Fitting small molecules to cryo-electron microscopy data

Fitting small molecules to cryo-electron microscopy data

Aaron Sweeney

A thesis submitted for the degree of Doctor of Philosophy

The Institute of Structural and Molecular Biology,
Birkbeck college, University of London,
United Kingdom

Declaration

I, Aaron Sweeney, declare that this thesis describes my own work. Suitable citations have been provided where information that has been obtained from other sources. Additionally, work conducted in collaboration with others has been indicated.

Signed

……

Date

……

Acknowledgements

Firstly, I would like to thank my supervisor Prof Maya Topf for all her help, support and guidance throughout my project. Your support enabled me to develop as a scientist and complete this thesis. Additionally, I would like to thank my second supervisor Prof Carolyn Moores for all her help and advice relating to the kinesins.

A huge thank you is extended to all the members of the Topf lab group past and present. Including, Dr Joe Newcombe who introduced me to the concepts of small molecule docking, Dr Tristan Cragnolini, and Dr Mauro Maiorca whose advice was invaluable for the completion of my thesis. I'd also like to thank the collaborators that enabled this thesis including Dr Nigel Unwin and Dr Alejandro Peña for supplying cryo-EM maps and advice.

I would like to extend a big thank you to the computer support team at Birkbeck, in particular Dr David Houldershaw who helped me numerous times with installing software and facilitating my experiments on the Birkbeck servers. I would also like to thank Charise Rawlinson for always offering to help me when I had no idea what I was doing regarding administration matters.

I would like to thank my friends in particular Shawn for helping me reset and recharge when my project was not going my way. Finally, I would like to thank my family including my beautiful wife for putting up with me during this project, especially during the last few months of writing. My sister for listening when I needed to rant and my mum, dad and step-dad for always believing in me and for enabling me to get an education.

Abstract

Recent innovations in the field of cryogenic-electron microscopy (cryo-EM) has enabled the visualisation of biological systems at atomic resolutions that rival that of X-ray crystallography. This is increasing the relevance of cryo-EM in the field of drug discovery, as it is now possible to solve high-resolution structures of biological complexes that may not have been amenable to crystallisation [1] and also in a more "native-like" state. However, it is not always possible to obtain structures to atomic resolutions with cryo-EM, currently only 16.28 % of structures deposited in the electron microscopy database [2] are at resolutions better than 3.0 Å, with the majority (45.05 %) at resolutions between 3.0 and 4.0 Å (*correct as of December 2021)*. A vast body of work has been conducted with the aim of fitting biological macromolecules into cryo-EM at various resolutions [3–6]. However significantly less has been reported regarding the fitting of small molecules into cryo-EM maps. The work presented in this thesis aimed at developing methodologies that enable the fitting of small molecules to cryo-EM maps at resolutions from near atomic to 4.5Å.

First, I used a fitting methodology that utilised consensus docking [7] in conjugation with a local difference mapping technique [8] to model the complex of the Eg5 kinesin motor domain with a novel inhibitor (GSK-1) in the presence of tubulin, into a 3.8 Å cryo-EM map (Chapter 2). The arrangement of structural elements within the protein allowed inferences to be made as to the mechanism of action of the drug [9].

Next, I present a new empirical molecular docking score for identifying correct ligand conformations within protein ligand complexes (Chapter 4). This score was integrated with goodness-of-fit scores commonly used for assessing the fit of biological molecules to cryo-EM maps [10]. Furthermore, we assessed the utility of this integrated score for fitting small molecules using simulated full maps and density difference maps (Chapter 4). This integrated score was then developed into a full methodology for fitting small molecules into cryo-EM maps, where its effectiveness was evaluated with experimental data at high (≤ 3.0 Å) and low (3.0 to 4.5 Å) resolution (Chapter 5).

The accurate identification of protein ligand interactions from atomic models is an important consideration for drug discovery. To this end, a new software is presented that predicts protein ligand interactions using geometric parameters (Chapter 3). This software was benchmarked using 35 high resolution protein-ligand complexes and compared to current state-of-the-art available software [11, 12].

Finally, I present the refined protein model of a *Torpedo* nicotinic acetylcholine receptor including the MX helix in a 6.6 Å cryo-EM map (Chapter 6). A combination of fitting software and bioinformatics identified the position of the MX helix relative to the cellular membrane. Our investigation suggested that the MX may function to entrap cholesterol, imposing rigidity to the receptor around the narrowest point of the central pore.

Publications

Peña, Alejandro, **Aaron Sweeney**, Alexander D. Cook, Julia Locke, Maya Topf, and Carolyn A. Moores. 2020. “Structure of Microtubule-Trapped Human Kinesin-5 and Its Mechanism of Inhibition Revealed Using Cryoelectron Microscopy.” *Structure* 28 (4): 450–57.e5.

Cragnolini, Tristan*, **Aaron Sweeney ***, and Maya Topf. 2021. “Automated Modeling and Validation of Protein Complexes in Cryo-EM Maps.” *Methods in Molecular Biology* 2215: 189–223.

Cragnolini T, Sahota H, Joseph AP, **Sweeney A**, Malhotra S, Vasishtan D, Topf M. 2021 TEMPy2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. Acta Crystallogr D Struct Biol. 77(Pt 1):41-47.

** joint first author*

Table of contents

List of figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

List of Tables

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Table 3. Sequence identities between Torpedo subunits and corresponding T. californica subunits.
Table 4. Average SMOC score for each subunit at the different stages of refinement

Abbreviations

2D – 2-dimensional
3D – 3-dimensional
5-HT3 – 5-Hydroxytryptamine
6D – 6-dimensional
ADP – Adenosine Diphosphate
ADP-AlFX – Adenosine Diphosphate/Aluminium fluoride complex
AMPPNP – Adenylyl-imidodiphosphate
ATP – Adenosine triphosphate
Br – Bromine
Ca – Calcium
Cα-RMSD – α-Carbon Root Mean Square deviation
CαBLAM –C-Alpha Based Low-resolution Annotation Method
CASF – Comparative assessment of scoring functions
CCC – Cross correlation coefficient
CDMD – Correlation driven molecular dynamics
Cl – Chlorine
Cryo-EM – Cryogenic electron microscopy
CTD – C-terminal domain
DA – Donor-acceptor
DHA – Donor-hydrogen-acceptor
DNA – Deoxyribonucleic acid
DOPE – Discrete optimised protein energy
EM – Electron microscopy
EMDB – Electron microscopy data bank
Fe – Iron
FN – False negative
FP – False positive
FSC – Fourier shell correlation
GA – Genetic algorithm
GABA – Gamma-aminobutyric acid
H-Bonds – Hydrogen bonds
HeLa – Henrietta Lacks (immortal cell line)
I – Iodine
IC_{50} – Inhibitory concentration 50
KSP – Kinesin spindle protein
LMB – Laboratory of molecular biology
MC – Monte-Carlo
MD – Molecular dynamics
MDFF – Molecular dynamics flexible fitting
Mg – Magnesium

MI – Mutual information
MICRO-ED – Micro-crystal electron diffraction
Mn – Manganese
MOE – Molecular optimisation environment
mRNA – Messenger Ribonucleic acid
MT – Microtubule
MT-GMPCPP – Microtubule-Guanosine-5'-[(α,β)-methyleno]triphosphate complex
NAChR – Nicotinic acetylcholine receptor
NMR – Nuclear magnetic resonance
NNP – Neural network potential
NTD – N-terminal domain
PDB – Protein data bank
Pi – Phosphate ion
*p*Ki – negative log of the inhibitor constant (Ki)
PLIP – protein–ligand interaction profiler
ProPLID – Program for Protein-Ligand interaction detection
PSO – Particle swam optimisation
QM/MM-MD – Quantum mechanics/molecular mechanics-molecular dynamics
RMSD – Root mean square deviation
SBDD – Structure based drug discovery
SCCC – Segment-based Cross Correlation
SMOC – Segment based Manders' Overlap Coefficient
SSEs – Secondary structure elements
STD – Standard deviation
TEM – Transmission electron microscopy
TEU – Torsional energy units
TM – Transmembrane
TP – True positive
vdW – van der Waals
Zn – Zinc

Chapter 1

Introduction

Proteins are the molecular machines necessary for life

Proteins are a class of biomolecules that facilitates nearly all of the processes that are necessary for life. The blueprints for these molecular machines are encoded in the DNA of an organism. Generally, production of proteins involves the DNA 'instructions' being transcribed into mRNA before being translated into proteins.

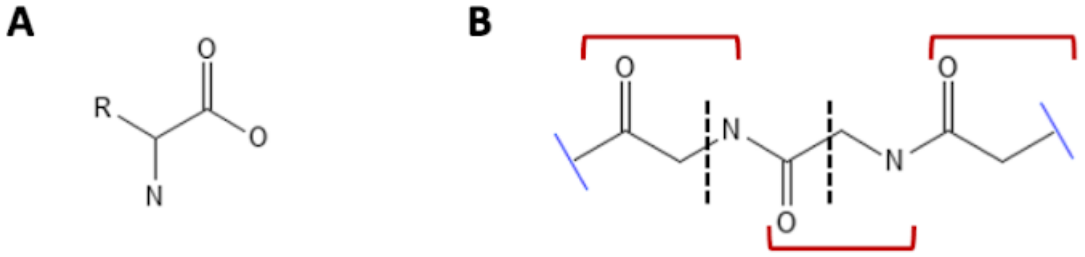


Figure 1. *(**A**) A chemical schematic representation of the common structure shared by amino acids, the R group represents where the amino acid sidechains are bonded to the basic structure. (**B**) A chain of three amino acids (glycine) linked by peptide bonds. The boundaries of the three amino acids are indicated (red). Additionally, a single peptide bonded between two residues is indicated (black dashed lines).*

The basic building blocks of proteins are amino acids. Each amino acid is composed of the same basic structure with an amino group and a carboxylic acid moiety (Figure 1). The 20 essential amino acids get their distinct molecular properties from differences in their organic sidechains (Figure 2). Proteins are composed of strings of amino acids chained together by amide bonds. Groups of amino acids residues within these chains form secondary structures. There are three general classes of protein secondary structures, helices, sheets, and loops (Figure 3), usually referred to as secondary structure elements (SSEs). Groups of secondary structures then come together and form domains of 3D structure that give the protein its overall fold. The protein folds are what allow a particular protein to carry out its function. Furthermore, groups of folded proteins can come together in a complex to carry out more complex tasks. Complexes of proteins interacting in this way are termed the protein quaternary structure.

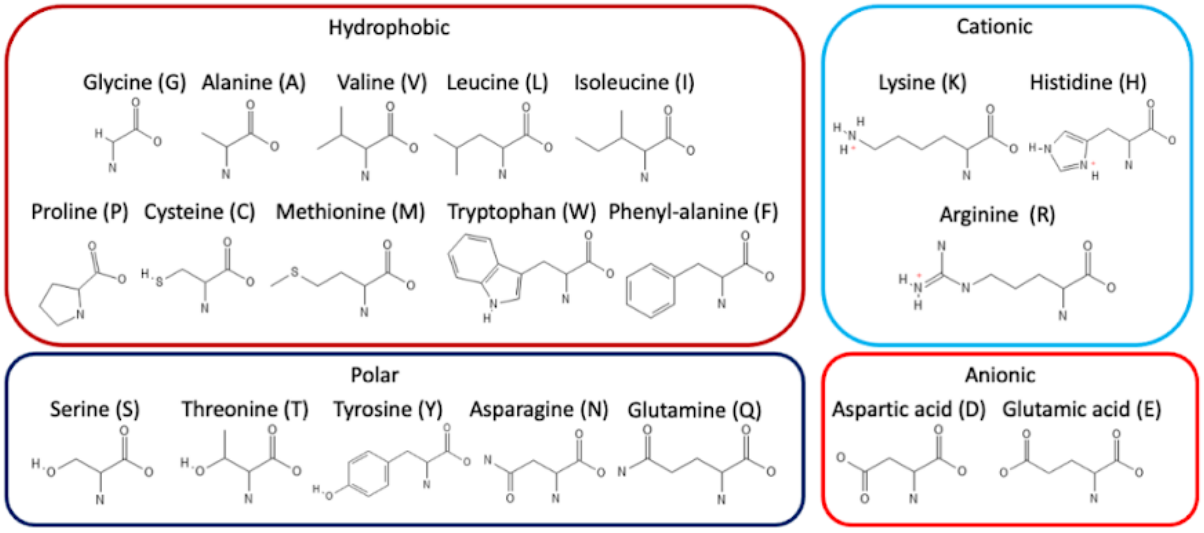


Figure 2. *A chemical schematic depiction of the 20 essential amino acids that proteins are formed from and their general chemical properties.*

Protein atomic models

The field of structural biology endeavours to study the structure of biological molecules, including proteins. One highly important resource for any structural biologist are the atomic models of such structures (Figure 3). These atomic models allow researchers to visually inspect the structure of biological complexes and facilitate further investigation. Much of this study can be done *in silico* and therefore methods of storing [13, 14] and visualising [15] atomic models computationally have been increasingly developed over the last two decades. One major resource that aids such study is the Protein Data Bank (PDB) [13, 14]. The PDB boasts an impressive collection of more than 184,700 (correct as of December 2021) atomistic models of biomolecular complexes. As this is such an important resource file formats have been developed to store this large collection of molecules including the '.pdb' and '.mmcif' file format. These files contain the information necessary to build atomic models, including atom positions and types.

Proteins targets for therapeutic intervention

As the proteins underpin most biological processes within an organism, those failing to complete their assigned tasks can lead to disastrous consequences for the organism, usually manifesting as some sort of pathology. There are many ways in which a protein may fail to fulfil its duties, one of the most common ways comes from mutations within the DNA that encodes a protein. These mutations propagate into the 3D-structure of the protein affecting the protein's function.

Another mechanism by which disease can arise is through the invasion of one organism by another. In humans this can occur by the body becoming infected with microorganisms such

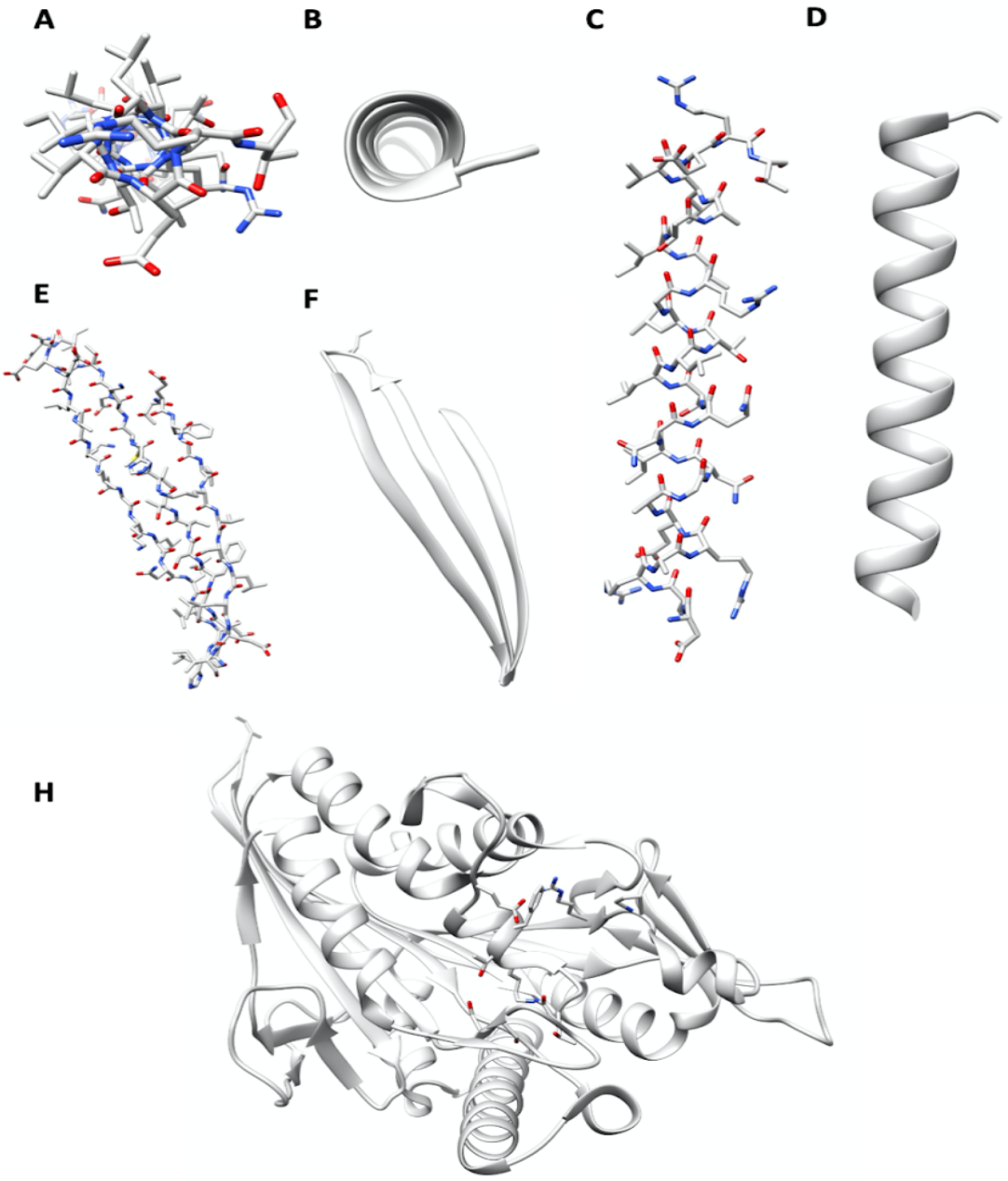


Figure 3. *An example of atomic models of secondary structural elements. A side on view of a typical a-helix model, all atoms and bonds between them is shown (**A**) along with a 3D schematic representation (**B**). The same is shown from a side-on view for all atoms (**C**) and 3D schematic (**D**). An example of a typical β-sheet is shown with an all atom (**E**) and 3D schematic (**F**) model. (**H**) The full atomic model of a kinesin motor domain. Secondary structure elements are shown with a 3D schematic representation and the atoms of some residue sidechains can also be seen.*

as bacteria or viruses. For example the current COVID-19 pandemic is caused by the invasion of host cells by coronavirus particles (*reviewed in* [16]). The virus then hijacks the host cellular machinery to produce more viral particles, which are released into the

bloodstream to infect other cells. The viral entry to human cells is mediated by a viral protein called the spike glycoprotein (*reviewed in* [17]).

The importance of proteins in the fundamental mechanisms of disease, has made them the target of therapeutic intervention by pharmaceutical companies and academic researchers worldwide. One branch of pharmacology to treat illness involves using small molecule drugs to perturb protein function. Small molecules (or ligands) are relatively simple chemical compounds that bind to proteins and can exert effect over the protein function. Small molecules are used natively by organisms as a method to facilitate the natural function of biomolecules. One such molecule being adenosine triphosphate (ATP), the "molecular currency" of life. These molecules bind to proteins and the energy released, through the hydrolysis of ATP to adenosine diphosphate (ADP), can be used by proteins to fulfil their functions. One example being the hydrolysis of ATP by the kinesin motor domain proteins that has been shown to drive the 'walking' of kinesins along microtubules [18, 19].

Man made therapeutics are often chemically optimised derivatives of molecules found in nature. Generally speaking, such drugs will bind to proteins to exert their effects. This binding can occur *othosterically* (i.e. at sites where naturally occurring molecules bind), competing with naturally occurring molecules for binding site occupancy, or *allosterically*, binding at a site distinct from that of natural products.

The binding of drugs to proteins can bring about different effects. *Agonists* will bring about a functional response in proteins, whilst *antagonists* will result in a reduction of biological response upon binding. Additionally, drugs can be classified as "inverse *agonists*". These drugs bind to proteins reducing any constitutive receptor activity. In addition to the type of effect a drug can produce, small molecule drugs can be classified as partial or full, defined by the effect they bring about. For example a "full *agonist*" will have the ability to elicit the maximal biological response the system can bring about, whilst a "partial agonist" will only have the ability to bring about a submaximal biological response.

Furthermore, drugs can be classified as *competitive* or *non-competitive* based on their mechanism of action. Competitive drugs bind *orthosterically* and compete with native small molecule ligands, whilst non-competitive drugs bind *allosterically*.

Molecular recognition by small molecules

Regardless of a drug's classification or the effects it brings about, a biomolecule needs a mechanism of recognising small molecules. This molecular recognition is achieved by interactions between proteins and small molecules. Mostly these interactions are non-covalent in nature. A small minority of drugs can bind molecules covalently for example penicillin, however, these are outside of the scope of this work. The following section gives an overview of common non-covalent bond types involved in protein ligand interactions.

Van der Waals and hydrophobic interactions

The van der Waals (vdW) shell of an atom is derived from a theoretical sphere that approximates the space an atom occupies, with the vdW radius defined as the distance from the centre to the circumference of the sphere. Electrons are in constant motion around atomic nuclei and are not distributed evenly around the atom. This uneven distribution results in dipoles (local regions of opposing charge) around atomic nuclei. These di-poles can be induced by neighbouring atoms or occur instantaneously via the stochastic distribution of electrons about an atom [20]. vdW interactions can broadly be described as the interactions between atomic dipoles of neighbouring atoms. However, at distances where vdW shells of two non-bonded atoms begin to overlap, the forces felt by atoms become repulsive in nature.

A mathematical model for estimating the attractive and repulsive forces felt by two atoms has been proposed [21]. However, for static models, vdW interactions may be estimated using the distance between two atomic centres. A single vdW interaction is relatively low energy, however the energy of such interactions are taken to be summative and many interactions can have a profound effect on stabilising protein ligand interactions.

Hydrophobic interactions are a related, but distinct phenomenon, driven by entropy. Certain hydrophobic atoms, such as carbon, in water will group together with other hydrophobic atoms to the exclusion of water. Water molecules then form a hydration shell around groups of hydrophobic atoms. In terms of protein-ligand interactions hydrophobic interactions occur when groups of hydrophobic atoms from the protein and small molecule ligand group together. Water then forms a larger structure around both groups of atoms, where the surface area is smaller than the sum of that formed by both groups of atoms individually. Hydrophobic interactions are critical in stabilising protein-ligand interactions and have been shown to correlate well with small molecule binding affinities [22].

Hydrogen bonds

The term hydrogen bond describes the formation of a ‘bridge’ between a lone pair of electrons from an electronegative ‘acceptor’ atom and a ‘donor’ hydrogen bonded to another electronegative atom (Figure 4). Such bonds are instrumental in the formation of SSEs such as α-helices and β-sheets within proteins. Hydrogen bonds can be described geometrically, by the distance between the hydrogen bond acceptor atom and donor hydrogen, along with the angle about the donor atom, donor atom hydrogen, and acceptor atom [23]. The energy involved in a hydrogen bond is higher than that of a single vdW interaction [24], with donor hydrogen-acceptor distances frequently seen at distances with overlapping vdW radii [23, 25]. However, the energy of a hydrogen bond is highly dependent on the atom types involved and the geometry of the bond. Furthermore, for a hydrogen bond to form, both donor and acceptor must first be desolvated. This process nearly cancels out the energy of formation of hydrogen bonding, indicating that hydrogen bonds may play a more important role in conferring specificity to a molecule for its protein binding site.

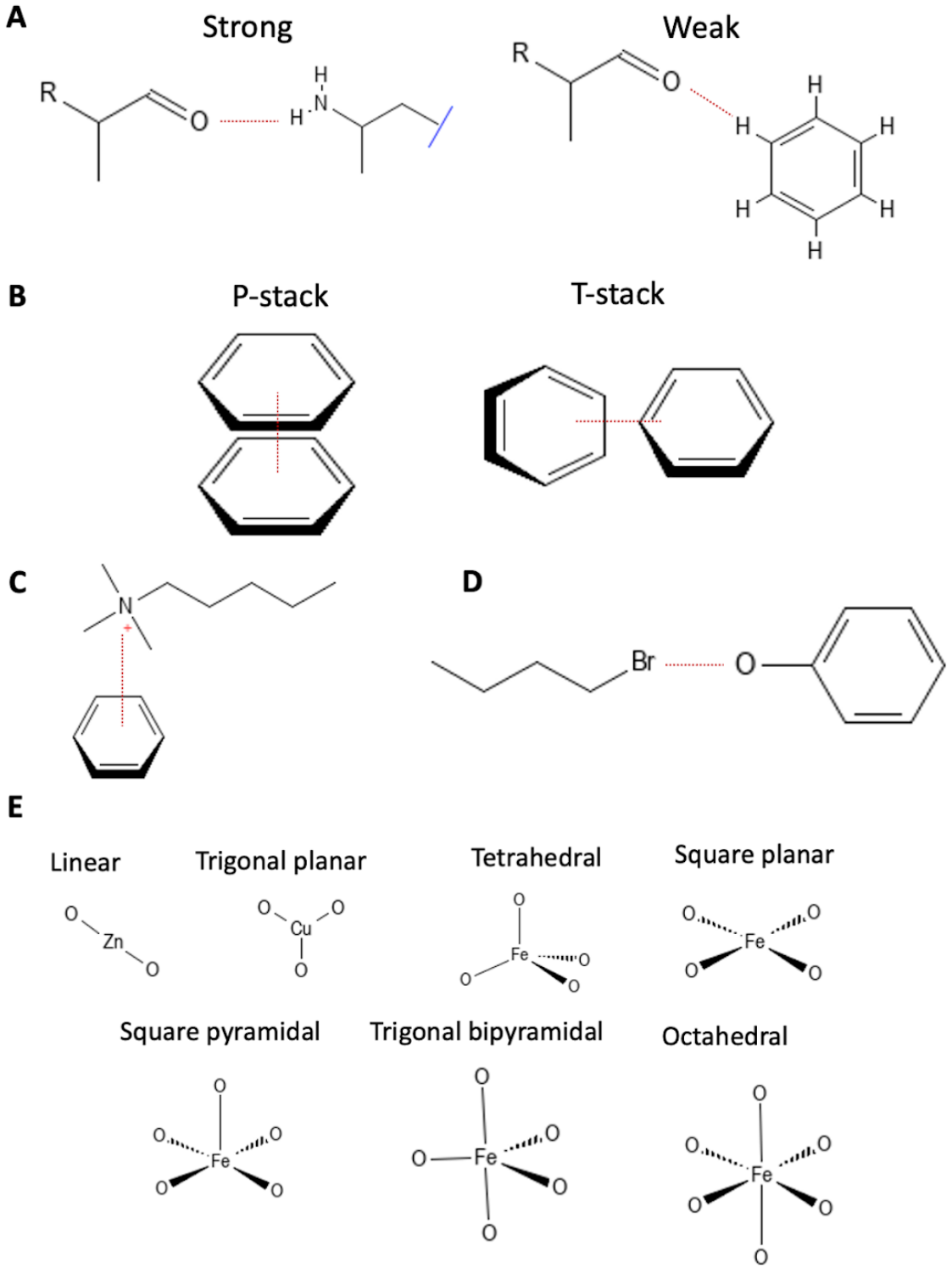


Figure 4. *Chemical schematics illustrating some common non-covalent interactions.* ***(A)*** *Schematics of strong and weak hydrogen bonds, the bond between acceptor atoms and donor hydrogens is highlighted (red dashed line).* ***(B)*** *Examples of* π-π *stacks, both P- and T-stack interactions are shown.* ***(C)*** *A schematic of a cation-*π *interaction the interaction of the cation with the delocalised* π *electron system is highlighted (red dashed line).* ***(D)*** *An example of a halogen bond the interaction between the halogen and the electronegative 'acceptor' atom is indicated (red dashed line).* ***(E)*** *schematics of common metal ion complex geometry. Black*

lines between oxygen atoms and metal ions indicate dative covalent bonds. The plane that the atoms are in are represented with wider lines, a solid wider line indicates the atoms lay in the plane towards the reader. Dashed wider lines indicate atom positions are in a plane away from the reader.

The aforementioned hydrogen bonds usually occur between highly electronegative atoms such as nitrogen, oxygen or fluorine. More recently, hydrogen bonding has been observed between hydrogens attached to weakly electronegative atoms such as carbon and electronegative acceptors [23, 26] (Figure 4). These 'weak' hydrogen bond interactions can be described by the same geometric criteria as 'strong' hydrogen bonds. However, they frequently occur over larger distances, typically unable to overcome repulsive forces associated with vdW radii overlap [23, 26]. The weak hydrogen bonds have been shown to be important for ligand affinity in multiple studies [27, 28].

Halogen bonds

All of the halogen atoms with the exception of fluorine have the potential to form halogen bonds [29]. Halogens bonds are dependent on the formation of a local region of positive charge about the halogen atom, termed the σ-hole, formed by an anisotropic polarisation of halogen atoms bonded to carbon atoms [30]. This region can form a directional electrostatic bond with available lone pairs of hydrogen bond acceptor atoms (with the exclusion of fluorine) (Figure 4). The strength of halogen bonds increases with the molecular weight of the halogen atom and in general the halogen bond is weaker than typical strong hydrogen bonds [31].

The halogen bonds can generally be described by geometric features including the, halogen-acceptor atom distance and bond angles around halogen-, acceptor- and the atoms that flank these [25]. Furthermore, halogen bonding has been an important consideration in the design of many small molecule therapeutics [32]

Interactions with aromatic rings

Aromatic ring systems contain a delocalised region of π-electrons above and below the plane of the ring. These electron dense regions can carry a negative charge and interact with surrounding chemical moieties. One of the most prevalent protein-ligand interactions involving aromatic rings are the π-π stacking interactions (Figure 4). This interaction occurs when the electron rich regions of one aromatic ring interact with electron deficient regions lining the ring. The geometry of such interactions can be generally described by the ring centre-ring centre distances and the relative plane angles of the rings [33]. Two favoured geometries of π-π stack predominate: the edge-to-face (or T-stack) orientation and the face-to-face (or P-stack) orientation (Figure 4). The energy of a typical π-π stack is approximately half that of the strong hydrogen bond ~2-3 kcal/mol [34], and such interactions have been utilised in the drug design process [35].

A second common interaction with aromatic rings is the cation-π interaction [36]. This interaction is electrostatic in nature and occurs between the π-electrons of an aromatic ring with a positively charged atom, such as an ammonium ion (Figure 4).

Metal ion complexes

Metal ion complexes refers to the dative covalent interactions formed between a positively charged metal ion and groups of electronegative atoms with available lone pairs (e.g. oxygen and nitrogen). Complexes can take on a distinct number of dative covalent bonds around the ion (Figure 4), referred to as the coordination number. Such complexes can be described by the geometry of angles of bonds that form a complex, and the ion-acceptor atom distances [37].

Methods for in silico drug discovery

The drug discovery pipeline (Figure 5) is the process by which pharmaceutical companies identify and develop new therapeutic agents. Once a target protein is identified there are four steps that are taken before a drug enters pre-clinical testing. Hit identification is the process of screening a large database of small molecule compounds for ones that exhibit the desired biological activity against the target. Once hits are identified, the hit-to-lead step identifies compounds that bind with high affinity, and exhibit good 'drug-like' properties. Drug-like properties dictate the bioavailability of a drug, properties such as molecular weight or log p (a measure of hydrophobicity) can determine how much of a drug will reach its target [38]. The last step involves optimising the lead compounds to improve their specificity, bioavailability and toxicity before entering clinical trials.

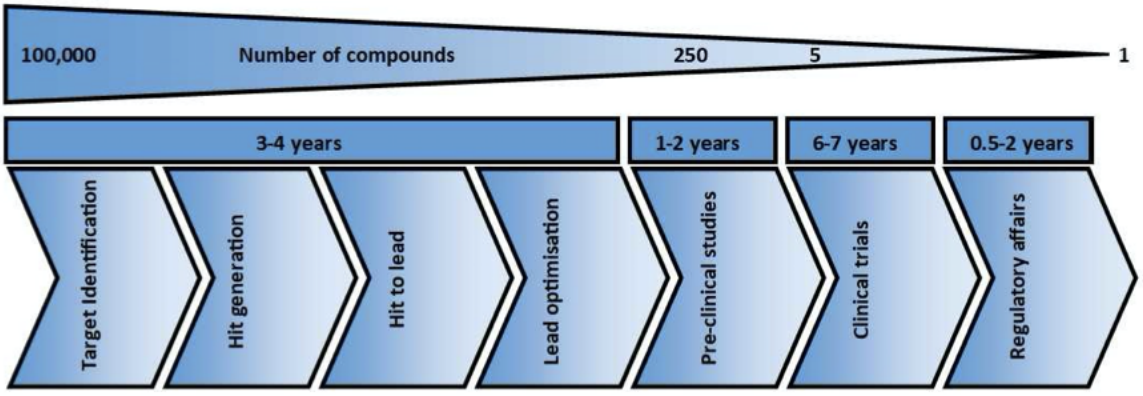


***Figure 5**. An overview of the drug discovery pipeline. Approximate timelines and number of compounds are shown for each stage.*

Classically, pharmaceutical companies have conducted the drug discovery pipeline with high throughput experimental screens, testing a large number of compounds with experimental assays. However, this is associated with a large cost and more companies are now turning to *in silico* methods to direct experimental methods.

The most common computational methods used are virtual high-throughput screening (Figure 6) ("virtual screening", also known as molecular docking) methods for hit identification. These methods utilise data from atomic models to computationally identify hits from a large library of compounds. Computational experiments typically rely on high-resolution (≤ 3.0 A) determination of atomic structure.

Binding site identification

In order to conduct molecular docking it is usual to first identify the binding site. There are two main challenges that the software must solve to predict a binding site. The first is a way to define and delineate a binding site. The second involves scoring predicted binding sites as more than one may be predicted.

The most accurate methods for determining sites involve using a representation of a 3D macromolecule. One subtype of these methods are the grid-based methods, which involve rendering of the protein in a 3-D grid and identifying points in the grid not included in the protein that satisfy certain conditions.

Software implementing these methods include POCKET [39] and $\text{Ligsite}_{\text{csc}}$ [40]. $\text{Ligsite}_{\text{csc}}$ renders a protein on a 3D grid and classifies potential binding site points as being within 3 Å of an atom point. Ray tracing is then used where points are encoded as a vector of X, Y, Z dimensions, and whether it appears within a protein or not. The idea is to identify vectors that start and end with protein points with 'solvent' (non-protein points) in between. Binding sites are built from vectors that satisfy these conditions.

Another method, Surfnet [41], delineates the protein by representing atoms by their vdW radii, then places probe spheres between at least two atoms, and reduces the radii until the sphere contains no atoms. Binding sites are built up from sets of probe spheres with a radius larger than a threshold radius.

CAST [42] and FPocket [43] are software that utilise alpha shapes (3D-euclidean curves representing a cluster of points). FPocket uses alpha spheres (the circumspheres of the tetrahedral in the alpha shape subset of the Delaunay tessellation) to build pockets by clustering the spheres using distance constraints.

A multitude of criteria exist to score and rank binding site predictions. Some are based in geometry such as the size of a binding site, as it has been assumed that ligand binding sites will occupy the largest volume. This was exemplified by a report showing that from 67 crystal structures of enzymes and ligands, 83.5 % were found bound to the largest binding site, 9 % to the second largest cavity and only 7.5% of ligands were found not bound to either the largest or second largest cavity [44].

However, exceptions exist to this for example allosteric ligands bind at sites distinct from the active site. To overcome this, methods such as LISE [45] introduce an energy-based scoring

function. This works by rolling a probe sphere around the surface of a molecule, the probe sphere represents a molecule or moiety with a feature such as hydrophobicity. Binding sites are composed of probe spheres, which show a favourable interaction with atoms in their surroundings.

The predictive power of some common programs were compared on a set of 48 structures where ligand bound and unbound models were available [45]. It was found that the methods that employed energy-based scoring, namely LISE and fPOCKET, were able to identify binding sites with much higher accuracy than those relying on geometric based scoring. One important consideration is the change in conformational state upon ligand binding. In the majority of cases, the programs were more accurate at predicting the correct binding site from the ligand-bound structure. However, one drawback the binding site prediction softwares mentioned here have is that they attempt to derive predictions from a single conformation of a protein. Upon binding of a drug to a protein the protein may undergo significant structural rearrangement, if the protein model is in the *apo* state drug binding sites may not be apparent in the model. One program, Provar, showed that using pocket predictions across an ensemble of protein models in a range of related conformations to the program was able to identify protein binding sites that may not have been evident within a single model [46].

Molecular docking

Once a suitable binding site is known, this data can be used to dock a virtual library of ligand molecules into the target model binding site (Figure 6). The aim of such an experiment is to identify ligands which are good binders to the target binding site. It is assumed that ligands that are good binders will also have some biological effect on the target, it is important to note that these types of experiments make no assumptions as to the 'drug-like' properties of ligands.

A multitude of protein-ligand docking software currently exists, each is composed of two parts; a search algorithm and a scoring function. The search algorithm generates multiple solutions (confirmations of the docked protein/ligand complex) and the scoring function provides a way to rank these solutions.

Scoring functions

Scoring functions are designed to correlate with the binding affinity between the docked ligand and protein. Within a virtual screening campaign, they function to identify potential lead compounds from the docked solutions. There are four general classes of scoring functions used: force-field-based, knowledge-based, empirical-based and machine-learning scores.

Knowledge based scoring functions are derived from large databases of protein/ligand interactions. These functions assign energy scores to docking features based on the frequency of their occurrence in a training set database. The most common docking feature used is a statistical preference for the distance between atoms pairs. This data is then transformed to give a probability distribution of 'free energies' for atom pairs. The knowledge-based score is

calculated as the sum of the 'free energies' in the docked ligand/protein model. One such scoring function that uses this approach is DrugScore [47].

Force field-based scoring functions are derived from equations describing nonbonded atomic interactions such as hydrophobic, electrostatic and H-bond interactions to estimate the affinity between ligand receptor binding. The scoring terms used in force field based scores are designed to approximate the scoring functions used in molecular dynamics (MD), for example the DOCK scoring function uses a Lennard-Jones function to describe vdW interactions and a coulombic function describing electrostatics [48].

Empirical scoring functions expand on force field-based scores by weighting interactions terms. This weighting coefficient is derived from fitting theoretical values to known experimental values. Thus, empirical scoring functions require calibration with a training data set.

One of the most well known empirical scoring functions is the AutoDock Vina scoring function [49]. This scoring function contains terms for steric interactions, repulsion, hydrogen bonding and hydrophobic interactions, and was calibrated on a subset of the PDBBind dataset [50, 51]. The AutoDock Vina scoring function is a relatively simple docking score. An example of a more advanced empirical scoring function is the GLIDE scoring function [52, 53]. The GLIDE scoring function is an extension of the ChemScore empirical scoring function that not only contains terms for hydrogen bonding and hydrophilic interactions but also includes terms for metal ion complexes [54]. The GLIDE score extends this scoring function with multiple hydrogen bonding terms depending on the charge states and atom types of donor and acceptor atoms, improved metal interactions terms, and a solvation term [52].

Machine learning scores that aim to predict the free energy of a protein ligand complex have been introduced more recently. One such score is the RF-Score-VS shown to outperform the AutoDock Vina scoring function when identifying potential lead compounds and estimating free energy [55].

Search algorithms

Once an appropriate scoring function has been identified a number of protein-ligand complex confirmations must be generated to be scored. Due to the large number of degrees of freedom most ligands have (3 translational, 3 rotational and *n* torsional degrees of freedom for each rotatable bond) there is a large number for possible conformations a ligand could adopt. Molecular docking programs aim to generate a large enough number of conformations to adequately search this space.

There are three general classes of search algorithms: shape complementarity, systematic search, and stochastic algorithms. The shape complementarity methods attempt to optimise the geometric fit of a rigid ligand in the binding site. Systematic search algorithms treat the ligand flexibly around its rotatable bonds and optimise geometric ligand protein complementarity in a large search space.

The shape complementarity methods are the simplest methods employed. The original version of the DOCK algorithm [56] utilised this method. The ligand is fit into the target binding site in a number of geometrically permissible conformations and the solutions that show the highest degree of complementarity to the binding site are kept. The systematic searching algorithms generate a large number of possible conformations within the binding site. There are three main subgroups of this type, exhaustive searching, conformational assemblies and fragment based. Since its inception, DOCK has gone through 6 iterations of improved algorithms. The latest algorithm DOCK 6 [57] utilises a fragment-based approach, where the ligand is broken down into its constituent moieties where atoms are connected by rotatable bonds. The largest of these fragments are then docked into the binding site based on their shape complementarity and the smaller fragments sequentially added until the ligand is fully reformed.

Exhaustive searching methods such as GLIDE [52] aim to search the conformational, positional and orientational space for a ligand docked into a target site, by allowing the ligand to be flexible about its rotational bonds. For a ligand with a large number of rotational bonds however this can be computationally expensive. Conformational assembly-based docking solves this problem, pre-defined ligand conformations are docked as rigid bodies into target binding sites.

Stochastic algorithms employ an element of probability to generate binding solutions. There are many different types of these algorithms, some of the most common are Genetic Algorithms (GA), Monte-Carlo simulations (MC) and Particle Swarm Optimization (PSO).

MC simulations involve making random changes to the solution of a docked ligand and assessing if these changes will be accepted using a Boltzmann probability distribution. The AutoDock Vina algorithm utilised a type of MC simulation called stochastic hill climbing to generate docking solutions [49].

The program GOLD [58] uses a genetic algorithm to search the ligand conformational space. The genetic algorithms are based on population dynamic principles such as inheritance, mutation and fitness. First a population of docked ligand conformations is randomly generated. Then genetic operators (such as random mutation, crossover, migration) are applied to the solutions to generate a new ‘generation’ of conformations. The fitness of these ‘children’ is assessed using a scoring function and the ‘weaker’ members of the parent population are replaced with the child conformations.

The PSO algorithms are very similar to the genetic algorithms. The algorithm starts with a population of randomly generated solutions. Through different generations the algorithm attempts to move the solutions towards the energy minima for the ligand. Further solutions are generated and each solution carries a vector value (a direction in which it moves towards its solution). The fitness of each solution is calculated and subsequent changes to the population follow the vector of the fittest solution. The program PLANTS [59] utilises a PSO algorithm based on the ability of ants to find the shortest path to their food.

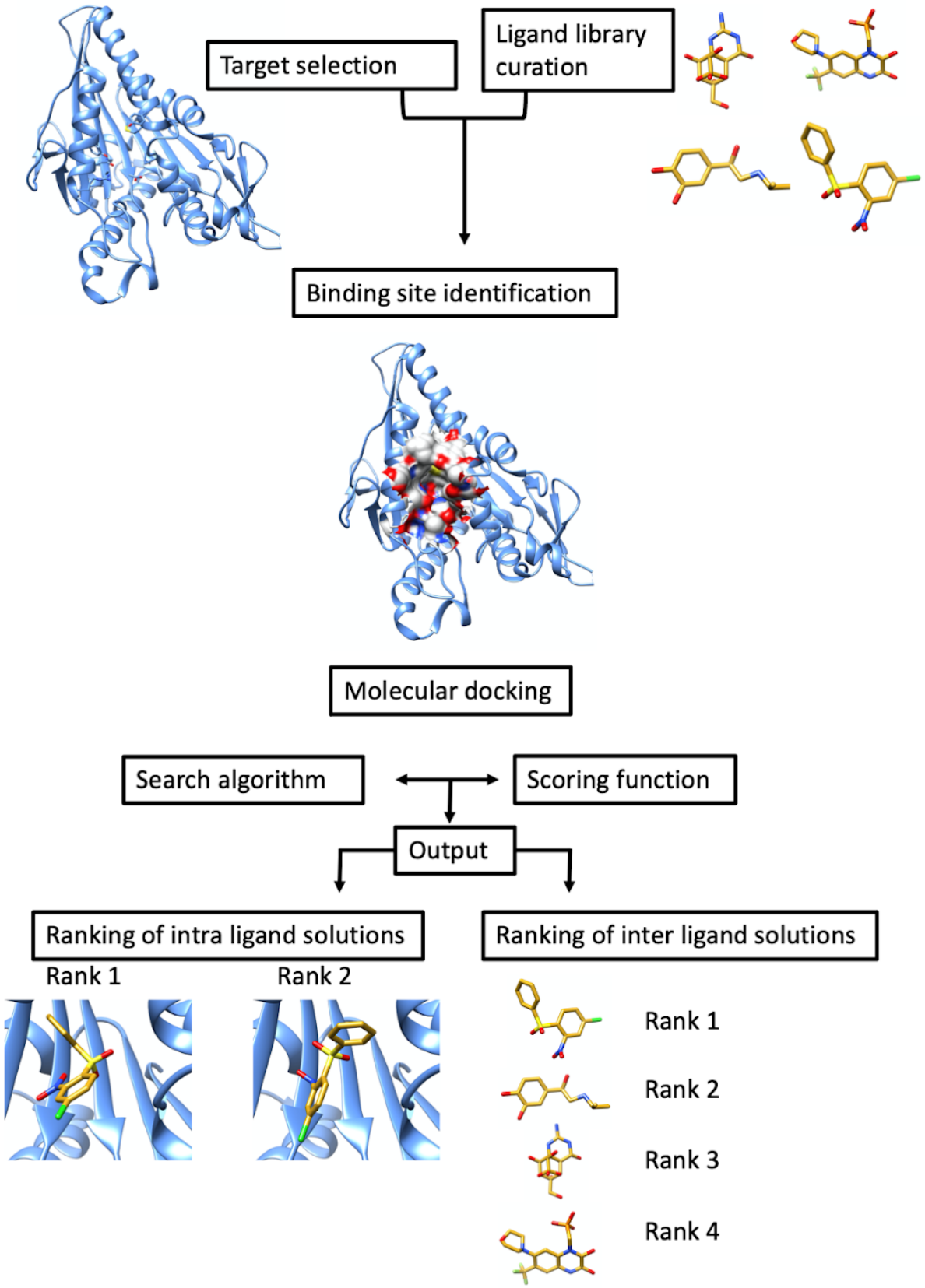


Figure 6. *An overview of a virtual screening pipeline. The pipeline begins with target selection and the curation of a virtual ligand library. The binding site within the target is then identified, before molecular docking is executed. In molecular docking solutions are*

constantly generated and scored. The output of molecular docking aims to both identify a correct ligand conformation within the binding site (intra ligand solutions) and rank all the ligands using the docking score (inter ligand solutions).

Consensus docking

One recent report evaluated the performance of ten docking programs for their performance [60]. A benchmark of 2002 ligand protein complexes, for which crystal structures and experimental binding data were available were docked and the power of the programs to generate the correct conformers and approximate free binding energies was assessed. To assess whether the correct pose was generated by the programs, the best pose (assessed by the conformation with the closest heavy atom root-mean-square deviation (RMSD) to the crystal structure) was scored as correct if its heavy atom RMSD to the native ligand was ≤ 2 Å. It was seen that the ten programs were able to generate a correct solution 60-80% of the time. Furthermore, the study showed that the programs predicted a correct ligand conformation as the top ranked solution only 40-60% of the time. This is an important finding as it indicated that the top scored pose does not always correspond to the correct pose. Taken together the results indicated that the ability of docking programs to produce the correct conformation is relatively good, however distinct conformations can yield similar ranked scores.

Most structural based virtual screening strategies would usually combine a number of scoring functions to try and increase the number of true hits produced. A more recent idea proposed involves using multiple docking methods to increase the number of hits identified during virtual screening [7]. This idea was born from the observation that correct prediction of the binding affinity was dependent on the correct ligand conformation being identified by the search algorithm. It was seen that using a database of 228 ligands for which crystal structures were available, when docked into their respective receptors using either AutoDock or AutoDock Vina the correct binding conformations were predicted 55 % and 64 % of the time, respectively. When only the consensus results from both were considered the accuracy was increased to 82% [7].

Molecular dynamics

MD aims to simulate how complex systems will behave over time. This can be applied to whole proteins and protein-ligand complexes. Unlike molecular docking, MD commonly simulates the movement of all atoms within the system. To do this the energy functions used are more complex than those in molecular docking experiments. This is due to the need to simulate covalent and noncovalent interactions. Commonly these molecular mechanics force fields are fit to approximate experimentally derived quantum mechanical calculations and experimental data. Covalent bond features such as stretching, bending and torsion, can be modelled with spring-like terms that keep bond geometry close to preferred values. For example, bond stretching and bending (i.e. bond length and bond angle) can be modelled relatively well using the harmonic form of a morse equation (E.q 1, 2).

$$Eq\,1.\ E_{bond}(r) = Kb(r - r_0)^2$$

Where, *r* is the simulated bond length, *r0* is the reference equilibrium bond length and *Kb* is a spring (or force) constant. Bond bending (i.e. bond angles) can be modelled by (E.q. 2):

$$Eq\,2.\ E_{angle}(\theta) = Kb(\theta - \theta_0)^2$$

Where $\boldsymbol{\theta}$ is the simulated bond angle, $\boldsymbol{\theta 0}$ is the reference equilibrium bond angle and *Kb* is a constant. Bond torsion energies can be modelled with a cosine curve (E.q.3):

$$Eq\,3.\ E_{torsion} = \sum A\,[1 + cos(n\tau - \phi)]$$

Where *A* is a constant, n controls the periodicity of the curve, τ is the torsion angle and ϕ controls the phase shift. Non-bonded terms are usually described as in molecular docking force fields, for example dispersion and repulsion can be models with a Lennard-Jones term (E.q 4):

$$Eq\,4.\ E_{lj} = 4\epsilon\left[(\frac{\sigma}{r})^{12} - (\frac{\sigma}{r})^{6}\right]$$

Where, ε is the well depth, $\boldsymbol{\sigma}$ is the vdW radii and r is the distance between atoms. Furthermore electrostatic interactions can be estimated with a coulombic expression (e.q 5):

$$Eq\,5.\ E_{col} = \frac{q_i q_j}{4\pi\epsilon r}$$

Where, qi/qj is the charge of atom i and j, ε is the dielectric constant of the medium, and r is the distance between atom pairs. In Molecular mechanics force fields the potential energy of the system would be calculated as the sum of all of the outlined equations (E.q 1 to 5) for all pairs of atoms in the system (E.q 6):

$$Eq\,6.\ = \sum_{bonds} E_{bond} + \sum_{bond\,angles} E_{angle} + \sum_{torsions} E_{torsion} + \sum_{nb-atom\,pairs} E_{lj} + E_{col}$$

Where *nb atom pairs* is the set of all non-bonded atom pairs in the system.

Many molecular mechanics force fields have been developed for use in MD simulations with protein systems, such as AMBER [61, 62] others contain additional parameters for small molecules, such as the CHARMM [63] and OPLS3 [64, 65] force fields.

The MD approach is generally considered to be more accurate than molecular docking. However, due to the length of time MD experiments take to complete, at present, the approach alone is not directly applicable to large scale drug discovery.

Experimental determination of structure

The computational methods outlined in the previous section rely heavily on the acquisition of high quality protein models. Whilst computational methods for structure prediction exist, such as homology modelling [66], by far the most reliable protein models are generally derived from experimental methods. The most common of which is X-ray crystallography and more recently cryogenic-electron microscopy (cryo-EM). Other methods of structural determination exist, such as Nuclear Magnetic Resonance (NMR) and micro-electron diffraction (micro-ED), however, they were deemed outside the scope of this work and as such are not discussed further.

Resolution

Different experimental techniques have different methods of calculating the resolution of a structure. However, optically resolution can be defined as the smallest distance at which two features can be distinguished. Microscopes are limited by many factors related to the microscope itself such as the numerical aperture and correct alignment of lenses. One main factor which limits the resolving power of a microscope is the wavelength of light used as an illumination source. A general equation to describe the relationship between resolution and wavelength illustrates this point (Eq. 7):

$$Eq\,7.\ R = \frac{1.22\ \lambda}{D_{scope}}$$

This is the Rayleigh criterion for resolution, where lambda is the wavelength, and D_{scope} is the diameter of the objective aperture. It is clear from this equation that shorter wavelengths will give a better resolving power (smaller value of R). Thus, in principle, the resolving power of light microscopy is limited by the wavelength of light. The theoretical resolution limit is approximately 200 nm, although more advanced techniques such as near field scanning techniques [67] show resolutions of approximately 30 nm. This resolution is not fine enough to clearly view distinct features of biological macromolecules, which are generally measured in the order of Å, i.e. $1 * 10^{-10}$ metres.

X-ray crystallography

X-ray crystallography experiments aim to determine the 3D structure of a molecule, for example a protein or DNA, based on the diffraction patterns of X-rays passing through a crystal (where diffraction refers to the scattering of electromagnetic radiation by matter). The main concept behind this technique is that atoms in a crystal diffract X-rays in patterns dependent on their location in 3-dimensional space. The reason why X-rays are used is that their wavelengths (typically 10^{-8} to 10^{-12} m) are of the same order of magnitude as the bonds between atoms, thus allowing individual atoms in a molecule to be resolved.

For diffracted X-rays to be detected with a high enough intensity, an ordered array of molecules (i.e. a crystal) is needed, so that equivalent atoms contribute to the diffraction pattern. To obtain crystals ultra-pure biological samples must, in an ordered fashion, switch between a liquid and solid state. Crystals found in nature e.g. NaCl or diamond, are held together by strong forces such as ionic or covalent bonds, respectively. On the other hand, crystals formed from biological macromolecules such as proteins are much more delicate being held together mainly by vdW forces or hydrogen bonds.

Once a crystal is obtained a diffraction pattern is generated by firing an X-ray beam at the crystal. The beam is scattered by electrons in atoms to form a diffraction pattern. The crystal is then rotated to collect diffraction data in 3-dimensions. This data is then used to retrospectively construct the electron density. As the crystal is an ordered array of atoms in a lattice the beams are diffracted in a systematic way, which is representative of the electron density of the sample. To compute the electron density map three pieces of information are needed: the intensities of spots, the positions of all spots, and the phase of the X-ray waves corresponding to each spot.

The resolution of the structure is determined by the angle of the outermost spots in the diffraction pattern using the Bragg equation (E.q 8):

$$Eq\,8.\ d = \frac{\lambda}{2sin(\theta)}$$

Where d is the resolution (lattice spacing), λ is the X-ray wavelength, and θ is the angle of reflection.

The two main quality assessment scores used to determine how well the atomic model describes the experimental data from an X-ray crystallography experiment are the R and R_{free}. R measures how well a model derived from the simulated diffraction pattern agrees with the observed diffraction pattern. The R value is calculated based on structure factors which describe the scattering caused by a particular plane in a crystal from the recorded intensities.

These structure factors are used to generate an initial model and are changed through many iterations of the refinement process. R is calculated by (E.q 9):

$$Eq\,9.\ R = \frac{\sum_{h,k,l} ||F_{obs}|\ |F_{calc}||}{\sum_{h,k,l} |F_{obs}|}$$

A perfect match will yield an R value of 0.0, a random assortment of atoms will yield an R value of approximately 0.6. One main criticism of the R value is that the R value is used to guide the refinement process and thus can be 'artificially' lowered by for example the omission of experimental data or the addition of water molecules.

R_{free} provides a more objective way to assess model quality. It is calculated in the same way as R. However, before construction of the initial model and refinement process, 10% of the experimental data is removed. The model is built and refined using the remaining 90 % . The R_{free} value indicates how well the model can explain the 10% of omitted experimental data. This value avoids the bias associated with R values. However, R_{free} has its own limitations as low quality local data can be hidden by a good average structure.

The quality of local features of a model, such as the position of residue sidechains, can be assessed by B-factors; these B-factors describe the displacement of atoms from the average position within a crystal (Also known as the Debye-Waller factor and expresses the extent of thermal vibrations within a lattice). Thus if a side chain is largely flexible it will have a larger B-factor associated with it. There is no one measure of the quality of a model and factors describing both global and local quality must be taken into account when assessing model quality.

Experimental determination of structure with cryo-EM

Electron microscopy has been around since the 1930's where it was shown to break the resolution limit of light. Just like photons used for light microscopy, electrons also behave as waves, and thus can be used to image samples. The wavelength of an electron is given by the De Broglie relationship (Eq. 10):

$$Eq\,10.\ \lambda = \frac{h}{mass * velocity}$$

Where, h is the Planck's constant and λ is the the wavelength of an electron. Electrons are charged particles that can be accelerated by an electric field, this has the effect of decreasing the wavelength of the electron beam. The smaller this wavelength the higher the resolving power of the microscope.

The most common type of electron microscopy used for structural determination is transmission electron microscopy (TEM), where image formation is dependent upon the movement of electrons through a thin sample and collection at a detector.

The interaction between electrons with matter is a relatively strong one. To avoid interference from atoms in the air with electrons used for imaging, the samples are imaged in a vacuum and in a solid phase [68]. To obtain samples in a solid phase they can be dehydrated, however, it is well documented that hydration is important for the maintenance of the native 3D structure of macromolecules [69]. Therefore, to obtain biological molecules in a hydrated solid state samples in solution are snap frozen in a timeframe that occludes the formation of damaging ice crystals [70]. This forms the basis of cryogenic-electron microscopy (Cryo-EM). This vitreous ice suspension also has the effect of cryo-protecting the sample against the damaging effect of electrons.

An electron passing through a biological sample can have several fates. It may pass straight through about its original trajectory, or it may be scattered by the atoms of the sample. This scattering can either be inelastic, involving energy transfer between the electron and atoms of the sample (and can damage the sample), or elastic (no energy transfer), where the path of the atom path length is perturbed slightly (Figure 7) [71].

It is this elastic scattering that provides the contrast from which images are formed. For any given image, different regions are distinguished from each other by contrast. The more distinct the contrast between two features the clearer the image. Images are usually formed from amplitude contrast (Figure 7), this occurs where parts of the sample absorb electrons and leave a sort of 'shadow region' on the detector where no electrons reach, which forms an outline of the image.

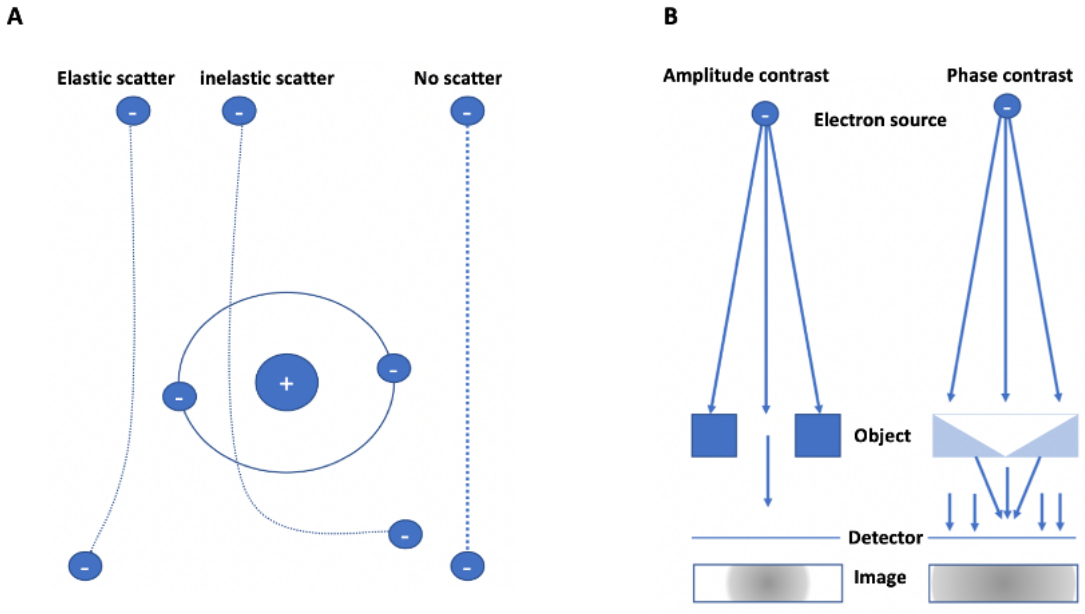


***Figure** 7. Electron scattering and contrast concepts. (**A**) shows the possible fates of electrons moving through a biological sample. Electrons can be elastically scattered or inelastically scattered. Both possibilities alter the pathlength compared to an unscattered electron. (**B**) shows the concepts of amplitude (left) and phase (right) contrast. In amplitude contrast*

electrons are blocked at regions of the object. Only electrons that pass through reach the detector, where amplitudes are recorded. This results in an image with a high contrast between where electrons have reached the detector and where they have not. In contrast elastic scattering as electrons pass through the sample cause pathlength to change and become out of phase with the incident beam. It is a lot harder to distinguish between phases in the image as the changes are slight. Image recreated from [71].

However, in the cryo-EM experiments used for structural determination there are very few electrons absorbed by the sample and images are formed from phase contrast (Figure 7). The elastic deflection of electrons by the sample changes the pathlength and hence phase of electrons passing through the sample. By introducing a phase shift of 90° relative to the incident beam this phase contrast can be converted to amplitude contrast from which 2D images can be produced [71].

3-dimensional density map formation

Sample molecules in a vitrified layer of ice will exhibit a large distribution of orientations from which 2D images can be collected. These 2D projection images are then used to construct a 3D density map. A digital 2D image can be thought of as a grid of pixels with each pixel containing a density value. A digital 3D image is an extension of this concept: an array of voxels (A 3D pixel) with associated density values. One of the most common ways of constructing 3D images from projection images uses Fourier space and the central section theory. This theory states that for a particular 3D image and projection images, a Fourier transform of the projection image is equivalent to taking a Fourier transform of the 3D image and taking the section through the origin that is perpendicular to the projection image.

The most common technique used for high-resolution structural determination in cryo-EM is single-particle reconstruction. In this technique a large number of 2D projection images are obtained, these images are then classified and averaged. There is one class per orientation of the particle, the averaging serves to increase the signal-to-noise ratio present in each image. A Fourier transform of all these images is calculated, and is combined with a reference 3D Fourier transform, by comparing experimental and reference projections. The inverse Fourier transform is then taken to retrieve the images in real space and the process is iterated using the newly generated reference map each time to refine the final map [72].

There are two main problems with this approach, the first arises when two different images contain mismatching information on the densities of the same voxel, the second is if there are significant regions of the 3D projection for which there is no projection image. The first problem can be handled by weighting the projection voxels before merging. Algorithms for this are contained within many image processing software, such as RELION [73]. The second problem is a lot harder to handle and can lead to anisotropic resolution in the final model.

Resolution estimates in cryo-EM

Resolution estimates in cryo-EM aim to determine at what resolution reliable data is present within the map. This is done using a method called the Fourier Shell Correlation (FSC). This method involves first splitting 2D projections into two half maps and refining a density from both. The Fourier transforms of both maps are compared by a cross-correlation at different spatial frequencies. There is currently no consensus as to what the correlation cut-off should be, early works use an FSC cut-off of 0.5 [74], however, further estimates of 0.143 [75] and 0.333 [76] have been proposed. Whilst the FSC provides a global estimate of the resolution of a map, it has now become clear that a global estimation of resolution is not sufficient to fully describe the distribution of quality of the experimental data. This can be due to many factors such as particles representing multiple conformations or missing data relation to projections around the Euler sphere. To this end methods have been developed that estimate the local resolution of a map. One such method, *blocres,* implemented in B-soft calculates the local FSC using a sliding window of size n^3 over the entire map [77]. Another method, ResMap, uses sinusoidal features to derive local resolution estimates across an EM density map [78].

The ‘resolution revolution’ in cryo-EM

Until recently the resolution of cryo-EM was limited to ~5 Å in most cases. Advances in technologies have pushed this resolution limit closer to that of the near-atomic resolution commonly seen in X-ray crystallography experiments. This is in part due to direct electron detectors used to capture images. It has been shown that using direct detectors in combination with improved imaging software, researchers are able to reduce motion blur [79] from microscopic movements of the sample staging area or induced by the electron beam itself [80]. The use of direct detectors with improved image processing software has resulted in a larger number of sub 5 Å structures being solved with cryo-EM.

Atomic Model building with cryo-EM data

For cryo-EM density maps to be useful for further experiments, atomic models must be calculated that represent the experimental data. To build and assess the quality of such models methods are needed that quantify how well an atomic model describes the experimental data (termed goodness-of-fit metrics).

Goodness-of-fit metrics

When calculating atomic models using cryo-EM density maps a metric is needed to assess the goodness-of-fit of the atomic model to the map, *i.e.,* how well the atomic model describes the experimental data. Since the exact positions that correspond to individual atoms are ambiguous a direct comparison of map to model is very complicated. The most common method to do this is to calculate a ‘synthetic’ density map using the atomic model and compare this map to the experimental data.

The method of calculating density maps from atomic models is usually referred to as blurring. A general method for achieving map blurring starts with transferring 'amplitudes' to a map of empty voxels, based on the position and atomic mass of atoms in the model. This map is then convolved with a Gaussian function. The level of blurring is controlled by altering the Gaussian σ variable that defines the width of the function. This σ is usually defined as the product of the resolution and a constant. The constant used can vary, two commonly used values are 0.225 [15] and 0.187 [81]. Map blurring in this way yields a map where the amplitudes of the densities are proportional to the atomic mass of atoms and the distribution proportional to the resolution [82, 83].

Methods to compare the fit of two maps generally fall into two broad categories, global or local. Global scores give an idea of how the model as a whole fits to the map, whilst local scoring metrics are able to give details regarding the goodness-of-fit of local regions of the model.

By far the most commonly used global metric is the cross correlation coefficient (CCC) (E.q 11):

$$Eq\,11.\ CCC = \sum \frac{x^i\ y^i}{|x^i|\ |y^i|}$$

Where x^i and y^i are the voxels values in the simulated and experimental maps, respectively.

A further metric that has been shown to be successful for the alignment of 2D images in cryo-EM is the mutual information (MI) score [84]. The MI score quantifies the information overlap between two probability distributions and is calculated as (E.q 12):

$$Eq\,12.\ MI = \sum_{x \in X} \sum_{y \in Y} \rho(x,y)\ log \frac{\rho(x,y)}{\rho(x)\rho(y)}$$

Where X and Y are the set of voxel values in the experimental and simulated maps, respectively. *p*(x) and *p*(y) are the probability of voxels x and y occurring given the values of the whole set of voxels. *p*(x,y) is the probability of the voxels x and y being aligned given the total set of aligned voxels between the sets X and Y. Due to the wide range of voxel values within a cryo-EM map it is necessary to bin the density values for calculation of the MI. Previous reports have shown that using 20 bins can yield good results when comparing 2D [84] and 3D protein maps [82].

Other global metrics for the goodness-of-fit include least squares function, difference-least squares function, Laplacian filtered CCC, Envelope score, Normal Vector score and Chamfer distance [82].

In order to probe the fit of local regions of the model to the map local scores have been introduced. One such score is the local segment-based cross correlation (SCCC). This score is calculated in much the same way as the CCC only over a local region of the map. For a given local region of the model, *e.g.*, an alpha helix, only the voxels in the simulated map containing density information from these local regions are used in the calculation. When the simulated map is aligned with the experimental map the CCC is calculated using the simulated local voxels and the experimental voxels to which they align. This concept was expanded on with the segment-based Manders' overlap coefficient (SMOC), that calculated the CCC across a sliding window of nine residues. The window is moved over one residue at a time iteratively to obtain a SMOC profile across the entire protein atomic model [10].

Resolution dependent refinement strategies

The refinement protocol is highly dependent on the resolution achieved experimentally. At high resolutions (~ ≤ 3.0 Å) the positions of individual atoms and chemical groups should be visible within the map. As the resolution worsens, visually identifiable features within the map begin to deteriorate. At resolutions around 5 Å the positions of bulkier sidechains and individual SSEs may be visible. In terms of SSEs the α-helices are generally visible at resolutions up to 10 Å (albeit only long ones), whilst β-sheets tend to disappear visually before this resolution (~5-6 Å and 4.5 Å for individual strands). At resolutions beyond 10 Å, visible features may only include domains, subunits and their respective boundaries (Figure 8).

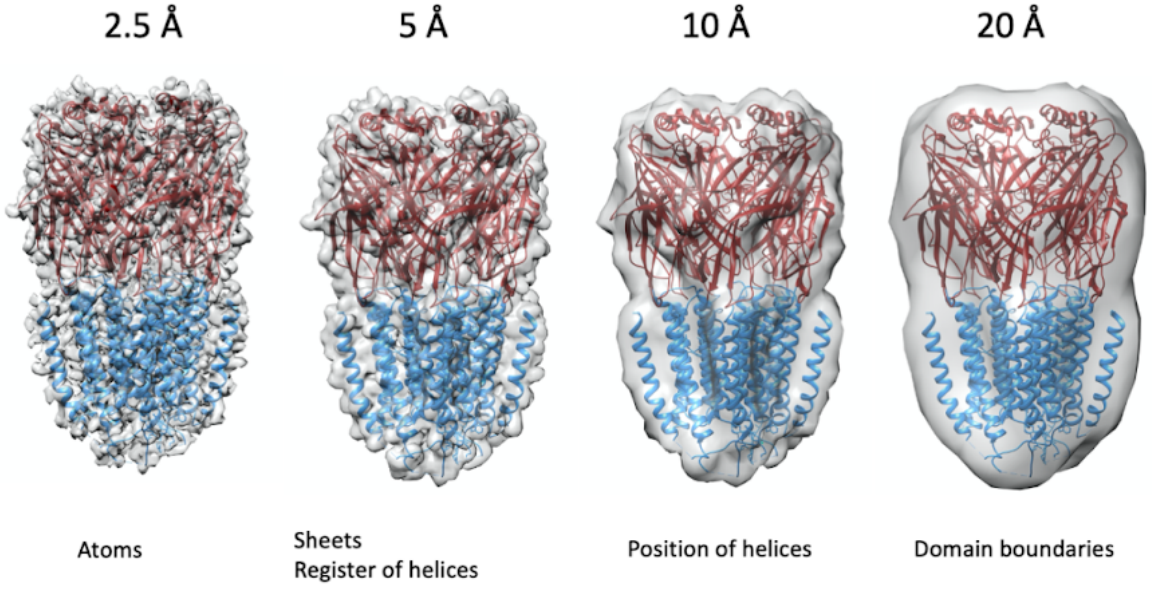


Figure 8. *Density maps simulated from an atomic model of the $GABA_A$ receptor at resolutions of 2.5, 5, 10, and 20 Å. In the atomic model the N-terminal domains of different*

subunits are colored red and the trans-membrane domains are colored blue. Also indicated underneath each image is the approximate map features visible at the defined resolution.

At atomic resolutions, where it is possible to delineate the positions of atoms in the protein backbones and sidechains, it is possible to directly build a protein model into a cryo-EM density map. The software C*oot* contains tools for the automated detection of SSEs from density that can be used to build an atomic model directly [85]. The software PHENIX was originally developed as a tool for model building using X-ray crystallography data, and has been extended for the refinement of models using cryo-EM data [86]. One common strategy for obtaining an initial model using high resolution cryo-EM data is to build an initial model with C*oot*, followed by a real-space refinement of the model in the map with PHENIX [5]. The *'phenix.realspace_refinement'* protocol uses a gradient driven minimisation of the model by default, however has options for simulated annealing and morphing. The scoring function contains terms that score the model fit with the map and protein model geometry. This refinement strategy has been used to calculate models with high resolution maps including the $GABA_A$ receptors [1].

However, at lower resolutions it is much more common to generate an initial model that represents an approximation of the protein model and refine this into the density map. This method is especially common with intermediate resolution reconstructions (3.5 Å - 15 Å) where the direct visualisation of side chain atoms is not possible and at low resolution (> 15 Å) where it is not possible to determine the position of SSE's.

It is common to use models of proteins previously derived using high-resolution X-ray crystallography data. However, when a model may be incomplete or unavailable, homology modelling can be used to generate an initial model using sequence or structural alignments with protein models that share a degree of sequence or structural similarity. The MODELLER package [66] calculates homology models using sequence alignments between the structure to be modelled and one or more template structures. This technique generates a number of possible solutions, the plausibility of solutions can be scored using the built-in discrete optimised protein energy (DOPE) score to identify the most relevant structure [87]. The DOPE score uses distance based statistical potentials that aim to predict the global free energy minimum of a protein. Other methods of obtaining initial models include using phyre^2 [88] that uses hidden Markov models to identify structurally homologous proteins for use as templates during model prediction. Other approaches include *ab initio* structure prediction (e.g. Rosetta [89]), and more recently, machine learning has been leveraged for the accurate prediction of protein structure [90]. All these methods produce multiple models at similar accuracies that can be constrained by the map.

Once an initial model has been obtained it must be fitted into the map. Initial fits can be done by keeping the internal coordinates of the protein model rigid and optimising with a goodness-of-fit score, methods to do this can be found in software such as Chimera [15] that conducts a 6D local search over the map whilst optimising the CCC. This procedure can also be accomplished with an exhaustive global search using software such as DOCK-EM [91].

However, much of the time the protein conformation imaged in the experimental map differs somewhat from the conformation in the initial model. Therefore, introducing an appropriate amount of flexibility into the fitting process at lower resolutions can improve the model-map goodness-of-fit.

The most common way to refine models at the intermediate-to-low resolution is with MD, or similar approaches, which will allow one to optimise the geometry of the model while improving its fit in the map (*flexible fitting*). This involves integrating classical MD force fields with a potential term derived from the density map itself. The software MDFF uses [92] a combination of MD potentials, potential derived from the density, and harmonic restraints to maintain appropriate secondary structure to fit models into cryo-EM maps. More recently the software ISOLDE was released [93], an interactive software implemented in ChimeraX [94], which includes restraints for dihedral bonds, atom distance and positions to aid with low resolution refinement. This protocol has been used to refine an atomic model of the pore forming unit of the ABC endotoxin at 4.4 Å [95].

At resolutions worse than atomic resolution, it is not advisable to allow all atoms to move independently due to ambiguity in their positions within the map. To avoid this, the protein model can be split into rigid bodies. Common ways to define rigid bodies include by domains and SSEs. The program RIBFIND [96, 97] has been reported for the automated identification of rigid bodies, by clustering groups of SSEs using user defined distance cutoffs.

Once rigid bodies have been identified, flexible fitting can be achieved by moving rigid bodies relative to each other, whilst restraining their internal coordinates. One such software developed for this purpose is Flex-EM [4, 97], where the atomic positions of atoms within rigid bodies are optimised using simulated annealing MD and a scoring function composed of terms for the CCC of the model with the map and stereochemical and non-bonded terms to ensure the model is sensible from a physico-chemical point of view. Using simulated annealing in combination with the CCC rather than density based potentials ensures atoms do not get stuck in local maxima during fitting. A hierarchical strategy for fitting was proposed with Flex-EM that involved multiple rounds of refinement with progressively smaller rigid bodies [97]. This strategy was used, for example, to model the kinesin-8 motor domain complexed with microtubules [98].

Another program that utilised the CCC, MD and correlation derived potentials for fitting at intermediate and low resolutions is the correlation-driven molecular dynamics (CDMD).To avoid atoms becoming stuck at local density maxima the refinement process is conducted iteratively, beginning with a simulated map blurred to a low resolution and constantly resampling at each iteration up to the best resolution available (determined by the resolution of the experimental map). Fitting in this way allows the atomic model to first adopt a good global fit, before refining local regions as the resolution of the blurred map increases. The CDMD methodology was validated using cryo-EM experimental data and shown to be able to fit protein models including that of the TRPV1 channel at 3.2 Å and tubulin at 4.1 Å [6]

Model validation

Once an appropriate fit has been produced, a method to validate the model that is independent from the fitting protocol is needed. Both the goodness-of-fit to the map and the chemico-physical reasonableness of the atomic model should be validated. The most common method to assess the fit to the map is with the global CCC. Additionally, local scores can often be informative at this stage to identify which regions of the map accurately or inaccurately describe the experimental data. Whilst, goodness-of-fit metric may be used in refinement it is good practice to not use the same metric for validation.

One of the simplest methods for the evaluation of protein model quality is by assessing the Ramachandran angles. Early work into the structure of proteins identified that the phi and psi dihedral angles have distinct values dependent on the SSE in which they exist [99]. Validation reports typically report the number of residues with either 'allowed', 'favoured' or 'outlier' values'. More recently assessment of Ramachandran scores has been converted to a Z-score [100].

With respect to the physico-chemical "reasonableness" of the atomic model the MolProbity [101, 102] software suite offers a plethora of validation tools. This includes validating protein torsion angles, bond angles, atom-atom distances, and non-bonded interactions. Additionally potential atom-atom clashes are identified. The MolProbity score combines clash scores, with Ramachandran and side chain rotamer scores into a single score, normalised to be on the same scale as resolution.

The original MolProbity suite was designed for use with atomic models derived from high-resolution structures. Since then further scores have been implemented for use with atomic models from lower resolution maps. The CαBLAM module uses the relative positions of Cα atoms to identify errors in the model backbone. The module assesses the model backbone for planar outliers, a common error in low resolution structures. Additionally, Cα dihedral angles are analysed. Furthermore, the MolProbity package identifies Ramachandran outliers along with side chain rotamer outliers [102].

Other common scores used for validation of geometric parameters of atomic models include the Qualitative model energy analysis (QMEAN) score [103]. This score was extended to assess the model quality of membrane proteins in the score QMEAN-Brane [104].

Small-molecule fitting in cryo-EM

Whilst a variety of methods have focused on fitting protein structures to cryo-EM maps, very few have focused specifically on methods for fitting small molecules to cryo-EM maps.

At atomic resolutions the fitting of small molecules follows a similar workflow to high resolution fitting for proteins. One of the most common strategies is to use C*oot* to build an initial ligand molecule into ligand density in the map. This protein ligand-complex can then

be further refined in real-space using the PHENIX software [5]. One key difference between the two workflows relates to the MD parameter functions for ligand atom and bond types. The atoms and bond types contained within proteins have been extensively studied and most MD force fields contain built-in parameters. Small-molecule ligands on the other hand have a much broader variety of atom and bond types and therefore it is often required that the parameters for handling the specific small molecules be added to the force field.

The derivation of bond and atom types in small molecules is non-trivial, especially if small molecules are supplied in '.pdb' format where bond type information is not provided. Software has been developed to automate this task, such as the electronic ligand builder and optimisation workbench (eLBOW) [105]. The software determines bond type information by comparing distances between ligand atoms with theoretically preferred values. Once bond types have been determined, a set of dihedral restraints can be assigned for each bond. This is needed by the MD software to ensure ligands maintain a meaningful geometry. Additionally, features include the assignment of partial charges and hydrogen positions.

More recently, the need for predetermined restraints was made redundant with the extension of the OPLS3 forcefield [64] to the OPLS3e forcefield [65]. The original OPLS force field contained harmonic energy terms for bond stretching and angles with torsion angle terms consisting of a truncated Fourier series summed over all tosions and a term representing vdW interactions [64]. The terms for bond angle and stretching were parameterised to quantum mechanical data, whilst the vdW term was parameterised to liquid state simulations. This OPLSe force field expanded on these parameters by generating molecule quantum chemical torsion energy profiles to score small molecule torsions, and assigning partial charges on-the-fly by fitting to quantum mechanical data [65]. This force field has since been added to the real-space refinement protocol contained in PHENIX to aid fitting small molecule ligands to cryo-EM maps [106].

