A practical guide to molecular docking and homology modelling for medicinal chemists

Lohning, Anna E; Levonis, Stephan M; Williams-Noonan, Billy; Schweiker, Stephanie S

Published in:
Current Topics in Medicinal Chemistry

DOI:
10.2174/1568026617666170130110827

Licence:
Other

A Practical Guide to Molecular Docking and Homology Modelling for Medicinal Chemists

Anna E. Lohning[*], Stephan M. Levonis, Billy Williams-Noonan and Stephanie S. Schweiker
Faculty of Health Sciences and Medicine, Bond University, Gold Coast, 4229, Queensland, Australia

Anna Elizabeth Lohning: Faculty of Health Sciences and Medicine, Bond University, Gold Coast, 4229, Queensland, Australia

Stephan M. Levonis: Faculty of Health Sciences and Medicine, Bond University, Gold Coast, 4229, Queensland, Australia

Billy Williams-Noonan: Faculty of Health Sciences and Medicine, Bond University, Gold Coast, 4229, Queensland, Australia

Stephanie S. Schweiker: Faculty of Health Sciences and Medicine, Bond University, Gold Coast, 4229, Queensland, Australia

Corresponding Author:

Anna Elizabeth Lohning
Faculty of Health Sciences and Medicine,
Bond University, Gold Coast, 4229, Queensland, Australia
Telephone: 07 5595 4779   Fax: (07) 5595 4538
Email: alohning@bond.edu.au

Running Title

Practical Guide to Molecular Docking for Medicinal Chemists

Byline

Anna E. Lohning. PhD, Stephan M. Levonis, PhD, Billy Williams-Noonan, Stephanie S. Schweiker, PhD.

Abstract:

Elucidating details of the relationship between molecular structure and a particular biological end point is essential for successful rationally-based drug discovery. Molecular docking is a widely accepted tool for lead identification however navigating the intricacies of the software can be daunting. Our objective was therefore to provide a step-by-step guide for those interested in incorporating contemporary basic molecular docking and homology modelling into their design strategy. Three molecular docking programs, AutoDock4, SwissDock and Surflex-Dock, were compared in the context of a case study where a set of steroidal and non-steroidal ligands were docked into the human androgen receptor (hAR) using both rigid and flexible target atoms. Metrics for comparison included how well each program predicted X-ray structure orientation via root mean square deviation (rmsd), predicting known actives via ligand ranking and comparison to biological data where available. Benchmarking metrics were discussed in terms of identifying accurate and reliable results. For cases where no three dimensional structure exists we provided a practical example for creating a homology model using Swiss-Model. Results showed an rmsd between X-ray ligands from wild-type and mutant receptors and docked poses were 4.15Å and 0.83Å (SwissDock), 2.69Å and 8.80Å (AutoDock4) and 0.39Å and 0.71Å (Surflex-Dock) respectively. Surflex-Dock performed consistently well in pose prediction (less than 2Å) while AutoDock4 predicted known active non-steroidal antiandrogens most accurately. Introducing flexibility into target atoms produced the largest degree of change in ligand ranking in Surflex-Dock. We produced a viable homology model of the P2X1 purireceptor for subsequent docking analysis.

Keywords: molecular docking, homology modelling, medicinal chemistry, lead identification

1. Introduction

A key goal in many medicinal chemistry projects is to synthesize a database of compounds which promote or inhibit a particular biological action, typically mediated by a receptor or enzyme. Furthermore, the added commercial necessity to facilitate this process within a minimum of synthetic steps, cost or timeframe expounds the difficulties synthetic chemists face. Development of a lead compound through to clinical trials can take up to fourteen years and at an expense of around US800 million.[1] Computational approaches hold a valid place in the overall strategy yet the ability to accurately predict binding affinity is still challenging.[2, 3] The availability of high resolution structural data facilitates receptor-based design providing valuable insight into the molecular interactions between potential drugs and their targets. Engaging in receptor-based design and its incorporation into drug design, though desirable, can be a daunting prospect for many researchers.

Advances in computational power have facilitated modelling of more complex biomolecular systems such as the temporal passage of ions through membranes.[4-6] More realistic representations of biological systems are always desired as is the ability to accurately predict the interactions between a biological target and a ligand. Key barriers to this problem include accurately accounting for entropic contributions to the overall binding free energy (derived from desolvation or solvent effects) and conformational changes of protein and/or ligand.[7, 8] Apart from its role as a biological solvent, water functions in many biological processes such as desolvation, binding, stabilization and catalysis. Including explicit solvent molecules in a calculation however exponentially increases CPU time thus it is unsurprising that many algorithms simulate implicit solvent with the most common model being the Generalized Born (GB)/Surface Area model.[9] A comparison of free energy predictions by various GB models incorporating implicit solvent with dielectrics ranging from five (non-polar) to eighty (aqueous) with previous calculations using explicit solvent demonstrated that for high dielectric field environments where hydrogen bonding is important, the results were much less favourable than for those in a binding pocket or in a membrane interior.[10] This clearly demonstrates the importance of including explicit solvent in calculations for aqueous or polar systems.

Techniques for molecular modelling can be broadly divided into those models that are based on quantum physics and those that are not. Of the former, *ab initio* methods are based solely on approximations to the Schrӧdinger equation, focus on electronic systems and involve no experimental data. Semi-empirical methods incorporate experimental parameters and extensive approximations to the Schrӧdinger equation while molecular mechanics focuses on atomic nuclei utilizing classical mechanics. Molecular mechanics (MM) treats molecules as a system of balls (atoms) and springs (covalent bonds). Parameters for the various atoms and bond types, derived from experiment or *ab initio* methods, contribute to the potential function for the system called a force field. MM methods are best suited therefore to intramolecular analyzes such as conformational analyzes and determination of dynamic properties of molecules. Quantum mechanics (QM) is preferred when intermolecular interactions are of interest since electrons govern these short-range interactions. At present, however, system sizes are limited to less than one or two hundred atoms which is miniscule for a biological system. Therefore while QM techniques are preferable, in many cases biomolecular systems, modelled in their entirety, are somewhat restricted to MM techniques. More recently, combined QM/MM techniques have been developed. In these models, QM is often used to model interactions between a ligand bound into a binding site and MM for the remainder of the biomolecule.[11]

1.1 Molecular Docking

Among the most common MM techniques, molecular docking (MD) provides a convenient way to leverage structure for ligand discovery. Compared to laboratory-based, serendipitous, high throughput screening (HTS), MD can access far more chemistry more quickly and with far less cost.[12] Molecular docking is an MM approach to 'fit' a ligand into a three-dimensional binding site. Two operations are involved. Firstly a search of conformational space available to a ligand followed by a scoring function representing binding affinity. Despite the differences in these operations, MD programs often use the same algorithm for both tasks. Some may include a facility to perform consensus scoring when a number of algorithms have been included. Algorithms differ in the weight given to particular non-covalent interactions and/or entropic parameters producing a diverse set of results for the same ligand database in the same target. Understanding the particular emphasis given to these parameters for a chosen algorithm helps the interpretation of this diversity. Consensus scoring has been shown to reduce the prevalence of false positives in a docking study.[13]

MD uses a stochastic, or random, searching algorithm and thus, can be limited by the time allocated to the search. There is an obvious trade-off between accuracy and time since the longer the process is allowed to proceed, the more likely it will be to find the global minimum conformation. This assumes this lowest energy conformation is indeed the biologically relevant orientation however this is not always the case as in transition state models. Additionally a deep, narrow energy well may not be the more preferred as it ignores entropic considerations.

The docking process is remarkably fast if target atoms are kept rigid though the more patient user can expect higher quality results when the target incorporates flexibility in protein side chains. Depending on the computer system, target biomolecule and number of ligands used in these calculations, a single flexible docking can be performed in around one minute.[14, 15] Algorithms which are capable of flexible backbone and sidechain atoms will be orders of magnitude more complex and therefore more time and computationally expensive.

Assessing an MD algorithm's ability to successfully search conformational space in which to fit a ligand is usually done by comparing the conformations of X-ray and docked ligand poses. This assumes however there has been no change in conformation of the ligand from solution state to crystal structure. Determination of root mean square deviation (rmsd) provides a quantitative measure of comparison. MD scoring functions are considered to adequately reproduce the bound conformations as the global (or local) minima if less than the accepted level of 2Å.[16, 17]

The docking algorithm can be further assessed by determining its capacity to find a known inhibitor amongst a large set of mostly random non-bonding compounds or decoys. Note that a program that finds the 'correct' pose for a single ligand may not also be good at comparing molecules.

Docking algorithms must rank compounds in terms of binding affinity which may or may not bear any relation to activity. Binding free energy is an ensemble property (dependent on all possible states) and a statistical mechanics problem. Algorithms estimating this property from a single state (that is, the crystal structure) will therefore be unsatisfactory. A workaround is to conduct virtual screening in which known binding compounds are seeded into a small database of compounds. Algorithms able to score these known binding compounds amongst the higher scoring compounds is desirable. There is of course an underlying assumption that the scoring actually infers binding affinity but, as mentioned above, this is not necessarily the case. Strategies for validating predictions include benchmarking with other programs and comparison to biological assay data if available. With over sixty different molecular docking algorithms available each employing one of over thirty different algorithms selecting an appropriate program appears the first of many questions to be answered. Biological endpoints of interest are quoted in diverse units from IC_{50} to K_i further limiting direct comparison.

Three key limitations of MD are the inability to accurately model solvent, entropy and target flexibility. Performance can be tracked by including known actives as well as other, mostly random non-bonding compounds (decoys). Large benchmarking sets which aid the ligand enrichment process are available for well-known targets.[18] Decoys should ideally possess a degree of structural similarity to the active compounds for fair comparison of at least be drug-like and can be sourced from a number of databases. The latter strategy of ligand enrichment among top-ranking hits is a key metric of molecular docking. While MD has a relatively low success rate in the translation of top-scoring molecules to hits with actual binding affinity however, it is an important tool since the generation of a few compounds with new chemistry remains of interest.

