data fusion.

In Figure 4.7B, we also present the results without using the 30% sequence identity cutoff, but the target GPCRs have been excluded from the ligand profile detection process. As expected, the RVS

performance becomes much better when homologous GPCRs are included in the ligand profile construction, where many of the points in the figure have been shifted to the upper two quadrants in the plots, as compared to Figure 4.7A. Interestingly, the synergistic effect as witnessed with the handicap is not as pronounced as without the handicap. As shown in Table 4.1 (Columns 4 and 5), the average and median $EF_{1\%}$ values of MAGELLAN are in general higher than that of the individual methods. Nonetheless, MAGELLAN overall achieved better performance than the individual component methods.

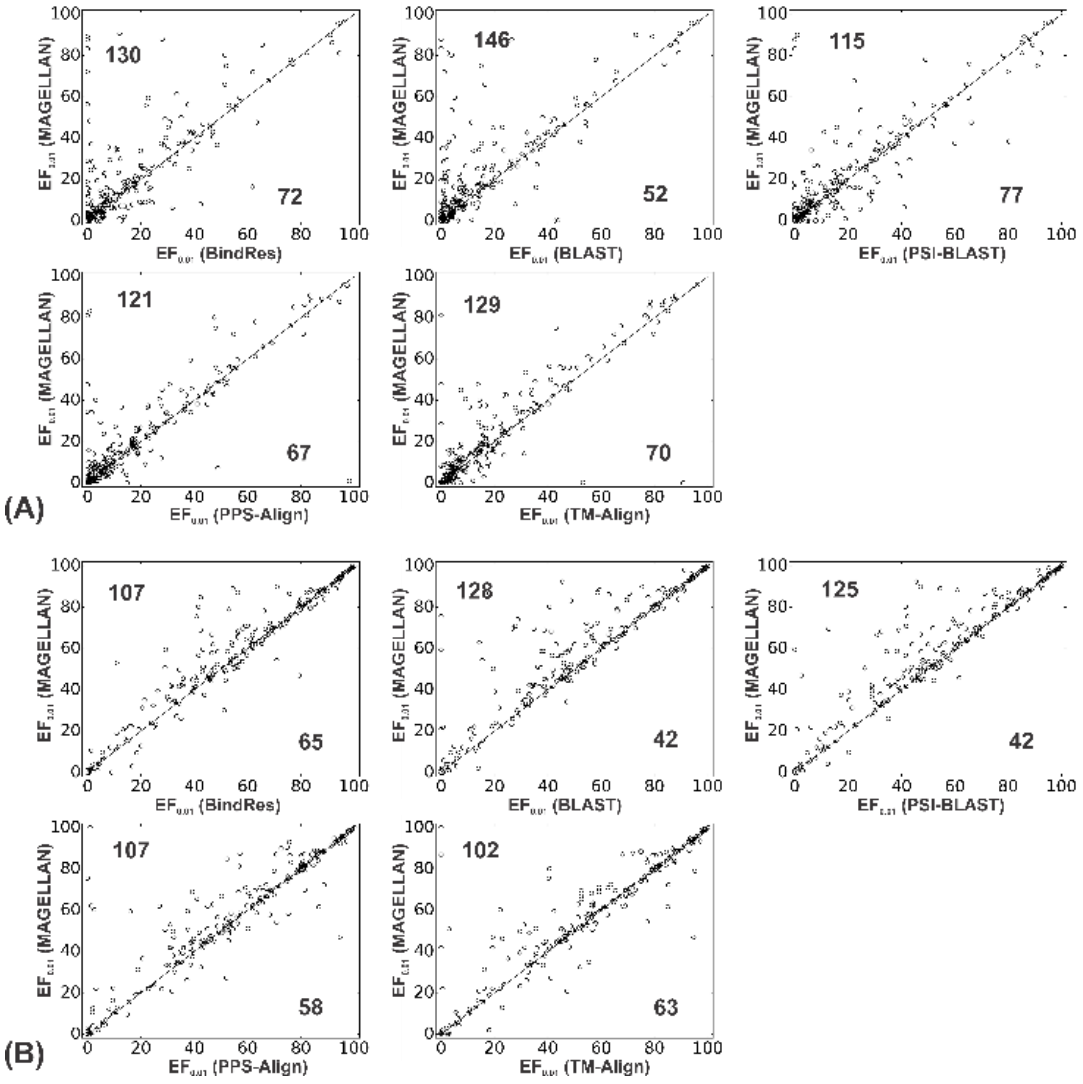


Figure 4.7 - Comparison of MAGELLAN and five component methods in the retrospective virtual screen experiment on 224 Class A GPCRs. Results are shown as $EF_{1\%}$ either with (A) 30% sequence identity cutoff or (B) no cutoff. Each point represents one GPCR. The numbers in each triangle represent the number of GPCRs for which the RVS method outperforms the comparison ones.

3.1.3 Both Sequence and Structural Alignments Are Essential to MAGELLAN Performance

To further examine the impact of the individual GPCR alignment methods on MAGELLAN's performance, we counted the highest-performing alignment method type for each GPCR under the sequence identity handicap, where the type was denoted as either sequence (BLAST, PSI-BLAST, BindRes) or structure based (TM-align, PPS-Align). Among the 224 Class A GPCRs, 89 had a structure-based method as their top-scoring method with the highest enrichment factors, while the sequence-based method does so in 135 cases. Out of the sequence-based methods, 60 of the cases were from PSI-BLAST. Overall, the number of GPCRs was relatively evenly distributed for each method, and no method stood out in particular, lending credence to the observation that both sequence and structural alignments aid in MAGELLAN's performance with the sequence identity handicap.

We also examined the effect of running MAGELLAN without the sequence identity handicap. Of the 224 Class A GPCRs examined, 101 resulted in a sequence-based method as their top-scoring method, while 123 were from structure-based methods. This was somewhat surprising, as the TM-Align method produced 94 of the best cases. However, the scoring function from that component was likely able to capture homologous GPCR's better than with sequence alone. Additionally, the contributions from BindRes were not menial, having 47 top-scoring GPCRs. Certainly, each method played a role in lending their predictive power to MAGELLAN.

3.2 Benchmark of MAGELLAN with Other Virtual Screening Approaches

To examine MAGELLAN with other state of the art approaches, we tested the performance in control with three widely-used virtual screening programs, including AutoDock Vina,[6] DOCK 6,[7] and PoLi.[5] The former two are receptor-docking based approaches, where the crystal structures of the target GPCRs were used as the input for molecular docking. For AutoDock Vina, all compounds were converted from Mol2 to PDBQT format. Additionally, the experimental GPCR PDB files were converted into the PDBQT format, whereby hydrogens and partial charges were added to all PDBQT files. A 30 x 30 x 30 $Å^3$ search space was defined on the receptor so that it centered upon the crystal ligand. Default settings were used for the virtual screen, with compounds ranked according to their docking scores. For DOCK 6, all GPCR PDB files were converted to the

Mol2 format, where hydrogens and partial charges were added. Spheres were selected within 5 Å of the crystal ligand, while the scoring grid enclosed the spheres with a 5 Å margin. Flexible docking was performed with the recommended settings, with compounds ranked according to their grid score. Finally, PoLi is a ligand-based virtual screening tool with the probe ligands detected by the binding-pocket structural comparisons between query and templates. As the software is not available for installation, the data was taken from the benchmark study of the original authors.[5] The benchmark tests of the methods were performed on two separate datasets from DUD-E[38] and GPCR-Bench.[39]

3.2.1 Tests on DUD-E Dataset

DUD-E[38] is a widely-used dataset specially designed for virtual screening benchmarks. It contains five Class A GPCR proteins, where each protein has on average 224 active ligands from ChEMBL. Each active ligand is paired with 50 molecular decoys (with similar chemistry but of different topology) drawn from ZINC. While the turkey beta-1 adrenergic receptor (P07700) was included in DUD-E, there is no pharmacological data in any of the ligand databases. Thus, the ligand clusters from the human orthologue (P08588) were used in its place. To examine the performance, we run MAGELLAN and the three control programs in an automated mode against the ligand dataset for each GPCR target, with the goal to pick up the active ligands using their scoring functions.

Table 4.2 - RVS results of $EF_{1\%}$ on five Class A GPCRs in DUD-E Dataset**.** Values out and in parentheses are the results for MAGELLAN with or without handicap cutoffs. Data for PoLi was taken from Roy et al.[5]

Gene	UniProt ID	MAGELLAN	PoLi	AutoDock	Dock 6
AA2AR	P29274	0.95 (39.03)	1.2	1.42	2.86
ADRB1	P07700	5.47 (36.11)	2.0	0.66	2.63
ADRB2	P07550	13.68 (34.75)	2.6	2.69	1.35
CXCR4	P61073	0 (23.97)	0	0	2.49
DRD3	P35462	28.65 (39.27)	5.2	2.64	1.26
Average		9.75 (34.63)	2.2	1.48	2.12

In Table 4.2, we list the $EF_{1\%}$ value of virtual screening for the five GPCRs, calculated by MAGELLAN, AutoDock Vina, and DOCK 6, respectively. Additionally, ROC curves are presented in Figure S4, and AUC values are given in Table S1. Here, the PoLi data are directly taken from Roy et al,[5] in which a similar sequence identity cutoff 30% was applied for the homologous GPCR filtering. The data show a better performance of MAGELLAN than the three control algorithms for four out of the five tested GPCRs, under the sequence identity cutoff of 30%. In particular, the dopamine receptor D3 (DRD3) performed exceptionally well with MAGELLAN, with an $EF_{1\%}$ 19.06, which is more than 3 times higher than that of the control methods (i.e., 5.2, 2.64, and 1.26 by Poli, AutoDock and Dock 6, respectively). With a number of related GPCRs selected (P25115 / P21728) from the alignments, the contribution of their ligands in the clusters was well established, accounting for 9 chemotypes (Figure 4.8A).

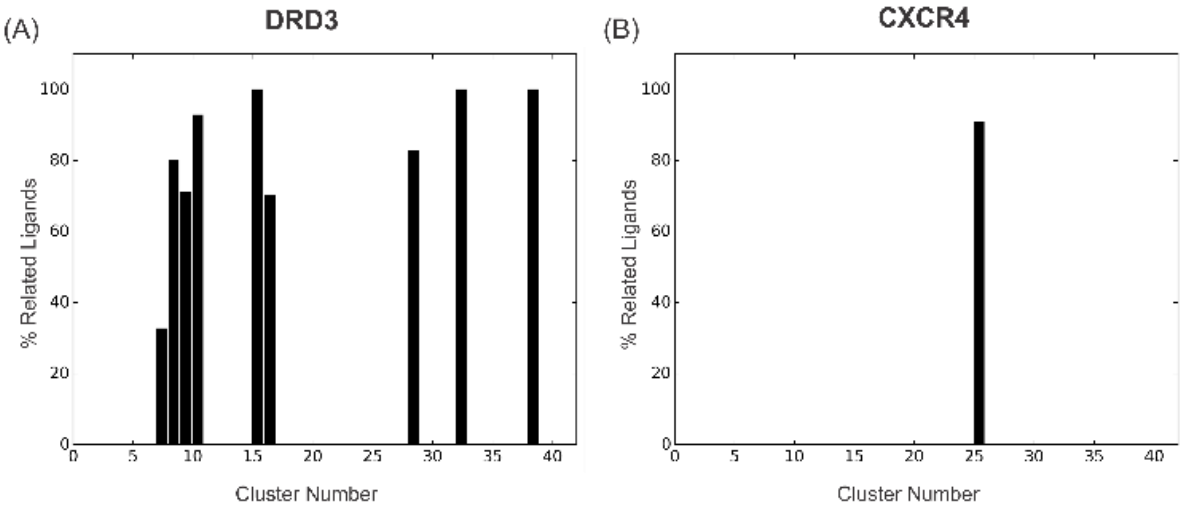


Figure 4.8 - Proportion of Ligands from Related GPCRs in Clusters from MAGELLAN with Handicap for (A) Dopamine Receptor D3 (DRD3) and (B) C-X-C chemokine receptor type 4 (CXCR4). The presence of chemotypes containing ligands from respective related GPCRs was examined, where one cluster represents one chemotype. The influence of related GPCRs resulted in 9 chemotypes for DRD3, while CXCR4 only had one.

The $EF_{1\%}$ of MAGELLAN is 4.59 and 5.83 for the beta-1 adrenergic receptor and beta-2 adrenergic receptor, respectively, which is moderately higher than the control methods. The main reason for the enhanced performance is due to complementary GPCR pipelines exploited in MAGELLAN which detected several closely-related GPCRs that bound with similar ligands (including P25100 / P18130 / O02824 for ADRB1, and Q01338 / P23944 for ADRB2), despite

the low sequence identity. The ligand profiles constructed from the closely-related ligands helped prioritize the active compound hits.

However, MAGELLAN yielded a low $EF_{1\%}$ of 0.95 and 0 for the adenosine A2A receptor (AA2AR) and the C-X-C chemokine receptor type 4 (CXCR4), respectively, where no closely-related GPCRs were selected for the former receptor. Consequently, both receptors could be rescued when run without the cutoff, in which related subtypes were correctly detected by both the structural and sequence-based GPCR alignment methods. The active ligands thus have significantly higher scores than that of the inactive ones for these two cases, which resulted in an $EF_{1\%}$ of 39.15. The C-X-C chemokine receptor type 4 had a few relatives selected (P51682 / P51684 / O54814), but despite their presence, their corresponding ligands were only present in one out of the top 40 clusters used in MAGELLAN (Figure 4.8B), suggesting the need for chemotype diversity of the set of clusters. Moreover, the ligand set sizes for the related GPCRs were very small (56, 65, and 34, respectively), lessening their influence overall.

Altogether, these results highlight the importance of the inclusion of homologous templates for ligand profile constructions. This phenomenon was observed for all the receptors, in which $EF_{1\%}$ was significantly increased by the inclusion of close homologous GPCRs in ligand profile construction. In particular, the number of chemotypes and size of ligand sets from related GPCRs appeared to play a role in performance. Overall, the average $EF_{1\%}$ is 9.75, which is 4.4 times higher than PoLi, and 6.6 and 4.6 times higher than AutoDock and Dock6 respectively. The $EF_{1\%}$ value will increase by 3.6 times if homologous GPCRs are included in the profile construction process.