A second strategy that has been shown to be able to fit small molecule ligands into cryo-EM maps at both high [107], and low [9, 98, 108] resolutions is to leverage molecular docking software. Multiple strategies using molecular docking have been proposed. The simplest of which compares the output of molecular docking software directly to the cryo-EM map [9, 98]. This strategy utilised some element of consensus docking to improve the quality of docking output. A more recent strategy included the CCC score directly in the docking score [107]. Multiple strategies have included post-docking refinement steps, one refined the output of molecular docking with the aforementioned OPLS3e/real-space refinement in PHENIX [107]. Another combined molecular docking with neural network potentials (NNPs) [108]. NNPs are neural networks that have been trained on sets of quantum mechanical data to estimate the energy of protein/ligand interactions. The NNPs have been shown to be nearly as accurate as quantum mechanical data with a significantly lower computational cost [109]. NNPs were combined with the output of docking data and a molecular mechanical force field describing the protein system in order to fit small molecules to low-resolution cryo-EM maps [108]

At low resolutions the density corresponding to ligands is not always evident. This can result in density from proteins bleeding into that of the ligand (Figure 9). One methodology has been proposed to deal with this situation involves density *difference mapping*. A difference map is essentially a subtraction of one map from another, theoretically leaving density that is unaccounted for in the map from which the subtraction is calculated (Figure 10). This technique requires scaling the amplitudes of both maps to be of the same magnitude. Scaling is usually conducted in Fourier space. Most methods usually scale the maps using the global average, by scaling amplitudes in defined resolution shells to a reference power spectrum [110–112]. Recently, a method of locally scaling two maps was reported [8]. This local

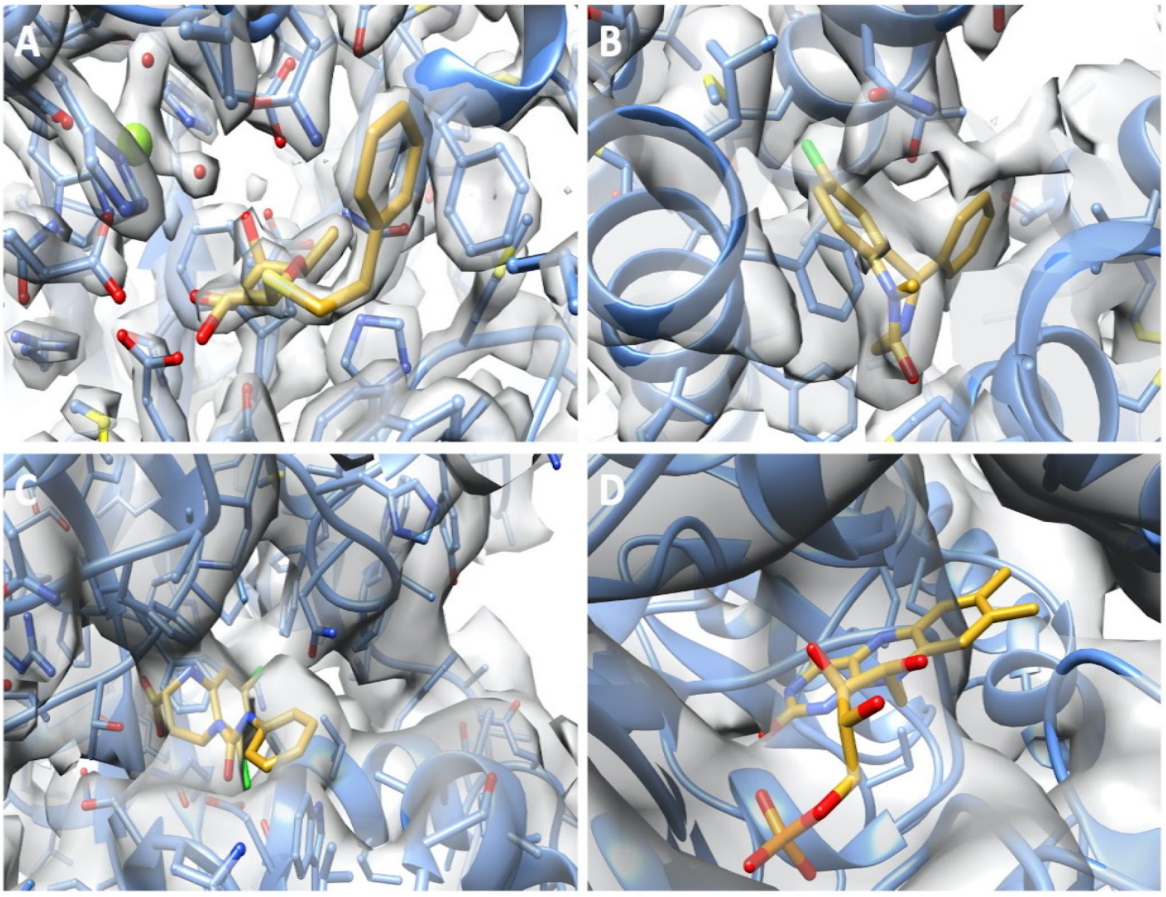


Figure 9. *Small molecule binding sites at various resolutions. (**A**) An atomic model of a molecule of 2-phenylethyl 1-thio-beta-D-galactopyranoside (yellow) within a binding site on the Beta-galactosidase protein (blue) within a cryo-EM density map (grey) at 2.2 Å. At this resolution the positions of protein sidechain atoms, solvent and ions in the map can be seen. (**B**) An atomic model of diazepam (yellow) bound in a subunit interface of the $GABA_A$ receptor protein (blue) in a cryo-EM map (grey) of 2.92 Å. At this resolution the position of the ligand molecule can be seen clearly in the density along with the approximate positions of protein residue sidechains. (**C**) An atomic mode of the $GABA_B$ receptor (blue) antagonist CGP54626 (yellow) in a cryo-EM map solved at 3.52 Å. At this resolution it becomes a lot harder to differentiate between protein and ligand density, although the approximate ligand position can be seen. Furthermore, the approximate positions of some larger sidechains is still visible in the map. (**D**) A molecule of flavin mononucleotide (yellow) bound to the*

respiratory complex I from Thermus thermophilus (blue) within a cryo-EM map of 4.25 Å resolution. At this resolution it becomes tough to identify protein density visually. Furthermore, no clear density corresponding to protein sidechains is visible.

scaling is conducted by scaling two aligned maps using a rolling cube of defined size, this is used to scale the voxels located centrally within the box. The amplitudes of both maps are scaled based on resolution shells. All the amplitudes in a particular resolution shell are scaled by a factor derived from the average amplitude in that shell. This protocol has been applied to the small molecules in cryo-EM maps, where the output of a molecular docking protocol was compared with a density difference map calculated between the experimental map and a map simulated using a refined protein model [98].

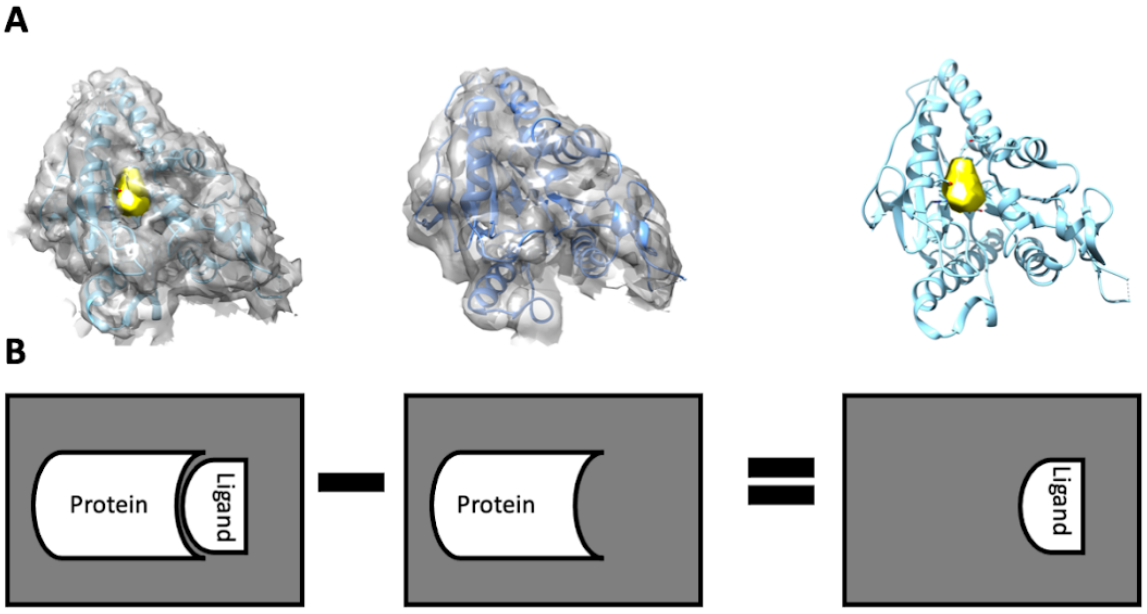


Figure 10. *A representation of the difference mapping method. (**A**) This method is represented using atomic models (blue) and cryo-EM density maps (grey), density corresponding to the ligand is highlighted (yellow). (**B**) A simplified illustration of the methodology where a cryo-EM map of a protein is subtracted from a cryo-EM map with density corresponding to a protein ligand complex. The end result theoretically leaves density differences between the two maps. In this case the ligand density.*

Cryo-EM in drug discovery

The contributions of cryo-EM to drug discovery have classically been limited to solving structures less amenable to crystallisation. One example of this are membrane proteins, exemplified by a number of structures of the $GABA_A$ Receptors in complex with various small molecule inhibitors [1].

However, one recent publication [113] reported one of the first examples of cryo-EM applied to fragment screening, a technique that until now has almost exclusively been applied to

X-ray crystallography. It involves incubating a protein drug target with a large number of molecule fragments, representing various chemical moieties, and solving the structures to identify bound fragments, before optimising these fragments into drugs in their own right. In the study it was shown that not only was cryo-EM able to unambiguously resolve the positions of small fragment molecules and side-chain conformations changes upon binding at resolutions of up to 3.2 Å, but also ligand density corresponding to fragments could be unambiguously assigned when protein targets were incubated with multiple fragment molecules. This study indicated that cryo-EM is fast becoming a relevant drug discovery technique in its own right, and a diverse range of accurate methodologies for the automated fitting of small molecules are needed.

References

1. Kim JJ, Gharpure A, Teng J, Zhuang Y, Howard RJ, Zhu S, et al. Shared structural mechanisms of general anaesthetics and benzodiazepines. Nature. 2020;585:303–8.

2. Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. New electron microscopy database and deposition system. Trends Biochem Sci. 2002;27:589.

3. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein Structure Fitting and Refinement Guided by Cryo-EM Density. Structure. 2008;16:295–307.

4. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42–9.

5. Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A, et al. Real-space refinement in PHENIX for cryo-EM and crystallography. Acta Crystallogr D Struct Biol. 2018;74 Pt 6:531–44.

6. Igaev M, Kutzner C, Bock LV, Vaiana AC, Grubmüller H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. Elife. 2019;8:e43542.

7. Houston DR, Walkinshaw MD. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. J Chem Inf Model. 2013;53:384–90.

8. Joseph AP, Lagerstedt I, Jakobi A, Burnley T, Patwardhan A, Topf M, et al. Comparing Cryo-EM Reconstructions and Validating Atomic Model Fit Using Difference Maps. J Chem Inf Model. 2020;60:2552–60.

9. Peña A, Sweeney A, Cook AD, Locke J, Topf M, Moores CA. Structure of Microtubule-Trapped Human Kinesin-5 and Its Mechanism of Inhibition Revealed Using Cryoelectron Microscopy. Structure. 2020;28:450–7.e5.

10. Joseph AP, Lagerstedt I, Patwardhan A, Topf M, Winn M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. J Struct Biol. 2017;199:12–26.

11. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015;43:W443–7.

12. Adasme MF, Linnemann KL, Bolz SN, Kaiser F, Salentin S, Haupt VJ, et al. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Res. 2021;49:W530–4.

13. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.

14. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007;35 Database issue:D301–3.

15. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.

16. Boopathi S, Poma AB, Kolandaivel P. Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. J Biomol Struct Dyn. 2021;39:3409–18.

17. Huang Y, Yang C, Xu X-F, Xu W, Liu S-W. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin. 2020;41:1141–9.

18. Rosenfeld SS, Fordyce PM, Jefferson GM, King PH, Block SM. Stepping and stretching. How kinesin uses internal strain to walk processively. J Biol Chem. 2003;278:18550–6.

19. Hyeon C, Onuchic JN. Internal strain regulates the nucleotide binding site of the kinesin leading head. Proc Natl Acad Sci U S A. 2007;104:2175–80.

20. Margenau H. Van der waals forces. Rev Mod Phys. 1939;11:1–35.

21. Jones JE, Chapman S. On the determination of molecular fields. —II. From the equation of state of a gas. Proceedings of the Royal Society of London Series A, Containing Papers of a Mathematical and Physical Character. 1924;106:463–77.

22. Böhm H-J, Klebe G. What can we learn from molecular recognition in protein–ligand complexes for the design of new drugs? Angew Chem Int Ed Engl. 1996;35:2588–614.

23. Panigrahi SK, Desiraju GR. Strong and weak hydrogen bonds in the protein-ligand interface. Proteins. 2007;67:128–41.

24. Emamian S, Lu T, Kruse H, Emamian H. Exploring nature and predicting strength of hydrogen bonds: A correlation analysis between atoms-in-molecules descriptors, binding energies, and energy components of symmetry-adapted perturbation theory. J Comput Chem. 2019;40:2868–81.

25. de Freitas RF, Schapira M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med Chem Commun. 2017;8:1970–81.

26. Sarkhel S, Desiraju GR. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. Proteins. 2004;54:247–59.

27. Pierce AC, Sandretto KL, Bemis GW. Kinase inhibitors and the case for CH...O hydrogen bonds in protein-ligand binding. Proteins. 2002;49:567–76.

28. Pierce AC, ter Haar E, Binch HM, Kay DP, Patel SR, Li P. CH...O and CH...N hydrogen bonds in ligand design: a novel quinazolin-4-ylthiazol-2-ylamine protein kinase inhibitor. J Med Chem. 2005;48:1278–81.

29. Auffinger P, Hays FA, Westhof E, Shing Ho P. Halogen bonds in biological molecules. Proc Natl Acad Sci U S A. 2004;101:16789–94.

30. Clark T, Hennemann M, Murray JS, Politzer P. Halogen bonding: the sigma-hole. Proceedings of "Modeling interactions in biomolecules II", Prague, September 5th-9th, 2005. J Mol Model. 2007;13:291–6.

31. Sarwar MG, Dragisic B, Salsberg LJ, Gouliaras C, Taylor MS. Thermodynamics of Halogen Bonding in Solution: Substituent, Structural, and Solvent Effects. J Am Chem Soc. 2010;132:1646–53.

32. Benjahad A, Guillemont J, Andries K, Nguyen CH, Grierson DS. 3-iodo-4-phenoxypyridinones (IOPY's), a new family of highly potent non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorg Med Chem Lett. 2003;13:4309–12.

33. Bissantz C, Kuhn B, Stahl M. A Medicinal Chemist's Guide to Molecular Interactions. J Med Chem. 2010;53:5061–84.

34. Grimme S. Do special noncovalent pi-pi stacking interactions really exist? Angew Chem Int Ed Engl. 2008;47:3430–4.

35. Stornaiuolo M, De Kloe GE, Rucktooa P, Fish A, van Elk R, Edink ES, et al. Assembly of a π–π stack of ligands in the binding site of an acetylcholine-binding protein. Nat Commun. 2013;4:1–11.

36. Biot C, Buisine E, Rooman M. Free-Energy Calculations of Protein−Ligand Cation−π and Amino−π Interactions: From Vacuum to Proteinlike Environments. J Am Chem Soc. 2003;125:13988–94.

37. Harding MM. Geometry of metal-ligand interactions in proteins. Acta Crystallogr D Biol Crystallogr. 2001;57 Pt 3:401–11.

38. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev. 2001;46:3–26.

39. Levitt DG, Banaszak LJ. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph. 1992;10:229–34.

40. Huang B, Schroeder M. LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol. 2006;6:19.

41. Laskowski RA. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph. 1995;13:323–30, 307–8.

42. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 1998;7:1884–97.

43. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10:168.

44. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci. 1996;5:2438–52.

45. Xie Z-R, Hwang M. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics. 2012;28:1579–85.

46. Ashford P, Moss DS, Alex A, Yeap SK, Povia A, Nobeli I, et al. Visualisation of variable binding pockets on protein surfaces by probabilistic analysis of related structure sets. BMC Bioinformatics. 2012;13:39.

47. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol. 2000;295:337–56.

48. Meng EC, Shoichet BK, Kuntz ID. Automated docking with grid-based energy evaluation. J Comput Chem. 1992;13:505–24.

49. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455–61.

50. Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. J Med Chem. 2004;47:2977–80.

51. Wang R, Fang X, Lu Y, Yang C-Y, Wang S. The PDBbind Database: Methodologies and Updates. J Med Chem. 2005;48:4111–9.

52. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J Med Chem. 2004;47:1739–49.

53. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J Med Chem. 2006;49:6177–96.

54. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des. 1997;11:425–45.

55. Wójcikowski M, Ballester PJ, Siedlecki P. Performance of machine-learning scoring functions in structure-based virtual screening. Sci Rep. 2017;7:46710.

56. DesJarlais RL, Sheridan RP, Seibel GL, Dixon JS, Kuntz ID, Venkataraghavan R. Using shape complementarity as an initial screen in designing ligands for a receptor binding site of known three-dimensional structure. J Med Chem. 1988;31:722–9.

57. Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, et al. DOCK 6: Impact of new features and current docking performance. J Comput Chem. 2015;36:1132–56.

58. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003;52:609–23.

59. Korb O, Stützle T, Exner TE. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In: Ant Colony Optimization and Swarm Intelligence. Springer Berlin Heidelberg; 2006. p. 247–58.

60. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016;18:12964–75.

61. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 2004;25:1157–74.

62. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015;11:3696–713.

63. Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, et al. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. J Comput Chem. 2010;31:671–90.

64. Harder E, Damm W, Maple J, Wu C, Reboul M, Xiang JY, et al. OPLS3: A Force Field Providing Broad Coverage of Drug-like Small Molecules and Proteins. J Chem Theory Comput. 2016;12:281–96.

65. Roos K, Wu C, Damm W, Reboul M, Stevenson JM, Lu C, et al. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. J Chem Theory Comput. 2019;15:1863–74.

66. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234:779–815.

67. Wabuyele MB, Culha M, Griffin GD, Viallet PM, Vo-Dinh T. Near-field scanning optical microscopy for bioanalysis at nanometer resolution. Methods Mol Biol. 2005;300:437–52.

68. J Jensen G. Cryo-EM, Part A: sample preparation and data collection. Preface. Methods Enzymol. 2010;481:xv – xvi.

69. Privalov PL, Crane-Robinson C. Role of water in the formation of macromolecular structures. Eur Biophys J. 2017;46:203–24.

70. Adrian M, Dubochet J, Lepault J, McDowall AW. Cryo-electron microscopy of viruses. Nature. 1984;308:32–6.

71. Orlova EV, Saibil HR. Structural Analysis of Macromolecular Assemblies by Electron Microscopy. Chem Rev. 2011;111:7710–48.

72. Penczek PA. Fundamentals of three-dimensional reconstruction from projections. Methods Enzymol. 2010;482:1–33.

73. Scheres SHW. RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J Struct Biol. 2012;180:519–30.

74. Böttcher B, Wynne SA, Crowther RA. Determination of the fold of the core protein of hepatitis B virus by electron cryomicroscopy. Nature. 1997;386:88–91.

75. Rosenthal PB, Henderson R. Optimal Determination of Particle Orientation, Absolute Hand, and Contrast Loss in Single-particle Electron Cryomicroscopy. J Mol Biol. 2003;333:721–45.

76. van Heel M, Schatz M. Fourier shell correlation threshold criteria. J Struct Biol. 2005;151:250–62.

77. Cardone G, Heymann JB, Steven AC. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. J Struct Biol. 2013;184:226–36.

78. Kucukelbir A, Sigworth FJ, Tagare HD. Quantifying the local resolution of cryo-EM density maps. Nat Methods. 2014;11:63–5.

79. Campbell MG, Cheng A, Brilot AF, Moeller A, Lyumkis D, Veesler D, et al. Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. Structure. 2012;20:1823–8.

80. Brilot AF, Chen JZ, Cheng A, Pan J, Harrison SC, Potter CS, et al. Beam-induced motion of vitrified specimen on holey carbon film. J Struct Biol. 2012;177:630–7.

81. Wriggers W, Milligan RA, McCammon JA. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. J Struct Biol. 1999;125:185–95.

82. Vasishtan D, Topf M. Scoring functions for cryoEM density fitting. J Struct Biol. 2011;174:333–43.

83. Cragnolini T, Sweeney A, Topf M. Automated Modeling and Validation of Protein Complexes in Cryo-EM Maps. Methods Mol Biol. 2021;2215:189–223.

84. Shatsky M, Hall RJ, Brenner SE, Glaeser RM. A method for the alignment of heterogeneous macromolecules from electron microscopy. J Struct Biol. 2009;166:67–78.

85. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 4:486–501.

86. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 2:213–21.

87. Shen M-Y, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006;15:2507–24.

88. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015;10:845–58.

89. Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods. 2020;17:665–80.

90. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.

91. Roseman AM. Docking structures of domains into maps from cryo-electron microscopy using local correlation. Acta Crystallogr D Biol Crystallogr. 2000;56 Pt 10:1332–40.

92. Trabuco LG, Villa E, Mitra K, Frank J, Schulten K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. Structure. 2008;16:673–83.

93. Croll TI. ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. Acta Crystallogr D Struct Biol. 2018;74 Pt 6:519–30.

94. Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. 2018;27:14–25.

95. Piper SJ, Brillault L, Rothnagel R, Croll TI, Box JK, Chassagnon I, et al. Cryo-EM structures of the pore-forming A subunit from the Yersinia entomophaga ABC toxin. Nat Commun. 2019;10:1–12.

96. Pandurangan AP, Topf M. RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps. Bioinformatics. 2012;28:2391–3.

97. Pandurangan AP, Topf M. Finding rigid bodies in protein structures: Application to flexible fitting into cryoEM maps. J Struct Biol. 2012;177:520–31.

98. Locke J, Joseph AP, Peña A, Möckel MM, Mayer TU, Topf M, et al. Structural basis of human kinesin-8 function and inhibition. Proc Natl Acad Sci U S A. 2017;114:E9539–48.

99. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7:95–9.

100. Sobolev OV, Afonine PV, Moriarty NW, Hekkelman ML, Joosten RP, Perrakis A, et al. A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. Structure. 2020;28:1249–58.e2.

101. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007;35 Web Server issue:W375–83.

102. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta

Crystallogr D Biol Crystallogr. 2010;66 Pt 1:12–21.

103. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 2008;71:261–77.

104. Studer G, Biasini M, Schwede T. Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinformatics. 2014;30:i505–11.

105. Moriarty NW, Grosse-Kunstleve RW, Adams PD. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. Acta Crystallogr D Biol Crystallogr. 2009;65:1074–80.

106. van Zundert GCP, Moriarty NW, Sobolev OV, Adams PD, Borrelli KW. Macromolecular refinement of X-ray and cryoelectron microscopy structures with Phenix/OPLS3e for improved structure and ligand quality. Structure. 2021;29:913–21.e4.

107. Robertson MJ, van Zundert GCP, Borrelli K, Skiniotis G. GemSpot: A Pipeline for Robust Modeling of Ligands into Cryo-EM Maps. Structure. 2020;28:707–16.e3.

108. Vant JW, Lahey S-LJ, Jana K, Shekhar M, Sarkar D, Munk BH, et al. Flexible Fitting of Small Molecules into Electron Microscopy Maps Using Molecular Dynamics Simulations with Neural Network Potentials. J Chem Inf Model. 2020;60:2591–604.

109. Lahey S-LJ, Rowley CN. Simulating protein–ligand binding with neural network potentials. Chem Sci. 2020;11:2362–8.

110. Tang G, Peng L, Baldwin PR, Mann DS, Jiang W, Rees I, et al. EMAN2: An extensible image processing suite for electron microscopy. J Struct Biol. 2007;157:38–46.

111. Heymann JB, Belnap DM. Bsoft: image processing and molecular modeling for electron microscopy. J Struct Biol. 2007;157:3–18.

112. Zivanov J, Nakane T, Forsberg BO, Kimanius D, Hagen WJH, Lindahl E, et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. Elife. 2018;7:e42166.

113. Saur M, Hartshorn MJ, Dong J, Reeks J, Bunkoczi G, Jhoti H, et al. Fragment-based drug discovery using cryo-EM. Drug Discov Today. 2020;25:485–90.

Chapter 2

A novel approach to fitting small molecules in cryo-EM maps reveals insights into GSK-1 inhibition of human kinesin 5.

Background

Kinesins are a superfamily of proteins with the striking ability to turn chemical energy into mechanical force that drives kinesin walking along microtubules, driven primarily by interactions between the kinesin motor domain and microtubules. This is exploited in a variety of cellular processes, such as axonal transport and mitosis. A structural understanding of these proteins is vital to our understanding of the cellular process in which kinesins take part in, as well as diseases arising from their dysfunction.

Kinesin structure, function and diversity

To date a large number of kinesin subfamilies and family members have been identified, with the subfamily classification approximately correlating with function [1]. It is apparent that they all share a fundamental set of structural elements that enables kinesin function. This includes a central-stalk domain, a tail domain, and arguably the most conserved and well studied domain is the kinesin motor domain (Figure 1) (the site of nucleotide binding and hydrolysis). To generate the mechanical forces necessary to traverse microtubules a nucleotide exchange/hydrolysis reaction takes place [2]. Where initially ADP is bound and the motor domain exhibits a low affinity for microtubules, this ADP is exchanged with an ATP, with the motor domain briefly transitioning through a no nucleotide state. The ATP bound motor domain has been shown to adopt a high affinity for microtubules. The ATP is then hydrolyzed to ADP and P^i releasing the kinesin motor domain from microtubules, and the cycle repeats.

Early structural data for the conformational characteristics of the kinesin motor domain came from structures solved by X-ray crystallography, of the Kinesin-1 motor domain (PDB ID: 1BG2, 1.8 Å) in complex with Mg^{2+} and ADP [3]. The structure revealed a core β-sheet structure composed of eight strands. ADP was seen to bind within a pocket at the front of the central β-sheet flanked by the α3-helix and the helix-loop-helix motif of the α2-helix. Sequence data revealed a phosphate binding loop (P-loop) located at the N-terminal end of the α2-helix. This P-loop was seen to contain the Walker-A motif (GxxxxGK(S/T)) necessary for coordination of α- and β-phosphates of bound nucleotides. Additionally, a Mg^{2+} ion was

seen to have coordinate interactions, with the β-phosphate of ADP, Thr-92 of the Kinesin-1 P-loop, and two water molecules.

Since the deposition of this structure a wealth of kinesin family members have been solved bound to ADP including, the kinesin-5 family member Eg5 (PDB ID: 1II6) [4], the Kinesin-7 family member CENP-E (PDB ID: 1T5C)[5] and the kinesin-8 family member KIF18A (PDB ID: 3LRE) [6], all of which were seen to have a conserved secondary structure element (SSE) architecture (Figure 1).

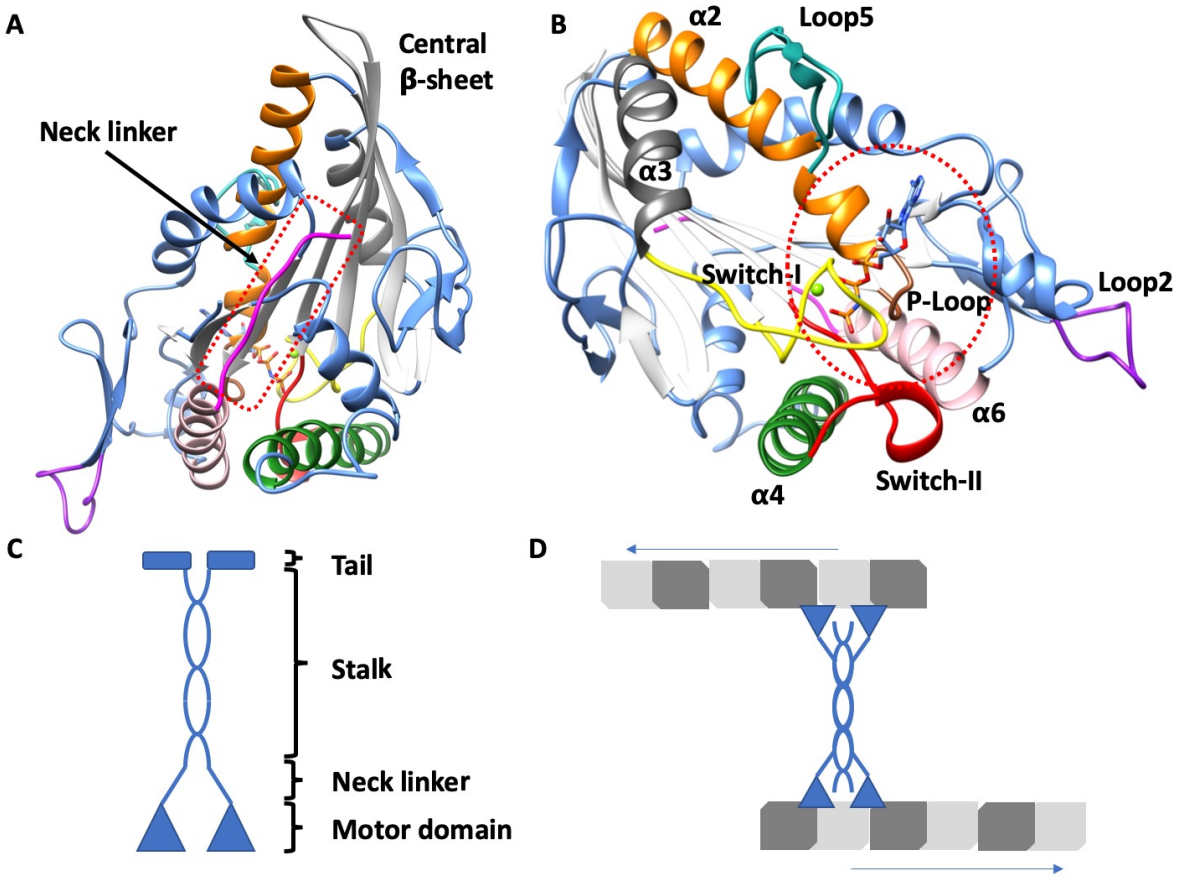


Figure 1. *A molecular model of the Kinesin Eg5 motor domain and neck linker in complex with AMPPNP and Mg^{2+}. **(A)** A front view of the Eg5 motor domain with SSEs highlighted. (a2-helix, orange; a3-helix, grey; a4-helix, green; a6-helix, pink; loop-2, purple; loop-5 turquoise; P-loop, brown; switch-I, yellow; switch-II, red;) The AMPPNP nucleotide and Mg^{2+} ion can be seen in the nucleotide binding pocket (red dashed circle). **(B)** An alternate view of the Eg5 motor domain with SSEs highlighted (Neck linker, magenta; central β-sheet, white). **(C)** A cartoon representation of a full length Kinesin with tail, stalk,neck linker, and motor domains shown. **(D)** A cartoon representation of the role of the Eg5 tetramer sliding microtubules (light and dark grey) apart during spindle formation. Figure partial recreated from [20].*

One of the first high-resolution structures of the kinesin motor domain in an ATP like state comes from X-ray crystallographic studies that solved the structure of Eg5 in complex with

Adenylyl-imidodiphosphate (AMPPNP, A non-hydrolysable ATP analogue) and Mg^{2+} at 2.2 Å (PDB ID: 3HQD) [7]. It is worth noting that this structure was solved in the absence of microtubules making it harder to relate structure to function.

This structure was seen to adopt a similar SSE arrangement, with a few subtle differences. Most notably was at the α3-helix, where compared with the ADP bound state, the C-terminal region of the helix was unwound and rotated approximately 11° relative to the central β-sheet. This appeared to have the effect of lengthening the switch-I loop, that was seen in a closed state interacting with the switch-II region, which in turn interacted with the bound nucleotide. These structural changes were seen to be propagated to the α4 helix, where the helix was shifted relative to the α6-helix allowing the neck linker region (that connects the motor domain to the central stalk) to dock along the side of the central β-sheet. Additionally, the helix was seen to be elongated in the AMPPNP bound state when compared to ADP bound structures [4].

Although the conformations of a multitude of kinesins in the ADP [3–6] and ATP [7–10] bound state have been characterised, the data reported for the no nucleotide (apo) state is sparse, with only four deposited structures reporting the apo motor domains, the *P. falciparum* Kin1 motor domain (PDB ID: 1RY6, 1.6 Å) [11], human kinesin-1 motor domain (PDB ID: 4LNU, 2.19 Å) [12], *M. musculus* KIF5C motor domain (PDB ID: 3J6H, 8.1 Å) [13], and the *C. elegans* Kinesin-6 motor domain (PDB ID: 5X3E, 2.61 Å) [14]. However, from the reported structures it is clear that the no nucleotide state differs from both ADP and ATP-like states. Most notably is that the extensive interaction network between switch-I, switch-II, P-loop and nucleotide is not possible due to the absence of a nucleotide, with most structures reporting the presence of only an ion within the nucleotide binding pocket. Additionally, the structures of switch-I and switch-II appear to adopt a more disordered conformation characteristic of an ADP bound state (with the exception of the *P. falciparum* Kin1 motor domain where switch-II adopts an unusual short helical arrangement [11]), whilst in all reported structures the α4-helices adopts an extended conformation similar to that of the ATP-like state.

For the processive march along microtubules it is necessary for kinesins to coordinate the action of two motor domains, connected via neckliner and stalk domains. The key step for force generation has been reported to be at the stage where both heads are bound to microtubules simultaneously [15], with the leading motor domain in the apo state and the trailing motor domain in the ATP bound state. In the ATP bound state the neck linker is docked along the central β-sheet creating tension between the two heads [16]. Upon ATP hydrolysis and P^i release the motor domain adopts a low affinity ADP bound conformation, where the neck linker exhibits a disordered conformation releasing the trialling motor domain and the neck linker tension. This is widely regarded as the general mechanism for microtubule force generation.

This mechanism along with sequence variability between sub-family members is the foundation that allows the kinesins to accomplish a wide array of molecular chores. Including

neuronal cell vesicle transport [17], intraflagellar transport [18], and spindle assembly [19] during cell division.

The mitotic Kinesins as drug targets

A subset of Kinesins have been shown to play a pivotal role within cell division, namely the regulation and organisation of spindle formation and dynamics. As such, there has been much interest in developing drugs that target these Kinesins as therapeutic interventions for cancer. One major driving force behind this is the severe side effects exhibited by drugs such as Taxol that targets microtubules directly [21]. Taxol will target tubulin in most cell types and has been associated with unwanted side effects such as neuro-cytotoxicity [22]. The hypothesis is that by targeting Kinesins that are specifically expressed during mitosis, kinesin inhibitors will be better tolerated.

One mitotic kinesin that has received much attention in recent years is kinesin-5 (Eg5, A.K.A Kinesin Spindle Protein (KSP)). Full length Eg5 is a 'dumbbell' shaped tetramer with a duo of motor domains at each end (Figure 1). It has been shown to possess the ability to slide antiparallel microtubules apart, and plays an essential part in the formation of the spindle during cell division [19]. The first mitotic kinesin inhibitors were identified for Eg5 [23]. Most of these drugs act by binding an allosteric site between the α2-helix, loop-5 and α3-helix SSEs [24–26] (Figure 2). Perhaps the most well known of these drugs is monastrol [23] (Figure 2). An inhibitor shown to slow down ADP release, where the kinesin motor domain has a low affinity for the microtubules disrupting Eg5 processivity [27].

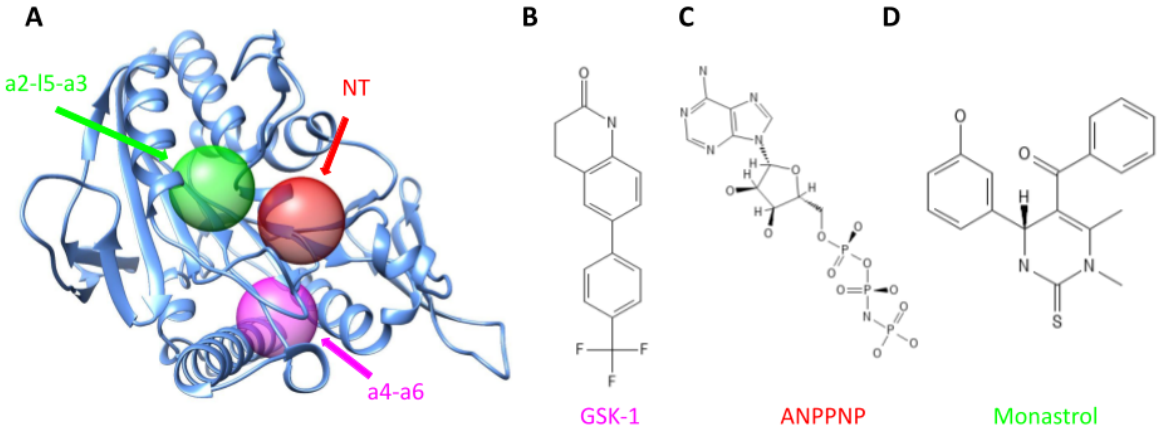


Figure 2. *Small molecule binding sites identified in the Eg5 kinesin motor domain.* ***A****. A model of the kinesin motor domain (blue) showing the nucleotide binding site (NT, red), along with the α2-Loop 5-α3- (a2-l5-a3, green) and the α4-α6-allosteric binding sites (a4-a6, fuchsia).* ***B****. A chemical schema of the drug GSK-1 known to bind the α4-α6-allosteric binding site.* ***C****. A chemical schema of the small molecule AMPPNP known to bind the*

nucleotide binding site. ***D****. A chemical schema of the drug monasterol known to bind the α2-Loop 5-α3-allosteric binding site.*

In vitro, monastrol showed promise resulting in mitotic arrest, yielding a characteristic monopolar spindle phenotype, based on which it is named, and reducing Eg5 motility in gliding velocity assays [23]. Since the discovery of monastrol many Eg5 inhibitors have been developed and undergone clinical trials [28, 29]. Despite being well tolerated none have shown efficacy as monotherapies [28, 30]. One hypothesis for the lack of efficacy when compared to drugs such as Taxol, which target microtubules directly, is that Eg5 inhibitors may only have a therapeutic effect when Eg5 is expressed during mitosis. A second hypothesis is that, due to the essentiality of the mitotic process, there may exist some mechanistic redundancy within the kinesins, such that when Eg5 is inhibited another kinesin may be able to compensate. Evidence for this comes from in vitro studies using a HeLa cell line showed that when Eg5 activity was inhibited using a small molecule inhibitor and KIF15 over expressed, KIF15 was able to compensate for the loss of Eg5 in spindle formation [31].

A second allosteric site has been identified on Eg5 at the kinesin/microtubule interface between the α4-,α6- helices (Figure 2). Drugs that bind at this site include the bi-aryl compounds GSK-1 (Figure 2) [32] and PVZB1194 [33, 34]. These drugs have the novel pharmacological property of being ATP competitive. This traps the motor in a state that occludes nucleotide binding and promotes tight microtubule binding. This mechanism may provide a way to perturb Eg5 activity whilst avoiding the problems inherent with mechanistic redundancy.

The structure of the Eg5 motor domain in complex with PVZB1194 was solved by X-ray diffraction to a resolution of 2.8 Å [33]. This structure revealed that the drug is bound within the α4-,α6- helices pocket, with no nucleotide occupying the nucleotide binding pocket. Structural insights suggested that PVZB1194 may interact with residue tyrosine 104 located at the bottom of the central β-sheet around the p-loop. The authors suggest that this interaction propagates significant structural rearrangement to the nucleotide binding site, occluding the binding of nucleotide and ‘trapping’ the kinesin in a state with high affinity for the microtubule. However, much of the structure surrounding the binding site was not visualised including the switch I and II loops, the N-terminal portion of the α4 helix, loop 5 and p-loop regions. Additionally, the structure was determined in the absence of microtubules and as such offers no structural insight into the interplay between the kinesin, drug and microtubule. Therefore there is a clear need to gain further structural insights into this class of inhibitor.

The following investigation aimed to model the structure of Eg5 bound to GSK-1 in the presence of microtubules using 3D reconstructions from Cryo-EM images solved in the lab of Prof. Carolyn Moores at Birkbeck College. This was achieved using a modified methodology for flexible refinement of an atomic model of kinesin that utilised density difference maps for fitting the Kinesin-8 (Kif18A) specific inhibitor BTB-1 into a cryo-EM map at a resolution of 4.8 Å [35]. The authors reported three density maps for tubulin-bound kinesin, representing

three states of kinesin in either the AMPPNP (Phosphoaminophosphonic acid-adenylate ester; A non-hydrolysable ATP analogue (Figure 2)) bound state, the no-nucleotide state or the BTB-1-bound state (EMD:3780, 3778, 3803; PDB: 5ocu, 5oam, 5ogc, respectively). A difference map was created between the BTB-1-bound kinesin map and the no-nucleotide state, utilising the difference mapping methodology in TEMPy [36]. The difference density corresponded to areas of conformational change in the vicinity of the nucleotide-binding pocket but it also included a prominent peak between helix-α2 and helix-α3, which was unoccupied by the fitted atomic model. Remarkably, this region corresponded to one of the well characterised allosteric inhibitor binding sites in Kinesin-5 (Kif11). To fit the atomic model of BTB-1 in the difference map at this region a two-stage docking protocol was used. First a global search for ligand binding site was conducted using HADDOCK [37], where the top scoring conformations were contained within the α2-,α3-binding pocket. A second stage focused on this binding pocket, and BTB-1 was docked by consensus docking using HADDOCK and AutoDock Vina [38]. From the top scoring conformations, two were chosen to be equally likely based on the CCC of the conformations with both the difference and original map. The position of the ligand corresponded well with the difference map density as well as biochemical mutation data of residues surrounding the binding site.

Results

Estimating the resolution of the map

Single particle Cryo-EM images and 3D density map reconstructions of Human Eg5 motor domain and neck linker in the presence of the small molecule inhibitor, GSK-1, were conducted by Dr Alejandro Peña of the Moores Group (see materials for further details) (Figure 3A). The Fourier shell correlation (FSC) from two half-maps, using a cutoff of 0.143 estimates that reliable information is contained within the map up to 3.8 Å resolution (Figure 3C). It would be expected that at this resolution the smallest visible structures would be α-helices, β-sheets, and perhaps bulkier sidechains. This correlates well with visible features in the tubulin density region in the map. Clear density was seen at the kinesin/microtubule interface that corresponded to the α4-,α5-, and α6-helices as well as the loop-8 and loop-12 regions (Figure 4B). Additionally, density that corresponded to the cover neck bundle and neck linker is well defined with the neck linker seen docked along the edge of the central β-sheet (Figure 4A). The β-sheet 1 subdomain is also visible with loop-2 region protruding from the flank (Figure 4A). Loop-2 density was seen to connect to the surface of α-tubulin at low density thresholds. Switch-I density is clearly visible, however only discontinuous density was seen for switch-II, and very little density corresponding to the P-loop (Figure 4C). Interestingly, no nucleotide was seen within the nucleotide binding pocket, and no density was seen for the Loop-5 region at any threshold, indicating this region is highly flexible under these conditions (Figure 4C).

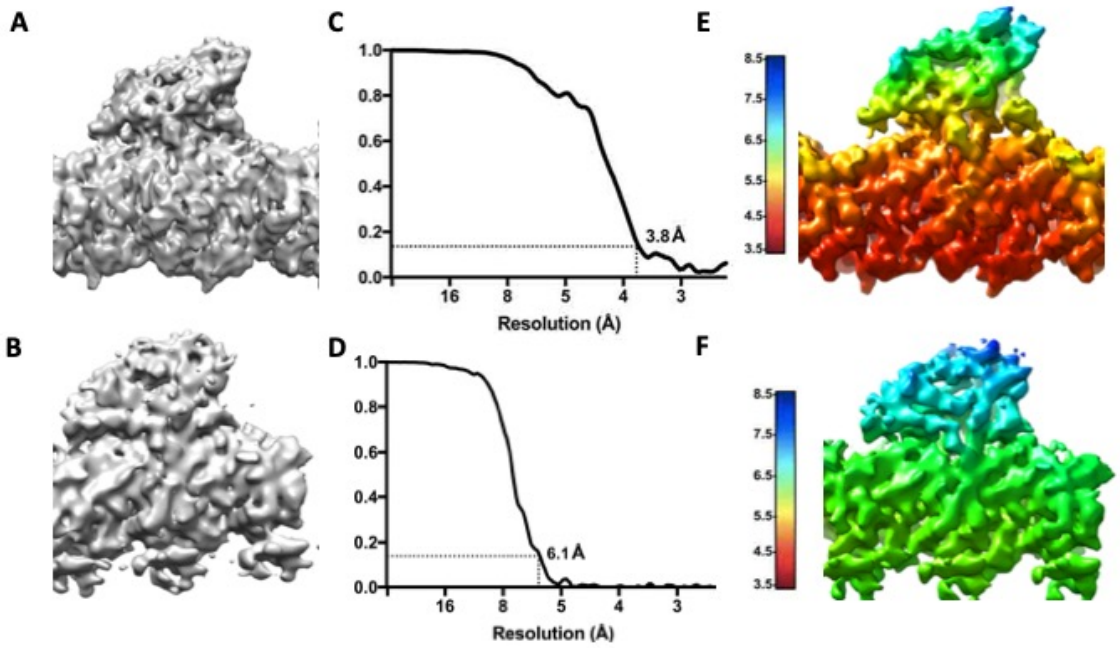


Figure 3. *The raw maps of Eg5 in complex with a- and β-tubulin and GSK-1 (**A**) or AMPPNP (**B**). The FSC plots of the two half maps for the GSK-1 bound kinesin (**C**) and AMPPNP bound kinesin (**D**). The local resolution of both maps is also shown for the GSK-1 map (**E**) and AMPPNP map (**F**), local resolution (Å) estimates for both maps are shown using a rainbow colour scheme and the key is located to the left of both images. Figure adapted from [20].*

A plot of local resolution estimates (derived with the RELION software [39]) shows that a resolution gradient exists within the map. Where the highest resolution data comes from the microtubules themselves (3.5 Å - 5.5 Å) with the worst resolution density being located towards the top of the central β-sheet (6.5 Å - 7.0 Å) (Figure 3E). Local resolution estimates for the core of the kinesin range from 5.5 Å - 6.5 Å within the area of the nucleotide binding pocket. Whilst helices α-2 and α-3 are visible, the fact that there was very little density corresponding to the p-loop and no density that corresponded to a nucleotide within this pocket, indicated that there was conformational variation within the nucleotide binding pocket. This is an unusual finding and although currently there is no structure of human Eg5 in the no nucleotide state, structures from homologous species indicate that the p-loop in this state adopts a relatively fixed conformation. One structure determined by X-ray crystallography to a resolution of the *C. Elegans* mitotic kinesin-6 family member zen4 motor domain [14] indicated that the no nucleotide state has 3 key features, an open ATPase catalytic site, an extended α4- helix and an occluded nucleotide binding site. However, the P-loop was not seen to adopt a defined conformation. This indicated that the lack of density at the P-loop region was due to drug binding.

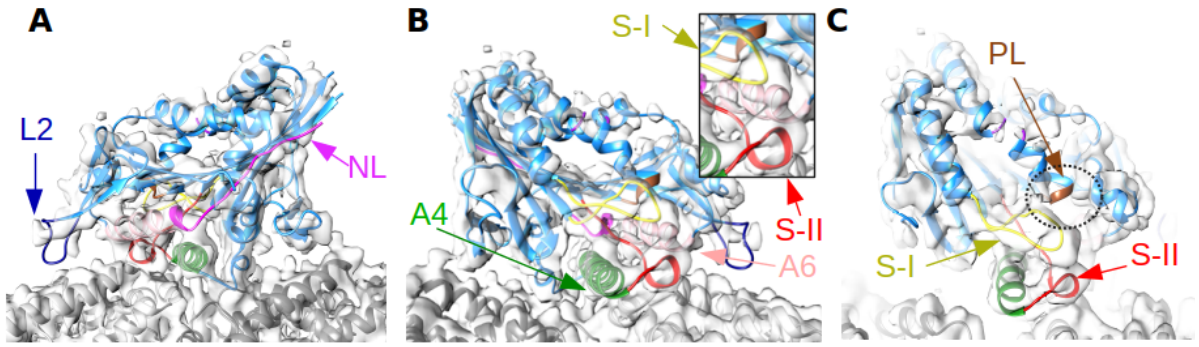


Figure 4. *Alternate views of an asymmetric unit from the GSK-1/Eg5 complex in the presence of microtubules (**A** and **B**) with the map contoured to a level that shows distinct SSEs. A refined model of the Eg5 motor domain is also shown in the map. Individual SSEs are colored fuchsia (NL, neck linker), green (A4, a4-helix), dark blue (L2, loop-2), pink (A6, a6-helix), red (S-II, switch-II), yellow (S-1, switch-I), and brown (PL, P-loop). Tubulin a and β subunits are also shown in light and dark grey, respectively. The neck linker can clearly be seen in a docked conformation (**A**), whilst at the microtubule interface (**B**) switch-I and switch-II density is clearly in a closed ATP-like conformation with connected densities from switch-I and switch-II (inset black box). Additionally the lack of nucleotide within the binding pocket is also shown (**C**, black dashed circle). Figure adapted from [20].*

A second reconstruction of the Human Eg5 in the Adenylyl-imidodiphosphate (AMPPNP; A non-hydrolyzable ATP analog) state in the presence of microtubules was also provided by Dr Alejandro Peña (Figure 3B). An FSC of the two half maps estimates that reliable information was contained in the map up to a resolution of 6.1 Å (Figure 3D). As with the GSK-1 bound map a resolution gradient was seen, with the best resolution being towards the bottom of microtubules and kinesin/microtubule interface (6.0 Å), towards the top of the kinesin central β-sheet the resolution drops off to between 7.0 Å and 8.0 Å (Figure 3F). Density corresponding to the α4-,α5-, and α6-helices is clearly defined (Figure 5B). The central β-sheet is seen clearly towards the microtubule/kinesin interface; however, towards the apex of the sheet, density is less well defined. It should be noted that this is in line with the drop in local resolution and indicates that this region is relatively flexible. The β-sheet 1 domain is clearly visible along with the protruding loop-2 (Figure 5A). Interestingly, in the AMPPNP state loop-2 appears further away from the surface of α-tubulin (Figure 5A, 5B). Clear density was visible for switch-I, discontinuous density was present for switch-II. Less obvious was clear density for the P-loop, however, clear density was seen within the nucleotide binding pocket corresponding to the nucleotide (Figure 5C). Additionally, density was seen connecting switch-I and switch-II, a feature characteristic of an ATP state (Figure 5C).

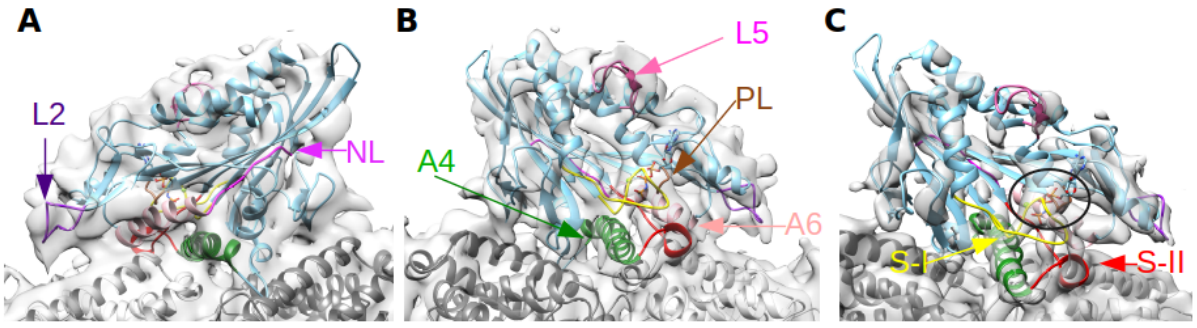


Figure 5. *Alternate views of an asymmetric unit from the AMPPNP/Eg5 motor domain reconstruction in the presence of microtubules (**A** and **B**). A refined model of the Eg5 motor domain is also shown in the map. Individual SSEs are colored fuchsia (NL, neck linker), green (A4, a4-helix), dark blue (L2, loop-2), salmon (A6, a6-helix), red (S-II, switch-II), yellow (S-I, switch-I), pink (L5, loop-5), and brown (PL, P-loop). Tubulin a and β subunits are also shown in light and dark grey, respectively. Additional density corresponding to AMPPNP within the nucleotide in the binding pocket was also seen (**C** black circle). Figure modified from [20].*

Calculation of atomic models

The aim of this study was to gain structural and functional insights into the mechanism of GSK-1 inhibition on Eg5. One way to do this is to obtain an atomic model of the Eg5 protein in complex with the inhibitor GSK-1, that accurately describes the data seen in the reconstructed Cryo-EM density map. Fortunately, the structure of the human Eg5 motor domain has been well characterised. High resolution X-ray crystallography structures, have identified the atomic coordinates of sidechain atoms, in the ATP [7] and ADP [4], whilst, structures obtained by Cryo-EM have clarified the relative positions of SSE's in a number of different nucleotide states, and in the context of the microtubule [40–42]. Additionally, a number of structures have been solved with the Eg5 motor domain complexed with small molecules. The most relevant being the biaryl inhibitor PVZB1194 in complex with the human Eg5 motor domain solved by X-ray diffraction at a resolution of 2.8 Å (PDB ID: 3WPN) [33]. As PVZB1194 shares some similarity to GSK-1 (both being biaryl inhibitors of Eg5), this structure seemed like a good starting point for modelling the GSK-1 bound structure.

An initial placement of the 3WPN atomic model into the section of the density map that corresponded to the motor domain of Eg5 was conducted using the '*fit-in-map*' function from Chimera [43]. This structure was seen to fit relatively well into the map with a CCC of 0.81. However, the structure of 3WPN was not complete in areas including, the helix-α3, helix-α4, switch I, switch II, loop5, P-loop and loop2 regions. It has been shown that having a starting conformation closer to the target structure can improve the speed and accuracy of density-based model refinement [44]. Therefore, a second template model that represented human kinesin-5 in an AMPPNP (adenylyl-imidodiphosphate, a non-hydrolysable ATP homolog) bound state (PDB ID: 3HQD) [7] was selected to model the missing regions.

This structure was chosen as the Eg5 motor domain has been shown to display distinct conformations as it undergoes the nucleotide cycle. Each state also displays a distinct kinetic profile, with the ADP bound state having a low affinity for microtubules, whilst the ATP and no-nucleotide state both show a higher affinity for the microtubule. Biochemical characterisation of GSK-1 showed that it exerts its effects by increasing the affinity of the Eg5 motor domain for microtubules. Thus, it was reasoned that the higher affinity AMPPNP bound state would be more representative of the drug bound state, with the two models having an average Cα-RMSD of 0.79 Å. Additionally, features within the map indicated the GSK-1 bound kinesin adopted a state similar to an ATP like state, for example density was seen connecting switch-I and switch-II (Figure 4C).

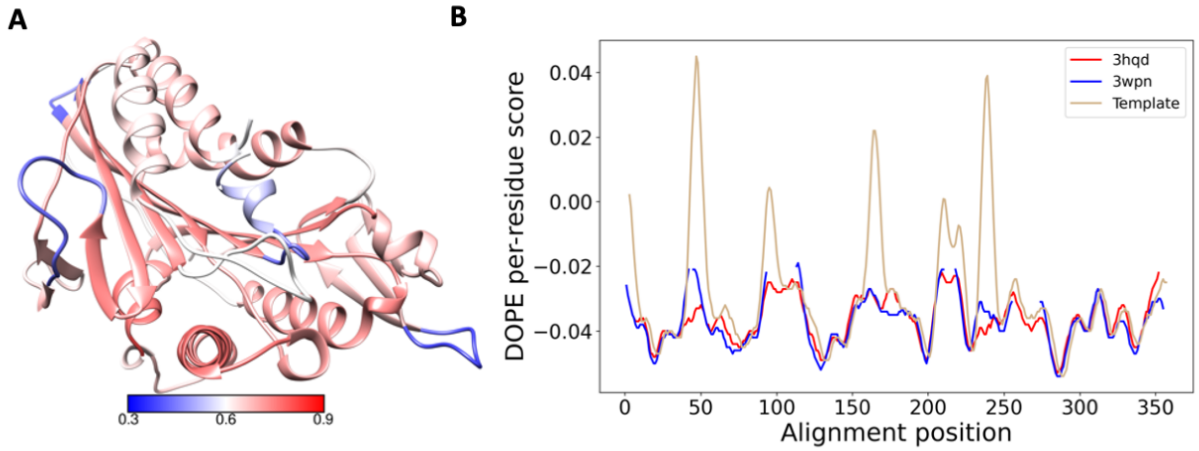


Figure 6. *(**A**) A 3D-representation of the SMOC plot of the initial hybrid model with the reconstructed GSK-1 bound Eg5 motor domain map. The SMOC score is shown per residue on the model by colour with a colour key underneath. It was seen that regions of low quality within the model mainly correspond to loop regions. (**B**) The per residue DOPE score for the AMPPNP bound Eg5 motor domain template (3HQD, red), the PVZB1194 bound Eg5 motor domain (3WPN, blue), and the calculated hybrid model (Template, Tan). A lower DOPE score represents better quality regions. Five regions of poor quality can be seen from the DOPE plot, all of which represent disordered regions within the model and correlated well with the SMOC plot.*

A hybrid model was built using MODELLER v9.21 [45], in which the local conformation in 3WPN was conserved, except missing regions, which were modelled based on 3HQD . The best initial model was selected using MODELLER DOPE score [46]. A plot of the MODELLER DOPE score per residue (Figure 6B) showed the calculated model adopted much of the fold characteristics from the aligned structures. However, several regions of low quality were identified from this plot. These regions were seen to correspond to disordered regions within the model and correlated well with regions that showed a low agreement with the map when it was assessed locally (Figure 6A).

For the AMPPNP atomic model the 3HQD kinesin motor domain structure was used as an initial model for refinement into the AMPPNP map. This model already had a good fit to the density with a CCC of 0.90 with the AMPPNP map.

For the modelling of microtubules in both models, a cryo-EM derived MT-GMPCPP structure of α- and β-tubulin was used (PDB: 6EVW) [47]. This is a standard model of microtubules, of the same species (Sus scrofa) and preparation protocol used in the Eg5/GSK-1 experiments.

Model refinement strategy

To more accurately identify density corresponding to Eg5, α-, and β-tubulin each protein was rigidly fit into the density map, using the Chimera '*fit-in-map'* function. The CCC of the hybrid model for the GSK-1-bound kinesin motor domain correlated well with the CCC of 3WPN alone, 0.81 and 0.81, respectively.

Density corresponding to each protein was segmented using the Segger tool [48] implemented in Chimera [43], and segmented densities were used to refine the models.

Once an initial rigid fit had been identified, the hierarchical flexible refinement protocol was used [44] to flexibly refine proteins into the map. In order to assess the local fit of the model in the density map the SMOC score [44] implemented in TEMPy [36] was used at each iteration.

One notable issue present during initial rounds of flexible fitting of the GSK-1 bound model into the map, was the over-fitting of regions with no corresponding density in the map. This was most noticeable at loop-5, where there was little density in the map that could reasonably account for this region of the atomic model, especially when compared to the AMPPNP bound model. Instead upon visual inspection it appeared that loop-5 was moving into density that corresponded to helix α-2. Additionally, this effect was seen with the four residue turn at the top of the central beta-sheet, whereupon flexible fitting the atomic model adopted a relatively compacted conformation in an effort to accommodate this region into the density map, this compacted conformation was thought to be overfit due to the presence of multiple clashing atoms. It was hypothesised that these regions were disordered, adopting multiple conformations in the GSK-1 bound state and thus do not appear in the final 3D reconstruction. To avoid these errors being introduced into the final model, both regions were removed from the model and the analysis was run again.

At each successive step of the refinement, the model was seen to agree better with the map compared to the previous step, with the average SMOC score of the initial model seen to be 0.80, the SSE rigid body step 0.82, and the final model 0.85 (Figure 7C). This increase in map/model agreement was more marked within loop regions that showed initial poor fits. In

particular the loop 2 region (Figure 7A, 7B) which, at low density thresholds, appears to contact the surface of α-tubulin.

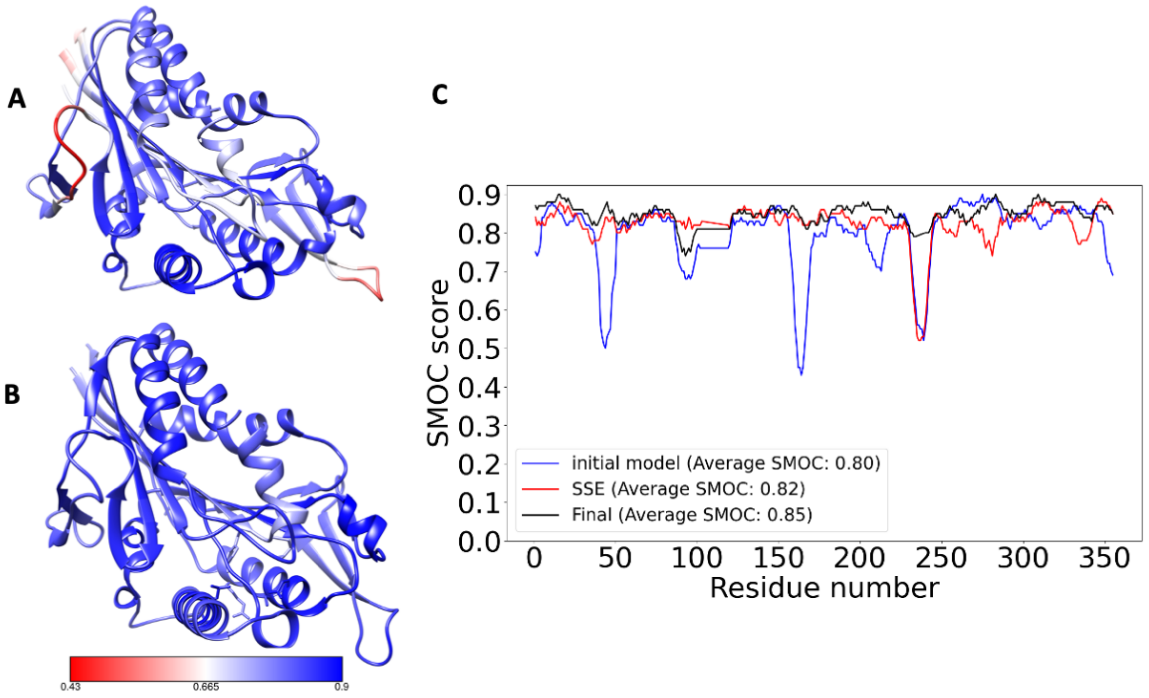


Figure 7. *The initial (**A**) and final (**B**) models for the GSK-1 bound Eg5 motor domain. The per residue SMOC plots are mapped onto both models with the colour key for both shown underneath **B.** It was seen that areas that showed the greatest improvements regarding the correlation with the map were the disordered loop regions. (**C**) The 2D-per residue SMOC plot for each stage of the refinement (initial model, Blue; SSE refinement red (solid); Final model, black). The average SMOC score at each stage is also shown in the legend.*
Figure adapted from [20].

For the AMPPNP model, at each refinement step the model map agreement increased. The initial, SSE and final model average SMOC scores of 0.86, 0.87, and 0.88, respectively (Figure 8C).

The final AMPPNP model correlated very well with the X-ray structure of the AMPPNP bound motor domain previously published. The Cα-RMSD between the two models was seen to be 3.38 Å (Figure 8D). The areas of worst agreement between the two models were seen to be the relative positions of alpha helices most notably α2-,α3-,α4-,α5-,α6-helices and the loop-5 region. One reason for this could be that the 3HQD model was derived from the Eg5 motor domain in the absence of microtubules. A second reason could be that the Cryo-EM reconstruction was at a relatively low resolution, and map errors from the reconstruction process lead to inaccurate SSE densities. However, the general agreement between the two models adds confidence to the AMPPNP reconstruction and atomic model. A similar trend in SMOC score was seen when MT's were subjected to the same flexible fitting protocol.

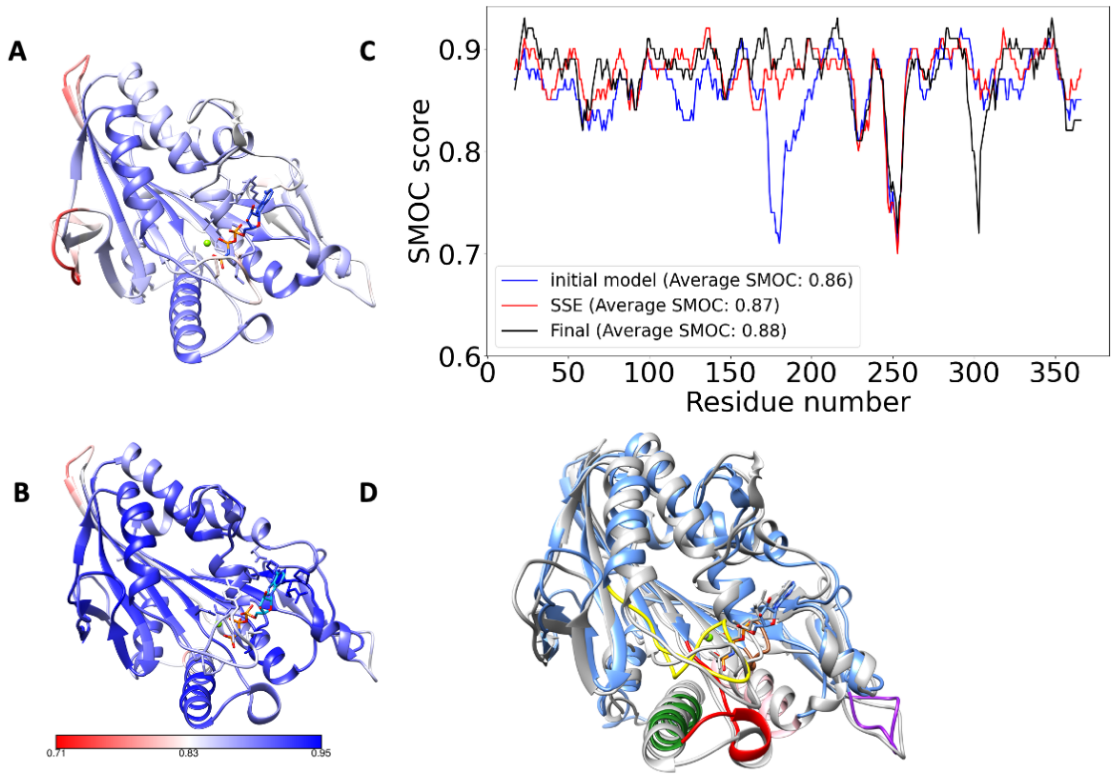


Figure 8. *A 3D representation of the SMOC score of the initial (**A**) and final (**B**) refined AMPPNP-bound Eg5 models with the reconstructions. Each residue is colored by its relative SMOC score. The colour key is also shown underneath the final model (**B**). (**C**) A 2D per residue SMOC plot for each step of the refinement, initial model (blue line), SSE refinement (red solid line), and final model (black line) is shown. The average SMOC scores for each refinement step is also shown within the legend. (**D**) shows a structural alignment between the 3HQD (grey) and final refined AMPPNP model (blue). SSEs are also highlighted (a4-helix, green; switch-II, red; switch-I, yellow; a6-helix, pink; loop-2, purple; p-loop, brown). Figure partially adapted from [20].*

Identification of the GSK-1 binding site

In order to identify the GSK-1 binding site itself, three independent methods were used, namely, density difference mapping, Meta-Pocket 2.0 [49] pocket prediction software, and blind docking with autodock vina [38]. Difference maps between the Eg5-GSK-1 reconstruction and the Eg5-AMPPNP reconstruction were calculated by using a difference mapping methodology implemented in TEMPy. However, due to the difference in resolution between the two maps, the Eg5-GSK-1 map was first blurred to the resolution of the Eg5-AMPPNP map, 6.1 Å. This resulted in a loss of some higher resolution features, which is especially pertinent when trying to identify the small molecule GSK-1.

Most differences identified corresponded to conformational differences between the two states, such as the loop 2, loop 5 regions, α3-helix and p-loop. One relatively large peak that

could not be accounted for by model, was seen between the α4- and α6- helices (Figure 9A). As a sanity check the reverse difference map was calculated by subtracting the GSK-1 map from the AMPPNP reconstruction. It was seen that differences in the reverse map correspond to areas of difference between the AMPPNP map and GSK-1 map, namely, the loop-2 region, loop-5 region (missing from the GSK-1 map) and the nucleotide within the binding pocket (Figure 9B). This gave confidence that the difference map methodology was producing meaningful results.

To further investigate this site as a potential drug binding pocket, Meta-pocket 2.0 [49] binding site prediction software was used. Three top clusters of pockets were seen, covering two previously known binding sites. The first was at the α-2-, α-3- binding pocket, shown to be the binding site for most known Eg5 small molecule inhibitors such as monastrol. The second located just below, around the p-loop is known to be the site of nucleotide binding and hydrolysis. Interestingly, two clusters of predictions overlapped with the site between the α-4-, α-6-helices, indicating that the site was solvent accessible (Figure 9C).

To ascertain which site, if any, could accommodate GSK-1 binding, blind docking with AutoDock Vina was conducted. Due to the stochastic nature of the Vina search algorithm and the relatively large search space, it was necessary to run the calculation a large number of times to fully explore the search space. A previous report that used AutoDock vina to identify the number of runs needed to precisely reproduce the correct binding mode suggested using at least 100 runs [50]. Thus, the docking was carried out a hundred times, each time with a search space encompassing the entire protein. The docking run yielded a total of 2000 conformations. The total number of conformations was filtered to only yield unique conformations by hierarchical clustering that favours the best scoring conformation (i.e. from the list of conformations in descending order from the best scoring: (1) the best score is set as an initial cluster centre. (2) the RMSD between the cluster centre and all other conformations is calculated.(3) Conformations that have an RMSD below the cutoff (< 2 Å) are removed from the list and said to be represented by the cluster centre. (4) The next best scoring conformation is set as a new cluster centre and the process is iterated again until the end of the list has been reached. For the Eg5 motor domain alone, 172 unique poses were seen with 2 % of these being located at the α-4-, α-6-helices pocket (Figure 9E). Interestingly, when the process was repeated with in the presence of MT's , 171 unique poses were identified, of which 32 % were seen to be located within the α-4-, α-6-helices pocket (Figure 9D).

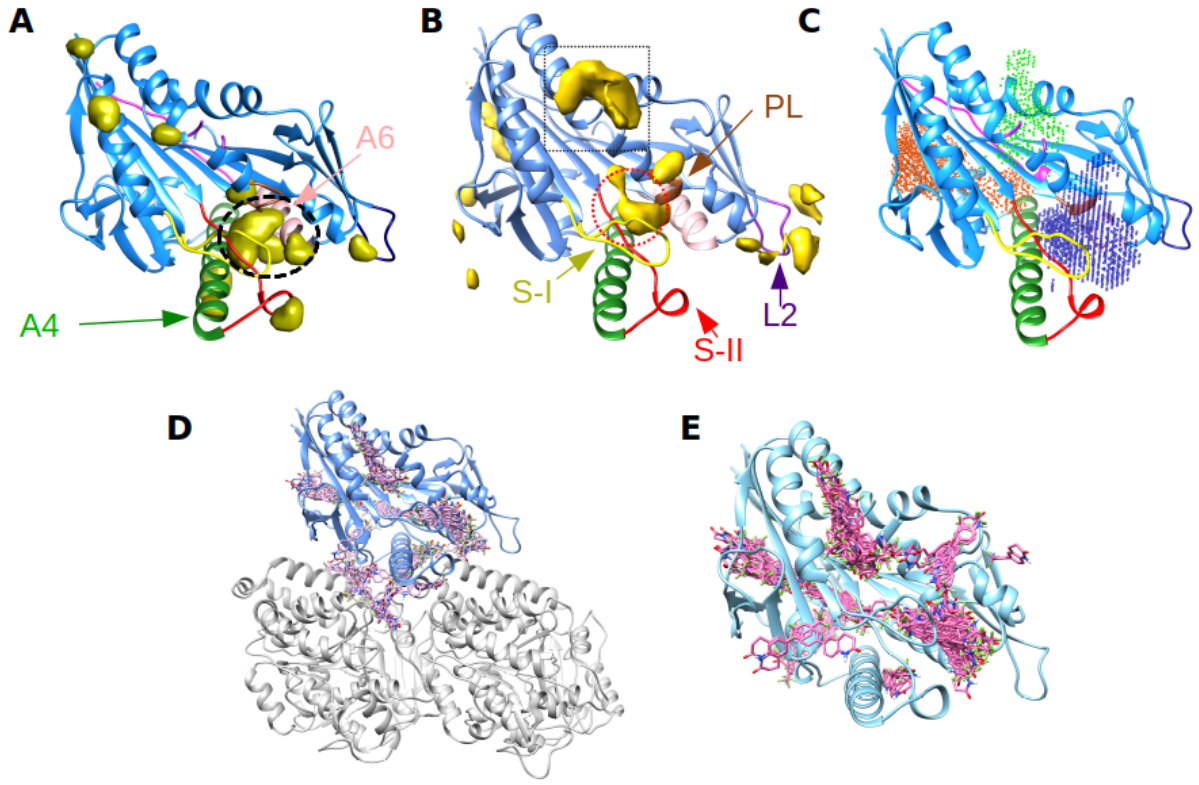


Figure 9. A. *The difference map calculated by subtraction of the AMPPNP map from the GSK-1 bound map. Several smaller densities can be seen that corresponded to variation in SSE conformation between the two models. One relatively large density peak (black dashed circle) can be seen between the a4- helix (A4, green) and the a6-helix (A6, pink).* ***(B)*** *The reverse map calculated as a subtraction of the GSK-1 map from the AMPPNP map. Several differences can be seen that correspond to SSEs that adopted different conformations following model refinement within the loop-2 region (L2, purple) and loop-5 region (black dotted box). Additionally, density corresponding to the AMPPNP ligand can be seen within the nucleotide binding pocket (red dotted circle) surrounded by the p-loop (PL, brown), switch-I (S-I, yellow), and switch-II (S-II, red).* ***(C)*** *The top three clusters from the MetaPocket-V2 binding site prediction server. Each point represents an individual prediction within a cluster. Clusters were seen to be located at the classical Eg5 inhibitor site (green cluster), at the nucleotide binding site (purple cluster), and behind the central β-sheet (orange cluster). Two clusters (purple and orange) were seen to contain points within the a4-,a6- binding pocket. The results of blind docking with the GSK-1 model using (pink) AutoDock Vina in the presence* ***(D)*** *and absence* ***(E)*** *of microtubules (grey). It is clear that in the presence of microtubules more conformations are predicted at the a4-,a6- binding pocket. Figure partially adapted from [20].*

The results of the blind docking correlated well with that of the Meta-pocket 2.0 assignment of pockets. With a majority of the solutions being located at either the α-2-, α-3- binding pocket, the nucleotide binding pocket or a solvent accessible region behind the central beta sheet. Only a small number of solutions were located between the α-4-, α-6-helices. However, this number increased when the MT models were added to the input structure. This indicated that the presence of MT's changes the chemical environment of this site. Therefore MT's

were added to all further binding site calculations. Other small clusters of GSK-1 binding were seen, however, did not correlate well with the difference density or meta-pocket 2.0 prediction software and thus were not investigated further.

Taken together the results showed that the site between the α-4-, α-6-helices was solvent accessible, and able to accommodate binding of GSK-1 especially in the presence of MT's. However, most importantly density was present within this site that couldn't be accounted for by the atomic model. Due to this, the α-4-, α-6-helices pocket was taken forward as the potential GSK-1 binding site.

Modelling the GSK-1 binding site

Using the information gained from the binding site identification step, a two stage consensus docking protocol was performed based on a previous methodology for modelling the BTB-1 binding site of human kinesin 8 [35]. In order to refine an atomic model of GSK-1 into the undescribed density within the binding site, GSK-1 was initially docked into the binding site using the three different scoring functions [51–53] implemented in GOLD [54] (see methods). For each run, a binding site radius of 12 $Å^3$ was used, the 'generate diverse solutions' option was on and the output was set to yield 100 conformations. For individual runs, redundant docking conformations (≤ 2 Å RMSD) were grouped and represented by the conformation with the best score (as in the blind docking step). Since it has been shown that consensus predictions can increase the accuracy of docking [55], only conformations predicted by all three scoring functions were analysed. Consensus conformations (≤ 2 Å RMSD) were then clustered and the CCC between each conformation and both the full map and difference map was calculated using TEMPy. The best conformation was selected as having the highest average CCC to both maps.

The best scoring conformations did not adequately fit the density (Figure 10A). It was hypothesised that this was due to the side chain positions within the binding site being incorrectly placed in the initial model, since the 3WPN structure was missing a large portion of this region. Therefore, the fit of the best scoring conformation was slightly refined into the density around the binding site using the Chimera *fit-in-map* function (Figure 10B), and the side chain atoms of residues that lined the binding site (within 5 Å of the ligand) were refined in the presence of the ligand using an all-atom refinement Flex-EM, while keeping the ligand rigid (Figure 10C).

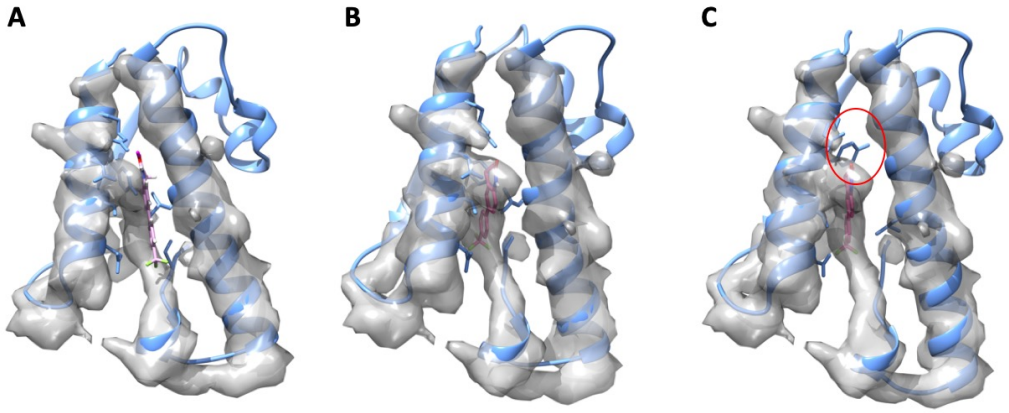


Figure 10. *(**A**) The top result from the first stage of consensus docking with GOLD. A section of the density corresponding to the a4-,a6-helices binding site is shown with the corresponding section of refined model fit to the map. The GSK-1 conformation (pink) can be seen located just outside the density unaccounted for by the model. This density can be seen occupied by the sidechains of the a6-helix residues. (**B**) The top conformation was better fit into the unaccounted density using the Chimera 'fit-in-map' function. However, this introduced severe clashes with the sidechains of the a6-helix. (**C**) To relieve clashes the side chains were refined with Flex-EM, whilst the ligand was held fixed. An example of where side chain and GSK-1 clashes have been resolved is highlighted (red circle).*

The second stage of ligand docking aimed to identify a ligand conformation that correlated with the density map. Again, three of the scoring functions implemented in GOLD were used along with AutoDock Vina to dock the ligand into the model from the previous step. For each GOLD run a radius of 6 Å was used, the 'generate diverse solutions' option was on and output was set to yield 100 conformations. For Vina a box-size of 12 $Å^3$ was used, and *num_modes* set to 20, all other settings were used as default. The results were analysed as in the focused docking stage. Conformations predicted by all four programs were individually assessed for the CCC to the full map using Chimera. From the 17 conformations predicted by all four softwares, two conformations were seen to account equally for the ligand density. One from AutoDock Vina (Figure 11A) (conformation 1) and one from GOLD Chemscore (conformation 2) (Figure 11C), showed a CCC of 0.62 (conformation 1) and 0.57 (conformation 2) with the difference map and 0.82 (conformation 1) and 0.80 (conformation 2) with the full map.