This review provides a practical guide to MD and homology modelling for those wishing to integrate computational methodologies into their design process. Three molecular docking programs, AutoDock4[19], SwissDock[20, 21] and Surflex-Dock[22], are compared in a test case involving the human androgen receptor (hAR) target and androgen receptor inhibitor database. The hAR is an important therapeutic target and several 3D structures are available. We chose to compare the wild type hAR with a mutant containing a commonly observed mutation leading to castration resistant prostate cancer. In this case, the mutation creates a larger binding pocket into which ligands may interact. Most MD tutorials contain instructions on simply producing MD data which is often a relatively rapid process. However we provide a more complete workflow including interpretation of the data. We also compare the programs' functionalities as demonstrated within this case study. Additionally, for those cases where an X-ray structure is unavailable, a homology modelling case study is provided using Swiss-Dock to build a model of the P2X1 purinergic receptor. Tips and limitations are provided as well within each technique. It is beyond the scope of this guide to provide a comprehensive review of computer-aided drug design in general and a detailed description/comparison of forcefields and other parameters has been purposely left out for simplicity although links to relevant resources are provided. The reader is directed to the following comprehensive reviews.[23, 24]

1.2 Selecting a Suitable Program

Selecting a suitable MD program depends on a number of considerations such as cost, ease of use, computational capacity, etc. Once selected it is important to conduct appropriate validations and benchmarking against other docking algorithms. At the end of this review, we have included a comparison of features of the selected MD programs as an example of key features to consider.

Many docking programs are available and their performance recently reviewed.[25] We have selected three programs which represent various user preferences for freeware, online server access only and commercial software. SwissDock[20] represents a choice for those wanting to conduct MD on an online server. One of the gold standard MD programs commonly used is the freeware package, AutoDock4[19, 26] (AD4), developed by the Scripps Institute [27]. Finally we've selected the MD package Surflex-Dock included as part of the commercially available molecular modelling suite of applications, Sybyl-X [28].

1.3 Resources for Structural Data

The Protein Data Bank (PDB) (www.rcsb.org) is one of a few indispensable global repositories archiving the experimentally-derived atomistic models of biological entities, including proteins and nucleic acids.[29] Curation and management of the world wide PDB is jointly performed by the four partners: Research Collaboratory for Structural Bioinformatics (RCSB), PDB from Europe and Japan and the Biological Magnetic Resonance Bank (BMRB). The database contains structures of biomolecules produced by X-ray crystallography, NMR, electron microscopy of a hybrid approach. The X-ray structures are an interpretation of the electron density of a static molecular system. Though X-ray crystallography produces the most highly resolved structures, biological systems are inherently dynamic and this remain a limitation of the technique. At the time of writing there are almost 113,000 biological structures in the database, around 30,000 of which are protein structures of human sequences. In stark contrast there are around 560 unique membrane proteins currently in the database of Membrane Protein of Known Structure and around 2,600 transmembrane proteins in the Protein Databank of Transmembrane Proteins (PDBTM). One reason for this is that soluble biomolecules are experimentally more facile to crystallize. Unfortunately it is the latter that many medicinal chemists aim to target since, for example, nearly half of all drugs target membrane-bound G-protein coupled receptors. [30]

Many of the entries contain small molecules that may be non-covalently incorporated ligands, cofactors or ions or modified or uncommon amino acids within the polymer. Within the PDB, these small molecules have been annotated and collated into a chemical database (CCD). Small, biologically-relevant peptide-like antibiotic and inhibitor molecules present in the PDB have recently been collated the Biologically Interesting molecule Reference Dictionary (BIRD) (similar to Chemical Components Dictionary) which can be searched or downloaded for analysis. BIRD entries may appear as a polymer or ligand (or both) providing sequence or chemical information respectively. Annotations have been extensively classified and searches within the PDB can produce a great deal of useful information.[31]

Knowledge of the three dimensional (3D) structure of biological targets provides the platform from which receptor-based modelling techniques can be performed. Due to the disparity in available structural target information other computational approaches are required. Under these circumstances 3D homology models can be generated from amino acid sequences of the target of interest if, for example, the 3D structure of a similar target, perhaps of the same family, is known. Another approach, in the absence of a 3D structure, is the ligand-based approach producing a pharmacophore model prepared from a library of ligands with known binding affinities with a particular target.

Preparation of target and ligands prior to analysis is a key aspect to the success of an MD analysis. Although high resolution structures are available, it is essential to keep in mind that X-ray crystallography data is subjective in the interpretation of electron density and other interpretations may exist. Furthermore, conformations in the solution state may differ markedly from those in the crystalline form required for crystallography. While errors in interpretation of entire protein structures are rare, careful inspection is required. For instance, many flexible regions in a protein are often undefined and their respective x,y,z coordinates are therefore omitted. Chirality of ligands can sometimes be interpreted incorrectly as has been recently highlighted for oligosaccharide [32]. Correct positioning of water, sidechains of glutamine, asparagine and histidine is often challenging for crystallographers.[20] Protonation state of any residues within the binding site should be assessed prior to MD. A report to check target structure can be obtained at the PDB. Careful attention to ligand topology is essential to assess the treatment of tautomers, protonation states or other physicochemical features.

1.4 Selected MD programs

1.4.1 SWISSDOCK server

SwissDock, developed by the Molecular Modellers Group at the Swiss Institute for Bioinformatics, is a free service for academic users wanting to dock a set of ligands to a target biomolecule using a docking server. While there is a vast amount of information published regarding SwissDock as a docking tool, to the best of our knowledge there is no information available which adequately explains its use to the average chemist.

SwissDock presents an intuitive graphical user interface (GUI) clearly setting out tabs for submitting docking runs providing access to some general algorithm information and examples. SwissDock's algorithm is based on the dihedral space sampling (DSS) in EADock which, depending on the nature of the target and ligand, generally performs a fast, single step process, with little user-controlled input. Results can be used as a seed generator for further docking. Energy calculations are performed using the CHARMM (Chemistry at HARvard Macromolecular Mechanics) forcefield on the Vital IT cluster computer. Cluster groups of docked poses are scored and the results can be downloaded. MD runs are queued and delays can occur as a result. Results are stored for one week on the server. An option for private use is available.

SwissDock performs single ligand docking. For ligand databases, each must be docked individually. This can be burdensome for large ligand libraries. Scripts are available elsewhere for automating multiple-ligand docking however this is beyond the scope of this review. SwissDock is also unable to dock single ligands into multiple targets.

In the absence of a grid box definition, SwissDock server will perform a 'Blind Docking' whereby the algorithm searches for thermodynamically favourable sites into which to bind the ligand. However, since many binding sites are located within a crevice or channel and therefore may be ignored by the program. Toggling between accuracy levels from 'Very Fast', to 'Fast' or to 'Accurate', marginally increases the likelihood of obtaining a binding mode that correctly predicts promising binding conformations such as that found from a crystal structure from 62%, to 63% and to 64% respectively.

1.4.2 AUTODOCK 4 (AD4)

AD4 and AutoDockTools (ADT) are freely available automated docking software packages developed by the Department of Molecular Biology at the Scripps Research Institute, La Jolla, CA and the Department of Cognitive Science at the University of California, San Diego, La Jolla, CA.

AD4 (version 4.2) is a standalone cross-platform application operating on Linux, Mac OS, Windows and Sun Solaris operating systems. Intel i86 (32-bit), x86_64 (64-bit), and PowerPC processors are supported. ADT is the graphical front-end, python molecular viewer for using AD4 and is included as part of a package known as MGLTools, also provided by the Scripps Research Institute.[33] [34] A key benefit to using AD4-based programs is that they are Industry standards, robust and cost free. By using different front-end software, it is possible to complete the entire analysis within a GUI environment. Virtual screening is also possible using a software package known as PyRx.[35] Section 2.2.2 focuses on guiding the reader through the use of AD4 via the ADT front-end to dock the ligand dataset into the hAR target. We also demonstrate how to do multiple ligand screening using the PyRx front-end GUI.[19]

1.4.3 SURFLEX-DOCK

Surflex-Dock[36], developed by Tripos, is a commercially available ligand-receptor docking and virtual screening program and is part of the SYBYL-X suite of molecular modelling package.[28] Although users of SYBYL have access to extensive help documentation and tutorials, we provide a basic workflow using Surflex-Dock to perform the same docking analysis as shown with the previous two packages. The main benefits of Surflex-Dock include a robust algorithm, easy target/ligand preparation, comprehensive user-control over the docking process; consensus scoring, protomol guided docking, rigid and flexible docking including ring flexing, ability to dock large libraries of ligands and parallelization.

Setting up an MD run with Surflex-Dock can be easy when default settings are maintained. We recommend initially leaving settings for the protomol size and docking parameters at default values until the user becomes more familiar with the various functionalities and how they affect the results. Experienced users can achieve a higher level of accuracy and specificity via the high level of user control. Surflex-Dock makes use of a protomol rather than a grid box functionality to define conformational space when bound ligands are available. During docking the ligands are fragmented and each fragment used to search available space. High scoring fragments are retained and the final ligand reconstructed from those high scoring fragments. In the absence of a bound ligand key target residues can be selected otherwise the software is able to predict potential binding sites in a third option.