In Figure 4.9, we present the receiver operating characteristic curves (ROC) for the retrospective virtual screen results by MAGELLAN and the control methods for all targets, where the corresponding area under the curve (AUC) values are listed in Table 4.3. It should be noted that we were unable to acquire a ROC curve for PoLi, as we did not have access to their data. Overall, MAGELLAN achieved slightly higher AUC values with the handicap cutoffs (AUC=0.69) as compared to PoLi (AUC=0.58), AutoDock Vina (AUC=0.64), and DOCK 6 (AUC=0.61). Without the handicap cutoff, MAGELLAN was able to attain much better performance (AUC=0.92). This

suggests that MAGELLAN is able to better correctly select compounds that would potentially bind the GPCR of interest as compared to the benchmarked methods.

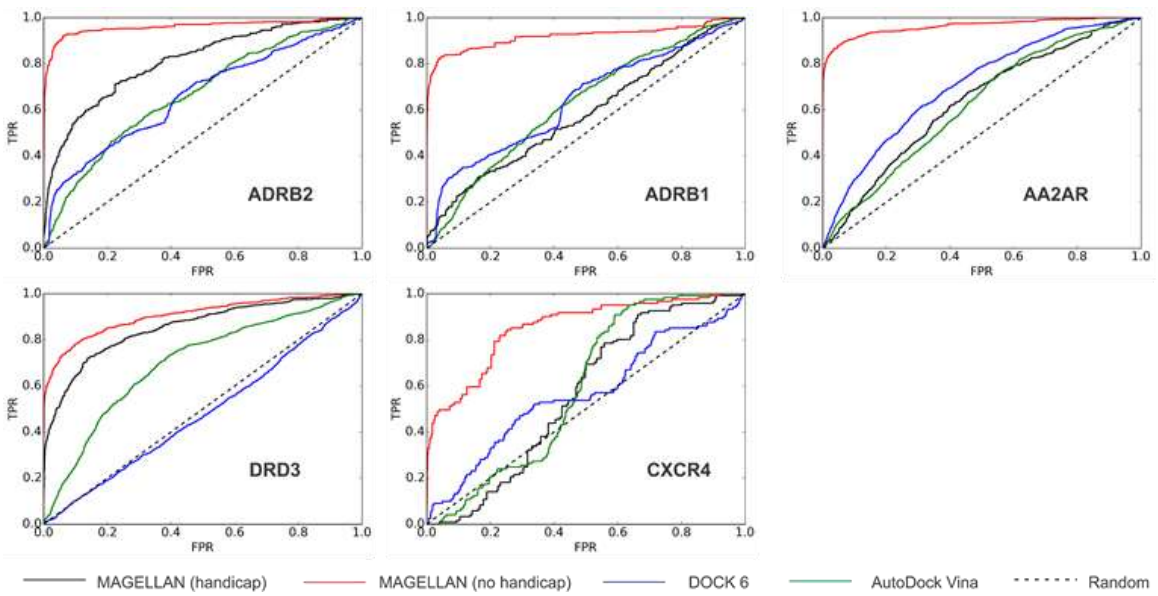


Figure 4.9 - ROC Curves for Retrospective Virtual Screen Results of MAGELLAN, AutoDock Vina, and DOCK 6 with DUD-E Dataset. The following GPCRs were tested: beta-2 adrenergic receptor (ADRB2), beta-1 adrenergic receptor (ADRB1), adenosine receptor A2A (AA2AR), dopamine D3 receptor (DRD3), and C-X-C chemokine receptor type 4 (CXCR4).

Table 4.3 - Comparison of receiver operating characteristic (ROC) values by MAGELLAN, PoLi, AutoDock Vina, and Dock 6 on 5 Class A GPCRs in DUD-E. Values out and in parentheses are the results for MAGELLAN with or without handicap cutoffs.

Gene Name	UniProt ID	MAGELLAN	PoLi	AutoDock Vina	Dock 6
AA2AR	P29274	0.63 (0.96)	0.53	0.62	0.71
ADRB1	P07700	0.58 (0.92)	0.62	0.63	0.64
ADRB2	P07550	0.81 (0.96)	0.56	0.67	0.66
CXCR4	P61073	0.56 (0.85)	0.49	0.59	0.57
DRD3	P35462	0.85 (0.91)	0.70	0.70	0.48
Average		**0.69 (0.92)**	**0.58**	**0.64**	**0.61**

3.2.2 Tests on GPCR-Bench Dataset

The second benchmark dataset on which we conducted experiments is GPCR-Bench;[39] it contains 20 Class A GPCRs, where each GPCR has between 100 to 600 active ligands accompanied by 50 decoys per active ligand. The RVS results on this benchmark are summarized in Table 4.4.

Table 4.4 - Summary of EF_(1%) results on 20 Class A GPCRs in GPCR-Bench. Values out and in parentheses are the results for MAGELLAN with or without handicap cutoffs.

Gene	UniProt ID	MAGELLAN	AutoDock	Dock6
GPR40	O14842	22.04 (48.92)	24.28	21.84
OX2R	O43614	7.92 (34.65)	1.82	0
ADRB2	P07550	24.64 (60.87)	0.20	9.40
ADRB1	P07700	1.03 (54.36)	0	2.73
ACM2	P08172	23.00 (36.00)	7.76	7.12
ACM3	P08483	24.88 (48.26)	2.31	8.97
S1PR1	P21453	0.50 (51.24)	0.47	0.16
PAR1	P25116	0.00 (0.00)	13.39	2.00
5HT1B	P28222	16.83 (60.89)	1.54	2.79
AA2AR	P29274	0.48 (34.13)	0	0.64
OPRD	P32300	27.93 (65.77)	3.77	0.75
HRH1	P35367	21.39 (50.75)	3.28	0.22
DRD3	P35462	46.27 (60.70)	1.24	1.03
OPRK	P41145	12.94 (47.76)	0.65	0.22
OPRX	P41146	2.99 (24.38)	0.25	2.85
5HT2B	P41595	18.41 (20.40)	0.98	0.98
OPRM	P42866	2.44 (58.54)	0	1.08
CCR5	P51681	4.06 (23.86)	0.53	4.36
CXCR4	P61073	4.26 (70.21)	0	0.26
P2Y12	Q9H244	11.94 (13.43)	0.76	2.03
Average		13.70 (43.26)	3.16	3.47

In total, MAGELLAN performed favorably (average $EF_{1\%}$ = 13.70) in this benchmark, as compared with AutoDock Vina (average $EF_{1\%}$ =3.16) and DOCK 6 (average $EF_{1\%}$ =3.47). AutoDock Vina and DOCK 6 achieved the best enrichment for the free fatty acid receptor 1 (GPR40) with an $EF_{1\%}$=24.28 and 21.84 respectively. Since all of its active ligands belong to the same chemotype,[39] the binding pocket of this target does not have as much variation compared with other more challenging targets, and thus makes it easier for docking (Figure 4.10).

MAGELLAN attained a comparable enrichment on this target with $EF_{1\%}$ =22.04; if the homologous templates are included, however, the performance is significantly improved to $EF_{1\%}$ =48.92. Additionally, AutoDock Vina achieved decent enrichment with the protease-activated receptor 1 (PAR1) at $EF_{1\%}$=13.39, while MAGELLAN resulted in $EF_{1\%}$=0.00 with a handicap. In fact, MAGELLAN detected several protease-activated receptor subtypes (Q63645, P55085, Q96RI0), but their respective ligand sets were of a very small size (2, 59, 10, respectively). As a result, their related ligands were not present in the top 40 clusters because of the minority of binding ligands, which resulted in the reduced performance. While most of the successful examples of GPCRs are found to have at least one related subtype that had a sizeable number of ligands, the data suggests that the number of ligands in the ligand sets of closely-related members is essential to the success of MAGELLAN, in addition to its ability to detect homologous GPCRs.

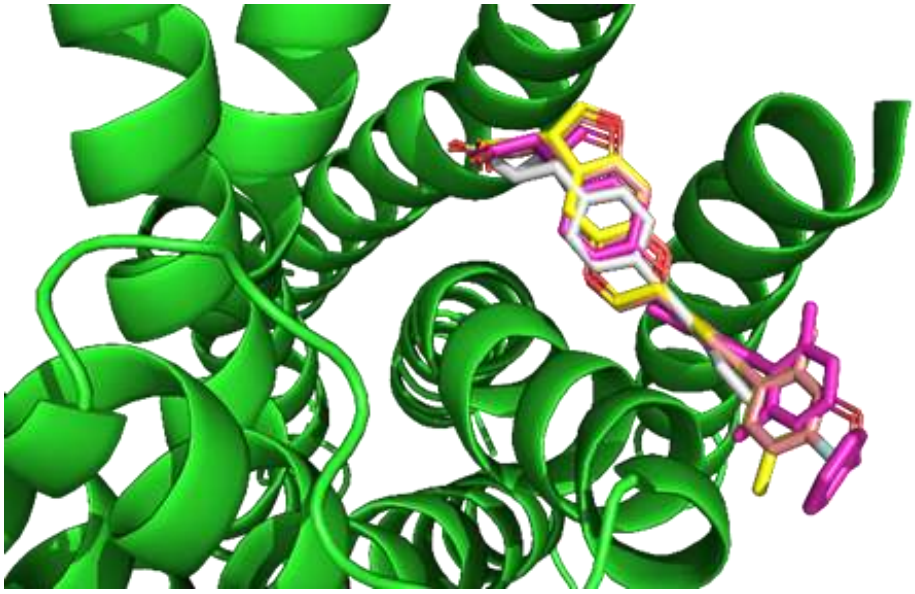


Figure 4.10 - Top 5 Active Compound Results for Free Fatty Acid Receptor 1 Using AutoDock Vina. Note that the binding pocket resides between two of its transmembrane domains, extending from the lipid bilayer to the inner cavity of the receptor. As is evident from the structure of the docked ligands, they share an obvious chemotype that also extends to the rest of the active set.

In Figure 4.11, we also present the log receiver operating characteristic curves (ROC) for the retrospective virtual screen results by MAGELLAN and the two control methods for all targets, where the corresponding Boltzmann-enhanced receiver operating characteristic (BEDROC, α=20) values are listed in Table 4.5. Here, as opposed to the conventional receiver operating characteristic (ROC), BEDROC has the advantage to better assess early enrichment,[40] which is important as

compounds selected for experimental validation are always chosen from top-ranked candidates. Overall, MAGELLAN exhibited a higher average early enrichment both with (BEDROC=0.32) and without (BEDROC = 0.68) the handicap cutoffs, as compared to AutoDock Vina (BEDROC=0.16) and DOCK 6 (BEDROC=0.14). This suggests again that MAGELLAN has a higher propensity to correctly select compounds that would potentially bind the GPCR of interest.

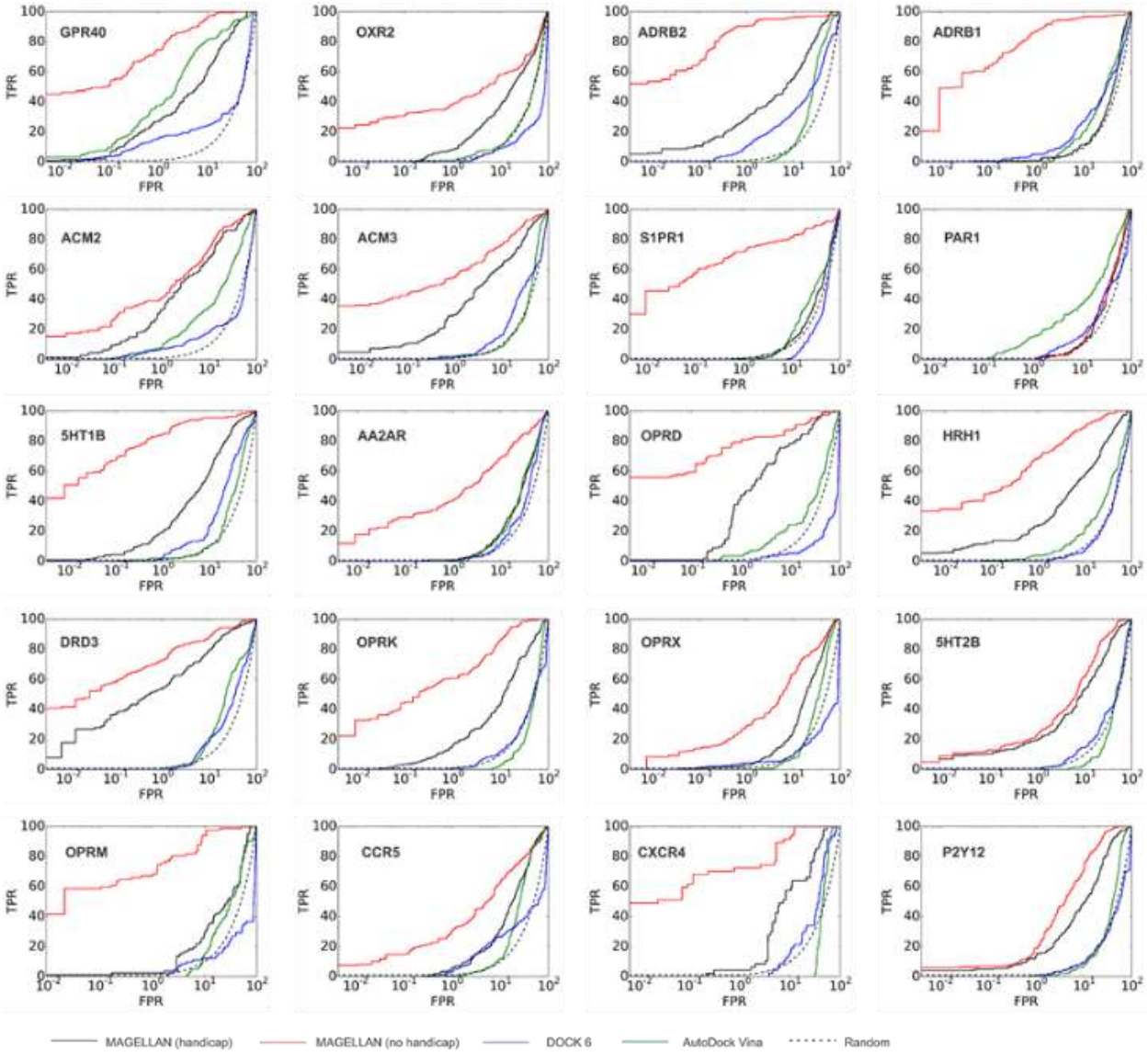


Figure 4.11 - Log ROC Curves for Retrospective Virtual Screen Results of MAGELLAN, AutoDock Vina, and DOCK 6 with GPCR-Bench Dataset. The following GPCRs were tested: free fatty acid receptor 1 (GPR40), orexin receptor 2 (OXR2), beta-2 adrenergic receptor (ADRB2), be ta-1 adrenergic receptor (ADRB1), muscarinic acetylcholine receptor 2 (ACM2), muscarinic acetylcholine receptor 3 (ACM3), sphingosine 1-phosphate receptor (S1PR1), proteinase-activated receptor 1 (PAR1), 5-hydroxytryptamine receptor 1B (5HT1B), adenosine receptor A2A (AA2AR), delta opioid receptor (OPRD), histamine H1 receptor (HRH1), dopamine D3 receptor (DRD3), kappa opioid receptor (OPRK), nociception receptor (OPRX), 5-hydroxytryptamine receptor 2B (5HT2B), mu opioid receptor (OPRM), C-C chemokine receptor type 5 (CCR5), C-X-C chemokine receptor type 4 (CXCR4), and purinergic receptor (P2Y12).