The resolution of the map and the pseudo-symmetry of GSK-1 (being a long planar molecule with electronegative groups at each end) makes distinguishing the correct conformation challenging.

Since the side chain positions within the binding site in the GSK-1 were not known, they were modelled mainly from X-ray crystallography derived models of Eg5 motor domain in complex with AMPPNP or PVZB1194. Due to this and in an attempt to better understand the differences between the two conformations, residues within 5 Å of any atom within GSK-1

were subjected to energy minimization in Chimera. Chimera uses AMBER force field parameters [56] for standard residues and assigns ligand parameters with the AMBER antechamber module. To reduce severe clashes 100 iterations of steepest descent minimization were calculated. This was followed by 10 steps of conjugate gradient minimisation to reach the energy minima (for both the initial step length was 0.02 Å).

Following this, the GSK-1/Eg5 interactions were assigned using both LigPlot+ [57] ,which uses HBPLUS [58] to assign H-bonds and hydrophobic interactions, and the Protein ligand interaction profiler (PLIP) [59], which uses geometric constraints to assign bond types. Ligplot+ predicted 'conformation 1' to make hydrogen-bonds with Gln106 and Arg355 (Figure 11B), whilst 'conformation 2' was predicted to make no specific interactions (Figure 11D). This correlated well with the assignment from PLIP, with the exception that the specific interactions with Gln106 in conformation 1 are predicted to be a halogen bond and a H-bond is predicted between GSK-1 and Arg402 of α-tubulin. Conformation 2 is predicted to make a π-Stacking interaction with Tyr352. Although these assignments appear to give an insight into possible drug mechanisms it is worth noting that the resolution of the map is not high enough to accurately determine the position of sidechains. However, computational analysis gives a strong indication of the location of the GSK-1 binding site.

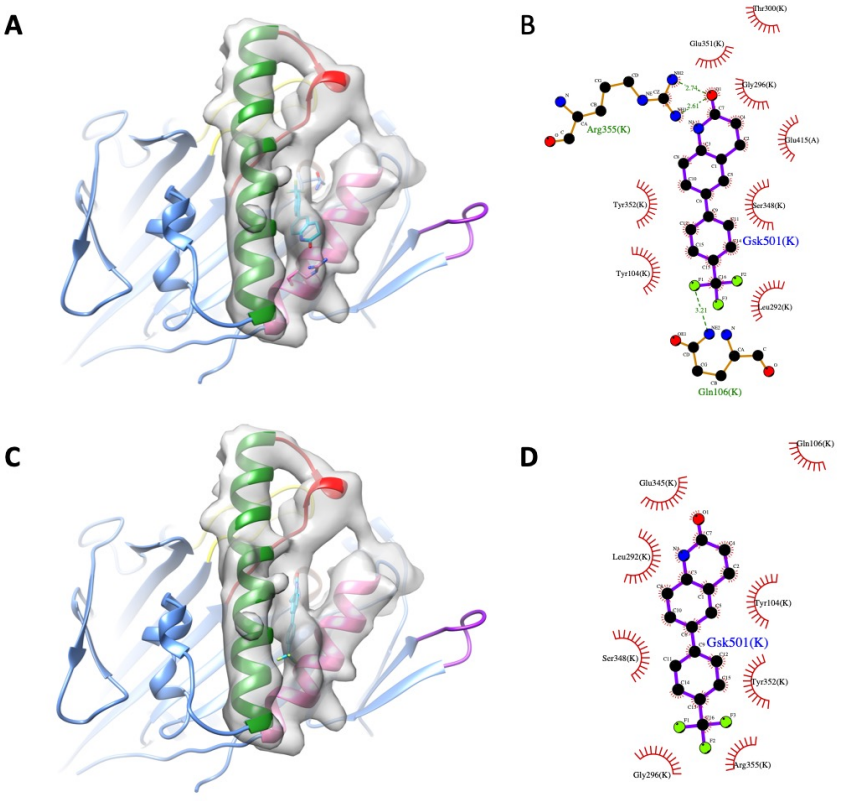


Figure 11. *(**A**) The final GSK-1 conformation predicted from the second stage of docking by AutoDock Vina (blue molecule). GSK-1 is within a section of density representing the binding site. Sidechains of ARG355 and GLN106 predicted to make specific interactions with GSK-1 are also shown. The Eg5 motor domain elements are highlighted, a4-helix (green),a6-helix (pink), switch-I (yellow), switch-II (red), p-loop (brown) and loop-2 (purple). (**B**) A 2D representation of the ligand (Ball and stick model) within the binding site. Residues predicted*

*by LigPlot+ to interact with GSK-1 specific interactions are shown as green dashed lines and hydrophobic interactions as red spiked semi-circles. (**C**) The final GSK-1 conformation (blue molecule) predicted by GOLD Chemscore. Again GSK-1 is shown within the binding site density and the Eg5 motor domain model is colored as in (**A**). (**D**) The 2D-representation of the GSK-1 conformation 2 within the binding site. Ligplot plus predicted no specific interactions. Hydrophobic interactions are represented by red spiked semi-circles. Figure adapted from [20].*

Discussion

Small molecule inhibitors that increase the affinity of Eg5 for the microtubule offer a new perspective with respect to perturbation of spindle function, and their use is still being investigated for efficacy in cancer therapy. However, there is little structural data with regard to these inhibitors and Eg5. This study presents a Cryo-EM structure of an Eg5 inhibitor of this type in the full context of the microtubule bound environment, and is the first structure for which GSK-1 has been visualised.

The importance of the structure can be exemplified by previous structures of Eg5 in complex with the biaryl compound PVZB1194 (PDB: 3WPN) [33], where much of the binding site was seen to be flexible and as such is not present in the corresponding atomic model. This includes much of the α4-helix, switch-I/II and the p-loop (Figure 12A), all of which are SSEs involved in the nucleotide cycle and can provide important mechanistic insights for small molecule inhibitors.

A comparison of the GSK-1/Eg5 structure with the atomic models in 3HQD [7] and the improved AMPPNP reconstruction indicates that in the presence of GSK-1 Eg5 adopts an ATP-like conformation. Most notably, is that density corresponding to the neck linker can be seen docked along the edge of the central β-sheet (Figure 4A). Additionally, density can be seen connecting switch-I with switch-II (Figure 4C), characteristic of an ATP like state. Interestingly, there is a difference in the position of loop-2 in the presence of GSK-1 (Figure 4A) or AMPPNP (Figure 5A), where in the presence of GSK-1, the loop 2 region was seen to be directed towards the top of α-tubulin, a feature not present in the AMPPNP reconstruction. The ATP state of the Eg5 motor domain has been shown to exhibit a relatively high affinity for MT's [2]. This may, in broad terms, explain the mechanism of action of GSK-1, whereby in the presence of GSK-1 the motor domain becomes 'trapped' in a conformational state with a high affinity for MT's and becomes static.

This proposed mechanism correlates well with previous kinetic characterization of GSK-1 in the presence of MT's indicating that GSK-1 was an ATP-competitive, MT-uncompetitive inhibitor with a K_i of 1.8 ±0.2 nM [32]. It was also seen that GSK-1 had no effect on the basale ATP-hydrolysis rate of Eg5 in the absence of MT's and the rate of inhibition increased as the concentration of MT's increased. This result was shown to be consistent with biochemical experiments conducted by Dr Alejandro Peña, using the purified Eg5 motor domain used for the Cryo-EM reconstructions presented in this investigation. The basal rate

of ATP-hydrolysis of Eg5 in the presence of MT's was seen to have a V_{max} of 1.22 ± 0.07 ATP/s and a $K_{0.5}$MT of 13.5 ± 3.3 nM (Figure 12C). This construct was seen to be inhibited in the presence of GSK-1 and MT's with an IC_{50} of 0.8 nM (Figure 12D). Taking this analysis one step further Dr Alejandoro Peña observed that in a multi-motor gliding assay, increasing concentrations of GSK-1 corresponded to a decrease of Eg5 motor domain driven MT motility, with an IC_{50} of 1.8 nM. It is also interesting to note that even with multiple washes MT gliding did not resume.

The reasons for this could be two-fold, the simplest of which is that the relatively high affinity of the drug for Eg5 could mean that washing was insufficient to displace the drug from the complex. Alternatively, it could be a consequence of where the drug binds, strong evidence was presented for the drug binding to the Eg5 motor domain between the α4- and α6- helices. Additional biochemical data conducted by Dr Alejandro Peña and a previous group [32] using mutant Eg5 motor domains supported this assertion (Figure 12B). Mutating residues within the binding site (Ile 299 to Phe and Ala 356 to Thr), reduced the basal rate of microtubule gliding, however conferred resistance to inhibition by GSK-1. However, since the α4-,α6- helices site is located at the MT/Kinesin interface, evidence provided from the blind docking experiments with AutoDock Vina indicated that the microtubule may contribute to the chemical environment of the binding site and may also be necessary for drug action, as GSK-1 was shown not to inhibit the basal rate of ATP hydrolysis in the absence of MT's. If this were the case it is hard to see how GSK-1 would escape the binding site during washing, although, this would now raise the question of how GSK-1 binds the site initially. Two possibilities arise from this, perhaps GSK-1 binds to Eg5 at some point during the mechanochemical cycle, most likely at the point nucleotide exchange period where the motor domain has a low affinity for microtubules, or the drug could first bind microtubules at the interface of the **α**-,**β**-tubulin interface and trap the motor domain during microtubule 'stepping', where upon binding of the Eg5 **α**4-helix to the microtubule/GSK-1 complex the Eg5 motor domain then becomes trapped in the microtubule bound conformation. The latter could have negative consequences for GSK-1 as a possible therapeutic intervention related to side effects. Drugs such as paclitaxel that bind and stabilise MT ends, favouring growth as a method to disrupt cell division [21], have been shown to have severe side effects [22], particularly in relation to neurotoxicity, as they can readily cross the blood brain barrier and stabilise the complex network of microtubules within neurons (and neuronal apoptosis leads to a wide variety of undesired effects). However experiments conducted by Dr Alejandro Peña on MT stability in the presence and absence of GSK-1 showed that GSK-1 alone had no effect on the stability of MTs. It was seen that when paclitaxel stabilised MTs were washed in non-paclitaxel containing buffer, MT's slowly depolymerised (1.50 ± 0.07 nm/s) this effect was abeited when paclitaxel was added to the wash buffer (0.24 ± 0.05 nm/s). When the motor domain was added to the wash buffer nucleotide state dependent stabilisation effects were seen (ATP 1.2 fold-, No nucleotide 2.3 fold-, AMPPNP 3.8 fold- increase in stability compared to wash buffer alone). The addition of GSK-1 with motor domain resulted in an 8 fold increase in stabilisation, however, no stabilisation effects were seen when GSK-1 alone was added to the buffer. These results indicated that GSK-1 did not have any stabilising effect on MTs. From this, it may be assumed that GSK-1 would not exhibit the same side effect

issues as paclitaxel. However, more work is needed to identify how GSK-1 binds and exerts its effects.

Although the resolution of the structure was not high enough to observe specific interactions. Following energy minimisation both PLIP [59] and Ligplot^{+} [57] predicted that the GSK-1 oxygen atom within the azacyclohexanone moiety of conformation 1 interacts with arginine 355 of the kinesin motor domain. This prediction is interesting as it precedes alanine 356, which when mutated to threonine was shown to confer resistance to GSK-1 inhibition (Figure 12B). One interesting speculation could be that the change in charge, as alanine is mutated to the more polar threonine residue, prevents GSK-1 forming the specific interactions necessary for inhibition. However, further experiments would be needed to confirm this.

The final GSK-1 model of kinesin showed a global CCC with the map of 0.85 indicating a good agreement of the data in the map, and correlated well with the average SMOC score, 0.85. However, the SMOC plot shows that two local regions are less well modelled compared to the rest of the structure. The first corresponds to the N-terminal region of the α2-helix and the P-loop. A visual inspection of the map showed that there was little density within this region with which to fit a model. The second region corresponds to the apex of the central β-sheet. Again there is little density within this region with which to fit the model. Additionally this is the region with the worst local resolution estimation. As such the confidence in the accuracy of these model regions is low.

The final AMPPNP model showed a global CCC to the map of 0.90 indicating that the final model is an accurate description of the map. The average SMOC score was seen to be 0.88 which correlated well with the global CCC. From the SMOC plots two regions appear to show a poor local fit with the model. The first being the apical turn at the top of the central β-sheet (residues 240-260). This was to be expected as the local resolution at this point falls off at this area in the map. Additionally, there is very little density at this point indicating that it may be a rather flexible region. A second area of poor fit was seen at the second region corresponding to loop-12. A visual inspection of this region in the map shows there is very little unambiguous density for which to fit the region. As such the confidence that these two regions are correctly modelled is low. However, the model does appear on a global level to represent the data reasonably well.

In addition to how well the model fits the data it is important to assess the geometry of the model to ensure it represents a model that is physically possible. MolProbity [60] is a commonly used validation software for such a purpose. The CαBLAM is useful for investigating the quality of backbone geometry especially of SSEs in models derived from lower resolution structures (> 2.5 Å, low resolution when compared to X-ray crystallography). When the GSK-1 Eg5 motor domain model was run through MolProbity no CαBLAM outliers were identified. Consistent with this, it was seen that there were no Ramachandran outliers seen, with 97.83 % residues having a favoured conformation and 2.17 % within acceptable limits. This indicated that the overall geometry of the SSE's both described the map well and was physically plausible. However, the geometry for sidechain

rotamers was not great with only 66.48 % of sidechains in favoured conformations, 12.67 % within an acceptable range and 20.89 % having unfavored rotamers. However, 39.06 % of these appear within loop regions which may be more likely to become distorted during the refinement process. As would be expected at this resolution there is low confidence in the side chain rotamers geometry within the model.

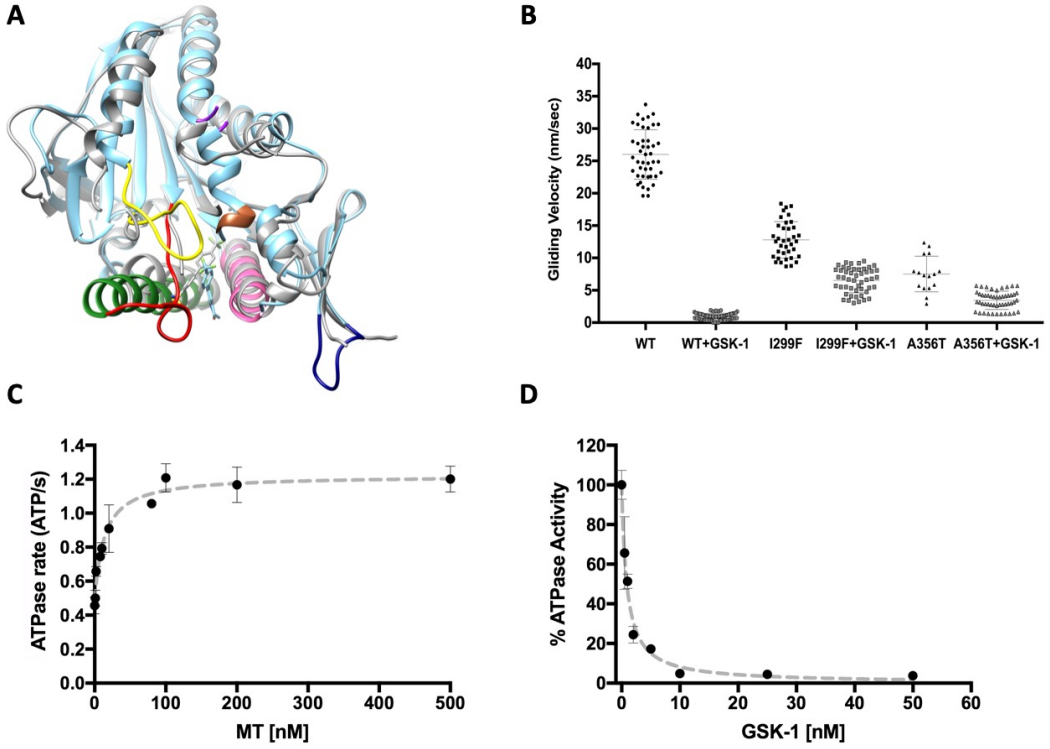


Figure 12. *(**A**) Shows a structural alignment between the refined GSK-1 model calculated in this investigation and the model of the Eg5/PVZB1194 complex (3WPN, grey). The general fold of our model correlated well with that of 3WPN. The SSEs of the GSK-1 bound Eg5 motor domain are highlighted (a4-helix, green; a6-helix, pink; switch-I, yellow; switch-II, red; p-loop, brown; loop-2, purple). (**B**) The results of gliding activity of wild type Eg5 and mutant constructs in the presence and absence of GSK-1. (**C**) The rate of ATP hydrolysis of the Eg5 motor domain protein in response to increasing concentrations of microtubules. Error bars = ± SD, n = between 4 and 12 for each point. (**D**) The ATP hydrolysis rate for the Eg5 motor domain construct in response to increasing concentrations of GSK-1. Error bars = ± SD, n = between 4 and 7 for each point. Figure adapted from [20].*

When the lower resolution AMPPNP model geometry was analysed with MolProbity a similar trend was seen with only 1.44 % of residues being CαBLAM outliers. In line with this 98.56 % of residues were seen to be Ramachandran favoured with only 0.29 % adopting an unfavored conformation. Again, the quality of side chain geometry was poor with only 66.45 % of side chain rotamers being favoured, 17.59 % being acceptable and 15.96 % adopting a poor conformation. However, 60.0 % of these occurred within loop regions.

Taken together the data presented here suggests that GSK-1 binds at a pocket between the α4-,α6- helices, located at the microtubule/Eg5 interface. This finding, based on Cryo-EM density fitting and molecular docking, is also supported with biochemical data (produced in the Moores lab. It was also seen that GSK-1 trapped the Eg5 motor domain in an ATP like conformation, whilst destabilising secondary structure elements within the nucleotide and 'classical' α2-,α3-helices binding site. This work provides an insight into the structural basis of how inhibitors such as GSK-1 trap Eg5 on microtubules, and demonstrated how molecular modelling integration with cryo-EM data can help identify small molecule binding sites.

Future perspectives

The methodology presented here represents the first steps in showing that consensus docking and experimental data from cryo-EM experiments can be utilised to accurately fit small molecules to intermediate resolution maps when high-resolution data is not present.

However, the workflow did highlight some important considerations. Firstly, the CCC alone was unable to distinguish between two conformations (approximately flipped 180° along the long axis). It is unclear from this investigation whether this is specific to GSK-1 (A long planar pseudo-symmetric ligand) or specific to the methodology as a whole. One reason this could occur as a result of the methodological principle is due to the fact that CCC alone does not take into account whether the position of the ligand is energetically favourable (in terms of electrostatics/hydrophobic interactions). This idea behind the original methodology [35] was that the docking software would take into account the energetic favorability of ligand conformations, guided by CCC. Although there is some evidence that docking programs are relatively good at generating correct or near-correct, solutions, especially when a consensus is taken into account [55]. Conversely, evidence suggests that the generation of correct solutions is dependent on model accuracy, of particular importance is the accurate positioning of side chains [61]. It is not always possible to achieve this level of accuracy when fitting an initial model to a cryo-EM map. Evidenced in this investigation, where although the structure of human Eg5 was solved to a high resolution by x-ray crystallography, with a similar ligand, much of the binding site was missing. Additionally it is unclear whether the binding site side chains for the inhibitor PVZB1194 would adopt the same positions as in the Eg5/GSK-1 complex.

In order to address these questions and further develop this methodology of small molecule fitting it will be necessary to systematically investigate the efficacy of docking programs in generating correct solutions, estimating free energy of the complex and ranking of solutions, in combination with and without CCC. Furthermore, it will be beneficial to use a method that refine ligand conformations into intermediate resolution density maps that also takes into account the chemical environment and uncertainty in the surrounding side chain positions.

Methods and software

Acquisition of the Cryo-EM maps used in this investigation was carried out by Dr Alejandor Peña. For further details see methods in [20].

Atomic model calculation

The initial model for the Eg5 motor domain/GSK-1 bound structure was derived from the atomic model for the human Eg5 motor domain in complex with PVZB1194 and in the absence of microtubules (PDB ID: 3WPN) [33]. This structure was seen to be incomplete in regions surrounding the PVZB1194 binding pocket including the helix-α3, helix-α4, switch I, switch II, loop5, P-loop and loop2 regions. Since it has been shown that having a starting conformation closer to the target structure improves the speed and accuracy of density based model refinement [44], and visual inspection of the cryo-EM map key features of an ATP like state were identified, missing regions within the model were modelled using an AMPPNP bound model of the Eg5 motor domain in the absence of microtubules (PDB ID: 3HQD) [7]. A sequence alignment was created between the Eg5 sequence to be modelled, 3WPN and 3HQD, such that 3HQD would only contribute missing regions to the model. MODELLER v9.21 [45] was used to generate 100 models. The best model was selected using MODELLER's DOPE score [46].

For the AMPPNP reconstruction the X-ray crystallography derived structure of the Eg5 motor domain in the AMPPNP bound state was used (3HQD) [7]. In both reconstructions, a cryo-EM derived MT-GMPCPP structure of *S. scrofa* α- and β-tubulin was used (PDB: 6EVW) [47].

As density corresponding to individual protein subunits could be easily determined in the map, the density corresponding to kinesin, α-, and β-tubulin was segmented using the Segger tool [48] implemented in Chimera [43]. Initial models were fit into the map using the Chimera '*fit-in-map'* function, and real space refinement was conducted hierarchically with Flex-EM [62]. Initially SSEs were used as rigid bodies with a cap_shift of 0.15 Å, followed by an all atom refinement with a lower cap shift of 0.1Å. At each iteration the local fit was assessed using the SMOC [44] score implemented in TEMPy [36].

Computation of difference maps

A difference map between the Eg5-GSK-1 reconstruction and Eg5-AMPPNP reconstructions and the reverse difference map were calculated using a local difference mapping method [63] implemented in TEMPy [36]. Briefly, the two maps were aligned along the microtubule density. Following this the method implemented in TEMPy, low pass filtered the resolution of the GSK-1 map to match that of the AMPPNP map (6.1Å) using a gaussian convolution.

Following this the amplitude of both maps were scaled at each resolution shell by a factor calculated as:

$$FT_{sc} = FT\,\frac{A1 + A2}{2 \cdot A1}$$

Where FT_{sc} is the scaling factor, FT is the original amplitudes in the shell, and A1 and A2 are the average amplitudes in the resolution shell for both maps.

Following scaling, the difference maps are calculated in real space by subtracting one map from another. A density threshold for the produced maps was also applied calculated by:

$$D_{frac} = \frac{\sum D1 - 2/\rho 1}{n}$$

Where Dfrac is the density threshold, D1-2 is the density difference in a given voxel, p1 is the density of scaled map 1 in the voxel and *n* is the number of voxels above the threshold.

A dusting step to remove smaller densities was also applied. Here the size of disconnected densities is calculated from the sum of voxels containing the density. The sizes are binned and bins that contain densities that satisfy:

$$\frac{n}{N} > 0.1$$

Where *n* is the number of disconnected densities in a bin and N is the total disconnected densities. Bins where the average density value is less than the mean density are removed.

The density map produced by substituting the AMPPNP map from the GSK-1 map was used to identify potential differences corresponding to the GSK-1 binding site. Additionally, the CCC calculated between the output conformations from molecular docking software and the difference map was used to identify possible GSK-1 binding modes.

Identification of GSK-1 binding sites

To identify possible binding sites for GSK-1 within the map the consensus between three methods was used. Initially the MetaPocket-V2 [49] server was used to identify possible druggable regions within the Eg5 motor domain. The refined atomic model of the Eg5 motor domain was used as input. For each output the individual scores are converted to Z-scores for comparison purposes, and the top three binding sites from each are pooled and clustered by spatial proximity to yield the consensus output clusters.

The second method was “blind” docking with AutoDock Vina [38]. The atomic model for GSK-1 was created from the canonical SMILES string
(O=C3CCc2cc(c1ccc(C(F)(F)F)cc1)ccc2N3) using Chimera [43] to assign 3D-coordinates. The input pdbqt files for the refined Eg5 motor domain with and without microtubules was created using the Chimera ‘*Vina’* function with the ‘*prep’* option set to True to ensure only the preparation steps were run. For each Vina run the exhaustiveness option was set to 10, the number of modes was set to 20 and the maximal energy difference set to 3 Kcal/mol. The box size was set to 65 $Å^3$ centred such that the whole Eg5 motor domain was encompassed by the box. To adequately explore the search space, the protocol was run 100 times using an in house bash script.

The output of each run (2000 conformations) were merged and redundant conformations (< 2.0 Å) were removed using a modified hierarchical clustering algorithm (in house python script) that weights each cluster by the conformation with the best energy score (lowest). Briefly, conformations were ordered by their energy score from best to worst. The algorithm moves iteratively down the list, assigning the first conformation as a cluster centre. Any conformation with a RMSD less than a cut-off value of 2.0 Å to the cluster centre is removed from the list and now represented by the cluster centre with the best energy score. Once the bottom of the list has been reached, the cluster centre is now set to the next conformation in the list and the process repeated again until the end of the list is reached.

A third method, used a visual inspection of the difference map calculated by subtracting the AMPPNP map from GSK-1 map to identify density not accounted for by the refined Eg5 motor domain model.

Modelling the GSK-1 binding site

In order to fit the GSK-1 model into the identified binding site a two-stage consensus docking protocol was used. In the first stage GSK-1 was docked into the identified binding site using three scoring functions provided within GOLD [54], namely, Chemscore, Goldscore, and ChemPLP. The input files for the receptor in the presence of microtubules were prepared in Chimera [43], using the ‘*DockPrep’* command. Hydrogens were added to the model. Chimera aims to add hydrogens to the model based on local environment and protonation states reasonable for physiological pH. Here, all histidines were protonated based on their local chemical environment, whilst all other standard residues were protonated based on expected protonation states at physiological pH regardless of the chemical environment. Partial charges were assigned using AMBER ff14SB parameters [56]. The output of this process was converted to mol2 format as used as the input for running GOLD. For each GOLD run a binding site radius of 12 Å was used, centred on a point that encompassed the secondary structure elements (α4-,α6-helices) that made up the binding site. The ‘generate diverse solutions’ setting was on (exclude redundant conformations < 1.5 Å from all others) and the output was set to generate 100 solutions.

The output from the individual scores were grouped with a cutoff ≤ 2.0 Å, using the same algorithm as in blind docking. It has previously been shown that consensus predictions can increase the accuracy of docking. Therefore, only conformations predicted by all three scoring functions (≤ 2.0 Å RMSD) were then clustered and the CCC between each conformation and the full map and difference map was calculated using an 'in-house' python script with the TEMPy package [36]. The best scoring conformation was seen as having the highest average CCC with both maps.

The best-scoring conformations did not adequately match the experimental data due to incorrectly fit sidechain positions within the binding site. The best scoring conformation was further refined into the density around the binding site using Chimera, and binding site atoms (≤ 5 Å of the ligand) were further refined using Flex-EM, with the ligand kept rigid.

The second focused stage of docking aimed at identifying a ligand conformation with the best possible CCC with the full and difference maps. To do this, again GOLD was used with the scoring functions: chemscore, goldscore and chemPLP. However, AutoDock vina was also run. Each GOLD run was executed as before; however the search radius was set to 6 Å, centred on the centroid of the refined ligand. For running Vina a box size of 12 $Å^3$ was used, and *num_modes* was set to 20. All other options were used as default. The results were initially analysed as in the first stage. Once consensus conformations had been identified, which were predicted by all four programs, the conformations were individually assessed by their CCC with the full map and difference maps, as in the previous step.

Once a viable conformation was identified, since the binding site region was a hybrid of the AMPPNP bound model 3HQD [7] and the PVZB1194 bound model 3WPN [33], the placement of sidechains would not have been optimally placed for GSK-1. Therefore, residue sidechains that had atoms within 5 Å of any atom of GSK-1 were energy minimised in Chimera using the '*minimise structure*' tool, with AMBER force field parameters [56] for standard residues and ligand parameters assigned with the AMBER antechamber module. To reduce severe clashes 100 iterations of steepest descent minimization was calculated. This was followed by 10 steps of conjugate gradient minimisation to reach the energy minima, for both the initial step length was 0.02 Å. Following this protein-drug interactions were assigned using both Ligplot+ [57] and the PLIP web server [59].

References

1. Miki H, Okada Y, Hirokawa N. Analysis of the kinesin superfamily: insights into structure and function. Trends Cell Biol. 2005;15:467–76.

2. Sadhu A, Taylor EW. A kinetic study of the kinesin ATPase. J Biol Chem. 1992;267:11352–9.

3. Kull FJ, Sablin EP, Lau R, Fletterick RJ, Vale RD. Crystal structure of the kinesin motor

domain reveals a structural similarity to myosin. Nature. 1996;380:550–5.

4. Turner J, Anderson R, Guo J, Beraud C, Fletterick R, Sakowicz R. Crystal structure of the mitotic spindle kinesin Eg5 reveals a novel conformation of the neck-linker. J Biol Chem. 2001;276:25496–502.

5. Garcia-Saez I, Yen T, Wade RH, Kozielski F. Crystal structure of the motor domain of the human kinetochore protein CENP-E. J Mol Biol. 2004;340:1107–16.

6. Peters C, Brejc K, Belmont L, Bodey AJ, Lee Y, Yu M, et al. Insight into the molecular mechanism of the multitasking kinesin-8 motor. EMBO J. 2010;29:3437–47.

7. Parke CL, Wojcik EJ, Kim S, Worthylake DK. ATP hydrolysis in Eg5 kinesin involves a catalytic two-water mechanism. J Biol Chem. 2010;285:5859–67.

8. Chang Q, Nitta R, Inoue S, Hirokawa N. Structural basis for the ATP-induced isomerization of kinesin. J Mol Biol. 2013;425:1869–80.

9. Cochran JC, Sindelar CV, Mulko NK, Collins KA, Kong SE, Hawley RS, et al. ATPase cycle of the nonmotile kinesin NOD allows microtubule end tracking and drives chromosome movement. Cell. 2009;136:110–22.

10. Ogawa T, Nitta R, Okada Y, Hirokawa N. A common mechanism for microtubule destabilizers-M type kinesins stabilize curling of the protofilament using the class-specific neck and loops. Cell. 2004;116:591–602.

11. Shipley K, Hekmat-Nejad M, Turner J, Moores C, Anderson R, Milligan R, et al. Structure of a kinesin microtubule depolymerization machine. EMBO J. 2004;23:1422–32.

12. Cao L, Wang W, Jiang Q, Wang C, Knossow M, Gigant B. The structure of apo-kinesin bound to tubulin links the nucleotide cycle to movement. Nat Commun. 2014;5:5364.

13. Morikawa M, Yajima H, Nitta R, Inoue S, Ogura T, Sato C, et al. X-ray and Cryo-EM structures reveal mutual conformational changes of Kinesin and GTP-state microtubules upon binding. EMBO J. 2015;34:1270–86.

14. Guan R, Zhang L, Su QP, Mickolajczyk KJ, Chen G-Y, Hancock WO, et al. Crystal structure of Zen4 in the apo state reveals a missing conformation of kinesin. Nat Commun. 2017;8.

15. Hyeon C, Onuchic JN. Internal strain regulates the nucleotide binding site of the kinesin leading head. Proc Natl Acad Sci U S A. 2007;104:2175–80.

16. Rosenfeld SS, Fordyce PM, Jefferson GM, King PH, Block SM. Stepping and stretching. How kinesin uses internal strain to walk processively. J Biol Chem. 2003;278:18550–6.

17. Hurd DD, Saxton WM. Kinesin mutations cause motor neuron disease phenotypes by disrupting fast axonal transport in Drosophila. Genetics. 1996;144:1075–85.

18. Scholey JM. Intraflagellar transport motors in cilia: moving along the cell’s antenna. J

Cell Biol. 2008;180:23–9.

19. Kapitein LC, Peterman EJG, Kwok BH, Kim JH, Kapoor TM, Schmidt CF. The bipolar mitotic kinesin Eg5 moves on both microtubules that it crosslinks. Nature. 2005;435:114–8.

20. Peña A, Sweeney A, Cook AD, Locke J, Topf M, Moores CA. Structure of Microtubule-Trapped Human Kinesin-5 and Its Mechanism of Inhibition Revealed Using Cryoelectron Microscopy. Struct Lond Engl 1993. 2020;28:450-457.e5.

21. Rowinsky EK, Donehower RC. Paclitaxel (taxol). N Engl J Med. 1995;332:1004–14.

22. Lipton RB, Apfel SC, Dutcher JP, Rosenberg R, Kaplan J, Berger A, et al. Taxol produces a predominantly sensory neuropathy. Neurology. 1989;39:368–73.

23. Mayer TU, Kapoor TM, Haggarty SJ, King RW, Schreiber SL, Mitchison TJ. Small Molecule Inhibitor of Mitotic Spindle Bipolarity Identified in a Phenotype-Based Screen. Science. 1999;286:971–4.

24. Yan Y, Sardana V, Xu B, Homnick C, Halczenko W, Buser CA, et al. Inhibition of a mitotic motor protein: where, how, and conformational consequences. J Mol Biol. 2004;335:547–54.

25. Talapatra SK, Anthony NG, Mackay SP, Kozielski F. Mitotic kinesin Eg5 overcomes inhibition to the phase I/II clinical candidate SB743921 by an allosteric resistance mechanism. J Med Chem. 2013;56:6317–29.

26. Talapatra SK, et al.. The structure of the ternary Eg5-ADP-ispinesib complex. Acta Crystallogr D Biol Crystallogr. 2012 Oct;68(Pt 10):1311-9.

27. DeBonis S, Simorre J-P, Crevel I, Lebeau L, Skoufias DA, Blangy A, et al. Interaction of the mitotic inhibitor monastrol with human kinesin Eg5. Biochemistry. 2003;42:338–49.

28. Shah JJ, Zonder JA, Cohen A, Bensinger W, Kaufman JL, Orlowski RZ, et al. The Novel KSP Inhibitor ARRY-520 Is Active Both with and without Low-Dose Dexamethasone in Patients with Multiple Myeloma Refractory to Bortezomib and Lenalidomide: Results From a Phase 2 Study. Blood. 2012;120:449–449.

29. Owens B. Kinesin inhibitor marches toward first-in-class pivotal trial. Nat Med. 2013;19:1550–1550.

30. Khoury HJ, Garcia-Manero G, Borthakur G, Kadia T, Foudray MC, Arellano M, et al. A phase 1 dose-escalation study of ARRY-520, a kinesin spindle protein inhibitor, in patients with advanced myeloid leukemias. Cancer. 2012;118:3556–64.

31. Tanenbaum ME, Macůrek L, Janssen A, Geers EF, Alvarez-Fernández M, Medema RH. Kif15 Cooperates with Eg5 to Promote Bipolar Spindle Assembly. Curr Biol. 2009;19:1703–11.

32. Luo L, Parrish CA, Nevins N, McNulty DE, Chaudhari AM, Carson JD, et al. ATP-competitive inhibitors of the mitotic kinesin KSP that function via an allosteric

mechanism. Nat Chem Biol. 2007;3:722–6.

33. Yokoyama H, Sawada J, Katoh S, Matsuno K, Ogo N, Ishikawa Y, et al. Structural Basis of New Allosteric Inhibition in Kinesin Spindle Protein Eg5. ACS Chem Biol. 2015;10:1128–36.

34. Matsuno K, Sawada J, Sugimoto M, Ogo N, Asai A. Bis(hetero)aryl derivatives as unique kinesin spindle protein inhibitors. Bioorg Med Chem Lett. 2009;19:1058–61.

35. Locke J, Joseph AP, Peña A, Möckel MM, Mayer TU, Topf M, et al. Structural basis of human kinesin-8 function and inhibition. Proc Natl Acad Sci U S A. 2017;114:E9539–48.

36. Farabella I, Vasishtan D, Joseph AP, Pandurangan AP, Sahota H, Topf M. TEMPy: a Python library for assessment of three-dimensional electron microscopy density fits. J Appl Crystallogr. 2015;48 Pt 4:1314–23.

37. van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016;428:720–5.

38. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem. 2010;31:455–61.

39. Scheres SHW. A Bayesian View on Cryo-EM Structure Determination. J Mol Biol. 2012;415:406–18.

40. Goulet A, Major J, Jun Y, Gross SP, Rosenfeld SS, Moores CA. Comprehensive structural model of the mechanochemical cycle of a mitotic motor highlights molecular adaptations in the kinesin family. Proc Natl Acad Sci U S A. 2014;111:1837–42.

41. Goulet A, Behnke-Parks WM, Sindelar CV, Major J, Rosenfeld SS, Moores CA. The structural basis of force generation by the mitotic motor kinesin-5. J Biol Chem. 2012;287:44654–66.

42. Bodey AJ, Kikkawa M, Moores CA. 9-Angström structure of a microtubule-bound mitotic motor. J Mol Biol. 2009;388:218–24.

43. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.

44. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42–9.

45. Šali A, Blundell TL. Comparative Protein Modelling by Satisfaction of Spatial Restraints. J Mol Biol. 1993;234:779–815.

46. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures.

Protein Sci. 2006;15:2507–24.

47. Manka SW, Moores CA. The role of tubulin-tubulin lattice contacts in the mechanism of microtubule dynamic instability. Nat Struct Mol Biol. 2018;25:607–15.

48. Pintilie GD, Zhang J, Goddard TD, Chiu W, Gossard DC. Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. J Struct Biol. 2010;170:427–38.

49. Huang B. MetaPocket: a meta approach to improve protein ligand binding site prediction. Omics J Integr Biol. 2009;13:325–30.

50. Hetényi C, Spoel D van der. Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci. 2002;11:1729–37.

51. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267:727–48.

52. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des. 1997;11:425–45.

53. Korb O, Stützle T, Exner TE. Empirical scoring functions for advanced protein-ligand docking with PLANTS. J Chem Inf Model. 2009;49:84–96.

54. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003;52:609–23.

55. Houston DR, Walkinshaw MD. Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context. J Chem Inf Model. 2013;53:384–90.

56. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015;11:3696–713.

57. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model. 2011;51:2778–86.

58. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol. 1994;238:777–93.

59. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res. 2015;43 Web Server issue:W443–7.

60. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 1:12–21.

61. Tuccinardi T, Poli G, Romboli V, Giordano A, Martinelli A. Extensive Consensus Docking Evaluation for Ligand Pose Prediction and Virtual Screening Studies. J Chem Inf

Model. 2014;54:2980–6.

62. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein structure fitting and refinement guided by cryo-EM density. Struct Lond Engl 1993. 2008;16:295–307.

63. Joseph AP, Lagerstedt I, Jakobi A, Burnley T, Patwardhan A, Topf M, et al. Comparing Cryo-EM Reconstructions and Validating Atomic Model Fit Using Difference Maps. J Chem Inf Model. 2020;60:2552–60.

Chapter 3

A new Program for Protein-Ligand interaction detection (ProPLID) utilising bond geometry

Background

The Protein data bank (PDB) [1] represents a wealth of structural information containing atomic models derived from X-ray crystallography, NMR, and cryogenic-electron microscopy experiments (cryo-EM). This data is important for advancing the field of structure based drug discovery, as atomic models of protein-ligand complexes can elucidate the mechanism of action of drugs by indicating the specific protein-ligand interactions necessary for drug action. This information can then be used to identify ways of improving drug potency, or designing new therapeutic interventions. Additionally, the importance of having high quality protein models in the apo state should not be understated. These models can reveal possible druggable sites within protein cavities, *in-silico* techniques such as molecular docking can then be used to identify small molecules that can bind such cavities.

A vast range of non-covalent interactions are important for molecular recognition of a drug to its protein target. One common method for the identification of individual interactions is to use geometric criteria. One recent study conducted a large scale statistical analysis using 11016 high resolution (≤ 2.5 Å) protein-ligand complexes reporting geometric data for a wide variety of interactions types including: hydrogen bonds, **π-π** stacks, cation-**π** interactions, halogen bonds and hydrophobic interactions, to name a few [2].

For such data to be useful rapid and accurate methods that identify the protein-ligand interactions that enable molecular recognition are needed. Many early softwares developed for this use were limited in the breadth of interactions predicted, such as HBPLUS [3] that predicts hydrogen bonds within an atomic model based on statistical observations on donor-hydrogen-acceptor (DHA) distance and angles using high-resolution structures from the PDB. Another early interaction prediction program LigPlot+ [4] built on the ideas of HBPLUS, by automatically generating 2-dimensional plots highlighting putative protein-ligand hydrogen bonds and hydrophobic interactions. A further method implemented in the scientific software MOE utilised empirical scoring functions to predict hydrogen bonds, hydrophobic interactions, metal ion complexes along with distance based approach for predicting **π-π** stacking interactions and cation-**π** interactions [5]. This approach was shown to be relatively successful on a benchmark of 12 protein-ligand complexes from the PDB,

however was implemented in commercial software adding an extra barrier to utility for researchers.

More recently, more open source software has been made available to the scientific community. One such method Arpeggio [6], used distance and geometric features to analyse a wide variety of protein-ligand interactions including, strong and weak hydrogen bonding, **π-π** stacking interactions, cation-**π** interactions, halogen bonds and metal ion complexes. One drawback to the methodology is that there is no easy way for researchers to adjust the parameters for detecting individual protein-ligand interaction types. This is an important consideration, since as the resolution of the data from which atomic models are derived worsens, the probability of modelling errors increases. Therefore, a method for relaxing criteria when detecting protein-ligand interactions is needed. One such software that addresses this problem is the Protein Ligand Interaction Profiler (PLIP) [7, 8]. PLIP is able to detect protein ligand interactions for hydrogen bonds, hydrophobic interactions, **π-π** stacking interactions, cation-**π** interactions, halogen bonds, and metal ion complexes. PLIP uses geometric criteria relating to bond types to identify interactions. It was shown to be reliable in predicting protein ligand interactions with a benchmark of 30 high resolution protein-ligand complexes. However, it is not clear how this reliability was assessed. In the paper true positive, false positive and false negative interactions for each complex in their benchmark are reported, derived based on interactions reported in associated papers. However, there were many interactions mentioned in the associated papers not reported in either category [7]. Additionally, the program identifies metal ion complexes however, no methodology as to how this is achieved is reported and the initial report states that metal ion complexes are not identified [7].

The importance of interactions such as strong hydrogen bonds [9, 10], **π-π** stacking interactions [11], halogen bonds [12] and hydrophobic interactions [13] in molecular recognition and lead optimization has been extensively studied. However, any reliable protein-interaction tool should include a wide variety of interactions, some of which are less well studied. One such example is the weak hydrogen bonds that can occur between hydrogens bonded to carbon and oxygen atoms with available lone pairs [9, 10]. Unlike strong hydrogen bonds that bridge strongly electronegative atoms via hydrogen, weak hydrogen bonds have been shown to involve atoms that are weakly electronegative such as carbon [14]. The importance of these interactions at the protein ligand interface was investigated in two studies that compared the occurrence of CH-O bonds with strong hydrogen bonds in 28 high resolution protein ligand complexes and small molecules in the Cambridge small molecule database [10] and data extracted from 251 high resolution X-ray structures from the PDB [9]. It was reported that CH-O bonding geometry was distinct from that of the strong hydrogen bonds, occurring over larger distances. Additionally, a larger number of weak hydrogen bonds were formed with binding site water molecules. Furthermore, the weak CH-O bond has been shown to play a significant role in the molecular recognition of kinase inhibitors [15]. A recent survey of 13,600 protein ligand CH-O bonds from protein-ligand complexes obtained from the PDB, reported general statistics for the

geometry of these bonds, with mean donor-acceptor distances of ≤ 3.6 Å and DHA angles of ≥ 130 ° [2].

This investigation reports a new software the Program for Protein Ligand Interaction Detection (ProPLID). ProPLID utilised geometric parameters to identify bonds taken from a systematic analysis of bond types from 11,016 protein ligand complexes from the PDB representing 750,873 independent interactions [2]. The software is implemented in a way that allows researchers a large amount of flexibility in setting geometric parameters, an important consideration when analysing interactions in lower resolution models. Additionally, it is pertinent to have a diverse range of softwares for most tasks in computational biology. Most current open source software is written with OpenBabel used to handle ligand models. ProPLID is the first example to use the small molecule package RDKit to achieve this. This report shows the accuracy of ProPLID for the identification of a wide variety of bond types using a benchmark of 35 high resolution protein-ligand complexes. Finally, the performance of ProPLID is compared with the open source software PLIP [7, 8].

The ProPLID Algorithm

The ProPILD software takes in at the very minimum a protein file in PDB format. Ligands can be supplied in multiple formats, either included as het atoms within the PDB, or separately as '.sdf' or '.mol2' files. By default the algorithm will identify all protein residues from the PBD files along with any solvent atoms and het groups. Binding sites are defined as a given cutoff radius from the centroid of supplied ligands. Atom typing for proteins is done using predefined criteria for each residue, for ligands atom typing is handled by the RDKit python package.

Hydrogen Bonds

Strong hydrogen bonds

To calculate hydrogen bonds, acceptors and donors within the binding site were first identified. The software identifies all oxygen and nitrogen atoms as potential donors/acceptors. Following this, the software calculates the available lone pairs and hydrogens on the atom available for accepting and donating to a hydrogen bond, respectively.

Nitrogen atoms that form ammonium ions are excluded from being hydrogen bond acceptors since the lone pairs necessary for accepting hydrogen bonds are not available. Additionally, Nitrogen atoms in amide groups are discounted as acceptors as their lone pairs are known to conjugate with the carbonyl system.

A list of candidate hydrogen bonds for each combination of acceptor/donor pairs is constructed. The distance between the acceptor and donor atoms is calculated, along with the bond angle around the donor, donor hydrogen and acceptor atoms. In the case where multiple

hydrogen atoms are bonded to the donor atom, the distance is calculated for each hydrogen atom and the bonds are treated individually. For bonds to make it to the candidate list, they must have geometric values greater than the cutoff values for hydrogen bonds. The geometric criteria for hydrogens bonds are the distance, and donor/hydrogen/acceptor angle.

Once a candidate list is generated, 'duplicate' bonds are removed i.e. bonds where atoms have the potential to be both donor and acceptors, and two candidate bonds are generated with donor/acceptor atom assignment flipped. During this step the bond with the greatest angle is kept. A second filter step is applied to ensure the predicted hydrogen bonds comply with the number of hydrogen bond donors and acceptors within the system. During this stage, the candidate bond list is sorted by distance with bonds with less distance between them given priority. The list is traversed by applying the bonds where there are available lone pairs and hydrogens. Once a bond is set, the lone pairs and hydrogens used to form the bond are no longer available to form other bonds with the rest of the system. Bonds are added to the system iteratively until the end of the candidate bond list is reached.

It is worth noting that for protein atoms any bonds that form within the protein are factored into the available lone pairs and hydrogens for an atom. For example for a backbone nitrogen in an α-helix, priority is given to the hydrogen bond formed to maintain the helical structure before considering bonds with ligand atoms. As far as we are aware this is a novel feature not implemented in other software such as PLIP.

Weak hydrogen bonds

The ProPLID algorithm also considers weak hydrogen bonds. Currently only weak hydrogen bonds that exist between carbon atoms with hydrogens and oxygen atoms are considered.
Weak hydrogen bonds were calculated in much the same way as for strong hydrogen bonds. Acceptors were any oxygen atom with an available lone pair, donors were any carbon atoms with an available hydrogen. The geometric criteria for weak hydrogen bonds was the same as for strong hydrogen bonds, with the exception of a minimum DHA angle criteria, and the algorithm progressed exactly as for strong hydrogen bonds.

When calculating the set of weak hydrogen bonds in the system, preference is given for strong hydrogen bonds to form first. This was done as weak hydrogen bonds between ligands and proteins are a much rarer occurrence than strong hydrogen bonds, the reason being that the bond itself is of lower energy, and the formation of a strong hydrogen bond will stabilise the complex more so than the formation of weak hydrogen bonds. As such weak hydrogen bonds are much more commonly formed between the ligand and binding site solvent.

Hydrophobic interactions

There are little geometric criteria in respect to hydrophobic interactions for which to consider, apart from the fact that the atom Van der Waals (vdW) radii must not overlap due to repulsive

forces between atoms. In ProPLID we employ information regarding the atom type and the local environment for which the atom is placed.

Potential hydrophobic atoms are considered to be carbon and hydrogen atoms that are only bound to other hydrophobic atoms. This list is restrictive and does not include other potentially, hydrophobic atom types such as the halogens. This was done to restrict the number of hydrophobic interactions reported to an accurate, but manageable level. As the number of hydrophobic interactions can easily outnumber the number of specific interactions reported for a protein ligand complex, this can make downstream analysis tricky.

Binding sites have local regions of heterogeneity with respect to hydrophobicity, with some regions being more hydrophobic and some more hydrophilic, hydrophobic interactions between ligand and protein atoms generally occur within these regions of hydrophobicity. In the absence of much meaningful geometric criteria for hydrophobic interactions, additional information regarding the local environment that ligand atoms are placed in was incorporated into the algorithm. To do this we used logP (a measure of the hydrophobicity of an atom) values of atoms within 6 Å of the ligand atom. LogP values were taken from XLOGP3 [16] and a cutoff for hydrophobicity was applied when considering possible hydrophobic interactions.

The algorithm first identifies a list of hydrophobic atoms within the protein and ligand. A candidate list of hydrophobic interactions is populated by passing through the list of ligand atoms and identifying all protein hydrophobic atoms within a hydrophobic distance cutoff. All identified interaction pairs are then added to the candidate list.

The vdW overlap between atom pairs is then calculated using the equation (Eq 1):

$$Eq\,1.\ VDW_{overlap} = d_{ij} - r_i - r_j$$

Where *dij*, is the distance between atom pair centres, and *ri* and *rj* are the vdW radii of atoms i and j. Values for vdW radii were taken from the United Atom Radii [17]. Maximum and minimum cutoff values for the vdW overlap are then applied, where the Maximum overlap being 0.0 Å and minimum values being negative, thus avoiding classifying overlapping atoms. Next, the logP of the local ligand environment is calculated as the sum of logP values for all protein atoms within 6 Å of the ligand atom, and a logP cutoff for hydrophobicity applied. Atom pairs meeting this criteria are then added to the system as identified hydrophobic interactions.

Furthermore, there is the option to add a condensed set of interactions to the system. If this option is applied, in the event where ligand atoms make multiple hydrophobic interactions to atoms within a single residue, only the interaction with the vdW overlap closest to 0.0 Å is given.

π-π stacking interactions

Aromatic rings for π-π stacking interactions are determined by the residue type in proteins, with phenylalanine, tyrosine, tryptophan and histidine residues being designated as aromatic. Within ligands there are three ways to determine aromatic rings. The default method is using RDKit that implements Hückel's rule in determining if a ring is aromatic. However, this can be problematic if ligands are extracted from a PDB file as the bond types must be inferred from atom-atom distances. Therefore two alternative methods of setting ligand aromatic rings are given. The first is by planarity where any ligand not in a boat or chain conformation (i.e. a planar ring) is set as aromatic, while the second is for the user to manually set aromatic rings.

For determining π-π stacking interactions a candidate list of π-π stacks is generated for every possible pair of ligand and protein aromatic rings in the system. A distance cutoff between ring centres is first applied to filter this list, followed by a cutoff for the plane angles between the two rings. A third criteria involving the offset of ring centres (set to 2 Å) is applied as in PLIP [7]. Two distinct types of π-π stacking geometries can be determined by the program. P-stacking where both face-to-face stacks and offset stacks are classified, and T-stacks where edge-to-face stacks are classified.

Cation-π interactions

Determination of aromatic rings for cation-π interactions used the same criteria as for π-π stacking. Protein cationic atoms were determined by the protein residue type. Cationic centres for arginine residues are defined as the centroid of the NH1 and NH2 nitrogens, or the side chain nitrogen position for lysine residues. Histidine residues by default are not protonated, however there is the option for the user to manually set histidine protonation states, or for histidine states to be inferred from the protein PDB or for both histidine protonation states to be checked automatically. Ligand cations are determined by RDKit by the number of bonded atoms, e.g, a ligand nitrogen that forms four bonds is considered a cation.

All combinations or pairs of cation atoms and aromatic rings are considered as candidate bonds. A distance cutoff criteria is used to filter the candidate list and identify cation-π interactions within the system.

Halogen Bonds

By default chloride, bromide and iodide halogen atoms are searched for in ligands, as it is assumed that protein residues contain no halogens. Here we introduce the concept of ‘root’ atoms, in the context of the ProPLID algorithm the ‘root’ atom refers to the atom covalently bonded to the halogen bond donor atom (chlorine, bromine, iodine atoms) or the atom/s covalently bonded to halogen bond acceptor atoms (oxygen, nitrogen, sulphur). Only halogens atoms bonded to carbon ‘*root’* atoms are considered as halogen bond donors.

Halogen bonds acceptors within the protein are considered to be oxygen, nitrogen and sulphur atoms bonded to either carbon, nitrogen or sulphur 'root' atoms. For halogen bond acceptors present within other ligands of the system the *'root'* atoms may also be phosphorous atoms.

The geometric criteria applied to identify halogen bonds are as follows: a donor acceptor distance cutoff, a donor root-donor-acceptor angle minimum and maximum values, and a donor-acceptor-acceptor root atom angle minimum and maximum values. Should an atom pair meet the criteria for a halogen bond, and the acceptor have an available lone pair to form the interaction, then the bond will be identified within the system. Since the success of identifying a halogen bond is dependent on available lone pair electrons in the acceptor during interaction assignment, halogen bonds are assigned once metal ion complexes and strong and weak hydrogen bonds have been assigned. Whilst this is not a perfect way to do this, it is justified by the fact that halogen bonds are reported to occur at much lower rates than hydrogen bonding.

Metal ion complexes

Currently, only calcium, manganese, magnesium, iron, copper and zinc metal ions complexes are considered for metal ions complexes. However, more may be added independently by researchers should appropriate geometric values be known.

Ions are identified from any PDB files or small molecule files supplied when creating the complex. All atoms within a defined cutoff distance of ion atoms that are designated as atoms with the ability to form coordinate interactions with ions are identified (by default only oxygen, nitrogen and sulphur atoms are taken). The angle and distance geometry are computed for all unique combinations of atoms, for all the supported coordination types. Supported coordination types are linear, trigonal planar, tetrahedral, square pyramidal, square planar, trigonal bipyramidal, octahedral and pentagonal bipyramidal. The RMSD of computed angles are compared against preferred theoretical values. If the RMSD is less than a defined RMSD cutoff the ion complex is appended to a list of putative ion-complex interactions.

Putative interactions are then sorted based on a hierarchy of bond orders and a tolerance for the deviation from theoretical angle and distance values. This step prevents the solutions from becoming dominated by ion complexes with low coordinate numbers. For example, should an octahedral complex be predicted, that complex will inherently contain a square planar complex and multiple linear complexes that will also be predicted. For complexes with lower coordination numbers, such as linear complexes, the chance of deviating greatly from prefered geometry is less than that for complexes with higher coordination numbers. Sorting in this way, allows for square planar complexes to be ranked higher than linear complexes and octahedral complexes to be ranked higher than both square planar and linear complexes,

even if the deviation from preferred geometry is greater for the higher coordination complex than the lower. Of course higher coordination complexes only jump up the order if their geometry is within an acceptable deviation compared to preferred values, referred to in the algorithm as the 'ion order relation RMSD' (where the RMSD is calculated from the preferred theoretical angles).

Implementation

The ProPLID algorithm has been implemented in python 3. A novel feature of the software is that multiple ligands may be added to the protein system, for example the output of a docking run, and the interactions of each ligand conformation with the protein system can be processed as a batch or individually. Small molecules, ions, residues, and solvent are dealt with using the python RDKit API within the system, as such it is possible to add molecules to the system using an RDKit *'Mol'* object. Furthermore, all the methods for manipulating small molecules in RDKit are available for use with individual ions, residues, and solvent molecules. In this way the software can be easily inserted into any current RDKit based workflows.

The software was designed to be ready to use out of the box, as well as allowing advanced customisation of most parameters to suit the users needs.

To this end minimal code is required to set up and run an analysis and write the output:

```
from ProPLID import *

ligand_path = './Main_ligand.sdf'
protein_path = './protein.pdb'
out_file = './ligand_interactions.txt'

#Initialise a protein system with a ligand supplied by sdf file
interaction_object = Interaction(protein_path, ligand = ligand_path)

#calculate ligand interactions
interaction_object.calculate_interactions()

#isolate the ligand molecule
ligand = interaction_object.get_residue("Main_ligand")

#write the interactions and geometric data to a .txt file
ProteinTools().write_interactions_to_file(ligand, out_file)
```

A more advanced customisation of the software is available. The interaction object created is linked with an interaction data object containing all of the interaction parameters necessary for the analysis, an example of the interaction data class containing all parameters that are customisable is given in the appendix (Figure A1).

To add custom values to the interaction object the interaction object can be changed directly once initialised or a custom interactions data object supplied to the interaction object upon initialization, and of course at any point the current interactions data being used can be overwritten with a new interactions data object.

```
#modify an existing parameter
interaction_object = Interaction(protein_path, ligand = ligand_path)
interaction_object.interactions_data.hbond_dist_max = 3.5

#supply a modified interaction data object upon initialization
interactions_data =  Interaction_data()
interactions_data.hbond_dist_max = 3.5
interaction_object = Interaction(protein_path, ligand = ligand_path, interactions_data =
interactions_data)

#overwrite all existing interactions data with an interactions data object
interaction_object.interactions_data = interactions_data
```

The output of the program is a python dictionary object of all interactions, between the protein residues, ions and ligands supplied. This can be used for downstream analysis or written to a '.txt' or '.json' file. Additionally, interactions for individual ligands, ions, residues or atoms can be pulled from all interactions for downstream analysis. A text file output example of interactions for a given ligand is shown below:

```
----------Metal complex----------
Metal complex: 1
Complex: Square Planar
Ion: MG 300:A
    Atom: Main_ligand 0 O 2, distance: 2.33
    Atom: GLU 166 OE2 8, distance: 2.26
    Atom: HOH 1009 O 0, distance: 2.33
    Atom: HOH 1018 O 0, distance: 2.36

----------Hydrogen Bond----------

Hydrogen Bond type: strong: 1
donor atom: Main_ligand 0 N 15
acceptor_atom: ASP 162 O 3
DA distance, DHA angle: 2.88, 168.14

Hydrogen Bond type: weak (CHO) 1
donor atom: Main_ligand 0 C 7
acceptor_atom: ASP 162 O 3
DA distance, DHA angle: 3.59, 147.18

----------Halogen bond----------

Halogen bond: 1
Donor atom: Main_ligand 0 Cl 4
Acceptor atom: GLY 61 O 3
Atom-atom distance, root/donor/acceptor angle, root/acceptor/donor angle :3.02 171.03 147.25

----------Pi-Pi stack----------
Pi-Pi stack P-stack: 1
Ring 1 :
    Atom: Main_ligand 0 C
    Atom: Main_ligand 0 N
    Atom: Main_ligand 0 C
    Atom: Main_ligand 0 C
```

```
    Atom: Main_ligand 0 C
    Atom: Main_ligand 0 N
Ring 2 :
    Atom: PHE 70 CG
    Atom: PHE 70 CD2
    Atom: PHE 70 CE2
    Atom: PHE 70 CZ
    Atom: PHE 70 CE1
    Atom: PHE 70 CD1
Ring-centre distance, ring plane angle, ring centre offset: 3.70, 157.58, 0.25

----------Pi-Cation----------
Pi-Cation: 1
Cation: LYS 65 NZ 8
Ring centre: [-29.343, -6.726, 8.824]
    Atom: Main_ligand  0 N 0
    Atom: Main_ligand  0 C 15
    Atom: Main_ligand  0 C 4
    Atom: Main_ligand  0 C 3
    Atom: Main_ligand  0 C 2
    Atom: Main_ligand  0 C 1
Cation-Pi distance: 4.26

----------Hydrophobic interactions----------

Hydrophobic interactions: 2
atom 1: Main_ligand 0 C 26
atom 2: LEU 69 CD2 7
ClogP, atom-atom distance, delta VDW: -0.19, 3.83, -0.19
```

The code base can be accessed at: https://github.com/SweeneyAaron/ProPLID.git

Assessing the accuracy of ProPLID

Generating an experimental benchmark

To generate an experimental benchmark[1] to assess the accuracy of the algorithm the PDB was searched for atomic models of proteins with small molecule ligands bound. The following inclusion criteria was used:

- ➔ A resolution of ≤ 2.5 Å
- ➔ An R-value of ≤ 0.25
- ➔ An R_{free} -value of ≤ 0.3
- ➔ Models containing nucleic acid molecules were excluded
- ➔ Publications contained significant descriptions of protein-ligand interactions

Having atomic models derived from high resolution experimental data decreased the chances of introducing errors from modelling of the atomic models when identifying interactions. Models containing nucleic acid in the binding site were excluded as it cannot be assumed that

[1] *The experimental benchmark was generated during a period where I was supervising a masters student in the lab of Prof. Maya Topf. I would like to thank and acknowledge Ms Luca Genz for her contributions to the curation of this benchmark.*

the geometry of ligand-nucleic acid interactions would mirror that of protein-ligand interactions. Furthermore, the final selection criteria relating to '*significant*' protein-ligand descriptions was rather subjective and represents a limitation of this investigation. What constituted a detailed description was decided by the researcher and as such was open to bias. One of the reasons this criteria was necessary was due to a trend of under-reporting interactions in publications. There were many reasons for this including the focus of the report being on a few specific interactions that highlighted the mechanism of action of drugs and the small molecules not being the subject of the report. Every effort was taken to apply this criteria in a meaningful way, for example, models were discounted when papers neglected to report an interaction type that should have been contained, such as vdW interactions.

The resulting benchmark was composed of 35 structures, containing 321 interactions, consisting of 176 hydrogen-bonds, 122 hydrophobic interactions, 13 **π-π** stacks, 5 halogen bonds, 3 metal ion complexes and 2 cation-**π** interactions (Table 1). The distribution of interactions correlated well with the ratios of interactions seen in the PDB [2], with the exception of hydrophobic interactions that were reported to be present at rates higher than hydrogen bonds. This was likely due to under-reporting of the hydrophobic interactions in papers.

Furthermore, it was necessary to ensure the benchmark represented a set of chemically diverse structures. To this end each chemical structure was converted to a 'fingerprint' using the RDKit python module. Briefly, a chemical fingerprint is a method of creating a vector representing the chemical characteristics of a molecule. The vector is represented by a bit array where each position in the array represents a specific chemical moiety. If the moiety is present on the molecule a 1 (i.e. 1 bit) is added to the corresponding position within the array, if not a value of 0 is determined at that position. One or more arrays can be compared as a way to identify chemically similar or dissimilar molecules. Here we compared molecules using the Tanimoto coefficient (Eq. 2) with a similarity cutoff of 0.85, where a value of 1.0 indicated a perfect match and a value of 0.0 reflects no similarity.

$$Eq\,2.\,T = \frac{N_c}{(N_a + N_b - N_C)}$$

Where T is the Tanimoto coefficient, N_c is the number of bits shared by both molecule vectors (a and b), N_a/N_b is the number of bits in the vector for molecule a/b.

It was seen that the tanimoto coefficients were rarely above 0.5 (Figure 1) with an average value over all pairs of 0.25. This indicated that the benchmark represented a chemically diverse set of molecules. Additionally, it was also pertinent to ensure the set of protein structures were diverse. To this end, the protein sequences were aligned using the clustal omega multiple sequence alignment algorithm [18]. It was seen that generally the protein sequences were diverse (Figure 1) with an average % sequence identity of 16.2 %.

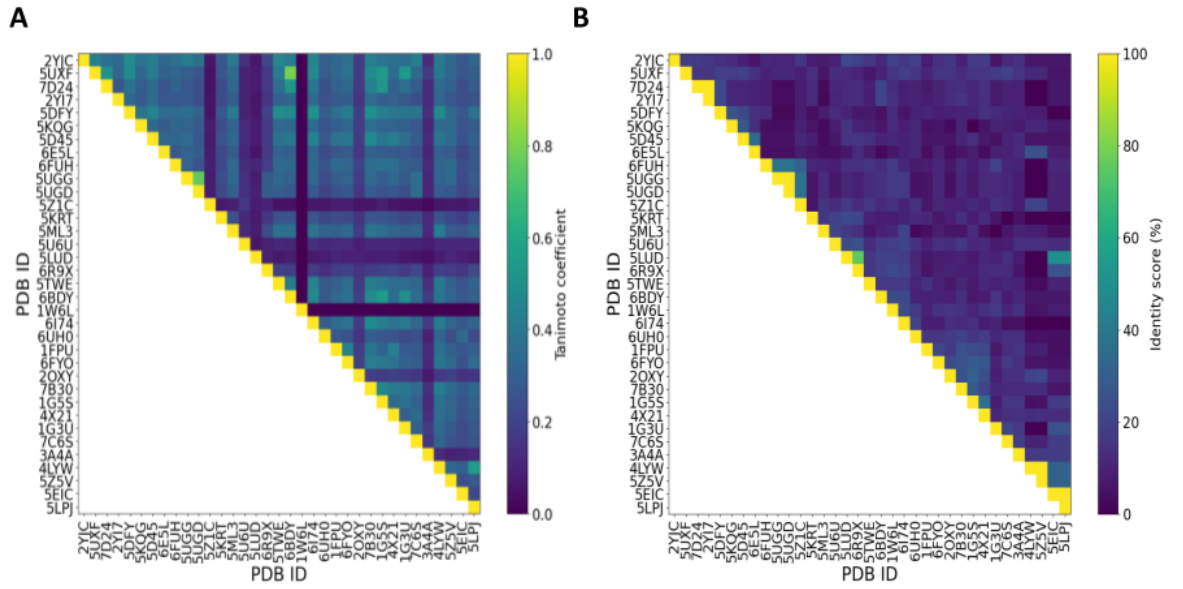


Figure 1. A. *A 2D heat map representing the tanimoto coefficients between pairs of ligand vectors.* ***B.*** *A 2D heat map representing the identity scores (%) between protein sequence pairs. In both images the PDB IDs are given on the X- and Y-axis, and the corresponding colour key is shown to the right of the image.*

Table 1. The resolutions, validation and interaction statistics of benchmark complexes

PDB code	Resolution	R	R_{free}	Hydrogen bonds	Hydrophobic	**π-π** stacks	Cation-**π**	Halogen bonds	Metal ion complexes
1fpu	2.4	0.233	0.264	6	10	-	-	-	-
1g3u	1.95	0.216	0.250	12	1	1	-	-	1
1g5s	2.61	0.201	0.242	5	17	1	-	-	-
1w6l	2.0	0.148	0.173	1	-	-	-	-	1
2oxy	1.81	0.207	0.269	-	-	-	-	2	-
2yi7	1.40	0.175	0.176	5	8	-	-	-	-
2yjc	1.14	0.147	0.184	2	2	-	-	1	-
3a4a	1.60	0.159	0.174	9	-	1	-	-	-
4lyw	1.95	0.235	0.197	4	3	1	-	-	-
4x21	1.95	0.20	0.224	5	-	-	-	1	-
5d45	1.65	0.188	0.217	2	3	1	-	-	-
5dfy	1.60	0.155	0.181	17	-	-	-	-	-
5eic	1.50	0.173	0.209	3	13	-	-	-	-
5kqg	1.50	0.167	0.194	8	6	3	-	-	-

5krt	1.65	0.185	0.195	4	9	-	-	1	-
5lpj	1.65	0.167	0.196	6	-	-	-	-	-
5lud	1.25	0.193	0.220	4	-	-	-	-	-
5ml3	1.40	0.189	0.203	7	-	2	-	-	-
5twe	1.5	0.150	0.169	7	-	-	-	-	-
5u6u	1.79	0.152	0.202	4	-	-	-	-	-
5ugd	1.38	0.160	0.182	3	14	1	-	-	-
5ugg	1.20	0.164	0.176	1	14	1	-	-	-
5uxf	1.50	0.168	0.186	7	-	1	-	-	-
5x1c	1.45	0.118	0.143	5	-	-	-	-	-
5z5v	1.66	0.179	0.219	5	6	-	-	-	-
6dby	1.51	0.144	0.186	9	-	-	-	-	-
6e5l	1.17	0.118	0.146	3	-	-	-	-	-
6fuh	1.37	0.134	0.180	5	-	-	-	-	-
6fyo	2.32	0.203	0.234	5	1	-	-	-	-
6i74	0.96	0.132	0.147	-	-	-	1	-	-
6r9x	1.66	0.144	0.193	4	-	-	-	-	-
6uh0	1.31	0.147	0.164	2	9	-	-	-	1
7b30	2.10	0.177	0.196	5	-	-	1	-	-
7c6s	1.60	0.207	0.222	6	6	-	-	-	-
7d24	1.55	0.180	0.198	5	-	-	-	-	-
Total Interactions (321)				176	122	13	2	5	3

Defining algorithm parameters

To assess the accuracy of the ProPLID software for detecting protein ligand interactions using bond geometry, initial parameters for the geometry of interaction types were taken from a recent publication that characterised the bond geometry of 11,016 protein ligand complexes deposited in the PDB [2] (Table 2). Additional parameters for logP were taken based on observations in the X-SCORE molecular docking score [19]. Values for metal ion geometry were taken from a report that analysed the geometry of metal complexes for Ca, Mg, Mn, Fe and Zn ions, in high resolution structures from the PDB and the Cambridge Structural Database (CSD) [20].

Table 2. Initial geometric parameters used for running the ProPLID software

Parameter	Value	Source

Binding site radius	9.0 Å	-
Strong hydrogen bonds		
Maximum DA distance	3.9 Å	[2]
Minimum DHA angle	90 °	[2]
Weak hydrogen bonds		
Maximum DA distance	3.6 Å	[2]
Minimum DHA angle	130 °	[2]
Maximum DHA angle	180 °	
Hydrophobic interactions		
Δ VDW overlap max	-0.4 Å	[21]
Δ VDW overlap min	0.0 Å	[21]
LogP cutoff	-0.5	[19]
Local environment cutoff	6.0 Å	[19]
π-π stacking interactions		
Ring centre distance max	4.0 Å	[2]
Plane angle deviation	∓ 30 °	[2]
Cation-π interactions		
Cation-ring distance max	4.0 Å	[2]
Halogen bonds		
Maximum DA distance (Cl)	3.5 Å	[2]
Maximum DA distance (Br)	3.6 Å	[2]
Maximum DA distance (I)	3.73 Å	[2]
Donor angle min	130 °	[2]
Donor angle max	180 °	[2]
Acceptor angle min	90 °	[2]
Acceptor angle max	150 °	[2]
Metal ion complexes		

Ion distance tolerance	0.5 Å	[20]
Max ion angle deviation	18.0 °	[20]
Ion geometry RMSD deviation	0.5 °	-
Ion order relation RMSD deviation*	1.5 °	-

** see metal ion complexes section above for a detailed description.*

Using these parameters and our software, predictions for each bond type in the benchmark of protein-ligand complexes were compared with the interactions reported in their respective papers.

Comparisons of calculated interactions with published interactions was conducted using the F-measure (Eq 3.):

$$Eq\,3.\ F_{measure} = \frac{True\ positives}{True\ positives\ +\ 0.5(False\ positives\ +\ False\ negatives)}$$

The F-measure is a metric of accuracy and recall of predictions. An F-measure of 1.0 is perfect and an F-measure of 0.0 indicates the lowest level of accuracy and recall. Here, true positives were defined as interactions mentioned in the associated paper and by the ProPLID algorithm, false positives are defined as interactions predicted by the ProPLID algorithm that were not mentioned in the associated paper, and false negatives are described as interactions mentioned in the associated paper that are not predicted by the ProPLID algorithm. The F-measure was chosen over other metrics as the true negative data was not needed to calculate the score.

The accuracy and recall of the ProPLID algorithm

The mean F-measure over all 35 cases was seen to be 0.386, with the highest F-measure being 0.762 and the lowest at 0.0 (Table 3, Figure 2).

A total of 138 true positive interactions were found by the ProPLID software. This included 99 hydrogen bonds representing 56.25 % of the total reported hydrogen bonds. A total of 31 reported hydrophobic interactions were found representing 25.41 % of the total reported. Only 2 of the 13 reported **π**-**π** stacking interactions were found, with only one of the two reported cation-**π** interactions. Of the 5 halogen bonds only 3 were found, along with 2 of the 3 metal-ion complexes.

It was seen that 238 interactions were reported as false positives, of these hydrogen bonds were seen to account for 65.12 % (155) , hydrophobic interactions for 33.61 % (80), halogen bonds for 0.42 % (1), cation-**π** for 0.0 % (0), **π-π** stacking interactions for 0.42 % (1), and metal ions for 0.42 % (1).

The program resulted in 183 false negative predictions, of these 42.07 % (77) were hydrogen bonds, 49.72 % (91) hydrophobic interactions, 6.01 % (11) **π-π** stacking interactions, 0.54 % cation-**π** (1) ,1.08% halogen bonds (2), and 0.54 % % (1) metal ion complexes.

These results indicated that whilst ProPLID had the most success at predicting hydrogen bonds, it was also the interaction type that resulted in the highest number of false positive predictions. Hydrophobic interactions were the second most successful prediction made by ProPLID; however, there was still a high number of false positives being predicted. The results indicated that the parameters used for predicting hydrogen bonds and hydrophobic interactions may be too broad.

For the **π-π** stacking interactions, cation-**π** and halogen bonds, the number of true positives was very low. This indicated that the parameters used to predict these interactions may have been too stringent.

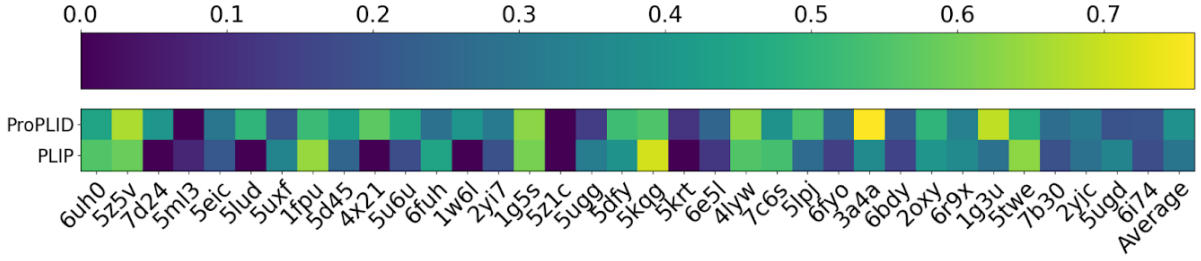


Figure 2. *A heatmap comparison of the F-measures calculated for ProPLID and PLIP for all 35 protein ligand complexes in the benchmark. The PDB ID of each case is given on the X-axis, along with the average F-measure over all cases. A colour key for the heat map is shown above the plot.*

Table 3. The F-measures, true positives (TP), false positives (FP) and false negatives (FN) predicted for each case in the benchmark using the ProPLID and PLIP algorithms.

PDB ID	F-measure ProPLID	TP ProPLID	FP ProPLID	FN ProPLID	F-measure PLIP	TP PLIP	FP PLIP	FN PLIP
6uh0	0.444	4	2	8	0.556	5	1	7
5z5v	0.667	7	3	4	0.588	5	1	6
7d24	0.400	3	7	2	0.0	0	11	5
5ml3	0.0	0	18	9	0.083	1	14	8
5eic	0.300	3	1	13	0.211	2	1	14

5lud	0.5	2	2	2	0.0	0	5	4
5uxf	0.194	3	20	5	0.343	6	20	3
1fpu	0.519	7	4	9	0.643	9	3	7
5d45	0.429	3	5	3	0.25	2	8	4
4x21	0.571	4	4	2	0.0	0	6	6
5u6u	0.462	3	6	1	0.182	1	6	3
6fuh	0.286	2	7	3	0.444	4	9	1
1w6l	0.4	1	2	1	0.0	0	1	2
2yi7	0.316	3	3	10	0.19	2	6	11
1g5s	0.629	11	1	12	0.606	10	0	13
5z1c	0.0	0	2	5	0.0	0	2	5
5ugg	0.138	2	11	14	0.32	4	5	12
5dfy	0.524	11	14	6	0.387	6	8	11
5kqg	0.552	8	4	9	0.710	11	3	6
5krt	0.118	1	2	13	0.0	0	5	14
6e5l	0.25	2	11	1	0.125	1	12	2
4lyw	0.632	6	5	2	0.556	5	5	3
7c6s	0.387	6	13	6	0.538	7	7	5
5lpj	0.545	6	10	0	0.267	2	7	4
6fyo	0.267	2	7	4	0.143	1	7	5
3a4a	0.762	8	3	2	0.364	4	8	6
6bdy	0.231	3	14	6	0.154	2	15	7
2oxy	0.500	1	1	1	0.4	1	2	1
6r9x	0.333	2	6	2	0.364	2	5	2
1g3u	0.690	10	4	5	0.286	4	9	11
5twe	0.476	5	9	2	0.632	6	6	1
7b30	0.273	3	13	3	0.19	2	13	4
2yjc	0.308	2	6	3	0.286	2	7	4
5ugd	0.194	3	10	15	0.345	5	6	13
6i74	0.200	1	8	0	0.182	1	9	0
Mean F-measure ProPLID = 0.386					Mean F-measure PLIP = 0.296			

Comparison of ProPLID with PLIP

To compare the performance of ProPLID to that of other software, the benchmark was run using the open source interaction software PLIP [7]. The mean F-measure across all cases was seen to be 0.296 (Table 3, figure 2), a result that was significantly less than the mean F-measure of 0.386 seen with ProPLID (Table 3, Figure 2) (p=0.03). ProPLID performed better on 23 of the 35 cases, whilst PLIP performed better on 11 cases. In 1 case both PLIP and ProPLID had an F-measure of 0.0.

PLIP identified 113 correct interactions of which, 65 were hydrogen bonds, 37 were hydrophobic interactions, 2 were halogen bonds, 8 were **π-π** stacking interactions, and 1 was a cation-**π** interaction.

A total of 233 false positive interactions were predicted by PLIP, of these 136 were hydrogen bonds, 82 were hydrophobic interactions, 3 were halogen bonds, 6 were **π-π** stacking interactions, 3 were cation-**π** interactions and 3 were metal ion complexes.

In terms of false negatives, a total of 209 interactions were not predicted by PLIP, 112 hydrogen bonds, 85 hydrophobic interactions, 3 halogen bonds, 5 **π-π** stacking interactions, 1 cation-**π** interaction and 3 were metal ions complexes.

The PLIP software managed to predict a total of 113 true interactions. This was 25 less than the ProPLID algorithm. However, using the PLIP software resulted in the prediction of 233 false positive interactions, 5 less than the ProPLID software.

Of the true positive interactions that PLIP predicted 65 were hydrogen bonds compared with 99 predicted by ProPLID. However, the number of false positive hydrogen bonds predicted by PLIP was 19 more than the number predicted with ProPLID.

In terms of hydrophobic interactions PLIP predicted 6 more true interactions than did ProPLID whilst both had a comparative level of false positive predictions.

PLIP was more accurate at predicting **π-π** stacking interactions than ProPLID. However, ProPLID had better success predicting metal ion complexes. Both softwares predicted the same amount of cation-**π** and halogen bonds.

The results indicated that whilst the ProPLID software had a better F-measure than PLIP, there was room for improvement especially with respect to predicting halogen bonds, **π-π** stacking and cation-**π** interactions.