1.5 Homology Modelling

Protein sequences are the fundamental determinants of biological structure and function.[37] The field of structural biology has provided a plethora of knowledge on protein structure enabling prediction for the translation of a protein's primary sequence, through secondary structure (such as alpha helices, beta sheets, turns etc) to common motifs, domains, folds, tertiary and even quaternary arrangement of subunits.[38] Homology modelling is a technique which builds a 3D structure from a protein sequence of interest based on the 3D structure of a similar protein. In the absence of an available crystal structure for a protein of interest, a homology model can provide an alternative for subsequent receptor-ligand analyses such as molecular docking or dynamics as long as a three dimensional structure exists for a similar protein. Recent results from the 10th Critical Assessment of Structure Prediction (CASP) showed a dramatic increase in accuracy of homology models.[39]

An inherent problem in homology modelling is that two proteins may be almost identical structurally yet share very little sequence homology.[40] This has been somewhat overcome by fold-recognition technologies.[41] A sequence identity of 35% or higher is considered a rule of thumb for reliable homology modelling [42, 43] although such a cutoff may miss some structural or evolutionarily-related template sequences. Inaccuracies between template and models can also result from difficulties modelling loop regions caused by insertions or deletions in the sequence. Loop regions are often functionally important as sites for ligand attachment or as part of a regulatory mechanism. Loops can be treated either by comparison to a database of loop conformations of similar sequence or by energy minimization techniques.[44] Other functional domains however, such as active sites, are generally well conserved.[45] Despite a number of other limitations, including side chain conformations, resulting in sources of error in the model structures even low-accuracy structures can still be useful for investigating hypotheses about binding site location, substrate specificity and drug design.

Swiss-Model is an automated comparative protein modelling server freely accessible to non-commercial users via the Expasy server.[46-48] The basic workspace provides users the opportunity to save their work and develop their models. Most often you will have a clear idea of the molecular target to which you are designing molecular modulators. The National Centre for Biotechnology Information (NCBI), (http://www.ncbi.nlm.nih.gov/) is a good place to start your search for an amino acid sequence for your target protein.

Sequences are input in FASTA format, ubiquitous in molecular biology.

SwissProt initially searches the PDB for three dimensional structures that match your query sequence. The user selects a template from the returned matches. In general, the most suitable template will be that with the highest level of sequence identity. SwissProt operates in three modes, automated, alignment and DeepView Project mode.

Accuracy of 3D template structure is reflected in the resolution (in Å) whereby the lower the better. This ranges from excellent (~1Å) to poor (>3.5Å). Resolution represents the average uncertainty for all atoms. Uncertainty increases with disorder in the protein crystal during the crystallography process. Note that temperature (or B) factors represent the uncertainty for individual atoms. A high temperature factor represents a low empirical electron density for the atom. A range of acceptable values signifying reasonable position confidence would be 30Å^2-60Å^2.

SwissProt conducts a sequence similarity search of the PDB using the BLOSUM62 similarity matrix function. Homology models are assessed by a QMEAN4 scoring function for the estimation of the global and local model quality.[49] QMEAN4 consisting of four structural descriptors. A torsion angle potential over three consecutive amino acids, two pairwise distance-dependent potentials and a solvation potential describes the burial status of the residues. A Global Model Quality Estimation (GMQE), a number between 0 and 1, is used to reflect the reliability of the model with one reflecting the most reliable.

One of the main limitations of homology modelling includes a lack of suitable/reliable templates. Most structures in the PDB are of crystallographic origin and the majority of those represent fragments of the full-length proteins - often not more than 30% of the query sequence of interest. NMR solution structures are generally smaller monomeric proteins with an average of 90 amino acids. Highly disordered X-ray data can result in missing residues in template structures propagating further errors in the final model. Errors or uncertainties in the sequence alignment result in erroneous homology models. The quality of the alignment is crucial for a reliable model. In many cases models cannot correctly predict sidechain rotamer positions since correct geometries are not part of the homology model generation algorithms. (Rotamers are different conformations of a sidechain in three-dimensional space) Errors are likely with respect to steric clashes, electrostatic repulsions etc. These effects may or may not be minimized by subsequent energy minimizations.

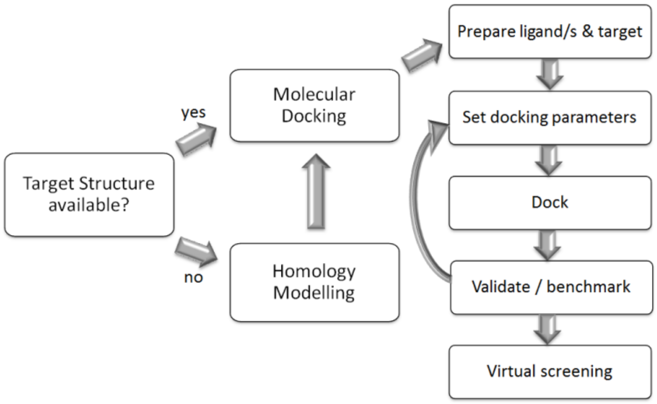
1.6 Scope

The purpose of this review was to provide a practical guide on incorporating MD into a design strategy by way of comparing three commonly used MD programs. Additionally, in cases where no crystal structure exists, a practical guide to homology modelling has also been included. It was assumed the reader possesses a basic familiarity with computers operating on either a Windows or Linux platform and write access for program installation.

The case study involved docking a set of androgen modulators into the wild type and mutant human androgen receptor sites. The database comprised known actives, decoys and a novel set of androgen receptor inhibitors.[50] Large datasets of active compounds as well as decoys specifically for the androgen receptor, designed for 3D virtual screening, were sourced from the database of useful decoys (DUD).[51]

Figure 1: Basic workflow for molecular docking.

A basic molecular docking workflow is presented in Figure 1.



2. Method

2.1 Ligand Database and Target Preparation

A small subset of steroidal and non-steroidal androgen receptor inhibitors along with native steroids and decoy (n=11) were prepared in SYBYLx-2.1 with hydrogens and Gasteiger-Hűckel charges added prior to energy minimization. Correct protonation was used where applicable. A structural similarity map was prepared which aids in grouping by structural features (Supplementary Figure S1). Biological, physical and chemical properties were determined for each ligand for subsequent correlation analysis and set out in Supplementary Table S1.

The wild type and mutant hAR structures, 2PNU and 2OZ7 respectively, were prepared with hydrogens and Gasteiger-Hűckel charges and ligands removed prior to energy minimization.

2.2 Molecular docking of androgen receptor inhibitors

2.2.1 MD using SwissDock

For non-specialists, setup of the target/small molecule docking process using SwissDock is intuitive and fast. While a command line option is available for experienced molecular modelers we recommend reading the associated documentation and practicing with provided tutorials to become more familiar with parameter functionality as default values may not always be appropriate. From the online server, results can often be returned within around thirty minutes depending on server queues, grid box size, exhaustive search parameters and flexibility options. Automated docking can be facilitated via the programmatic SOAP (Simple Object Access Protocol) interface supplying template scripts which can be downloaded in Perl, Python and PHP.

Table 1: Practical Guide to MD using SwissDock

Step	Task	Comments/recommendations
1	Target preparation.	Remove ligand and any other non-protein molecules. This may be performed in JMol [52] or other molecular visualization program.
1	Upload crystal structure via '**Target Selection** 'tab	PDB or CHARMM file formats can be accepted. CHARMM input files are also accepted including coordinate file (PDB), extra topology (RTF) and parameter files (PAR).
2	Upload ligand files via the '**Submit Docking**' tab. (see note below on multiple ligand docking)	Prepare single or multiple ligand files in **.mol2** format. Alternatively ZINC database identifiers can be used to access a large ligand library. Add hydrogens, check chirality, protonation state and topology. Prior energy minimization of the ligand is not required.
3	Add '**Extra Parameters**'	Select one of three levels of accuracy, Very Fast, Fast or Accurate, reflecting parameters such as binding mode (BM), sample size, number of minimization steps and number of BMs. Note for ligands with less than 15 rotatable bonds or those likely to fit exactly into a particular pocket, the first two modes may be suitable.
4	Specify '**Region of Interest**' to limit docking to a specific site (Local Docking). Select x,y,z coordinates (in Å) for the centre point and size of the Grid Box.	Units must be in Angstroms, Å, and correct to ***two decimal places***. Coordinates for a grid-box can be obtained from AutoDock Tools, however, *x,y,z* sizes for this grid-box generated by AutoDock Tools will be given in Grid-Box Units, which are 0.375Å each, and so a box with dimensions of 40×40×32 Grid-Box Units generated in AutoDock Tools must be converted to the equivalent dimensions of 15×15×12Å for use in SwissDock.
5	Specify residues for flexible docking via the '**Flexibility**' tab.	Select side-chains within 0, 3 or 5Å of a ligand.
6	'**Submit docking**'	Results sent by email with a link to SwissDock where the results may be viewed.
7	Assessment of docked poses.	A more comprehensive results assessment can be carried out using UCSF Chimera [53]. A direct link is provided with the results where poses can be assessed and compared in relation to the protein target. Scores are provided in units of estimated free energy or full fitness.

2.2.2 MD Using AUTODOCK

Download and install both AD4 and ADT from the Scripps Research Institute's website. Calculations are performed in several steps. First is the preparation of coordinate files using ADT followed by pre-calculation of atomic affinities using AutoGrid prior to the docking of ligands using AD4. Analysis of results is performed using ADT.

The python molecular viewer's main screen provides access to various computational features. On the left is a list of molecules including macromolecules and ligands. For single ligand docking click on the 'AutoDockTools' icon which brings up the docking menu. Work from left to right across the ADT toolbar. After reading in your macromolecule and ligand, there are a number of preparative steps required to prepare both the macromolecule and ligand for a docking analysis. A key step is the preparation of a new file format (*.pbdqt). It is then optional to define any flexible residues in the macromolecule. After this, the user needs to define a grid box to define an area of the macromolecule over which the docking analysis will take place. If key binding residues are known, this is a relatively simple process and is performed with a three dimensional graphical aid. Too large a grid box may result in a long processing time, and too small may prevent accurate results. For our analysis in both SwissDock and AutoDock we defined the Gridbox with centroid for *x*, 27.00Å; *y*, 2.00Å and *z*, 2.75Å and dimensions *x* and *y*, 15.00Å: *z*, 12.00Å (Figure 2).