Table 4.5 - Comparison of Boltzmann-enhanced receiver operating characteristic (BEDROC) values by MAGELLAN, AutoDock Vina and Dock 6 on 20 Class A GPCRs in GPCR-Bench. Values out and in parentheses are the results for MAGELLAN with or without handicap cutoffs.

Gene Name	**UniProt ID**	**MAGELLAN**	**AutoDock Vina**	**Dock 6**
GPR40	O14842	0.46 (0.87)	0.59	0.35
OX2R	O43614	0.26 (0.51)	0.12	0.03
ADRB2	P07550	0.45 (0.94)	0.17	0.23
ADRB1	P07700	0.07 (0.92)	0.07	0.23
ACM2	P08172	0.54 (0.62)	0.11	0.21
ACM3	P08483	0.49 (0.69)	0.07	0.13
S1PR1	P21453	0.06 (0.80)	0.22	0.13
PAR1	P25116	0.06 (0.06)	0.34	0.26
5HT1B	P28222	0.42 (0.92)	0.07	0.02
AA2AR	P29274	0.10 (0.58)	0.29	0.13
OPRD	P32300	0.60 (0.84)	0.08	0.19
HRH1	P35367	0.43 (0.80)	0.11	0.18
DRD3	P35462	0.69 (0.83)	0.09	0.07
OPRK	P41145	0.33 (0.75)	0.13	0.08
OPRX	P41146	0.15 (0.46)	0.18	0.04
5HT2B	P41595	0.38 (0.44)	0.16	0.06
OPRM	P42866	0.16 (0.82)	0.12	0.12
CCR5	P51681	0.20 (0.49)	0.04	0.05
CXCR4	P61073	0.28 (0.81)	0.06	0.09
P2Y12	Q9H244	0.33 (0.49)	0.09	0.14
Average		**0.32 (0.68)**	**0.16**	**0.14**

3.3 Case Studies: Using MAGELLAN Prediction for Common Drug Targets and De-orphanization

3.3.1 Mu Opioid Receptor

One of the main purposes of MAGELLAN is to predict compounds that could potentially be developed into clinically useful drugs. As an illustration towards this goal, we examined here a medically-important target, the mu opioid receptor (UniProt ID: P35372), which is closely involved in the reduction of pain. Common drugs in pain reduction include morphine and heroin, which are both strong opioid agonists. However, a major side effect of their consumption often results in various unwanted side effects, such as nausea, constipation, respiration depression, and addiction; as a result, many current research efforts have been trying to develop drugs with the analgesic effect while trying to reduce or eliminate the aforementioned maladies.[41]

From the retrospective virtual screen experiment performed without sequence identity cutoff, a high $EF_{1\%}$ of 87.81 was achieved for the human mu opioid receptor. This is not surprising, as many opioids are notorious for their promiscuous binding to the opioid family of receptors, which also comprise additionally of the kappa opioid, delta opioid, and nociceptin receptors. All GPCR alignment methods exploited were able to select at least one of these receptors. Using a similarity ensemble approach analysis of the GPCR test sets as described in Methods, we have constructed a minimum spanning tree based on the similarity of the test sets identified by MAGELLAN, where all the opioid receptors were found pharmacologically related (Figure 4.12, blue nodes).

From these results, it is clear that the application of MAGELLAN to known drug targets, such as the mu opioid receptor, can yield accurate virtual screening results. More importantly, a fraction of top-ranking results could be used sequentially as a targeted library in a structure-based virtual screen campaign, such as docking, in order to address the shortcomings of ligand-based virtual screening methods.[8] As docking can be computationally expensive, MAGELLAN can thus serve as a relatively-quick filtration step of the database of interest.

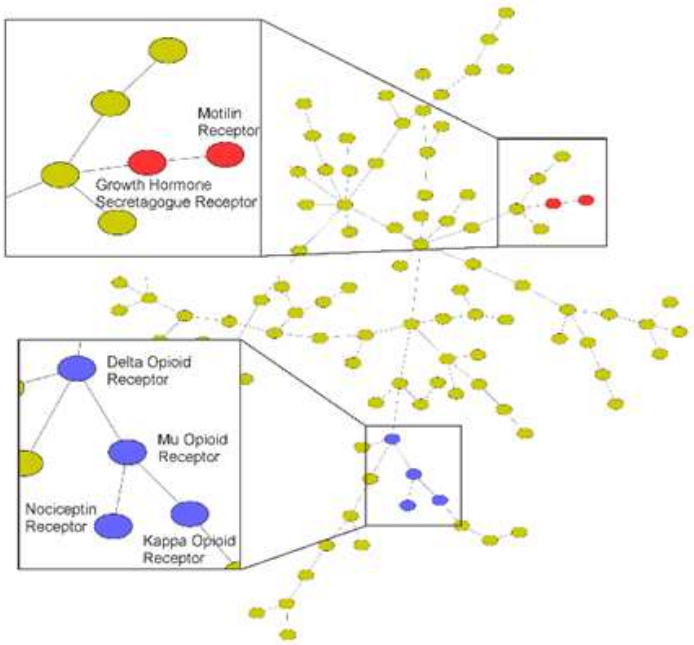


Figure 4.12 - Ligand set similarity map constructed from MAGELLAN predictions on the GPCR test sets. Each node corresponds to a GPCR ligand set. The map was created using Kruskal's algorithm based on E-values of the ligand sets calculated by the similarity ensemble approach (SEA).

3.3.2 Motilin Receptor

Apart from drug discovery, MAGELLAN can be extended to GPCR de-orphanization. As orphan GPCRs have either no identified endogenous ligand or unknown function, they likely have an undiscovered physiological role likely to be intertwined with disease and remain potential therapeutic targets in drug discovery. For the purposes of this study, we have chosen to examine a former orphan GPCR, the human motilin receptor (UniProt ID: O43193), to illustrate the de-orphanization process of MAGELLAN.

It is known that the closest-related GPCR to the motilin receptor is the growth hormone secretagogue receptor type 1 (also known as ghrelin receptor), with a sequence identity of 51% between the human variants. With such a high sequence identity, it was predicted that MAGELLAN would be able to enrich for reference compounds. With no sequence identity cutoff, an $EF_{1\%}$ of 33.33 was achieved. Every alignment method was able to select multiple orthologues of growth hormone secretagogue receptor type 1. An example BindRes alignment between the human motilin receptor and pig pig growth hormone secretagogue receptor type 1 is given in Figure 4.13, where it is apparent that there is high sequence identity (~70%) between the aligned

binding site residues. Once again using SEA analysis, it was shown that the ligand sets of this receptor and the motilin receptor were pharmacologically similar (Figure 4.12, red nodes) and likely to have contributed to the high enrichment. It should be noted that there were no ligand sets corresponding to any orthologues of the motilin receptor; there was ligand set data for only the human variant. Taken together, these data suggest the feasibility for the application of MAGELLAN in efforts to deorphanize GPCRs.

BW:	1.35	1.39	1.42	1.46	2.53	2.57	2.58	2.61	2.62
Query:	V	C	L	G	I	L	P	L	Y
	:	:	:	:	:		:	:	
Target:	V	C	L	G	I	M	P	L	F
BW:	2.64	2.65	3.28	3.29	3.32	3.33	3.36	3.37	3.39
Query:	L	W	S	L	G	E	T	Y	T
	:	:				:	:	:	:
Target:	L	W	F	Q	S	E	T	Y	T
BW:	3.40	4.56	4.57	4.60	4.61	5.38	5.39	5.42	5.43
Query:	L	S	A	F	L	V	M	V	T
		:	:			:	:	:	
Target:	V	S	A	I	F	V	M	V	S
BW:	5.46	5.47	6.44	6.47	6.48	6.51	6.52	6.55	6.58
Query:	Y	F	F	C	W	F	H	R	Y
		:	:	:	:	:	:	:	
Target:	F	F	F	C	W	F	H	R	F
BW:	6.59	7.35	7.39	7.42	7.43	7.45	7.46		
Query:	I	N	L	F	Y	S	A		
		:		:	:	:	:		
Target:	S	N	F	F	Y	S	A		

Figure 4.13 - BindRes Alignment for Human Motilin Receptor. BW refers to the Ballesteros-Weinstein numbering scheme for the binding site residues, while the query and target are the human motilin receptor (UniProt: O43193) and pig growth hormone secretagogue receptor type 1 (UniProt: Q95254), respectively. The colons depict residues that are identical.

4. Discussion

Built on the assumption that similar receptors bind with similar ligands, we have developed a new hierarchical, ligand-profile based approach, MAGELLAN, to virtual drug screen targeting Class A GPCRs. Starting from amino acid sequence of the target proteins, MAGELLAN first utilizes GPCR-I-TASSER[23] to generate tertiary structure prediction for the target protein. Next, five pipelines of structure, sequence and orthosteric binding-site based alignment methods are extended to the detection of homologous and analogous proteins, where all known ligands bound with the proteins are clustered for the construction of a set of chemical profiles; these are finally used to

match through the compound libraries for screening putative ligands and drugs for the target receptor.

The method was first tested on a comprehensive set of 224 Class A GPCRs and achieved a median enrichment factor $EF_{1\%}$ of 15.31 after excluding all homologous templates in both structure prediction and GPCR template detection processes, which is significantly higher than the pipelines using individual GPCR alignment methods. In addition, MAGELLAN was tested on two independent benchmark sets from DUD-E[38] and GPCR-Bench,[39] consisting of 5 and 20 Class A GPCRs, and compare favorably with other state-of-the-art docking and ligand-based virtual screening approaches, including AutoDock Vina,[6] DOCK 6,[7] and PoLi.[5] Detailed data analysis shows that the major advantage of MAGELLAN lies at the utilization of both structure (including global and local) and orthosteric binding-site based comparisons for GPCR template detections, whereas the ligand profiles constructed from the multiple resources of data fusion help enhance the sensitivity and specificity of the virtual screening through the compound databases.

Apart from the favorable benchmark performance, several advances may help future MAGELLAN developments. First, MAGELLAN is a ligand-profile based approach utilizing only ligand-GPCR associations. This is different from other ligand-oriented approaches, such as PoLi which relies on known ligand-protein complex structures from the BioLip.[42] Currently, the number of non-redundant ligand-GPCR associations from GLASS with experimental data is 533,470, which is over 7,500 times higher than the number of known, non-redundant ligand-GPCR complexes in BioLiP; this is part of the reason for the significant improvement of MAGELLAN over PoLi. The gap between the ligand-receptor association and the protein binding structure databases are rapidly increasing[15-17, 43], which should give additional advantage and potential to the future development of the ligand-based methods such as MAGELLAN.

Compared to the docking-based approaches, MAGELLAN has the advantage in utilizing low-resolution predicted structures, since high-resolution experimental structures are often unavailable to many important drug targets. Technically, it is also a benefit to exploit the global fold comparison for GPCR template detection because many experimentally--solved structures are in unbound apo form, which can significantly impact the accuracy of the docking-based approaches

that often have difficulty in modeling the ligand-induced conformational changes. In addition, docking a large library of compounds is very computationally expensive and time consuming. As experienced in this study, it typically took days to weeks for AutoDock Vina or DOCK 6 to complete a docking screen for a single GPCR, depending on the target; conversely, ligand-based virtual screening using MAGELLAN only takes about an hour. Nevertheless, docking based programs have the advantage to generate 3D model of binding structures that is often useful for additional function and drug-based analyses. Meanwhile, we also found that there are several cases (such as GPR40 and PAR1) for which the docking-based approach achieve a much higher enrichment. Thus, a combination of MAGELLAN with structure-based docking should further improve the functionality and accuracy, which is currently under development.

It is important to note that one feature of MAGELLAN is its potential in discovering new ligands for orphan GPCRs. Since it is not limited by the PDB and utilizes related GPCRs to infer potential ligands, compounds ranked from a screening database by MAGELLAN could be selected and experimentally validated with biological assays. Currently, there are 87 Class A, 8 Class C, and 26 adhesion orphan GPCRs, according to a recent overview,[44] meaning that there is much more to discover. As illustrated in Figure 7, the motilin receptor, a former orphan Class A GPCR, achieved a moderately-high enrichment in a retrospective virtual screen. While this and other examples can only be validated with experimental, pharmacological studies in prospective screens, the data demonstrated an additional aspect of applications of MAGELLAN to the deorphanization of GPCRs.

5. Conclusion

Overall, we believe that MAGELLAN has surpassed the status quo, given the benchmark results against other state-of-the-art virtual screening algorithms. Based on the aforementioned case studies, our algorithm is readily applicable to the numerous orphan GPCR's whose functions and endogenous ligands remain to be elucidated. Moreover, it can also be used as a ligand-based virtual screening method and would likely prove useful in prospective virtual screening studies where information about the receptor is scant. Lastly, a user-friendly web server is provided for researchers interested in virtually screening a GPCR of interest against ZINC database.