An Improved parameter value set

One of the key features of the ProPLID software is the ability to customise the parameters for identifying interactions. To examine the usefulness of this feature we experimented with varying the parameter set to increase the mean f-measure for the benchmark set.

The aim was to find values for each interaction type that were broad enough to account for small errors in modelling, but stringent enough not to report a high level of false positives. For hydrogen bonds and hydrophobic interactions the set included enough interactions to generate meaningful results. However, for other interaction types the number of interactions was low, therefore the best parameters found were compared with literature values to fully assess their reasonableness.

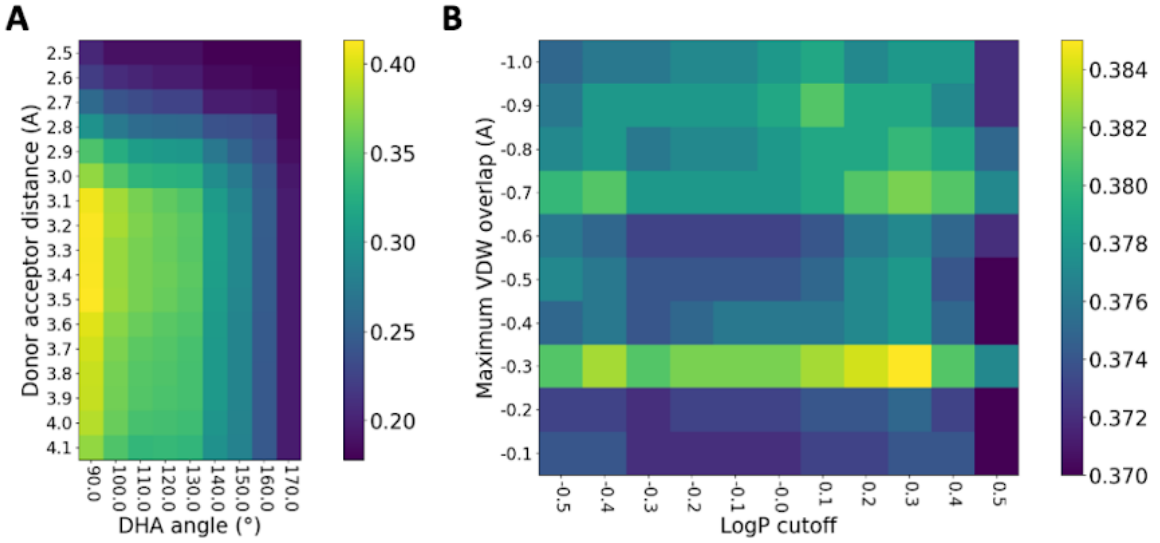


Figure 3. *The mean F-measure calculated when running the ProPLID algorithm with hydrogen bond distance parameters between 2.5 Å and 4.1 Å, and angle criteria between 90° and 170 ° (**A**). The mean F-measures calculated when changing the values for calculating hydrophobic interactions (**B**) with a vdW overlap between -0.1 Å and -1.0 Å, and logP cutoff values between -0.5 and 0.5. The mean F-measure is indicated by colour and colour keys are shown to the right of plots.*

Strong Hydrogen bonds

To improve the parameters for strong hydrogen bonds the mean F-measure was calculated using values for the maximum donor-acceptor atom distance between 2.0 and 4.1 Å and minimum DHA angles between 90° and 180°. At all distance values, increasing the stringency of the minimum DHA angles only served to decrease the mean F-measure (Figure 3). However, when the maximum donor-acceptor atom distance was reduced from 3.9 to 3.5 Å there was a marked increase in the mean F-measure from 0.386 to 0.406. Using these parameters ProPLID had a comparable predictive power to the original values, predicting 97 correct hydrogen bonds, compared to 98 with the original values. The reason for the marked

increase in mean F-measure was mostly due to the number of false positives predicted with these values, down from 155 with the original values to 117. This indicated there may be some value in restricting the maximal donor-acceptor atom to 3.5 Å. This value was at the higher end of the range of acceptor-donor distances reported in a study of bond geometries for protein-ligand complexes of the PDB [2].

Weak hydrogen bonds

Along with strong hydrogen bonds, ProPLID also introduced geometric analysis for weak hydrogen bonds. To find optimal parameters for predicting weak hydrogen bonds, the mean F-measure was calculated for the benchmark using values of donor-acceptor atom maximal distances from 2.9 to 3.7 Å, and DHA angles from 150 to 180°. The value pair that gave the best mean F-measure was at a donor acceptor atom maximal distance of 3.5 Å and a DHA minimum angle of 150°. Three protein-ligand complexes in our benchmark reported weak CH-O hydrogen bonds (1G5S, 4X21, 6FYO). Of these the original parameter value set predicted 2 correctly. This was increased to 3 using the new parameters, evidenced by 98 true positive hydrogen bonds found. Furthermore, adapting the parameters reduced the number of false positive hydrogen bonds predicted from 117 to 96. This led to an increased mean F-measure of 0.418 compared to using the best value pairs found for the strong hydrogen bonds alone. The distance parameter correlated well with the range of distances seen in a systematic survey of protein-ligand complexes in the PDB [2]. The minimum DHA angle values correlated well with observations from high resolution protein ligand complexes, where two maxima are seen at 150 ° and one at 180° [9].

Hydrophobic interactions

To assess the predictive power of the ProPLID algorithm for hydrophobic interactions, the algorithm was run using maximal vdW overlap distances between -0.1 and -1.0, and logP cutoff values ranging from -0.5 to 0.5 (Figure 3).

The values that produced the best results were difficult to determine as the increase in mean F-measure was relatively evenly distributed. The mean F-measure was seen to be more dependent on the vdW overlap cutoff than the logP. Two values for the vdW overlap were seen to produce the best results. The first band was seen with a maximal overlap of -0.3 Å. Within this band the mean F-measure peaked at two values of logP one at -0.4 and one at 0.3 (Figure 3).

However, both were seen to be more restrictive in assigning hydrophobic interactions than the original values with one predicting 27 true positives and 53 false positives (-0.3 Å and 0.3), and the other predicting 27 true positives and 59 false positives (-0.3 Å and -0.4). Both these value pairs significantly reduced the predictions of hydrophobic interactions compared to the original values and the PLIP software. However, the improved F-measure scores of

0.396 and 0.392 were deemed to be the result of predicting significantly less false positives than both the original values and the PLIP software.

From the plot, a second region where the score improved, albeit to a lesser extent, was using the vdW overlap cutoff of -0.7 Å (Figure 3). Again at this value there were two values of logP where the mean F-measure peaked at -0.4 and 0.3. Both of these value pairs predicted more true hydrophobic interactions (40) than when using the original values or the PLIP software. However, the number of false positives predicted increased to 104 when a logP cutoff of 0.3 was used and 114 when a logP cutoff of -0.4 was used. This indicated that the inclusion of logP was having a small effect on predicting hydrophobic contacts and restricting its value to 0.3 was advantageous. It was less clear which value of vdW overlap cutoff should be used as the value of both true positives and false positives increased as the maximum overlap increased. This led to the conclusion that the value used for this cutoff will most likely be problem-specific. However, for the purposes of this investigation we chose the value that predicted the most true positives, -0.7 Å.

π-π stacks

The performance of ProPLID when identifying **π-π** stacking interactions using the original values was poor. To this end the mean F-measure was calculated over the whole benchmark using ring centre-centre distances from 4.0 to 6.0 Å and plane angle deviations from 5 to 30°.

The mean F-measure was seen to be the highest when using a ring centre-centre distance of 5.2 Å and a plane angle deviation of 30 °. The mean F-measure improved from 0.386 to 0.393, at these values. A total of 8 correct **π-π** interactions were predicted along with 5 false positive interactions. The number of correct **π-π** stacking interactions was comparable to that of the PLIP software. However, less false positive results were produced using ProPLID. When the paper results were searched to identify why ProPLID could not identify the last 5 **π-π** stacking interactions it was seen that 4 of the 5 interactions represented aliphatic-aromatic ring stacks and were not true **π-π** stacking interactions. Neither ProPLID nor PLIP had geometric parameters that could identify these interactions, as less information is available regarding the geometry of the aliphatic-aromatic ring stacks when compared to the plethora available for the **π-π** stacking interactions. However, it was seen as a positive that these interactions were not labelled as true **π-π** stacking interactions by ProPLID. The final false negative interaction had ring plane-plane angles greater than the allowed deviation, a plane plane angle of 131 ° .

The values for the ring centre-centre distances correlated well with previously published literature. The **π-π** P-stack, where rings stack face-to-face is reported as having a ring centre-centre distance approximately equal to twice the vdW radii of carbon (~3.4 Å) [2]. Whilst the **π-π** T-stack where rings line up edge-to-face, have been reported to have a larger ring centre-centre distance, with benzene dimers reported at a distance of 4.96 Å [22]. The

value for the ring centre-centre distance was very close to this value at 5.2 Å, approximately 3 times the vdW radii of carbon, as such was deemed reliable.

Cation-π interactions

The cation-π interactions had the least geometric parameters of any interaction type with only ring centre-cation distance available. Additionally, the number of cation-π interactions in the benchmark was low. When attempting to optimise the values for the prediction of cation-π interactions the aim was to correctly predict both interactions in the benchmark whilst keeping the number of false positives predicted to an acceptable level. The ProPLID software was run at ring centre-cation distance values from 4.0 to 6.1 Å. It was seen that a ring centre-cation distance value of 4.3 Å was the minimum distance required to correctly predict both cation-π interactions, the mean F-measure at this distance was 0.387. However, this resulted in the prediction of 2 false positive interactions. They were included by the program as they fulfilled the geometric criteria,however, were to the side of the plane of the ring and therefore would not interact with the conjugate system. This highlighted an area of improvement for further iterations of the program. However, the false positive rate was deemed to be reasonable therefore these values were taken. The value of 4.3 Å did not deviate too far from the distance criteria reported in analysis of protein-ligand interactions within the PDB [2]. Additionally, this distance was within the 4.6 Å cutoff that was shown to be where most aromatic-amino sidechain groups clustered within a benchmark of 57 high resolution proteins [23]. Therefore, the distance cutoff of 4.3 Å was determined to be reliable.

Halogen bonds

The mean F-measure was not increased in any meaningful way by increasing the individual geometric values for halogen bonding. The values for the two interactions that were not predicted showed why this was the case. The first was a Br-O halogen bond with an acceptor-donor distance of 4.16 Å, 0.59 Å larger than the expected value for such a bond [2]. Additionally, the second halogen bond was a Cl-O bond with an acceptor-donor distance of 4.0 Å, 0.53 Å larger than expected for this bond type [2]. Reduction of the tolerances of the values for acceptor-donor distance resulted in a decrease in the mean F-measure due to a high level of false positives predicted. Since the default literature values showed an improved performance for ProPLID over PLIP when predicting true positives and limiting the prediction of false positives, the default halogen bonding parameters were left unchanged.

Metal ion complexes

The parameters for predicting metal-ion complexes were not improved upon. As any changes were detrimental to the mean F-measure, and no further true positives were predicted.

New parameters increased the accuracy and recall of ProPLID

When the new geometric parameters for predicting bonds were combined, the mean F-measure increased to 0.429 (Table 4). The best scoring case had an F-measure of 0.8 and the lowest 0.0 (Figure 4). It was seen that 148 true positive interactions were predicted, with 201 false positives and 173 false negatives. All three of these values were better than when the benchmark was scored using ProPLID with the original values or the PLIP software.

The ProPLID software outperformed PLIP in 24 cases, whilst PLIP was better in 11 cases, and in one case neither PLIP or ProPID was able to achieve a mean F-measure greater than 0.0. Furthermore, the new parameters improved the quality of interaction predictions in 21 cases compared to the original values, 6 cases were better with the original values and 8 cases remained the same.

Table 4. The F-measures, true positives, false positives and false negatives predicted for each case in the benchmark using the ProPLID algorithm with improved parameters

PDB ID	F-measure	True positives	False positives	False negatives
6uh0	0.632	6	1	6
5z5v	0.8	8	1	3
7d24	0.4	3	7	2
5ml3	0.071	1	18	8
5eic	0.3	3	1	13
5lud	0.571	2	1	1
5uxf	0.24	3	14	5
1fpu	0.48	6	3	10
5d45	0.444	4	8	2
4x21	0.571	4	4	2
5u6u	0.462	3	6	1
6fuh	0.333	2	5	3
1w6l	0.5	1	1	1
2yi7	0.444	4	1	9
1g5s	0.686	12	0	11
5z1c	0.0	0	2	5
5ugg	0.077	1	9	15
5dfy	0.512	11	15	6
5kqg	0.667	9	1	8
5krt	0.111	1	3	13

6e5l	0.222	2	13	1
4lyw	0.8	8	4	0
7c6s	0.522	6	5	6
5lpj	0.6	6	8	0
6fyo	0.25	2	8	4
3a4a	0.8	8	2	2
6bdy	0.273	3	10	6
2oxy	0.5	1	1	1
6r9x	0.333	2	6	2
1g3u	0.741	10	2	5
5twe	0.526	5	7	2
7b30	0.333	2	5	3
2yjc	0.333	2	5	3
5ugd	0.276	4	7	14
6i74	0.2	1	8	0
Mean F-measure = 0.429				

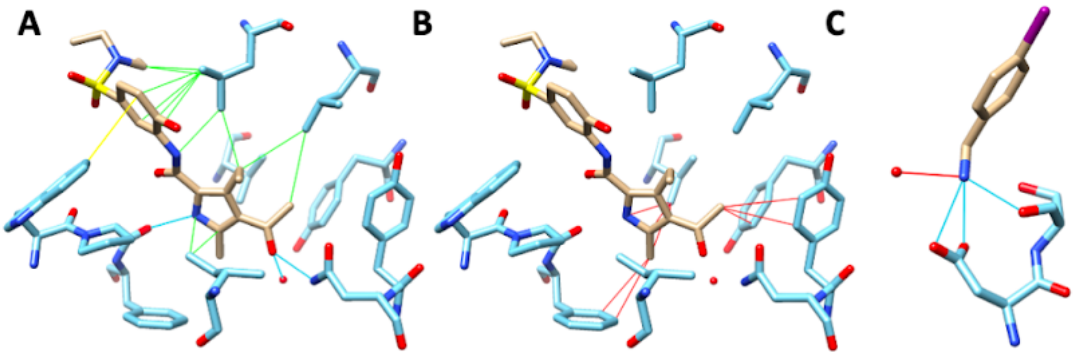


Figure 4. *The ProPLID protein (light blue)- ligand (tan) interaction predictions for the case with an F-measure of 0.8 (PDB ID: 4LYW). True positive interactions are shown (**A**) where blue lines represent hydrogen bonds, yellow lines represent **π-π** stacking interactions, and green lines represent hydrophobic interactions. False positive predictions made by ProPLID (**B**) are indicated with red lines. (**C**) Also shown is the case where the F-measure was 0.0 (PDB ID: 5Z1C), hydrogen bonds indicated in the associated paper are indicated with blue lines, false positive interactions predicted by ProPLID are indicated with red lines.*

Discussion

For analysis of protein ligand interactions it is important to have a diverse range of software able to accurately predict non-covalent interactions between the two. The software ProPLID achieved this using geometric parameters. It was seen to perform better than current software for the detection of interactions using bond geometry . Additionally, parameters for predicting weak CH-O hydrogen bonds were introduced in ProPLID. ProPLID was shown to be able to accurately predict these interactions and therefore has an improved complement of interactions types it can detect. Finally, the ProPLID software has a comprehensive list of parameters to customise the detection of all interaction types. The usefulness of this functionality was shown when initial parameters were customised and able to improve the quality of predictions whilst yielding values that correlated well with previously published literature.

The benchmark chosen consisted of 35 high-resolution structures from the PDB (Table 1). The distribution of interactions correlated well with that seen in the PDB [2] with the exception of hydrophobic interactions. However, the number of interactions for cation-π, halogen bonds and metal ion complexes was low.

Using literature values from a recent systematic review of protein-ligand interactions in the PDB [2] the ProPLID algorithm was more accurate at identifying protein-ligand interactions than PLIP software (Table 3, Figure 2). These results were encouraging, however, there were areas where ProPLID could improve. We attempted to improve the results of predictions by modifying the geometric parameters used to identify interactions.

This was mainly to show the utility of being able to modify parameter values. This is an important consideration as models obtained from low resolution experiments may not be as accurate, and protein ligand interaction geometry may not fit with what is observed in high resolution structures. For a number of interactions better values for geometric interactions were found. It is fully acknowledged that this methodology may have just fit parameters to structures in our benchmark. However, every care was taken to relate the identified values to reported literature values and further work would include assessing these parameters using a distinct benchmark.

We identified geometric criteria to better predict cation-π interactions. Whilst our benchmark only included 2 cation-π interactions the analysis included the whole benchmark. The reason being was to prevent overfitting of the parameters to the benchmark. By including the whole benchmark in the calculation there were multiple opportunities for protein-ligand pairs to detect false positive cation-π interactions. By monitoring the level of false positives predicted we could ensure that the values found were reasonable. The value of 4.3 Å found was seen to correlate well with literature values [2, 23]. Additionally, only two false positive interactions were detected. These false positives were seen to be aligned to the edge of aromatic groups. One way to improve the results would be to include parameters relating to the position of the cation group above the plane of the aromatic ring, such a strategy has previously been useful

in clustering cation-π interactions in protein models [23]. However, it may also be the case that these represent true interactions. In the case that the aromatic ring contains highly electron negative -R groups, the charge distribution can deviate from the classical polar-π model [24, 25], with the electron cloud at the edge of the ring adopting a more negative charge. In this case the cation may interact with the edge of the ring. However, the charge distribution and electron cloud topology of an aromatic system is complex to model, and further investigation is needed to improve geometric parameters for classifying cation-π interactions.

Furthermore, due to the low number of halogen and metal-ion complexes in the benchmark we could not improve upon the parameters for identifying these interactions. However, the values used showed a better performance than the open source software PLIP [7] for this is most likely due to the parameters the two softwares use. ProPLID included geometric parameters for each halogen (Cl,Br,I)-acceptor (O,N,S) pair, whereas PLIP uses a generic distance cutoff of 4.2 Å for all halogen interactions. Additionally, ProPLID was able to correctly identify two of the three metal-ion complexes within the benchmark, only failing to include a single ligand-metal ion interaction in the last. In contrast, PLIP was unable to identify any of the complexes correctly, the reason for this likely to be incorrect assignment of ligand atoms to the complexes. However, in two published papers, the authors of PLIP are yet to describe how metal ion complexes are calculated [7, 8].

It was difficult to accurately determine parameters related to detecting hydrophobic interactions. Benefits were seen when using a more stringent cutoff of the logP at 0.3. This mainly had the effect of reducing false positive predictions rather than increasing true positive predictions. However, two distance cutoff values -0.3 and -0.7 Å were shown to have roughly equivalent mean F-measures.The difference was that at the lower cutoff of -0.3 Å, there were less true positives along with less false positives. At the higher cutoff of -0.7 Å, there was a higher level of both true and false positives. This was an interesting finding, and can be explained by how hydrophobic interactions are typically calculated. Most papers that report hydrophobic interactions do so using a single distance cutoff for non-bonded atoms in the vicinity of the ligand, whereas in other papers hydrophobic interactions were ignored entirely. This led to the situation where, if the distance cutoff was increased to include interactions mentioned in papers, the false positive interactions increased due to papers that under reported hydrophobic interactions, and *vice versa*. It was seen that the PLIP software was slightly better at predicting hydrophobic interactions than ProPLID, however, suffered from the same under/over reporting problem as ProPLID. Thus it made it difficult to determine what constituted a true hydrophobic interaction and what was a true false positive and thereby confidently identify geometric criteria for the hydrophobic interactions.

ProPLID was shown to have the greatest success at predicting hydrogen bonds. Additionally, we were able to predict weak hydrogen bonds. For both hydrogen bonds types we identified distance and angle cutoff values that correlated well with previous reported values [2, 9, 10]. Particularly interesting was the value found for the minimum DHA angle of weak hydrogen bonds, 150°. This correlated well with a previous report that showed the distribution of CH-O

DHA values peaks at 150° and 180° [9]. Using a minimum DHA angle of 150 ° the ProPLID software was able to identify all the weak hydrogen bonds present in the benchmark and a significant improvement of the F-measure was seen. This was mainly due to a decrease in false positives predicted. The addition of weak hydrogen bonds represented a step forward in non-covalent interaction detection using geometric parameters. This is especially important as it relates to the lead optimisation process in drug discovery as it has been shown that optimisation of such interactions with the surrounding protein environment results can be an important consideration during lead optimisation of drugs [26].

Taken together the investigation showed that using geometric criteria to predict protein-ligand interactions resulted in reasonably accurate predictions. Furthermore, the ProPLID software was shown to be more accurate and identify a greater complement of interactions than other available software. Finally, the ProPLID algorithm has been implemented in such a way that gives researchers flexibility in its use. It will give reasonable results 'out of the box', however, interaction parameters can be fully customised with minimal code.

Future Directions

Future work will focus on the addition of new interaction types to the detection software. Additionally, adding an energy function layer to aid predictions would be explored. Finally, the ProPLID software would be added to commonly used molecular visualisation tools such as the USFC-chimeraX visualisation software [27]. This would aid researchers to visually inspect results and allow for publication quality figures to be produced using the software.

Methods and software

PDB files for the benchmark structures were downloaded from the PDB [1, 28] , protein files were downloaded in PDB format, ligands were downloaded as '.sdf' files as they contained more accurate information on ligand bond orders. To prepare structures for ProPLID analysis the ligands were removed from PDB files before analysis and supplied separately in '.sdf' file format.

The ProPLID software was developed using Python 3. The Python package RDKit was used to handle small molecules and Biopython was used to handle protein files. LogP values for calculating the hydrophobicity of local environments were taken from XLOGP3 [16].

Calculation of the F-measures, true positives and false positives was achieved using an in house python script. The PLIP software [7] was used via the interactive web server available at: https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index . The results were downloaded in '.xml' format and converted to '.json' format for calculation of the F-measure using an in house python script.

References

1. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 2007;35 Database issue:D301–3.

2. de Freitas RF, Schapira M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med Chem Commun. 2017;8:1970–81.

3. McDonald IK, Thornton JM. Satisfying Hydrogen Bonding Potential in Proteins. J Mol Biol. 1994;238:777–93.

4. Laskowski RA, Swindells MB. LigPlot+: Multiple Ligand–Protein Interaction Diagrams for Drug Discovery. J Chem Inf Model. 2011;51:2778–86.

5. Clark AM, Labute P. 2D Depiction of Protein−Ligand Complexes. J Chem Inf Model. 2007;47:1933–44.

6. Jubb HC, Higueruelo AP, Ochoa-Montaño B, Pitt WR, Ascher DB, Blundell TL. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. J Mol Biol. 2017;429:365–71.

7. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015;43:W443–7.

8. Adasme MF, Linnemann KL, Bolz SN, Kaiser F, Salentin S, Haupt VJ, et al. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Res. 2021;49:W530–4.

9. Panigrahi SK, Desiraju GR. Strong and weak hydrogen bonds in the protein-ligand interface. Proteins. 2007;67:128–41.

10. Sarkhel S, Desiraju GR. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: strong and weak interactions in molecular recognition. Proteins. 2004;54:247–59.

11. Stornaiuolo M, De Kloe GE, Rucktooa P, Fish A, van Elk R, Edink ES, et al. Assembly of a π–π stack of ligands in the binding site of an acetylcholine-binding protein. Nat Commun. 2013;4:1–11.

12. Benjahad A, Guillemont J, Andries K, Nguyen CH, Grierson DS. 3-iodo-4-phenoxypyridinones (IOPY's), a new family of highly potent non-nucleoside inhibitors of HIV-1 reverse transcriptase. Bioorg Med Chem Lett. 2003;13:4309–12.

13. Böhm H-J, Klebe G. What can we learn from molecular recognition in protein–ligand complexes for the design of new drugs? Angew Chem Int Ed Engl. 1996;35:2588–614.

14. Taylor R, Kennard O. Crystallographic evidence for the existence of CH.cntdot..cntdot..cntdot.O, CH.cntdot..cntdot..cntdot.N and CH.cntdot..cntdot..cntdot.Cl hydrogen bonds. J Am Chem Soc. 1982;104:5063–70.

15. Pierce AC, Sandretto KL, Bemis GW. Kinase inhibitors and the case for CH...O hydrogen

bonds in protein-ligand binding. Proteins. 2002;49:567–76.

16. Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, et al. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. J Chem Inf Model. 2007;47:2140–8.

17. Tsai J, Taylor R, Chothia C, Gerstein M. The packing density in proteins: standard radii and volumes. J Mol Biol. 1999;290:253–66.

18. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

19. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des. 2002;16:11–26.

20. Harding MM. Geometry of metal-ligand interactions in proteins. Acta Crystallogr D Biol Crystallogr. 2001;57 Pt 3:401–11.

21. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.

22. Arunan E, Gutowsky HS. The rotational spectrum, structure and dynamics of a benzene dimer. J Chem Phys. 1993;98:4294–6.

23. Singh J, Thornton JM. SIRIUS. An automated method for the analysis of the preferred packing arrangements between protein groups. J Mol Biol. 1990;211:595–615.

24. Hunter CA, Sanders JKM. The nature of .pi.-.pi. interactions. J Am Chem Soc. 1990;112:5525–34.

25. Cozzi F, Ponzini F, Annunziata R, Cinquini M, Siegel JS. Polar interactions between stackedπ systems in fluorinated 1,8-diarylnaphthalenes: Importance of quadrupole moments in molecular recognition. Angew Chem Int Ed Engl. 1995;34:1019–20.

26. Pierce AC, ter Haar E, Binch HM, Kay DP, Patel SR, Li P. CH...O and CH...N hydrogen bonds in ligand design: a novel quinazolin-4-ylthiazol-2-ylamine protein kinase inhibitor. J Med Chem. 2005;48:1278–81.

27. Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. 2018;27:14–25.

28. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.

Chapter 4

Integrating goodness-of-fit metrics with an empirical scoring function for fitting small molecules to density maps

Background

Advances in the field of cryogenic-Electron Microscopy (cryo-EM) are increasing the relevance of the technique in structure based drug discovery (SBDD). SBDD utilises the 3D structure of small molecule drugs (ligands) bound to their targets to understand the drug's mechanism of action and further develop compounds to bring about desired physiological effects. The 3D structures for this technique have classically come from X-ray crystallography experiments, where the resolving power allows for the structural determination of complexes at atomic resolutions (approximately ≤ 2.5 Å) and the accurate placement of individual atoms within the binding site when building atomic models.

More recently, structures have been obtained from cryo-EM experiments that have reached atomic or near-atomic resolutions (< 3.0 Å). This has had advantageous consequences for the field of SBDD, as the cryo-EM technique is not associated with crystallisation issues common with X-ray crystallography, which can make obtaining structures of certain proteins more difficult (e.g. membrane proteins). Furthermore, cryo-EM has the benefit of obtaining structures that may be closer to the native structure or a range of transition structures compared to the relatively static structures resolved in X-ray crystallography.

However, even though there may be no technical barrier to resolving structures at atomic resolutions with cryo-EM, statistically it is still rather uncommon. As of 2021 only 15.15 % of cryo-EM maps deposited in the electron microscopy data bank (EMDB) [1] were at resolutions < 3.0 Å, with the vast majority (44.01 %) being within the 3-4 Å range. A vast body of work has been reported on methods to fit proteins into cryo-EM maps at various resolutions [2–4]. Substantially less has been published with regards to the fitting of small molecules into binding sites within proteins. Therefore, new *in-silico* methods are currently needed for the fitting of small molecules to cryo-EM derived maps at near-atomic (≤ Å 3.0) resolutions and above.

Methods for fitting atomic models in cryo-EM density maps are built of two broad parts, an optimisation/search algorithm for exploring the conformational space and a scoring function. The scoring functions generally contain terms that score the goodness-of-fit of a candidate model to the density map, and terms that score the physico-chemical reasonableness of the

atomic model (e.g. with respect to bond geometry, non-bonded interactions or atom-atom distances).

The most common way to score the goodness-of-fit of an atomic model to a cryo-EM map is to compare an experimental map to a density map simulated from the atomic model, which has been blurred to match the resolution of the experimental map.

One metric that has almost exclusively been used for scoring the goodness-of-fit is the cross correlation coefficient (CCC). This metric has been successfully applied to scoring functions for the fitting of proteins both in high [3, 4] and low resolution [2, 5] cryo-EM density maps, and more recently small molecule ligands [6]. Another goodness-of-fit metric, the mutual information (MI) score, has been applied within the context of fitting protein complexes to experimental maps of GroEL at resolutions of 6, 10, 15 and 22 Å [5]. In this investigation 1000 atomic models were produced at each resolution that varied in root-mean-square deviation (RMSD) of atoms from the deposited models. The correlation between RMSD and either the MI or CCC score was calculated. It was seen that in all cases the MI score produced better or comparable results to the CCC score. However, the MI score was seen to be more robust to worsening resolution than the CCC score.

More recently the local variations of the MI and CCC scores that scored the fit of local regions within maps were examined [7]. The experiments involved investigating the ability of the local MI and CCC scores to align cryo-EM density map pairs from three categories, ribosomes, virus, and ‘other’ (other included maps from various protein complexes and RNA polymerases), at resolutions ranging from 4.3 to 18.0 Å. For each map pair 100 alignments were generated at various distances from a reference alignment and the alignments ranked with either score to ascertain where a score can discriminate between a map-map alignment close to the reference alignment. For viral and ‘other’ maps the local MI score was seen to be better able to discriminate a good from a bad alignment than the local CCC-based methods. However, the CCC-based methods had a better discriminatory power than the MI-based methods with respect to ribosomal map-map alignments. The authors postulate that this is due to the ribosomal map pairs having a greater number of features in common. Taken together the literature indicates that better fitting metrics may be available than the standard CCC. However, to date none have been applied to fitting small molecules into cryo-EM maps.

In parallel to improved goodness-of-fit metrics, it is also necessary to develop and investigate scores relating to the physico-chemical reasonability of the placement of ligand atoms within target binding sites. A common method when fitting is to combine such a score with the goodness-of-fit metric, with proteins. More recently a method that integrates the GLIDE molecular docking score [8] with the CCC has been reported [6].

The GLIDE score [8] and subsequent GLIDE-XP score [9] are both empirical scoring functions based on the Chemscore [10]. Both have shown to be relatively precise in reproducing the ligand conformations from deposited protein/ligand complexes with average RMSDs from the reference conformations of 1.95 Å and 1.75 Å when benchmarked with 282

and 198 protein ligand complexes, respectively for GLIDE and GLIDE-XP [8, 9]. Other empirical scoring functions commonly utilised for molecular docking include the AutoDock Vina scoring function [11] based on X-score [12]. Docking with AutoDock Vina was reported to achieve a 78% success rate, defined as generating a ligand conformation ≤ 2.0 Å from the reference ligand of the deposited atomic model, using a benchmark of 198 protein ligand complexes. Additionally, the empirical scoring functions Chemscore [10], and Goldscore [13] implemented in the docking software GOLD are commonly used for molecular docking, and were seen to achieve success rates of 68.4 % and 63.9 %, respectively, defined as a top ranked conformation ≤ 2.0 Å from the reference ligand, using a benchmark of 305 protein ligand pairs [14].

However, as with most molecular docking software the success is highly dependent on both the search algorithm and scoring function. Therefore, it is difficult to directly compare the effectiveness of scoring functions in isolation from the optimisation algorithms from reported results. One recent study, the comparative assessment of scoring functions 2016 (CASF-2016) [15], investigated the effectiveness of molecular docking scoring functions in isolation from their respective optimisation algorithms. The CASF-2016 report divides the success of scoring functions into four categories: scoring power, ranking power, docking power, and screening power. *Scoring power* is defined as the ability of the scores to linearly correlate with experimentally determined binding affinities. The *ranking power* is the ability of scoring functions to rank ligands by binding affinity. There is a subtle difference between the scoring and ranking power in that a linear correlation is not required in the assessment of ranking power. *Docking power* is defined as the ability of the scores to identify a correct conformation from a decoy set containing correct and incorrect ligand conformations. Lastly, the *screening power* is defined as the ability of the scoring functions to identify true binders from a pool containing ligands known to bind and random molecules.

The CASF-2016 test set consisted of 285 protein-ligand complexes with models derived from X-ray crystallography experiments, with a resolution of ≤ 2.5 Å and an R-factor of less than 0.25. Additionally, each protein-ligand complex was required to have reliable binding data relating to the affinity of the compounds. The scoring power of the aforementioned empirical scoring functions were assessed by the average Pearson correlation coefficient with the experimental binding affinities, the scoring functions ranked from best scoring power to worse scoring power were: X-score, AutoDock Vina, Chemscore, GLIDE, GLIDE-XP , and Goldscore with correlation coefficients of 0.631, 0.604, 0.574, 0.513, 0.467, and 0.416, respectively. In terms of ranking power, the scores were assessed using the Spearman's rank coefficient the scores were reported to be 0.604, 0.528, 0.526, 0.414, 0.284, and 0.257, for the X-score, AutoDock Vina, Chemscore, GLIDE, Goldscore, and GLIDE-XP, respectively.

Screening power was assessed with cross docking experiments and assessing the ability of each score to identify known binders from random molecules using the enrichment factor$_{1\%}$ to rank them in the top 1% of scores. The enrichment factors ranked best to worse were 11.44, 8.83, 8.65, 7.7, 4.27, and 2.68, for GLIDE, GLIDE-XP, Chemscore, AutoDock Vina, Goldscore and X-score, respectively [15]. Taken together these results indicated that the

X-score and AutoDock Vina score had the best ability to estimate the free energy of binding and rank ligands accordingly, whilst the GLIDE and GLIDE-XP scores were the best at identifying true binders from decoys.

In terms of scoring functions for ligand fitting in cryo-EM, the most relevant score is the docking power, i.e. a score's ability to identify a correct ligand conformation from an incorrect one. To generate a benchmark for the assessment of ranking power, three molecular docking softwares were used (GOLD, Surflex and MOE) to generate various binding poses for each individual protein ligand complex. Up to 100 ligand decoys were given for each case evenly spread out in RMSD from 0.0 to 10.0 Å from the deposited reference ligand. Each score was assessed by its ability to rank a correct conformation (≤ 2.0 Å of the reference ligand) as top, second or third (Figure 1). The most effective scores were seen to be the GLIDE and AutoDock Vina scores, with comparable ranking power, followed by GLIDE-XP, Chemscore, Goldscore and X-score. The AutoDock Vina and GLIDE scores were able to identify a correct ligand conformation within the top three ranked conformations 92.6 % of the time [15]. Taken together this data indicated that molecular docking scores may be accurate enough for the purposes of fitting small molecule ligands to cryo-EM data.

One method commonly used in X-ray crystallography is to use density difference mapping to identify density corresponding to the ligand once a protein model has been fit. Briefly, this technique involves the subtraction of one map from another, leaving positive and negative differences between the two maps. Recently a technique to do this based on local scaling the amplitudes of the cryo-EM map was reported [16]. In the work, the authors used two maps of a glycine receptor bound to strychnine (at 3.9 Å resolution) and to glycine/ivermectin (at 3.8 Å resolution) to show that local scaling of the map was better able to identify density corresponding to ligands than methods that globally scaled the amplitudes of the maps. The technique was also shown to identify ligand density when the two maps had a resolution mismatch, using a model of kinesin-6 motor protein bound to ADP-ALFx at 4.4 Å and a kinesin-6 motor protein in the *apo* form at a resolution of 6.1 Å. Furthermore, using a difference density map generated using a 3.2 Å map of haemoglobin and a map blurred to the same resolution using the atomic positions of the PDB deposited atomic model, showed that the technique correlated well with a local density based scoring function, the SMOC score [7], in identifying poorly fit residues in the map.

Additionally, this technique has been used to identify density corresponding to the BTB-1 bound to a kinesin-8 motor domain at 4.8 Å resolution [17], and GSK-1 bound to Eg-5 at 3.8 Å resolution [18]. In both studies molecular docking was employed to generate candidate ligand conformations that were fitted using the CCC to the difference density map.

The literature suggests that a small molecule fitting software that takes advantage of locally scaled difference maps may be advantageous. However, there is no data that compares these procedures to using the full map. Additionally, an integrative workflow that has been optimised for fitting small molecules into density difference maps has not been established.

This investigation compared the CCC and MI scores for the purposes of fitting small molecule ligands to cryo-EM maps using both difference and full maps simulated at resolutions from 2.5 to 8.5 Å. Following this a new empirical scoring function was developed and trained using a subset of the CASF-2016 dataset. The ranking power of the new empirical score was then tested and compared with the Autodock Vina score. Finally, the MI score was integrated with the empirical scoring function and the ranking power with both full and difference maps compared.

Results

Correlation with difference maps can identify correct conformations

Previous reports have shown that, when fitting small molecules into an experimental density map, combining a chemio-physical based scoring function with the fit to an experimental density map offers a high success rate [6, 19].

Fitting in this way has the effect of reducing the space needed to be searched when compared to docking alone. The experimental density acts as a spatial restraint meaning it is only required that the space around the ligand density is extensively searched, whilst solutions that provide a poor fit to the map can be instantly rejected. However, using the full density map has its drawbacks, for example as the resolution worsens it is more difficult to identify the density corresponding to the ligand.

One way to overcome this limitation is to employ the aid of difference density maps to fit the ligand in. This has the added benefit of theoretically removing density corresponding to the protein atoms, leaving only the ligand density to be searched for a good fit. Several recent publications have employed this technique for fitting small molecule ligands [17, 18].

To highlight the effect of using a difference map as a spatial restraint to fit a ligand, we utilised the CASF-2016 database containing 285 high resolution protein/ligand pairs [15]. Each protein/ligand pair had a decoy database of up to 100 ligand conformations, varying in RMSD to the deposited native structure equally spaced between 0.0 Å to 10.0 Å.

For each protein/ligand pair, density maps were calculated using the ligand only and the decoy ligand conformations were ranked by either their docking score or CCC with the calculated density maps at resolutions of 2.5, 3.5, 4.5, 5.5, 6.5, 7.5, or 8.5 Å. The ranked ligands were searched to see if a correct conformation (defined as a conformation with a heavy atom RMSD less than 2.0 Å from the native ligand) appeared in the top 1, 2 or 3 conformations.

The results showed that the docking scores were relatively powerful at ranking the conformations so that a correct conformation was ranked within the top 3 (Figure 1). However, when the conformations were ranked solely based on the CCC with the calculated density map, the percent of cases with a correct conformation in the top 3 increased compared

to the docking scores alone. Interestingly, this trend was seen at all resolutions tested, even though as the resolutions got worse, it became harder to visually identify any discriminating features regarding the correct ligand conformation (Figure 1). This effect was most likely due to the maps being simulated with no noise and free of errors common with density maps generated experimentally. However, this experiment demonstrates the potential for map fit metrics to identify correct conformations from incorrect. It is important to note here that the CCC metric in this experiment gave no indication of the plausibility, with respect to physical and chemical principles, of the ligand conformations. It does however illustrate the broader aims of this project to identify and develop a fitting algorithm that combines a physio-chemical scoring function with the fit to a difference map to refine a ligand into the binding site of an experimental density map.

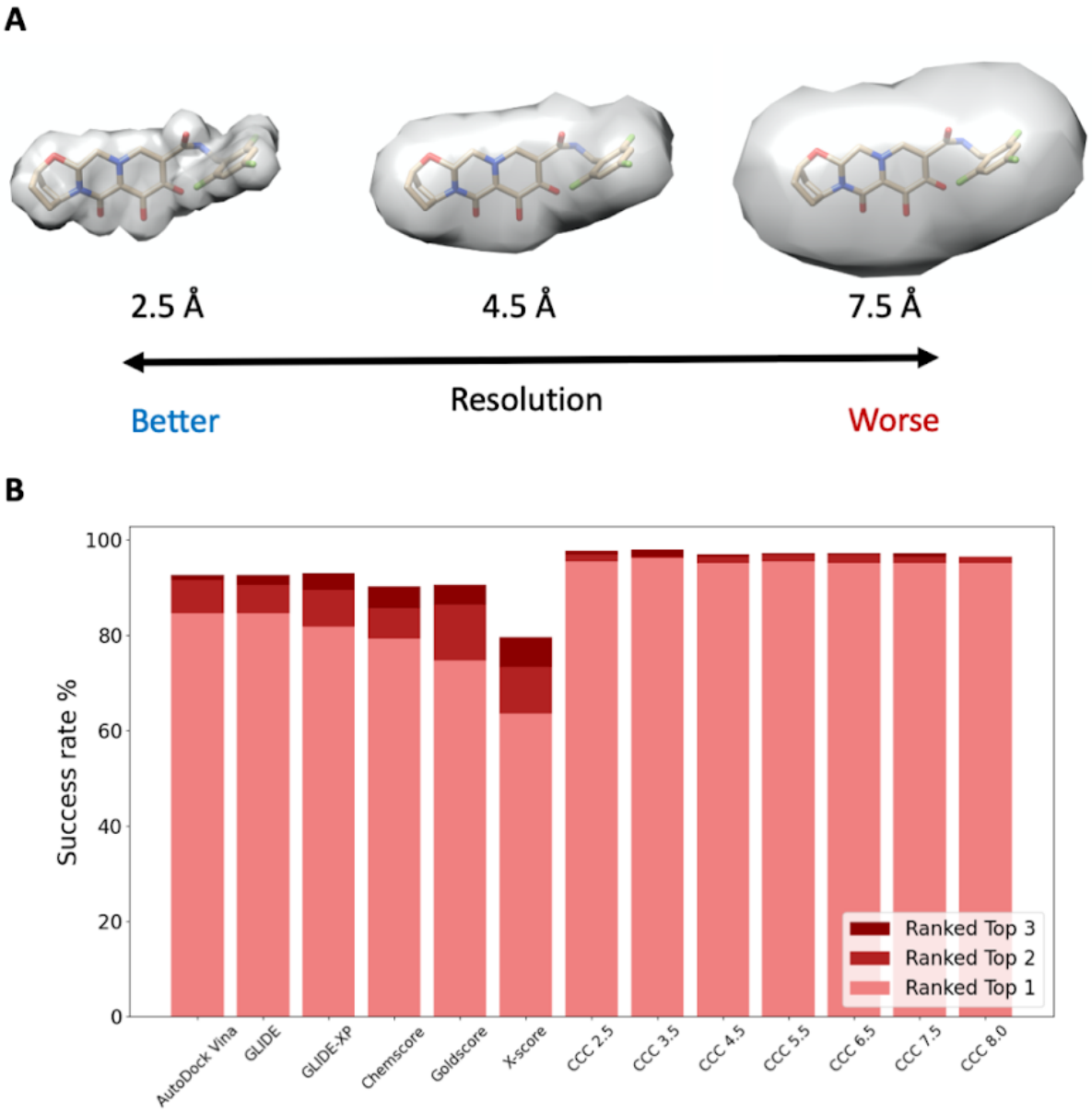


Figure 1. ***A*** *shows simulated density maps for a given ligand at 2.5, 4.5 and 7.5 Å resolution. The ligand conformation used to calculate the density is also shown for clarity.* ***B*** *shows the results of ranking ligand decoy sets of the CASF-2016 benchmark of 285 protein-ligand*

complexes with docking scores for AutoDock Vina, GLIDE, GLIDE-XP, Chemscore, Goldscore, X-score or with the CCC with a simulated density map of the deposited ligand at resolutions of 2.5, 2.5, 4.5, 5.5, 6.5, 7.5 and 8.5 Å. Bars show the percent of cases where a correct conformation (an RMSD < 2.0 Å from the deposited conformation) were ranked as the top scored (light red), second ranked scored (red), or third ranked scored (dark red) ligand.

Goodness-of-fit metrics with simulated data at various resolutions

To develop a scoring function for fitting it was necessary to identify a goodness-of-fit metric to compare the fit of ligand conformations to difference density maps. Here two of the most common goodness-of-fit metrics, the CCC and MI were investigated.

For initial experiments density maps were simulated for each of the 285 protein/ligand complexes in the CASF-2016 database [15]. Difference maps were generated prior to fitting using the protein models provided with the CASF-2016 protein/ligand pairs, and density maps were calculated according to the methodology outlined in TEMPy [20].

The CCC and MI scores between ligand conformations and simulated difference maps were evaluated for each ligand conformation in the CASF-2016 decoy set for the respective protein/ligand complex [15]. The mean Pearson correlation coefficient between CCC/MI and the RMSD was calculated across all cases for both metrics. The average Pearson correlation coefficients for MI are given in Table 1. It was seen that for both MI and CCC at each resolution tested there was a strong negative correlation between the RMSD and goodness-of-fit metric. However, at each resolution tested MI had a higher correlation. At better resolutions (2.5 to 4.0 Å) the difference between mean correlations was not statistically significant. At worse resolutions (4.5 to 8.5 Å) the Pearson correlation coefficient for the MI score remained relatively consistent, whilst the Pearson correlation coefficient for the CCC scored groups decreased. The average correlation coefficient between the two groups was seen to be statistically significant at all resolution ranges from 4.5 to 8.5 Å, indicating that the MI score was more robust to changes in the resolutions of the difference density maps.

Table 1. A Table of average Pearson correlation coefficients between the MI/CCC of simulated difference maps and RMSD of the CASF-2016 decoy sets at resolutions between 2.5 and 8.5 Å.
/home/aaron/Documents/refined_set/1qb1/1qb1_protein.mol2

Resolution (Å)	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5
MI	-0.802	-0.805	-0.808	-0.812	-0.812	-0.812	-0.813	-0.814	-0.813	-0.812	-0.810	-0.810	-0.809
CCC	-0.792	-0.789	-0.787	-0.787	-0.782	-0.777	-0.773	-0.772	-0.768	-0.766	-0.762	-0.761	-0.759
*P value**	0.41	0.24	0.14	0.067	0.038	0.017	0.008	0.005	0.0035	0.003	0.002	0.002	0.0016

** P values are calculated with a Student's T-test between the results of individual groups of MI and CCC at each resolution.*

A deeper dive into the data revealed some key differences between the two metrics. When the conformations were ranked by either the CCC or MI score it was seen that in all but one case (case at resolution 7.0 Å, Figure 2) the scores showed an equivalent amount of cases where a correct conformation appeared in the top three solutions. However, a closer look showed that at 9 of the 13 resolutions tested the MI score had a higher percent of cases with a correct solution within the top 2 solutions. This was increased to 11 of the 13 resolutions when looking for a correct solution ranked as top.

Table 2. A Table of average Pearson correlation coefficients between the MI/CCC of simulated density maps and RMSD of the CASF-2016 decoy sets at resolutions between 2.5 and 8.5 Å.

Resolution (Å)	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0	8.5
MI	-0.799	-0.799	-0.769	-0.737	-0.695	-0.664	-0.615	-0.577	-0.531	-0.503	-0.471	-0.435	-0.425
CCC	-0.797	-0.803	-0.795	-0.786	-0.771	-0.757	-0.739	-0.722	-0.700	-0.684	-0.665	-0.650	-0.635
*P value**	0.815	0.707	0.057	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

** P values are calculated with a Student's T-test between the results of individual groups of MI and CCC at each resolution.*

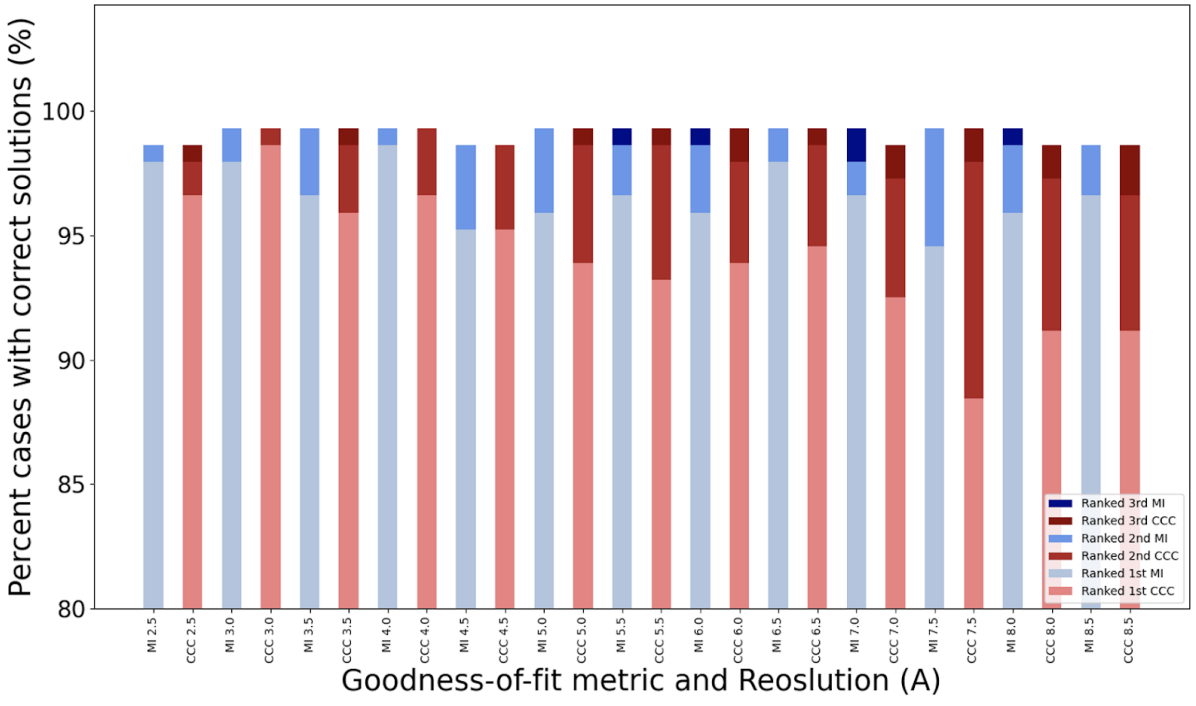


Figure 2. *The results of identifying a correct conformation (defined as a conformation with an RMSD < 2.0 Å from the deposited conformation) using the MI (blue) or CCC (red) between a difference map and simulated maps for CASF-2016 decoy ligand conformations. The percent of cases that identified a correct conformation as the top ranked (MI: light blue, CCC: light red), second ranked (MI: blue, CCC: red), or third ranked solution (MI: dark blue, CCC: dark red) is indicated at resolutions between 2.5 and 8.5 Å.*

The difference between the number of cases where a correct solution was ranked in the top 1 or 2 solutions for the MI or CCC, generally became larger as the resolution worsened. Once again indicating that the MI score was more robust at worse resolutions than the CCC, in agreement with Vasishtan *et al.* [5]. One explanation for these results is that at lower resolutions MI can differentiate between conformations that are close to each other, whilst this information is lost with the CCC and a larger distribution of conformations were being scored with a similar CCC. This effect is exemplified by looking at the top three results for an individual case (PDB ID: 3B65) , where at a resolution of 2.5 Å, both scores ranked the same decoy as top and all three of the top ranked solutions for both scores had an RMSD to the deposited ligand conformation of 0.6 Å or less (Figure 3). However, as the resolution worsened to 4.5 Å the top predicted conformation had an RMSD of 2.48 Å to the deposited conformation when the decoys were ranked by CCC. Conversely, at this resolution the top three decoys ranked by MI all had RMSDs of less than 0.53 Å to the deposited structure (Figure 3). Furthermore at a resolution of 6.0 Å the top two predicted decoy conformations by the CCC score were seen to be incorrect with RMSDs of 2.48 and 2.31 Å, whilst the MI score was able to rank three correct conformations within the top three with RMSDs less than 0.71 Å (Figure 3). It is worth noting that at all three resolutions the CCC did find a correct conformation within the top three. However, the results indicated that at low resolutions the MI score was better able to discriminate between small changes in the decoys conformations. For example, in the case of 3B65 at 6.0 Å resolution the CCC for the top three ranked decoys conformations were 0.92, 0.90, 0.89 with RMSDs of 2.48, 2.128, 0.27 Å, respectively (Figure 3). Whilst the MI scores for these same decoys at 6.0 Å resolution were seen to be 0.052, 0.052, and 0.057, respectively.

To compare the effect of using a simulated difference map in the docking process vs using a simulated full map, the experiment was repeated using full maps generated with using the protein and ligand from the deposited structures in EMDB. At all resolutions the Spearman correlation coefficient was better when the simulated difference maps were used compared to using the simulated full map (Table 2). Interestingly, when using the full maps the ability of the CCC and MI score to rank conformations by their RMSD began to steeply decline at resolutions worse than 4.0 Å. Furthermore, the CCC was seen to perform significantly better than the MI at resolutions equal to or worse than 4.0 Å. Nevertheless, the overall results indicated that using a density difference map may be beneficial to fitting small molecules compared to fitting to the full map directly.

The success rates of the MI score at identifying correct conformations were encouraging. However, the MI score alone was not capable of scoring the chemical and physical plausibility of a ligand conformation.

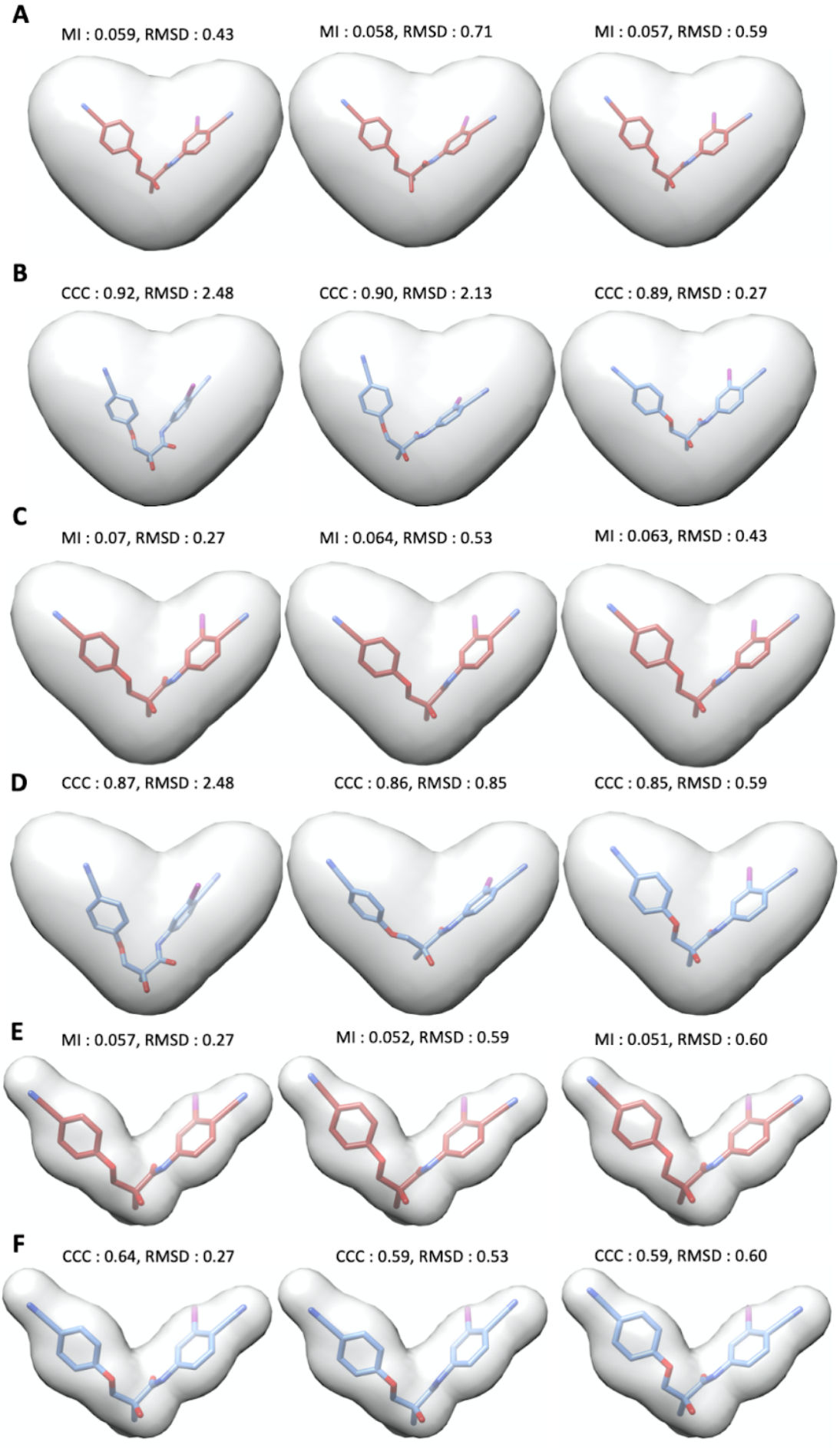

A
MI : 0.059, RMSD : 0.43
MI : 0.058, RMSD : 0.71
MI : 0.057, RMSD : 0.59
B
CCC : 0.92, RMSD : 2.48
CCC : 0.90, RMSD : 2.13
CCC : 0.89, RMSD : 0.27
C
MI : 0.07, RMSD : 0.27
MI : 0.064, RMSD : 0.53
MI : 0.063, RMSD : 0.43
D
CCC : 0.87, RMSD : 2.48
CCC : 0.86, RMSD : 0.85
CCC : 0.85, RMSD : 0.59
E
MI : 0.057, RMSD : 0.27
MI : 0.052, RMSD : 0.59
MI : 0.051, RMSD : 0.60
F
CCC : 0.64, RMSD : 0.27
CCC : 0.59, RMSD : 0.53
CCC : 0.59, RMSD : 0.60

Figure 3. *The top three ranked decoy conformations for the MI and CCC at resolutions of 6.0 Å (**A, B**), 4.5 Å (**C** , **D**) and 2.5 Å (**E, F**). The MI/CCC scores are indicated above each conformation along with the RMSDs from the reference ligand. The density maps are shown at each resolution (Grey) along with the atomic models of decoy conformations (MI: Red, CCC: Blue). Additionally, in all panels CCC/MI ranking goes from first (left most), to second (centre), to third (left most).*

Development of a empirical scoring function to identify correct conformations

The AutoDock Vina score has been shown to be one of the most accurate molecular docking scores for the estimation of free energy of binding [15]. However, it was trained on the deposited ground truth structures from the PDB-Bind dataset [21, 22] specifically for estimating the free energy of a protein-ligand complex [11]. Additionally, the score does not explicitly consider geometric or interaction terms regarding **π-π** stacking interactions, shown to be the third most common protein-ligand interaction type behind hydrogen bonds and hydrophobic interactions [23].

Therefore, a new score was derived based on the Vina score with the aim of optimising protein-ligand binding geometry. The aim was to integrate the new score with the MI score to aid in identifying chemically and physically plausible ligand conformations in cryo-EM maps.

Scoring terms

The score kept the general term for scoring atom-atom steric interactions for Vina [11]. This score is the combination of two gaussian functions (Eq 1 and Eq 2) and a repulsion function (Eq 3).

$$Eq\,1.\; gaussian^1(d) = e^{-\frac{d}{0.5}^2}$$

$$Eq\,2.\; gaussian^2(d) = e^{-\frac{d-3}{2}^2)}$$

$$Eq\,3.\; repulsion(d) = \left\{d^2\; if\; d < 0,\; else\; 0\right\}$$

$$where,$$

$$Eq\,4.\; d = d_{ij} - r_i - r_j$$

Where, d_{ij} is the distance between atomic centres of protein atom *i* and ligand atom *j*, *ri/rj* is the Van der Waals (vdW) radii of atom *i* or *j*, respectively.

An additional term for scoring hydrogen bond distance geometry was included where (Eq 5, if atom *i* is a hydrogen bond acceptor and atom *j* is a hydrogen bond donor, a potential interaction between them was scored as a fuzzy logic function of distance, as in X-SCORE [12] and AutoDock Vina [11].

$$Eq\,5.\ HBond_{distance}(d) = \left\{ \begin{array}{c} -1\ if\ d \leq -0.7 \\ 0\ if\ d > 0 \\ else,\ -1 \cdot \frac{d}{-0.7} \end{array} \right\}$$

This accounts for the increase in allowed vdW overlap caused by the formation of a hydrogen bond.

Additionally, a term to describe the hydrogen bond angle was investigated (Eq 6), where the score was a linear interpolation between 0 and 1 for hydrogen bond angles between 90 and 180.

$$Eq\,6.\ HBond_{angle}(\theta) = \left\{ \begin{array}{c} 0\ if\ \theta < 90 \\ else,\ \frac{\theta - 90}{90} \end{array} \right\}$$

To score hydrophobic interactions the hydrophobic matching algorithm was used, first outlined in SCORE [24]. This scored the local environment for a given hydrophobic atom. Hydrophobic atoms were defined as any carbon, sulphur, bromine, chlorine and iodine atoms that were bound exclusively to hydrophobic atoms or hydrogen.

The environment that a hydrophobic atom is placed in was determined as the sum of the logP of all atoms within 6.0 Å of the hydrophobic atom as in X-SCORE [12] and SCORE [24], with values for logP were taken from XlogP3 [25]. However, unlike X-SCORE and SCORE the hydrophobic score was the sum of the logP values for each hydrophobic atom in the ligand as opposed to a binary 1 or 0 score if the hydrophobic atom is or is not within a local hydrophobic environment based on a cutoff. Using the score in this way had the added benefit of punishing penalising atoms placed in a local hydrophilic environment.

To investigate the preferred geometry of π-π stacking interactions the Tough-D1 dataset [26] was used. This data set contained 3079 protein ligand complexes where the ligand was stabilised by at least 1 aromatic ring. For all protein-ligand aromatic π-π stacking interactions, data regarding the angle between interacting aromatic rings and the distance between ring centres was extracted. The data was split into two sets, one set containing distance and angle data for π-π P-stacks (angle > 45 °) and one for π-π T-stacks (angle < 45

°) (Figure 4). From this data the aim was to estimate the probability of a π-π stacking interaction being true given the ring plane angle and ring centre-ring centre distance values.

To do this probability density functions were derived by fitting the data to 106 common distributions with the python package Fitter, each fit was scored using the residual sum-of-squares criteria. The distribution that best described the angle data for both P-stack and T-stack data sets was seen to be a beta normal distribution (Figure 4C, 4D), with sum of the square residuals of 0.0037 and 0.0023, respectively.

For the distance data in the P-stack set an exponential normal distribution was seen to fit the data best with the sum of the squared residuals being 0.29. Whilst for the T-stack distance data, a skewed normal distribution was seen to fit the data best, where the sum of the squared residuals was seen to be 0.13 (Figure 4E, 4F).

The geometry score for a given π-π stack during fitting was the product of the probability density functions for the angle and distance. The probability density functions to use were determined by the plane angle of the two rings. Where the plane angle was less than 45 ° the probability density functions derived from data for P-stacks were used, otherwise the probability density functions for T-stacks were used.

The results showed that the aromatic fitting term was a reasonable approximation of the experimental data for **π-π** stacks (Figure 4G, 4H). A 2-dimensional histogram of the experimental data showed that there were two distinct regions where most of the interactions were clustered (Figure 4G). One where the ring centre-centre distance was between 3.5 Å and 4 Å with a ring plane angle between 0 ° and 15 °, a second high density cluster was seen where the distance was between 4.25 Å and 5.0 Å with ring plane angles between 75 ° and 90 °. The first of these clusters representing the P-stacks correlated well with the distance geometry for **π-π** stacks previously reported [23], where the distance between ring centres was seen to be approximately equal to twice the value of the vdW radii for carbon (~3.4 Å). The distance values were slightly larger for T-stacks but correlated well with values observed in benzene ring T-stacks with ring centre-centre distances measured at 4.96 Å [27]. The scoring function for the geometric parameters of **π-π** stacking interactions correlated well with the experimental data where the best scored geometric parameters for a P-stack were between 3.5 Å and 4.0 Å with angles between 0° and 5 °, and for a T-stack distances between 4.5 Å and 5.0 Å with angle deviations between 85 ° and 90 ° scored highest. One deviation in the calculated scores from the experimental data was that the angle deviation was more stringent with allowed angle deviations from the 'perfect' geometry (0 °, 90 ° for P- and T-stacks, respectively) of 5 °, compared to the 15 ° deviation seen in the experimental data. This increased stringency comes from the exponential portion of the beta distribution probability density functions used to fit the angle values. Furthermore, the experimental values come from a large selection of aromatic ring types, with varying ring size and -R groups thereby making the calculated scoring term is a general term for aromatic interactions, it may be possible that different ring types have slightly different preferred geometries that we do not account for in our scoring term.

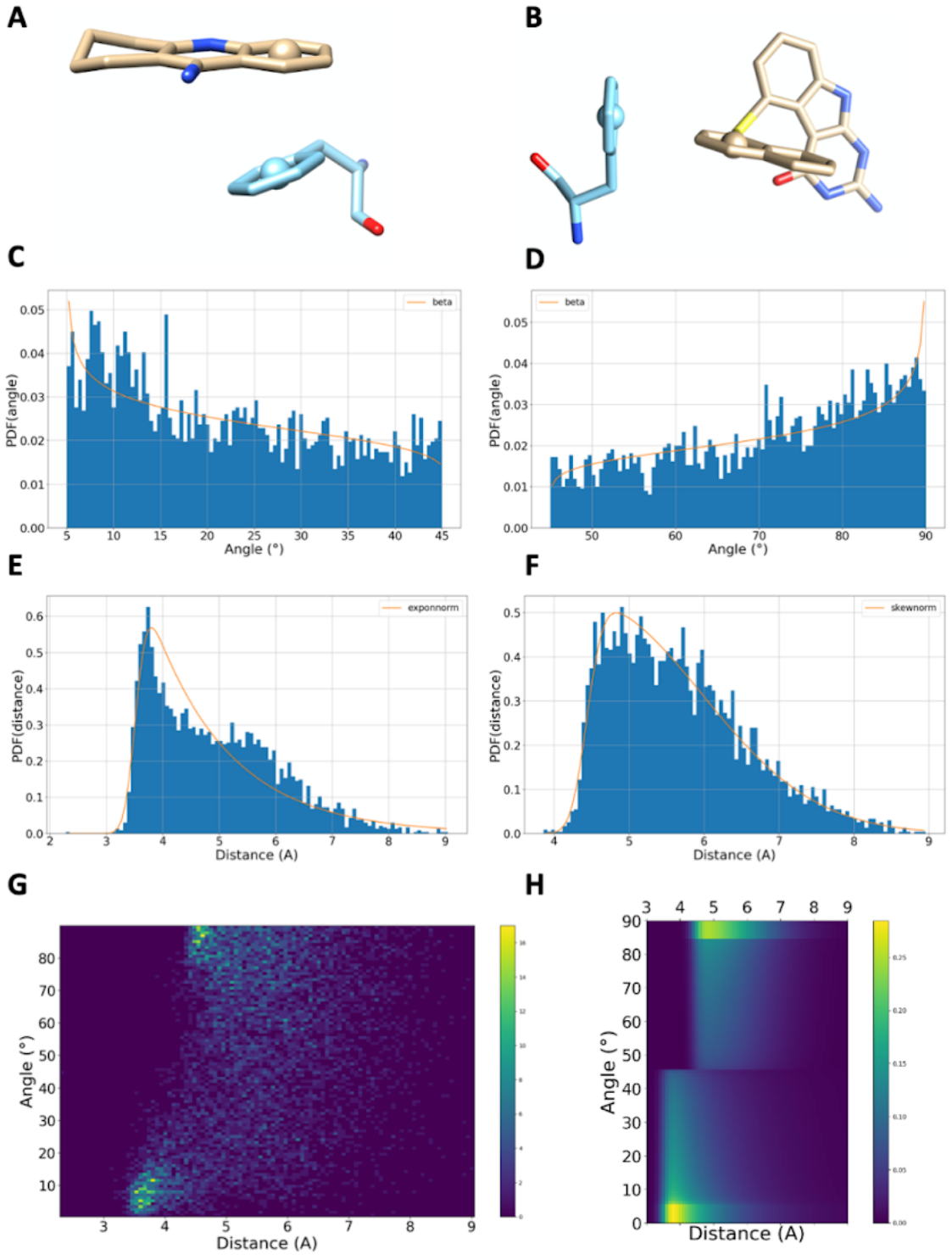


Figure 4. *The results of developing an aromatic scoring term for* ***π-π*** *stacking geometry. Two types of* ***π-π*** *stack were explicitly considered, the P-stack (**A**) and T-stack (**B**). histograms of the plane angles between pairs of aromatic rings for the P-stack (**C**) and T-stack (**D**) datasets are shown fit to Probability density functions (orange line). Additionally, histograms of aromatic ring-ring centres for P-stacks (**E**) and T-stacks (**F**) are also shown fit to their respective probability density functions (orange lines). A 2-dimensional histogram of angle*

and distance data for ***π-π*** *stacks in the Tough-D1 dataset is shown (**G**). The colour bar is shown to the right where the colour represents the number (n) of interactions in the dataset with corresponding distance and angle data, yellow indicating a higher n (~ 36) whilst blue indicates a low value of n (~ 0). A 2-dimensional plot of the output values of the aromatic scoring term (**H**) is shown for distances between 3 and 9 Å, and angles between 0 and 90 °. A colour bar shown to the right of the plot corresponds to the output values of the aromatic term, yellow values indicate a higher score (~0.3) and blue values indicate a lower score (~0.0).*

Finally, to ensure atoms within the ligand do not clash with themselves, a Lennard Jones 8-4 potential was included and calculated as the sum over all ligand atoms that were not covalently bonded to each other (Eq 7).

$$Eq\,7.\, ligand_{intraVDW} = \sum_{i}^{ligand} \sum_{j}^{ligand_{nb}} \left[\left(\frac{d_{i_r,j_r}}{d_{i,j}} \right)^8 - 2 \left(\frac{d_{i_r,j_r}}{d_{i,j}} \right)^4 \right]$$

Where i is the set of ligand atoms, *j* is the set of ligand atoms not covalently bonded to ligand atom *i*. d ir,jr is the sum of the vdW radii of ligand atoms i and j. Di,j is the Euclidean distance between atoms centres of ligand atom i and j.

Optimising the scoring weights

The ability of four scoring functions to identify a correct ligand conformation from an incorrect one was investigated. The Lennard-Jones 8-4 potential, the Vina steric and gaussian terms, the hydrogen bond distance term and the hydrophobic terms made up the basis of the scoring function (*score 1*), to this we added the, aromatic scoring term (*score 2*), the Hydrogen bond angle term (*score 3*) , or both the Hydrogen bond angle term and the aromatic scoring term (*score 4*).

$$Score\ 1 = w_1 L_{VDW} + w_2 V_{g1}(d) + w_3 V_{g2}(d) + w_4 V_s(d) \\ w_5 HB_d(d) + w_6 HPI(lp)$$

$$Score\ 2 = w_1 L_{VDW} + w_2 V_{g1}(d) + w_3 V_{g2}(d) + w_4 V_s(d) \\ w_5 HB_d(d) + w_6 HPI(lp) + w_7 Arm(\theta, d)$$

$$Score\ 3 = w_1 L_{VDW} + w_2 V_{g1}(d) + w_3 V_{g2}(d) + w_4 V_s(d) \\ w_5 HB_d(d) + w_6 HPI(lp) + w_8 HB_\theta(\theta)$$

$$Score\ 4 = w_1 L_{VDW} + w_2 V_{g1}(d) + w_3 V_{g2}(d) + w_4 V_s(d) \\ w_5 HB_d(d) + w_6 HPI(lp) + w_7 Arm(\theta, d) + w_8 HB_\theta(\theta)$$

Where, w1-8 are the various weights that each term contributes to the total score. LVDW is the Lennard-Jones 8-6 potential term, Vg1/Vg2 were the AutoDock Vina gaussian terms 1 and 2, Vs is the AutoDock Vina steric term, HBD is the hydrogen bond distance term, HPI is the hydrophobic interaction term, HBA is the hydrogen bond angle term, and Arm is the aromatic **π-π** stacking term.

To determine the weights for each term 101 structures from the CASF-2016 dataset were used [15]. From each structure 50 near-native ligand conformations (< 1.0 Å RMSD from the deposited conformation) were generated using PLANTS docking software [28]. A further 950 random conformations with RMSDs ranging from 1.0 Å to 10.0 Å from the deposited ligand conformation were generated using an in-house python script. The Nelder-Mead algorithm was used to minimise the average Pearson correlation coefficient between the square-root of the RMSD and the scoring function for each of the 101 protein ligand sets. The average Pearson correlation coefficients were -0.747, -0.749, -0.735, and -0.739 for score 1, 2, 3, and 4, respectively (Table 3).

Table 3. A table of weighted values for each term in scores 1-4, along with average Pearson correlation coefficients for the training set.

	Score term									
Score	Intra VDW	Vina gauss 1	Vina gauss 2	Vina steric	Hbond dist	Hbond angle	Hydrophobic	Aromatic score	converged	Correlation
1	-0.014	0.343	0.037	-0.678	-2.48	-	0.166	-	True	-0.747
2	-0.004	0.10	0.011	-0.205	-0.79	-	0.047	7.03	True	-0.749
3	-0.054	1.93	0.180	-3.478	-1.09	2.303	0.241	-	True	-0.735
4	-0.085	1.25	0.110	-2.52	-2.16	2.65	0.80	1.98	True	-0.739

The values for the weights were seen to be slightly different for each score group. This was assumed to be due to each score containing different terms and the relationship between the weights and terms changing with the addition of new terms. One pattern we observed that supports this hypothesis was the relationship between the Vina gaussian and steric terms. For score groups 3 and 4 the values were seen to be higher by a magnitude of 10x when compared to score groups 1 and 2 (Table 3). However, the relationship between the three variables was conserved in all three groups. Where the values of the Vina gaussian 1 weights were approximately of the order of 10x the magnitude of the weights for Vina gaussian 2 terms. Furthermore, in all four cases the values of the Vina steric weights were approximately of the opposite sign of the vina gaussian 1 and 2 weights and approximately -1 * 2x the magnitude of the Vina gaussian 1 weights.

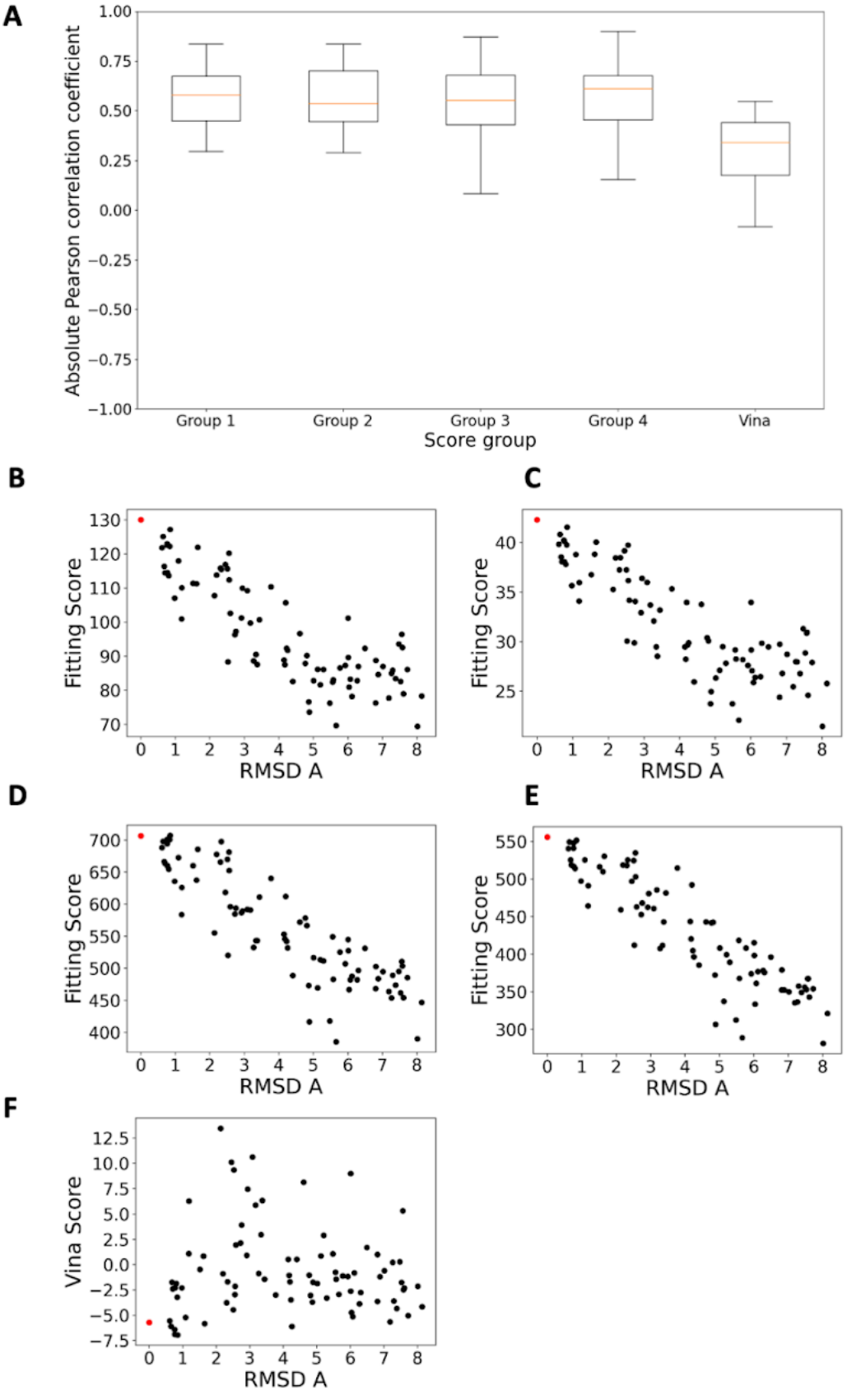

A
Absolute Pearson correlation coefficient
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Group 1
Group 2
Group 3
Group 4
Vina
Score group
B
Fitting Score
RMSD A
C
Fitting Score
RMSD A
D
Fitting Score
RMSD A
E
Fitting Score
RMSD A
F
Vina Score
RMSD A

Figure 5. *The absolute values of Pearson correlation coefficient for scores 1-4 and the AutoDock Vina scoring function (**A**). The individual scores for all the decoys in the individual case of 2CET for the test set used are shown for score 1 (**B**), 2 (**C**), 3 (**D**), 4 (**E**), and the AutoDock Vina scoring function (**F**). Scores representing the reference ligands are also indicated (Red dots).*

To assess the robustness of the weighted scoring functions the Pearson correlation coefficient between the RMSD and the weighted scoring function for 23 additional protein ligand complexes and decoy sets were assessed. The decoy sets for each protein ligand complex contained up to 100 ligand conformations spaced out in RMSD from the deposited structure from 0.0 Å to 10.0 Å . However, in contrast to the datasets used to train the scores, the conformations were generated using the GOLD docking program with ligands scored using Chemscore. Thus, the conformations were chemically reasonable and contained few errors with regards to atom-atom clashes or unreasonable geometry, as such this benchmark provided a much sterner test for the scoring functions than the benchmarks they were trained on.

The results showed that it was harder for each scoring function to identify a correct conformation, with average absolute Pearson correlation coefficients of 0.51, 0.51, 0.50, 0.52 for scores 1, 2, 3 and 4, respectively (Figure 5A). This is a marked decrease in the accuracy seen during fitting of the weights. However, the dataset represented a much harder test for the scoring functions. When the same dataset was scored with the AutoDock Vina scoring function the correlation was seen to be 0.30 (Figure 5A).

The individual scores per case, indicated a reason for the improved correlation seen compared with the AutoDock Vina scoring function. It seemed Vina had more trouble in identifying conformations far (with respect to RMSD) from the deposited binding conformation, when the decoys conformations were chemically reasonable. Whilst scoring functions 1-4 still had this issue to some extent, conformations far from the RMSD of the deposited solution were generally scored lower than conformations closer to the deposited solutions. An example for the test protein ligand complex 2CET from the benchmark is shown in Figure 5, the full results of the benchmark are shown in Figures A2-6.