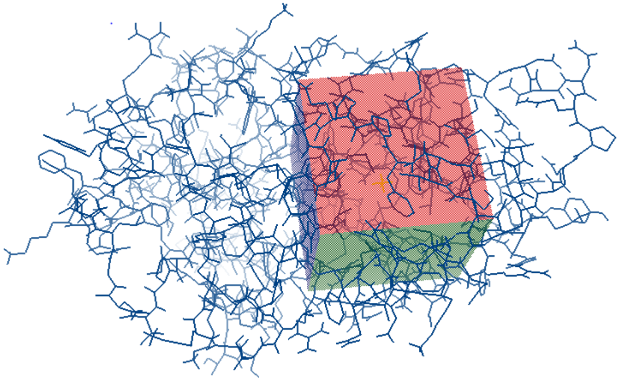


Figure 2: Grid box display in ADT.

After the grid box has been defined the user can now save this (as a *.gpf file) and choose both the macromolecule and ligand file to be used for docking (these must be pre-prepared as described earlier). Also the user must now define other docking parameters, these are also prefilled with default values for novice users. After this, the docking parameter file must be generated as a *.dpf file which will be used for the docking run.

At this stage the user is ready to firstly run AutoGrid and then AutoDock from within the GUI. AutoDock relies on the results generated by AutoGrid and expects to find these in the working directory. AutoDock can now be run. After the process has competed, the docking results can be reviewed using the "analyze" features of ADT.

Table 2: Practical Guide to MD using AD4.

Step	Task	Comments/Recommendations
1	Prepare directories for use	Create new folders for .pdb and ligand files. (avoid spaces in the pathname).
2	Prepare software for use	Open ADT, set working directory to the folder created in step 1. **File>Preferences>Set**

3	Prepare the macromolecule	Delete water molecules. Add hydrogens. Use '**grid**' menu to save the molecule as a **.pbdqt** file. Gasteiger charges are added automatically. **Edit>Delete Water** **Edit>Hydrogens>Add** **Grid>Macromolecule>Choose**
4	Prepare the ligand	Open ligand file from "ligand" menu. Gasteiger charges are automatically added. Ligands (in.*mol2*format) must have hydrogens already added. **Ligand>Input>Open**
5	Prepare the ligand	Use "**torsion tree**" to define a central atom (root) before "**choose torsions**" to choose rotatable (active) bonds. Save ligand as a **.pdbqt** file. **Ligand>Torsion Tree>Detect Root** **Ligand>Torsion Tree>Choose torsions** **Ligand>Output>Save as PDBQT**
5	Define grid box	Grid box represents search space for docking process. Define area using GUI by dragging box over the molecule, or type in the parameters manually. **Grid>Macromolecule>Choose** **Grid>Grid Box...**
6	Set map types	By choosing a ligand, the map types used for the AutoGrid calculations can be automatically identified. A *.gpf* file can now be saved that will define the parameters for the AutoGrid process. **Grid>Set Map Types>Choose Ligand** **Grid>Output>Save GPF**
7	Run AutoGrid	AutoGrid will run using the **.gpf** file. Output files are found in working directory. **Run>Run AutoGrid**
8	Prepare target and ligand	Choose the macromolecule and the ligand. Set the docking parameters, these are prefilled with default values. Output a "Lamarckian GA" to produce the *.dpf* file that contains the parameters for AD4. **Docking>Macromolecule>Set Rigid Filename** **Docking>Ligand>Choose** **Docking>Search Parameters>Genetic Algorithm** **Docking>Output**
9	Run AD4	Using the .dpf file AD4 will run and produce output files in the working directory. **Run>Run AutoDock**
10	Visualize Results	From the "Analyze" menu it is possible to show the conformations, ranked by energy, and visualize these with the ligand and macromolecule.

After AD4 has been used in the manner described above to dock a single ligand, it is possible to then use PyRx for the automated screening of a ligand library. PyRx can use the .pdbqt files produced by ADT, or can generate its own. This case study will assume the prior tasks have already been completed, and files from the same working folder can be used.

Table 3: Practical Guide to Virtual Screening with PyRx.

Step	Task	Comments/Recommendations
1	Open PyRx	PyRx generates its own working folder each time. To reset PyRx, delete this folder. Copy this folder as a backup.
2	Select AD4 wizard	AutoDock Wizard>**Start Here**
3	Choose local installation	Checkbox "local". Click "**start**" button.
4	Select Molecules	Select the ligand files for the library search (these can be unprepared *.mol2* files). Select the macromolecule file, it is convenient to use the pre-prepared *.pbdqt* file from the previous exercise. Click "**Add Ligand**" Click "**Add Macromolecule**" Click "**Forward**"
6	Run AutoGrid	Using the GUI, select the box size making sure to cover the target area. Click "**Run AutoGrid**"
7	Run AutoDock	Options exist to change the docking parameters. "Lamarckian (GA)" and "Maximum number of energy evaluations: short". A short setting for maximum number of energy evaluations hastens the docking process at the cost of accuracy. Click "**Run AutoDock**"
8	Analyze Results	Results are provided in the form of a table ranked by binding affinity. Clicking on a table entry highlights the ligand in the three dimensional view.

2.2.3 MD using SURFLEX-DOCK

Since SYBYL users have access to a high level of supporting material, we have refrained from reproducing it here. Instead we have included that which is necessary to perform a basic docking run for comparative purposes with the other programs.

Protomol Generation:

The Surflex-Dock protomol is a computational representation of the intended binding site to which putative ligands are aligned. The protomol is not meant to be an absolute docking envelope (Figure 3). Its purpose is to direct the initial placement of the ligand during the docking process. Docked ligands are scored in the context of the receptor, not in the context of the protomol. Protomol generation in Surflex-Dock is described by Jain.[54]

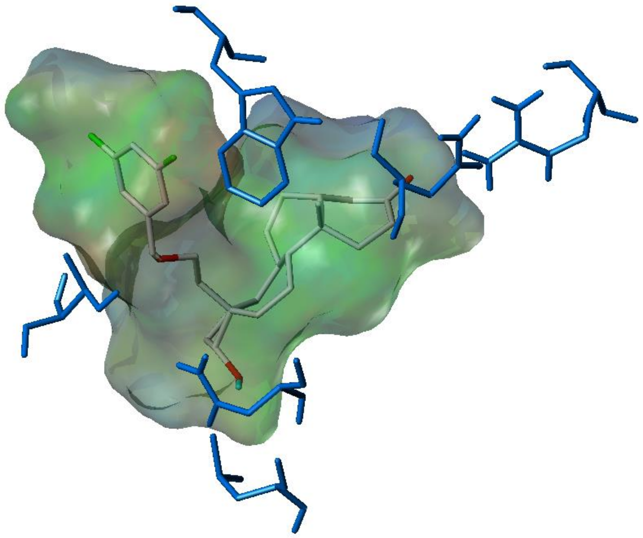


Figure 3: Surflex-Dock protomol: transparent surface coloured by lipophilic character (bloat 10, threshold 0.5). Wild type hAR (PDB ID: 2PNU). Key binding site residues blue; X-ray ligand, EM7544, atom types.

Scoring Functions:

Surflex-Dock incorporates a scoring function, Total Score, expressed as pK_D. An advantage of the Surflex-Dock algorithm is that it includes consideration of hydrophobic, polar, repulsive, entropic and solvation terms. The results also provide a breakdown of the total score into polar and repulsive contributions in terms of ‘crash’ and ‘polar’ scores. The smaller the crash score, the better Surflex-Dock is at screening out false positives. However, this may discard true positives that fit tightly in the pocket.[55, 56]

Protein flexibility can be incorporated at various levels of complexity from hydrogens to heavy atoms. Protein movement takes place in a second Surflex-Dock run, producing additional score set and accessible in the results browser. More consistent results can be produced by increasing the number of starting conformations.

Table 4: Practical Guide to MD using Surflex-Dock.

Step	Task	Commands	Comments
1	Prepare ligand database	Create a ligand dataset. If using ChemDraw save in **.sdf** format to be read by SYBYLx-2.1. Import into Sybylx-2.1.1 and save in a new molecular database. Save as single **.mol2** files and translate to **.sln** format.	*Ensure correct atom Types. Add hydrogens and charges. Minimize energy. (Our parameters:- Gasteiger-Hűckel charges, Amber7FF99 force field,[16] Conjugated. Termination conditions: ΔE<0.050 kcal/mol)
2	Prepare Target	**>Applications>docking suite>dock ligands** >Select docking mode*. >Define protein 1. Extract ligand from .pdb 2. Add charges and hydrogens to target	We recommend using GeomX for more exhaustive & accurate docking. Trade off with time.
3	Define Protomol	Select from three modes to define conformational search space. *Ligand* (if crystal ligand exists) *Residues* (if site is known) or *Multichannel* where potential binding sites are predicted. Determine optimal values for parameters threshold and bloat.	**Threshold** (buriedness) – between 0.01-0.99. (Default 0.50). Tip: Increasing threshold decreases volume. **Bloat** (inflates protomol). Default 0Å otherwise between (0-10Å.)
4	Set up Docking parameters	Options for flexible target docking mode, either ‘hydrogens’ only or key ‘heavy atoms’ Select appropriate ligand.	Flexible sidechains vastly increases run time. Test system firstly without flexibility.
5	Select Reference Ligand	**.mol2** file or from a molecular area	Automatically extracted in ***Ligand Mode.***
6	Select Number of Starting Conformations	Select 4 if using flexible ‘Hydrogens’ or 6 if selecting heavy atom flexibility.	Produces more consistent results especially in flexibility mode.
7	Run	Choose a file name and run docking.	Include total cores desired for run if applicable on your system.
8	Analysis of results	Select Results file and browse results. **>Applications>Docking Suite>Analyze results** Display and analyze within Sybylx2.1.1. Convert to spreadsheets to view and compare parameters, conduct bioinformatics, QSAR, correlations etc.	Ligands can be displayed in/with the protein; site; reference ligand; or protomol.