6. References

1. O'Hayre, M.; Vazquez-Prado, J.; Kufareva, I.; Stawiski, E. W.; Handel, T. M.; Seshagiri, S.; Gutkind, J. S., The emerging mutational landscape of G proteins and G-protein-coupled receptors in cancer. *Nat Rev Cancer* **2013,** *13* (6), 412-24.
2. Rompler, H.; Staubert, C.; Thor, D.; Schulz, A.; Hofreiter, M.; Schoneberg, T., G protein-coupled time travel: evolutionary aspects of GPCR research. *Molecular interventions* **2007,** *7* (1), 17-25.
3. Tanrikulu, Y.; Kruger, B.; Proschak, E., The holistic integration of virtual screening in drug discovery. *Drug Discov Today* **2013,** *18* (7-8), 358-64.
4. Geppert, H.; Vogt, M.; Bajorath, J., Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J Chem Inf Model* **2010,** *50* (2), 205-16.
5. Roy, A.; Srinivasan, B.; Skolnick, J., PoLi: A Virtual Screening Pipeline Based on Template Pocket and Ligand Similarity. *J Chem Inf Model* **2015,** *55* (8), 1757-70.
6. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010,** *31* (2), 455-61.
7. Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C., DOCK 6: Impact of new features and current docking performance. *J Comput Chem* **2015,** *36* (15), 1132-56.
8. Drwal, M. N.; Griffith, R., Combination of ligand- and structure-based methods in virtual screening. *Drug Discov Today Technol* **2013,** *10* (3), e395-401.
9. Civelli, O.; Saito, Y.; Wang, Z. W.; Nothacker, H. P.; Reinscheid, R. K., Orphan GPCRs and their ligands. *Pharmacology & Therapeutics* **2006,** *110* (3), 525-532.
10. Klabunde, T., Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br J Pharmacol* **2007,** *152* (1), 5-7.
11. Brylinski, M.; Skolnick, J., A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A* **2008,** *105* (1), 129-34.
12. Zhou, H. Y.; Skolnick, J., FINDSITEx: A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. *Molecular Pharmaceutics* **2012,** *9* (6), 1775-1784.
13. Shoichet, B. K.; Kobilka, B. K., Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol Sci* **2012,** *33* (5), 268-72.
14. Rose, P. W.; Prlic, A.; Altunkaya, A.; Bi, C.; Bradley, A. R.; Christie, C. H.; Costanzo, L. D.; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y. P.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K., The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic acids research* **2017,** *45* (D1), D271-D281.
15. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012,** *40* (Database issue), D1100-7.

16. Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K., BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic acids research* **2007,** *35* (Database issue), D198-201.
17. Chan, W. K.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Ozgur, A.; Zhang, Y., GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* **2015,** *31* (18), 3035-42.
18. Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P., Comparability of Mixed IC50 Data - A Statistical Analysis. *Plos One* **2013,** *8* (4), e61007.
19. Zhang, Y.; Skolnick, J., TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic acids research* **2005,** *33* (7), 2302-9.
20. Hu, J.; Li, Y.; Yu, D. J.; Zhang, Y., PSS-align: Sequence-order independent comparison of protein binding pocket structures. **2018**, submitted.
21. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J., Basic local alignment search tool. *J Mol Biol* **1990,** *215* (3), 403-10.
22. Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* **1997,** *25* (17), 3389-402.
23. Zhang, J.; Yang, J.; Jang, R.; Zhang, Y., GPCR-I-TASSER: A Hybrid Approach to G Protein-Coupled Receptor Structure Modeling and the Application to the Human Genome. *Structure* **2015,** *23* (8), 1538-49.
24. Yang, J.; Roy, A.; Zhang, Y., Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics* **2013,** *29* (20), 2588-95.
25. Sievers, F.; Higgins, D. G., Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. *Multiple Sequence Alignment Methods* **2014,** *1079*, 105-116.
26. Gloriam, D. E.; Foord, S. M.; Blaney, F. E.; Garland, S. L., Definition of the G protein-coupled receptor transmembrane bundle binding pocket and calculation of receptor similarities for drug design. *J Med Chem* **2009,** *52* (14), 4429-42.
27. Ballesteros, J. A.; Weinstein, H., [19] Integrated methods for the construction of three-dimensional models and computational probing of structure-function relations in G protein-coupled receptors. *Methods in neurosciences* **1995,** *25*, 366-428.
28. Landrum, G., RDKit: Open-source cheminformatics. *Online). http://www. rdkit. org. Accessed* **2006,** *3* (04), 2012.
29. Taylor, R., Simulation Analysis of Experimental-Design Strategies for Screening Random Compounds as Potential New Drugs and Agrochemicals. *Journal of Chemical Information and Computer Sciences* **1995,** *35* (1), 59-67.
30. Butina, D., Unsupervised data base clustering based on Daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences* **1999,** *39* (4), 747-750.
31. Dalke, A., The FPS fingerprint format and chemfp toolkit. *Journal of Cheminformatics* **2013,** *5* (1), P36.
32. Keiser, M. J.; Roth, B. L.; Armbruster, B. N.; Ernsberger, P.; Irwin, J. J.; Shoichet, B. K., Relating protein pharmacology by ligand chemistry. *Nat Biotechnol* **2007,** *25* (2), 197-206.
33. Keiser, M. J.; Hert, J., Off-target networks derived from ligand set similarity. *Methods Mol Biol* **2009,** *575*, 195-205.

34. Jnr, J. K. In *On the shortest spanning subtree and the traveling salesman problem*, Proc. Amer. Math. Soc, 1956; pp 48-50.
35. Smoot, M. E.; Ono, K.; Ruscheinski, J.; Wang, P.-L.; Ideker, T., Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* **2010,** *27* (3), 431-432.
36. Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E. L., Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **2001,** *305* (3), 567-80.
37. Bemis, G. W.; Murcko, M. A., The properties of known drugs. 1. Molecular frameworks. *J Med Chem* **1996,** *39* (15), 2887-93.
38. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** *55* (14), 6582-94.
39. Weiss, D. R.; Bortolato, A.; Tehan, B.; Mason, J. S., GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J Chem Inf Model* **2016,** *56* (4), 642-51.
40. Truchon, J. F.; Bayly, C. I., Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model* **2007,** *47* (2), 488-508.
41. Jordan, B. A.; Cvejic, S.; Devi, L. A., Opioids and their complicated receptor complexes. *Neuropsychopharmacology* **2000,** *23* (4), S5-S18.
42. Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research* **2013,** *41* (Database issue), D1096-103.
43. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic acids research* **2000,** *28* (1), 235-42.
44. Alexander, S. P.; Davenport, A. P.; Kelly, E.; Marrion, N.; Peters, J. A.; Benson, H. E.; Faccenda, E.; Pawson, A. J.; Sharman, J. L.; Southan, C.; Davies, J. A.; Collaborators, C., The Concise Guide to PHARMACOLOGY 2015/16: G protein-coupled receptors. *Br J Pharmacol* **2015,** *172* (24), 5744-869.

CHAPTER 5.

Development of a Combined Ligand- and Structure-Based Virtual Screening Approach for the Discovery of Novel Bifunctional μ-Opioid Agonist/δ-Opioid Antagonist Compounds

1. Introduction

Opioids are a class of molecules commonly used in drugs for regulating pain. Despite their effectiveness as pain killers, available opioids commonly result in deleterious side effects, such as constipation and respiratory depression.[1] Compounded with their addictive nature, the North American continent has been experiencing an ever-worsening opioid epidemic, resulting in increasing numbers of deaths from drug overdose in recent years. However, opioids still remain commonly prescribed for pain relief.

Long-term opioid usage has become more prevalent, but data establishing the efficacy of its long-term efficacy is scant.[2] Additionally, many physicians have begun to refer patients to specialists, which are typically neurologists; given the small specialist-to-patient ratio, this could potentially delay the management of pain.[2] With the long-term use of opioids comes the increased risk of addiction and misuse. In 2015 alone, there were reported to be about 2.4 million Americans classified as having abused opioids.[3] Compounding this is the ever-increasing amount of deaths resulting from opioid-related overdoses, increasing from 28,647 in 2014 to 33,091 in 2015.[4] With the opioid epidemic declared a public health emergency by the president of the United States of America in late 2017, it is imperative that safer, more efficacious opioids be developed to address the current status in pain management.

The opioid receptors are a family of G protein-coupled receptors (GPCR) responsible for the modulation of pain, motor control, and mood. Three subtypes (μ, δ, and κ) are pharmacologically mediated by opioids. The μ-opioid receptor (MOR) is the classical target associated with the analgesic effect of opioids. Apart from this intended effect, a plethora of side effects, such as respiratory depression, nausea, sedation, addiction, and constipation, accompany the use of

opioids.[1] As a result, the δ-opioid receptor (DOR)[5] and κ-opioid receptor (KOR)[6] have increasingly been seen as potential targets in drug development for novel opioids with reduced side effects.

Historically, there have been numerous efforts to develop an opioid that could exert analgesia while reducing or eliminating all associated side effects. In the late 19th century, Bayer marketed heroin as a non-addictive alternative to morphine in an effort to combat opioid addiction, though years afterwards it was revealed that users were ironically becoming addicts.[7] More recent research has tended to focus on the phenomenon of biased agonism, whereby a ligand could stabilize the conformation of a GPCR so that it preferentially goes through one signal transduction pathway by recruiting a heterotrimeric G protein over β-arrestin or vice versa.[8] Taking advantage of this, there have been several success stories with various GPCRs, especially in the development of opioids with reduced side effects; one of the most promising compounds found currently is oliceridine (TRV130) from Trevena,[9] which is awaiting NDA review at the time of writing. As with oliceridine, the vast majority of drug discovery efforts employ high throughput screening in order to sift through a gargantuan sea of chemical diversity to find functional compounds. However, this process is usually costly, time consuming, and laborious, so thus *in silico* methods, such as molecular docking, can be utilized to reduce the chemical space by screening a virtual compound library and selecting only prioritized compounds for pharmacological assays.[10]

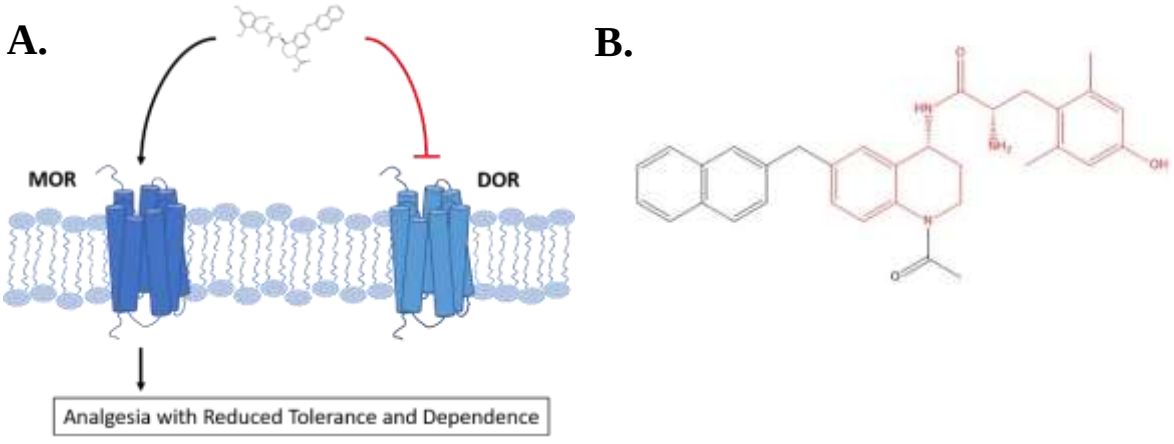


Figure 5.1 – (A.) Compounds that act as a MOR agonist and DOR antagonist can both elicit analgesia and reduce tolerance and dependence in animal models. The black and red lines represent agonist and antagonist function, respectively. (B.) Compound **14a** is a bifunctional peptidomimetic opioid with a tetrahydroquinoline (THQ) scaffold, which is shown in red.

There have been a number of computational studies reporting their findings on discovering biased agonists for MOR[11] and KOR[12-13] through *in silico* methods. While probing the concept of biased agonism remains a fertile area of research, an alternate mode to eliciting analgesia with reduced side effects has been steadily coming into the spotlight, as well. More specifically, studies have shown that co-administration of a MOR agonist and a DOR antagonist produces such a response in animal models.[14] Moreover, mixed-efficacy (a.k.a. bifunctional) MOR agonist/DOR antagonist compounds have been developed by several groups at the University of Michigan and shown to bind to both MOR and DOR while only activating MOR (Figure 5.1A).[15-16] In particular, the synthetic peptidomimetic opioid, compound **14a** (Figure 5.1B), was shown to have a very promising safety profile.[16] Additionally, other groups have concurrently developed bifunctional ligands with different scaffolds and similar effects.[17-18] Despite the potential of these bifunctional opioids as therapeutics, the compounds from these studies were based on traditional opioid structures, and thus, novel scaffolds that can potentially be developed into more efficacious, safer analgesics remain to be discovered. Importantly, one gap that has yet to be filled in this area of research is the application of *in silico* methods to the discovery of such compounds.

Most virtual screening campaigns performed today, such as the aforementioned virtual screening campaigns with biased agonists, tend to focus on a single target and a single screening method, but these can provide challenges for the discovery of bifunctional opioids. First, screening methods can typically be categorized as ligand based or structure based; the former utilizes pure chemical information in its search process, whereas the latter docks and scores compounds into a receptor structure. As with any method, there are advantages and disadvantages to each. Ligand-based methods are computationally inexpensive and can screen millions of compounds within a short time but have the drawback of being biased towards the known ligands used to build the model. Conversely, structure-based methods inherently have no bias, but they are extremely computationally expensive. A trend in recent years has culminated in the combination of these methods to address their respective shortcomings in various studies[19] and thus would be meaningful to pursue. Second, instead of screening one target, both MOR and DOR would have to be screened. A study employing a dual-target virtual screening approach against the serotonin 1B and 2B as target and anti-target, respectively, found selective agonists, which lends credence to this concept.[20]

In the present study, I present a combined ligand-based and structure-based virtual screening pipeline for the discovery of novel bifunctional opioids. The druglike subset of ZINC database[21] was screened against MOR and DOR sequentially using MAGELLAN, followed by docking with AutoDock Vina.[22] Docking results were then post-processed on the basis of rescoring, chemical novelty, and co-existence in top-ranked compounds of both MOR and DOR. Compounds selected for experimental validation were chosen from a combination of clustering and visual inspection.

2. Methods

All of the following procedures were performed with custom scripts in Perl and Python under the Red Hat Linux operating system. Retrospective virtual screens were run on a local computing cluster, while the prospective virtual screens were submitted to the Comet computing resource from XSEDE.[23] The chemical novelty filter was written in the C programming language.