The scoring functions were trained specifically to identify a good conformation from a bad one, and not specifically to estimate the binding affinity of a compound, however, the correlation of the scores with experimentally determined binding affinities for each of the deposited ligands in the test set was investigated (Figure 6A-D). Interestingly, each score showed some measure of correlation with experimentally determined affinities, with Pearson correlation coefficients of 0.695, 0.696, 0.666, and 0.678 for the score groups 1,2,3 and 4, respectively. The correlation of the Vina score with experimental binding affinities for the test set deposited structures was seen to be -0.392 (Figure 6E). Taken together the results indicated that our scores were better able to discriminate between correct and incorrect ligand conformations and rank ligands by binding affinity than the AutoDock Vina score.

All four scoring functions showed comparable docking power with respect to identifying correct and incorrect conformations within the test set. The addition of the aromatic scoring term increased the accuracy of score 4 compared to score 3 in our test set whilst it had little effect when added to score 1. Due to this, score 4 was taken forward for further analysis and integration with the MI score.

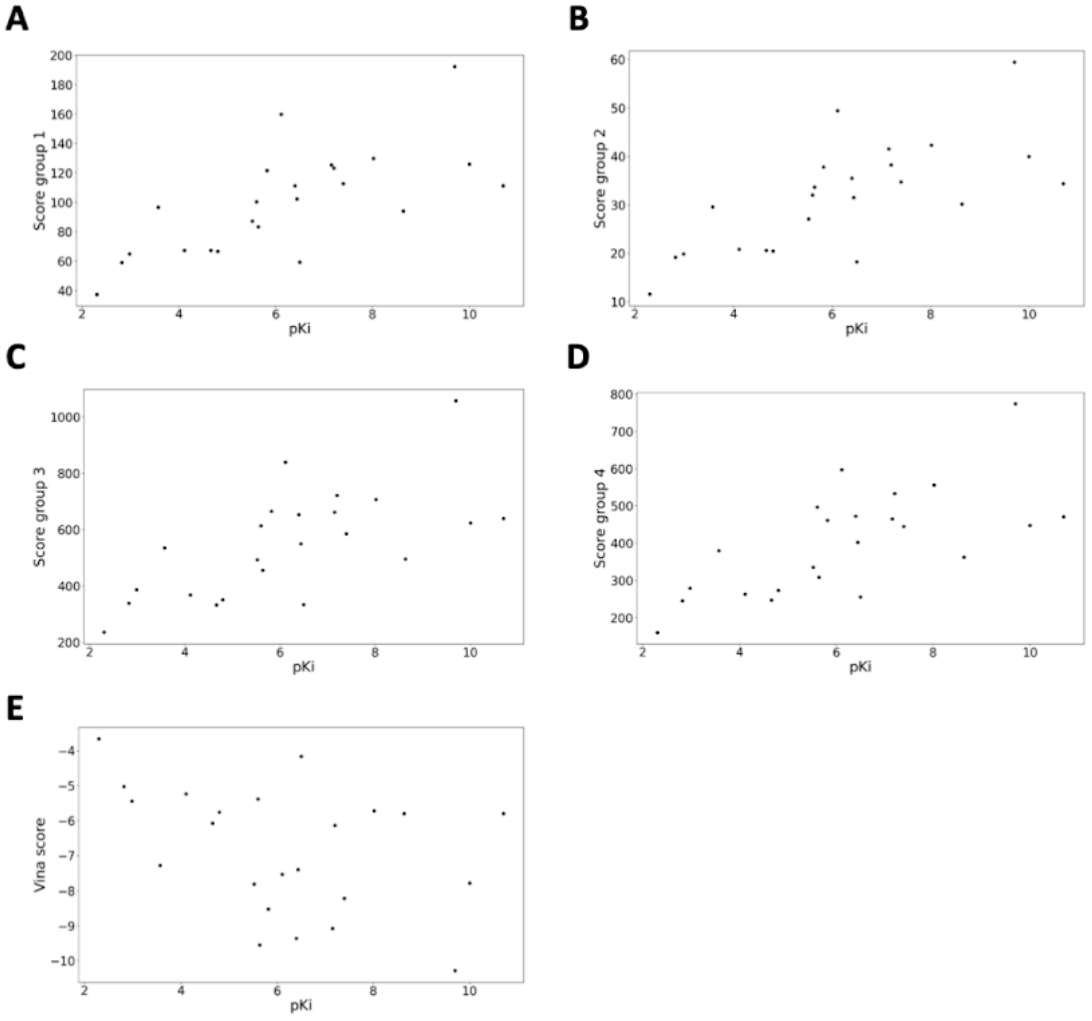


Figure 6. *Correlation of scores 1 (**A**), 2 (**B**), 3 (**C**), 4 (**D**) and the Vina score (**E**) with experimentally determined binding affinity (pKi) for the deposited ligands in the test set. Scores 1-4 are arbitrary units, whilst the Vina score is given in units of kcal/mol.*

Integrating the MI score with the empirical scoring function

To integrate the empirical scoring function (score 4) with the MI score, simulated difference density maps and full maps for the 23 test cases were calculated at resolutions ranging from 2.5 Å to 8.5 Å. Each decoy conformation in the individual test cases were scored by the empirical scoring function and the MI score against the simulated density difference map and full map. Scores were then integrated as the sum of the two scores and where the MI score

was weighted by approximate factors of 0.1, 0.5, 1.0, 5.0 and 10.0x the magnitude of the empirical scoring function (Eq 8):

$$Eq\,8.\ Integrated\ score = (w_{MI} \cdot MI + empirical\ score)$$

Where w_{mi} was the weight for the MI score.

For the empirical scoring function, the scores were usually of a magnitude within the range 0 to 1000. For the density difference maps, the MI scores were within the range of 0.1 and 0.01, therefore for this group the MI scoring weights were 1000, 5000, 10000, 50000, and 100000. In the case of the full map, the MI scores were approximately an order of magnitude lower than for the density difference maps. Thus, the weights were 10000, 50000, 100000, 500000 and 1000000.

Following this, the decoys were ranked by the integrated scores and the Pearson correlation coefficient was calculated for the integrated score and the RMSDs to the deposited ligand. For each experimental condition, the average Pearson correlation coefficient across all test cases was used as an indication of the docking power of the integrated score under that respective condition.

At all resolutions investigated, the density difference map results showed that any combination of the MI score with the empirical scoring function was able to significantly improve the correlation between the score and RMSD compared to the empirical scoring function alone (Table 4). Additionally, at each resolution when the MI was weighted at 5x and 10x the empirical scoring function a slight increase in the average Pearson correlation coefficient was seen. However, this effect was not statistically significant.

The results using the full maps showed a similar trend where at each resolution range tested at least one integration score was better than the empirical scoring function or the MI score alone (Table 5). At resolutions better than 4.0 Å the best integrated score was with an MI weight of 10x. This was similar to what was observed in the density difference map groups, and is most likely a consequence of having very well defined atom densities at this resolution range.

However, at resolutions worse than 4.0 Å, the integrative score that gave the best average Pearson correlation coefficient became more and more dependent on information from the empirical scoring function. Where, the best integrative scores at 4.5 Å had an MI score weight of 5x. At 5.0-5.5 Å this dropped to a weight of 1x, and at 6.0-6.5 Å was 0.5x. Below this resolution range the best integrative scores had an MI weight of 0.1x. This indicated that there is less and less information within the map and physico-chemical information from the empirical scoring function was needed.

Interestingly, this effect was not seen with the density difference maps indicating the presence of protein density at low resolutions was making it more difficult for the MI score to differentiate between correct and incorrect conformations. This correlated well with previous experiments comparing difference density maps with full maps using the MI score alone (Table 4 , Table 5).

Table 4. The average Pearson correlation coefficients of the test set with the integrated scores at resolutions from 2.5 to 8.5 Å using the difference maps

	weight of the MI score to the empirical scoring function						
Resolution	0x	0.1x	0.5x	1x	5x	10x	MI only
2.5	-0.520	-0.642	-0.767	-0.786	-0.791	-0.790	-0.789
3.0	-0.520	-0.643	-0.772	-0.791	-0.797	-0.796	-0.795
3.5	-0.520	-0.641	-0.776	-0.797	-0.805	-0.804	-0.804
4.0	-0.520	-0.636	-0.776	-0.800	-0.809	-0.808	-0.807
4.5	-0.520	-0.635	-0.777	-0.802	-0.810	-0.810	-0.808
5.0	-0.520	-0.632	-0.777	-0.803	-0.811	-0.811	-0.810
5.5	-0.520	-0.627	-0.777	-0.806	-0.816	-0.816	-0.814
6.0	-0.520	-0.624	-0.776	-0.808	-0.819	-0.818	-0.816
6.5	-0.520	-0.621	-0.774	-0.808	-0.819	-0.818	-0.816
7.0	-0.520	-0.616	-0.771	-0.808	-0.820	-0.819	-0.818
7.5	-0.520	-0.612	-0.767	-0.805	-0.818	-0.817	-0.815
8.0	-0.520	-0.609	-0.764	-0.805	-0.821	-0.820	-0.818
8.5	-0.520	-0.608	-0.764	-0.805	-0.820	-0.819	-0.817

Table 5. The average Pearson correlation coefficients of the test set with the integrated scores at resolutions from 2.5 to 8.5 Å using the full maps

	weight of the MI score to the empirical scoring function						
Resolution	0x	0.1x	0.5x	1x	5x	10x	MI only
2.5	-0.520	-0.546	-0.629	-0.685	-0.767	-0.774	-0.772
3.0	-0.520	-0.558	-0.659	-0.714	-0.773	-0.776	-0.776
3.5	-0.520	-0.568	-0.666	-0.706	-0.740	-0.743	-0.743
4.0	-0.520	-0.575	-0.662	-0.689	-0.708	-0.709	-0.709
4.5	-0.520	-0.568	-0.633	-0.648	-0.652	-0.651	-0.649
5.0	-0.520	-0.564	-0.617	-0.626	-0.621	-0.618	-0.614

5.5	-0.520	-0.555	-0.587	0.587	-0.570	-0.565	-0.558
6.0	-0.520	-0.550	-0.575	-0.571	-0.547	-0.541	-0.533
6.5	-0.520	-0.543	-0.549	-0.536	-0.500	-0.492	-0.482
7.0	-0.520	-0.540	-0.535	-0.513	-0.462	-0.451	-0.439
7.5	-0.520	-0.533	-0.515	-0.485	-0.423	-0.410	-0.396
8.0	-0.520	-0.526	-0.488	-0.447	-0.374	-0.361	-0.347
8.5	-0.520	-0.529	-0.491	-0.454	-0.388	-0.376	-0.363

Discussion

For fitting small molecules in cryo-EM maps it is first necessary to evaluate the performance of fitting metrics related to both the goodness-of-fit to the experimentally derived map and the chemio-physical reasonability of the ligand conformations. The work presented here identified the MI score as having a comparative power to the commonly used CCC in assessing the goodness-of-fit to a density map. Furthemore, it was shown at lower resolutions the MI score was more robust than the CCC when using density difference maps. Following this, a new empirical scoring function was parameterized specifically for identifying correct conformations. It was shown to have a better *scoring power* and *docking power* than the autodock Vina scoring function. Additionally, it was one of the first examples of empirical scores to include parameters relating to the geometry of aromatic interactions explicitly. The results and insights presented here will be beneficial for the development of future small molecule fitting algorithms.

First we showed that the ranking power of the CCC was significantly better than that of five common docking scores (Figure 1). Whilst this experiment was a relatively obvious result as we used simulated density maps from the deposited ligand, it illustrated a key point of this investigation quite well where density maps essentially act as spatial restraints to reduce the search space when compared to purely using computational techniques such as molecular docking. This effect has recently been shown in one report that combined the GLIDE docking software [8, 9] with the CCC for fitting small molecules into cryo-EM maps [6] where it was seen that adding the CCC to the GLIDE docking score resulted in solutions significantly closer in RMSD to the deposited ligand conformation compared to using the GLIDE software alone.

We, however, investigated here the use of two common metrics the CCC and MI score for ranking the fit of small molecules across a larger range of resolutions using simulated difference density maps (rather than whole maps in the previous studies [5, 7]) (Figure 2). It was seen that at all resolutions the MI score had a better Pearson correlation coefficient with the RMSD of the deposited molecule than did the CCC. However, this result was only seen to be significant at resolutions worse than 4.5 Å (Table 1).

This result correlated well with previous reports that have compared the MI score with the CCC for fitting protein into cryo-EM maps, where it was seen that the MI score had better or comparable correlation with the RMSD of proteins for both simulated and experimental maps [5]. Furthermore, the MI score was seen to be more resistant to changes in resolution than the CCC.

When the experiment was repeated using simulated full maps it was seen that at resolutions up to 4.0 Å the MI and CCC had comparative ranking power. However, at resolutions worse than 4.0 Å the ranking power of the CCC was statistically better than the MI score. This was unexpected when taken together with our results using the simulated difference maps and previous published literature. One explanation of these results may be related to the process of binning density values when calculating the MI score. The process of binning densities is subjective and arbitrary but necessary to successfully use the MI score. For difference maps the previously used bin value of 20 [5] bins was used. When using difference maps this value appears to yield good results as there are fewer voxel amplitude values to consider as density values of the protein are removed due to the difference mapping process. However when using full maps there is a greater distribution of density values from the ligand and protein densities. Additionally, at lower resolutions the difference in amplitudes of densities are much closer together. The small molecule represents a very small part of the density with a closer distribution of density values. It may be the case that the number of bins was not large enough for the MI score to pick up on subtle changes in the positions of ligands. This is evidenced by the fact that at lower resolutions where the difference in amplitudes between density values is at its highest the MI score performance was comparable to the CCC.

Whilst the MI score was shown to be superior to the CCC at identifying a correct fit of small molecule ligands to a difference density map, it does not explicitly give information regarding the chemio-physical reasonability of the ligand conformation. Therefore, a new empirical scoring function was parameterized, that with our test set was seen to correlate better with the RMSD and experimentally determined binding affinity than AutoDock Vina.

A new term for scoring **π-π** stacking interactions based on statistical data from experimentally determined protein ligand structures (Figure 4) was introduced. It's worth pointing out that the aromatic interaction term did not aim to directly estimate the interaction energy of aromatic rings. Instead interaction energy was handled implicitly by the hydrophobic and steric terms. The aromatic term was more of a correction term aimed at ensuring the preferred geometry of **π-π** stacking interactions was scored . The aromatic scoring term was able to replicate the geometry seen in the experimentally determined structures of the Tough-D1 benchmark very well (Figure 4) for both P- and T-stacks. However, when the weights were determined for the scores the weights for the aromatic terms whilst of the same magnitude were different from each other being 7.05 and 1.98 for score groups 2 and 4, respectively. Considering that the weights for the steric, vina gaussian 1 and vina gaussian 2 terms were an order of magnitude lower than the weights found for score group 4, this was rather unusual and indicated that score group 2 would weight aromatic interactions much higher than for group 4. It may also have indicated that the addition of the

aromatic scoring term to score 2 may have had little effect. During the training the addition of the aromatic scores to score 1 and score 3 resulted in a marginal increase in the average Pearson correlation coefficients (Table 4) and scores 1 and 2 showed slightly better performance compared to scores 3 and 4. When the scores were investigated with the test set a significant decrease in performance was seen compared to the training set (Figure 5). This decrease in performance may indicate that the scores generalised poorly. However, the decoy set used was a much harder test for the scoring functions, since the decoy set represented chemically reasonable conformations, it was less clear which conformation was incorrect and was an alternative explanation of the decrease seen in the performance.

To evaluate the docking power of the scores, the test set was scored with the AutoDock Vina scoring function, shown to be one of the best empirical scores for docking power [15]. All four scores were seen to outperform the AutoDock Vina scoring function (Figure 5). This indicated that the decrease in performance was most likely due to the tougher benchmark than the scores generalising poorly.

With the test data set scores 3 and 4 showed better average Pearson correlation coefficients than scores 1 and 2, however, the difference was marginal. Additionally the addition of the aromatic term to score 2 showed no improvement over score 1. Whilst a slight improvement was seen when the aromatic term was added in score 4 (Figure 5). Due to this score 4 was taken forward for further investigation.

Our results then indicated that combining the MI score with the empirical scoring function, score 4, could increase the ranking power when compared to the MI score and empirical scoring function alone (Table 4, Table 5). When density difference maps were used at all resolutions, the best MI weights were seen to be either 5x or 10x the magnitude of the empirical scoring function. This indicated that the MI score contained most of the information regarding the correct conformation of the ligand. One limitation of this investigation was that it was conducted using simulated maps. These maps were completely free from errors in the calculation of the maps. This is not the case when experimental density maps are calculated, where errors can be introduced at many stages including, alignment of 2D classes, calculation of 3D densities and protein modelling errors, all of which lead to maps with a significantly higher signal-to-noise ratio. Therefore, it may be the case that the results seen here are somewhat artificially inflated due to the lack of noise in the maps.

Interestingly, a trend was seen where the average Pearson correlation coefficient actually increased as the resolution became worse, with integrative scores where the MI score was weighted above 1x and the MI score alone. This effect was also seen to a lesser extent with our investigation into the MI score alone (Table 1). One explanation for this could be related to the volume of the density and lack of noise. At worse resolutions the volume of the density is increased (Figure 6). This could lead to the situation where the top fitting conformations are still being ranked correctly, however, decoy conformations further apart from the ground truth, with respect to RMSD, are being artificially ranked better at lower resolutions. This would be due to the increased size of the volumes at worse resolutions resulting in regions of

decoy conformations far from the RMSD of the ground truth overlapping with regions of density not present in the difference maps calculated at lower resolutions.

Evidence to support this hypothesis comes from the results of the integrative scores calculated with the full maps (Table 5), where it was seen that at all combinations of MI and empirical score and the MI alone the average Pearson correlation coefficient decreased as the resolution worsened. The only major difference between these experiments was the presence of density corresponding to protein atoms. Therefore at lower resolutions the situation does not arise where decoys further away are now overlapping with the increased volume of ligand density due to the presence of protein density. If this were the case then it would be expected that as the resolution worsened the integrative scores would rely on the geometrical and chemical information contained within the empirical scoring function, which is exactly the trend observed (Table 5).

Taken together the results suggest that the integrated score developed here was effective at fitting small molecules into simulated density maps. It was also seen that there was an advantage in performance of ranking power when simulated difference maps were used over simulated full maps.

However, the simulated maps used were free of noise and it would be necessary to test the scoring function using maps containing noise that represented the experimental situation more accurately.

The 2D images used to derive 3D density maps during cryo-EM experiments have a low signal-to-noise ratio. It is also reasonable to assume that most of the noise in these 2D-micrographs is gaussian. Therefore one of the simplest ways to incorporate noise is to add a 2D-micrograph of gaussian noise to a back projected image derived from the noise free simulated 3D map (ensuring the amplitudes of both are scaled until the desired signal-to-noise ratio is achieved). These 2D-micrographs containing noise can then be reconstructed to form a 3D-map with a noise level similar to that seen with experimentally derived maps. Such a protocol was used to generate benchmark images with known quantities of noise and signal for aligning particularly noisey micrographs [29].

An alternative to generating pure noise micrographs computationally is to use 2D-micrographs consisting of pure noise taken from experimental images. Noisey images can be constructed from regions adjacent to particles, where it is assumed no particles are present, and added to back projected micrographs from the simulated density map. Such a protocol was developed in a reported protocol for reducing overfitting of high resolution frequencies when computing 3D-density maps [30].

Whilst the two methods presented above represent an accurate way to determine background noise they fail to incorporate noise from other sources, such as structural noise coming from sample conformational heterogeneity. One report has suggested a methodology for obtaining a benchmark dataset that includes structural noise. First back projected images were obtained

from multiple atomic models that vary slightly in conformation. The images are then combined and gaussian noise added. To simulate noise introduced by the microscope the images were subjected to modulation by a contrast transfer function, before another layer of gaussian noise was added. The 2D micrographs were shown, visually, to be a good representation of experimentally determined micrographs, with a similar signal-to-noise ratio [31]. The future direction for this investigation would be to assess the power of the scoring function on a benchmark of simulated density maps that accurately represented experimental noise.

Methods and software

Preprocessing of the CASF-2016 benchmark

The software presented here was mostly implemented in python using the RDKit module to handle small molecule ligand operations. Thus the ligands included in the CASF-2016 benchmark set were pre-processed for use with RDKit. The deposited and decoy ligands supplied with the CASF-2016 benchmark were supplied in mol2 format. The RDKit does not have great support for reading '.mol2' files, therefore OpenBabel was used to convert '.mol2' files to '.sdf' file format. Hydrogens were added to each ligand and corresponding decoys set using RDKit. During this procedure a number of ligands failed to load into the rkit module, this was due to the ligands coming initially from pdb file formats then being converted to mol2 files formats in the CASF-2016 dataset. PDB file formats do not contain adequate information regarding ligand atom-atom bond information and must be inferred from distances when converted to mol2 files. The most common reason for ligands failing to be loaded into the RDKit module was related to unexpected valencies on atoms. This indicated that there was inadequate bond typing information when they were converted from mol2 to sdf file format. From the 285 initial test cases 124 were successfully loaded into the RDKit module, which was the CACSF-2016 subset used in our investigations.

Calculating density maps

Density maps were calculated using the Chimera [32] molmap function. Either protein, ligand or protein ligand complexes were loaded into chimera and hydrogens removed. Simulated maps were generated using the '*molmap'* command. The molmap command calculates density maps as the sum of gaussian functions for each atom.Where the width of the gaussian is proportional to the resolutions and the amplitude is proportional to the atomic weight.

Calculating difference density maps

Difference density maps were calculated using the local scaling method [16] implemented in TEMPy [20]. The input to the software were the simulated full maps generated with Chimera,

the protein pdb file in the *apo* state, and the resolution. For all options the default setting was used.

Scoring function implementations

The scoring functions used here (Score 1-4) were implemented in python in a new software. The software takes an input of a protein file in PDB format, a ligand file or list of ligand files in sdf format, and a binding site centroid and radius. For all experiments the centroid of the deposited ligand was used to centre the binding sites, along with a radius of 12.5 Å. All residues within the protein that had at least one atom within the binding site were taken as binding site residues, and scores were calculated using this set of atoms and the equations described in the main body of the text.

Calculations

RMSD calculations were completed using functions that were built into the software. The RMSD indicated here was calculated across all heavy atoms (i.e. excluding hydrogens), using the formula (Eq 9):

$$Eq\,9.\ RMSD = \sqrt{\frac{1}{n}\sum \|x_i - y_i\|^2}$$

Where, xi, yi are ligand atom pairs in the ligands x and y, respectively.

Implementations of the CCC and MI scores found in TEMPy [20] were used for goodness-of-fit calculations.

All t-tests were conducted using the python 3 package scipy [33] (V1.7.2) with the 'scipy.stats.ttest_ind' method. Pearson and Spearman rank correlations were calculated with the same package, using the 'scipy.stats.personr' and 'scipy.stats.spearmanr' methods. The Neadler-Mead minimisation was conducted using the 'scipy.optimise.minimise' method with the 'method' argument set to 'Neadler-Mead'.

References

1. Tagari M, Newman R, Chagoyen M, Carazo JM, Henrick K. New electron microscopy database and deposition system. Trends Biochem Sci. 2002;27:589.

2. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein Structure Fitting and Refinement Guided by Cryo-EM Density. Structure. 2008;16:295–307.

3. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42–9.

4. Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A, et al. Real-space refinement in PHENIX for cryo-EM and crystallography. Acta Crystallogr D Struct Biol. 2018;74 Pt 6:531–44.

5. Vasishtan D, Topf M. Scoring functions for cryoEM density fitting. J Struct Biol. 2011;174:333–43.

6. Robertson MJ, van Zundert GCP, Borrelli K, Skiniotis G. GemSpot: A Pipeline for Robust Modeling of Ligands into Cryo-EM Maps. Structure. 2020;28:707–16.e3.

7. Joseph AP, Lagerstedt I, Patwardhan A, Topf M, Winn M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. J Struct Biol. 2017;199:12–26.

8. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J Med Chem. 2004;47:1739–49.

9. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J Med Chem. 2006;49:6177–96.

10. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des. 1997;11:425–45.

11. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455–61.

12. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des. 2002;16:11–26.

13. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. J Mol Biol. 1997;267:727–48.

14. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins. 2003;52:609–23.

15. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model. 2019;59:895–913.

16. Joseph AP, Lagerstedt I, Jakobi A, Burnley T, Patwardhan A, Topf M, et al. Comparing Cryo-EM Reconstructions and Validating Atomic Model Fit Using Difference Maps. J Chem Inf Model. 2020;60:2552–60.

17. Locke J, Joseph AP, Peña A, Möckel MM, Mayer TU, Topf M, et al. Structural basis of human kinesin-8 function and inhibition. Proc Natl Acad Sci U S A. 2017;114:E9539–48.

18. Peña A, Sweeney A, Cook AD, Locke J, Topf M, Moores CA. Structure of Microtubule-Trapped Human Kinesin-5 and Its Mechanism of Inhibition Revealed Using Cryoelectron Microscopy. Structure. 2020;28:450–7.e5.

19. van Zundert GCP, Moriarty NW, Sobolev OV, Adams PD, Borrelli KW. Macromolecular refinement of X-ray and cryoelectron microscopy structures with Phenix/OPLS3e for improved structure and ligand quality. Structure. 2021;29:913–21.e4.

20. Cragnolini T, Sahota H, Joseph AP, Sweeney A, Malhotra S, Vasishtan D, et al. TEMPy2: a Python library with improved 3D electron microscopy density-fitting and validation workflows. Acta Crystallogr D Struct Biol. 2021;77 Pt 1:41–7.

21. Wang R, Fang X, Lu Y, Wang S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. J Med Chem. 2004;47:2977–80.

22. Wang R, Fang X, Lu Y, Yang C-Y, Wang S. The PDBbind Database: Methodologies and Updates. J Med Chem. 2005;48:4111–9.

23. de Freitas RF, Schapira M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med Chem Commun. 2017;8:1970–81.

24. Wang R, Liu L, Lai L, Tang Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. Molecular modeling annual. 1998;4:379–94.

25. Cheng T, Zhao Y, Li X, Lin F, Xu Y, Zhang X, et al. Computation of octanol-water partition coefficients by guiding an additive model with knowledge. J Chem Inf Model. 2007;47:2140–8.

26. Brylinski M. Aromatic interactions at the ligand-protein interface: Implications for the development of docking scoring functions. Chem Biol Drug Des. 2018;91:380–90.

27. Arunan E, Gutowsky HS. The rotational spectrum, structure and dynamics of a benzene dimer. J Chem Phys. 1993;98:4294–6.

28. Korb O, Stützle T, Exner TE. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In: Ant Colony Optimization and Swarm Intelligence. Springer Berlin Heidelberg; 2006. p. 247–58.

29. Radermacher M, Ruiz T. On cross-correlations, averages and noise in electron microscopy. Acta Crystallogr Sect F Struct Biol Cryst Commun. 2019;75:12–8.

30. Chen S, McMullan G, Faruqi AR, Murshudov GN, Short JM, Scheres SHW, et al. High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. Ultramicroscopy. 2013;135:24–35.

31. Baxter WT, Grassucci RA, Gao H, Frank J. Determination of signal-to-noise ratios and spectral SNRs in cryo-EM low-dose imaging of molecules. J Struct Biol. 2009;166:126–32.

32. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.

33. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17:261–72.

Chapter 5

A genetic algorithm for the flexible fitting of small molecules

Background

Over the last decade, advancements in cryogenic-electron microscopy (cryo-EM) have increased the relevance of the method in structure-based drug discovery. This is evidenced by an increase in the number of EM-maps deposited within the EMDB at 3 Å or better resolution, up from 2.34 % of all deposited maps in 2015 to 15.15 % in 2021 (*correct as of October 2021)*. Typically, at these resolutions an accurate estimation of ligand atoms, ions, protein sidechains and possibly water molecules within the binding site can be clearly identified. This advancement in the resolution limits of cryo-EM has facilitated the identification of drug target structures that were previously challenging with other techniques such as X-ray crystallography or NMR (reviewed in [1]).

However, there are still challenges associated with obtaining accurate structures with cryo-EM. For instance, structures obtained by cryo-EM typically show a variable resolution throughout, this can be due to a variety of factors including heterogeneity or flexibility within the sample, incorrect angle assignment during 3D reconstruction, preferred orientation or radiation damage of particles (reviewed in [2]). This leads to some local regions of the density map with a relatively higher resolution than the global resolution given and others with a relatively lower resolution.

Drug binding sites are often resolved at a worse resolution than the surrounding protein structure. Reasons for this include small fluctuations in the positions of small molecules within a binding site, partial occupancy of the ligand, multiple ligand binding states or a combination of these. It is important to take this into account when calculating atomic models of protein/ligand complexes to avoid overfitting.

Fitting small molecules into Cryo-EM maps

Fitting small molecule ligands into cryo-EM maps has two main steps: generating an initial fit of the ligand in the map, followed by further refinement in the context of the protein binding site.

This first step is crucial to the success of any refinement strategy, as having a starting ligand conformation closer to the correct solution increases the chance of generating a high quality fit to the map. Furthermore, the strategy taken is dependent on the resolution of the map, as lower resolution maps contain less data with respect to atomic positions within the binding site.

When a map has been derived at sufficiently high resolution (typically < 3.0Å resolution), the drug can be modelled directly into the map using software previously developed for modelling with X-ray crystallography data that has been adapted for Cryo-EM data. One of the most common methods for doing this is to first do model building in C*oot [3]* followed by real-space refinement with the software PHENIX [4, 5]. Once an initial protein model has been obtained in Coot, an initial placement of the ligand in the density can be generated by either manually modifying the ligand rotatable bonds or using the built-in automated ligand fitting protocols in Coot. One such commonly used protocol is the 'Jiggle fit' protocol [6]. Briefly, an initial conformation is subjected to small changes in rotation and translation parameters and rigidly fit into the density. Where the conformation showing the best cross-correlation with the density is then refined in real-space. The quality of this initial fit is largely dependent on the quality of the map and decreases as the resolution worsens.

Once an initial model is obtained, this model is further refined into the map using the PHENIX 'phenix.real_space_refine' protocol [5]. The aim of this protocol is to increase the fit to the map quantified with the CCC whilst ensuring the model maintains a meaningful geometry. To ensure meaningful geometry restraints are used, at resolutions of 3 Å or better, all atoms are moved simultaneously to improve the fit with the map, whilst restricting the values of covalent bond lengths and angles, dihedral angles, planarity and chirality restraints, and atom-atom repulsion. Such restraints are added to avoid overfitting the model to the map. Geometric restraints for ligands can be determined beforehand from either a PDB file or smiles string. The software eLBOW [7] is included within the PHENIX package and can determine allowed bond angles/distances, dihedrals and planarity restraints for ligands to be supplied to the refinement protocol. This protocol has successfully been used in model-fitting pipelines for numerous protein ligand complexes obtained by Cryo-EM at high resolutions including the GABAA Receptors in complex with various anaesthetics and benzodiazepines [8], the TRPC6 receptor bound to the antagonist AM-1475 [9], and the GLP-1 receptor complexed with agonist PF-06882961 [10], to name a few.

More recently, the need for pre-determined restraints in the 'phenix.real_space_refine' protocol were made redundant with the addition of the OPLS3e forcefield [11] , which combines fixed partial charges, on-the-fly semi-empirical quantum mechanical calculations with torsion restraints found by searching a built-in database.

Whilst fitting ligands at high resolution is a fairly straightforward task, the resolution range between 3.0 and 4.5 Å is more challenging. There are a few reasons for this. In terms of ligand density, there is a large difference between the extremes of this range, as when close to 3.0 Å there may be enough information contained within a density map for a modeller to

visually discriminate density corresponding to the ligand positions and perhaps the approximate positions of the atoms therein. Whilst at the lower end of this range, it may not even be possible to decipher the approximate positions of chemical moieties, and only an approximate position of the ligand can be determined visually with any accuracy. This issue is compounded further by the variable resolutions contained within maps. It is possible that even when the global resolution is very close to atomic resolution (better than 3.0 Å), the local resolution of the drug and surrounding binding site can be much lower. This effect can be even more intensified as you move towards the upper limits of this intermediate resolution range (4.5Å).

This was the case with the structure of Eg5 in complex with α- and β-tubulin and GSK-1 solved in chapter 2. The global resolution of the map was seen to be 3.8 Å; however, local resolution estimates showed the resolution of the GSK-1 binding site to be between 5.5 and 6.5 Å. Furthermore, at these resolutions, cryo-EM is unable to resolve the positions of water molecules within the binding site and the exact positioning of sidechains that may interact with the ligands, further hindering the ability to identify chemically reasonable ligand conformations with a good fit to the density. However, assessing how accurately ligands can be fit at this range is an important challenge. Even though it is becoming more common to solve structures with cryo-EM at resolutions that rival X-ray crystallography, these structures only represent a small proportion of maps currently deposited in the EMDB. Density maps within the 3-4 Å range still make up the bulk of maps deposited in the EMDB (44.01 %, correct as of October 2021) and therefore, being able to derive accurate atomic models from these maps is still an important endeavour.

One such study that assessed the accuracy of fitting ligands within this range was the updated ‘phenix.real_space_refine’ with the OPLS3e force field [12]. A real space re-refinement of 15 Cryo-EM models deposited in the PDB into their deposited maps, with resolutions ranging from 1.9 to 4.3 Å, using a combination of the OPLS3e force field [11] with an implicit solvation term showed a comparable model quality to that obtained with the standard ‘phenix.real_space_refine’. Additionally, no noticeable improvement was seen in model quality when refining models with the OPLS3e protocol compared to the standard PHENIX protocol, when model quality was assessed by MolProbity scores [13, 14] and PHENIX-Ramachandran Z-scores [15]. However, a majority of cases showed significantly lower ligand strain energies [12], a measure of the energetic cost of adopting a bound conformation.

Furthermore, the study highlighted the interplay between weighting the chemical/physical scoring function and the goodness-of-fit metric to the experimental data (CCC in this case). It was seen that there was a wide range of weights that could generate atomic models within approximately the same model quality; however, a boundary was reached where further increases in weighting for the goodness-of-fit to the experimental data resulted in overfit models when model quality was assessed by MolProbity [13, 14] and PHENIX-ramachandran Z-scores [15].

Since generating a good initial conformation of the ligand increases the success of a ligand fitting strategy, pipelines have been developed that focus on generating a better initial fit of the ligand by leveraging molecular docking software.

The program GemSpot [16] combines the search algorithm and scoring function of the molecular docking software GLIDE [17, 18] with a real space CCC of the cryo-EM map, to guide approximations of ligand fitting. In contrast to the aforementioned 'jiggle fit' protocol, by utilising the scoring function of the molecular docking software rather than solely relying on the CCC of atoms with the map, GemSpot generates chemically meaningful approximations of the ligand position prior to refinement. This significantly reduces the search space needed to refine candidate solutions, as solutions with a good docking score but low CCC and conversely solutions with good CCC and low docking scores can be immediately discounted.

Following molecular docking in this manner, candidate ligand conformations were minimised in the map using the 'phenix.real_space_refine' with ligand restraints generated by the OPLS3e force field [11] implemented in PHENIX [12]. The final stage of ligand fitting involves simulating the placement of water molecules within the binding site using a Monte-Carlo simulation with the software JAWS [19].

The GemSpot pipeline utilised the same benchmark of 15 cryo-EM maps with associated models at resolutions between 1.9 Å to 4.3 Å, as was used in the PHENIX/OPLS3e report [12]. When six map model protein/ligand complexes were tested with a resolution of 3.0 Å or better, the ligand CCC's were seen to be similar to the deposited CCC in nearly all cases. Additionally, the pipeline achieved accurate replication of the placement of water molecules within the maps. However, the only assessment of the quality of the fit ligands was the CCC and no assessment of ligand overfitting was given.

In the report, the aforementioned GemSpot methodology [16] was applied to a further 10 maps at resolutions of 3.0 to 4.5 Å. At this resolution, it was not possible to place water molecules within the maps. However, the poses generated generally agreed with the CCC of the deposited conformations. Once again at this resolution, the CCC correlated well with the CCC seen in the deposited models, and in six cases increased. However, once again no assessment of overfitting or model quality was made and it cannot be ruled out that ligands were fit in energetically strained conformations. Although, when taken together with the results of the PHENIX/OPLS3e report [12] on the same benchmark, it can be assumed the conformations are relatively accurate, as the same refinement protocol is applied post docking with GLIDE. Unfortunately, the authors make no comment on the effect of utilising the docking software to generate initial conformations.

This protocol has since been further developed for the specific task of fitting ligands into density maps at lower resolutions (3.0 - 5.0 Å) using molecular dynamics (MD) and neural network potentials [20]. The methodology uses a protein model initially fit into a density map in real space, along with a low energy conformation of the ligand docked into the initial

model using GLIDE. This model is refined further with flexible fitting molecular dynamics, in NAMD. Additional refinement was conducted using flexible fitting MD using CHARMM force field parameters for protein atoms and neural network trained potentials for the ligand. The final stage of refinement involved refinement with the neural network derived potentials with quantum mechanical/molecular mechanical-molecular dynamics (QM/MM-MD).

The study utilised three map model systems and initial experiments were conducted with the deposited structures and cryo-EM maps of nicotinamide adenine dinucleotide-bound horse liver alcohol dehydrogenase (6NBB), at 2.9 Å resolution. Refinement was conducted on the best scoring GLIDE docking pose for either the *apo* protein model deposited with the PDB or on where the sidechain atoms of the binding site had been displaced using a short MD run. Two density maps were used for analysis, the 2.9 Å deposited map and a map blurred to 5.0 Å using a gaussian function. Using only QM/MM - MDFF and the *apo* protein model for refinement the deposited conformation was recreated, with a ligand CCC to the map of 0.86 with the 2.9 Å map and 0.9 with the 5 Å map, compared to 0.94 in the deposited model. When the same analysis was run using the MD modified atomic model and the 2.9 Å map, the ligand CCC converged to only 0.32 with an RMSD relative to the deposited model of 3.2 Å. The results of the analysis were not reported for the 5 Å map. This result highlights one of the key issues when fitting ligands at lower resolutions, in that the accuracy of the refinement is highly dependent on both the initial placement of the ligand and the quality of the refined protein atomic model.

Following this, the authors attempt to improve the quality of fit using NNP-MDFF. However, since the NNP for the phosphate groups of the nicotinamide adenine dinucleotide ligand had not been calculated, a second system was introduced. Here they used the X-ray derived structure of the EGFR tyrosine kinase domain bound to erlotinib solved at 2.75 Å resolution. For these experiments, the X-ray derived map was recalculated at 4 and 5 Å resolution. The analysis was repeated as before, starting with the best docked conformation from GLIDE refined into either the deposited apo protein or an MD modified atomic model. At 4 Å, the NNP-MDFF was able to reproduce the binding site, with both models giving a CCC of 0.9 compared to 0.94 in the deposited structure. Whilst the CCC reported are lower than those reported in the deposited structure, this may be due to overfitting of the ligand in the deposited model, and indeed it was shown that the ligand strain energy was more stable by 3.6 kcal/mol compared to the deposited pose. This highlights the importance of taking the chemical information of the ligand/protein complex, as the best scoring pose may not be the one with the best CCC. At 5 Å resolution simulations for both, the apo- and MD-modified proteins failed to replicate the deposited structure. This was reported to be due to the fact that at 5 Å resolution, density from protein atoms begins to blend with the ligand density meaning the ligand begins to explore regions of the binding site containing non-ligand density and adopting unreasonable conformations. This report was a nice example of the challenges faced when modelling ligand binding sites at lower resolutions. However, the number of structures used were low and the relevance of utilising X-ray derived maps is debatable. Additionally, the authors state that NPP-MDFF increased the radius of convergence compared to QM/MM-MDFF (i.e. during QM/MM-MDFF, the starting RMSD of the binding site had to

be closer to the deposited model to accurately recreate the binding site when compared to NPP-MDFF); however, the QM/MM-MDFF and NPP-MDFF refinements were conducted on a single system and different systems at that, with maps derived from different experimental procedures. Therefore, whilst NPP-MDFF was certainly shown to be an interesting technique for fitting ligands at lower resolution, the direct comparison to QM/MM-MDFF with the available data seems an unfair one.

Another commonly used refinement software is ROSETTA [21] , which contains a collection of computational tools for the analysis of protein structures. An automated structural refinement protocol using ROSETTA was shown to have success at refining multiple ligands into a map at 3.4 Å [22]. The general refinement protocol reported occurred in three stages. First, potential errors in the initial model, such as local strain and poor fit to the map, were identified. These regions of poor model quality were then rebuilt in a fragment-based approach. Secondly, models that showed better fit to the map and model quality, assessed by MolProbity, were taken forward for a final all-atom refinement at the local level. To include ligands in this protocol, the proteins were first refined in the absence of ligand, whilst restraining ligand binding site geometry. The ligands were then added back in and the model re-refined. This protocol significantly improved the model quality when compared to the deposited model, assessed by MolProbity [13, 14], and EMRinger scores [23]. However, the FSC of the map to the model decreased when compared to the deposited structure from, 0.743 to 0.708. The authors hypothesise that this may be due to overfitting in the deposited model, which was supported by the deposited model having poor MolProbity clash scores and a large number of rotamer outliers.

One further study reported the use of correlation-derived MD to automate the fitting of atomic models at lower resolution [24]. The principle of this methodology is based on introducing a potential for the CCC to an MD force field. Briefly, the CCC between the experimental map and a map simulated from the model structure is calculated, which is then converted to a potential and added to a MD force field to derive atom potentials. Next, atoms are moved along their potentials and the structure updates. This process continues iteratively until convergence. The refinement is controlled by modifying various parameters during fitting. The first is the resolution of the simulated map, which begins at an artificially low resolution and drops to the given resolution of the map as the refinement continues. This is done to avoid getting stuck at a local minimum during refinement. Secondly, the weights of the CCC potential with respect to the force field potential are modified during fitting, where the strength of the CCC potential increases as the run continues. Furthermore, a high temperature for simulated annealing is used to enforce a better local fit of sidechains with the map, whilst leaving better fitting regions unmodified.

The study utilised this protocol to fit the atomic model of an N-ethylmaleimide sensitive factor complexed with ATP into a 3.9 Å cryo-EM density map. The result showed the fit model was of a better quality than the deposited model, evidenced by a significantly reduced clash score. Two high-resolution X-ray structures were used to validate the output model. The better model quality was seen to be down to the fit of ATP molecules. In the deposited model,

half of the six ATP molecules showed an inaccurate geometry, with respect to the dihedral rotation angles of the ATP adenosine and phosphate groups, relative to the high-resolution X-ray structures. Whilst the geometry of the ATP in the correlation-derived MD model appeared much closer to the high-resolution X-ray structures, with an average RMSD over the entire models of 0.2 Å.

Few studies have focused on the fitting of ligands into cryo-EM density maps at very low resolutions, greater than 4.5 Å, as it is assumed that there is nowhere near enough data to accurately determine ligand positions. Another problem is that at this resolution, density from protein atoms bleeds into the density of the ligand, compounding the issue of defining ligand density to refine to. However, some methods have attempted to circumvent this issue.

One strategy for fitting ligands at low resolutions involves the use of difference density maps. The general workflow of the protocol involves first calculating a difference map between the experimental protein/ligand complex density map and a simulated or experimental map of the apo protein. Candidate ligand conformations are generated using docking software and the agreement of the conformations with the difference map density. Finally, the best conformations are further refined by minimisation of the binding site. This protocol attempts to deal with the issues surrounding the merging of protein and ligand densities that occur at very low resolution (> 4 Å). By calculating the density that corresponds to the ligand with the use of difference mapping, the chances of the molecule moving into density corresponding to protein atoms during refinement is minimised. This protocol was first applied to the structure of Kinesin-8 in complex with the small molecule BTB-1 [25]. The Global resolution of the map was seen to be 4.8 Å; however, local resolution estimates indicated that the resolution of the binding site was approximately 6.5 Å. At this resolution, accurate placement of ligands directly into density is very difficult. After calculation of a difference map, the BTB-1 model was docked into the binding site using a three-step protocol. Initially, HADDOCK [26] was used for a global docking to identify the BTB-1 binding site. Once the binding site was identified the ligand was docked using AutoDock Vina [27] with the sidechains of the binding site allowed to be flexible. A final phase of docking was conducted with HADDOCK and ligands as rigid bodies. The top solutions from this final phase were then flexibly fit into the binding site using Flex-EM [28, 29]. Whilst this protocol attempted to overcome the limitations of fitting ligands at low resolutions, the final deposited structures showed that the protocol was unable to discriminate between two distinct conformations of the BTB-1 ligand within the binding site. Although the positions of these two ligands correlated well with site-directed mutagenesis data, the accuracy of protein ligand interactions seen in the model was questionable.

This protocol was further built on in a kinesin-Eg5 complex with GSK-1 [30]. The overall resolution of the map was seen to be 3.8 Å and the local resolution estimates of the binding site were seen to be approximately 5.5 Å resolution. This protocol utilised a consensus docking approach with GOLD [31] and Autodock Vina [27], as it has been shown that a consensus docking approach can increase the likelihood of identifying a correct conformation, followed by a binding site minimisation with the AMBER ff14SB force field

[32] parameters used from protein residues and the generalised amber force field used for ligand atoms [33]. However, although once again the approximate position of the ligand binding site could be identified and correlated well with site directed mutagenesis data, we were unable to distinguish between two equally likely conformations. As a result, the confidence in the accuracy of the protein ligand interactions within the atomic models was low.

Many protocols have been developed for assessing protein model quality such as MolProbity [13, 14]. However there are less methods designed for assessing the quality of ligands in such structures. One of the most common ways to do this is with the ligand strain energy. Ligand strain energy calculation can be achieved by either using quantum mechanical, molecular mechanical or statistical methods. One recent methodology for the latter method developed histograms of acceptable angles around torsions for a large number of small molecules in the Cambridge structural database [34] and the PDB [35] as a way to calculate ligand strain energy [36]. This methodology was built upon by one team that converted these databases of torsion angles into theoretical values for torsional strain energies namely, the torsion energy units (TEUs) [37]. Using a benchmark of 40 protein ligand complexes from the DUD-E database [38] and ligand benchmarks containing true binders and decoys, the team aimed to identify a cutoff for TEU for which true binders could be separated from decoys. The team identified a cutoff of 1.5 TEUs as a good cutoff for this task.

Presented in this investigation is a methodology for fitting small molecules into cryo-EM maps. The method builds on the difference mapping methodology previously used to fit ligands at lower resolutions [25, 30]. A genetic algorithm was developed to generate candidate ligand conformations within protein binding sites. The fits of generated ligand conformations with the difference maps were scored with the integrated scoring function (Chapter 4), that was shown to be able to accurately identify a correct conformation from decoy conformations when a simulated difference map was used for fitting.

The methodology was benchmarked using a set of protein ligand models derived from experimental data with resolutions ranging from 2.2 Å to 4.5 Å. To address the quality of small molecule fits we used the CCC metric to assess the fit to experimental cryo-EM maps. Additionally, the statistical method for calculator ligand strain energy [37] was employed with a cutoff of 1.5 TEUs to assess the quality of ligand conformations independently of the maps.

Results

The Genetic search algorithm

For fitting ligands into cryo-EM difference density maps a Genetic Algorithm (GA) was selected (Algorithm 1). The GA was developed to work in three stages, a global search, a

local search and a fine tuning stage (Figure 1). Each stage was a GA in its own right with parameters modified to progressively reduce the search space.

The global stage was a coarse search throughout the entire binding site to identify possible candidate conformations for further refinement. The binding site was defined from a given radius and centre point. For our investigations the binding site radii were set to 12.5 Å and the binding sites were centred around the centroid of the deposited ligands. Any residue that contained atoms within the binding site was considered a binding site residue.

Algorithm 1. Genetic Algorithm pseudo-code

```
Initialise_parameters()
S_0 ← generate_initial_ population()
for i = 1 to n_generations do
   for j =1 to n_population * crossover_rate do
      s_j ← generate_new_individual_by_crossover()
      S_i ← S_i ∪ s_j
   end for

   for j = 1 to n_population * mutation_rate do
      S_ij ← mutate_individual()
   end for

   S_i ← score_population()
   S_i ← sort_population_by_score()

   for j = 1 to n_population do
      S_i+1 = S_i+1  ∪ S_ij
   end for
return S_n_generations
```

To begin the search, a specific number of initial conformations were generated. By default 100 initial conformations were generated. Each conformation was encoded in a "chromosome" with at least 6 "genes", three values representing the x, y, z coordinates of the centroid of the molecule, three values representing a rigid rotation in the plane of x, y, z about the centre of the molecule, and *n* values representing the dihedral angles around all rotatable bonds within the molecule (where *n* was the number of rotatable bonds within a molecule) (Figure 1).

To generate a pool of acceptable values for the x, y, z points of the translational genes, the binding site was split into 1D sets where the initial centroid x, y or z value was increased and

decreased in 0.1 Å steps to the half maximal of the binding site radius. When generating initial conformations the x, y, and z values were randomly selected from their respective lists of acceptable values.

The set of acceptable values for both the x, y, and z rotation genes and dihedral angles were taken from a list of all the values between 0 and 360 at 5 ° intervals. For each gene a random value between 0 and 360 at 5 ° intervals was taken to generate initial conformations.

The first global stage of the algorithm was run for 200 generations with a mutation rate of 0.5 and a crossover rate of 0.5. The top 10 scored conformations after 200 generations were taken forward to the local stage of refinement.

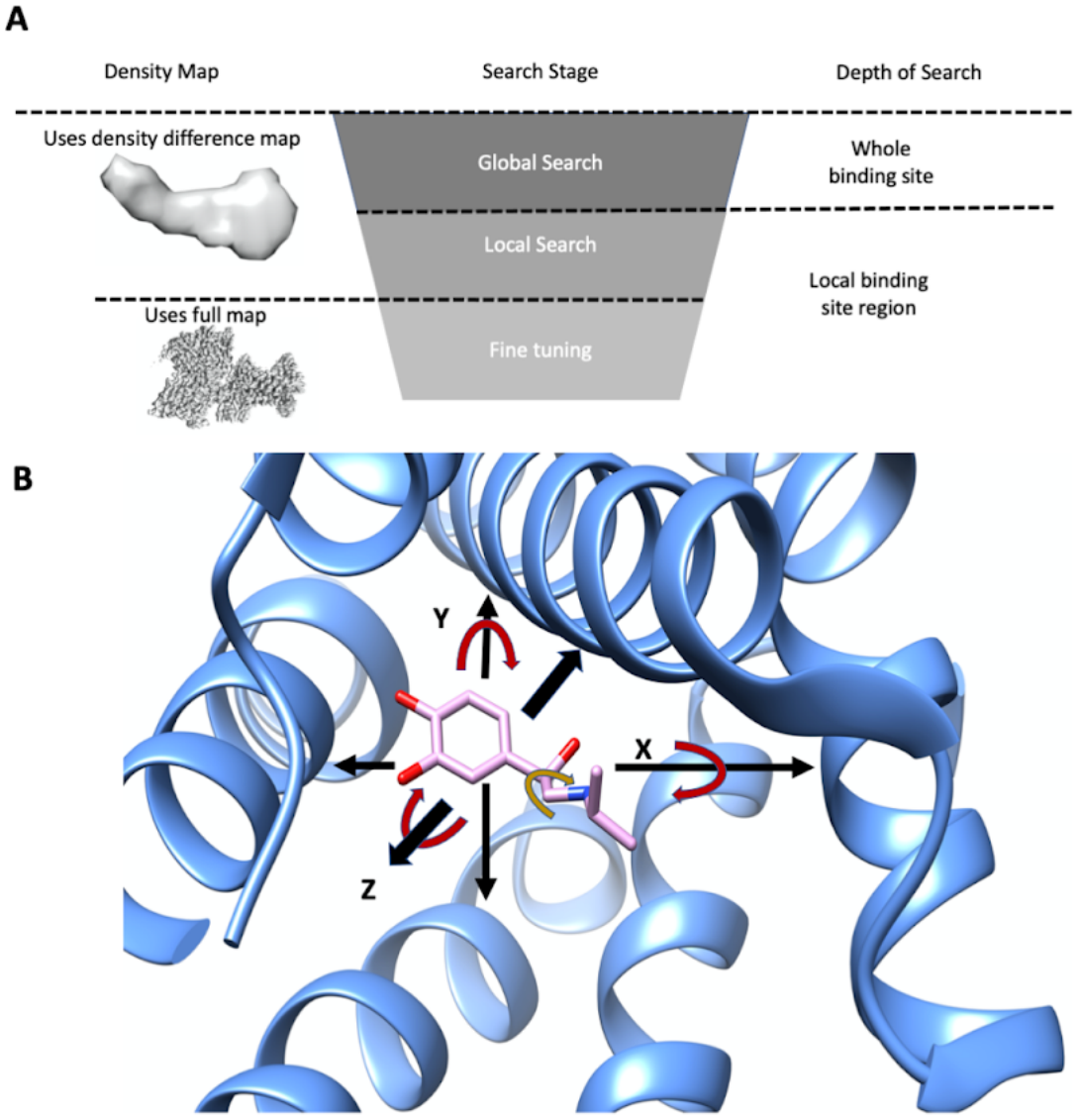


Figure 1. *(**A**) A schematic representation of the three stages of the search algorithm. The maps used and the search space for each stage are indicated. (**B**) The degrees of freedom used in the search algorithm. The translational (black arrows), rotational (red arrow) and*

dihedral (yellow arrow) are indicated in the X, Y and X dimensions. For clarity the dihedral degree of freedom is indicated only about a single dihedral bond.

The second stage of refinement attempts to locally refine the best conformations from the global stage. For each ligand taken forward, a new genetic run was set up, with 100 initial conformations and 100 generations. However, the possible values for the centroid x, y, z genes were constrained to be ∓ 1 Å from the centroid of the global conformation to be optimised, with a step of 0.05 Å. The rigid x, y, and z rotation along with the dihedral angle gene values were constrained to be within ∓ 30 ° of the values found in the global search stage, with a step size of 0.25 °. The best conformation for each individual run was taken forward to the final stage of refinement.

Both the Global and local stages of refinement were designed to fit a ligand to a difference density map (Figure 1A) and whilst the map provides good approximation of the ligand density it may be lacking in some of the finer details. To this end a third fine-tuning stage was implemented that fits the locally refined ligand to the full map. Again a new genetic run was set up for each ligand refined from the local stage, with 100 initial conformations and 100 generations. This time however, centroid x, y and z values were constrained at ∓ 0.5 Å in steps of 0.01 Å and rotations and dihedral angles restrained at ∓ 15 ° in steps of 0.1 °.

Since the GA involves an element of randomness, it is not guaranteed that it will converge at the same results each time. Due to this, each condition was run three times and the results from each run were combined, yielding 30 final conformations per fit.

A high resolution experimental benchmark

A high-resolution benchmark was constructed for docking from 26 protein-ligand complexes from cryo-EM maps with resolutions between 2.2 Å and 3.0 Å (Table 1). This set represented a diverse range of ligands with rotatable bond numbers ranging from 1 to 15 (Figure A7).

Difference maps for each protein ligand complex were calculated using a local scaling method [39] implemented in TEMPy [40]. All density difference maps were inspected visually for quality. Most difference maps were of a high quality (Figure A8), assessed with the CCC overlap with deposited ligands (Table 1). A minority of maps were assessed as being of low quality, an example being the 6TTI difference map for the ligand NXE, where density for the trifluoromethyl group was missing when compared to the full map (Figure 2). These cases were still included in the benchmark as they represented what was achievable with real world data.

Table 1. The Benchmark PDB ID, ligand ID, resolutions and difference map CCC.

PDB	Ligand	resolution	CCC		PDB	ligand	resolution	CCC
6X40	RI5	2.86	0.591		6KPF	8D0	2.9	0.666

6TTE	PTQ	2.2	0.522		6QM7	J6E	2.8	0.590
6UDP	IMP	2.95	0.760		6TW1	M4H	2.7	0.541
6TTQ	FBP	2.7	0.682		6TTI	NXE	2.5	0.343
6A95	9SR	2.6	0.694		6X1A	UK4	2.5	0.729
6OAX	AGS	2.9	0.738		7JJO	5FW	2.6	0.754
6PEQ	ZK1	2.97	0.786		6VFX	ATP	2.9	0.770
6UZ8	R0D	2.84	0.756		6REY	IHP	3.0	0.407
6WVR	TA1	2.9	0.674		6NYY	ANP	3.0	0.751
6UQE	ADP	3.0	0.580		7C7Q	2C0	3.0	0.470
6OO3	6EU	2.9	0.528		6PUW	KLQ	2.9	0.594
6PUZ	XXJ	2.8	0.512		7CFM	FWX	3.0	0.732
6X3T	PFL	2.55	0.479		6X3X	DZP	2.92	0.650

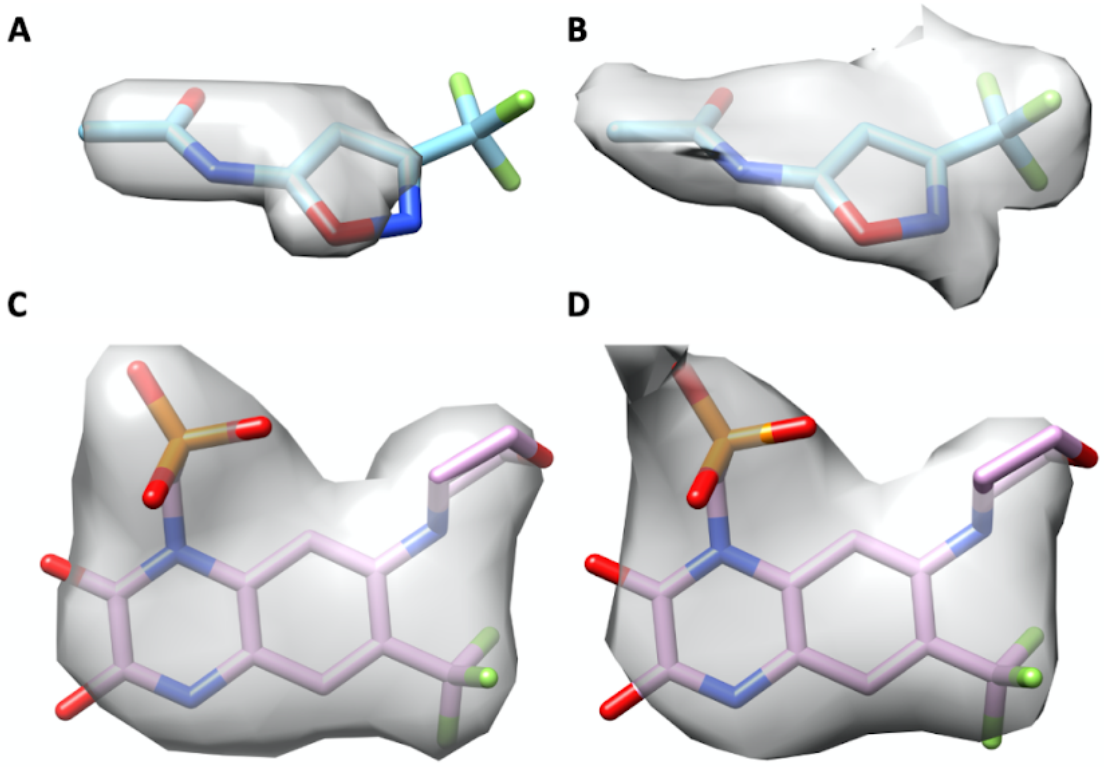


Figure 2. *An Example of a low-quality density difference map (**A**) and a comparison to ligand density in the full map (**B**). Also shown is an example of a high-quality difference map (**C**) and a comparison to the deposited full map (**D**). The deposited ligand models are also shown for the case of NXE (PDB ID: 6TTI A,B) and ZK1 (PDB ID: 6PEQ, C, D).*

Integrating the MI score into the flexible fitting algorithm.

Previous experiments identified an empirical scoring function that was shown to have a better docking power than the AutoDock Vina scoring function (Chapter 4). Furthermore, using simulated maps, it was shown that integrating this empirical scoring function with the MI score resulted in an increase in the docking power of the scoring function at resolutions from 2.5 Å to 8.5 Å. Here, the integrated scoring function was examined in conjunction with the GA for fitting small molecules to experimental data.

To this end the GA and the integrated score were used to fit small molecules to difference maps generated using the high-resolution experimental benchmark (Table 1). The MI score was weighted by approximate factors of 0.1x, 0.5x, 1.0x, 5.0x and 10.0x the magnitude of the empirical score (as in chapter 4). The success rate was defined as whether the software was able to generate a ligand conformation ≤ 2.0 Å from the deposited reference conformation. A cutoff of 2.0 Å is a common cutoff and has previously been used to describe correct solutions in the evaluation of molecular docking results [41].

Additionally, the quality of solutions was investigated by calculating the number of cases that generated a solution ≤ 1.5 Å or ≤ 1.0 Å from the reference ligand. Furthermore, a second metric, the mean minimum RMSD, gave further indication regarding the quality of solutions. The mean minimum RMSD was calculated as the RMSD averaged over the benchmark from the ligand conformations that had the lowest RMSD to the reference ligand in each case.

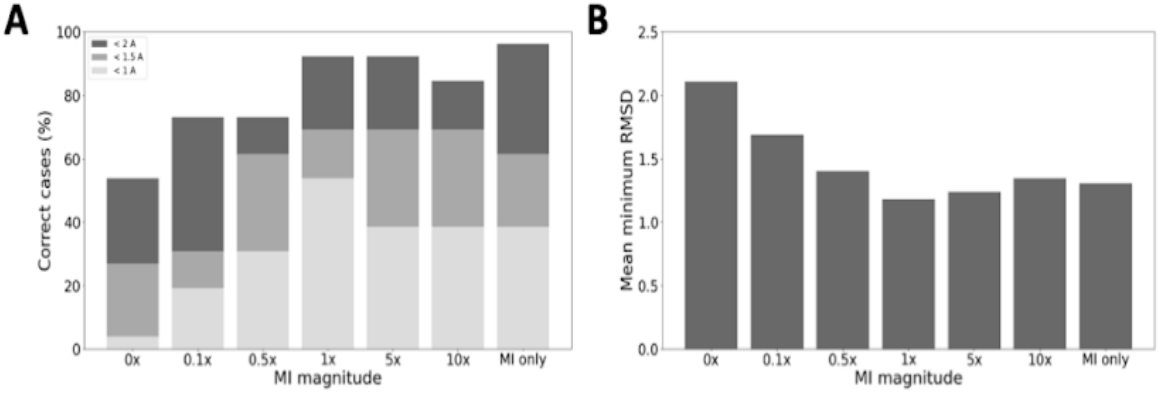


Figure 3. *The results of flexible fitting of ligands for the 26 cases of the high resolution experimental benchmark, using an integrated scoring function composed of the empirical scoring function and MI score weighted at 0x, 0.1x, 0.5x, 1.0x, 5.0x, 10.0x the magnitude of the empirical scoring function, or MI alone. (**A**) Percent of cases where the fitting algorithm generated a solution ≤ 2.0 Å (dark grey), ≤ 1.5 Å (grey), or ≤ 1.0 Å (light grey). (**B**) The mean minimum RMSD values of all cases for the various scoring functions.*

The results indicated that when the MI score was integrated with the empirical scoring function at a magnitude of 1x or 5x, the amount of correct cases found was one less than when using the MI alone at 92.30 %, 92.30 % and 96.15 %, respectively (24/26, 24/26 and 25/26 cases) (Figure 3). At all other weight combinations a lower amount of correct cases were seen.

This initially indicated that the addition of the empirical score had no advantageous effect on fitting small molecules, with respect to identifying a correct conformation. However, the quality of solutions identified with the integrated score was seen to be better than with the MI score alone: the percent of correct solutions identified at ≤ 1.5 Å and ≤ 1.0 Å was 69.23 % (18 of 26 cases) and 53.84 % (14 of 26 cases) for the 1x integrated score, compared with 61.53 % (16 of 26 cases) and 38.46 % (10 of 27 cases) for the MI score alone (Figure 3).

Additionally, the mean minimum RMSD of the integrated score was 1.18 Å (∓ 0.53 STD) compared to 1.30 Å (∓ 0.68 STD) for the MI score alone. However, this decrease was not statistically significant ($p = 0.47$). Taken together the results indicated that most of the information regarding the conformation of the ligand is contained within the MI score. However, the addition of the empirical scoring function can improve the quality of the fit compared to the MI score alone.

The previous experiments were carried out using the deposited protein models; these models had been optimised with the deposited ligand conformations. This may introduce a level of model bias, to account for this residues in the binding sites of protein models were subjected to a single round of flexible fitting in the absence of the ligand with Flex-EM [28, 29]. The analysis was run again in the same way using the re-refined protein models and scoring with the 1x integrated scoring function. It was seen that the software was able to produce a correct fit (≤ 2.0 Å) 88.46 % of the time (23 of 26 cases). In 73.07 % of cases (19 of 26 cases) a ligand conformation ≤ 1.5 Å RMSD to the reference ligand was produced, and in 46.15 % of cases a solution with an RMSD of ≤ 1.0 Å from the reference ligands (12 of 26 cases) was given. The mean minimum RMSD was 1.22 Å (∓ 0.62 STD), a result that showed no statistically significant difference when compared to using the deposited models (p=0.81). Using re-refined protein models showed a slight decline in the number of correct cases and the quality of solutions, however, it indicated that the effect of model bias in our results was small and confirmed that the integrated scoring method for fitting small molecules produced meaningful results. All further experiments and analysis were conducted on the solutions and models in the re-refined set.

To assess the effect of fitting with difference maps versus full maps ligands were fitted into the re-refined protein set, using the deposited full maps and the integrated score with MI weighted at 1x. Fitting with the full maps resulted in correct fits only 69.23 % of the time (18/26 cases). The quality of fits was seen to be lower than when using the difference maps with solutions having an RMSD ≤ 1.5 Å to reference ligands 53.86 % of the time (14/26 cases) and ≤ 1.0 Å 26.92 % of the time (7/26 cases). Additionally, the mean minimum RMSD was 1.92 Å (∓ 1.54 STD), a statistically significant decrease compared to fitting with

the difference maps (p=0.04). This indicated that fitting with difference maps produced more correct fits of higher quality than fitting with the full map directly.

Quality of generated solutions

The quality of the best ligand conformations (assessed by RMSD to the reference ligand) was further probed by calculating the ligand strain energy. The ligand strain energy was assessed using the statistical approach presented by Gu *et al [37]*. In 8 cases the best ligand conformation had a better strain energy to the deposited reference confirmation. In 7 cases the strain energy was worse but still within an acceptable range ≤ 1.5 torsion energy units (TEU) to the reference ligand, and in 11 cases the best solution had a much worse ligand strain energy than the deposited solution (Table 2).

Table 2. The CCC and strain energies of the deposited reference ligand conformations and solutions that gave the lowest RMSD to the reference ligands.

PDB	Ligand ID	RMSD	CCC	CCC deposited	Strain energy	Strain energy reference
6x40	RI5	0.925	0.156	0.130	1.23	0.65
6tte	PTQ	1.321	0.134	0.149	6.186	5.712
6udp	IMP	0.704	0.210	0.207	3.73	4.867
6ttq	FBP	0.819	0.248	0.237	12.64	10.36
6a95	9SR	0.470	0.122	0.125	3.287	1.955
6oax	AGS	1.589	0.162	0.182	3.469	7.434
6peq	ZK1	0.796	0.183	0.185	5.853	3.79
6uz8	R0D	0.662	0.108	0.112	2.22	1.809
6wvr	TA1	1.942	0.134	0.162	41.12	12.053
6uqe	ADP	0.879	0.398	0.403	8.27	5.35
6oo3	6EU	1.429	0.082	0.112	16.911	18.464
6puz	XXJ	2.29	0.145	0.178	9.767	17.549
6x3t	PFL	0.724	0.082	0.091	1.44	0.618
6kpf	8D0	1.037	0.124	0.117	23.62	13.24
6qm7	J6E	0.641	0.083	0.084	7.28	6.25

6tw1	M4H	1.156	0.212	0.179	3.14	3.916
6tti	NXE	1.46	0.243	0.281	6.29	1.12
6x1a	UK4	3.141	0.089	0.168	18.17	12.77
7jjo	5FW	0.787	0.308	0.290	8.821	5.10
6vfx	ATP	1.042	0.162	0.182	5.725	7.725
6rey	IHP	1.797	0.520	0.569	2.28	9.629
6nyy	ANP	1.703	0.114	0.131	10.569	7.34
7c7q	2C0	0.653	0.057	0.056	3.27	5.20
6puw	KLQ	2.018	0.218	0.241	8.93	2.16
7cfm	FWX	1.167	0.120	0.135	14.740	10.739
6x3x	DZP	0.574	0.105	0.095	1.088	0.966

The software was able to identify a solution within 2.0 Å of the reference conformation 88.88 % of the time (Figure A9, Table A1). Failing on three occasions with the ligands UK4 (PDB ID: 6X1A), KLQ (PDB ID: 6PUW), and XXJ (PDB ID: 6PUZ), the best solutions had RMSD values to the reference ligands of 3.141, 2.018, and 2.29 Å, respectively. To assess why this was the case the re-refined protein models were inspected. In the case of the ligand UK4 it was found that during re-refinement the sidechain of TRP203 had moved into the density corresponding to the ligand (Figure 4). It was assumed that this would prevent the ligand adopting a correct conformation due to steric hindrance. To ascertain if this was the case the side chain was manually moved out of the density using a rotamer library (Figure 4) and the analysis ran again. The results showed that moving the side chain of tryptophan 203 out of the ligand density reduced the RMSD of the best solution to the reference ligand from 3.141 Å to 2.46 Å. This was an improvement however a correct conformation for the ligand UK4 still could not be found (Figure 4).

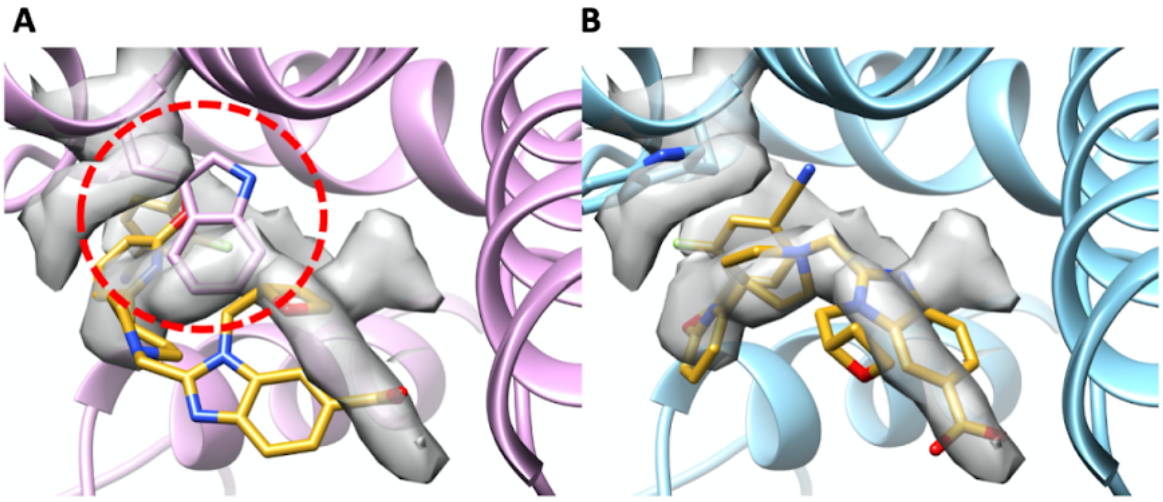


Figure 4. *(**A**) A re-refined protein model (PDB ID: 6X1A) where the side chain atoms of a binding site tryptophan (TRP203) had moved into the density corresponding to the ligand (red circle). (**B**) The manually modified protein model where the side chain atoms were moved out of ligand density (grey). For both models the best fit ligand conformations (assessed by RMSD to the reference ligand) are shown (yellow). For clarity only density around the ligand and the tryptophan residue is shown.*

Upon examination of the ligands KLQ and XXJ, the binding sites were seen to contain nucleotides and the ligand solutions had moved into the density corresponding to nucleotides. The scoring function did not contain any scoring terms for nucleotide atoms and thus ligands were not penalised for exploring this space. It cannot be assumed that ligands will make specific interaction with nucleotides in the same way as proteins, and it would require further investigation to include them in the scoring function, it is reasonable to assume that ligand atoms cannot occupy the same space as atoms from nucleotides. To make the algorithm and scoring function aware of these atoms being there, the vdW overlaps of ligand and nucleotide atoms were added into the scoring function using the Autodock Vina steric term of the empirical scoring function (Chapter 4).

Following this change the analysis was run again and the RMSD decreased from 2.29 Å to 1.86 Å and for KLQ and from 2.01 Å to 0.956 Å. This indicated that making the algorithm aware of the position of nucleotide atoms resulted in more correct solutions, however, the quality of the solution was only high quality for the KLQ ligand (Figure 5).

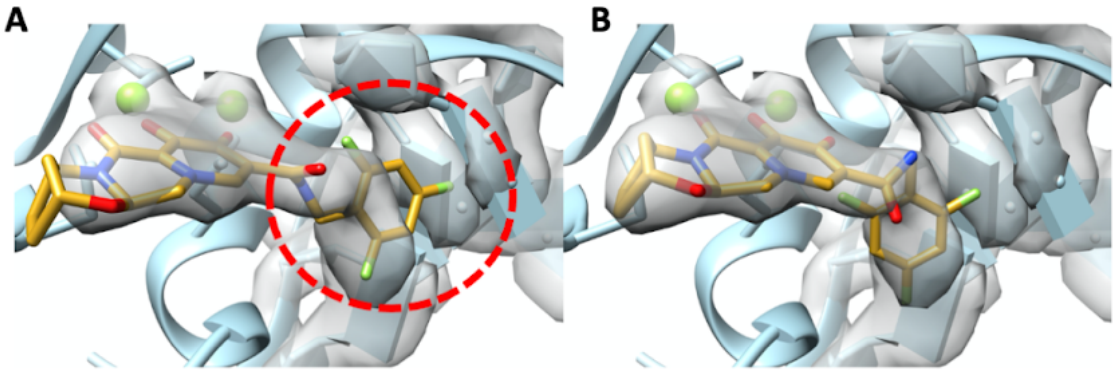


Figure 5. *(**A**) A case where the best ligand solution (yellow model) (assessed by RMSD to the reference solution) explored regions of the density map (grey) occupied by nucleotide atoms (red circle). (**B**) The best solution (yellow model) once the algorithm was made aware of the nucleotide atoms. For clarity only density near the ligand and nucleotide atoms are shown.*

Binding site minimisation

The GA did not handle side chain flexibility. As the proteins were re- refined in the absence of ligands it was necessary to minimise the atoms that made up the binding sites. Once solutions were generated the surrounding binding sites were minimised using the open source MD program OpenMM, implemented in python. The importance of this step is highlighted in the case of the ligand FBP (PDB ID: 6TTQ) binding site (Figure 6). During re-refinement with Flex-EM a loop within the binding site was moved towards the ligand density, and began to clash with the ligand. Following binding site minimisation this loop moved back into the density that it corresponded to (Figure 6). The effect of this movement could be seen in the MolProbity scores that decreased from 2.30 to 2.23. In most cases the binding site minimisation improved the quality of the models, evidenced by a decrease in MolProbity score (Table 3).

Table 3. Table of MolProbity scores before and after binding site minimisation.

PDB ID	MolProbity initial	MolProbity Min
6x3x	1.76	1.68
7c7q	1.91	1.88
7jjo	1.60	1.35
6tti	1.75	1.64
6qm7	2.39	2.37
6x3t	1.90	1.86

6uz8	1.93	1.90
6peq	2.18	2.18
6ttq	2.30	2.23
6udp	1.59	1.59
6x40	1.88	1.80
7cfm	1.85	1.74
6x1a	2.61	2.58
6oo3	2.01	2.02
6tte	1.57	1.51
6a95	2.46	2.46
6tw1	2.35	2.32
6vfx	1.51	1.45
6nyy	2.05	2.05
6puw	1.71	1.70
6oax	2.31	2.31
6wvr	0.95	0.95
6uqe	2.18	2.18
6puz	1.70	1.63
6kpf	1.85	1.85
6rey	1.67	1.67

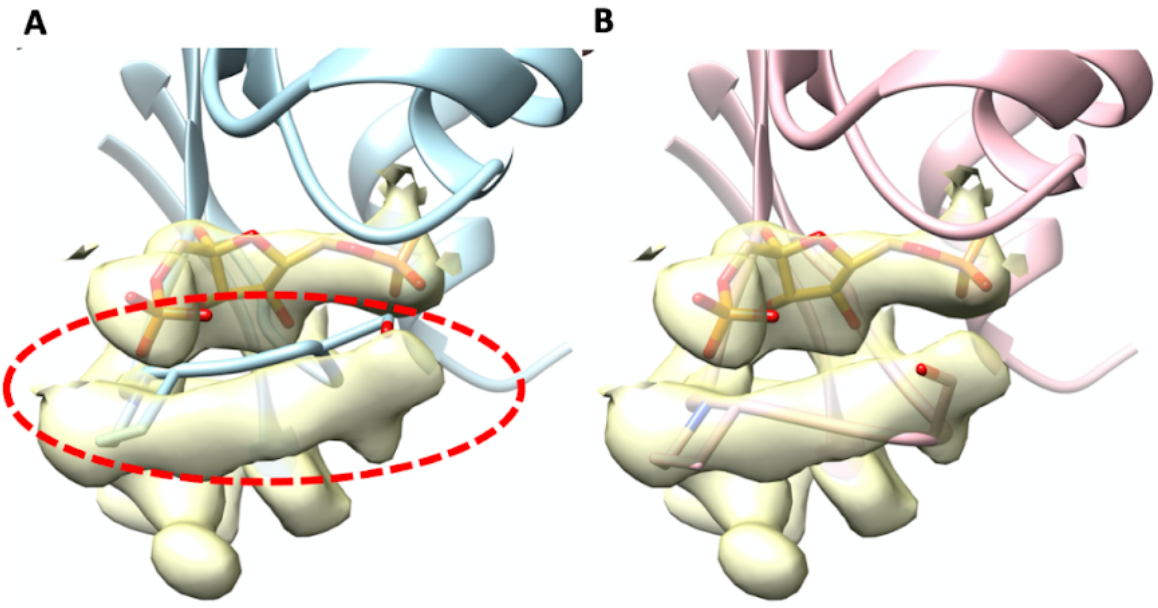


Figure 6. *(**A**) A re-refined protein model (PDB ID: 6TTQ) where a loop within the binding site had moved out of corresponding density (red circle). (**B**) The fit of the binding site to the density (grey) after binding site minimisation. In both models the best ligand solution is shown (yellow). For clarity only density corresponding to the ligand and the poorly fit loop regions are shown.*

Assessment of the fitting power of the genetic algorithm

To further assess the fitting power of the integrated scoring function the average Pearson correlation coefficient was calculated to identify any relationship between the integrated score and the RMSD to the reference ligand. The average Pearson correlation coefficient was seen to be -0.41 (∓0.25 STD).

Additionally, the previously assessed solutions (Table 2), which were assumed to be the best solutions generated based on the RMSD with the reference ligand, appeared in the top 5 solutions for 19 of the 26 cases. In 7 cases the solution with the best RMSD appeared outside the top 5 solutions, for 5 of these cases a correct solution was still identified within the top 5 solutions. For the remaining 2 cases 2 had no correct ligand solution.