2.3 Homology Modelling using Swiss-Model

For this section we chose the sequence of the human ATP-gated, ion channel P2X1 receptor as no crystal structures currently exist and thus a relevant example. A model was built using Swiss-Model, based on the closest sequence available, that of the zebrafish P2X4 (PDB ID: 4DW1) receptor with a sequence identity of 45.8%.

Search results for the human P2X1 purinergic receptor sequence revealed 10 potential templates with the highest sequence identity with the P2X4 receptor with a sequence identity of 45.8 (Supplemental Figure S2). A clear picture of how close the sequences are to the query sequence can be displayed via the sequence similarity tab which depicts a cluster graph of the sequences.

We selected the 4DW0 for a template based on the highest global model quality estimation (GMQE) score of 0.68. This is a quality estimation which combines properties from the target-template alignment. The resulting 3D structure, shown in

Figure 4 (and corresponding sequence alignment in Supplemental Figure S3) shows sections of secondary structure coloured blue for high quality areas and orange for low quality areas. Residues in the alignment can be selected interactively which automatically identifies their 3D position in the model.

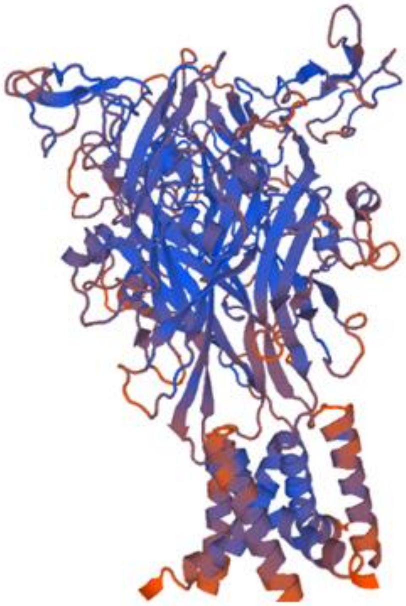


Figure 4: Template produced from query sequence of human P2X1 and target structure of 4DW0.

In this case, all three gaps shown in the alignment correspond to surface loops of the extracellular domain of the receptor rather than amid core β-sheets or α–helices. Thus the model could reasonably be considered for further molecular modelling projects.

Table 5: A Practical Guide to Homology Modelling.

Steps	Task	Comments
1	>Retrieve query sequence (see notes on 'Obtaining Target (query) Sequences)'.	Connect to NCBI[a], enter protein name including species. Select 'protein' in the database window. Select appropriate sequence, save in FASTA format. (Tip: save as a .txt file).
2	Conduct a multiple sequence alignment (MSA) (optional) (see notes on MSA).	The following tools provide options for conducting MSAs: ClustalOmega[b] or T-Coffee[c].
4	Select **>Start Modelling**	Navigate to the SwissModel[d] site to start a new modelling project.
5	Upload sequence (FASTA format) then select **>Search for Templates**	Templates will provide a number of possible 3D structures and their sequence identities to the query sequence.
6	View Output. Compare by selecting the top few matches.	Select templates to compare superimpositions in the viewer. Note the %Coverage column. Many sequences and targets contain only fragments of the entire sequence. It is important to match relevant domains of your protein for your model such as that containing the binding site.
7	Inspect the alignment and check for gaps. To analyse gaps more closely:- **>View Project in DeepView**. (see notes on Deepview)	(See Notes – Dealing with Gaps in Sequences) Analyse superimposition for any unresolved residues in the template. Identify positions of any gaps in either template/query sequence that may impact on the overall reliability of the model. Tip: Comparing secondary structure alignment is useful. For example, gaps in loops are usually permissible. Gaps in β-sheets may be more hazardous to the overall 3D model structure.
8	Select appropriate template and select **>Build Model**	Comparing multiple models (via superimpositions) is useful to identify the most appropriate model.
	Save model in *.pdb* file format. Additionally save the model file.	A simple viewer is available in SwissModel but more powerful freeware programs are available to visualize, display and manipulate your models, including superimpositions. We recommend Chimera. [53]

[a][57], [b][58], [c][59], [d][60]

Obtaining Target (query) Sequences

When searching the NCBI Protein database for amino acid sequences, retrieval of relevant information is fast and specific if using appropriate Boolean expressions. For example, to search for a sequence for the human P2X1, enter 'human+P2X1+receptor AND "Homo sapiens"'. Four entries were retrieved at the time of writing. Note that links are available for downloading the sequence in FASTA format which should be saved as a .txt file. (Edit in Notepad+ if required.)

Multiple Sequence Alignments (MSAs)

It may be the case that there are a number of sequences for relevant proteins with which to compare with your target protein or template. MSAs are common tools in many molecular biology projects for visualizing differences in amino acid sequences which translate to structural difference that may impact on function.

DeepView

In other cases where sequence alignment is not straight forward and insertions or deletions may be present, Swiss-Model has a structure viewer called DeepView which can be useful for making alterations prior to building the final model. Users with a deeper understanding of protein structure will find this process easier.

Dealing with Gaps in the Sequences

Gaps in the template sequence are due to insertions (or deletions) during the sequence alignment process whereas gaps in the query sequence means the two residues flanking the gap, while peptide bonded in the 3D model, could be far apart in the sequence.

When residues in the query sequence are untemplated, that is, there are no corresponding template residues, these are given a high temperature factor in the 3D model and can be easily selected for and visualized. Luckily long regions of untemplated residues can be visually easy to spot as they are often represented by hairpin loops extending out from the more compact protein structure.

In some cases, highly mobile residues which result in unresolved electron density and thus residues without coordinates in the template are not included in the sequence alignment nor indicated by a gap. This causes a splice in the template sequence effectively shifting the query-template alignment. The resulting homology model then will not reflect the presence of these residues. One way to avoid this is by first analyzing the structural alignment in DeepView. After visual inspection, the sequence of target/template may be carefully altered if needed or utilized in an MD protocol.

3. Results & Discussion

The MD case study aimed to compare the docked poses of a series of selective androgen receptor modulators (SARMs) into the hAR followed by a brief analysis of the results. Antiandrogens antagonize the actions of testosterone or 5α-DHT by competing for AR binding sites and may be steroidal or non-steroidal in structure. A small subset of what is otherwise a vast array of known SARMs was selected for this comparative demonstration. These include the natural substrate, testosterone and its active form, dihydroxytestosterone (DHT), steroidal androgen with high affinity, R1881, cyproterone acetate, a weak antiandrogen with high binding affinity, a number of non-steroidal antiandrogens including the clinically relevant, flutamide and bicalutamide and a number of other potential hAR modulators. A non-steroidal decoy of similar structure to the non-steroidal hAR modulators was obtained from the ZINC database [61] which reportedly does not bind the hAR.

The biological, physical and chemical properties of the ligands set out in Supplementary Table S1 are helpful in the interpretation of docking results to assess potential correlations relating to size, shape, polarity etc. These will be discussed in the next section.

3.1 Comparison of Docking Results

Analysis of docking results is often a complex task and is usually conducted by visual screening of the docked poses in their respective binding sites. Some programs can display the results including intermolecular hydrogen bonding which aids in lead identification. For example, Figure 5 shows the docked pose of EM7544 from wild type hAR with three hydrogen bonds to key binding site residues. Importantly, one key residue, T877, is the mutated residue displayed in the mutant T877A hAR. The replacement of threonine’s hydroxyl group with the small side chain of alanine creates a larger binding site in the mutant enabling a larger array of compounds to enter leading to receptor promiscuity.

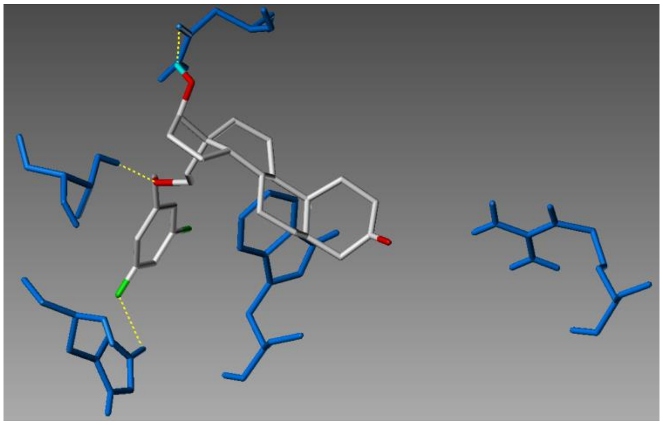


Figure 5: Docked pose of EM7544 (atom colours) from 2PNU (wild type hAR) depicting three intermolecular hydrogen bonds (yellow). Key binding site residues are shown in blue.

A comparison of each docking program was conducted to assess how well each was able to reproduce the orientation of the crystal ligand in the native and mutant receptor. A root mean square deviation (rmsd) of less than 2Å is generally acknowledged as acceptable performance. Note that this guideline assumes that the crystallized orientation is the active orientation, however; it is arguable that this may rarely be the case.

Figure 6 shows the orientations of the extracted ligand, EM5744, from 2PNU and respective superimposed docked ligands. It is clear that SwissDock performed least well on the basis of the above criteria. Both AD4 and Surflex-dock reproduced the orientation of the crystal ligand more closely although superimposition shows the Surflex-dock result to be an order of magnitude closer.

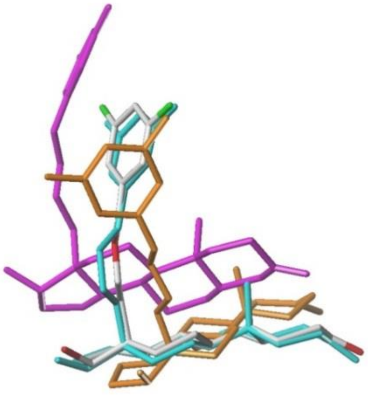


Figure 6: Conformation superimposition of docked EM5744 using: SwissDock (magenta, RMSD 4.15Å), AD4 (orange, RMSD 2.69Å) and Surflex-Dock (cyan, RMSD 0.39Å) with the conformation of EM5744 extracted from PDB ID: 2PNU (atom colours).