2.1 Virtual Libraries

The druglike subset of ZINC12 database[21] served as the virtual library, consisting of over 13 million compounds prefiltered by Lipinski's rule of five.[24] For the ligand-based portion of the virtual screen, the database was downloaded as SMILES strings and converted into 1,024-bit Morgan fingerprints with a radius of 2 using RDkit.[25] For the structure-based portion of the virtual screen, the corresponding compounds in MOL2 format were downloaded and converted into PDBQT files using relevant Python scripts from AutoDockTools,[26] in which hydrogens and partial charges were re-added.

Active and decoy sets for MOR and DOR from GPCR-Bench[27] were downloaded from the DUD-E database[28] in the MOL2 and SMILES formats. For every active compound, there were 50 decoys with similar physicochemical properties but dissimilar 2D topology. Like before, all compounds were converted to PDBQT format, with re-added hydrogens and partial charges. Furthermore, all SMILES strings were converted into 1,024-bit Morgan fingerprints with a radius of 2 using RDkit.[25]

2.2 GPCR Structure Preparation

The agonist-bound MOR (PDB: 5C1M) and antagonist-bound DOR (PDB: 4EJ4) structures were downloaded from Protein Databank in PDB format. These were structures were selected because of their activation states, as the goal of the study is to predict bifunctional MOR agonist / DOR antagonist compounds. All small molecules and ions were removed from the PDB structures. In the original MOR structure, His54 on the N-terminus forms an interaction with the secondary amine of the agonist, BU72, though it was found not to be physiologically relevant by experiment.[29] Therefore, the N-terminus was removed to prevent the occlusion of the rest of the binding pocket. Conversely, the DOR structure did not require such a modification. Similar to the virtual libraries, the receptor structures were converted to PDBQT format, where hydrogens and partial charges were added with relevant Python scripts from AutoDockTools.[26] For AutoDock Vina,[22] 16 Å x 16 Å x 16 Å boxes were centered over the binding pockets for each receptor according to their crystallographic ligands.

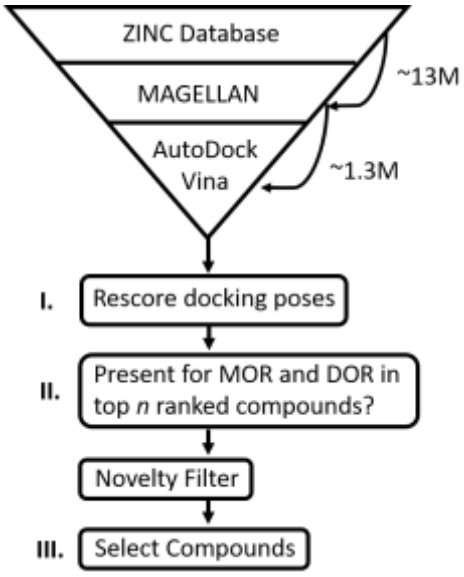
2.3 Retrospective Virtual Screen

MAGELLAN and AutoDock Vina, respectively, were the ligand- and structure-based algorithms used in present virtual screen campaign. In the former algorithm, known ligands from sequentially- and structurally-related GPCRs are used to infer what ligands a target GPCR would bind using a consensus chemical similarity approach, while in the latter, a compound is docked into the binding pocket of a receptor, conformationally sampled for the most optimal pose, and assigned a score. Retrospective virtual screening is typically employed to validate the performance of the virtual screening method and to gauge how likely it would succeed in the wet lab. To do so, a virtual library is spiked with known active ligands, and the algorithm of interest would aim to acquire as many active ligands as possible within a set number of top-ranked compounds. Enrichment factors are then calculated to provide a quantitative metric of how many fold over random the algorithm is operating. The calculation for the enrichment factor for the top 1% of the top-ranked compounds is shown as follows:

$$EF_{1\%} = \frac{N_{act}^{1\%} / N_{select}^{1\%}}{N_{act} / N_{tot}}$$

where N_{act} and N_{tot} are the total numbers of the active and all compounds, respectively. $N_{act}^{1\%}$ and $N_{select}^{1\%}$ are, respectively, the numbers of active ligands and the number of all candidates in the top 1% of the ranked database.

Receiver operating characteristic (ROC) curves were generated to evaluate how well it was able to discriminate between active and decoy compounds. The false positive rate (FPR) and true positive rate (TPR) are defined as the percent actives and decoys found, respectively. The scored compounds are ranked, then the FPR and TPR are calculated for each compound starting from the top-ranked compound and ending at the last one. Additionally, semi-log ROC curves were produced to emphasize the results from the top-ranked compounds of the database. This region of the database is the pool from which compounds typically get selected for experimental validation. An accompanying metric, Boltzmann-enhanced discrimination of the receiver operating characteristic (BEDROC), was used that correspondingly measures enrichment at this region.[30] An α=20 was used in the calculation, where the top 8% of the database accounted for 80% of the BEDROC score.[31]



Strategy	I.	II.	III.	# Cpds
A	X-Score (Top)	20,000	Manual Inspection	81
B	X-Score (Top)	5,000	Clustering	144
C	X-Score (Top9)	20,000	Manual Inspection	67
D	X-Score (Top9)	5,000	Clustering	166

Figure 5.2 – Prospective virtual screening pipeline with post-processing. (Left Panel) The main virtual screening portion lies within the inverted triangle. Roman numerals indicate the variations in post-processing procedures used. (Right Panel) The letters represent the four different post-processing strategies employed, including the number of compounds selected for experimental validation.

2.4 Prospective Virtual Screen Pipeline

An illustration of the prospective virtual screen pipeline is shown in the left panel of Figure 5.2. First, ~13 million compounds from ZINC12 database were first screened against both MOR and DOR with MAGELLAN. The top 1 million resulting ranked compounds from each receptor were pooled together into a targeted library, which consisted of ~1.3 million compounds. AutoDock Vina was then run against both MOR and DOR using the targeted library, which was carried out using 8 CPU cores and an exhaustiveness of 12. Subsequently, 9 docked poses were generated for each compound.

Following docking, the compounds underwent various stages of post-processing in order to improve the predictive accuracy and to provide a knowledge-based means to select compounds for experimental validation. First, the docked poses were rescored with an independent scoring function. Using the docking score from a docking program is typically not advised, as it was very likely optimized to find the best docking poses but not the best predicted affinities.[32] Therefore, many studies have increasingly relied upon rescoring or prioritizing docking poses through a separate means.[33-35] Both the top-scoring compound (top) or all 9 docked poses (top9) were rescored with either the knowledge-based scoring function, DSX,[36] or the empirical scoring function, X-Score.[37] In the case of top9, the poses were reranked on the basis of rescoring, and the top-scoring compound and its corresponding score was selected. It should be mentioned that X-Score was chosen over DSX on the basis of retrospective virtual screening performance and will be discussed later.

For both rescoring strategies, compounds were then selected for clustering or visual inspection (Figure 5.2, right panel). Compounds colocalizing in the top 5,000 ranked compounds for both MOR and DOR were selected and used in a chemical novelty filter. Here, each compound was compared against a master list of 5,153 MOR- and DOR-associated compounds with at most 10 μM in K_i, K_d, EC_{50}, or IC_{50} from GLASS database.[38] If a compound was at least 30% similar by Tanimoto coefficient to any ligand in the list, then it was discarded. The compounds were then subjected to hierarchical substructure clustering using LibMCS, JChemSuite 18.18.0, 2018, ChemAxon (http://www.chemaxon.com). The compounds from the lowest-level clusters were rescored by taking the average score from MOR and DOR, and the top-scoring from that cluster

was chosen for experimental validation. If there were 10 or more compounds in the cluster, then the top one-tenth of the compounds were chosen. Additionally, compounds colocalizing in the top 20,000 ranked compounds for both MOR and DOR were selected as with the top 5,000 and filtered for chemical novelty. PLIP[39] was utilized to filter out compounds that had interactions with Asp147 of MOR and Asp128 of DOR.[11] This was then followed by visual inspection of the resulting docked poses using PyMol.

3. Results and Discussion

3.1 Comparison Between Docked and Experimental Ligand Poses

As one of the first controls used in docking, reproduction of the experimental pose can provide a good initial quality control assessment of the procedure. To do so, the crystallographic ligands were stripped from the MOR and DOR structures. Subsequently, they were redocked with AutoDock Vina, and the top pose was evaluated. It was shown that pose reproduction with both structures produced good overlap between the crystallographic and docked poses, resulting in RMSD values with less than 1 Å deviation (Figure 5.3). Typically, RMSD values less than 2 Å indicate success in pose reproduction.[40] Taken together, this suggests that the docking procedure with AutoDock Vina performed adequately with the native ligand.

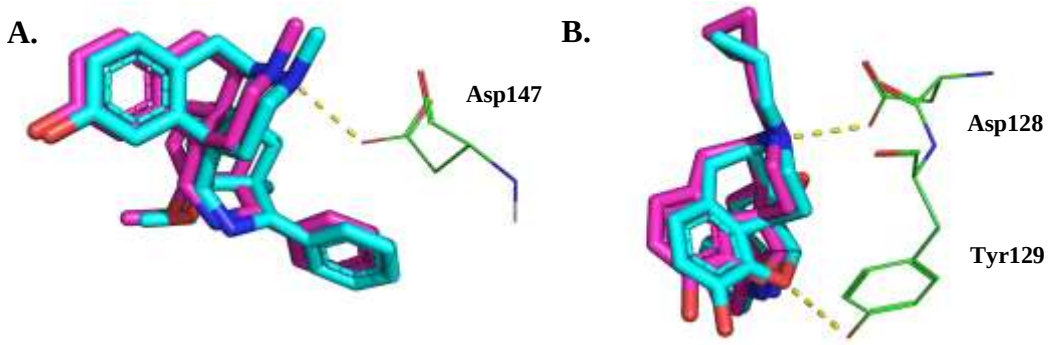


Figure 5.3 – Pose Reproduction of Co-Crystallized Ligands. For both (A) BU72 from MOR and (B) naltrindole from DOR, the crystallographic ligands (cyan) are overlain with the top predicted pose (magenta) by AutoDock Vina. Calculated RMSDs were 0.7 Å and 0.6 Å, respectively.

3.2 Model Validation of Ligand- and Structure-Based Methods

In order to examine and evaluate the ligand-based virtual screening component of the pipeline, retrospective virtual screens were run against MOR and DOR using MAGELLAN on the GPCR-

Bench dataset. A summary of screening statistics is given in Table 5.1. Overall, both MOR and DOR were able to achieve $EF_{1\%}$ values greater than 50-fold over random. Its ability to discriminate actives over decoys was very favorable (Figure 5.4A). More importantly, high BEDROC values (MOR: 0.824 / DOR: 0.842) were indicative of early enrichment in the area of the database where researchers would select compounds for experimental validation. Altogether, this suggests that MAGELLAN is able to achieve a high level of performance with both MOR and DOR in the GPCR-Bench dataset, and thus validate this component in the prospective virtual screening pipeline.

Table 5.1 – Retrospective Virtual Screening Statistics for MAGELLAN

	$EF_{1\%}$	AUC	BEDROC (α=20)
MOR	58.537	0.975	0.824
DOR	65.766	0.956	0.842

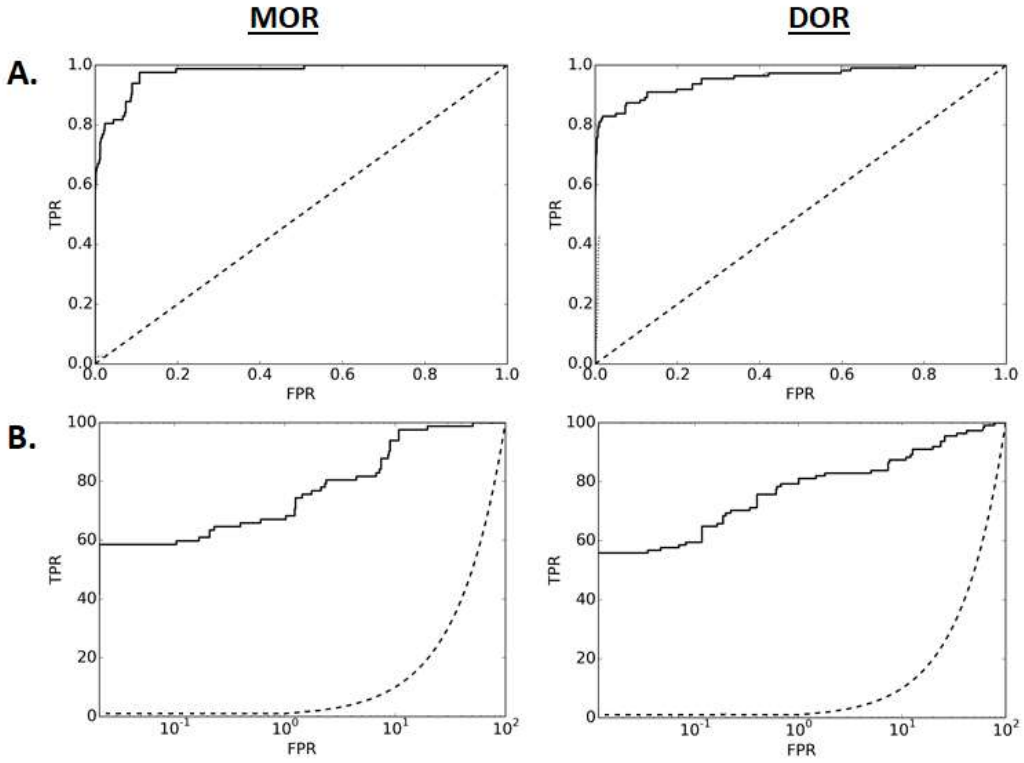


Figure 5.4 – Performance of MAGELLAN against GPCR-Bench Dataset. (A.) ROC curves are given for both MOR and DOR. (B.) Additionally, semi-log ROC curves are shown to emphasize the impact on early enrichment in virtual screening. The dashed lines represent random.