Whilst a comparison to the reference ligand is an important consideration, it does not give a full picture of the fitting power. To assess the power of the algorithm for ranking solutions by the fit to the map, the mean Pearson correlation coefficient between the integrated score and the CCC of the ligand with the deposited full map was calculated. The mean Pearson correlation coefficient was seen to be 0.59 (∓ 0.2 STD). This indicated that the scoring function was generally scoring solutions with better fits to the map higher than those with a worse fit.

No significant correlation between the integrated score and the ligand strain energy was seen with a Pearson correlation coefficient of 0.01. However, within the top 5 ranked solutions for most cases a solution existed with a comparable (∓ 1.5 TEU) strain energy to that of the deposited ligand (Table A1). Indicating that the algorithm and scoring function was producing meaningful results in most cases. There were four cases where no solution within the top 5 had a comparable strain energy to the deposited ligand.

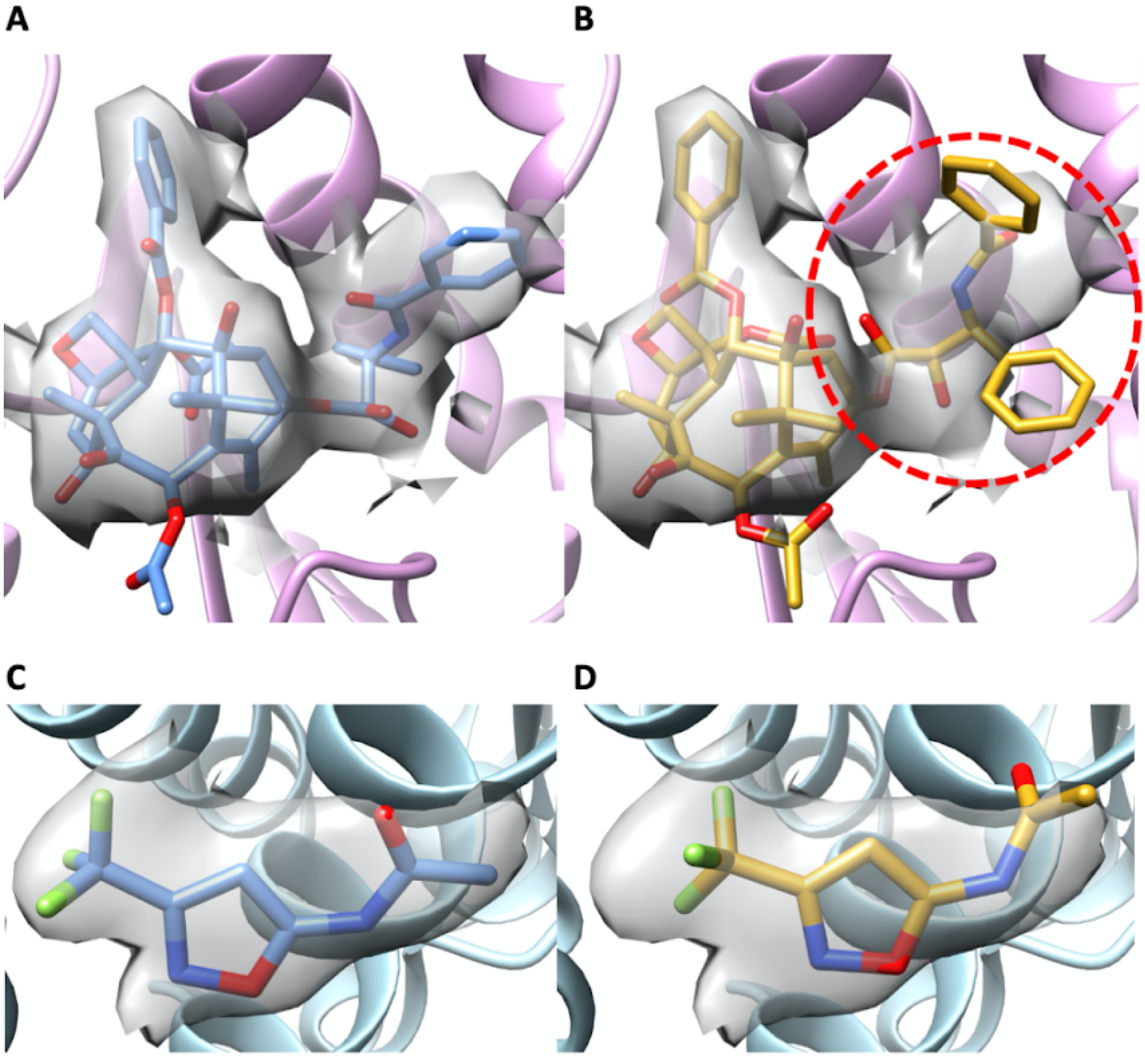


Figure 7. *(**A**) The deposited ligand conformation for the ligand TA1 (PDB ID: 6WVR, dark blue). (**B**) The best solution generated by the GA for the TA1 ligand. An area of poor fit to the density (grey) is highlighted (red circle). (**C**) The deposited ligand confirmation for the ligand NXE (PDB ID: 6TTI, dark blue). (**D**) The best solution generated by the genetic algorithm (yellow). For clarity only density close to the ligand of deposited solutions is shown.*

To investigate the cause of this the solutions for each case were visualised, one case was a large ligand, TA1 (PDB ID: 6WVR) with 15 rotatable bonds and no solution within the top 5 was correct. One solution, ranked outside of the top 5, was within the cutoff for being considered correct, however, still had a significantly higher strain energy than the reference

conformation. Most of the ligand was seen to be fitted correctly, however one region was fitted incorrectly (Figure7). This may explain the high levels of ligand strain seen in this case.

The second case was a small ligand, NXE (PDB ID: 6TTI), with only 2 rotatable bonds. Whilst correct conformations were present in the top 5 solutions the best had an RMSD to the reference ligand of 1.46 Å, which for such a small ligand was quite far from the optimal solution (Figure 7).

The final two cases are the two cases identified previously that contained nucleotides within the binding site. In the case of the ligand XXJ (PDB ID: 6PUZ), the solution was not a correct one and as such was low quality. In the case of the ligand KLQ (PDB ID: 6PUW) the solution was relatively high quality. However, as the interactions between ligand and nucleotide were not taken into account in the scoring function in a meaningful way, the trifluorobenzene ligand moiety that should interact with the ligand was fit incorrectly (Figure 5).

Generally, the GA produced meaningful results that were comparable to deposited structures in terms of CCC and ligand strain energy (Figure A9, Table A1), one example being the case for ligand IMP (PDB ID: 6UDP) where the algorithm identified a ligand conformation with a better CCC to the full map and a lower ligand strain energy (Figure 8) .

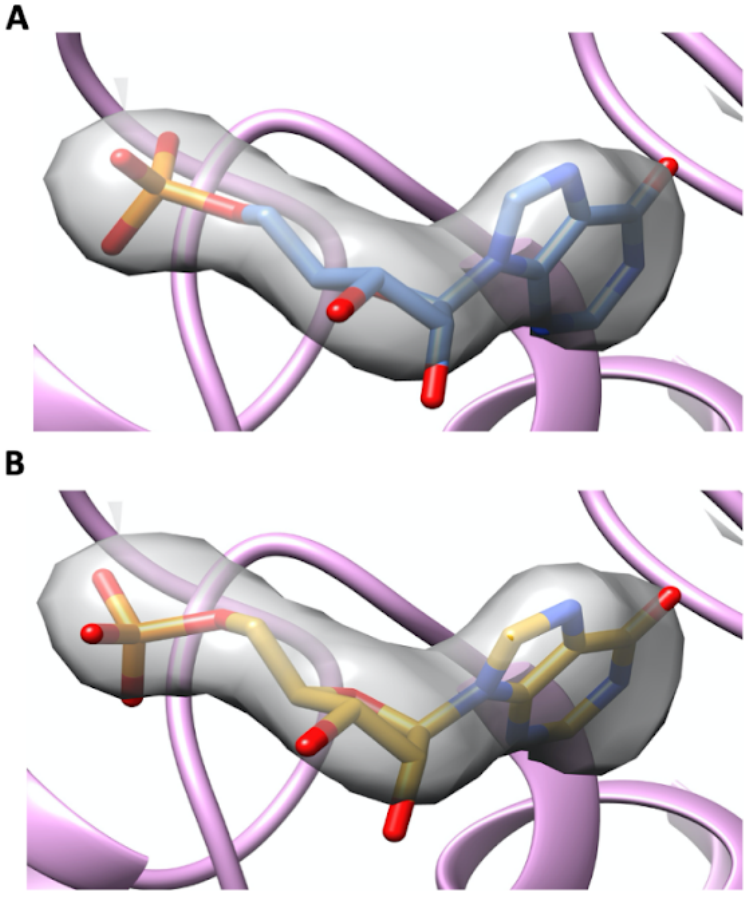


Figure 8. *(**A**) The deposited ligand conformation for the ligand IMP (PDB ID: 6UDP, dark blue). (**B**) The top scoring ligand conformation generated by the GA (yellow). For clarity only density surrounding the ligand is shown (grey).*

A lower resolution benchmark

The results for the high resolution benchmark were encouraging, however, many of the cryo-EM maps deposited in the EMDB are at resolutions between 3.0 Å and 4.5 Å. Therefore, a benchmark of 15 protein ligand complexes derived from cryo-EM maps at resolutions between 3.0 and 4.5 Å was curated (Table 4, Ligand chemical structures are given in Figure A10). For each of the cases, care was taken to ensure that a structure derived from high resolution data existed to compare solutions to.

Table 4. The PDB ID, ligand ID , resolution and CCC with density difference maps for benchmark structures in the resolution range 3.0 Å and 4.5 Å.

PDB	ligand	resolution	CCC		PDB	ligand	resolution	CCC
6U8S	IMP	3.14	0.591		5OAF	ADP	4.06	0.650
7CUM	2BV	3.52	0.646		6K42	CZX	4.1	0.704
6IP2	ATP	3.7	0.521		6RZB	TA1	4.1	0.677
6T24	9ZK	3.7	0.585		6R4O	FOK	4.2	0.712
6N57	1N7	3.7	0.707		6ZIY	FMN	4.25	0.326
6U8R	IMP	3.91	0.673		5WEL	ZK1	4.4	0.807
5W3J	GTP	4.0	0.569		7CKQ	P4P	4.4	0.261
6WHV	QGP	4.05	0.587					

For each protein-ligand complex, a difference map was generated as before and each inspected visually (Figure A11). The CCC of the deposited ligand conformation with difference maps (Table 4) was on average lower than that of the high-resolution benchmark. It was reasoned that this was due to the worse resolution of these maps not containing the finer details of the ligand conformations that the high resolution maps did.

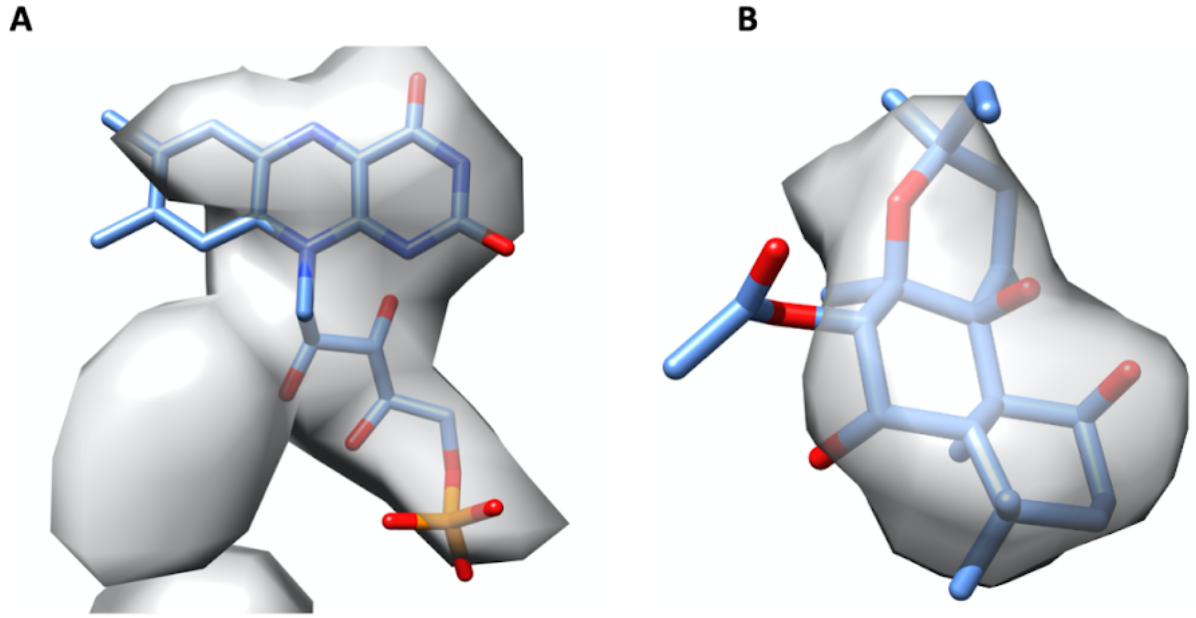


Figure 9. *(**A**) A low quality difference map (grey) obtained for the ligand FMN (blue) (PDB ID: 6ZIY). (**B**) A higher quality difference map for the ligand FOK (PDB ID: 6R4O).*

Furthermore, some maps had a significantly lower CCC with the deposited ligand than the rest. These difference maps were defined as low quality. A visual inspection of one case of ligand FMN (PDB ID: 6ZIY) showed that the difference map contained extra density that did not correspond to the ligand (Figure 9). However, these cases were kept in the benchmark as they represented the quality of difference maps achievable with real world data.

Fitting small molecules with the genetic algorithm at resolutions between 3.0 Å and 4.0 Å

Once again the empirical scoring function was integrated with the MI score weighted at 0.1x, 0.5x, 1x, 5x and, 10x the magnitude of the empirical scoring function, and the GA was run for all cases in the lower resolution benchmark (Figure 10). At all magnitudes tested the combination of MI score with empirical scoring function was better than using the MI or empirical scoring function alone. Furthermore, as with the high-resolution benchmark (Figure 3), the MI score outperformed the empirical scoring function alone.

Interestingly, at this lower resolution the best combination weight was 0.5x, compared to 1x at better resolutions. At this weight a correct solution was identified 86.66 % of the time (13/15 cases), these solutions were ≤ 1.5 Å 60.0 % of the time (9/15 cases), and ≤ 1.0 Å 20 % of the time (3/15 cases). This correlated well with the mean minimum RMSD (Figure 10) of 1.45 Å (∓ 0.54 STD), the lowest seen of any combination. This indicated that there was less information pertaining to the ligand positions within the map at lower resolutions and the empirical scoring function was needed to make up for this shortfall.

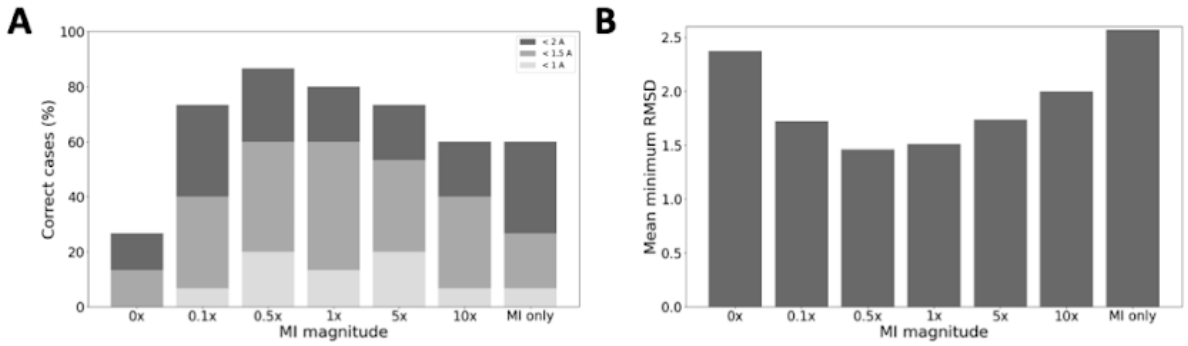


Figure 10. *The results of the genetic algorithm for the 26 cases of the benchmark at 3 - 4.5 Å resolution. (**A**) The percentage of cases where the fitting algorithm generated a solution ≤ 2.0 Å (dark grey), ≤ 1.5 Å (grey), or ≤ 1.0 Å (light grey). (**B**) The mean minimum RMSD values of all cases for the various scoring functions. For both plots the weighting of the MI score with the empirical scoring function is indicated (MI magnitude).*

To account for model bias all ligands were extracted from the deposited protein models and all residues in the binding site refined into the full maps with Flex-EM [28, 29]. The number of correct solutions declined slightly to 80.0 % (12 / 15 cases). The quality of the solutions also decreased, where only 53.33 % of them were found at ≤ 1.5 Å from the reference ligand, and only 6.67 % (1/15 cases) were less than 1.0 Å . This decline in quality correlated with a decline in the mean minimum RMSD from 1.45 (∓ 0.54 STD) to 1.68 Å (∓ 0.54 STD), however, this was not seen to be statistically significant (p=0.26).

The results using the density difference maps proved better than when the analysis was conducted using the full maps rather than density difference maps. Using full maps with the re-refined protein models resulted in a correct solution found only 46.66 % of the time (7/15 cases). The quality of solutions using the full map was drastically reduced with only 20.0 % (3/15 cases) of solutions found at ≤ 1.5 Å, and 6.6 % found at ≤ 1.0 Å to the reference ligands. The mean minimum RMSD was seen to be 2.25 Å (∓ 1.15 STD), a statistical decrease from using the difference maps (p=0.02).

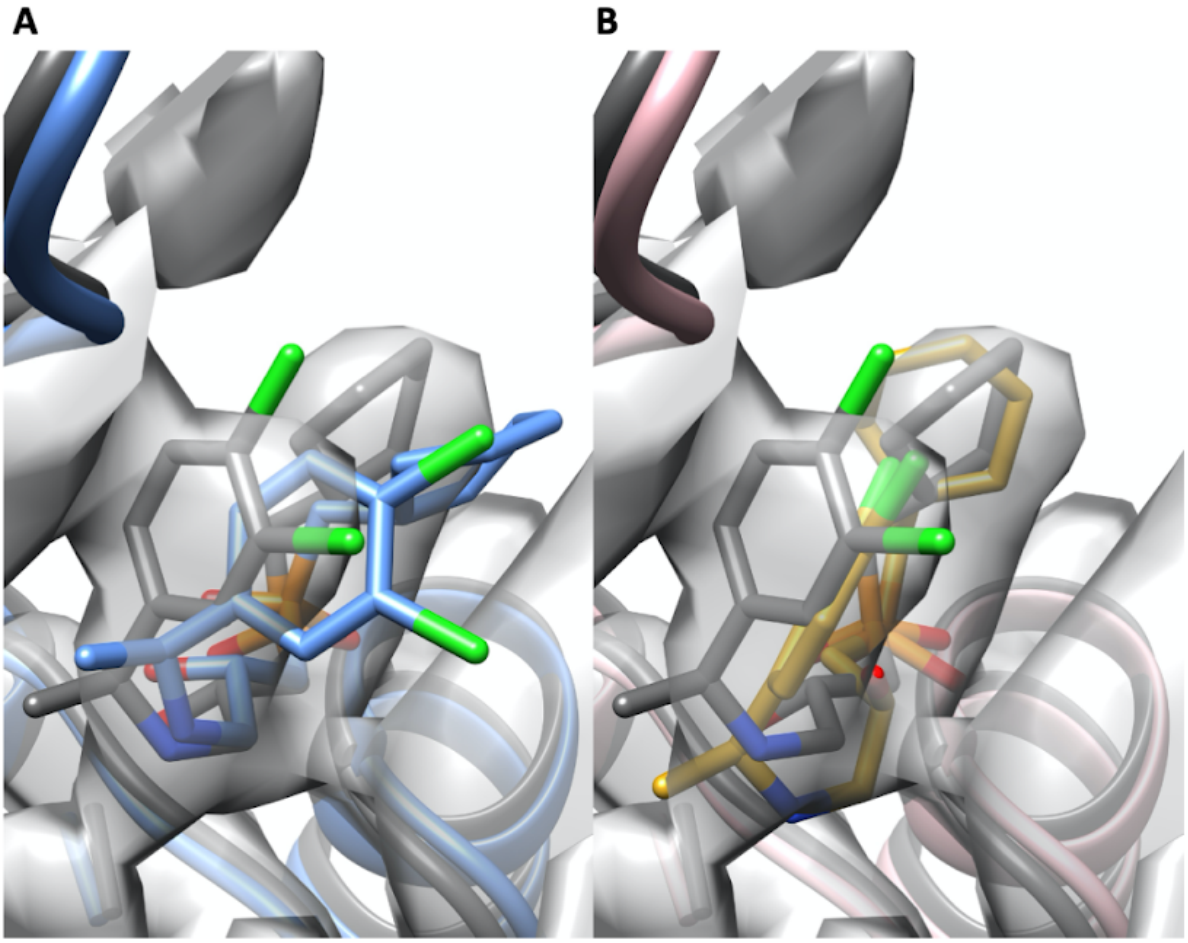


Figure 11. *(**A**) The deposited reference ligand 2BV (PDB ID: 7CUM, blue) fitted to the deposited density map (grey). Also shown is a structural alignment of the high resolution control model (PDB ID: 7C7S, dark grey) with the deposited protein model. (**B**) The second ranked solution from the GA (yellow) that showed the best CCC with the deposited density map (grey). Also shown is a structural alignment of the re-refined protein model used to fit the ligand (pink) with the high resolution control structure (PDB ID: 7C7S, dark grey).*

The algorithm failed to find correct conformations in three cases. The first of these cases was for the 2BV ligand (PDB ID: 7CUM) fit at 3.52 Å. The deposited reference solution was compared to a high resolution control structure (PDB ID: 7C7S, 2.9 Å) and was seen to fit poorly to the map (Figure 11). The top 5 solutions generated by the GA (Table A2, Figure A12) were seen to have a higher CCC to the full map than the reference ligand, the highest of which was seen to be more representative of the conformation contained within the high resolutions structure (Figure 11). However, the strain energy was to be significantly higher than for both the low resolution and high resolution conformations. This may be due to incorrect torsion angles of the acyl chain region or difluorobenzene moiety in the solution, which did not match the high resolution control (Figure 11).

The second case was for the ligand TA1 (PDB ID: 6RZB) fit in a map of 4.1 Å, for which the deposited ligand model corresponds well to that seen in the high-resolution control (PDB ID 6WVR) (Figure 12). The best solution generated in the top 5 (with respect to RMSD to the reference ligand), had an RMSD to the reference ligand of 2.85 Å and was seen to correlate poorly with the high resolution control (Figure 12). This was possibly due to the lack of defining density for the ligand in the full map and the large number of rotatable bonds within the ligand (15 rotatable bonds).

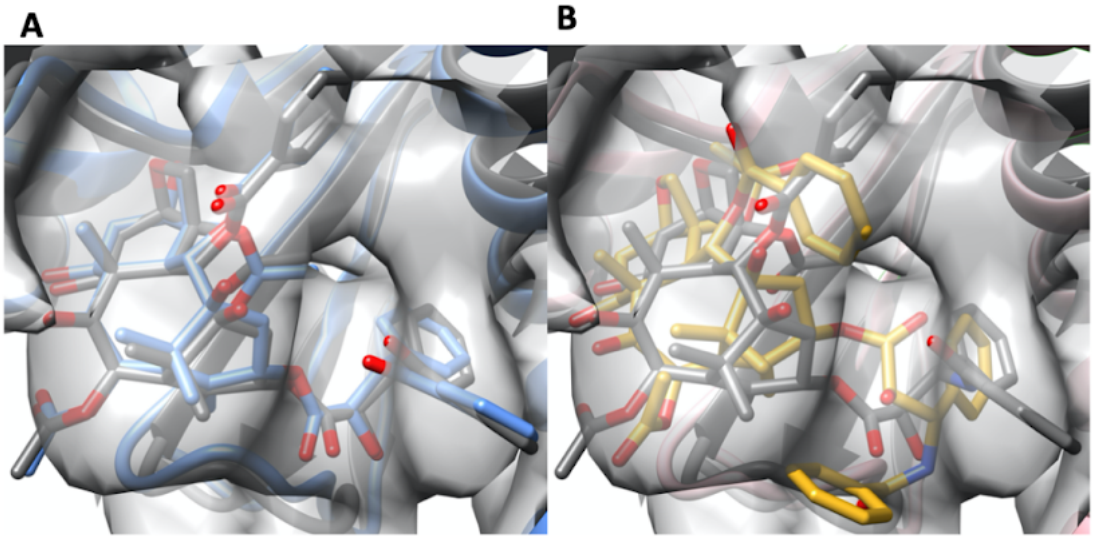


Figure 12. *(**A**) structural alignment between the deposited ligand (blue) and high resolution control (dark grey).(**B**) A structural alignment between a solution given by the GA (ligand: yellow, protein: pink) and the high resolution control (dark grey). The full map density around the ligand binding site deposited with the model 6RZB is also shown in both panels (grey).*

The final failed case was for the ligand ZK1 (PDB ID: 5WEL) deposited with a density map at the upper resolution limit of the benchmark (4.4 Å). Upon inspection of the re-refined protein model it was seen that the protein side chain of TYR450 had moved into density corresponding to the ligand (Figure 13). When this side chain was manually moved out of the density and the analysis rerun the top solution had an increased CCC with the full map and a lower ligand strain energy than the deposited solution (Figure 13).

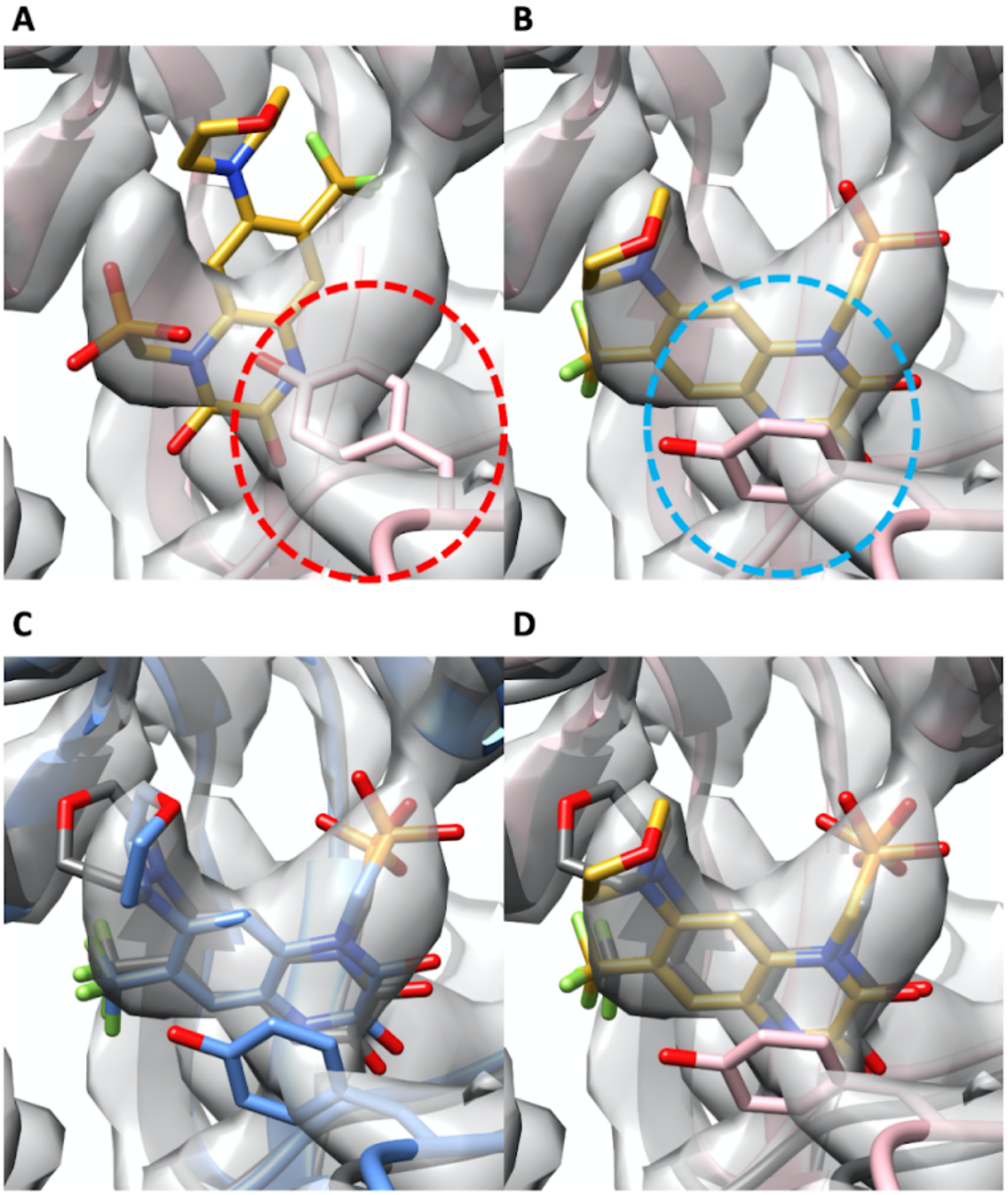


Figure 13. *(**A**) The re-refined protein model for 5WEL (pink) where the TYR450 can be seen occupying density (grey) corresponding to the position of the ligand (red circle). The solution with the best RMSD to the reference ligand is also shown (yellow). (**B**) The re-refined structure after manual intervention (pink) to move the tyrosine 450 residue (blue circle). The top ranked ligand conformation given by the GA is also shown (yellow). (**C**) A structural alignment of the deposited protein model (PDB ID: 5WEL) with the ligand ZK1 and the high-resolution reference structure (PDB ID: 6FQK, dark grey). (**D**) A structural alignment between the top ranked solution given by the genetic algorithm (yellow) with the modified re-refined protein model (pink) and the high resolution control (dark grey).*

The fitting power of the GA at 3.0 Å to 4.5 Å

The results using the re-refined protein models and difference maps had an average Pearson correlation coefficient with the ligand CCC (with the full map) of 0.231 (∓ 0.32 STD). This was significantly lower than what was seen with the high-resolution benchmark ($p < 0.001$), and correlated with the reduction in the number of correct cases and quality of solutions.

Once again ligand strain energy was employed to analyse the quality of ligand solutions. For all the solutions generated, there were 4 cases where no solution existed in the top 5 scored conformations that were within the 1.5 TEU strain energy cut off, when compared to reference ligands. Two of these cases have already been addressed; one represented a case with a large amount of rotatable bonds and a correct solution was not found (TA1, PDB ID: 6RZB). The other was the 2BV ligand case (PDB ID: 7CUM) where the deposited ligand had a poor fit. A comparison of the torsion angles with that in the high-resolution control showed whilst this solution represented a better fit to the map than the deposited structure, the torsion angles were incorrect (Figure 11).

A comparison of the case with ligand ADP (PDB ID: 5OAF) with the deposited ligand showed that the solutions given by the GA were not correlated well with the reference solution (Figure 14). However, interestingly the top scored solution, which was deemed incorrect with respect to the RMSD to the deposited reference ligand, shared some similarities with the high-resolution ligand (PDB ID: 2C9O) that the deposited model did not. ARG76 in the deposited structure had moved out of the binding site. In the high resolution control, the β-phosphate group makes a hydrogen bond with this arginine that is not present in the deposited model. This moves the phosphate chain to a position that does not correlate well with the high resolution structure (Figure 14). In the re-refined protein this arginine is moved back into the binding site, allowing the β-phosphate to make the missing hydrogen bond and moving the position of the β-phosphate and ribose moiety of the adenosine group closer to that of the high resolution structure (Figure 14). However, the torsion angles of the phosphate chain appear incorrect when compared to the high-resolution structure; this and the incorrect torsion of the ring system in the adenosine group may explain the relatively high strain energy of the solution.

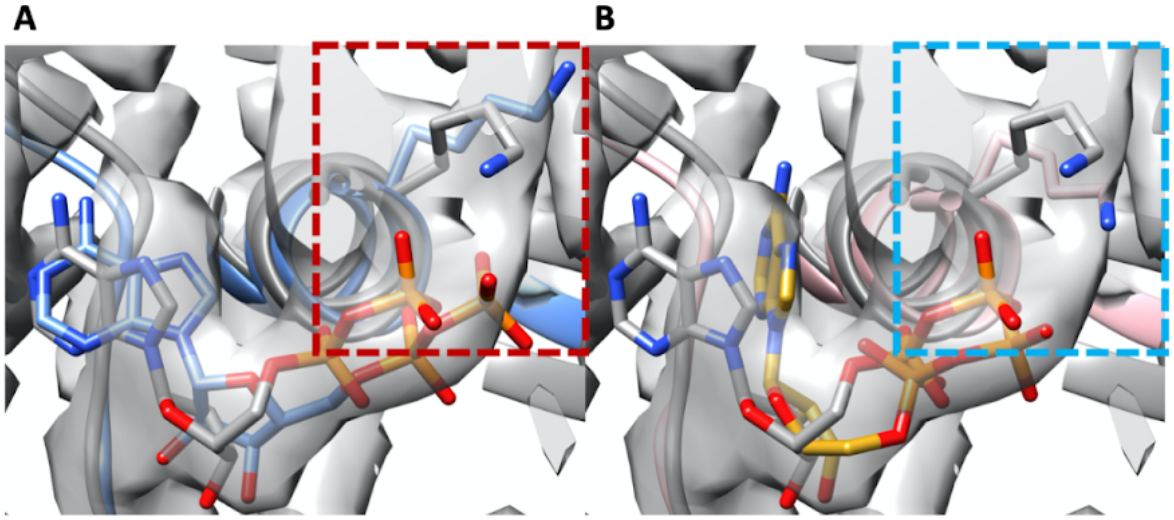


Figure 14. *(**A**) A structural alignment of the high-resolution control model (PDB ID: 2C9O, dark grey) and the deposited atomic model (PDB ID: 5OAF, blue). The positions of ARG76 and β-phosphate groups of the ATP molecules are shown (red box). (**B**) A structural alignment of the top ranked solution given by the GA (yellow) fitted into the re-refined protein model (pink), and the high-resolution control (dark grey). The positions of β-phosphate groups and ARG76 sidechains are also shown (blue box). In both panels the density deposited with the 5OAF model is shown (light grey).*

The final case where the top 5 ranked solutions were of poor quality with respect to strain energy was for the ligand 9ZK (PDB ID: 6T24). The top ranked ligand solution was relatively well fitted in the density, with an RMSD of 1.3 Å to the reference ligand. However the strain energy was more than double that of the reference ligand (21.21 TEU and 8.3 TEU, respectively). Inspection of the fits and a comparison to the high-resolution control, indicated two areas that were of a poor quality fit (Figure 15), one ring system that had a flipped conformation and a long acyl chain had incorrect torsions. The poor torsions of these chemical groups most likely account for the increased ligand strain energy. It should be pointed out that there was no defining density for the acyl group in the deposited map.

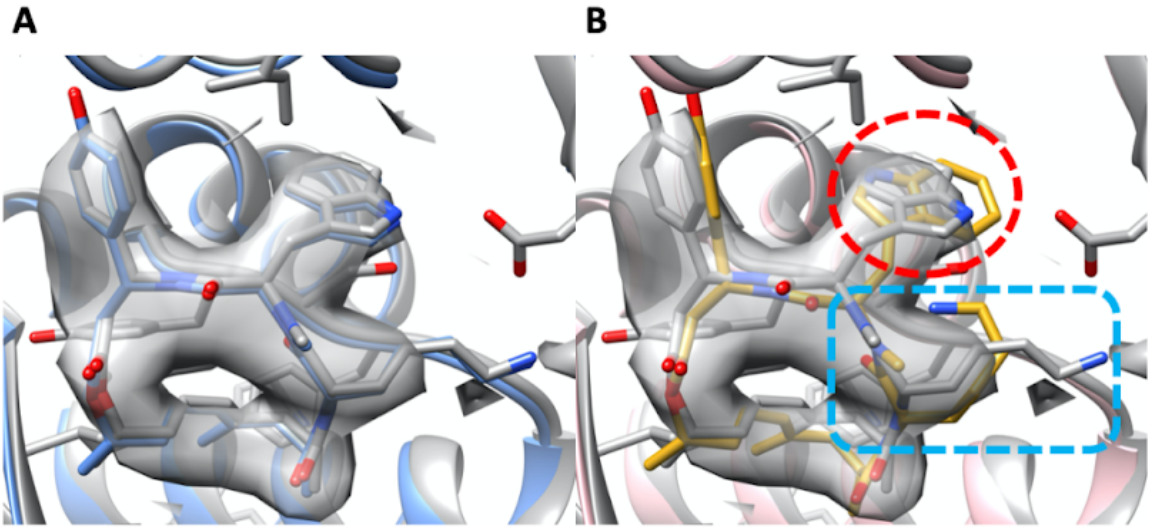


Figure 15. *(**A**) A structural alignment between the deposited model in the case with ligand 9ZK (PDB ID: 6T24, blue) and a high resolution control (PDB ID: 6T23, dark grey). The deposited density map is also shown (grey). (**B**) A structural alignment between the re-refined protein model for 6T23 (pink) and the top solution given by the GA (yellow). Areas of poor fit to the map are indicated (red circle, blue box).*

Overall the quality of ligand solutions in the lower resolution benchmark was lower than that of the high-resolution benchmark; however, a large amount of correct fits were found along with a few high-resolution solutions. One such case is for the ligand FOK (PDB ID: 6R4O), where the third highest scoring solution given by the GA had a higher CCC of the ligand with the full map and a lower strain energy than the deposited reference model, 0.17 and 3.8 TEU, and , 0.16 and 6.97 TEU, respectively (Figure 16). A comparison to the high-resolution control did not correlate well with either solution when a structural alignment was made. However, the structures were seen to match poorly and the reference and the GA solutions were clearly seen to fit the density better. Furthermore, many of the same bonds made by the high-resolution control were made in the deposited structure and the GA model.

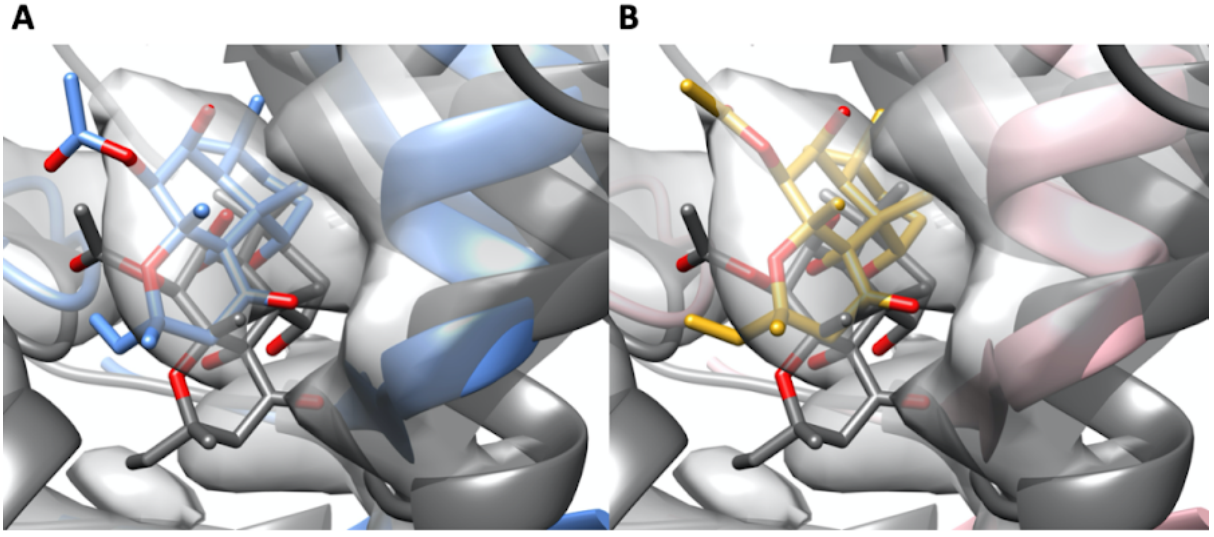


Figure 16. *(**A**) A structural alignment between the deposited model with the ligand FOK (PDB ID: 6R4O, blue) and a high-resolution control (PDB ID: 1CUL, dark grey). (**B**) The re-refined protein model (pink) and a solution generated by the GA (yellow). The density map deposited with the atomic model 6R4O is shown in both panels (grey).*

Binding site minimisation

Binding site minimisation was conducted after ligand solutions were generated. MolProbity scores for the minimised structures with ligands that showed the best RMSD to the reference ligands indicated that this minimisation improved the quality of protein models (Table 5).

Table 5. MolProbity scores for the re-refined protein models and after binding site minimisation.

PDB ID	MolProbity score re-refined	MolProbity score minimised		PDB ID	MolProbity score re-refined	MolProbity score minimised
6U8S	2.79	2.64		5OAF	1.86	1.88
7CUM	1.99	1.99		6K42	2.0	1.83
6IP2	2.23	2.20		6RZB	2.53	2.53
6T24	1.10	0.86		6R4O	2.27	2.20
6N57	2.86	2.86		6ZIY	2.14	2.06
6U8R	2.69	2.67		5WEL	2.09	2.09
5W3J	2.17	2.04		7CKQ	2.74	2.74
6WHV	2.91	2.90				

Discussion

Drug discovery with cryo-EM is becoming more prevalent as more and more high resolution structures of protein targets bound to small molecules are deposited in the EMDB. However, is it still commonplace to solve such structures to resolutions between 3.0 and 4.5 Å. Thus, moving forward tools for the accurate and automated fitting of small molecules into cryo-EM density maps are needed. Here we presented a novel algorithm that utilises density difference mapping for such a task. It was shown to be able to accurately fit ligands at resolutions up to 4.5 Å.

At high resolution the algorithm produced many meaningful results with comparable fits to the deposited cryo-EM maps as the reference ligands (Figure A9). However, it was seen that for a few cases no correct solution was found. One of these cases had nucleotide present in the binding site, the benchmark also included a second case with nucleotide in the pocket (6PUW). For the second case, a relatively high-quality correct solution was found, however, in this solution the interactions between ligand and nucleotide were absent from the best

solution (Figure 5). This indicated that when nucleotides were present within the binding sites, the solutions generated may require further refinement.

The second incorrect case represented a case where during re-refinement ligand atoms had moved into the binding site. This brings up an important point regarding the algorithm. For a good fit to be identified one must check that the binding site corresponds well with the map before fitting. The effect of this is exemplified by comparing the quality of the solutions generated using deposited structures than with the re-refined output, where although a comparable number of correct solutions were generated, the quality of the output with the deposited protein models was much higher than for the re-refined models (Figure 3). Furthermore, for the ligand ZK1 in the low resolution benchmark, manually moving a side chain that had moved into the binding site during refinement gave rise to a solution that had a better CCC with the deposited full map and a lower ligand strain energy (Figure 13). To improve on this work future iterations of the algorithm should take into account rearrangements of protein atoms within the binding site. One way to achieve this would be to integrate the algorithm with flexible-fitting MDs. However, this was considered outside the scope of the investigation.

To evaluate the quality of solutions a statistical approach was used to calculate ligand strain energies [37]. The ligand strain energy correlated well with solutions far in RMSD from the reference ligands. Additionally, it was also able to identify when solutions that were close in RMSD to the reference ligands had regions that were poorly fit to the full map. This indicated that the statistical approach to calculate strain energy was valid. Furthermore, one way to improve on further iterations of the algorithm may be to directly include the strain energy calculations when generating solutions. For both high-resolution and low resolution benchmarks there was no significant correlation between the CCC or RMSD and the ligand strain energy. The strain energy only evaluates the quality of the ligand conformation in isolation from the fit to the map or the protein/ligand interface. The reason no correlation was seen is most likely due to the optimisation algorithm. The GA optimised the positions of ligands over many iterations. Therefore, it is entirely possible for a ligand to have a high quality conformation with a poor fit to the map. Thus, no correlation between ligand strain energy and the RMSD or CCC was to be expected.

One trend that was noticeable was that the solutions with larger ligands (greater number of rotatable bonds) were generally of a lower quality than those with a lower number of rotatable bonds. This indicated that the default search parameters for the GA may need to be increased (i.e. number of genetic runs or initial solutions) to account for the increase in chemical space needed to be searched.

With the high-resolution benchmark the scoring function alone was able to identify a correct ligand conformation 53.85 % of the time (14 / 26 cases). This was lower than what has been shown for molecular docking algorithms such as AutoDock Vina [27]. That was seen to produce solutions $\leq$ 2.0 Å to reference ligands, 78 % of the time when tested on 190 high

resolution complexes. However, the same paper showed that the AutoDock 4 [42] algorithm was only able to produce solutions ≤ 2.0 Å 49 % of the time.

This indicated that the scoring function had some power for molecular docking in its own right. Although, it should be noted that the GA was designed to fit molecules into density difference maps, therefore it was optimised to search a large number of conformations in a small area, with the use of the local and fine tuning stages (Figure 1). However, optimising the algorithm specifically for molecular docking was outside the scope of this study.

For both the high and lower resolution benchmarks the solutions calculated when using the density difference maps were higher than using the full maps. The greatest decline in performance was seen with the lower resolution benchmark with only 46.66 % of cases identified correctly with the full maps compared to 80.0 % with the difference maps. The decline in performance correlated with the resolution of the maps. This was to be expected as higher resolution maps have more defined density. The fact that the algorithm was able to identify correct fits at resolutions between 3.0 Å and 4.5 Å correlated well with two cases reported in literature where density difference mapping has been used to overcome the limitations associated with fitting ligands at low resolution [25, 30].

Nevertheless, the quality of ligand conformations solutions were dependent on the quality of density difference map obtained, at both high and low resolution. Examples of low quality difference maps in the high resolution benchmark included the difference map for the ligand NXE (PDB ID: 6TTI), with an example of a high quality difference map being for the ligand ZK1 (PDB ID: 6PEQ) (Figure 2). The cross correlations of the deposited reference ligands and density difference maps were seen to be 0.343 and 0.786 for the ligands, NXE and ZK1, respectively. The quality of difference map was seen to have an effect on fitting, was evident, where for the case of the NXE ligand only one solution in the top 5 was considered correct (Figure 7). This solution had a lower CCC with the full map and the ligand strain energy was significantly higher (Table 2). Conversely, in the case of ZK1, one ligand ranked in the top 5 solutions had a comparable cross correlation and ligand strain energy to the reference structure (Table 2, Figure 17).

The overall quality of the difference maps in the low-resolution benchmark were less than that of the high-resolution benchmark. Within the low resolution difference maps there were various degrees of map quality. An example of a low quality difference map was for the ligand FMN (PDB ID: 6ZIY), where the reference ligand had a CCC to the difference map of 0.326 (Figure 9). An example of a higher quality difference map was seen for the ligand FOK (PDB ID: 6R4O), where the reference ligand had a CCC of 0.712 with the difference map (Figure 9).

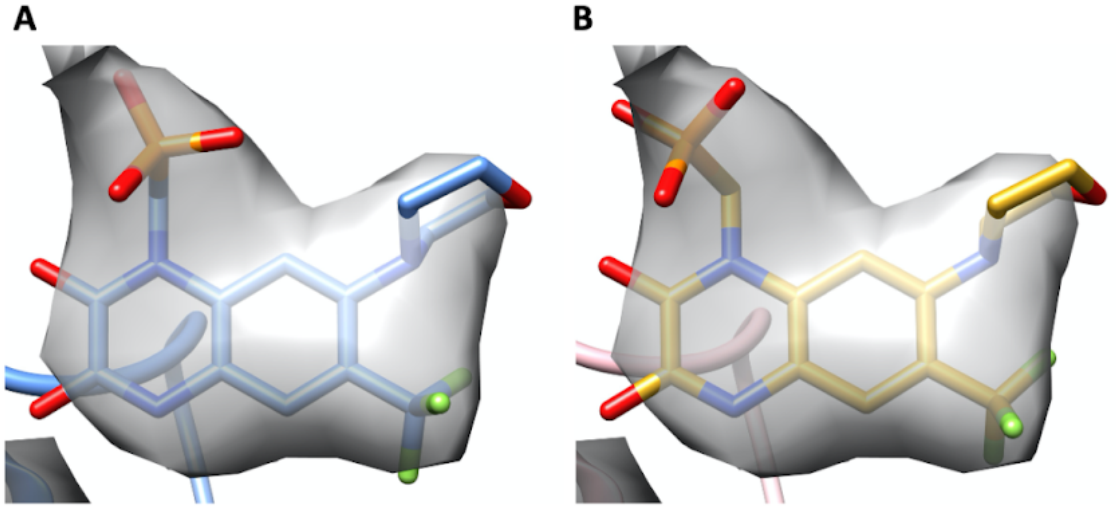


Figure 17. *(**A**) The deposited reference ligand ZK1 (PDB ID: 6PEQ) (**B**) A top ranked solution given by the genetic algorithm for the ligand ZK1. The deposited full map is also shown in both panels (grey). For clarity only density surrounding the deposited ligand is shown.*

The effect this had on fitting was that for the FMN ligand no solution in the top 5 predicted was correct. One solution ranked outside the top 5 had an RMSD of less than 2.0 Å to the reference solution. However, visual inspection of the solution shows that a flexible region of the ligand has moved into density present in both the difference and full map that did not correlate to the ligand (Figure 18). However, the fitted solution for the fitting ligand FOK showed an increased CCC with the full map than the deposited ligand (Figure 16). These results indicated that obtaining a high quality difference map for fitting increases the chances of generating a high quality solution.

Recently, a software GemSpot was reported that utilised the GLIDE molecular docking software to fit ligands into cryo-EM maps [16]. A direct comparison between the two methods was not conducted in this research however, the high resolution benchmarks for both had two overlapping ligands S9R (PDB ID: 6A95) and J6E (PDB ID: 6QM7). The GemSpot software was able to produce fits with comparable CCC to the deposited experimental models, as did the GA presented here. However, the ligand strain energies for each solution is not reported for GemSpot. Furthemore, both Gemspot and the GA were able to produce reasonable fits at resolutions above 3.0 Å. Another recent publication that utilised the GLIDE docking software and neural network potentials (NNPs) for fitting small molecules into cryo-EM maps, suffers from drawbacks relating to the database of NNP parameters. The NNP database used in the report did not contain information for all chemical groups, for example phosphate groups [20]. Our methodology does not suffer from such drawbacks and was able to fit ligands at the same resolutions shown in the NNP methodology report.

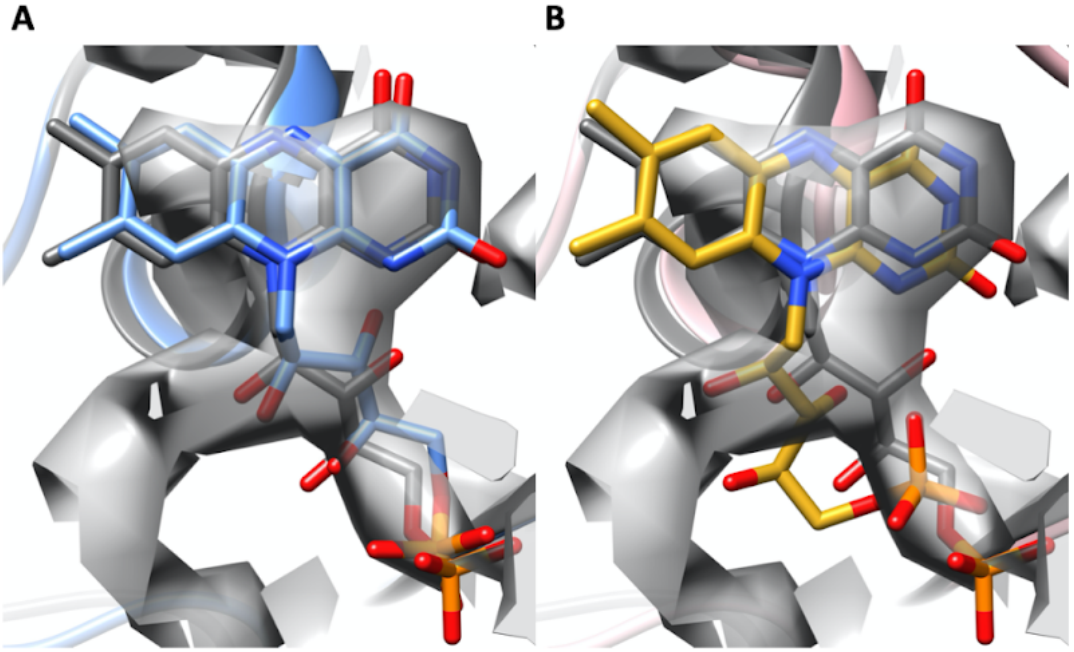


Figure 18. *(**A**) A structural alignment between the deposited protein model for the PDB 6ZIY (blue) with the ligand FMN, against a high resolution control (PDB ID: 3I9V, dark grey). (**B**) A structural alignment for the re-refined protein model (pink) and a solution from the GA (yellow) against the high-resolution control. The full map deposited with the atomic model 6ZIY is shown in both panels (grey). For clarity only density around the deposited ligand is shown.*

The GA presented here represents the first software specifically for fitting small molecules to cryo-EM maps that directly utilised difference mapping. The results showed that the solutions were comparable to current state-of-the-art software in the field for fitting ligands in terms of the resolution that we were able to produce correct fits [16, 20]. Additionally, the inclusion of difference mapping in our method represents a novel step distinct from other current methodologies.

Methods and software

Benchmark atomic models were obtained from the PDB [35] along with '.sdf' files of deposited small molecules. Cryo-EM maps were obtained from the EMDB. The GA software was written in python-3 and individual test cases were run on a single cpu of machine with 128Gb of memory and a AMD 3950X cpu (16 cores).

Difference density maps were calculated using the local scaling method [39] implemented in TEMPy . The input to the software were the deposited full maps, the protein pdb file in the *apo* state, and the resolution. For all options the default setting was used.

Protein re-refinement was conducted using a single round of MD-flexible fitting with Flex-EM [28, 29] Briefly, ligands were removed from atomic models and any atoms not within 15 Å of the binding site were restrained from moving during refinement using a rigid body file. The number of MD cycles was set to 1 all other options were used as default.

RMSD calculations were completed using functions that were built into the software. The RMSD indicated here was calculated across all heavy atoms (i.e. excluding hydrogens), using the formula (Eq 1):

$$Eq\,1.\ RMSD = \sqrt{\frac{1}{n}\sum \|x_i - y_i\|^2}$$

Where, xi, yi are ligand atom pairs in the ligands x and y, respectively.

References

1. Sugiki T, Furuita K, Fujiwara T, Kojima C. Current NMR Techniques for Structure-Based Drug Discovery. Molecules. 2018;23. doi:10.3390/molecules23010148.

2. Orlova EV, Saibil HR. Structural Analysis of Macromolecular Assemblies by Electron Microscopy. Chem Rev. 2011;111:7710–48.

3. Emsley P, Lohkamp B, Scott WG, Cowtan K. Features and development of Coot. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 4:486–501.

4. Adams PD, Afonine PV, Bunkóczi G, Chen VB, Davis IW, Echols N, et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 2:213–21.

5. Afonine PV, Poon BK, Read RJ, Sobolev OV, Terwilliger TC, Urzhumtsev A, et al. Real-space refinement in PHENIX for cryo-EM and crystallography. Acta Crystallogr D Struct Biol. 2018;74 Pt 6:531–44.

6. Debreczeni JÉ, Emsley P. Handling ligands with Coot. Acta Crystallogr D Biol Crystallogr. 2012;68 Pt 4:425–30.

7. Moriarty NW, Grosse-Kunstleve RW, Adams PD. electronic Ligand Builder and Optimization Workbench (eLBOW): a tool for ligand coordinate and restraint generation. Acta Crystallogr D Biol Crystallogr. 2009;65:1074–80.

8. Kim JJ, Gharpure A, Teng J, Zhuang Y, Howard RJ, Zhu S, et al. Shared structural mechanisms of general anaesthetics and benzodiazepines. Nature. 2020;585:303–8.

9. Bai Y, Yu X, Chen H, Horne D, White R, Wu X, et al. Structural basis for pharmacological modulation of the TRPC6 channel. Elife. 2020;9:e53311.

10. Zhang X, Belousoff MJ, Zhao P, Kooistra AJ, Truong TT, Ang SY, et al. Differential GLP-1R Binding and Activation by Peptide and Non-peptide Agonists. Mol Cell.

2020;80:485–500.e7.

11. Roos K, Wu C, Damm W, Reboul M, Stevenson JM, Lu C, et al. OPLS3e: Extending Force Field Coverage for Drug-Like Small Molecules. J Chem Theory Comput. 2019;15:1863–74.

12. van Zundert GCP, Moriarty NW, Sobolev OV, Adams PD, Borrelli KW. Macromolecular refinement of X-ray and cryoelectron microscopy structures with Phenix/OPLS3e for improved structure and ligand quality. Structure. 2021;29:913–21.e4.

13. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007;35 Web Server issue:W375–83.

14. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 1:12–21.

15. Sobolev OV, Afonine PV, Moriarty NW, Hekkelman ML, Joosten RP, Perrakis A, et al. A Global Ramachandran Score Identifies Protein Structures with Unlikely Stereochemistry. Structure. 2020;28:1249–58.e2.

16. Robertson MJ, van Zundert GCP, Borrelli K, Skiniotis G. GemSpot: A Pipeline for Robust Modeling of Ligands into Cryo-EM Maps. Structure. 2020;28:707–16.e3.

17. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. J Med Chem. 2004;47:1739–49.

18. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, et al. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes. J Med Chem. 2006;49:6177–96.

19. Michel J, Tirado-Rives J, Jorgensen WL. Prediction of the Water Content in Protein Binding Sites. J Phys Chem B. 2009;113:13337–46.

20. Vant JW, Lahey S-LJ, Jana K, Shekhar M, Sarkar D, Munk BH, et al. Flexible Fitting of Small Molecules into Electron Microscopy Maps Using Molecular Dynamics Simulations with Neural Network Potentials. J Chem Inf Model. 2020;60:2591–604.

21. Leaver-Fay A, Tyka M, Lewis SM, Lange OF, Thompson J, Jacak R, et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. Methods Enzymol. 2011;487:545–74.

22. Wang R, Fang X, Lu Y, Yang C-Y, Wang S. The PDBbind Database: Methodologies and Updates. J Med Chem. 2005;48:4111–9.

23. Barad BA, Echols N, Wang RY-R, Cheng Y, DiMaio F, Adams PD, et al. EMRinger: side chain-directed model and map validation for 3D cryo-electron microscopy. Nat Methods. 2015;12:943–6.

24. Igaev M, Kutzner C, Bock LV, Vaiana AC, Grubmüller H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. Elife. 2019;8:e43542.

25. Locke J, Joseph AP, Peña A, Möckel MM, Mayer TU, Topf M, et al. Structural basis of human kinesin-8 function and inhibition. Proc Natl Acad Sci U S A. 2017;114:E9539–48.

26. Dominguez C, Boelens R, Bonvin AMJJ. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc. 2003;125:1731–7.

27. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455–61.

28. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein Structure Fitting and Refinement Guided by Cryo-EM Density. Structure. 2008;16:295–307.

29. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42–9.

30. Peña A, Sweeney A, Cook AD, Locke J, Topf M, Moores CA. Structure of Microtubule-Trapped Human Kinesin-5 and Its Mechanism of Inhibition Revealed Using Cryoelectron Microscopy. Structure. 2020;28:450–7.e5.

31. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. J Mol Biol. 1997;267:727–48.

32. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015;11:3696–713.

33. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field. J Comput Chem. 2004;25:1157–74.

34. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge Structural Database. Acta Crystallogr B Struct Sci Cryst Eng Mater. 2016;72 Pt 2:171–9.

35. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.

36. Schärfer C, Schulz-Gasch T, Rarey M. TorsionAnalyzer: exploring conformational space. J Cheminform. 2013;5 Suppl 1:P3.

37. Gu S, Smith MS, Yang Y, Irwin JJ, Shoichet BK. Ligand Strain Energy in Large Library Docking. J Chem Inf Model. 2021;61:4331–41.

38. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J Med Chem. 2012;55:6582–94.

39. Joseph AP, Lagerstedt I, Jakobi A, Burnley T, Patwardhan A, Topf M, et al. Comparing Cryo-EM Reconstructions and Validating Atomic Model Fit Using Difference Maps. J Chem Inf Model. 2020;60:2552–60.

40. Cragnolini T, Sahota H, Joseph AP, Sweeney A, Malhotra S, Vasishtan D, et al. TEMPy2:

a Python library with improved 3D electron microscopy density-fitting and validation workflows. Acta Crystallogr D Struct Biol. 2021;77 Pt 1:41–7.

41. Bursulaya BD, Totrov M, Abagyan R, Brooks CL 3rd. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des. 2003;17:755–63.

42. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30:2785–91.

Chapter 6

The MX helix of the *Torpedo marmorata* nicotinic acetylcholine receptor in its native membrane

Background

The Nicotinic Acetylcholine receptors (NAChR) are ion channels belonging to the Cys-loop pentameric ligand gated ion channel superfamily of proteins. The cys-loop receptors are defined by a characteristic loop formed from a region of conserved residues flanked by cysteine residues that interact via a disulphide bond. Family members of the Cys-loop receptor superfamily include the NAChRs, 5-HT3 receptors, **γ**-Aminobutyric acid ($GABA_A$) receptors, Glycine receptors, and Glutamate receptors. Each member of the Cys-loop superfamily shares a similar domain architecture with five subunits interacting to form the mature receptor [1]. Each subunit is composed of three main domains, the N-terminal domain (NTD) composed of two main **β**-sheets, a transmembrane domain (TM) with four membrane spanning helices TM1-TM4, and a C-terminal domain (CTD) that varies in size depending on the receptor (Figure 1). Multiple structural studies have confirmed the domain architecture of the Cys-loop receptors with structural models being reported for the, NAChRs [2–4], 5-HT3 [5, 6], Glutamate [7, 8], glycine [9], and $GABA_A$Rs [10, 11]. The NAChRs are relevant drug targets having been associated with a number of pathologies, including epilepsy [12], Alzheimers's disease [13], and Parkinson's disease [14]. Therefore, the development of therapeutic agents relies on a comprehensive understanding of the structure of NAChRs and relating this to NAChR function.

NAChR structure and function

Some of the earliest structural studies were completed on the *Torpedo Marmorata* (Herein will be referred to as *Torpedo*) commonly known as the electric ray [4]. The electric organs of these animals harbour membranes where NAChRs are highly concentrated making them an ideal system for purification and study of NAChRs [15]. Early studies identified four distinct subunits contributing to the formation of the channel [4]. Two **α**-subunits, and one **β**-, **δ**,- and **γ**-subunit were arranged to form the central pore (Figure 1C). The TM2 helices were seen to line the pore with 5 TM2 helices contributing to pore formation. The TM1 and TM3 helices from each subunit were seen to form a 10 helix ring between the TM2 and TM4 helices. The TM4 helix was seen to line the outside of the receptor at the interface with the lipid environment of the membrane.

Structural models of the human NAChR revealed the presence of a small amphipathic helix, the MX helix, within the transmembrane domain [2]. This feature was also seen present in other family member structures such as the 5-HT3 receptors [5, 6] (Figure 1D). The MX helices in these models were seen to be approximately 3 helical turns and positioned close to the intracellular membrane boundary with the cytoplasm. The MX helix was seen to have a tilt with respect to the theoretical plane of the membrane of between 20 ° and 40 ° (Figure 1D). However, the exact position of this helix relative to the rest of the receptor as well as the positions of the MX loop connecting the MX helix to the TM3 helix vary between models. One possible reason for this could be that the receptors were purified in detergent and would have therefore lost potentially stabilising interactions with the lipid environment. Thus, making characterising the function of the MX helix with respect to the lipid environment of the membrane a challenge.

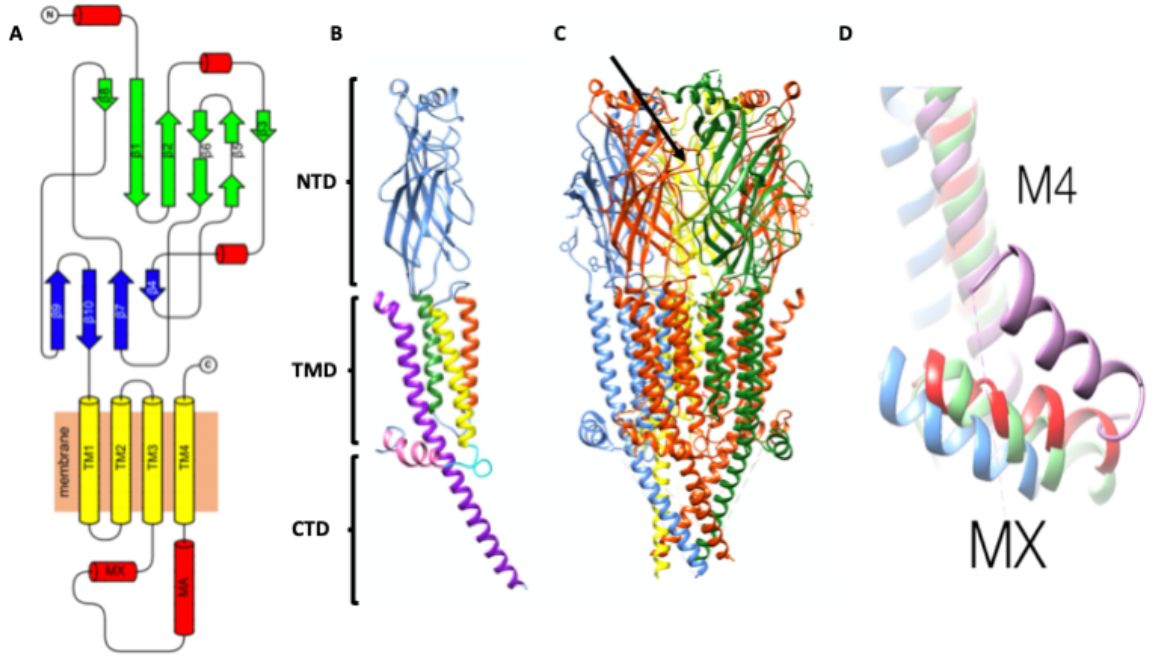


Figure 1. *(**A**). A topology diagram of the NAChR subunit secondary structure elements. (**B**). A structural model of an individual NAChR subunit. The N-terminal domain (NTD), transmembrane domain (TMD), and C-terminal domain (CTD) are indicated alongside. The helices that make up the transmembrane domain are colour coded, TM1 (green), TM2 (red), TM3 (yellow), TM4/MA (Purple), and the MX helix (Pink). The Mx loop is also highlighted (Cyan). (**C**). An atomic model showing the full length receptor of the Torpedo NAChR. Subunits are colored red (**α**), blue (**β**), yellow (**δ**) and green (**γ**). The agonist binding site is also indicated (black arrow). (**D**). An alignment of the MX helices from different Cys-loop models deposited in the PDB; **β**2- (red) and **α**4-helix (green) of the human **α**4**β**2 receptor (PDB ID: 6CNJ) and the mouse 5HT-3 receptor from the deposition 6HIQ (blue) or 6DG8 (purple).*

The NAChRs are found on the postsynaptic side of the synapse. In the presence of their endogenous neurotransmitter, acetylcholine, channels rapidly open becoming permeable to

small cations, Ca^{2+}, K^{+} or Na^{+}, bringing about a rapid change in membrane potential. In most cases there exists two acetylcholine binding sites per receptor located within the NTD at the interfaces of subunits [16], in the *Torpedo* structure these are located between the **α**/**δ** subunit interface and the **α**/**γ** subunit interface [17]. Binding of the agonist to this site forces the C-loop to close over and initiates receptors opening [16].

The ion channel itself is gated by three rings of hydrophobic residues within the TM2 helix [18]. In the closed state these rings create a hydrophobic environment that is energetically unfavourable with respect to ion permeability. This gate is allosterically coupled to the agonist binding site. Agonist binding induces conformational changes in the M2 helix enabling the channel to become permeable to ions. Evidence suggests that the *Torpedo* M2 helix moves from a kinked conformation to a more straight conformation upon agonist binding [18]. This has the effect of moving the narrowest section of the pore from the hydrophobic ring approximately midway up the channel, to a region close to the intracellular membrane, here residues that line the pore are hydrophilic, this creates a more favourable energetic environment for the passage of ions through the pore.

It has been shown that the lipid environment influences receptor function with a minimum number of lipids needed for receptor ion permeability in response to agonist, with this activity decreasing with the ratio of lipid to protein within the membrane [19]. In addition to the correct lipid:protein ratio, it has been shown that cholesterol is required to maintain channel gating properties of *Torpedo* NAChRs within the membrane environment [20]. A Molecular Dynamics (MD) simulation of the NAChR within its membrane environment identified TM4 as a possible lipid sensing motif, where during the simulation TM4 flipped between making contacts with the TM1/TM3 helices and the lipid environment [21]. Further evidence for the lipid sensing role of the TM4 helix was seen in experiments that used single-molecule kinetic analysis to show that the TM4 helix undergoes significant structural rearrangements during the channel open to close transition [22]. Taken together this indicates that there is significant interplay between channel gating the lipid environment and the TMD.

Early experimental evidence indicated that two pools of cholesterol interacted with NAChRs, one of which was seen to bind tightly to the receptor [23]. The cholesterol binding CARC motif has been identified and studied within the context of the *Torpedo* NAChR. Photoaffinity labelling experiments have shown that cholesterol interacts with the TM1, TM3 and TM4 helices of the *Torpedo* NAChR [24]. Furthermore, MD studies found that cholesterol interacted with the *Torpedo* NAChR in both a superficial manner as well as at sites more deeply buried within the TMD. It was seen that the absence of these cholesterol molecules caused the receptor to collapse, indicating cholesterol plays a critical role in receptor structure [25].

In order to gain further structural insight of the MX helix within the native membrane, this investigation aimed to build an atomic model of the *Torpedo* NAChR into a density map of a receptor in its native membrane. To achieve this goal we used a method of purification for the NAChRs from their native membrane that did not involve the use of detergent thus

maintaining their native structure within the lipid environment [26]. Homogenates were extracted from the *Torpedo* electric organ in and purified by centrifugation and incubated with an appropriate buffer that facilitated the formation of tubular crystals with receptors sitting side by side in a 3-dimensional (3D) helical array, from which different helical families can arise [27]. 3D density maps were created by averaging the members belonging to a single helical family around the tube [27]. Here we show that the MX helices of the *Torpedo* NAChR are positioned at the intracellular surface of the lipid bilayer, excluding larger phospholipid head groups from these regions. Our investigation suggested that the MX may function to entrap cholesterol, imposing rigidity to the receptor around the narrowest point of the central pore.

Methods and software

Maps and specimen preparations were carried out by Dr Nigel Unwin (LMB, Cambridge).
The methods used as described here briefly:

Tissue from *Torpedo* ray was homogenised to release NAChR containing vesicles. Vesicles were purified by centrifugation and converted into tubes by incubation with a buffer as in Kubalek et al., 1987 [26]. Specimens were applied to holey carbon supports and imaging was conducted with a FEI Titan Krios. Micrographs were drift corrected with MotionCorr2. All subsequent map processing was carried out with RELION on maps with the appropriate (-16,6) helical symmetry. The FSC between the two halfmaps was carried out using the EMAN2 package.

Calculation of atomic models

Initial alignments

An initial alignment was inherited from Dr Joseph Newcombe who used the Human **α**4**β**2 NAChR structure [2] (PDB ID: 6CNJ) sequence as the primary alignment to model the *Torpedo* N-terminal domain and Transmembrane domain (excluding the TM4 and MA helices) [28]. The mouse 5-HT3 receptor structure (PDB ID: 6BE1) [5] sequence was aligned against the *Torpedo* TM4 and MA helices. The *Torpedo* structure [3] (PDB ID: 2BG9) sequence was aligned against the *Torpedo* C-terminal domain. This alignment was updated to generate an initial alignment for homology modelling. The sequence of the mouse 5-HT3 receptor structure was replaced with a more recent structure [6] (PDB ID: 6HIQ), this new structure sequence was also aligned against the *Torpedo* MX helix as a secondary template. Sequences were aligned using the Clustal Omega algorithm [29] accessed via UniProt (uniprot.org/align). A small error in the *Torpedo* sequence to be modelled was fixed manually within the alignment. This final alignment was used as an input for homology modelling.

Homology modelling

Homology modelling was carried out using the MODELLER v9.23 [30] software. Initially 200 models were generated and scored with the built-in DOPE score [31]. The DOPE score uses an atomic distance based statistical potential to predict the global free energy of a protein model based on a comparison with a training set of native structures. The top 10 structures as assessed by DOPE were further assessed using the QMEANBrane web server (https://swissmodel.expasy.org/qmean/). The QMEANBrane is a local score derived from multiple statistical potentials trained solely on native transmembrane models and gives a per-residue score for the quality of the protein model [32]. The top scoring model as assessed by DOPE and QMEANBrane was taken forward for further analysis.

Model refinement into the cryo-EM map

The model generated from MODELLER was initially rigidly fit into the density map using the *'fit-in-map'* function implemented in Chimera [33]. This was used as an initial input for the flexible fitting software Flex-EM [34]. A previously published hierarchical fitting procedure was used to prevent overfitting to noise in the map [35]. This involved using progressively smaller rigid bodies as input. This had the advantage of maintaining the internal geometry of rigid bodies whilst fitting them into the map relative to each other. Initially the rigid bodies were set as the individual subunits of the model. The fitting runs were set up for four iterations of MD or until convergence was reached (i.e. further iterations produced no increase in fit), the '*cap_shift*' parameter (maximal atom displacement) was set to 0.39 Å. Following this rigid bodies were then defined as individual secondary structure elements (SSEs) as determined by RIBFIND [36]. Flexible fitting was conducted as before, however a '*cap_shift*' of 0.15 Å was applied. At each round of refinement the global fit to the map was assessed using the cross correlation coefficient (CCC), an indicator of the global fit of the model to the map. To assess the local quality of the models the Segment-based meranders coefficient (SMOC) was used [35]. The SMOC score is related to the CCC however is calculated over a sliding window of 9 residues, by this the local fit of segments within the model are assessed. Additionally, the QMEANBrane score was used to assess local quality of the model independently of the fit to the map.

Bioinformatics

In order to further understand the correct conformations of the MX, TM4 and MA helices bioinformatic analyses were conducted. All sequences of the *Torpedo* and Human NAChR subunits along with the mouse 5-HT3A receptor were extracted from either the Swissprot [37] or TrEMBL databases [38] (accessed via: uniprot.org). Alignments of all the subunits were carried out using the Clustal Omega [29] multiple sequence alignment software via the UniProt interface (uniprot.org/align). To predict the secondary structure of the *Torpedo*, TM4 and MA helices domain the JPRED 4 algorithm that uses a neural network trained to predict secondary structure [39].

An updated homology model

During the course of this investigation (15/4/2020) a 2.69 Å structure of the *Tetronarce californica* NAChR bound to **α**-bungarotoxin was released [40]. This structure provided a much more suitable starting point for model building. Due to this a new homology model was built. Initially each subunit from the *Tetronarce californica* structure was aligned against the *Torpedo* sequence using the Clustal Omega algorithm [29] (accessed via: uniprot.org/align). This sequence alignment and the *Tetronarce californica* structural model were used as input for homology modelling with MODELLERv 9.23 [30]. Homology modelling was carried out as before, with 200 models generated and ranked by the DOPE score [31]. The top ten scoring models were then ranked using the QMEANBrane software [32]. Since the *Tetronarce californica* structure was in a closed conformation distinct from the resting state conformation, flexible fitting with Flex-EM [34] was used to improve the fit of the model with the map. Two methods of refinement were performed, one used only the MX helix and the rest of the subunit as distinct rigid bodies, the second used all individual SSEs as defined by RIBFIND [36] as rigid bodies. As before, the CCC was used to assess the global fit of the model to map with the local fit assessed by the SMOC score [35], the local quality of the model was assessed using QMEANBrane [32]. Final model geometry was assessed using the MolProbity suite [41]. The MolProbity suite scores model quality based on geometric parameters such as preferred rotamers, bond geometry (i.e. length, angle, dihedral), and Ramachandran angles of residues.

Results

Map resolution estimates and features

Map acquisition and single particle reconstruction of the *Torpedo* NAChRs were conducted by Dr Nigel Unwin. Micrographs of tubes of NAChRs in a cholinergic membrane were seen to pack tightly and resemble the structural organisation seen in native membranes [15]. A single helical family (-16,6) was assessed in this investigation. Receptors were seen to form a basic asymmetric unit of two receptors packed together along *α*-subunits, with a pair of MX helices orientated parallel to one another (Figure 2A).

In order to assess the map quality, it is common to use the FSC between two half-maps. The FSC between two half-maps was seen to be 6.6 Å, at a cut-off of 0.143 (Figure 2B). This indicated that reliable information was present in the map regarding the position of helices, sheets and domains within the map.

This correlated well with features visible within the map, including transmembrane helices and N-terminal domain beta sheets. Most importantly the MX helices are visible within the

map (Figure 2C). However, whilst the location of helices can be seen clearly, the register of these helices was unclear.

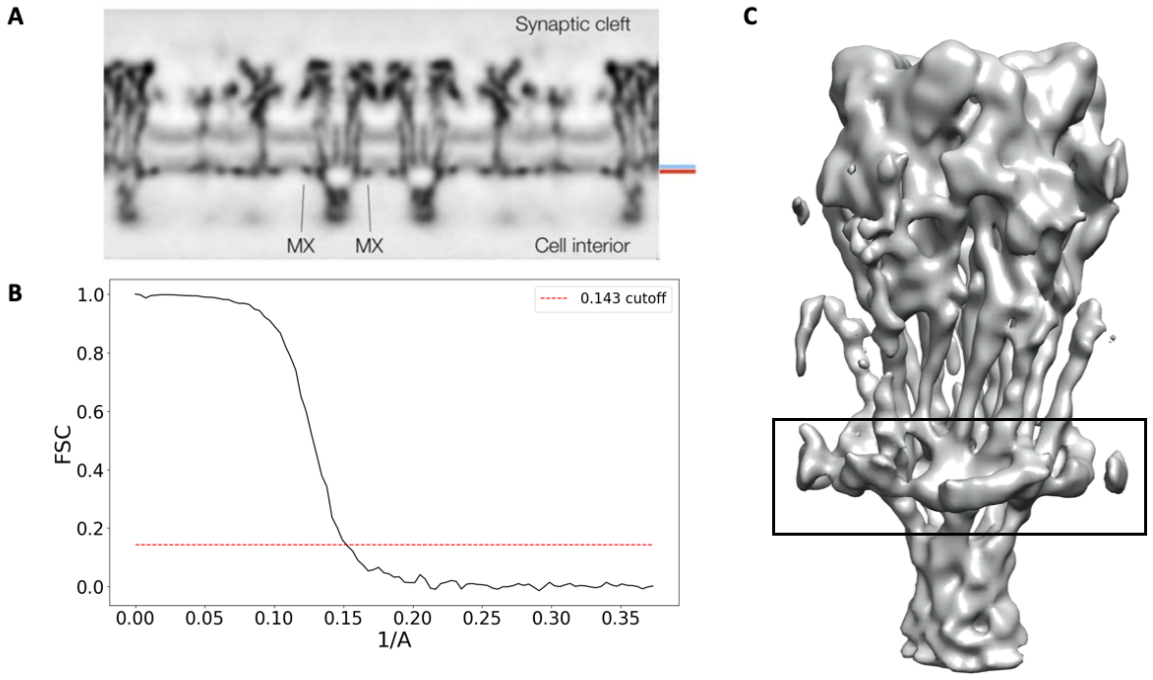


Figure 2. *(**A**). A cross section of the membrane densities after reconstruction by averaging densities belonging to a single helical family. The basic asymmetric unit of two receptors can be seen. Density peaks corresponding to the MX helix are indicated in the image. (**B**). The FSC curve of the agreement between two half maps used in the 3D-reconstruction. The dashed line indicates a cut-off value used to estimate the resolution of 0.143. (**C**). The 3D reconstruction of the density corresponding to a single receptor. The MX helix density is highlighted (black box). Figure adapted from [42] with permissions.*

The MX helices were seen to create an interrupted ring around the TM domain. A cross section of the map showed that the MX helix was situated at the intracellular side of the membrane in-line with regions of density attributed to the phospholipid head groups of the membrane (Figure 2A). The membrane itself can be seen in the map presented as a ~12 Å band of uniform density with more intense bands of density lining the top and bottom, most likely from the phospholipid head groups. In regions above the MX helices this intense density of the phospholipid head group is absent or weakened on the intracellular side of the membrane, indicating that the phospholipid head groups are somehow excluded from this region. This exclusion may be explained by the insertion of the MX helix into the intracellular side of the membrane, which indicated a steric method of exclusion at these regions where the polar head groups of the long acyl chain fatty acids of the membrane would be displaced by the presence of the MX helix.