For the mutant hAR (PDB ID: 20Z7), however, SwissDock outperformed AutoDock in terms of comparison of rmsd of docked versus X-ray ligand conformations. Figure 7 shows the orientation of the extracted X-ray ligand, cyproterone acetate, from 2OZ7 superimposed with respective docked ligands using the 3 programs. Surflex-dock’s showed a consistently high level of performance.

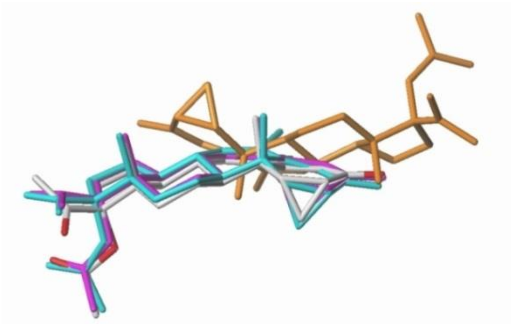


Figure 7: Conformation superimposition of docked cyproterone acetate using: SwissDock (magenta, RMSD 0.83Å), AD4 (orange, RMSD 8.80Å), Surflex-Dock (cyan, RMSD 0.71Å) and 2OZ7 X-ray ligand (atom colours).

Inter-program variations between algorithms are highlighted in Table 6 and 7 where we compared ligand ranking within the steroid binding sites of wild type hAR (2PNU) and mutant hAR (2OZ7) with both flexible and rigid target atoms. By comparing ligand ranking we avoid the complexity of non-uniformity in units of scoring between the various programs. Overall, incorporating flexible target atoms 5Å or less from any atom of the ligand made little difference in the SwissDock analysis although introducing flexibility boosted both wild type and mutant X-ray ligands top ranking positions from midfield. A similar trend was observed for the AutoDock analysis with little change in ranking by introducing flexibility. Interestingly Surflex-Dock produced more significant changes to ranking on the introduction of flexibility with the triazole compounds and flutamide showing a much higher ranking.

The grid box dimensions were kept consistent for AutoDock and SwissDock. Since Surflex-Dock utilizes a protomol instead to define ligand conformational search space, every effort was made to keep the overall area consistent. Note the volume of the wild type hAR binding site is somewhat smaller than the T877A mutant due to the reduction in side chain size as well as removal of a hydrogen bonding group. This mutation contributes to the development of treatment resistance in prostate cancer by driving receptor activation by the same compounds that initially acted as antagonists to the hAR.[62]

The flexibility of target atoms appeared to make minimal difference to ranking by AD4 in wild type hAR though more noticeable changes in the mutant site was observed. The largest change in ranking going from rigid to flexible target was with cyproterone acetate and R1881 where their positions were reversed from first to fourth respectively.

Analyzing the data from a biological interest viewpoint, that is, comparing the ranking of each ligand from wild type to mutant, revealed no change for the majority of ligands in the SwissDock analysis with a preference for the steroidal ligands. A similar trend was observed for AutoDock while Surflex-Dock showed a high degree of rank change going from wild type to mutant sites. Here no clear preference for steroidal was observed with triazole compound 39S ranking third. The X-ray ligand, EM7544 (from wild type target) was ranked 10th in the wild type to first in the mutant site. Similar results were observed for the X-ray ligand, cyproterone acetate (from the mutant target). Whether this is simply an artefact of the Surflex-Dock protocol of ligand-based protomol generation or may have any biological significance, is not known.

Table 6: Comparison of Ligand Ranking in wild type hAR.

Ligand[a]	Type	Rigid			Flexible		
		SWD	AD4	SFD	SWD	AD4	SFD
testosterone	Native	1	2	4	3	2	6
DHT	Native	2	3	6	1	3	7
EM7544 (2PNU)	Steroidal high affinity	6	8	1	4	6	2
Cyproterone acetate (2OZ7)	Steroidal inhibitor	9	1	10	2	4	8
R1881	Strongly binding androgen	4	4	5	6	1	9
Bicalutamide	Non-steroidal inhibitor	5	6	8	7	5	10
Flutamide	Non-steroidal inhibitor	3	7	7	5	8	11
CHEM 366105 (active)	Non-steroidal inhibitor	8	9	2	9	10	1
Triazole cpd 39S	Non-steroidal inhibitor	11	11	11	11	9	3
Triazole cpd 39R	Non-steroidal inhibitor	10	10	9	10	7	5
Zinc12848678 (decoy)	Non-steroidal Decoy	7	5	3	8	3	4

Analysis of the SwissDock results showed the smaller ligands obtaining a higher ranking in the both sites whether in a rigid or flexible target. The only anomaly in this trend was cyproterone acetate whose ranking increased in a flexible site. The flexible alkyl tail of this ligand may partly account for this observation. The larger compounds tended to rank lower.

In the wild type site only Surflex-Dock ranked the X-ray ligand first or second top scoring pose for both flexible and rigid docking. It is worth noting that these ligands are used as reference ligands in the docking protocol and may, as a result, bias this ligand towards the top. For the larger mutant site however all three programs ranked the X-ray ligand top score when run in flexible mode but only Surflex-Dock did so in rigid mode.

The triazole compounds were ranked low by all programs in flexible and rigid modes for the larger mutant site. The same results was found for the rigid mode docking within the wild type site however Surflex-Dock ranked them much higher in flexible mode compared to the other programs.

Table 7: Comparison of Ligand Ranking in mutant hAR.

Ligand[a]	Type	Rigid			Flexible		
		SWD	AD4	SFD	SWD	AD4	SFD
testosterone	Native ligand	2	1	5	5	4	9
DHT	Native ligand	3	2	3	3	3	7
EM7544 (2PNU)	Steroidal high affinity ligand	failed	8	11	2	5	2
Cyproterone acetate (2OZ7)	Steroidal inhibitor	6	4	1	1	1	1
R1881	Strongly binding androgen	7	3	6	6	2	10
Bicalutamide	Non-steroidal inhibitor	4	7	7	7	6	5
Flutamide	Non-steroidal inhibitor	1	6	4	4	8	11
CHEM 366105 (active)	Non-steroidal inhibitor	8	11	9	9	10	8
Triazole cpd 39S	Non-steroidal inhibitor	9	10	8	11	11	3
Triazole cpd 39R	Non-steroidal inhibitor	10	9	10	10	9	6
Zinc12848678 (decoy)	Non-steroidal Decoy	5	5	2	8	7	4

Post MD processing capabilities within Surflex-Dock include the ability to convert MD data into a molecular spreadsheet in which other physical or chemical properties can be added for each ligand to find potential correlations. Figure 8 and 9 compared the results for ligands docked into the wild type hAR and mutant sites respectively in flexible mode. The highest degree of correlation was found between molecular weight and total score (Figure 8) and number of rotatable bonds with total score (Figure 9). In both bases we also have coloured by clogP. A trend in both sites was observed whereby higher scores were given for larger and more flexible compounds with a higher degree of lipophilicity.

The consensus between six algorithms of Surflex-Dock provides an additional layer of interpretation of the results. Total scores and ranking should only be assessed alongside a visual inspection of the docked poses where predicted binding interactions can be checked for validity, comparison to those that may be important for native ligands or existing inhibitors and also if the volume of the grid is sufficient to encapsulate the binding pocket. The Cscores are set out in Supplemental Table S2 where ligands can be assessed on the basis of consensus and compared to their ranking. Interestingly few of the ligand total scores showed a high level of consensus. Interestingly the triazole S enantiomer produced a Cscore of 6 (highest). This compound was shown to score highly in the mutant site. Note that for the Surflex-Dock analysis only the top scoring poses within each cluster was selected.

Further quantitative structure-activity relationship (QSAR) and CoMFA analyzes can be performed within the Surflex-Dock GUI if biological data is available.

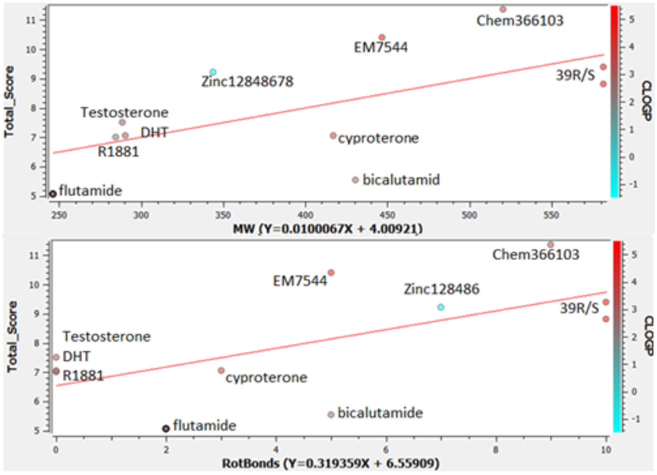


Figure 8: MD results from Surflex-Dock for wild type hAR site (PDB ID: 2PNU) in flexible mode. Partial least squares correlation of molecular weight (r^2=0.383) and rotatable bonds (r^2=0.412) with total score and coloured by clogP.