Molecular docking with AutoDock Vina was utilized as the structure-based virtual screening component of the pipeline. As with MAGELLAN, a retrospective virtual screen was run against both MOR and DOR using the GPCR-Bench dataset. A summary of screening statistics is given in Table 5.2. Using only the AutoDock Vina scores, enrichment of active compounds was poor for both MOR ($EF_{1\%}$=0.000) and DOR ($EF_{1\%}$=3.953). Furthermore, the ability to discriminate active compounds over decoys was close to random (Figure 5.5A). To account for this, two strategies were employed: 1.) Rescore the top pose from AutoDock Vina (Top), and 2.) Rescore all poses, re-rank, then select top pose again (Top9). The scoring function from docking algorithms is typically optimized for finding the best pose and is usually not a good substitute for approximation of the binding affinities. Additionally, the best pose is not always going to be the pose with the highest docking score, as it may be randomly-distributed among all the generated poses when ranked by the original docking score.[41] It has also been shown that rescoring docked poses with a more reliable predictor can result in better correlation between the predicted and experimental binding affinities.[42] Therefore, DSX and X-Score were utilized for rescoring.

DSX was able to slightly improve enrichment for MOR and DOR using both the Top and Top9 strategies. Discrimination between active compounds and decoys was also better (Figure 5.5A). More striking, however, was the increase in early enrichment (Figure 5.5B); performance with DOR saw a slight increase, while that with MOR increased from a BEDROC value of 0.078 to 0.221 (Top) and 0.201 (Top9). On the other hand, X-Score was able to improve the docking model even further than DSX. Enrichment was noticeably increased with this scoring method, where the Top9 strategy achieved $EF_{1\%}$ of over 10 for both receptors. Discrimination between active compounds and decoys was approximately the same for MOR between DSX and X-Score, though it improved to a greater extent for DOR (Figure 5.5A). Though early enrichment for MOR improved for X-Score as compared to DSX overall, it was extremely noticeable for DOR (Figure 5.5B). As a result, X-Score was chosen as the rescoring method for the prospective virtual screening pipeline.

Table 5.2 – Retrospective Virtual Screening Statistics for AutoDock Vina

	MOR			DOR		
	$EF_{1\%}$	AUC	BEDROC (α=20)	$EF_{1\%}$	AUC	BEDROC (α=20)
AutoDock Vina	0.000	0.637	0.078	3.953	0.611	0.139
DSX (Top)	3.239	0.742	0.221	3.765	0.707	0.202
DSX (Top9)	2.699	0.738	0.201	3.577	0.735	0.228
X-Score (Top)	7.826	0.750	0.284	8.660	0.812	0.380
X-Score (Top9)	10.253	0.750	0.283	10.166	0.847	0.424

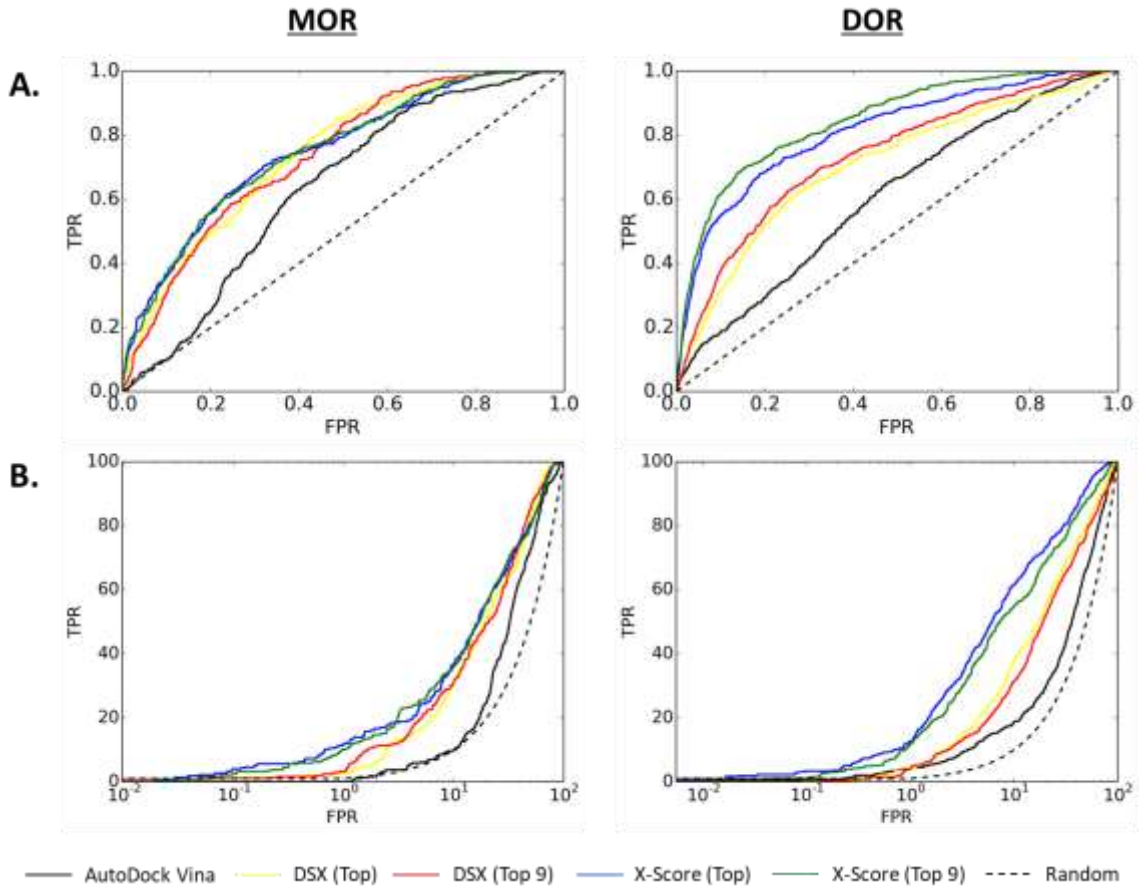


Figure 5.5 – Performance of Rescoring Docking Poses with GPCR-Bench Dataset. (A.) ROC curves are given for both MOR and DOR. (B.) Additionally, semi-log ROC curves are shown to emphasize the impact on early enrichment in virtual screening.

Upon further investigation of the docking scores generated from AutoDock Vina, it was observed that few active compounds achieved high ranks for both MOR and DOR (Figure 5.6A). However, upon rescoring with X-Score using either the Top (Figure 5.6B) or Top9 (Figure 5.6C) strategies, it was apparent that numerous active compounds were able to achieve higher docking scores

overall and relative to decoys. As such, this accounts for the increases seen in enrichment factors and AUC values for both receptors (Table 5.2). An interesting observation made from the data was that there was a slight dependence of the docking scores on the molecular weight of the compounds. This has historically been a common problem with scoring functions because larger molecules will likely possess higher scores due to their ability to establish a greater amount of interactions with the receptor.[43] Unfortunately, this means that the prospective virtual screen will likely have a slight bias towards higher-molecular weight compounds. Despite this, these larger compounds are equally viable as candidates for experimental validation and were treated as such in the current study.

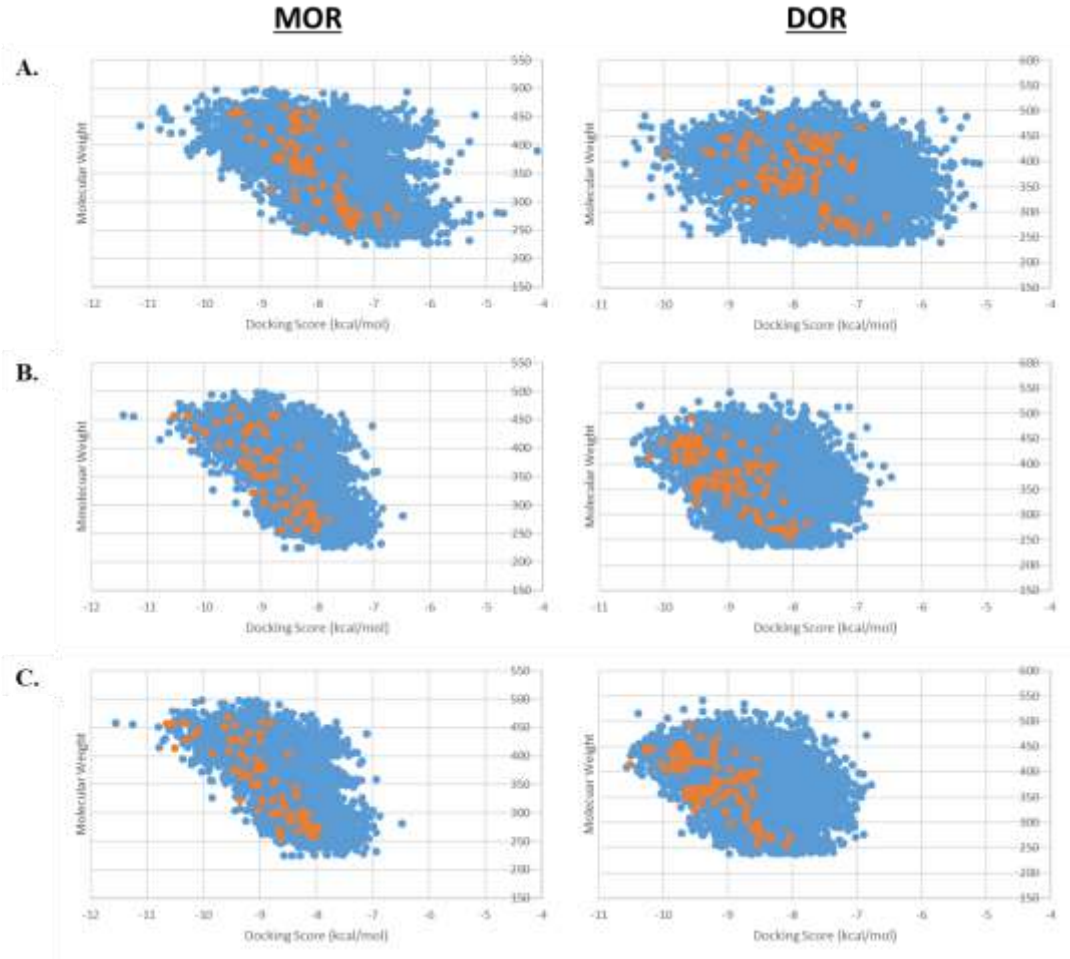


Figure 5.6 – Improvement of Docking Results with Rescoring Docked Poses. Docking scores from (A.) AutoDock Vina, (B.) X-score using the Top strategy, and (C.) X-score using the Top9 strategy are plotted against the molecular weights of the respective active compounds (orange) and decoys (blue). A shift of active compounds against decoys was apparent after rescoring, though a slight dependence of docking score on molecular weight was observed.

An important observation of the validation procedure is that docking appeared to perform worse quantitatively than MAGELLAN in the retrospective virtual screening tests against GPCR-Bench. While this is true, it must be reiterated that ligand- (i.e. MAGELLAN) and structure-based (AutoDock Vina) methods each have their own sets of advantages and disadvantages. While ligand-based methods are fast, they typically contain inherent biases towards known active ligands built into its algorithm. In contrast, structure-based methods contain no such biases towards any existing ligands whatsoever, though they are extremely computationally intensive. The prospective virtual screening pipeline was built in such a fashion that MAGELLAN could speedily reduce the chemical space of the virtual library, so that it could provide a targeted library for the more computationally-demanding AutoDock Vina. Moreover, docking can produce revelations about the conformation of a compound in the binding pocket that chemical similarity cannot. For example, a similar compound found by ligand-based virtual screening that has additional chemical moieties to known active compounds may not work experimentally, whereupon docking could account for this with a structure-based prediction on why it does not fit optimally into the binding pocket.

3.3 Evaluation of Prospective Virtual Screen

As a ligand-based method was used as the first step in the prospective virtual screening pipeline, there will be an inherent bias in acquiring compounds chemically-similar to the input in the top-ranking compounds. However, MAGELLAN does not operate solely with known active ligands of just MOR or DOR as inputs, as done with similarity-based virtual screening approaches; rather, ligands from homologous GPCRs are also used, potentially extending chemical coverage to opioid receptor relatives. Under the notion that similar receptors bind similar ligands,[44] it is possible that novel opioid chemotypes can be inferred from related GPCRs and found in the top-ranked region of the virtual library. This was observed from the virtual screening results, where numerous compounds passed a chemical novelty filter in which any compound at least 30% chemically similar to any known MOR or DOR ligands were excluded from the study. In fact, at least 35% of compounds subjected to the novelty filter passed for each of the post-processing strategies employed. Taken together, this suggests that MAGELLAN is able to discover novel chemotypes and does not entirely adhere to the classical notion of bias towards known active ligands.

Through the post-processing strategies involving manual inspection of the docking poses, 81 and 67 compounds were chosen through Top and Top9, respectively (Table 5.1). An example of selection based on interaction with Asp128 for MOR and Asp147 for DOR is given in Figure 5.7. It should be noted that the top 20,000 compounds for both receptors were considered, which is beyond the coverage of $EF_{1\%}$ metric that accounts for the top ~13,000 top-ranked compounds. However, it has been previously described that there is merit in looking further down a list of ranked compounds, when using visual inspection as a subjective criterion.[27] Using this notion, compounds were selected in this fashion to increase the chance of selecting active compounds based on classical interactions with the receptor. Additionally, clustering of the top 5,000 compounds for each receptor was performed and compounds selected from each cluster in order to generate a chemically-diverse set of compounds without having to require an inordinate amount of funding to experimentally validate. An example for this is shown in Figure 5.8, where ZINC02131167 was selected on the basis of its average docking score between MOR and DOR. Overall, unique 360 compounds were chosen from four different post-processing strategies (Table 5.1). Note that many compounds overlapped between the strategies, which explains why the numbers add up to greater than 360 in the table.