Building an atomic model of the NAChR at the cholinergic membrane

At the start of this investigation there was no reported structure for the full length *Torpedo* NAChR. Thus, the first step was to calculate an initial atomic model to be refined into the map using homology modelling.

Template selection

To generate an initial model for refinement, homology modelling was carried out using MODELLER v9.23 [30]. It was first necessary to identify template structures that share a high degree of homology with the *Torpedo* receptor. Since the NAChRs belong to the Cys-loop receptor family, structures of these receptors deposited in the PDB were investigated to identify suitable template structures.

There are two structures of the *Torpedo* NAChR currently deposited in the PDB both determined by Cryo-EM to a resolution of 4 Å. The first is the TM domain of the receptor (PDB ID: 1OED) [4]. This receptor was used as the basis for the model of the second model (PDB ID: 2BG9) [3] which included the N-terminal domain, TM domain, and C-terminal MA helix; however no structure corresponding to the MX helix was reported. In theory these two structures would be a great starting place to build an initial model. However, multiple papers and the authors of the original 1OED structure have acknowledged that there is a potential TM register error within this structure between the M1 and M2 helices and associated loop region. This error was carried over to the TM region of the 2BG9 receptor as it was used as a basis for modelling the TM region [28].

Two high resolution X-ray crystallography structures of the Glutamate gated chloride channels from *C. Elegans* have been deposited at resolutions of 3.8 Å (PDB ID: 3RIA) [8] and 3.2 Å (PDB ID: 4TNW) [7]. Whilst these receptors are at a high resolution, they were determined outside of the native membrane environment and neither contained structural information regarding the MX helix. Since the aim of this study was to characterise the interactions between the MX helix and its lipid environment, the suitability of this structure as a template structure was in doubt.

Three structures of the 5-HT3 receptor have been deposited in the PDB. The first was determined to a resolution of 3.5 Å by X-ray crystallography (PDB ID: 4PIR) [43], and was determined outside of the native membrane, however, it included structural data regarding the MX helix. The second structure was determined by cryo-EM to a resolution of 4.31 Å (PDB ID: 6BE1) [5] and contained structural information regarding the MX helix.

The third structure was also determined by cryo-EM to a higher resolution of 3.2 Å (PDB ID: 6HIQ) [6] and also contained structural information regarding the MX helix, and thus was a good candidate for a template model.

Three structures of the Glycine receptor in various states have been deposited in the PDB, determined by cryo-EM to a resolution of 3.9 Å (PDB ID: 3JAE, 3JAD) and 3.8 Å (PDB ID:

3JAF) [9]. However, the proteins were solubilized in detergent and hence their structure may not be representative of a membrane bound receptor. Additionally, neither structure contained information regarding the MX helix. A further point is that the channel is selective for anions, oppositely charged to the ions that enter the NAChR. One hypothesis in this investigation was that the MX helix may play a role in ion selectivity by controlling the width of the pore, and thus utilising a structure from a receptor with such a different ions selectivity was not an appropriate choice. For the same reasons the multiple structures of $GABA_A$ Receptors [10, 11] were excluded from the analysis.

Two structures of the human **α**4**β**2 NAChR have been deposited in the PDB differing in stoichiometry. Both were determined by cryo-EM to resolutions of 3.9 Å (PDB ID: 6CNK) and 3.7Å (PDB ID: 6CNJ) [2]. Both these structures contained information regarding the MX helices and represented potential viable candidates from homology modelling.

Table 1. Sequence identities of cys-loop receptor family subunits against Torpedo subunits.

Torpedo NAChR subunit	Mouse 5-HT3-A	Human NAChR **α**4	Human NAChR **β**2	Glutamate	Glycine	GABA **α**1	GABA **β**2	GABA **γ**2
alpha	24.65	52.07	47.88	19.29	21.88	19.91	19.41	21.14
beta	27.33	44.91	46.98	19.21	20.76	18.14	20.45	19.49
delta	25.12	39.64	44.58	19.39	22.46	20.05	17.87	19.35
gamma	23.34	40.16	44.99	19.69	22.70	17.91	18.96	19.22

The sequence identities of the most relevant structures from each category of Cys-loop receptor showed that the human **α**4 subunit had the closest sequence identity to the *Torpedo* **α**-subunit, with the human ***β***2 subunit showing the highest sequence identity to the *Torpedo* ***β***-,**δ**- and **γ**-subunits (Table 1). Due to this the Human **α**4***β***2 NAChR structure, 6CNJ was used to model the N-terminal and TM regions of the receptor including the MX region. The TM4 and MA helices were modelled using the 5-HT3-A receptor , 6HIQ, whilst the C-terminus was modelled using the *Torpedo* structure 2BG9.

Initial alignments

An initial alignment for homology modelling was inherited from Dr. Joseph Newcombe, a former PhD student in the lab (Figure A13) [28]. The alignment was created using Clustal Omega multiple sequence alignment program [29] using the Human **α**4***β***2 NAChR structure (PDB ID: 6CNJ) as the primary template for modelling, the Mouse 5-HT3-A receptor (PDB ID: 6BE1) to model the TM4 and MA helices, and the *Torpedo* structure (PDB ID: 2BG9) to model the receptor C-terminal region. However, it was deemed pertinent to use the high

resolution Cryo-EM structure of the 5-HT3-A receptor (PDB ID: 6HIQ) to model the MA and TM4 helices. Therefore, the relevant sections of the alignment were updated to include the higher resolution structure.

Before modelling, a small error in the inherited alignment had to be corrected. It was found that there was a two-residue deletion in the target sequence at a region that corresponded to the delta subunit TM4 helix (Figure 3A). These residues were inserted into the alignment and the alignment corrected using Clustal Omega and manual adjustment (Figure 3B). Homology modelling was carried out using MODELLER v9.23 and top models were identified using the built-in Modeller DOPE score [31]. Following initial fitting of these models to the map, it became apparent that the architecture of the MX helices using the X-ray structure 6CNJ as a primary template for this region resulted in an MX helix conformation that did not reflect what was seen in the cryo-EM map generated with receptors in their native membrane (Figure 4A). However, the Cryo-EM structure of the mouse 5-HT3-A receptor seemed closer to what was expected for the MX helix.

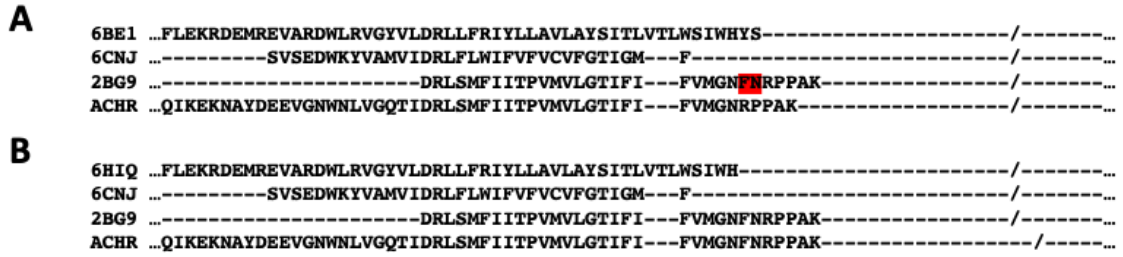


Figure 3. *A section of the initial TM4 helix of the δ-subunit from the inherited alignment. (**A**). Shows the deletion of two residues (Red) from the Torpedo sequence to be modelled (ACHR). (**B**). Shows the same section of the fixed alignment. 6HIQ and 6BE1: 5HT-3 receptor, 6CNJ: Human α4β2 NAChR, 2BG9 : Torpedo NAChR, ACHR: sequence to be modelled.*

Due to this, the mouse 5-HT3-A receptor part of the alignment was expanded to include the 5-HT3-A receptor MX helices as a secondary template for this region (Figure 4B). Once again homology models were created, and the top structures identified using the DOPE score.

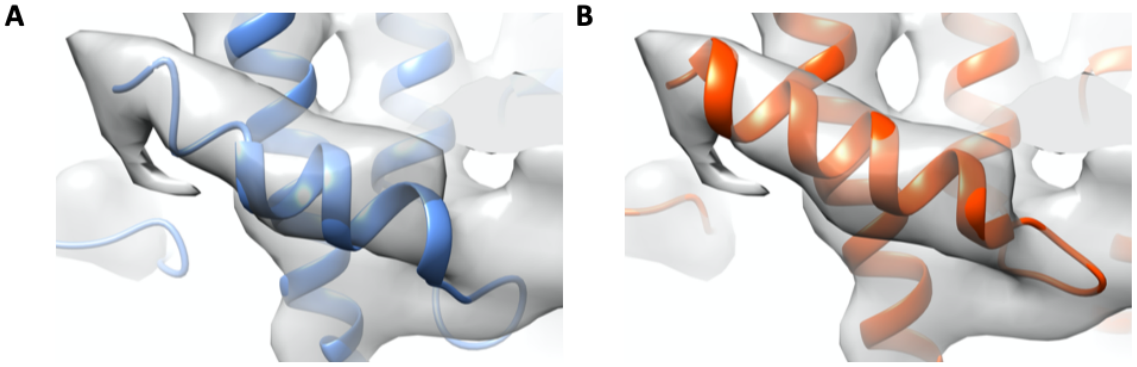


Figure 4. *The MX helix of homology models created using the Human* $\alpha 4\beta 2$ *NAChR (6CNJ) as a primary template for the MX helix (**A**) or both the Human* $\alpha 4\beta 2$ *NAChR (6CNJ) and the mouse 5-HT3 receptor as templates for this region (**B**).*

Generating homology models

200 initial models were generated and DOPE scores between them were fairly similar, ranging from the best scoring -238794.2 to the worst scoring model -233190.2. The DOPE score is a relatively good and generally accepted statistical potential score for identifying a good model from a bad model. However, the score was configured on monomers and a test set which was not directly designed to be used for membrane proteins [31]. Due to this the top ten models identified by the DOPE scores were subjected to analysis using QMEANBrane [32], a scoring function specifically designed to assess the local model quality of membrane protein. It was seen that the QMEANBrane scores generally correlated well with the DOPE score, with the exception of models 176 and 151, which were ranked higher by QMEANBrane than their respective DOPE scores (Table 2). The top scoring model by DOPE and QMEANBrane was taken forward for further analysis.

Table 2. DOPE and QMEAN Brane scores of the top 10 homology models generated.

Model	DOPE	QMEAN Brane
069	-238794.20312	0.714
065	-238686.90625	0.709
154	-238633.92188	0.711
192	-238553.34375	0.711
176	-238531.32812	0.714
151	-238480.51562	0.714
122	-238448.04688	0.712
149	-238444.01562	0.712
124	-238390.76562	0.710
075	-238312.67188	0.710

The QMEANBrane score is a local score that assigns a score per residue [32]. The distribution of the QMEAN Brane score for the best initial homology model (Figure 5) showed that the worst scoring regions of the molecule were the ends of the C-terminal MA helix, the N-terminal ends of the TM4 helix and loop regions located in the N-terminal

domain. The MX helices scored relatively well, with the worst sections being the flanks of the helices. This is most likely due to the MX helices not being in the correct orientation within the membrane. From the density map it appears that the MX helices lie flat against the intracellular side of the membrane, and may protrude slightly into the membrane (Figure 2). In the homology model the MX helices are tilted (Figure 4) somewhat similar to the orientations seen in receptors purified in detergent that are most likely incorrect (Figure 1).

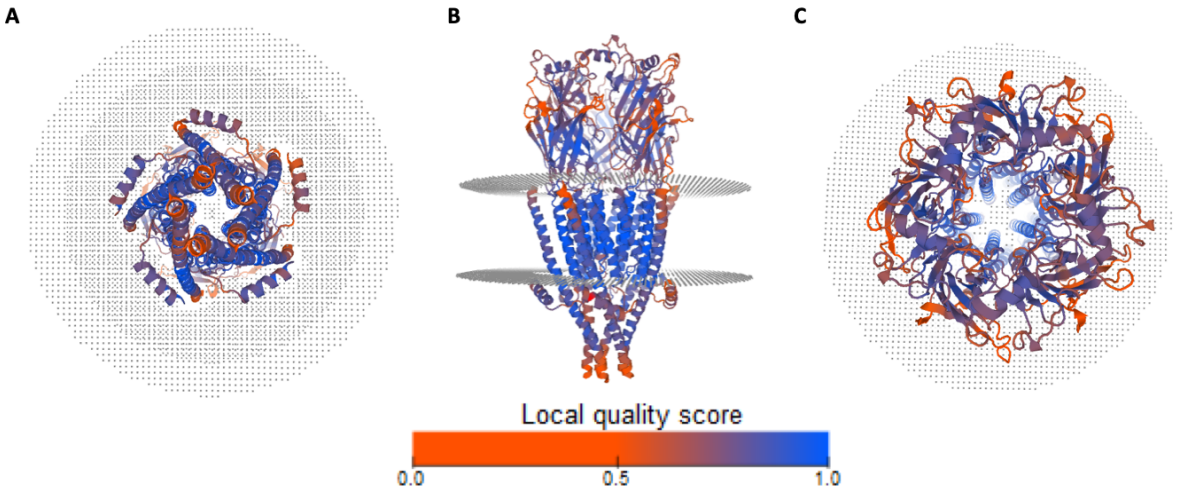
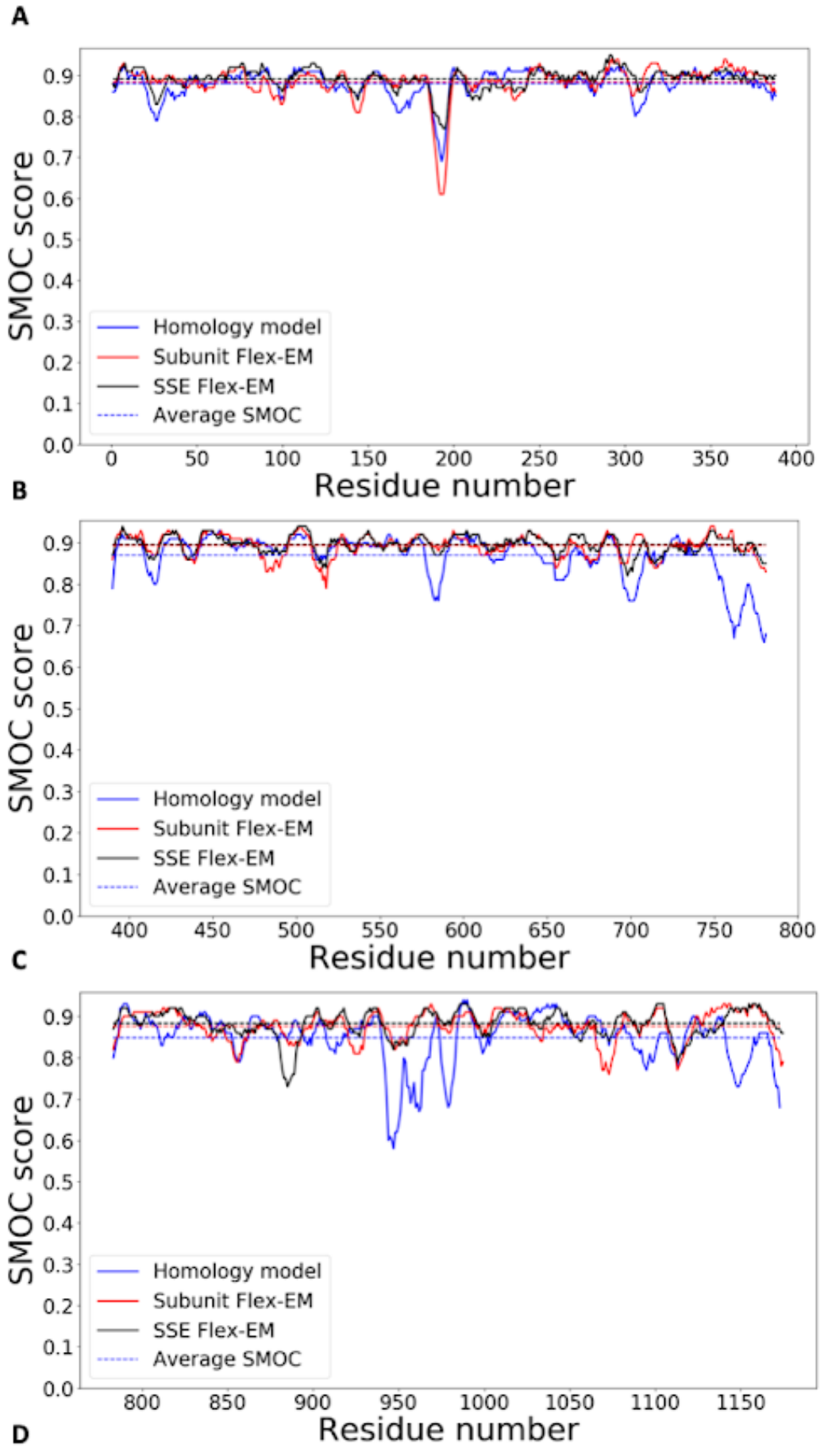


Figure 5. *The per-residue distribution of the QMEANBrane scores for the best ranked homology model. The model can be seen as a bottom up view from the intracellular side (**A**), side on view through the membrane (**B**), and from a top down view from the extracellular side (**C**). The legend shows the corresponding colour key from scores in the range 0.0 - 1.0, where 1.0 indicates a good score. The cellular membrane is represented by the grey dotted disks.*

Flexible refinement with flex-EM

The initial model generated with MODELLER was rigidly fitted into the cryo-EM map using the *fit-in-map* tool implemented in Chimera [33]. It was seen to fit the map well with a CCC of 0.875. Once the rigid fit had been identified the model was further refined into the EM map using a hierarchical refinement protocol [35] and the flexible fitting software Flex-EM [34]. This protocol uses progressively smaller rigid bodies as an input for flexible fitting and has been shown to improve the fit of models to intermediate resolution cryo-EM maps, whilst preventing overfitting to noise.

Initially, individual subunits were used as rigid bodies followed by a second round of fitting using individual SSE's. At each round of refinement, the global CCC of the model to the map was increased from 0.875 for the homology model, to 0.896 using individual subunits and finally, 0.897 using SSEs. The final increase in CCC at first glance appears to be a negligible increase. However, this highlights the need to utilise a local scoring function that assesses


A
SMOC score
Residue number
Homology model
Subunit Flex-EM
SSE Flex-EM
Average SMOC
B
SMOC score
Residue number
Homology model
Subunit Flex-EM
SSE Flex-EM
Average SMOC
C
SMOC score
Residue number
Homology model
Subunit Flex-EM
SSE Flex-EM
Average SMOC
D

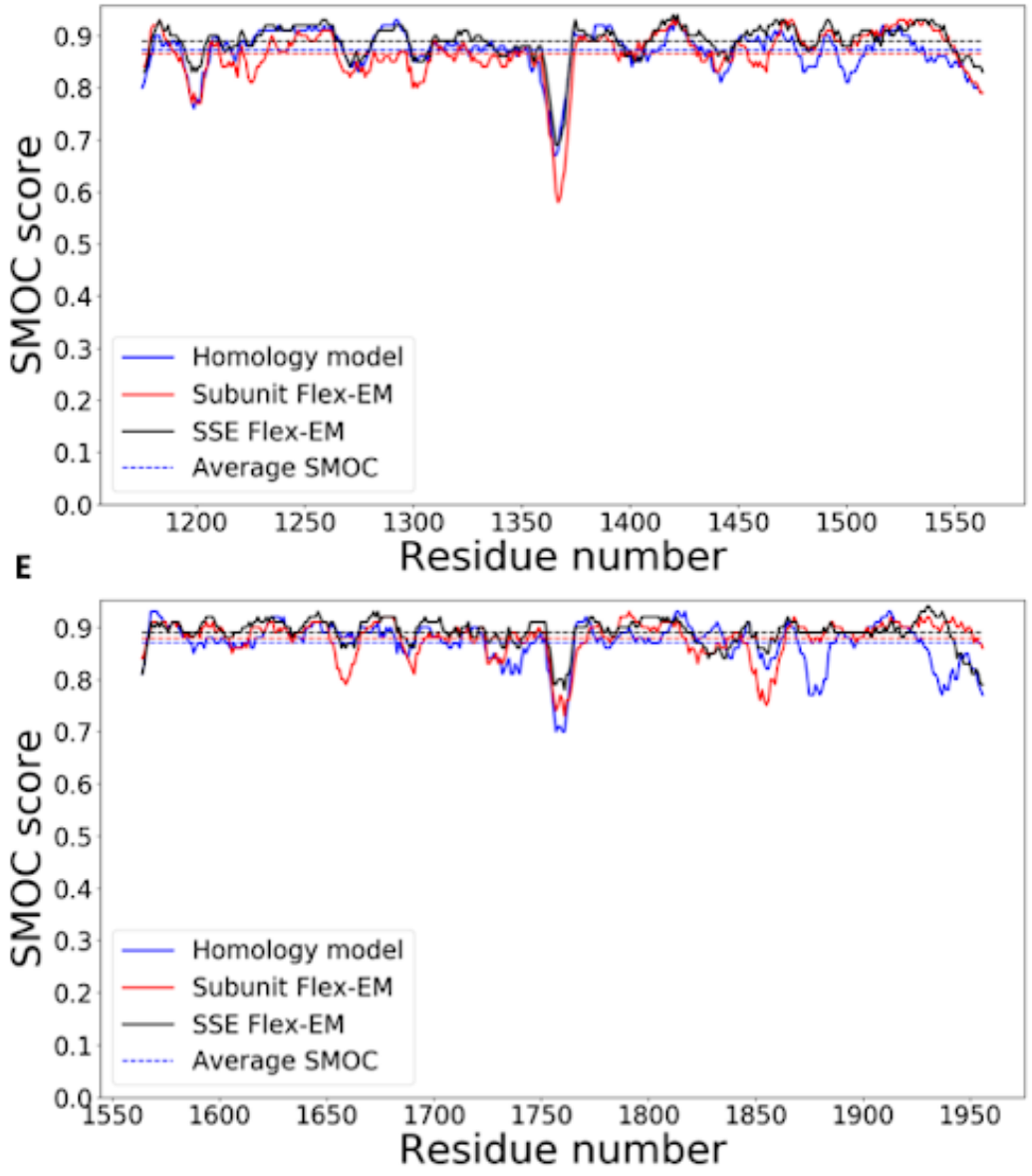


Figure 6. *SMOC plots for each subunit (**α** : **A**, **β** : **B**, **δ : C**, **α** : **D**, **γ : E**) describing the local agreement with the density map, at the various stages of the flexible refinement protocol. The local distribution of the SMOC score over a sliding window of 7 residues are shown with the solid lines for the initial homology model (blue), after flexible fitting using the individual subunits as rigid bodies (red), and after flexible fitting using SSEs as rigid bodies (black).*

the fit of individual regions of the model (i.e. SSEs or individual residues). At each round of refinement the SMOC score [35] was employed to assess the local fit to the map. The average SMOC score for individual subunits was seen to increase at each iteration with the exception of subunit D that was seen to decrease during the fitting of domains before increasing in the final SSE refinement (Figure 6). This was most likely due to the initial homology model having a different pore size and needing flexible refinement to fit accurately.

With a map at resolution of 6.6 Å it was expected that the smallest features present on the map would be helices and sheets and indeed, these features were clearly seen in the map

(Figure 2C). Any further fitting would require flexibly fitting individual residues to the map and there would be a real chance of overfitting residues to noise within the map. From the FSC curve it was seen that a significant amount of noise was present especially at higher spatial frequencies, therefore no further refinement was conducted.

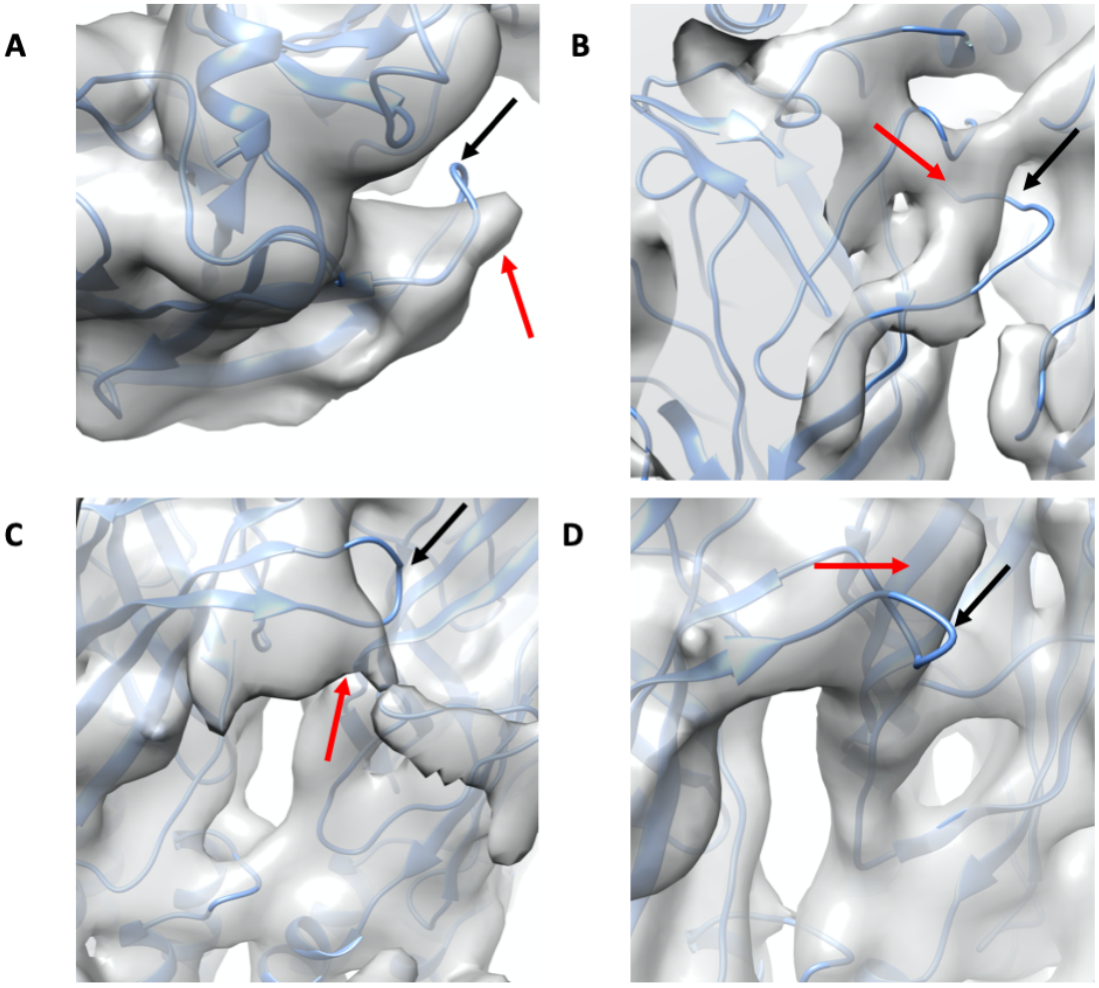


Figure 7. *Regions of poor fit in the N-terminal domain after flexible fitting with Flex-EM.* ***A.*** *The* ***α****-subunit (chain A) C-loop.* ***B.*** *The* ***δ****-subunit (chain C) NTD-loop* ***C.*** *The* ***α****-subunit (chain D).* ***D.*** *The* ***γ****-subunit (chain A) C-loop. Black arrows indicate areas where the model has a poor fit in these regions, red arrows indicate the density corresponding to poorly resolved regions.*

From the plots it was seen that in several areas within subunits (A 185-199, C 874-894, D 1359- 1374, E 1750-1767) where the SMOC score drops notably from the average (Figure 6). These regions were all found to be loop regions within the N-terminal domain present at regions of the map where the density was rather featureless (Figure 7). This could be the case due to inherent flexibility at these regions, meaning that the local resolution of the map at these regions were at lower than the surrounding regions. A second area that required further investigation was the C-terminal domain helices MA and TM4 regions (Figure 8). Visual

inspection, SMOC scores and a QMEANBrane analysis highlighted that these regions had a poor fit with the map. This could be a result of the difference in pore diameter between the map and initial model and thus the MA/TM4 helices were not optimised for a fit into a map with a smaller pore. This problem could be solved using flexible fitting of these helices with individual residue restraints, however this would most likely result in overfitting of the model to the map. Additionally, errors in this region could come from assigning an incorrect helix register within the map. It is clear from a visual inspection of the map that the helix register is ambiguous. This same problem was seen with the fit of the MX helices to the map; the helices within the model had a more helical character than that previously seen for the human $\alpha 4\beta 2$ NAChR structure and much more resembled that of the mouse 5-HT3 receptor (Figure 1). However, these regions failed to move into the density during flexible fitting. One reason for this could be steric clashes between sidechains of residues. Since more flexible fitting would have most likely resulted in overfitting of the model to the map, a bioinformatics approach was taken to investigate the fit of the MA/TM4 helices along with the expected position of the MX helices.

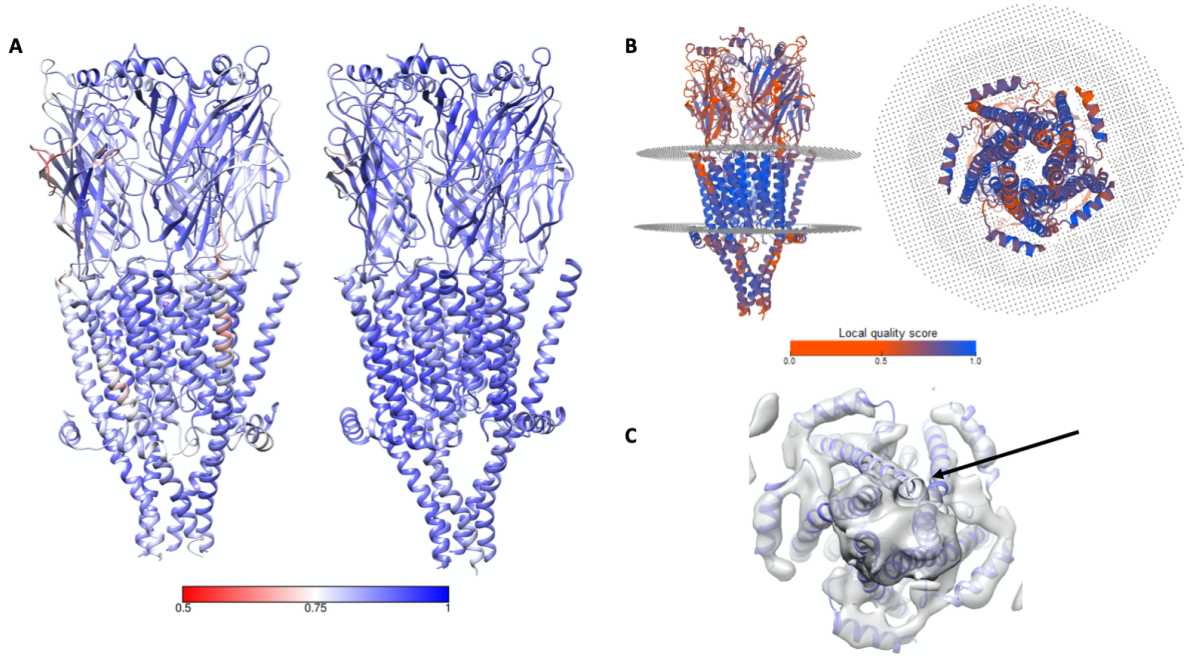


Figure 8. *SMOC and QMEANBrane analysis of the model after flexible fitting with Flex-EM.* ***A.*** *Models of the Torpedo NAChR before flexible fitting (left) and after (right). A legend is shown underneath showing the correspondence between model colour and SMOC score.* ***B.*** *The QMEAN Brane analysis after flexible fitting from a side view (left) and a bottom up view from the intracellular side of the membrane (right). The membrane is represented by the grey dashed disks. A legend showing the correspondence between QMEANBrane score and model colour is shown underneath.* ***C.*** *A bottom up view from the intracellular side of the model fitted into the map following flexible fitting into the map (grey). The black arrow indicates an area of problematic fit between map and model.*

Bioinformatic analysis

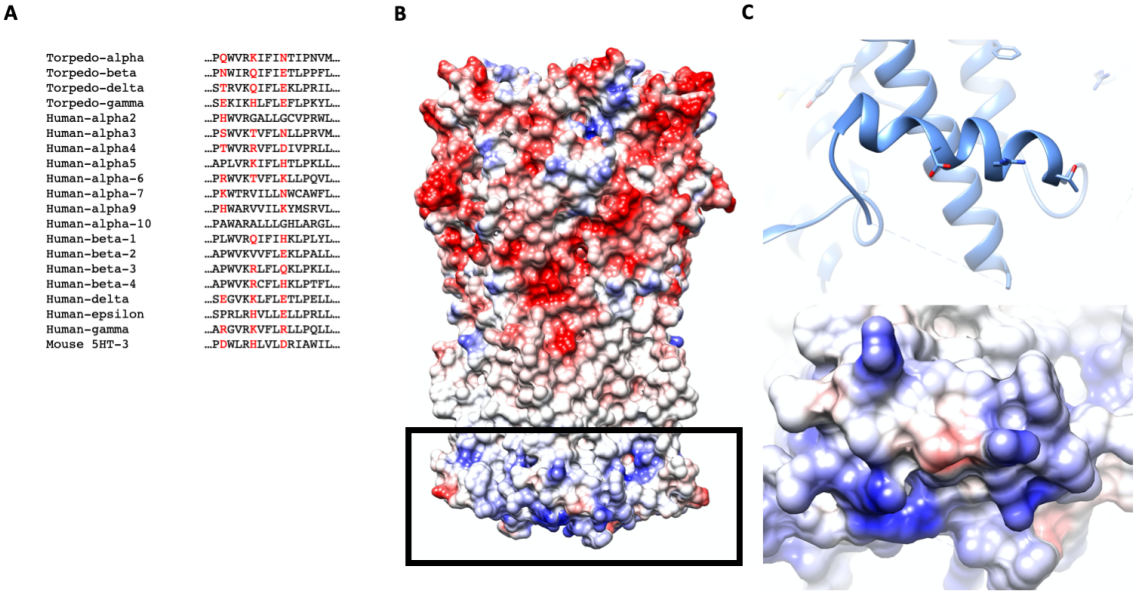


Figure 9. A. *A section of an alignment between the Torpedo NAChR, Human NAChR and mouse NAChR subunit MX helix sequences. The four residues highlighted red correspond to the four residue repeating pattern of polar residues.* ***B.*** *A surface representation of the Human* ***α****4****β****2 NAChR colored by coulombic potential. The MX helix regions are highlighted (black box).* ***C.*** *Show a zoomed in section of the human* ***α****4****β****2 NAChR* ***α****-subunit MX helix. The charged residues that align to the outer (membrane facing) region of the MX helix are seen in the model (top) with the corresponding Van Der Waals surface and coulombic potential shown below.*

In order to identify the correct orientation of the MX helix, a multiple sequence alignment was created using the Clustal Omega software [29], and conserved residues assessed. The closest functional and sequence homolog used to generate the initial model was the Human **α**4***β***2 NAChR. Therefore, all human NAChR sequences available in the Swiss-Prot database [37] were aligned against the different *Torpedo* subunit sequences. Since the 5-HT3 receptor was also used to generate the homology model this sequence was also included in the alignment. Immediately evident was a three residue repeating pattern of charged residues found in all but the Human **α**2-, **α**5-, **α**10-, **β**1-, **β**2-,**β**3-, **β**4-, and **ε**- NAChR subunits (Figure 9A). Where this occurs, the charged residues typically flank three hydrophobic residues, meaning the basic unit of the pattern (e.g. Charged, hydrophobic, hydrophobic, hydrophobic) corresponds to approximately one helical turn. Inspection of the Human **α**4***β***2 NAChR structure indicated that these residues align to the outside of the MX helix (i.e. facing away from the channel) (Figure 9B). This has the effect of creating a hydrophobic pocket between the TM region and the MX helix. Additionally having the charged residues aligned facing the membrane puts this region of the helix in a perfect place to interact with either the surrounding charged phospholipid head groups or the polar moieties on cholesterol.

Additionally, ths charge may prevent the MX helix from penetrating deeper into the hydrophobic region of the membrane.

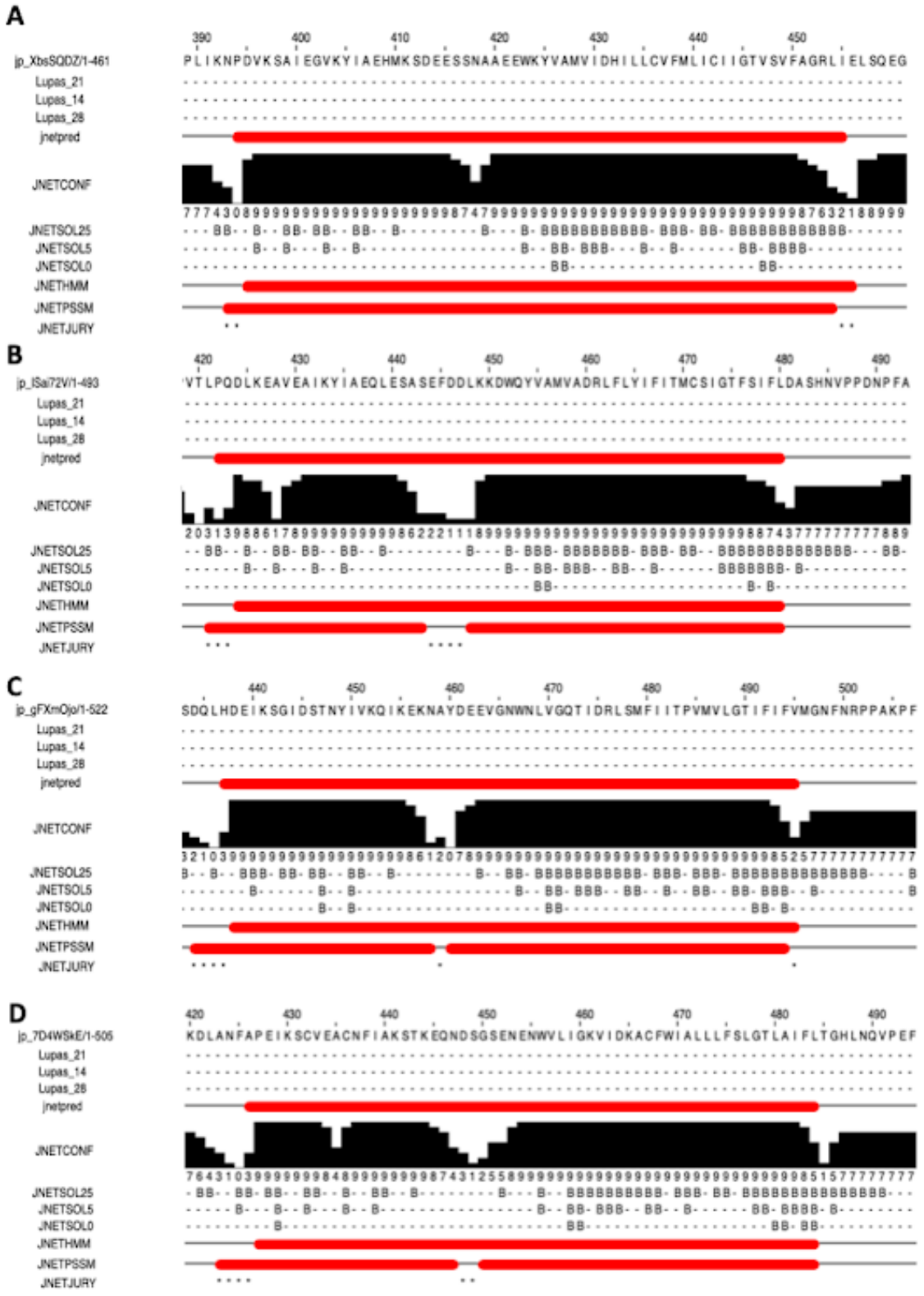


Figure 10. *The results of the JPred prediction of SSEs for the TM4 and MA helix of the Torpedo* ***α****- (A),* ***β****- (B),* ***δ****- (C), and* ***γ****-subunits (**D**). The sequence is shown atop each result.*

JNETCONF is the confidence of the secondary structure prediction on a scale from 1 -10 with 10 being the most confident. The red bars indicate the prediction of helices by the JPRED algorithms.

In the previous model of the *Torpedo* (PDB ID: 2BG9) the TM4/MA helices are modelled as two separate helices with a flexible loop in between [3]. If this is the case, modelling this flexible region would make identifying the register of these helices much easier. To investigate this possibility, the subunit sequences were run through the JPred secondary structure prediction server [39]. The results showed inconsistency within the predictions; for each subunit the secondary structure was predicted to be one continuous helix, although in each subunit the region located approximately at the boundary of the MA/MX helix had notably lower confidence scores for the SSE predictions (Figure 10). A visual inspection of the map at these regions showed no clear signs of any unordered region. However, the resolution of the map was relatively low and the presence of an unordered region connecting the MA and TM4 regions cannot be ruled out.

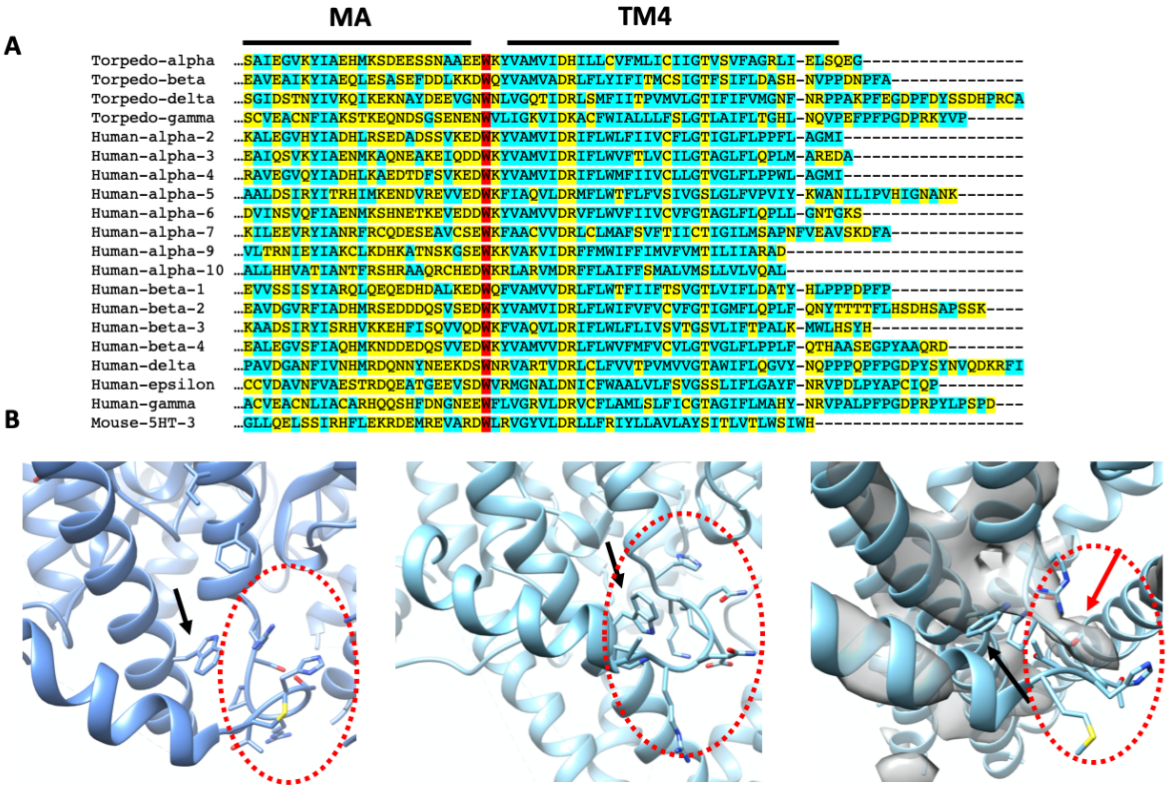


Figure 11. ***A.*** *A section of the alignment between the Torpedo NAChR subunit isoforms, Human NAChR subunit isoforms, and the mouse 5-HT3 receptor MA and TM4 sequences. Polar and charged residues are highlighted yellow and hydrophobic residues are highlighted blue. A tryptophan conserved within all sequences is highlighted in red. Above the alignments the predicted positions of the MA and TM4 helices are shown (black bars).* ***B.*** *The conserved tryptophan in the Human* $\alpha 4\beta 2$ *NAChR (left) and the mouse 5HT-3 receptor (centre) is indicated (black arrow). The loop preceding the MX helix is highlighted (red circle). The final*

model after flexible fitting for the Torpedo NAChR is shown (right) within the density map (grey). The position of the conserved tryptophan (black arrow) and the preceding MX loop (red circle) are highlighted. A conspicuous peak of density within the MX loop is also indicated (red arrow).

Since the MA helix is positioned at the extracellular side of the membrane with the TM4 helix being transmembrane, it was hypothesised that there would be a clear difference between the content of charged and hydrophobic residues within these regions that may give a better idea of the register of the helices. The previous alignment was searched for hydrophobic/charged residues within these regions. From the multiple sequence alignment, it was evident that the MA portion of the sequence had a notably higher charged residue content, before switching to a more hydrophobic region at residues predicted to form the TM4 helices (Figure 11A). This makes sense, considering the position of these structures within the cell. One interesting finding was the presence of a conserved tryptophan residue located approximately at the change from MA to TM4 helix (Figure 11A). In the Human $\alpha 4\beta 2$ NAChR and the mouse 5-HT3 models this residue can be seen pointing towards the loop that connects to the MX helix (Figure 11B), specifically directed towards a single helical turn within this region. Inspecting the map at a higher threshold revealed a conspicuous peak of density within this region approximately located at this helical turn (Figure 11B). It may be the case that this conserved tryptophan residue interacts with either the TM3 Helix or MX loop in order to orientate the MX helix in position. From the results it was hypothesised that the change from mainly hydrophobic to mainly hydrophilic within the TM4/MX helices, along with getting the position of the conserved tryptophan correctly placed, would orientate these regions with the correct register.

A new high resolution homologous structure

Soon after the completion of this investigation a structure of the *Tetronarce californica* NAChR bound to the snake venom inhibitor α-Bungarotoxin at a resolution of 2.69 Å was released (PDB ID: 6UWZ) [40]. The receptors were extracted from their native membrane with detergent before being reconstituted in nanodiscs. So, whilst not being present in their native membrane, the nanodiscs should mimic a mature receptor within the membrane. The 6UWZ model was seen to fit well into our density map with a CCC of 0.861. This model was seen to have a very high sequence homology with the *Torpedo* subunits (Table 3). A new homology model was generated using MODELLER and the *T. californica* as the input structure. Sequence alignments were conducted using Clustal Omega and homology models calculated using MODELLER v9.23 as before.

Table 3. Sequence identities between *Torpedo* subunits and corresponding *T. californica* subunits.

Torpedo Subunit	Sequence identity (%)
α	99.123
β	96.146
δ	97.510
γ	97.628

The initial homology model was seen to fit into the map with a CCC of 0.888, which correlated well with the average SMOC scores of 0.893, 0.891, 0.835, 0.877 and 0.878 for subunits A, B, C, D and E, respectively. Since the map was at a relatively low resolution, and the template model by comparison was at a very high resolution, it was hypothesised that too much flexible fitting would increase the chances of overfitting the model to noise within the map. Due to this, two separate methodologies were taken and compared. One model was fitted to the map using SSEs as rigid bodies, while a second model was fitted to the map with rigid bodies defined as the MX helix and the rest of the protein, with the MX loop as flexible.

From the SMOC plots it was seen that fitting with SSEs as rigid bodies notably increased the local score compared to using only the MX helix or main chain (Table 4, Figure 13).This correlated with the global CCC scores of 0.889 and 0.900 for the MX only and all SSE refinements, respectively. However, in all plots the regions corresponding to the MX helix exhibit a higher SMOC score for the model where only the MX helix is refined than when all SSEs were used for the refinement (Figure 13). This suggested that some or all SSEs may have been overfitted during refinement.

Table 4. Average SMOC score for each subunit at the different stages of refinement

	Chain A	Chain B	Chain C	Chain D	Chain E
Homology model	0.893	0.891	0.835	0.877	0.878
MX refinement	0.893	0.890	0.836	0.878	0.881
SSE refinement	0.899	0.898	0.869	0.889	0.896

This overfitting was evident when the model quality was assessed, the QMEANBrane score for the MX only refined model being 0.762, whilst the all SSE refinement score was notably lower at 0.739. Therefore, the MX-only refined model was considered as the final model.
The geometry of the final model was evaluated using the MolProbity suite of programs. It was seen that the final model contained 0.35 % of residues with geometry deemed as Ramachandran outliers with 97.12 % favoured, and no CA geometry outliers.

Model validation

A comparison with the previous model produced using the alignment against the human α4β2 NAChR, mouse 5-HT3 receptor and *Torpedo* structures, showed an RMSD of 2.822 Å (Figure 12 A). However, the individual subunits showed much smaller RMSD values when compared of 1.120 Å, 1.177 Å, 1.168 Å, 1.109Å, and 1.096 Å for subunits A, B, C, D and E (α, β, δ, α, and γ, respectively), respectively. This indicated that much of the secondary structure elements within the individual subunits were correct. In fact the main differences were seen to be within the MA/TM4 helix, MX helix and MX-loop regions (Figure 12 B). The register of the TM4 region was seen to differ between the subunits along with the end of the MA helix of the C-terminal domain. This was expected as these regions were problematic when modelling the original model. Furthermore, the positions of the MA helices and loops were seen to differ between the two models with the models having a different register and helix length. The helices of the final model based on the *Tetronarce californica* were seen to be one turn shorter than in the initial model. Both models showed similar profiles when the model geometry was checked with MolProbity, the initial model having slightly worse Ramachandran outliers, 0.57 %, and showed no CA geometry outliers. However, the initial model was scored notably worse by QMEANBrane, having a score of 0.702, compared to 0.762 scored by the final model (Figure 12 C). The difference in RMSD across the whole model appeared to be due to the size of the channel formed by the five subunits. The channel formed in the final model was narrower than in the initial model, with the TM4 helix closer to the inner 3 helix ring and notably more kinked. However, since the local QMEANBrane score for the final model was notably higher than in the initial model, as such this was deemed the more correct of the two.

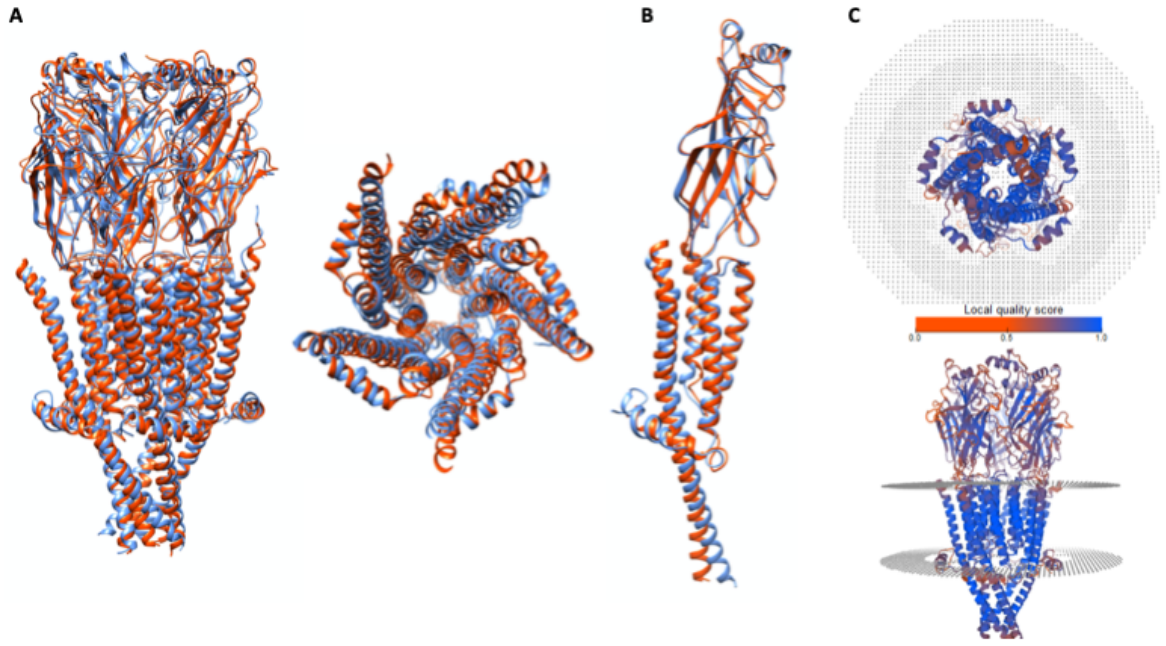
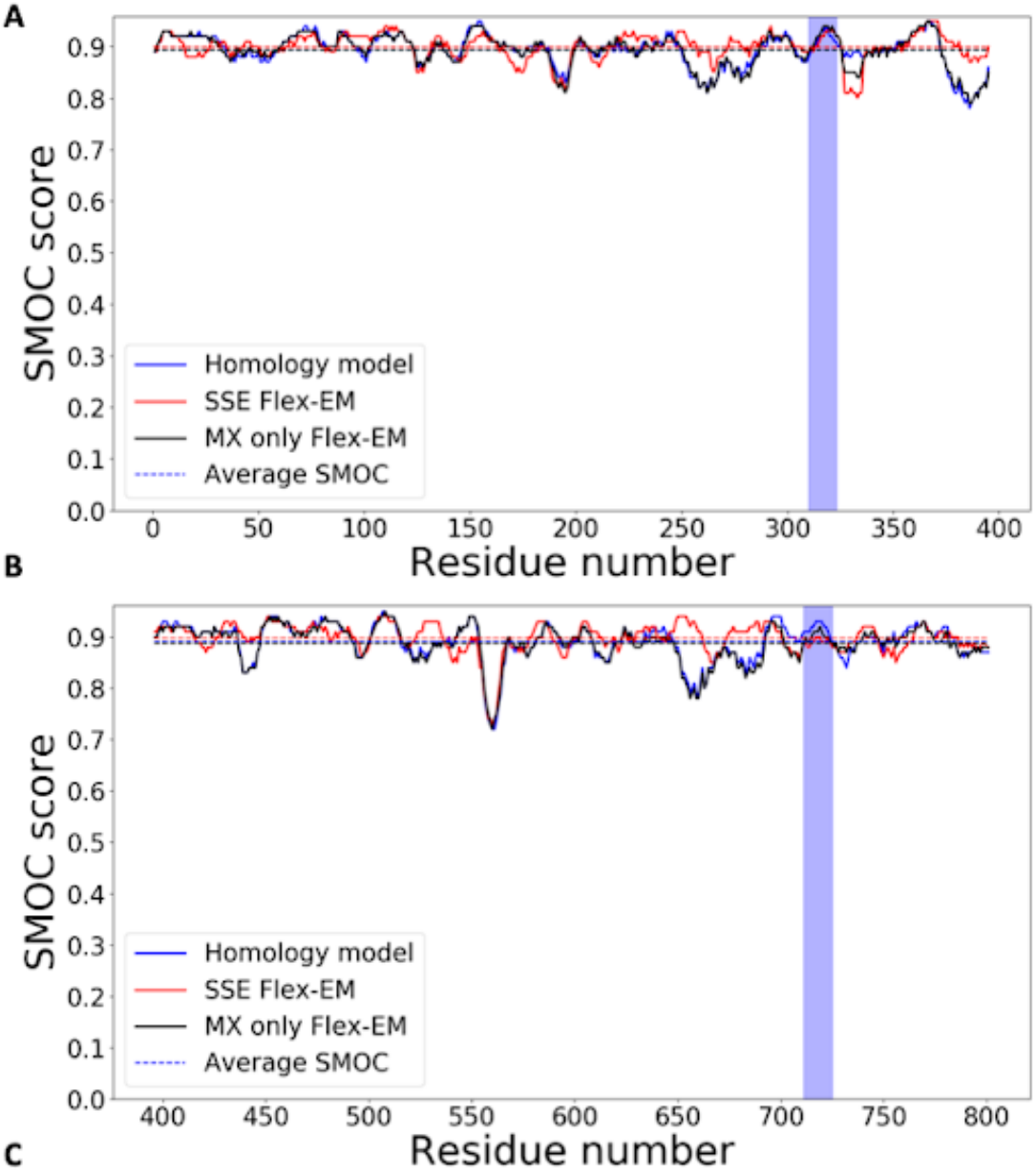
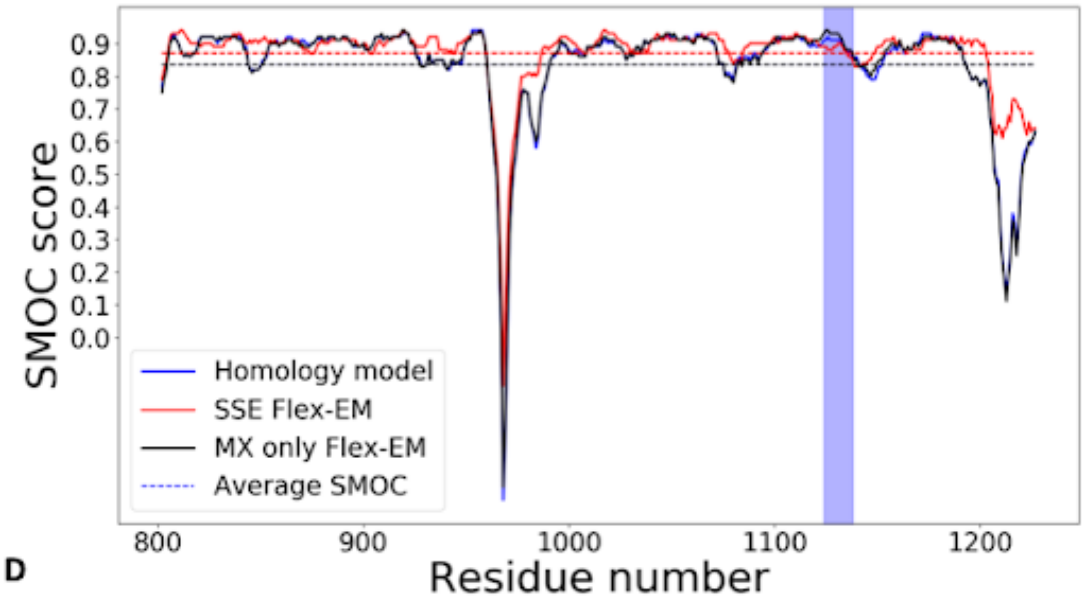


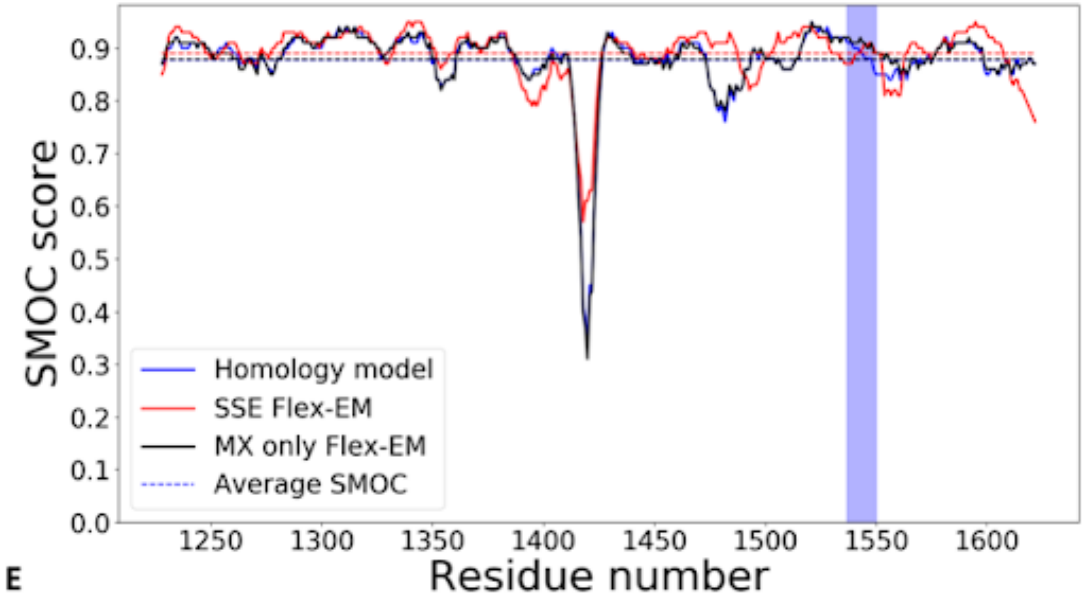
Figure 12. *(A). The agreement between the model derived from homology modelling with homologous structures human α4β2 NAChR, the mouse 5-HT3 receptor and the Torpedo*

NAChR (red) and the model derived from homology modelling with the T. californica structure (blue). The full receptor agreement is shown (left) and a section through the TMD (right). ***(B).*** *The agreement between individual alpha subunits of the receptors.* ***(C).*** *The QMEANBrane analysis of the final model produced from the T. californica homology modelling and refinement. The model is colored by the local score from 0.0 to 1.0. The approximate position of membrane boundaries are indicated by grey disks. Two views are shown one bottom up from the intracellular side (top) and one side view (bottom). A colour key corresponding to the QMENABrane score is also shown.*




SMOC score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Homology model
SSE Flex-EM
MX only Flex-EM
Average SMOC
800
900
1000
1100
1200
Residue number

D


SMOC score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Homology model
SSE Flex-EM
MX only Flex-EM
Average SMOC
1250
1300
1350
1400
1450
1500
1550
1600
Residue number

E

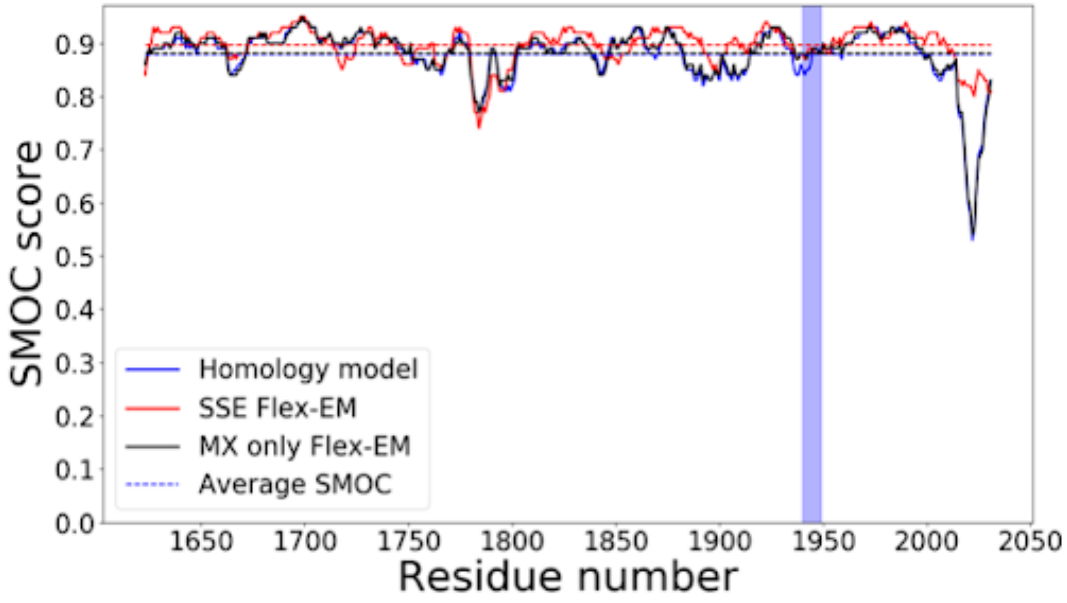

SMOC score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Homology model
SSE Flex-EM
MX only Flex-EM
Average SMOC
1650
1700
1750
1800
1850
1900
1950
2000
2050
Residue number

Figure 13. *SMOC plots of for each subunit (**α** : **A**, **β** : **B**, **δ** : **C**, **α** : **D**, **γ** : **E**) describing the local agreement with the density map at the various stages of the flexible refinement protocol. The local distribution of the SMOC score over a sliding window of 7 residues are shown with the solid lines for the initial homology model (blue), after flexible fitting using the SSEs as rigid bodies (red), and after flexible fitting using the MX helix only (black). Also shown are the positions of the MX helix for each subunit within the plot (purple block).*

Final model features

As predicted by the bioinformatic analysis, the C-terminal located helix MA was seen to exhibit a much more hydrophilic character, whilst the TM spanning helix TM4 was clearly seen to be hydrophobic in nature (Figure 14A). In the final model the TM4/MA helix was seen to be one continuous helix. This correlated with what was observed in the new *T. californica* structure [40]. Additionally, the bioinformatic analysis predicted that the MX helices of the new model to be aligned with the hydrophilic residues facing outwards (i.e. away from the channel), creating a hydrophobic pocket where the MX helix contacts the TM domain. The angle of the MX helix was seen to be in-line with previously reported structures of the MX helix, in the Human **α**4**β**2 structure and the mouse 5HT-3 receptor (Figure 14B) . The conserved tryptophan within the TM4/MA helix was seen to be pointing towards the MX loop in the final model. The single helical turn within this loop corresponds well with the position of a conscious peak of density seen in the map (Figure 14C). Prediction of interactions of the side chain of this tryptophan showed that multiple hydrophobic interactions were predicted with residues off the MX loop (Figure 14C). One interpretation of this is that the hydrophobic pocket formed between the residues of the loop and conserved tryptophan serve as a method of orientating the MX helix in the correct conformation within the membrane.

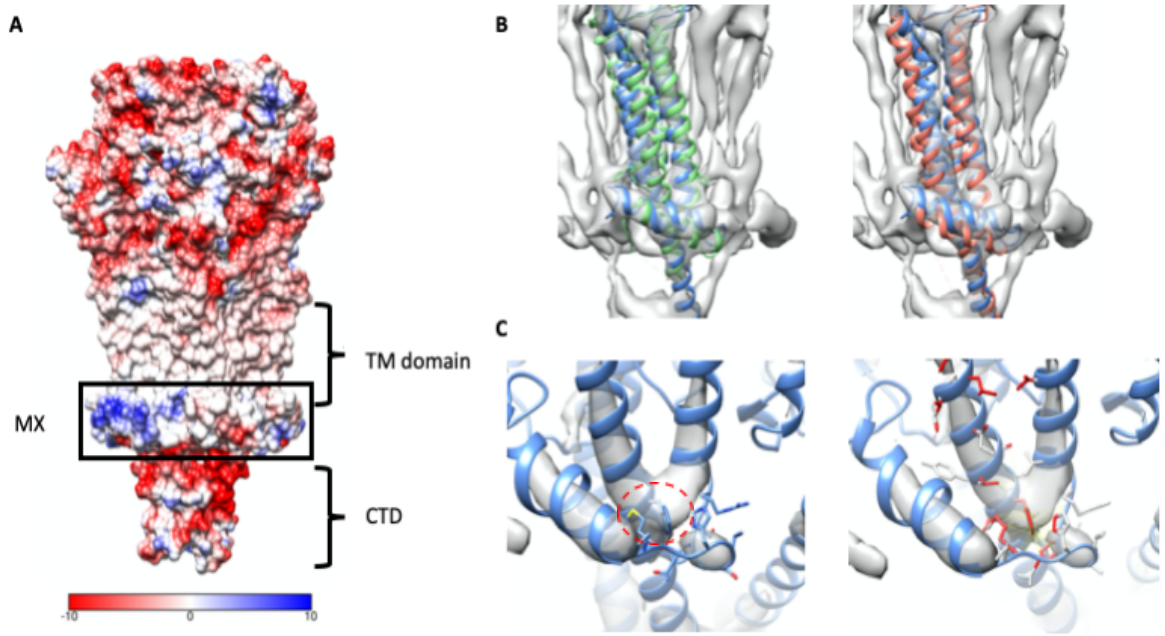


Figure 14. *(**A**) A calculated Van der Waals surface from the final model coloured by coulombic potential. The transmembrane domain (TM domain), MX helix regions (MX) and C-terminal domain (CTD) are indicated. A colour key corresponding to the coulombic potential is shown underneath. (**B**) A comparison of the MX helix angle in the final model (blue) with the human **α**4**β**2 structure (left, green) and the mouse 5HT-3 receptor (right , red). The density map is also shown (grey). (**C**) The conserved tryptophan (red dashed circle) in the final model is shown along with the MX helix loop (right). The results of a binding analysis using Chimera shows hydrophobic interactions (yellow and red) formed between this tryptophan and the MX helix loop (right). The density map is also shown (grey).*

Discussion

Nicotinic acetylcholine receptors cluster on the muscular side of the cholinergic synapse and are activated in response to the release of acetylcholine from the presynaptic nerve terminal. This investigation was the first of its type to analyse the structure of these receptors in their native cholinergic membranes. A short amphipathic helix, the MX helix, was seen to be a key element of this system, forming a discontinuous ring around individual receptors. The helix was seen to lie on the intracellular side of the membrane and appeared to regulate the receptor packing density via MX-MX interactions. This function may have been due to its interactions with the surrounding lipid environment, where density maps showed the MX helix interacting with surrounding lipids and controlling the local lipid environment evidenced by the absence of phospholipid head groups of membrane fatty acid chains at regions above the MX helix. The components of the lipid environment have been well documented and shown to influence NAChR structure and function, with one early study showing that a membrane containing cholesterol or anionic phospholipids was required for the receptor to exhibit a high affinity state for its agonist [20]. This effect was later shown to be caused by the decoupling of the agonist binding site and transmembrane region, with receptors adopting distinct conformations within membranes with different lipid constituents [44]. This study reported that NAChRs in native membranes or membranes containing phosphatidylcholine, cholesterol and phosphatidic acid, bound agonist and were seen to move from an activated conformation to an agonist desensitised conformation whilst membranes composed solely of phosphatidylcholine were seen to bind agonist. However no activation of receptors was seen due to the receptor already being stabilised in a desensitised conformation. These studies indicated the importance of membrane lipid composition in receptor function. Logically, for this to occur the receptor would require a method of ‘sensing’ the lipid environment. Due to the position of the MX receptor with respect to both the membrane, penetrating into the membrane at the intracellular boundary, and with respect to the ion channel, being located within a plane corresponding to the point of the channel with the smallest radius, places the MX helix in a perfect position to act as a lipid sensor and transmit this to the protein.

Evidence for the interactions between the MX helix and the lipid environment came from the ultrastructure. It was seen that at areas directly above the MX helix the densities corresponding to the phospho head groups of fatty acid chains were absent (Figure 2). The MX helix was seen to penetrate into the membrane. This, coupled with the charge distribution

seen in the model of the receptor (Figure 14A), may offer some insight into the displacement of the phospho heads within the membrane. The generation of the van der Waals isosurface and calculation of coulombic potentials in the final model (Figure 14A) indicated that the membrane facing region of the MX helix was relatively polar, whilst the interface between the MX helix and TM4/MA helix was seen to adopt a more hydrophobic character, consistent with the amphipathic nature of this helix. This may have the effect of placing the hydrophilic portion of the MX helix in a perfect place to interact with zwitterionic phospho-head groups of long chain fatty acids. While a hydrophobic pocket would prefer to interact with more hydrophobic molecules, in these regions small hydrophobic cholesterol molecules could replace fatty acid chains. However, in the density map reconstruction there was no evidence seen for lipids as discrete molecular densities. This was most likely due to averaging the structure over many images, and could lead to several conclusions. Either, there could have been a specific interaction and conformation of the MX helix and lipid molecules that was lost through small differences in the averaged images for the class, or the interactions between lipids and MX helix is not rigidly defined and hence the interactions would exhibit slight differences between micrographs that have been lost during the averaging process. It is also worth noting that cholesterol lacks a dominant polar moiety containing only a hydroxyl group at this resolution, as such, the probability of seeing this group as a discrete density is low and may explain the lack of discrete density observed above the MX helix. Whilst there is no direct evidence for the presence of cholesterol molecules above the MX helix, cholesterol is a necessary requirement for the function of NAChRs [20] and is present in the *Torpedo* membrane at a relatively high concentration of approximately 40 mol % whilst smaller phospholipids with relatively small head groups such as phosphatidic acid are present at much lower levels (<0.5 mol %) [45]. Due to this and the fact that the polar nature of the phosphatidic acid head groups would be disfavored at the MX regions, cholesterol seems to be the most likely candidate to occupy these regions.

The amphipathic MX helix in this investigation was seen to only penetrate into the phospholipid head groups within the membrane. This is contrary to other transmembrane amphipathic helices that have been observed to protrude into the fatty acid regions of the membrane [46]. Therefore, cholesterol molecules occupying the regions above the MX helix would be well positioned to align naturally against the acetyl chains of the surrounding membrane, possibly becoming trapped (Figure 15). The aligned cholesterol molecules would then be placed in space at the narrowest region of the channel at an area where it has been reported ion gating occurs. Thus, we suggest that the relative rigidity imparted by the cholesterol molecules would make ion selectivity more precise.

Previous reports have also highlighted important roles for cholesterols as auxiliary binding molecules for the function of NAChRs. Molecular dynamic studies have indicated that the TM4 helix acts as a lipid sensing element flipping between interactions with the TM1-4 bundle and the surrounding lipid environment during simulations [21], further *in silico* experiments identified two methods of interaction of cholesterol with the receptor, one superficial and another where cholesterol molecules were buried deeper within the receptor TMD [25]. However, these calculations were based upon the atomic model of the *Torpedo*

NAChR with the acknowledged TM2 register error (PDB ID: 2BG9). Additionally the model coordinates did not contain positions for the MX helix residues [3]. It may be the case that the MX helix also plays an additional role in the interaction between the receptor and cholesterol. A speculatory role for the MX helix could be to create an environment that favours cholesterol alignment to the receptor TMD, thus providing a readily available pool of cholesterol that may interact either specifically with the outer ring of the TMD or maintain the presence of cholesterol molecules buried deeper within the receptor. However, further experiments would be needed to confirm this.

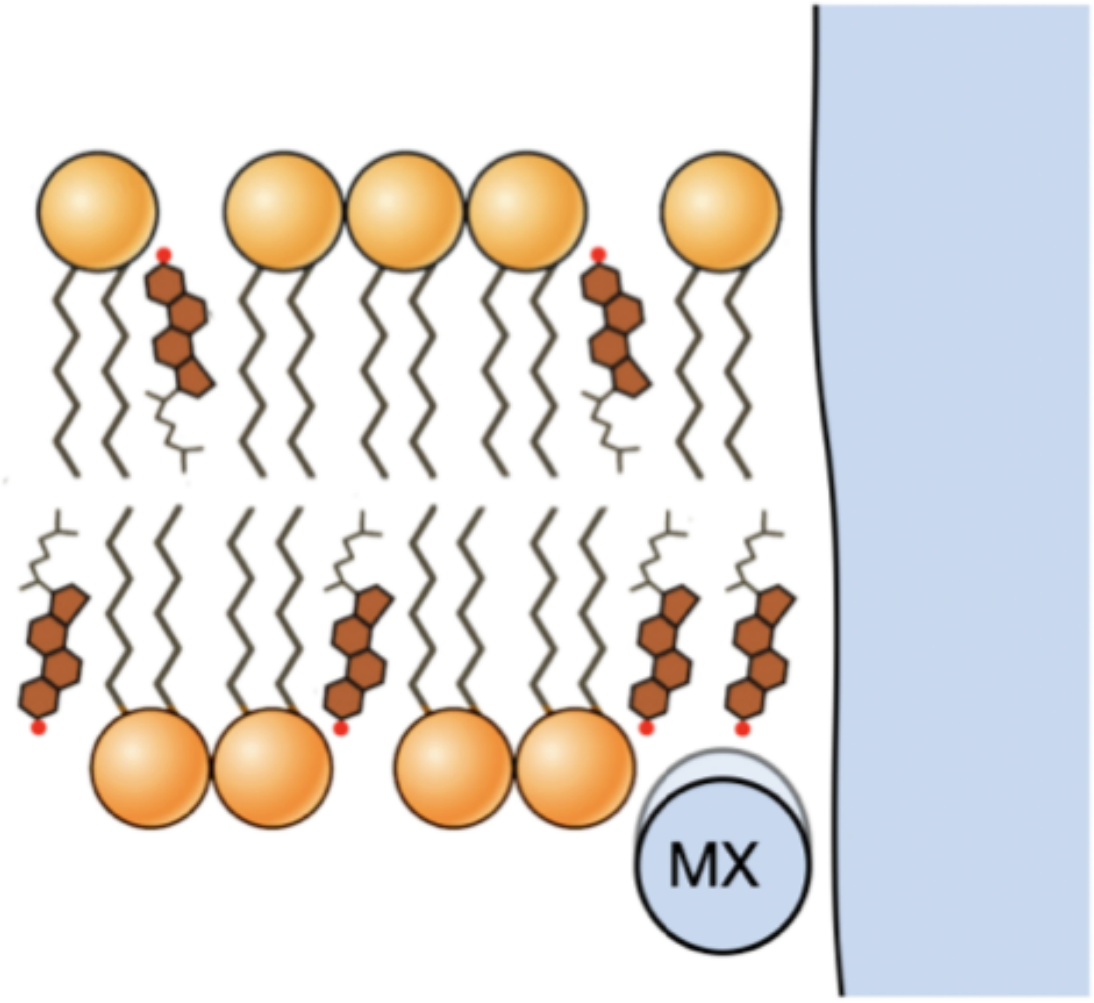


Figure 15. *A cartoon representation of the proposed mechanism of cholesterol sequestering at the MX helix. Long acyl chains of the lipid bi-layer are shown (orange ball: head group, black acyl chain: hydrophobic tails) where they are displaced above the MX helix (blue circle). Providing an energetically favourable environment for cholesterol (brown) within the membrane, cholesterol molecules would then align against the TMD (blue side wall) of the NAChR. Figure adapted from [42] with permissions.*

It must be acknowledged that during the course of this investigation a high resolution structure of the *Tetronarce californica* NAChR bound to the snake venom inhibitor **α**-Bungarotoxin at a resolution of 2.69 Å was released [40]. This structure had a very high

sequence homology to that of the *Torpedo* receptor (Table 3) this investigation was focused on and indeed provided a much better template for our modelling purposes. This report clearly takes away a lot of the impact of work done during this investigation and arguably the questions we set out to answer could have been answered by this structure. However, there are several key distinctions between the *Tetronarce californica* structure and our own. Namely, the *Tetronarce californica* structure was modelled in a deactivated state bound to a snake venom inhibitor. In their report the authors use early structures of the *Torpedo* NAChRs [3] (PDB ID: 2BG9) as a reference to identify the structural consequences of inhibition. The authors acknowledge that these structures contain a TM register error. Thus, our structure provides a more accurate reference point of the NAChR in the resting state within its native membrane. Secondly, to achieve the high resolution observed in their structure receptors were uncoupled from their native pentameric dimers and reconstituted as monomers. In order to fully consider the interactions of the MX helix with the lipid environment it is necessary to observe the regions where receptors align along the MX helix. Thus the structure we have determined provides a model of the *Torpedo* NAChR in the resting state, and provides structural insights regarding the interplay between the MX helix and native membrane.