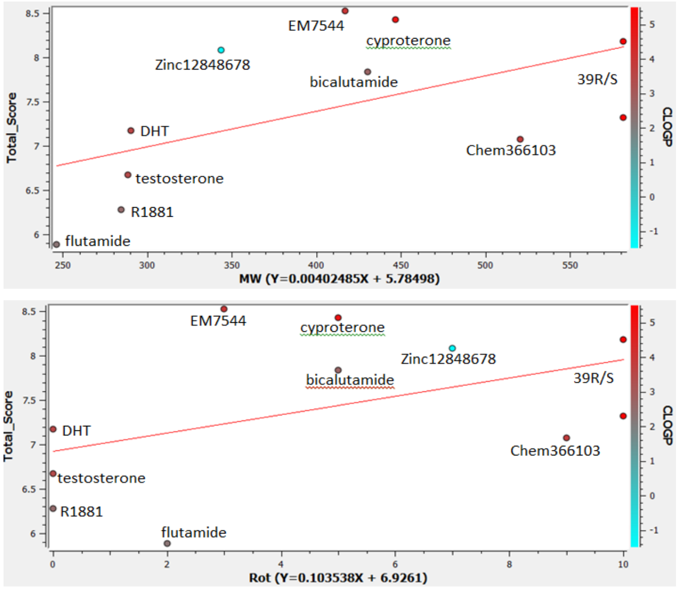


Figure 9: MD results from Surflex-Dock for mutant hAR site (PDB ID: 2OZ7) in flexible mode. Partial least squares correlation of molecular weight (r^2=0.307) and rotatable bonds (r^2=0.214) with total score and coloured by clogP.

Interestingly, the decoy ligand, ZINC12848678, scored second position after the X-ray ligand in the mutant hAR site. This relatively small, charged ligand appears to prefer the larger site created by the mutation and can take advantage of hydrogen bonding groups around the site providing plenty of possible polar interactions for ligand stabilization.

Figure 10 shows the orientation of testosterone in the wild type hAR site overlaid with the docked position of antiandrogen, bicalutamide, and also a triazole non-steroidal hAR inhibitor, compound 39. The lipophilic moieties are closely aligned in the steroid binding site. The more polar ends of the non-steroidal inhibitors orient towards a narrow channel, the end of which is capped by a key □ helix with a role in receptor activation.

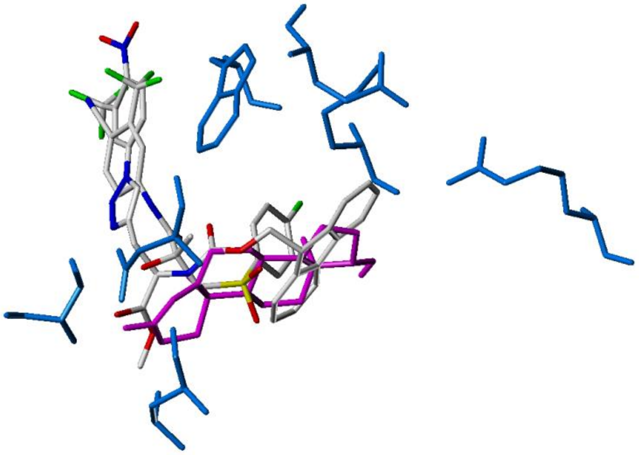


Figure 10: Superimposition of testosterone (magenta), bicalutamide (atom types) and triazole compound 39 (atom types with FMoc group)

Where available, biological activity data for relevant ligands was collected and set out in supplementary Table S1. This data, along with physical and chemical properties of the ligands, can be further used in a quantitative structure-activity relationship (QSAR) analysis – a useful technique in design optimization. It must be noted that a robust docking analysis would often incorporate the use of a large training dataset of actives and decoys which would then be further optimized through ligand enrichment and visualization of cluster orientations and ranking prior to use with a test dataset.

A limitation of this study was the small ligand database. This was specifically made small for illustrative purposes. Furthermore a thorough docking analysis of the hAR would take advantage of the entire available dataset of active and decoys for the hAR from which incorporating a ligand enrichment protocol would facilitate a higher likelihood of the algorithm’s capacity to select high affinity ligands.

3.2 MD program comparison

The docking case study presented here provided a means upon which to compare the usefulness of the three selected programs. To further facilitate choice of MD program we have summarized from key features which may be of assistance in the selection of a suitable MD program (Table 8). While the results from our MD case study showed variability in the metrics used to compare the various programs, produce a clear winner in terms of reliability of results. Benchmarking between the most commonly used algorithms, such as that used by AD4 and Surflex-Dock, has shown a good level of performance. From our results we found both AutoDock and Surflex-Dock produce high quality results. Our results showed AutoDock to be superior for active pose prediction while Surflex-Dock was more consistently able to predict X-ray ligand orientation. In overall usability however, especially for multiple ligand docking, we found Surflex-Dock to be most useful.

Table 8: Summary of selected MD programs.

Feature	**SwissDock**	**AD4**	**Surflex-Dock**
Availability	Freeware	Freeware	Commercially available as a module of SYBYLx2.1 suite
Beginner-freindly	Easy to Moderate	Moderate	Moderate
Min. Hardware Requirement	desktop PC.	Intel i86 (32-bit), x86_64 (64-bit) processor.	Intel i86 (32-bit), x86_64 (64-bit)/ PowerPC processor.
Operating System	Server/cross platform.	Cross platform	Cross platform.
Applicability	Single or multi target	Robust and accurate single target/single ligand (option for multiple ligand)	Robust and accurate single target– multiple ligand modelling.
ForceField	CHARMM	CHARMM	AMBER (a range of options are provided).
Advantages	Intuitive online server, widely utilised in literature.	Freeware, robust, gold standard. Benchmarking shows a good level of docking accuracy.	Good level of docking accuracy, short run times, handles large ligand libraries, GUI, high level of parameter control and support.
Limitations	Server-based. Min. parameter control, support via help forum.	Long run times, max. 32 torsion angles, limited support via forum. Separate GUI for virtual screening.	Commercially available. Less commonly used in Industry for comparative purposes.

3.3 Homology Model Results

Residues in the alignment can be selected interactively which automatically identifies the 3D position in the model. In this case, all three gaps were found to correspond to loops on the surface of the extracellular domain of the receptor rather than core α-sheets or β-helices regions thus the model could reasonably be considered for further molecular modelling projects.

A superimposition depicting ribbons through the backbone atoms of template and target is shown in Figure 11. Despite the relatively low sequence identity between the template and target, the rmsd between the backbone of the proteins was remarkably good. The process can be repeated with a number of the high scoring matches. These may be compared to determine the degree of general agreement in 3D structure. Once satisfied, the best homology model could be utilized for subsequent MD, molecular dynamics or other receptor-based analyzes.

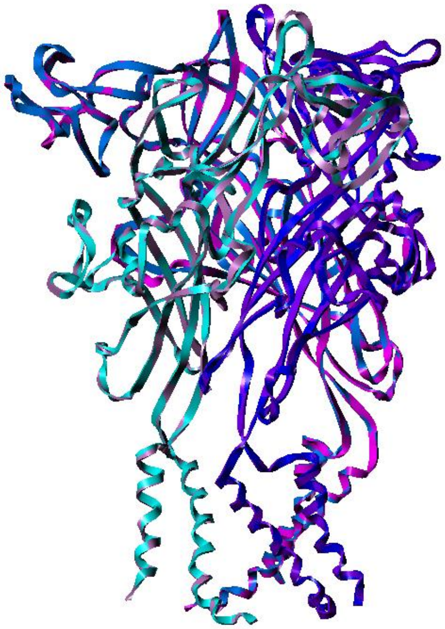


Figure 11: Superimposition of the template crystal structure of human P2X4 receptor (PDB ID: 4DW1) with the final homology model using the sequence of human P2X1.

4. Conclusion

This review aimed to serve as a practical guide to contemporary molecular docking by providing a detailed, step-by-step workflow for docking a set of ligands into a target. By selecting three MD programs covering a range of user preferences, we were also able to illustrate the necessity for validation and benchmarking in MD. Additionally, where the specific 3D target structure is unavailable, we guide the reader through a homology modelling case study to facilitate generation of a model structure upon which subsequent MD may be applied. We hope this tutorial review may be of assistance to those interested in incorporating these in silico techniques into lead identification strategies.

5. Conflict of Interest

The authors declare there is no conflict of interest.

6. Acknowledgements

The authors wish to acknowledge the valuable advice of Dr. Luke Henderson in the preparation of this manuscript.