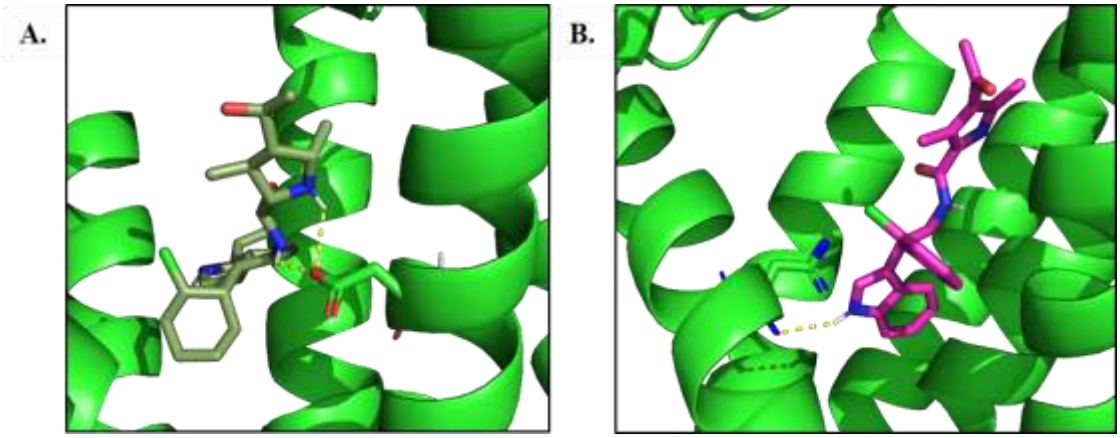


Figure 5.7 – ZINC Compound Selected by Visual Inspection of Docked Poses. (A.) Two secondary amines from ZINC09463337 (olive green) were found to form polar interactions with the one of the oxygens of Asp147 (dashed yellow lines) from MOR. (B.) A single secondary amine from ZINC09463337 (magenta) was observed to interact with the backbone amine of Asp128 (dashed yellow line) from DOR.

A final control for the prospective virtual screen was the inclusion of compound **14a** (Figure 5.1B) in the molecular docking validation. For both Top and Top9 strategies, it was able to achieve ranks

of 68 (0.005%) and 31 (0.002%) for MOR, respectively, and 634 (0.058%) and 878 (0.069%) for DOR, respectively; this is well within 1% of the ranked database and reflects very high predicted binding affinities. Interestingly, neither the Top or Top9 strategies produced interactions between the conserved aspartates (MOR: Asp147, DOR: Asp128) and compound **14a** (Figure 5.9). However, it should be noted that compound **14a** is 521 Da, a tad larger than the 500 Da threshold from the druglike subset of ZINC. Given the previous observation of a slight dependence of the docking score on the molecular weight (Figure 5.6), the use of compound **14a** as a positive control is called into question. Regardless, the usage of a larger active dataset, such as DUD-E, for model validation appears to be more reliable as a control, given the degree of uncertainty exhibited by virtual screening methods; the performance of a single compound should not hold sway over an entire virtual screening campaign. Though compound **14a** performed well in the current pipeline, it could potentially have failed with other strategies. Useful information could be gained from the inclusion of such a control, but careful consideration should be applied so that erroneous conclusions would not be met.

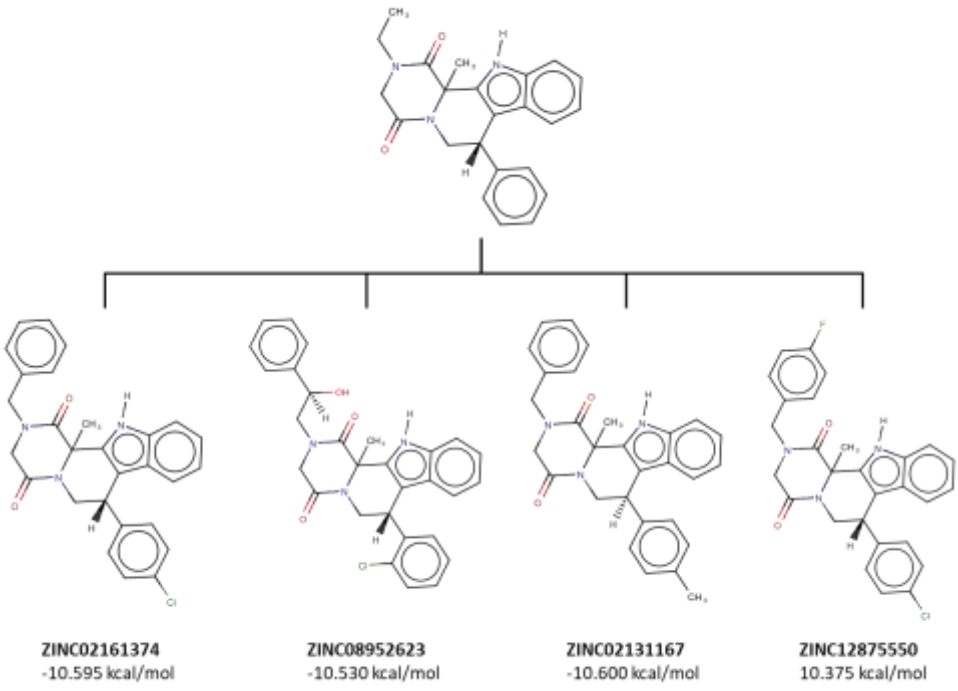


Figure 5.8 – Example of Clustering Results for Top9 Strategy. Shown here is the substructure (top) for one cluster that represents the four ZINC compounds (bottom). Docking scores are given as the average between MOR and DOR. ZINC02131167 selected from this cluster because it had the most favorable score. MarvinSketch was used for drawing and displaying the chemical structures, MarvinSketch 18.10.0, 2018, ChemAxon (http://www.chemaxon.com).

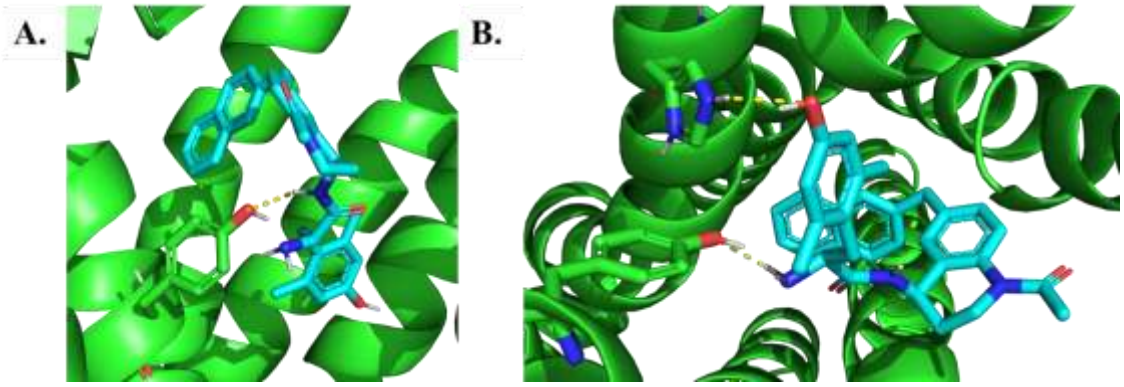


Figure 5.9 – Representative Docked Poses of Compound **14a** in MOR and DOR using Top9 Strategy. (A.) For MOR, a polar interaction was observed between a secondary amine and the hydroxyl group of Tyr148. (B.) For DOR, interactions were seen between the hydroxyl of the 2,6-dimethy tyrosine (Dmt) moiety and the secondary amine of the imidazole from His301 and between a secondary amine and the hydroxyl group from Tyr109.

4. Conclusion

In the present study, a virtual screening pipeline consisting of sequential ligand- (MAGELLAN) and structure-based (AutoDock Vina) methods was developed to predict bifunctional MOR agonist / DOR antagonist compounds. Benchmarking of MAGELLAN in retrospective virtual screening using the GPCR-Bench data set confirmed that ligand-based methods work well to retrieve active compounds when the known ligands are used in the model. Using the structures of both active MOR and inactive DOR, the crystallographic ligands were successfully redocked into the respective receptors, resulting in RMSD values of less than 2 Å. Furthermore, it was observed that docking with AutoDock Vina and ranking the compounds based on its score resulted in poor performance for both MOR and DOR in retrospective virtual screening. However, rescoring docking poses with DSX and X-Score comparatively improved performance. Overall, X-Score provided better discrimination of active compound over decoy and overall improved enrichment compared with DSX and was thus chosen for use in the prospective virtual screening pipeline.

Four post-processing strategies were used to refined the selection of compounds for experimental validation, which employed different rescoring strategies (Top and Top9), as well manual inspection and substructure clustering. This was done in order to ensure chemical diversity and structural insight in the selection procedure. A control bifunctional opioid, compound **14a**, was spiked into the virtual screen and produced in high-ranking results. Overall, 360 compounds were

found to fit the criteria. While the prospective virtual screening pipeline has been established, its success rate cannot be determined unless the compounds are experimentally validated. Thus, cAMP and β-arrestin assay will be run on a select subset of compounds to determine whether they function as opioids.

Among the insights gained from this study, one of the most important was the fact that docking scores appear to be biased towards higher molecular weight compounds in general. As larger compounds typically make more interactions, the docking score will reflect as such. To remedy this, ligand efficiency measures can be introduced. Further retrospective screens will need to be run in order to examine whether this type of post-processing would improve the model.

5. References

1. Ricardo Buenaventura, M.; Rajive Adlaka, M.; Nalini Sehgal, M., Opioid complications and side effects. *Pain physician* **2008,** *11*, S105-S120.
2. Rep, M. R., A persistent pain. *MMWR Recomm Rep* **2016,** *65*, 1-49.
3. Tetrault, J. M.; Fiellin, D. A., MOre beds or more chairs? using a science-based approach to address the opioid epidemic. *Annals of Internal Medicine* **2018,** *168* (1), 73-74.
4. Rudd, R. A., Increases in drug and opioid-involved overdose deaths—United States, 2010–2015. *MMWR. Morbidity and mortality weekly report* **2016,** *65*.
5. Spahn, V.; Stein, C., Targeting delta opioid receptors for pain treatment: drugs in phase I and II clinical development. *Expert opinion on investigational drugs* **2017,** *26* (2), 155-160.
6. White, K. L.; Robinson, J. E.; Zhu, H.; DiBerto, J. F.; Polepally, P. R.; Zjawiony, J. K.; Nichols, D. E.; Malanga, C.; Roth, B. L., The G protein–biased κ-opioid receptor agonist RB-64 is analgesic with a unique spectrum of activities in vivo. *Journal of Pharmacology and Experimental Therapeutics* **2015,** *352* (1), 98-109.
7. Pettey, G. E., The Heroin Habit, Another Curse. *Alabama Medical Journal* **1903,** *15*, 174-80.
8. Wisler, J. W.; Xiao, K.; Thomsen, A. R.; Lefkowitz, R. J., Recent developments in biased agonism. *Current opinion in cell biology* **2014,** *27*, 18-24.
9. Soergel, D. G.; Subach, R. A.; Burnham, N.; Lark, M. W.; James, I. E.; Sadler, B. M.; Skobieranda, F.; Violin, J. D.; Webster, L. R., Biased agonism of the μ-opioid receptor by TRV130 increases analgesia and reduces on-target adverse effects versus morphine: a randomized, double-blind, placebo-controlled, crossover study in healthy volunteers. *PAIN*® **2014,** *155* (9), 1829-1835.
10. Tanrikulu, Y.; Kruger, B.; Proschak, E., The holistic integration of virtual screening in drug discovery. *Drug Discov Today* **2013,** *18* (7-8), 358-64.
11. Manglik, A.; Lin, H.; Aryal, D. K.; McCorvy, J. D.; Dengler, D.; Corder, G.; Levit, A.; Kling, R. C.; Bernat, V.; Hübner, H., Structure-based discovery of opioid analgesics with reduced side effects. *Nature* **2016,** *537* (7619), 185.

12. Negri, A.; Rives, M.-L.; Caspers, M. J.; Prisinzano, T. E.; Javitch, J. A.; Filizola, M., Discovery of a novel selective kappa-opioid receptor agonist using crystal structure-based virtual screening. *Journal of chemical information and modeling* **2013,** *53* (3), 521-526.
13. Zheng, Z.; Huang, X.-P.; Mangano, T. J.; Zou, R.; Chen, X.; Zaidi, S. A.; Roth, B. L.; Stevens, R. C.; Katritch, V., Structure-based discovery of new antagonist and biased agonist chemotypes for the kappa opioid receptor. *Journal of medicinal chemistry* **2017,** *60* (7), 3070-3081.
14. Hepburn, M. J.; Little, P. J.; Gingras, J.; Kuhn, C. M., Differential effects of naltrindole on morphine-induced tolerance and physical dependence in rats. *Journal of Pharmacology and Experimental Therapeutics* **1997,** *281* (3), 1350-1356.
15. Bender, A. M.; Griggs, N. W.; Anand, J. P.; Traynor, J. R.; Jutkiewicz, E. M.; Mosberg, H. I., Asymmetric synthesis and in vitro and in vivo activity of tetrahydroquinolines featuring a diverse set of polar substitutions at the 6 position as mixed-efficacy μ opioid receptor/δ opioid receptor ligands. *ACS chemical neuroscience* **2015,** *6* (8), 1428-1435.
16. Harland, A. A.; Yeomans, L.; Griggs, N. W.; Anand, J. P.; Pogozheva, I. D.; Jutkiewicz, E. M.; Traynor, J. R.; Mosberg, H. I., Further optimization and evaluation of bioavailable, mixed-efficacy μ-opioid receptor (MOR) agonists/δ-opioid receptor (DOR) antagonists: balancing MOR and DOR affinities. *Journal of medicinal chemistry* **2015,** *58* (22), 8952-8969.
17. Healy, J. R.; Bezawada, P.; Shim, J.; Jones, J. W.; Kane, M. A.; MacKerell Jr, A. D.; Coop, A.; Matsumoto, R. R., Synthesis, modeling, and pharmacological evaluation of UMB 425, a mixed μ agonist/δ antagonist opioid analgesic with reduced tolerance liabilities. *ACS chemical neuroscience* **2013,** *4* (9), 1256-1266.
18. Gomes, I.; Fujita, W.; Gupta, A.; Saldanha, S. A.; Negri, A.; Pinello, C. E.; Eberhart, C.; Roberts, E.; Filizola, M.; Hodder, P., Identification of a μ-δ opioid receptor heteromer-biased agonist with antinociceptive activity. *Proceedings of the National Academy of Sciences* **2013,** *110* (29), 12072-12077.
19. Drwal, M. N.; Griffith, R., Combination of ligand- and structure-based methods in virtual screening. *Drug Discov Today Technol* **2013,** *10* (3), e395-401.
20. Rodríguez, D.; Brea, J.; Loza, M. I.; Carlsson, J., Structure-based discovery of selective serotonin 5-HT1B receptor ligands. *Structure* **2014,** *22* (8), 1140-1151.
21. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *J Chem Inf Model* **2012,** *52* (7), 1757-68.
22. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010,** *31* (2), 455-61.
23. Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D., XSEDE: accelerating scientific discovery. *Computing in Science & Engineering* **2014,** *16* (5), 62-74.
24. Lipinski, C. A., Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004,** *1* (4), 337-341.
25. Landrum, G., RDKit: Open-source cheminformatics. *Online). http://www. rdkit. org. Accessed* **2006,** *3* (04), 2012.
26. Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J., AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry* **2009,** *30* (16), 2785-2791.