7. References

1. DiMasi, J.A., R.W. Hansen, and H.G. Grabowski, *The price of innovation: new estimates of drug development costs.* J Health Econ, 2003. **22**(2): p. 151-85.
2. Blundell, T.L., *Structure-based drug design.* Nature, 1996. **384**(6604 Suppl): p. 23-27.
3. Stjernschantz, E.O., Chris, *Improved Ligand-Protein Binding Affinity Predictions Using Multiple Binding Modes.* Biophys J., 2010. **98**(1): p. 2682–2691.
4. Khalili-Araghi, F., Gumbart, James, Wen, Po-Chao, Sotomayor, Marcos. Tajkhorshid, Emad, Schulten,Klaus *Molecular dynamics simulations of membrane channels and transporters.* Curr Opin Struct Biol. , 2009. **19**(2).
5. Maffeo, C., Bhattacharya, Swati, Yoo, Jejoong, Wells, David, Aksimentiev,Aleksei, *Modeling and Simulation of Ion Channels.* Chem. Rev., 2012. **112**(12): p. 6250–6284.
6. Wereszczynski J.1, McCammon J.A., *Statistical mechanics and molecular dynamics in evaluating thermodynamic properties of biomolecular recognition.* Q Rev Biophys., 2012. **45**(1): p. 1-25.
7. Ross, G.A., Morris, G.M., Biggin,P. C., *One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery.* J. Chem. Theory Comput., 2013. **9**: p. 4266–4274
8. Matteo Aldeghi, A.H., Michael J. Bodkin, Stefan Knapp, Philip C. Biggin, *Accurate calculation of the absolute free energy of binding for drug molecules.* Chem. Sci., 2016. **7**: p. 207-218.
9. Kleinjung, J. and F. Fraternali, *Design and application of implicit solvent models in biomolecular simulations.* Curr Opin Struct Biol, 2014. **25**: p. 126-34.
10. Zhang, H., T. Tan, and D. van der Spoel, *Generalized Born and Explicit Solvent Models for Free Energy Calculations in Organic Solvents: Cyclodextrin Dimerization.* J Chem Theory Comput, 2015. **11**(11): p. 5103-13.
11. Gleeson, M.P. and D. Gleeson, *QM/MM Calculations in Drug Discovery: A Useful Method for Studying Binding Phenomena?* Journal of Chemical Information and Modeling, 2009. **49**(3): p. 670-677.
12. Anuradha Roy, P.R.M., Sitta Sittampalam, Rathnam Chaguturu, *Open Access High Throughput Drug Discovery in the Public Domain: A Mount Everest in the Making.* Curr Pharm Biotechnol., 2010. **11**(7): p. 764–778.
13. Charifson, P.S., Corkery, J., Murcko, M.A., Walters, W.P., *Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins.* J Med Chem., 1999. **42**(25): p. 5100-9.
14. Halperin, I., Ma, Buyong, Wolfson, Haim, Nussinov,Ruth *Principles of Docking: An Overviewof Search Algorithms and a Guide to Scoring Functions.* PROTEINS: Structure, Function, and Genetics, 2002. **47**: p. 409–443.
15. Steinbrecher, T., Labahn, Andreas *Towards Accurate Free Energy Calculations in Ligand Protein-Binding Studies.* Current Medicinal Chemistry, 2010. **17**: p. 767-785.
16. Warren GL, A.C., Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS., *A Critical Assessment of Docking Programs and Scoring Functions.* J Med Chem., 2006. **49**.
17. Philippe Ferrara, H.G., Daniel J. Price, Gerhard Klebe, Charles L. Brooks, *Assessing scoring functions for protein-ligand interactions.* J. Med. Chem., 2004. **47**.
18. Neves, M.A.C., *Docking and scoring with ICM: the benchmarking results and strategies for improvement.* J Comput Aided Mol Des., 2006. **26**(6): p. 675–686.
19. Morris, G.M., Huey, R., Lindstrom, W., Sanner, M. F., Belew, R. K., Goodsell, D. S. and Olson, A. J., *Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity.* J. Computational Chemistry, 2009. **16**: p. 2785-91.
20. Grosdidier, A., V. Zoete, and O. Michielin, *SwissDock, a protein-small molecule docking web service based on EADock DSS.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W270-7.
21. Grosdidier A1, Z.V., Michielin O., *Fast docking using the CHARMM force field with EADock DSS.* J Comput Chem., 2011. **32**(10): p. 2149-59.
22. Russell Spitzer, A.N.J., *Surflex-Dock: Docking Benchmarks and Real-World Application.* J Comput Aided Mol Des., 2012. **26**(2): p. 687–699.
23. Martin Karplus, A.M., Nature Structural Biology 9, 646 - 652 (2002) 2002. **9**: p. 646 - 652.
24. Ou-Yang, S.S., Lu, J., KONG, X., Liang, Z. Luo, C., Jiang, H., *Review - Computational drug discovery.* Acta Pharmacologica Sinica, 2012. **33**: p. 1131–1140.
25. Yuriev, E., *Challenges and advances in structure-based virtual screening.* Future Medicinal Chemistry, 2014. **6**(1): p. 5-7.
26. Trott, O. and A.J. Olson, *Software News and Update AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading.* Journal of Computational Chemistry, 2010. **31**(2): p. 455-461.
27. Goodsell, D.S. and A.J. Olson, *Automated Docking of Substrates to Proteins by Simulated Annealing.* Proteins-Structure Function and Genetics, 1990. **8**(3): p. 195-202.
28. International, T., *SYBYL-X 2.1.1.* 2014. p. Molecular Modelling Suite.
29. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.
30. Salon, J.A., Lodowski, D.T., Palczewski, K., *The Significance of G Protein-Coupled Receptor Crystallography for Drug Discovery.* Pharmacol Rev. , 2011. **63**(4): p. 901-937.
31. Sen, S., et al., *Small molecule annotation for the Protein Data Bank.* Database (Oxford), 2014. **2014**: p. bau116.

32. Lütteke, T., von der Lieth, C.W., *pdb-care (PDB CArbohydrate REsidue check): a program to support annotation of complex carbohydrate structures in PDB files.* BMC Bioinformatics, 2004. **5**(69).
33. Sanner, M.F., *Python: a programming language for software integration and development.* J Mol Graph Model, 1999. **17**(1): p. 57-61.
34. Sanner, M.F., A.J. Olson, and J.C. Spehner, *Reduced surface: an efficient way to compute molecular surfaces.* Biopolymers, 1996. **38**(3): p. 305-20.
35. Dallakyan, S. and A.J. Olson, *Small-molecule library screening by docking with PyRx.* Methods Mol Biol, 2015. **1263**: p. 243-50.
36. Russell Spitzer, A.N.J., *Surflex-Dock: Docking Benchmarks and Real-World Application.* J Comput Aided Mol Des., 2012. **26**(6): p. 687–699.
37. Lua, R.C., Wilson, S.J., Konecki, D.M., Wilkins, A.D., Venner, E., Morgan, D.H., Lichtarge, O., *UET: a database of evolutionarily-predicted functional determinants of protein sequences that cluster as functional sites in protein structures.* Nucleic Acids Research, 2016. **44**: p. D308–D312.
38. Schwede, T., *Protein Modelling: What Happened to the "Protein Structure Gap"?* Structure, 2013. **21**(9).
39. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T.,Tramontano, A., *Critical assessment of methods of protein structure prediction (CASP) — round x.* Proteins:, 2014. **82**(0 2): p. 1–6.
40. Frenkela, Z.M., Frenkela, Z.M., Trifonova, E.N. Snir, S., *Structural relatedness via flow networks in protein sequence space.* Journal of Theoretical Biology, 2008. **260**(3): p. 438–444.
41. di Luccio1 E. Koehl, P.K., *A quality metric for homology modeling: the H-factor.* BMC Bioinformatics 2011, 12:48, 2011. **12**(48).
42. Xiang, X., *Advances in Homology Protein Structure Modeling.* Curr. Protein Pept. Sci., 2006. **7**(4): p. 217–227.
43. Chothia, C., Lesk, A.M., *The relation between the divergence of sequence and structure in proteins.* EMBO J., 1986. **5**(4): p. 823–826.
44. Venselaar, H., Joosten, R.P., Vroling, B., Baakman, C.A. B., Hekkelman, M.L., Krieger, E., Vriend, G., *Homology modelling and spectroscopy, a never-ending love story.* Eur Biophys J., 2010. **39**(4): p. 551–563.
45. Fong, J.H., Marchler-Bauer, A., *Protein subfamily assignment using the Conserved Domain Database.* BMC Research Notes, 2008. **1**(114).
46. Marco Biasini, S.B., Andrew Waterhouse, Konstantin Arnold, Gabriel Studer, Tobias Schmidt, Florian Kiefer, Tiziano Gallo Cassarino, Martino Bertoni, Lorenza Bordoli, Torsten Schwede, *SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information.* Nucleic Acids Research, 2014. **42** (Web server Issue): p. W252-W258.
47. Arnold, K.B., L., Kopp, J.,Schwede, T., *The SWISS-MODEL Workspace: A web-based environment for protein structure homology modelling.* Bioinformatics, 2006. **22**: p. 195-201.
48. Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J. and Schwede, T., *Protein structure homology modelling using SWISS-MODEL Workspace.* Nature Protocols, 2009. **4**: p. 1-13.
49. Benkert, P., S.C. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins, 2008. **71**(1): p. 261-77.
50. Altimari, J.M., Niranjan, Birunthi, Risbridger, Gail P., Schweiker,Stephanie S., Lohning,Anna E., Henderson, Luke C., *Preliminary investigations into triazole derived androgen receptor antagonists.* Bioorganic & Medicinal Chemistry, 2014. **22**(9): p. 2692-2706.
51. Huang, N., B.K. Shoichet, and J.J. Irwin, *Benchmarking sets for molecular docking.* J Med Chem, 2006. **49**(23): p. 6789-801.
52. Herraez, A., *Biomolecules in the Computer Jmol TO THE RESCUE.* BIOCHEMISTRY AND MOLECULAR BIOLOGY EDUCATION, 2006. **34**(4): p. 255-261.
53. Pettersen EF1, G.T., Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE., *UCSF Chimera--a visualization system for exploratory research and analysis.* J Comput Chem., 2004. **25**(13): p. 1605-12.
54. Jain, A.N., *Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition.* J Comput Aided Mol Des, 2000. **14**(2): p. 199-213.
55. Kellenberger, E., Rodrigo, Jordi, Muller, Pascal, Rognan, Didier, *Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy.* PROTEINS: Structure, Function, and Bioinformatics, 2004. **57**: p. 225–242.
56. Spitzer, R., Jain, Ajay N., *Surflex-Dock: Docking Benchmarks and Real-World Application.* J Comput Aided Mol Des., 2012. **26**(6): p. 687–699.
57. National Centre for Biotechnology Informtation http://www.ncbi.nlm.nih.gov/ (Accessed Nov.21.
58. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J. , Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., *Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.* . Molecular Systems Biology, 2011. **7**: p. 1-6.
59. Notredame, C., Higgins, D.G., Heringa, J., *T-Coffee: A novel method for fast and accurate multiple sequence alignment.* . Journal of Molecular Biology, 2000. **302**(1): p. 205-217.
60. *Swiss-Model Homology Modelling Server*. [cited 2015 Dec.13]; Available from: http://swissmodel.expasy.org/.

61. Irwin, J.J. and B.K. Shoichet, *ZINC - A free database of commercially available compounds for virtual screening.* Journal of Chemical Information and Modeling, 2005. **45**(1): p. 177-182.

62. Eisermann, K., Wang, Dan, Jing, Yifeng, Pascal, Laura E.,Wang, Zhou, *Androgen receptor gene mutation, rearrangement, polymorphism.* Transl Androl Urol., 2013. **2**(3): p. 137–147.