27. Weiss, D. R.; Bortolato, A.; Tehan, B.; Mason, J. S., GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J Chem Inf Model* **2016,** *56* (4), 642-51.
28. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** *55* (14), 6582-94.
29. Huang, W.; Manglik, A.; Venkatakrishnan, A.; Laeremans, T.; Feinberg, E. N.; Sanborn, A. L.; Kato, H. E.; Livingston, K. E.; Thorsen, T. S.; Kling, R. C., Structural insights into μ-opioid receptor activation. *Nature* **2015,** *524* (7565), 315.
30. Truchon, J. F.; Bayly, C. I., Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model* **2007,** *47* (2), 488-508.
31. Sastry, G. M.; Inakollu, V. S.; Sherman, W., Boosting virtual screening enrichments with data fusion: coalescing hits from two-dimensional fingerprints, shape, and docking. *J Chem Inf Model* **2013,** *53* (7), 1531-42.
32. Guimarães, C. R.; Cardozo, M., MM-GB/SA rescoring of docking poses in structure-based lead optimization. *Journal of chemical information and modeling* **2008,** *48* (5), 958-970.
33. Anighoro, A.; Bajorath, J. r., Three-dimensional similarity in molecular docking: prioritizing ligand poses on the basis of experimental binding modes. *Journal of chemical information and modeling* **2016,** *56* (3), 580-587.
34. Kooistra, A. J.; Vischer, H. F.; McNaught-Flores, D.; Leurs, R.; De Esch, I. J.; De Graaf, C., Function-specific virtual screening for GPCR ligands using a combined scoring method. *Scientific reports* **2016,** *6*, 28288.
35. Alves, M. J.; Froufe, H. J.; Costa, A. F.; Santos, A. F.; Oliveira, L. G.; Osório, S. R.; Abreu, R.; Pintado, M.; Ferreira, I. C., Docking studies in target proteins involved in antibacterial action mechanisms: Extending the knowledge on standard antibiotics to antimicrobial mushroom compounds. *Molecules* **2014,** *19* (2), 1672-1684.
36. Neudert, G.; Klebe, G., DSX: a knowledge-based scoring function for the assessment of protein–ligand complexes. *Journal of chemical information and modeling* **2011,** *51* (10), 2731-2745.
37. Wang, R.; Lai, L.; Wang, S., Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *Journal of computer-aided molecular design* **2002,** *16* (1), 11-26.
38. Chan, W. K.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Ozgur, A.; Zhang, Y., GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* **2015,** *31* (18), 3035-42.
39. Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M., PLIP: fully automated protein–ligand interaction profiler. *Nucleic acids research* **2015,** *43* (W1), W443-W447.
40. Jiang, L.; Rizzo, R. C., Pharmacophore-based similarity scoring for DOCK. *The Journal of Physical Chemistry B* **2014,** *119* (3), 1083-1102.
41. Plewczynski, D.; Łaźniewski, M.; Augustyniak, R.; Ginalski, K., Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of computational chemistry* **2011,** *32* (4), 742-755.
42. Durrant, J. D.; Friedman, A. J.; Rogers, K. E.; McCammon, J. A., Comparing neural-network scoring functions and the state of the art: applications to common library screening. *Journal of chemical information and modeling* **2013,** *53* (7), 1726-1735.

43. Schulz-Gasch, T.; Stahl, M., Scoring functions for protein–ligand interactions: a critical perspective. *Drug Discovery Today: Technologies* **2004,** *1* (3), 231-239.
44. Klabunde, T., Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br J Pharmacol* **2007,** *152* (1), 5-7.

CHAPTER 6.
Conclusions and Future Directions

1. Conclusions

Altogether, this dissertation had the goal aiming for the construction of a virtual screening pipeline from start to finish. The topic of the first half was on GPCR-related database development, resulting in GPCR-EXP (structural data) and GLASS[1] (pharmacological data). However, the second half was concerned with algorithm and virtual screening pipeline development, producing MAGELLAN and the virtual screening pipeline for the discovery of bifunctional opioids.

GPCR-EXP was made in response to the dearth of GPCR experimental structure databases that had intuitive interfaces and consistent updating. This is not to say that similar resources, such as GPCRdb,[2] are not useful; they have their place amongst expert users that can benefit the most from their wealth of data. However, the lay user in GPCR structure biology would likely be most comfortable with an easy-to-use browsing interface for quick searches, which our database fulfills. If anything, GPCR-EXP would be considered a complementary resource. Furthermore, an entire roster of predicted GPCR structures of human origin were modelled by GPCR-I-TASSER[3] and provided on the database web server. Regarding GLASS database, it was originally developed to compensate for the now-defunct GLIDA,[4] which was a GPCR-ligand database that ceased updates in 2010. Pharmacological data was scraped from various external chemical databases and unified under one roof. As of the time of writing, GLASS contains over 500,000 unique GPCR-ligand associations, as opposed to about 39,000 from GLIDA's last update; this represents an over 12-fold increase in data. Both databases should impact the GPCR community for either structure- or pharmacology-related needs. In particular, GPCR-EXP has pre-processed structures containing only the GPCR, where fusion proteins and other associated proteins (i.e. nanobodies, G proteins, etc.) were filtered out; these would be convenient for molecular docking studies, saving the user the time in editing the PDB file. Furthermore, superposed structures of the same GPCR would be greatly useful in examining structural differences between different activation or ligand-bound

states. For ligand-based virtual screens, having a repository of pharmacological data in GLASS provides the foundations for algorithm development. Studies that have used this data in their pipelines include WDL-RF[5] and SwissSimilarity.[6]

The wealth of information in GLASS database paved the way for the development of MAGELLAN, a ligand-based virtual screening algorithm. After numerous revisions and refurbishments, the present form of MAGELLAN employs five sequence- and structure-based alignment algorithms, which gather related GPCR's. The corresponding ligand sets from these were clustered, where the top clusters were used in a ligand profile. In a large-scale retrospective virtual screening test against 224 Glass A GPCRs, MAGELLAN was able to achieve a median enrichment factor ($EF_{1\%}$) of 14.38, which is substantially higher than that detected using the individual GPCR-alignment methods. Furthermore, MAGELLAN was tested on two public virtual screening databases (DUD-E[7] and GPCR-Bench[8]) and achieved an average EF of 9.75 and 13.70, respectively, which compares favorably with other state-of-the-art docking- and ligand-based methods, including AutoDock Vina[9] (1.48/3.16), DOCK 6[10] (2.12/3.47) and PoLi[11] (2.2). A case study with the motilin receptor, a former orphan receptor, demonstrated its potential application in the de-orphanization of orphan GPCR's.

In the final portion of the dissertation, a combined virtual screening pipeline was developed, utilizing the previous works. Ligand- and structure-based methods each have their own set of advantages and disadvantages, and combining these together has been shown in various studies to compensate for the shortcomings.[12] Thus, a sequential virtual screening pipeline was constructed and applied to the discovery of novel bifunctional opioids with mu opioid receptor (MOR) agonist and delta opioid receptor (DOR) antagonist activity. Retrospective virtual screens against both MAGELLAN and AutoDock Vina were established, and both were reported to have over-random discrimination between actives and decoys. In particular, the docking-based screening did not perform as well as the ligand-based screening, but this was an expected sacrifice to enable the discovery of chemically-novel opioids. Furthermore, rescoring of docking results resulted in improved enrichment. Following a variety of post-processing procedures, 360 predicted bifunctional opioids were selected for potential experimental validation.

2. Future Directions

2.1 Additional Features for GPCR-EXP

The current form of the database includes superposed GPCR structures only of the same type. For example, all 4 structures for the mu opioid receptor have been structurally aligned. However, another feature that would be extremely useful for researchers would be the inclusion of the ability to align structures of whichever GPCR in which the user has an interest. This would enable convenient, custom access to aligned structures, as opposed to having one type provided. Additionally, being able to search for ligands using a molecular editor, such as JSME,[13] would be helpful for those users interested in GPCR ligands.

2.2 Increasing Data for GLASS

While massive in scale on its own right, there are additional sources of data that could be incorporated into GLASS database in order to bolster its content size. PubChem[14] has publicly-available sets of high throughput screening data from 1.25 million assays covering approximately 10,000 proteins. There is no doubt that this would be a valuable source of information, but sifting through, organizing, and cleaning up the data would be no easy feat. Nevertheless, it would greatly be worthwhile exploring this source of data for database integration.

2.3 Screening Orphan Receptors with MAGELLAN

Many orphan GPCR's have either no known function or endogenous ligand. In these cases, MAGELLAN could be applied to virtually screen chemical databases. A chemically-diverse set of top-scoring hits could be tested in the lab to check for receptor activation, which could potentially lead to further elucidation on function. Additionally, synthetic ligands that were experimentally found to bind to the orphan GPCR could be used as a scaffold for further investigation into what endogenous ligand it binds.

2.4 Experimental Validation of Predicted Bifunctional Opioids

As mentioned throughout, the only way to evaluate a virtual screen is to assay a set number of carefully-chosen compounds in the lab, which would be subjected to cAMP and β-arrestin assays. The former would measure the extent of receptor activation, while the latter would determine how much of the unwanted effector protein is being recruited. Both the mu opioid receptor (MOR) and

delta opioid receptor (DOR) would be screened in the hopes to find cAMP production in MOR but not DOR. However, inactivity at DOR does not imply antagonist properties. Thus, binding curves must be generated with known competitors to establish binding of a compound to both MOR and DOR. Collaboration with the Traynor lab is currently underway for the experimental validation of a few top-scoring predicted bifunctional opioids.

2.5 Alternative Metrics for Database Ranking in Virtual Screening

It was observed that there was a slight association of docking score with the molecular weight of the compounds during the retrospective virtual screen with the GPCR-Bench dataset. Therefore, many of the selected compounds were of higher overall molecular weight. It should be noted that they already constrained by Lipinski's rule of five,[15] as no compound was larger than 500 Daltons. An alternative way of rescoring compounds is to employ normalization using the ligand efficiency (LE) metric.[16] A common representation is shown as follows:

$$LE = \frac{\Delta G}{\#\ of\ Heavy\ Atoms}$$

A drawback to this calculation is that it does not scale linearly. Thus, modified metrics have been developed to account for this, such as fitness quality (FQ).[17] The equation is shown as follows:

$$FQ = \frac{LE}{LE_{scaled}}$$

where $LE_{scaled} = 0.0715 + \frac{7.5398}{\#\ of\ Heavy\ Atoms} + \frac{25.7079}{(\#\ of\ Heavy\ Atoms)^2} + \frac{361.4722}{(\#\ of\ Heavy\ Atoms)^3}$. This would be a very useful metric to utilize in order to acquire more compounds with smaller molecular weight in the top of the ranked database.

3. References

1. Chan, W. K.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Ozgur, A.; Zhang, Y., GLASS: a comprehensive database for experimentally validated GPCR-ligand associations. *Bioinformatics* **2015,** *31* (18), 3035-42.

2. Pándy-Szekeres, G.; Munk, C.; Tsonkov, T. M.; Mordalski, S.; Harpsøe, K.; Hauser, A. S.; Bojarski, A. J.; Gloriam, D. E., GPCRdb in 2018: adding GPCR structure models and ligands. *Nucleic acids research* **2017,** *46* (D1), D440-D446.
3. Zhang, J.; Yang, J.; Jang, R.; Zhang, Y., GPCR-I-TASSER: a hybrid approach to G protein-coupled receptor structure modeling and the application to the human genome. *Structure* **2015,** *23* (8), 1538-1549.
4. Okuno, Y.; Tamon, A.; Yabuuchi, H.; Niijima, S.; Minowa, Y.; Tonomura, K.; Kunimoto, R.; Feng, C., GLIDA: GPCR--ligand database for chemical genomics drug discovery--database and tools update. *Nucleic acids research* **2008,** *36* (Database issue), D907-12.
5. Wu, J.; Zhang, Q.; Wu, W.; Pang, T.; Hu, H.; Chan, W. K.; Ke, X.; Zhang, Y.; Wren, J., WDL-RF: predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest. *Bioinformatics* **2018,** *1*, 12.
6. Zoete, V.; Daina, A.; Bovigny, C.; Michielin, O., SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening. ACS Publications: 2016.
7. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012,** *55* (14), 6582-94.
8. Weiss, D. R.; Bortolato, A.; Tehan, B.; Mason, J. S., GPCR-Bench: A Benchmarking Set and Practitioners' Guide for G Protein-Coupled Receptor Docking. *J Chem Inf Model* **2016,** *56* (4), 642-51.
9. Trott, O.; Olson, A. J., AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* **2010,** *31* (2), 455-61.
10. Allen, W. J.; Balius, T. E.; Mukherjee, S.; Brozell, S. R.; Moustakas, D. T.; Lang, P. T.; Case, D. A.; Kuntz, I. D.; Rizzo, R. C., DOCK 6: Impact of new features and current docking performance. *J Comput Chem* **2015,** *36* (15), 1132-56.
11. Roy, A.; Srinivasan, B.; Skolnick, J., PoLi: A Virtual Screening Pipeline Based on Template Pocket and Ligand Similarity. *J Chem Inf Model* **2015,** *55* (8), 1757-70.
12. Drwal, M. N.; Griffith, R., Combination of ligand-and structure-based methods in virtual screening. *Drug Discovery Today: Technologies* **2013,** *10* (3), e395-e401.
13. Bienfait, B.; Ertl, P., JSME: a free molecule editor in JavaScript. *Journal of cheminformatics* **2013,** *5* (1), 24.
14. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic acids research* **2009,** *37* (Web Server issue), W623-33.
15. Lipinski, C. A., Lead-and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004,** *1* (4), 337-341.
16. Reynolds, C. H., Ligand efficiency metrics: why all the fuss? *Future medicinal chemistry* **2015,** *7* (11), 1363-1365.
17. Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D., Ligand binding efficiency: trends, physical basis, and implications. *Journal of medicinal chemistry* **2008,** *51* (8), 2432-2438.