Conclusions and future directions

This investigation produced a new atomic model for the resting state of the *Torpedo* NAChR within the context of the native membrane including the MX helix. The MX helix was seen to be located at the intracellular boundary of the lipid bi-layer and cytoplasm, penetrating into the membrane resulting in the exclusion of the hydrophilic heads of long chain fatty acids that make up the bilayer. The chemico-physical properties of the MX within the receptor setting showed that it maintains a hydrophobic environment well suited to interact with small hydrophobic molecules such as cholesterol. This may be critical for conferring rigidity to the receptor based on the position of the MX helices being at the narrowest section of the pore. Further experiments would be needed to further tease out a role for the MX in receptor lipid interactions. Previous *in silico* studies [21, 25] have relied upon receptor models that contain errors within the TMD and lack the MX helix. Thus, this model would be well suited to a molecular dynamic analysis aimed at investigating the equilibrium between the receptor and the surrounding lipid environment. Following on from this, a virtual screening campaign directed at the MX helix site may identify novel therapeutics to alter the function of NAChRs.

References

1. Thompson AJ, Lester HA, Lummis SCR. The structural basis of function in Cys-loop receptors. Q Rev Biophys. 2010;43:449–99.

2. Walsh RM, Roh S-H, Gharpure A, Morales-Perez CL, Teng J, Hibbs RE. Structural principles of distinct assemblies of the human α4β2 nicotinic receptor. Nature. 2018;557:261–5.

3. Unwin N. Refined structure of the nicotinic acetylcholine receptor at 4A resolution. J Mol Biol. 2005;346:967–89.
4. Miyazawa A, Fujiyoshi Y, Unwin N. Structure and gating mechanism of the acetylcholine receptor pore. Nature. 2003;423:949–55.

5. Basak S, Gicheru Y, Samanta A, Molugu SK, Huang W, Fuente M la de, et al. Cryo-EM structure of 5-HT3A receptor in its resting conformation. Nat Commun. 2018;9:514.

6. Polovinkin L, Hassaine G, Perot J, Neumann E, Jensen AA, Lefebvre SN, et al. Conformational transitions of the serotonin 5-HT3 receptor. Nature. 2018;563:275–9.

7. Althoff T, Hibbs RE, Banerjee S, Gouaux E. X-ray structures of GluCl in apo states reveal a gating mechanism of Cys-loop receptors. Nature. 2014;512:333–7.

8. Hibbs RE, Gouaux E. Principles of activation and permeation in an anion-selective Cys-loop receptor. Nature. 2011;474:54–60.

9. Du J, Lü W, Wu S, Cheng Y, Gouaux E. Glycine receptor mechanism elucidated by electron cryo-microscopy. Nature. 2015;526:224–9.

10. Miller PS, Aricescu AR. Crystal structure of a human GABAA receptor. Nature. 2014;512:270–5.

11. Masiulis S, Desai R, Uchański T, Serna Martin I, Laverty D, Karia D, et al. GABAA receptor signalling mechanisms revealed by structural pharmacology. Nature. 2019;565:454–9.

12. Bertrand S, Weiland S, Berkovic SF, Steinlein OK, Bertrand D. Properties of neuronal nicotinic acetylcholine receptor mutants from humans suffering from autosomal dominant nocturnal frontal lobe epilepsy. Br J Pharmacol. 1998;125:751–60.

13. Coyle JT, Price DL, DeLong MR. Alzheimer's disease: a disorder of cortical cholinergic innervation. Science. 1983;219:1184–90.

14. Aubert I, Araujo DM, Cécyre D, Robitaille Y, Gauthier S, Quirion R. Comparative alterations of nicotinic and muscarinic binding sites in Alzheimer's and Parkinson's diseases. J Neurochem. 1992;58:529–41.

15. Kistler J, Stroud RM. Crystalline arrays of membrane-bound acetylcholine receptor. Proc Natl Acad Sci U S A. 1981;78:3678–82.

16. Celie PHN, van Rossum-Fikkert SE, van Dijk WJ, Brejc K, Smit AB, Sixma TK. Nicotine and carbamylcholine binding to nicotinic acetylcholine receptors as studied in AChBP crystal structures. Neuron. 2004;41:907–14.

17. Chiara DC, Xie Y, Cohen JB. Structure of the agonist-binding sites of the Torpedo nicotinic acetylcholine receptor: affinity-labeling and mutational analyses identify gamma Tyr-111/delta Arg-113 as antagonist affinity determinants. Biochemistry. 1999;38:6689–98.

18. Unwin N, Fujiyoshi Y. Gating movement of acetylcholine receptor caught by

plunge-freezing. J Mol Biol. 2012;422:617–34.

19. Jones OT, Eubanks JH, Earnest JP, McNamee MG. A minimum number of lipids are required to support the functional properties of the nicotinic acetylcholine receptor. Biochemistry. 1988;27:3733–42.

20. Fong TM, McNamee MG. Correlation between acetylcholine receptor function and structural properties of membranes. Biochemistry. 1986;25:830–40.

21. Xu Y, Barrantes FJ, Luo X, Chen K, Shen J, Jiang H. Conformational dynamics of the nicotinic acetylcholine receptor channel: a 35-ns molecular dynamics simulation study. J Am Chem Soc. 2005;127:1291–9.

22. Mitra A, Bailey TD, Auerbach AL. Structural dynamics of the M4 transmembrane segment during acetylcholine receptor gating. Struct Lond Engl 1993. 2004;12:1909–18.

23. Leibel WS, Firestone LL, Legler DC, Braswell LM, Miller KW. Two pools of cholesterol in acetylcholine receptor-rich membranes from Torpedo. Biochim Biophys Acta. 1987;897:249–60.

24. Hamouda AK, Chiara DC, Sauls D, Cohen JB, Blanton MP. Cholesterol interacts with transmembrane alpha-helices M1, M3, and M4 of the Torpedo nicotinic acetylcholine receptor: photolabeling studies using [3H]Azicholesterol. Biochemistry. 2006;45:976–86.

25. Brannigan G, Hénin J, Law R, Eckenhoff R, Klein ML. Embedded cholesterol in the nicotinic acetylcholine receptor. Proc Natl Acad Sci U S A. 2008;105:14418–23.

26. Kubalek E, Ralston S, Lindstrom J, Unwin N. Location of subunits within the acetylcholine receptor by electron image analysis of tubular crystals from Torpedo marmorata. J Cell Biol. 1987;105:9–18.

27. Miyazawa A, Fujiyoshi Y, Stowell M, Unwin N. Nicotinic acetylcholine receptor at 4.6 A resolution: transverse tunnels in the channel wall. J Mol Biol. 1999;288:765–86.

28. Newcombe J, Chatzidaki A, Sheppard TD, Topf M, Millar NS. Diversity of Nicotinic Acetylcholine Receptor Positive Allosteric Modulators Revealed by Mutagenesis and a Revised Structural Model. Mol Pharmacol. 2018;93:128–40.

29. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

30. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234:779–815.

31. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006;15:2507–24.

32. Studer G, Biasini M, Schwede T. Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinforma Oxf Engl.

2014;30:i505-511.

33. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem. 2004;25:1605–12.

34. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein structure fitting and refinement guided by cryo-EM density. Struct Lond Engl 1993. 2008;16:295–307.

35. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42–9.

36. Pandurangan AP, Topf M. RIBFIND: a web server for identifying rigid bodies in protein structures and to aid flexible fitting into cryo EM maps. Bioinforma Oxf Engl. 2012;28:2391–3.

37. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, et al. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol Clifton NJ. 2016;1374:23–54.

38. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, Apweiler R. High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform. 2002;3:275–84.

39. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43:W389-394.

40. Rahman MM, Teng J, Worrell BT, Noviello CM, Lee M, Karlin A, et al. Structure of the Native Muscle-type Nicotinic Receptor and Inhibition by Snake Venom Toxins. Neuron. 2020;106:952-962.e5.

41. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 1:12–21.

42. Unwin N. Protein-lipid architecture of a cholinergic postsynaptic membrane. IUCrJ. 2020;7 Pt 5:852–9.

43. Hassaine G, Deluz C, Grasso L, Wyss R, Tol MB, Hovius R, et al. X-ray structure of the mouse serotonin 5-HT3 receptor. Nature. 2014;512:276–81.

44. daCosta CJB, Baenziger JE. A lipid-dependent uncoupled conformation of the acetylcholine receptor. J Biol Chem. 2009;284:17819–25.

45. Rotstein NP, Arias HR, Barrantes FJ, Aveldaño MI. Composition of lipids in elasmobranch electric organ and acetylcholine receptor membranes. J Neurochem. 1987;49:1333–40.

46. Hristova K, Wimley WC, Mishra VK, Anantharamiah GM, Segrest JP, White SH. An amphipathic alpha-helix at a membrane interface: a structural study using a novel X-ray

diffraction method. J Mol Biol. 1999;290:99–117.

Chapter 7

Thesis Summary

The following chapter provides a summary of the works presented in each chapter and the future directions that these projects could take.

Chapter 2

A novel approach to fitting small molecules in cryo-EM maps reveals insights into GSK-1 inhibition of human kinesin 5

Chapter 2 aimed to derive an atomic model of the human kinesin-5 family member Eg5 in complex with a novel inhibitor GSK-1. A cryo-EM map was derived by Dr Alejandro Peña to an average resolution of 3.8 Å. The atomic model was used to provide insights into the novel binding mode and mechanism of action of the GSK-1 inhibitor.

An initial model was built using homology modelling with MODELLER [1]. The template structure used was a model of the Human Eg5 motor domain in complex with the inhibitor PVZB1194 [2]. This model was not complete and incomplete regions were built using a second homology model of the Human Eg5 motor domain in the AMPPNP bound state [3]. The protein was fitted into the cryoEM density using hierarchical modelling with Flex-EM [4, 5]. The fit of the protein model to the map improved at each successive stage of modelling. To identify the GSK-1 binding site, a three pronged consensus method was used. This methodology included the use of Meta-Pocket 2.0 binding site prediction software [6], blind docking with AutoDock Vina [7], and density difference mapping [8]. Both AutoDock Vina and Meta-Pocket 2.0 identified multiple sites within the proteins where small molecules may bind, these predictions correlated well with binding sites of previous Eg5 motor domain inhibitors . The difference mapping protocol showed one region at the boundary where the Eg5 motor domain interacts with microtubules, correlating well with the results of AutoDock Vina, Meta-Pocket 2.0 and a previous report of the Eg5 motor domain bound to the PVZB1194 inhibitor [2]. The GSK-1 small molecule was fitted to the identified binding site using a consensus molecular docking protocol with GOLD [9] and AutoDock Vina [7]. The initial results did not correlate well with ligand density seen within the map and the sidechain residues of the binding site were further refined in the presence of the ligand. The consensus docking protocol was repeated using this refined model and two equally plausible binding modes were identified.

The final model shed light on a plausible mechanism of action of the small molecule inhibitor GSK-1. With the protein seen to adopt structural features similar to that of an ATP or

AMPPNP bound state, where the motor domain has been shown to have a high affinity for the microtubules [10]. In terms of the position of the GSK-1 binding site this correlated well with experimental data provided by Dr Aalejandro Peña [11]. However the resolution of the map was not sufficiently high to confidently identify a single binding mode for GSK-1. Future experiments in this project would aim to decipher a single binding mode for GSK-1. This could be achieved using molecular dynamic simulations. Furthemore, a virtual screening campaign would aim to identify further possible inhibitors that bind within this site.

Chapter 3

A new Program for Protein-Ligand interaction detection (ProPLID) utilising bond geometry

Chapter 3 aimed to build a new software for the identification of protein-ligand interactions in atomic models. The software used geometric parameters derived from a large scale analysis of protein-ligand interactions including strong and weak hydrogen bonds, hydrophobic interactions, $\boldsymbol{\pi}$-$\boldsymbol{\pi}$ stacking interactions, cation-$\boldsymbol{\pi}$ interactions, halogen bonds and metal ion complexes. Most interactions were identified by geometric criteria, however, since little geometric data exists for hydrophobic interactions, predictions were aided using a method to calculate the local hydrophobicity of the binding site [12]. The geometric parameters used for interaction detection cutoff values were taken from literature values including a large-scale analysis of protein-ligand interactions within the PDB [13] and metal-ion geometry seen within the Cambridge Structural Database [14].

To assess the accuracy of the software, a benchmark was curated composed of 35 high resolution atomic models of protein-ligand complexes deposited in the PDB [15]. This benchmark represented a total of 321 protein-ligand interactions, the distribution of which correlated well with literature reports [13]. The F-measure was chosen to assess the accuracy of the software based on the number of true positive, false negative, and false positive predictions made. The software had a mean F-measure over the whole benchmark of 0.386, and this was compared against the software PLIP [16, 17]. The PLIP software showed a mean F-measure of 0.296 over the same benchmark, a result that was significantly worse than the software presented in this chapter.

Future experiments would aim to add further interaction types to the software, as well as adding in energy terms for each interaction type to aid in increasing the number of true positive interactions identified, whilst simultaneously reducing the number of false positives predicted.

The following two chapters aimed at developing a method to fit small molecules into cryo-EM maps at both medium-to-low and high resolutions.

Chapter 4

Integrating goodness-of-fit metrics with an empirical scoring function for fitting small molecules to density map

Chapter 4 presented a novel scoring function that integrated the MI goodness-of-fit metric with an empirical scoring function for scoring protein-ligand interactions. We began by comparing the use of two common goodness-of-fit metrics, the MI and CCC, at identifying a correct ligand conformation, using a benchmark derived from the CASF-2016 decoy database [18]. Both scores were compared using simulated full maps and density difference maps. The Pearson correlation coefficient between goodness-of-fit scores and the RMSD of decoys to reference structures was chosen to assess the '*fitting power*' of both scores. It was shown that both scores had a better '*fitting power'* when the density difference maps were used. When the density difference maps were used, the MI score was better correlated with the decoy RMSD than the CCC. This result was only significant at resolutions worse than 4.5 Å, in-line with previous reports [19].

Following on from this a new empirical scoring function was parameterized based on the AutoDock Vina scoring function [7]. This empirical scoring function introduced terms for the scoring of $\boldsymbol{\pi}$-$\boldsymbol{\pi}$ interactions, an interaction frequently reported in protein-ligand complexes. The $\boldsymbol{\pi}$-$\boldsymbol{\pi}$ interaction terms were optimised using a benchmark of 3079 protein ligand complexes containing $\boldsymbol{\pi}$-$\boldsymbol{\pi}$ interactions, and were seen to be a reasonable approximation of the distribution of data found in this benchmark.

Following this, four separate scoring functions containing terms to score hydrogen bonding, steric interactions, atom-atom repulsion, hydrophobic interactions, and $\boldsymbol{\pi}$-$\boldsymbol{\pi}$ interactions were assessed. The aim of this was to derive an empirical scoring function with the ability to rank a correct ligand conformation from an incorrect one. To learn the weighting terms for each score, 101 protein ligand complexes were used, for each 950 decoy conformations were generated along with 50 correct conformations. The weights were optimised by minimising the average Pearson correlation coefficient between the docking scores and the RMSD for conformations to the reference (deposited) ligand conformation. For the training sets, the addition of the aromatic scoring term resulted in a better average Pearson correlation coefficient. The scoring functions were tested using a test set of 23 protein-ligand complexes and up to 100 RMSD decoys from the CASF-2016 data set [18]. The average Pearson correlation coefficient was lower for the test set than for the training set, most likely due to a harder benchmark being used during testing. However, the scores all showed a better '*docking pow*er' than the AutoDock Vina [7] scoring function on the same test set. All four scoring functions showed a comparable docking power, and the score that had the best average Pearson correlation coefficient on the test set was used in further experiments.

The last set of experiments aimed to integrate the MI score with the new empirical scoring function. This was done by weighting the MI score by a value that brought it in-line with the

empirical scoring function at a magnitude of 0.1x, 0.5x, 1x, 5x and 10x. The best weight of the MI score was assessed using the test set of 23 protein-ligand complexes and CASF-2016 decoys [18]. The average Pearson correlation between the integrated score and the RMSD of ligand conformations to the reference ligand was calculated for each weight using simulated full and difference density maps at resolutions between 2.5 and 8.5 Å. Using density difference maps the best weights identified were at 5x and 10x for all resolutions. This result was echoed using full maps at resolutions up to 4.5 Å. However, at lower resolutions the best weight dropped to 1x and 0.5x that of the empirical scoring function. This investigation showed a novel scoring function for fitting small molecules into cryo-EM maps and further work aimed at integrating this with a fitting algorithm and benchmarking using experimental data.

Chapter 5

A genetic algorithm for the flexible fitting of small molecules

This chapter aimed to build upon the work in chapter 4 and develop an automated method for fitting small molecules into cryo-EM maps. A three stage genetic algorithm (GA) was developed as a search algorithm for fitting small molecules. This algorithm optimised the position and orientation of ligands in the binding site, along with adjusting molecules around their rotatable bonds. Two experimental benchmarks were curated, the first was composed of 25 high resolution (≤ 3.0 Å) protein ligand complexes from the PDB [15] and EMDB with associated cryo-EM maps. The second was composed of 15 protein-ligand complexes and maps, with resolutions between 3.0 and 4.5 Å. For the lower resolution benchmark, control structures were identified for each complex with comparative models derived at high resolutions.

Ligands were removed from their associated complexes and fit into cryo-EM maps using the GA using the integrated score developed in chapter 5 and difference maps calculated using the deposited maps and protein models with a local difference mapping methodology implemented in TEMPy. In all cases and for both benchmarks, the integrated score using the empirical scoring function (Chapter 4) and the weighted MI score (at 0.1x, 0.5x , 1x , 10x the magnitude of the empirical scoring function) was used.

For each case a fit was considered correct if the resultant ligand was within 2.0 Å of the deposited ligand. For the high-resolution benchmark comparable rates of success were seen using weights of 1x, 5x, 10x or the MI score alone, with success rate between 92.30 % and 96.15 %. However, the best results for the integrated score at 1x were seen to be closer in RMSD to that in the deposited structures. For the low resolutions benchmark the best weight for the MI score was seen to be 0.5x, indicating that at lower resolutions more input from the empirical scoring function was needed to make up the shortfall information in the map. To account for model bias, the deposited protein models were re-refined into the deposited maps in the absence of the ligand. At both high and low resolutions a comparable success rate to

that of the deposited models was seen when using the re-refined protein models. Finally, the analysis repeated using the full deposited maps for both benchmarks showed a higher success rate when using the density difference mapping technique. Failed cases were mostly seen to be due to incorrect binding site side-chain orientations within the re-refined protein and low quality difference maps.

The methodology presented here showed a novel method of fitting small molecules distinct from previously published methodologies [20, 21]. The algorithm was shown to be able to accurately fit small molecules up to resolutions of 4.5 Å. Future investigation would aim to improve the scoring function by adding additional terms representing a wider variety of protein-ligand interactions. Additionally, integrating terms for scoring the quality of ligand conformation such as the torsional strain energy would be investigated.

Chapter 6

The MX helix of the Torpedo marmorata nicotinic acetylcholine receptor in its native membrane

Chapter 6 aimed to calculate an atomic model into a cryo-EM map of the *Torpedo* Nicotinic Acetylcholine Receptor (NAChR) at 6.6 Å in its native membrane (Map provided by Dr Nigel Unwin). This atomic model was used to investigate the position of the MX helix within this membrane. Due to the resolution of the map homology modelling was employed to build the atomic model. At the time of building the model (in collaboration with Nigel Unwin, LMB) two structures of the *Torpedo* NAChR [22, 23] were deposited within the PDB [15]. Both structures were discounted as candidates for the main template used in model building, due to a lack of information present regarding the MX helix and a TM register error in both structures [24]. An $\alpha 4\beta 2$ NAChR structure was chosen as the main template for modelling [25], along with a 5-HT3-A receptor [26] and a *Torpedo* NAChR structure [23]. An initial sequence alignment [24] was obtained from Dr. Joseph Newcombe, this alignment was updated and a small error fixed.

Initial models were generated using MODELLER [1] and the model fitness to the map using a hierarchical modelling protocol with Flex-EM [5, 19]. At each round of fitting the mode-map agreement was assessed with the CCC and SMOC score. The agreement of the model with the map improved with each successive round of flexible fitting. One problem area identified during fitting was the positions of the TM and MX helices relative to themselves and the membrane. The resolution of the map did not permit for further refinement without the possibility of overfitting. To overcome this, a bioinformatic analysis was undertaken. This analysis highlighted regions of hydrophobicity and hydrophilicity within both the MX and TM helices, which identified where these helices aligned within the context of the double-membrane receptors. At this point in the course of the investigation a high resolution structure of the *Tetronarce californica* NAChR bound to α-Bungarotoxin at a resolution of 2.69 Å was released [27]. This model provided a much better starting point for

homology modelling and a new model was built using MODELLER. A comparison of the position of the heavy atoms of both models showed an RMSD of 2.8 Å. Both models were seen to have a comparable mode quality when assessed using MolProbity scores [28, 29]. However, the model based on the high resolution structure had a better score when assessed with the QMEANBrane score [30]. The final model was seen to correlate well with the observations made during the bioinformatic analysis.

The atomic model allowed us to make inferences regarding the functional role of the MX helix. It was suggested that the position of the MX helix relative to the TM domain allows the MX helix to 'sense' the local lipid environment and 'transmit' this information to the protein channel. Evidence supporting this came from the ultra-structure of the density maps and the absence of membrane phospholipid head groups at regions above the MX helices, however, no discrete density for lipids was seen in the averaged density maps used to build the atomic model. Future experiments would aim to further understand the function of the MX helix, this would be done using the final model and molecular dynamics simulations within the context of the double membrane lipid environment. Furthermore, this region may provide a novel site for small molecules to bind. Considering the importance of the NAChRs in disease, a virtual screening campaign may identify novel inhibitors to perturb the function of the NAChRs. A recent study on the **α**7 NAChR demonstrated the utility of such approach [31].

References

1. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234:779–815.

2. Yokoyama H, Sawada J-I, Katoh S, Matsuno K, Ogo N, Ishikawa Y, et al. Structural basis of new allosteric inhibition in Kinesin spindle protein Eg5. ACS Chem Biol. 2015;10:1128–36.

3. Parke CL, Wojcik EJ, Kim S, Worthylake DK. ATP hydrolysis in Eg5 kinesin involves a catalytic two-water mechanism. J Biol Chem. 2010;285:5859–67.

4. Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A. Protein Structure Fitting and Refinement Guided by Cryo-EM Density. Structure. 2008;16:295–307.

5. Joseph AP, Malhotra S, Burnley T, Wood C, Clare DK, Winn M, et al. Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. Methods. 2016;100:42–9.

6. Zhang Z, Li Y, Lin B, Schroeder M, Huang B. Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics. 2011;27:2083–8.

7. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31:455–61.

8. Joseph AP, Lagerstedt I, Jakobi A, Burnley T, Patwardhan A, Topf M, et al. Comparing Cryo-EM Reconstructions and Validating Atomic Model Fit Using Difference Maps. J Chem Inf Model. 2020;60:2552–60.

9. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking11Edited by F. E. Cohen. J Mol Biol. 1997;267:727–48.

10. Rosenfeld SS, Fordyce PM, Jefferson GM, King PH, Block SM. Stepping and stretching. How kinesin uses internal strain to walk processively. J Biol Chem. 2003;278:18550–6.

11. Peña A, Sweeney A, Cook AD, Locke J, Topf M, Moores CA. Structure of Microtubule-Trapped Human Kinesin-5 and Its Mechanism of Inhibition Revealed Using Cryoelectron Microscopy. Structure. 2020;28:450–7.e5.

12. Wang R, Lai L, Wang S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J Comput Aided Mol Des. 2002;16:11–26.

13. de Freitas RF, Schapira M. A systematic analysis of atomic protein–ligand interactions in the PDB. Med Chem Commun. 2017;8:1970–81.

14. Harding MM. Geometry of metal-ligand interactions in proteins. Acta Crystallogr D Biol Crystallogr. 2001;57 Pt 3:401–11.

15. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42.

16. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015;43:W443–7.

17. Adasme MF, Linnemann KL, Bolz SN, Kaiser F, Salentin S, Haupt VJ, et al. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. Nucleic Acids Res. 2021;49:W530–4.

18. Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, et al. Comparative Assessment of Scoring Functions: The CASF-2016 Update. J Chem Inf Model. 2019;59:895–913.

19. Vasishtan D, Topf M. Scoring functions for cryoEM density fitting. J Struct Biol. 2011;174:333–43.

20. Robertson MJ, van Zundert GCP, Borrelli K, Skiniotis G. GemSpot: A Pipeline for Robust Modeling of Ligands into Cryo-EM Maps. Structure. 2020;28:707–16.e3.

21. Vant JW, Lahey S-LJ, Jana K, Shekhar M, Sarkar D, Munk BH, et al. Flexible Fitting of Small Molecules into Electron Microscopy Maps Using Molecular Dynamics Simulations with Neural Network Potentials. J Chem Inf Model. 2020;60:2591–604.

22. Miyazawa A, Fujiyoshi Y, Unwin N. Structure and gating mechanism of the acetylcholine receptor pore. Nature. 2003;423:949–55.

23. Unwin N. Refined structure of the nicotinic acetylcholine receptor at 4A resolution. J Mol Biol. 2005;346:967–89.

24. Newcombe J, Chatzidaki A, Sheppard TD, Topf M, Millar NS. Diversity of Nicotinic Acetylcholine Receptor Positive Allosteric Modulators Revealed by Mutagenesis and a Revised Structural Model. Mol Pharmacol. 2018;93:128–40.

25. Walsh RM, Roh S-H, Gharpure A, Morales-Perez CL, Teng J, Hibbs RE. Structural principles of distinct assemblies of the human α4β2 nicotinic receptor. Nature. 2018;557:261–5.

26. Polovinkin L, Hassaine G, Perot J, Neumann E, Jensen AA, Lefebvre SN, et al. Conformational transitions of the serotonin 5-HT3 receptor. Nature. 2018;563:275–9.

27. Rahman MM, Teng J, Worrell BT, Noviello CM, Lee M, Karlin A, et al. Structure of the Native Muscle-type Nicotinic Receptor and Inhibition by Snake Venom Toxins. Neuron. 2020;106:952–62.e5.

28. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res. 2007;35 Web Server issue:W375–83.

29. Chen VB, Arendall WB 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, et al. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr. 2010;66 Pt 1:12–21.

30. Studer G, Biasini M, Schwede T. Assessing the local structural quality of transmembrane protein models using statistical potentials (QMEANBrane). Bioinformatics. 2014;30:i505–11.

31. Smelt CLC, Sanders VR, Newcombe J, Burt RP, Sheppard TD, Topf M, et al. Identification by virtual screening and functional characterisation of novel positive and negative allosteric modulators of the α7 nicotinic acetylcholine receptor. Neuropharmacology. 2018;139:194–204.

Appendices

Chapter 3

```
class Interaction_data():
    def  __init__(self):

        #general parameters
        self.ligand_binding_site_radius = 9.0
        self.residue_ids = ["ALA", "ARG", "ASN" ,"ASP", "CYS", "GLU","GLN", "GLY", "HIS",
"ILE","LEU", "LYS", "MET", "PHE", "PRO", "SER","THR", "TRP", "TYR","VAL"]
        self.aromatic_sidechains = ['PHE', 'TYR','TRP', 'HIS']
        self.Cation_residues = ['LYS', 'ARG']
        self.Cation_atom_ids = ['NH1','NH2', 'NZ']


        #metal complex geometric parameters
        self.ion_ids = ["CA", "MN", "MG", "FE", "CU","ZN"]
        self.ion_protein_distences = {"CA": {"HOH": 2.39, "ASP": 2.36, "GLU": 2.36,  "SER": 2.43,
"THR":  2.43, "TYR": 2.20, "HIS":  2.38,  "CYS" :2.56, "O":2.36},
                      "MG": {"HOH": 2.07, "ASP": 2.26, "GLU": 2.26,  "SER": 2.10,  "THR": 2.10,
"TYR": 1.87, "HIS": 2.05, "CYS"  : 2.03, "O":2.26},
                      "MN": {"HOH": 2.21, "ASP": 2.21, "GLU": 2.21,  "SER": 2.25, "THR": 2.25, "TYR":
1.88, "HIS":  2.19, "CYS" :2.35, "O": 2.21},
                      "FE": {"HOH":  2.09, "ASP": 2.01, "GLU":  2.01, "SER":2.13, "THR":  2.13,
"TYR": 1.93, "HIS": 2.08, "CYS": 2.27,"O":2.01},
                      "CU": {"HOH":1.97, "ASP":  1.96, "GLU":  1.96,  "SER":  2.00,  "THR":   2.00,
"TYR": 1.90, "HIS":  1.99, "CYS": 2.17, "O":1.96},
                      "ZN": {"HOH": 2.09, "ASP":   2.04, "GLU":  2.04, "SER":  2.14, "THR": 2.14,
"TYR":  1.95,  "HIS":  2.00,  "CYS": 2.29, "O":2.04}}

        self.ion_ligand_distences= {"CA":{"O": 2.43 , "N":2.38, "ELSE": 2.56},
                           "MG": {"O": 2.10, "N": 2.05, "ELSE": 2.03},
                           "MN": {"O":  2.25, "N": 2.19, "ELSE":2.35},
                           "FE":{"O": 2.13, "N": 2.08, "ELSE":  2.27},
                           "CU": {"O":2.00, "N": 1.99, "ELSE": 2.17},
                           "ZN": {"O":2.14, "N": 2.00, "ELSE":  2.29}}

        self.ion_dist_tolerence =  0.5

        self.ion_residue_donor_ids  = {"ELSE":["O"],"HOH": ["O"], "ASP": ["OD2", "OD1"], "GLU":
["OE1", "OE2"], "SER":["OG"], "THR":["OG1"], "TYR":["OH"], "HIS":["NE2","ND1"], "CYS":["SG"] }

        self.ion_donor_symbols = ["N", "O", "S"]

        self.ion_angle_data = {7:{"name":"linear","NumAtoms": 2, "angles": [[180.0]] * 2},
                  6: {"name": "Trigonal Planar","NumAtoms": 3 , "angles": [[120.0,120.0]] * 3},
                  5: {"name": "Square Planar", "NumAtoms": 4 , "angles": [[90.0,90.0,180.0]] * 4},
                  4: {"name": "Square  Pyramidal","NumAtoms": 4 ,"angles":[[90.0,90.0,90,90.0]] +
[[90.0,90.0,90.0,180.0]] * 4},
                  3: {"name": "Tetrahedral", "NumAtoms": 4 ,"angles": [[109.5, 109.5]]  * 4},
                  2: {"name": "Trigonal Bipyramidal","NumAtoms": 5, "angles": [[90.0, 90.0, 120.0,
120.0]] * 3 +[[90.0,90.0,90.0,180.0]] *2 },
                  1: {"name": "Octahedral","NumAtoms": 6 , "angles": [[90.0,90.0,90.0,90.0,180.0]] *
6},#added another 180
                  0: {"name": "Pentagonal Bipyramidal", "NumAtoms": 7, "angles":[[72,72,90,90,144,
144]] * 5 + [[90,90,90,90,90,180]] * 2 }}

        self.ion_angle_deviation =  18.0

        self.ion_rms_deviation = 0.5
```

```
        self.ion_order_relation = {"Octahedral": ["Square  Pyramidal","Square Planar", "linear"],
                        "Square Planar":["linear"],
                        "linear": [],
                        "Trigonal Bipyramidal":["Trigonal Planar"],
                        "Trigonal Planar":[],
                        "Square  Pyramidal":["Square Planar","linear"],
                        "Tetrahedral" :  [],
                        "Pentagonal Bipyramidal": []
                        }
        self.ion_order_relation_rms = 1.5

        #hydrogen bond parameters
        #strong hydrogen bonds
        self.hbond_dist_max = 3.9
        self.hbond_don_angle_min_degree =  90
        self.hbond_electronegitive = ['O','N']
        self.Hbonds_smarts_to_exclude = ["[N][C](=[O])", "[N+]"]

        #weak hydrogen bonds
        self.weak_HBond_acceptors = ["O"]
        self.cho_hbond_dist_max = 3.6
        self.cho_hbond_dist_min = 3.0
        self.cho_hbond_don_angle_min_degree = 130

        #pi stack parameters
        self.pistack_dist_max = 4.0
        self.pistack_ang_dev = 30
        self.pistack_offset_max = 2.0

        #cation-pi parameters
        self.pication_dist_max = 6.0

        #hydrophobic interaction parameters
        self.hydrophobic_elements = ["C", "H"]
        self.contact_delta_hydro_max = -0.4
        self.contact_delta_hydro_min = 0.0
        self.contact_delta_clash = 0.6
        self.condense_hydrophobic_interactions = True
        self.hydrophobic_env_dist_cutoff = 6.0
        self.hydrophobic_env_logp_min = -0.5

        self.xlogp3_values = {'GLY': {'C': -0.8076, 'O':0.7148, 'N':-0.2610,'CA':-0.0821},
     'ALA': {'C': -0.8076, 'O':0.7148, 'N':-0.2610,'CA':-0.1426,'CB':0.5201},
     'VAL': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.1485,'CG1':0.5201,'CG2':0.5201},
     'ILE': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.1485,'CG2':0.5201,'CG1':0.3436,'CD1':0.5201},
     'LEU': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':0.1485,'CD1':0.5201,'CD2':0.5201},
     'MET': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':-0.0821,'SD':0.4125,'CE':0.0402},
     'PHE': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.2718,'CG':0.1911,'CD1':0.3157,'CD2':0.3157,'CE1':0.3157,'CE2':0.3157,'
CZ':0.3157},
     'TYR': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.2718,'CG':0.1911,'CD1':0.3157,'CD2':0.3157,'CE1':0.3157,'CE2':0.3157,'
CZ':-0.0112,'OH':-0.0381},
     'TRP': {'C': -0.8076, 'O':0.7148, 'N':-0.2610,'CA':-0.1426,'CB':0.2718,
'CG':0.1911,'CD1':-0.1039,'CD2':0.1911,'NE1':0.2172,'CE2':-0.0112,'CE3':0.3157,'CZ2':0.3157,'CZ3':0.31
57,'CH2':0.3157},
     'SER': {'C': -0.8076, 'O':0.7148, 'N':-0.2610,'CA':-0.1426,'CB':-0.0821,'OG':-0.4802},
     'THR': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':-0.1426,'CG2':0.5240,'OG1':-0.4802},
     'ASN': {'C': -0.8076, 'O':0.7148,
```

```
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':-0.8076,'OD1':0.7148,'ND2':-0.6414},
     'GLN': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':0.3436,'CD':-0.8076,'OE1':0.7148,'NE2':-0.6414},
     'ASP': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':-0.8076,'OD1':-0.4802,'OD2':-0.4802},
     'GLU': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':0.3436,'CD':-0.8076,'OE1':-0.4802,'OE2':-0.4802},
     'ARG': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':0.3436,'CD':-0.0821,'NE':-0.2610,'CZ':-0.8076,'NH1':-0.7445,
'NH2':-0.7445},
     'LYS': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.3436,'CG':0.3436,'CD':0.3436,'CE':-0.0821,'NZ':-0.7445},
     'HIS': {'C': -0.8076, 'O':0.7148,
'N':-0.2610,'CA':-0.1426,'CB':0.2718,'CG':-0.1874,'ND1':0.3181,'CD2':-0.1039,'CE1':-0.1039,'NE2':0.318
1},
     'CYS': {'C': -0.8076, 'O':0.7148, 'N':-0.2610,'CA':-0.1426,'CB':0.0821,'SG':0.4927},
     'PRO': {'C': -0.8076, 'O':0.7148, 'N':0.3333,'CA':-0.1426,'CB':0.3436,'CG':0.3436,'CD':-0.0821}
     }

        #halogen bond parameters
        self.halogen_donor_atoms = ['I','Br', 'Cl']
        self.halogen_donor_root_atoms = ['C']
        self.halogen_acceptor_atoms = ['O','N','S']
        self.halogen_acceptor_root_atoms = ['C','N','P','S']
        self.halogen_geometric_data = {"Cl":{"O":{"dist": 3.47, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0},
                                             "N":{"dist": 3.5, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0},
                                             "S":{"dist": 3.75, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0}},
                                       "Br":{"O":{"dist": 3.57, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0},
                                             "N":{"dist": 3.6, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0},
                                             "S":{"dist": 3.85, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0}},
                                       "I": {"O":{"dist": 3.7, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0},
                                             "N":{"dist": 3.73, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0},
                                             "S":{"dist": 3.98, "donor_angle_min": 130.0,
"donor_angle_max":180.0, "acceptor_angle_min":90.0, "acceptor_angle_max":150.0}}}
```

Figure A1. *The full complement of parameters that can be modified in ProPLID*

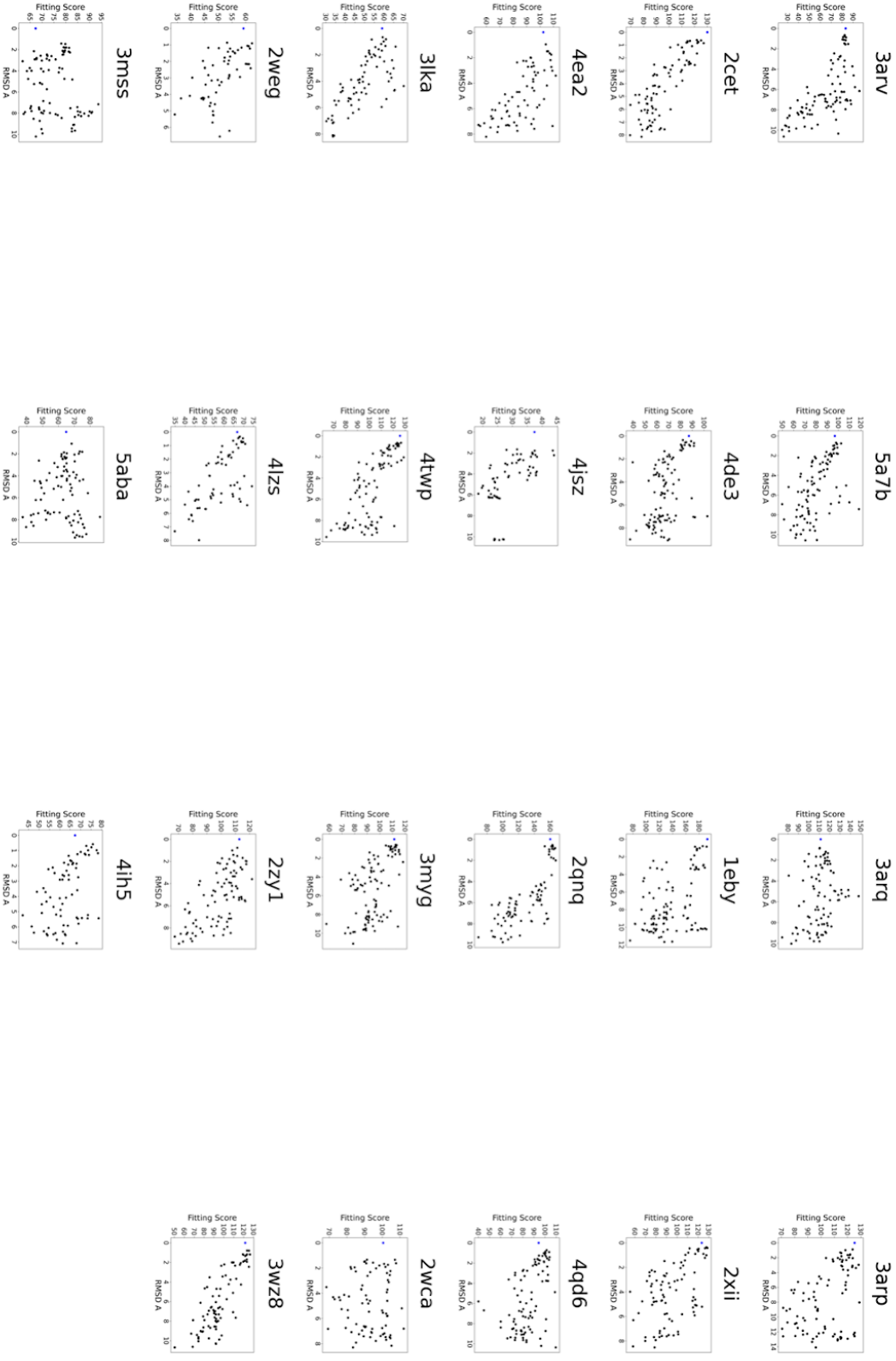


Figure A2 *Correlations of the Score 1 fitting score and the RMSD of decoy ligands to the reference ligand for each case in the test benchmark set. PDB ids are given above plots*

Figure A3. *Correlations of the Score 2 fitting score and the RMSD of decoy ligands to the reference ligand for each case in the test benchmark set. PDB ids are given above plots*

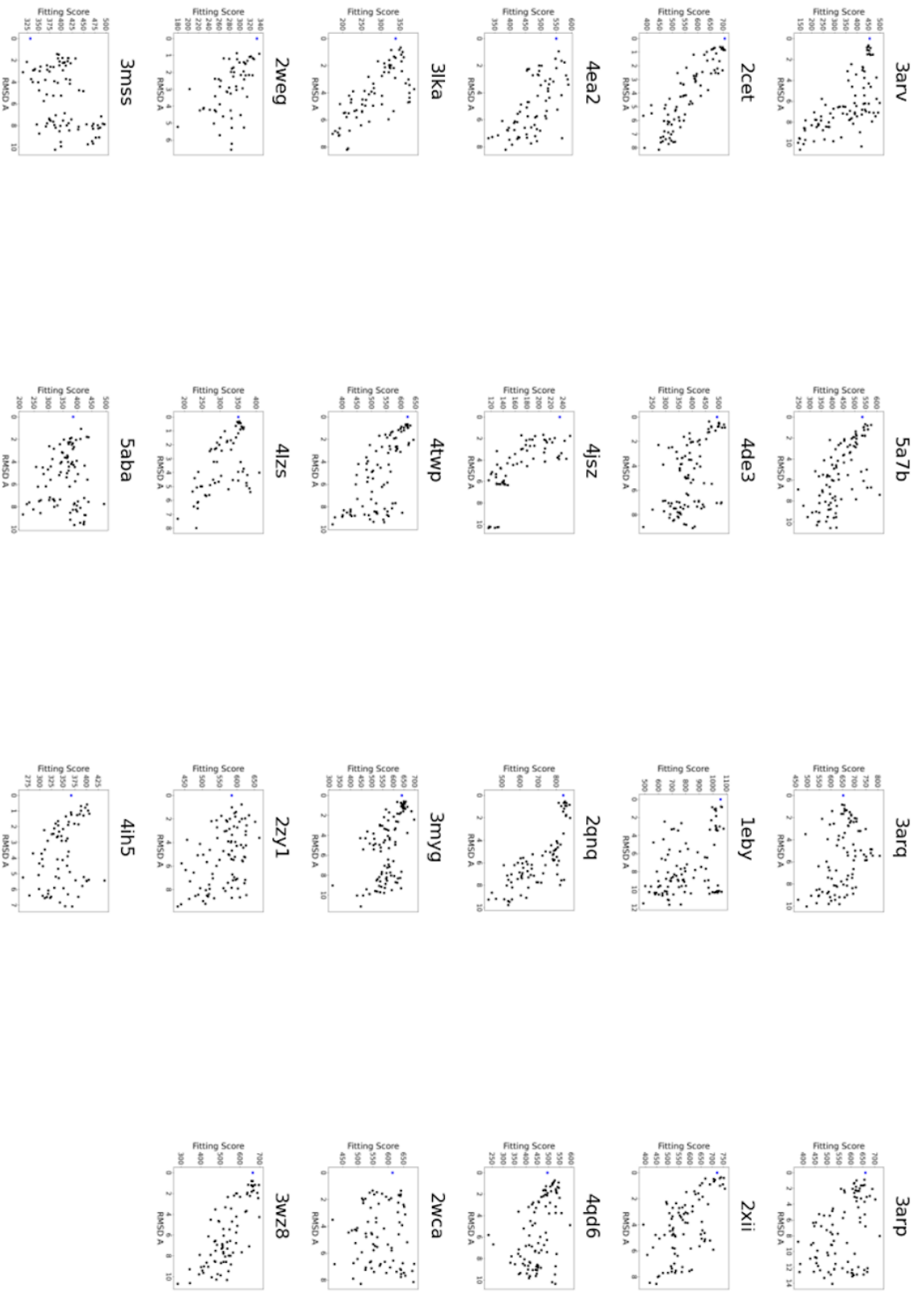


Figure A4. *Correlations of the Score 3 fitting score and the RMSD of decoy ligands to the reference ligand for each case in the test benchmark set. PDB ids are given above plots*

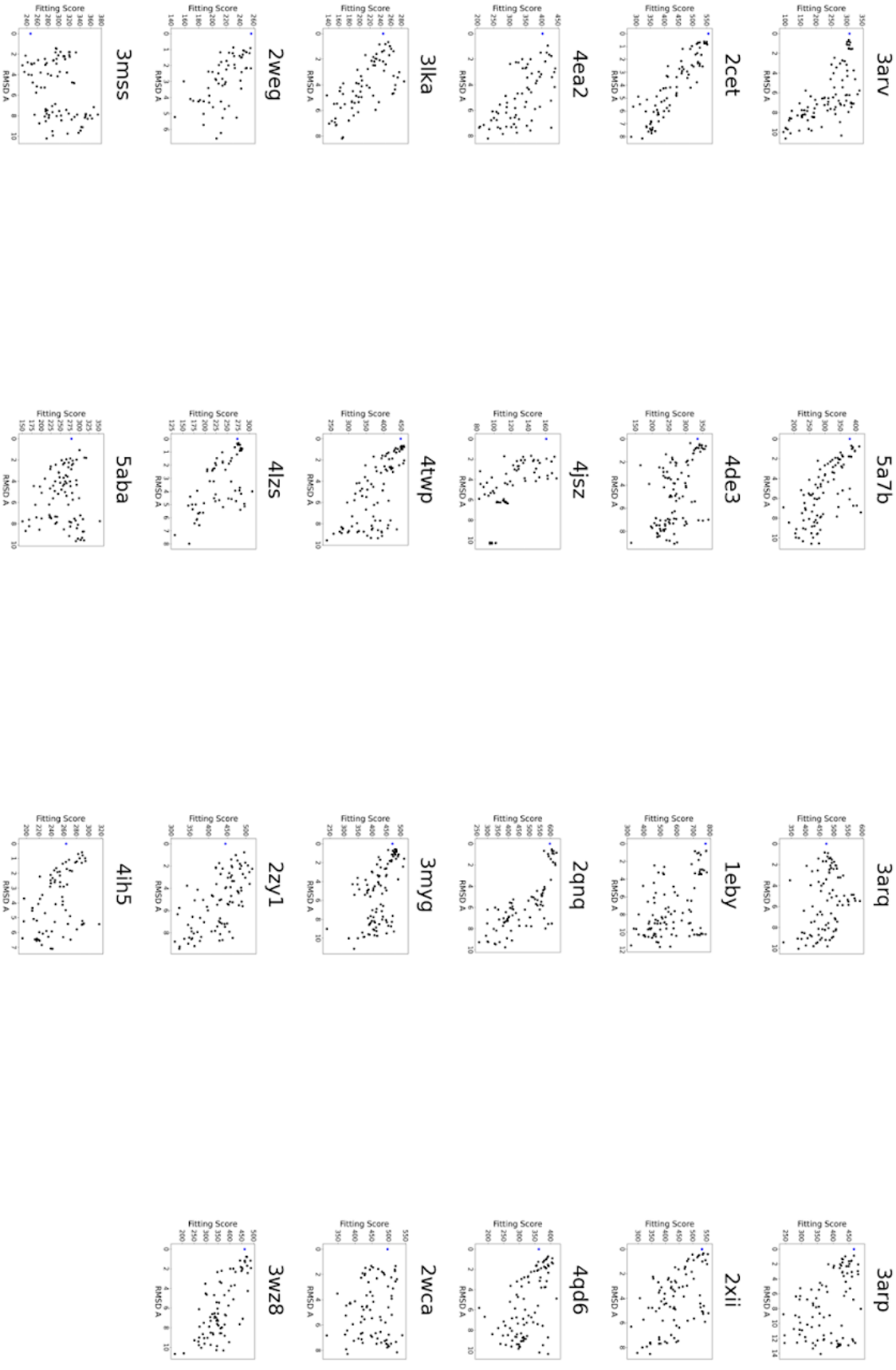


Figure A5. *Correlations of the Score 4 fitting score and the RMSD of decoy ligands to the reference ligand for each case in the test benchmark set. PDB ids are given above plots*

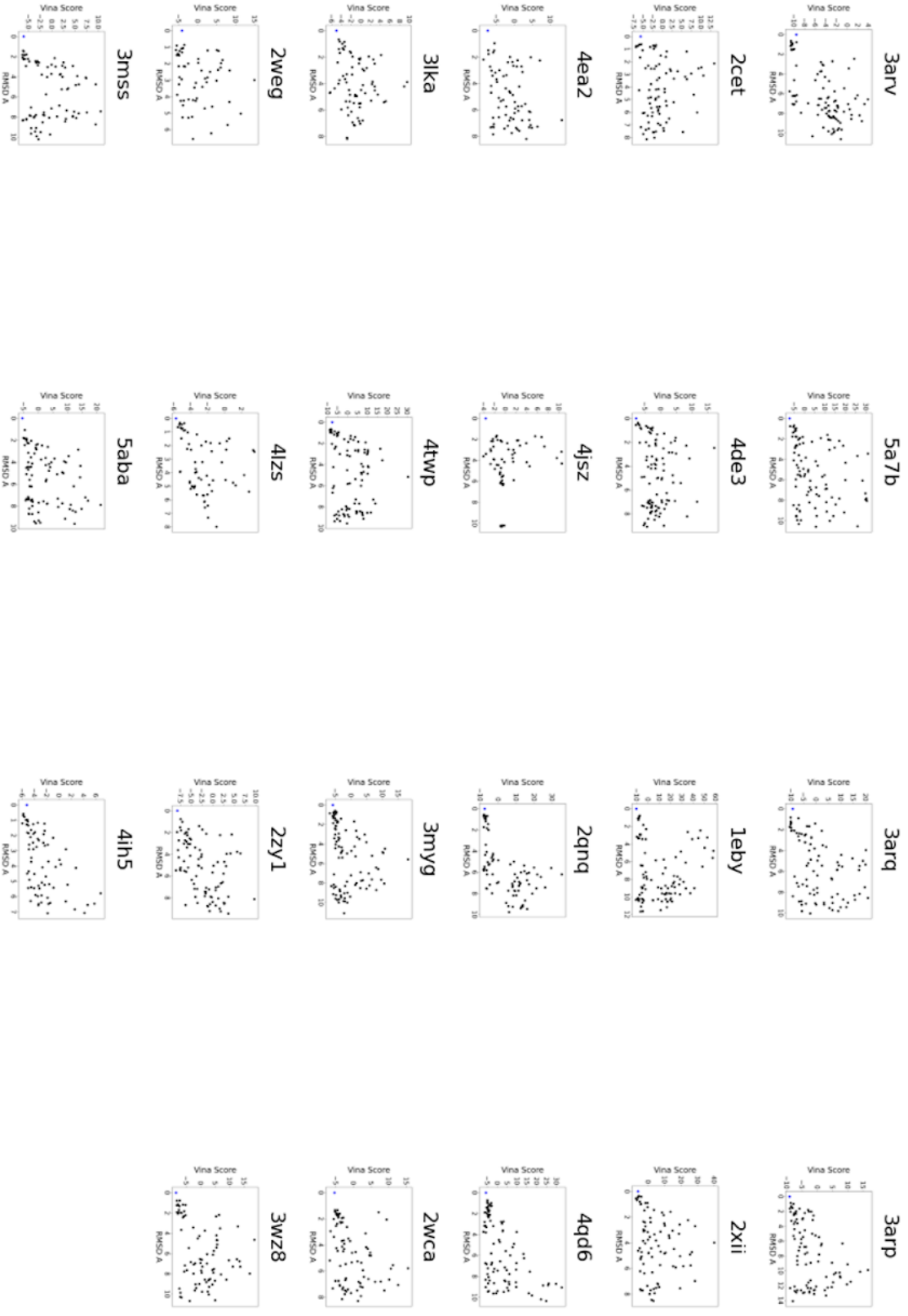


Figure A6. *Correlations of the AutoDock Vina score and the RMSD of decoy ligands to the reference ligand for each case in the test benchmark set. PDB ids are given above plots*

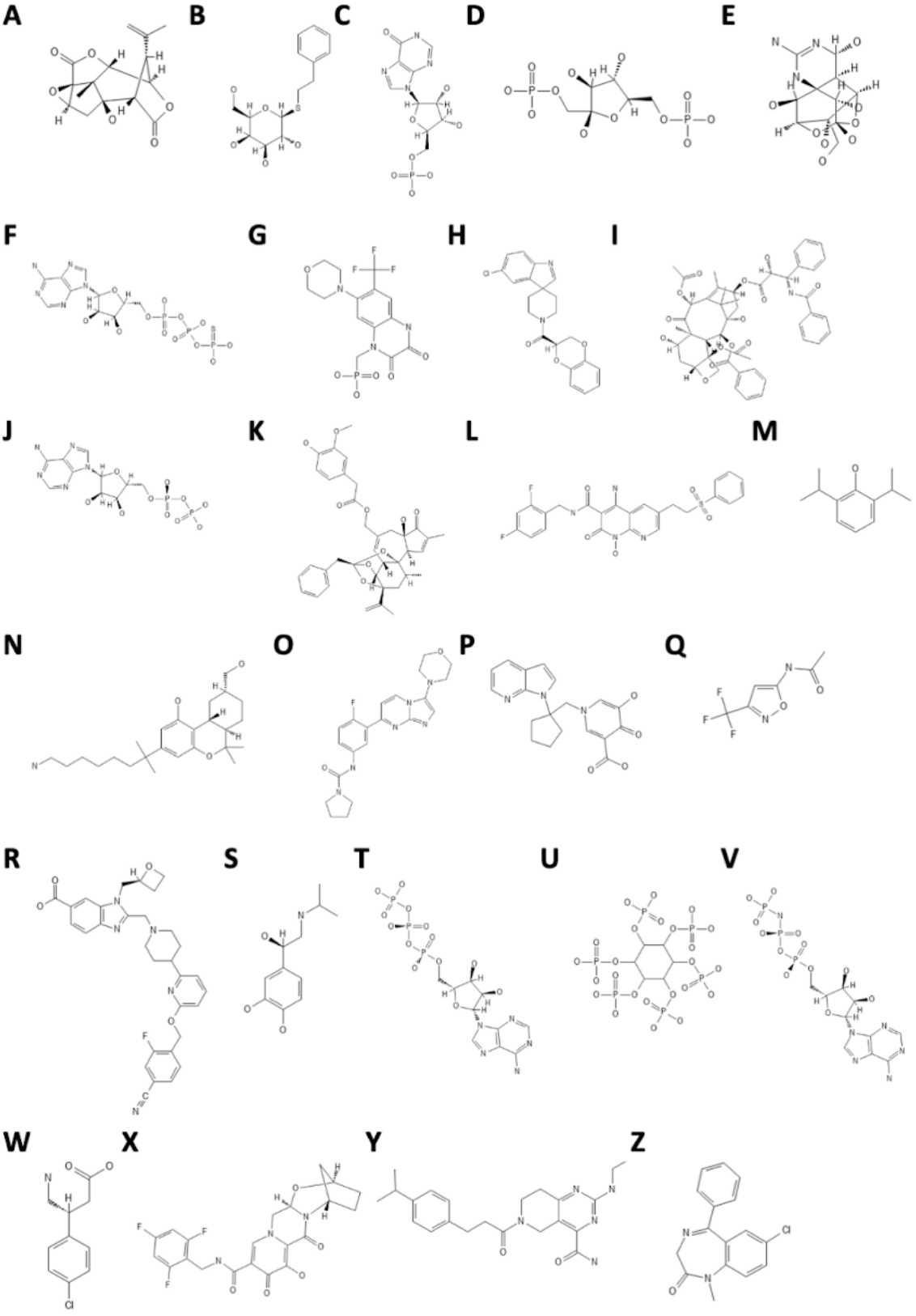
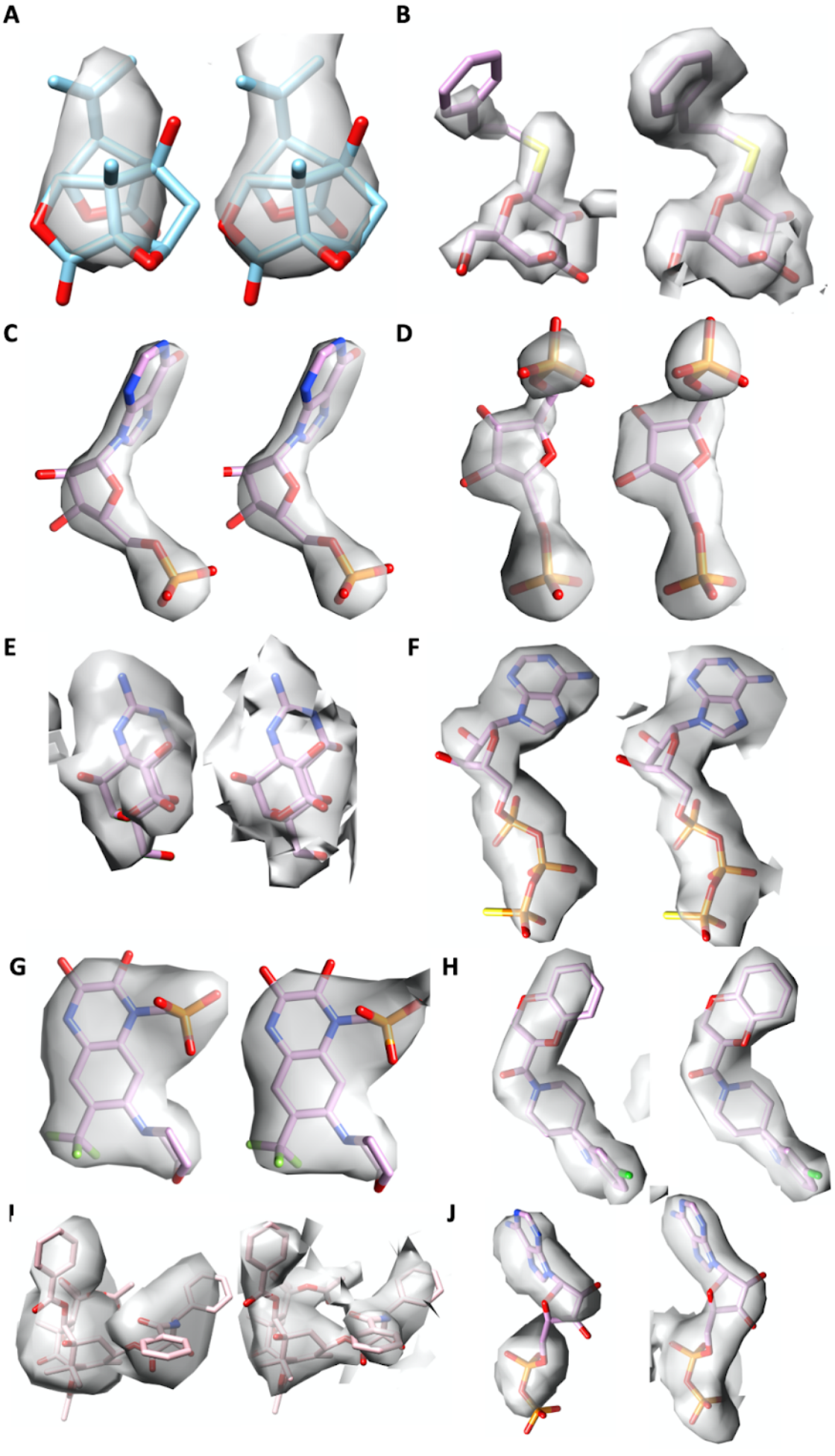
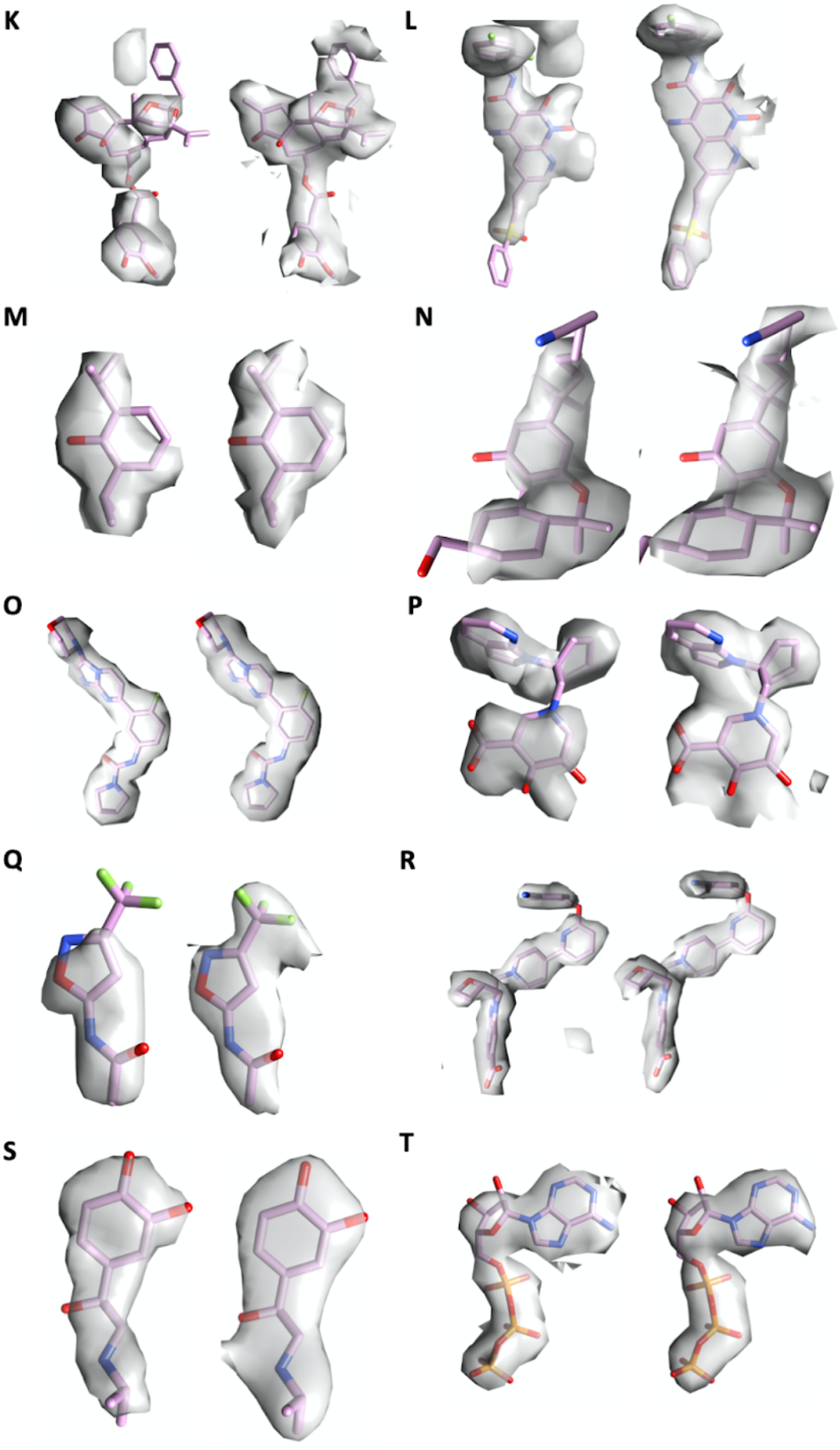


Figure A7*. Chemical schematics for ligands in the high resolution benchmark.* ***A****, PDB ID: 6X40 Ligand ID: RI5.* ***B****, PDB ID: 6TTE Ligand ID: PTQ.* ***C*** *PDB ID: 6UDP Ligand ID:*

IMP. ***D*** *PDB ID: 6TTQ Ligand ID: FBP.* ***E*** *PDB ID: 6A95 Ligand ID: 9SR .* ***F*** *PDB ID: 6OAX Ligand ID: AGS.* ***G*** *PDB ID: 6PEQ Ligand ID: ZK1.* ***H*** *PDB ID: 6UZ8 Ligand ID: R0D.* ***I*** *PDB ID: 6WVR Ligand ID: TA1.* ***J*** *PDB ID: 6UQE Ligand ID: ADP.* ***K*** *PDB ID: 6OO3 Ligand ID: 6EU.* ***L*** *PDB ID: 6PUZ Ligand ID: XXJ.* ***M*** *PDB ID: 6X3T Ligand ID: PFL.* ***N*** *PDB ID: 6KPF Ligand ID: 8D0.* ***O*** *PDB ID: 6QM7 Ligand ID: J6E.* ***P*** *PDB ID: 6TW1 Ligand ID: M4H.* ***Q*** *PDB ID: 6TTI Ligand ID: NXE.* ***R*** *PDB ID: 6X1A Ligand ID: UK4.* ***S*** *PDB ID: 7JJO Ligand ID: 5FW.* ***T*** *PDB ID: 6VFX Ligand ID: ATP.* ***U*** *PDB ID: 6REY Ligand ID: IHP.* ***V*** *PDB ID: 6NYY Ligand ID: ANP.* ***W*** *PDB ID: 7C7Q Ligand ID: 2C0.* ***X*** *PDB ID: 6PUW Ligand ID: KLQ.* ***Y*** *PDB ID: 7CFM Ligand ID: FWX.* ***Z*** *PDB ID: 6X3X Ligand ID: DZP*


A
B
C
D
E
F
G
H
I
J


K
L
M
N
O
P
Q
R
S
T

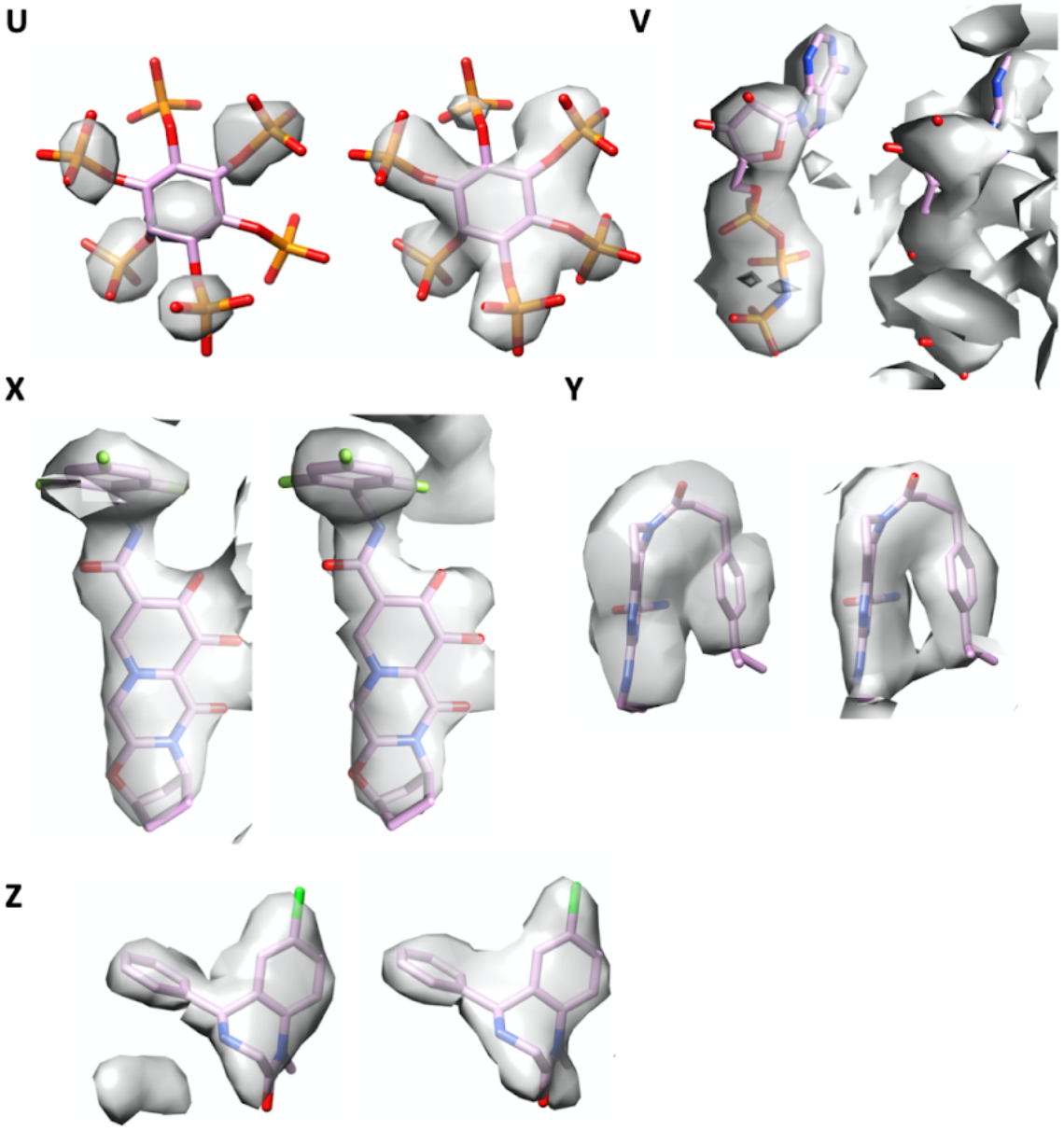
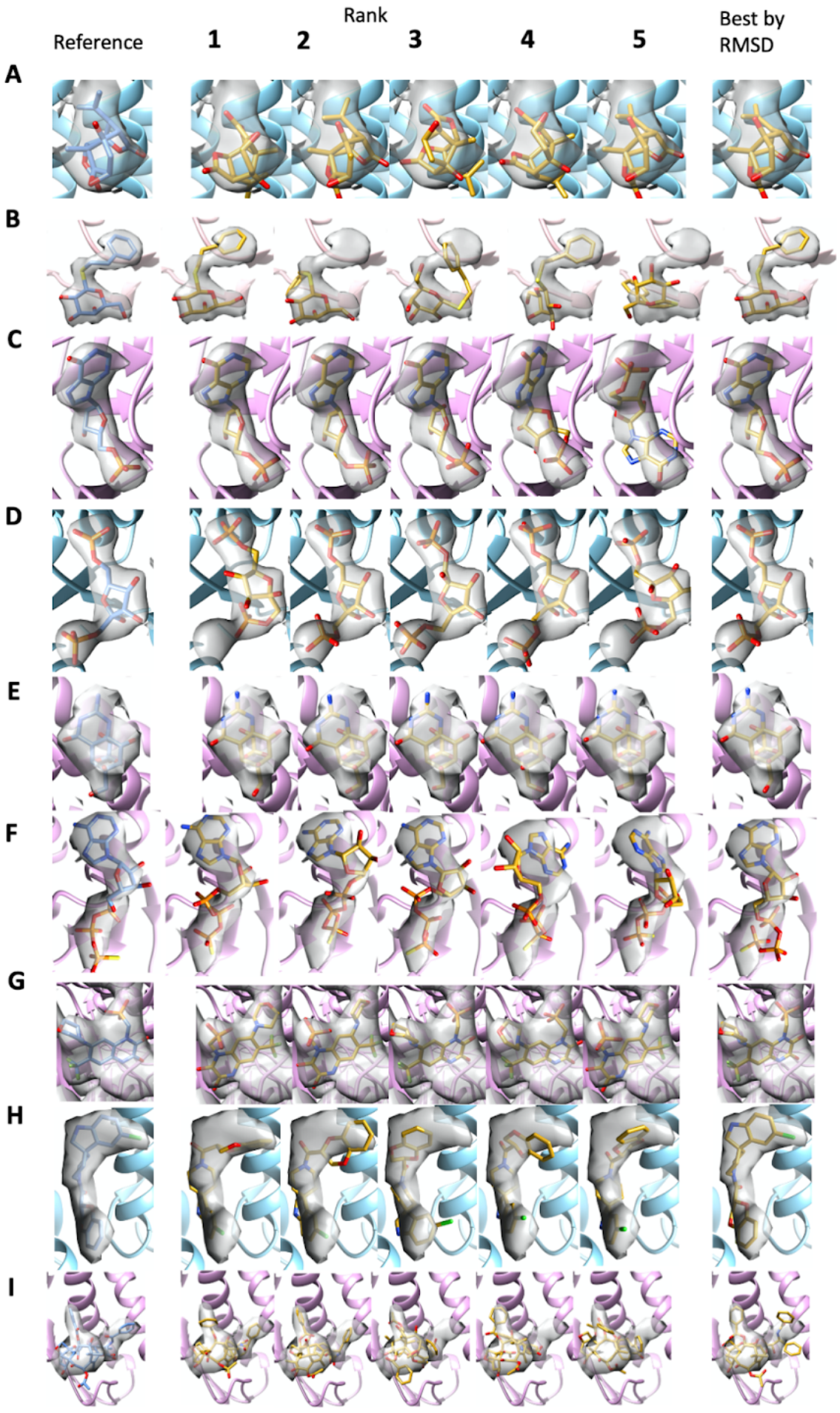
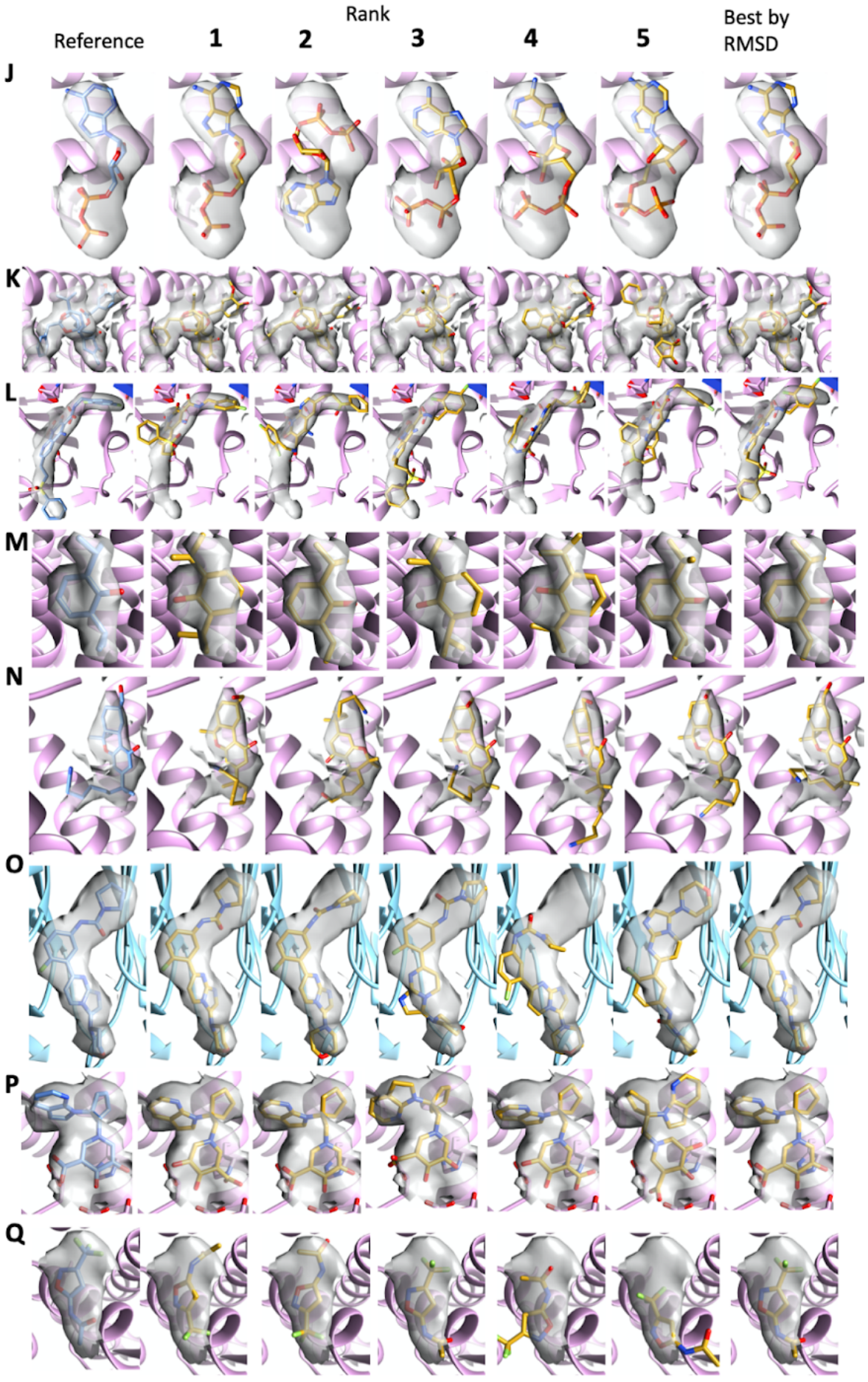


Figure A8 *The calculated density difference maps (left) with full maps (right) for each case in the high resolution benchmark.* ***A****, PDB ID: 6X40 Ligand ID: RI5.* ***B****, PDB ID: 6TTE Ligand ID: PTQ.* ***C*** *PDB ID: 6UDP Ligand ID: IMP.* ***D*** *PDB ID: 6TTQ Ligand ID: FBP.* ***E*** *PDB ID: 6A95 Ligand ID: 9SR .* ***F*** *PDB ID: 6OAX Ligand ID: AGS.* ***G*** *PDB ID: 6PEQ Ligand ID: ZK1.* ***H*** *PDB ID: 6UZ8 Ligand ID: R0D.* ***I*** *PDB ID: 6WVR Ligand ID: TA1.* ***J*** *PDB ID: 6UQE Ligand ID: ADP.* ***K*** *PDB ID: 6OO3 Ligand ID: 6EU.* ***L*** *PDB ID: 6PUZ Ligand ID: XXJ.* ***M*** *PDB ID: 6X3T Ligand ID: PFL.* ***N*** *PDB ID: 6KPF Ligand ID: 8D0.* ***O*** *PDB ID: 6QM7 Ligand ID: J6E.* ***P*** *PDB ID: 6TW1 Ligand ID: M4H.* ***Q*** *PDB ID: 6TTI Ligand ID: NXE.* ***R*** *PDB ID: 6X1A Ligand ID: UK4.* ***S*** *PDB ID: 7JJO Ligand ID: 5FW.* ***T*** *PDB ID: 6VFX Ligand ID: ATP.* ***U*** *PDB ID: 6REY Ligand ID: IHP.* ***V*** *PDB ID: 6NYY Ligand ID: ANP.* ***W*** *PDB ID: 7C7Q Ligand ID: 2C0.* ***X*** *PDB ID: 6PUW Ligand ID: KLQ.* ***Y*** *PDB ID: 7CFM Ligand ID: FWX.* ***Z*** *PDB ID: 6X3X Ligand ID: DZP*


Reference
Rank
1
2
3
4
5
Best by RMSD
A
B
C
D
E
F
G
H
I


Rank
Reference
1
2
3
4
5
Best by RMSD
J
K
L
M
N
O
P
Q

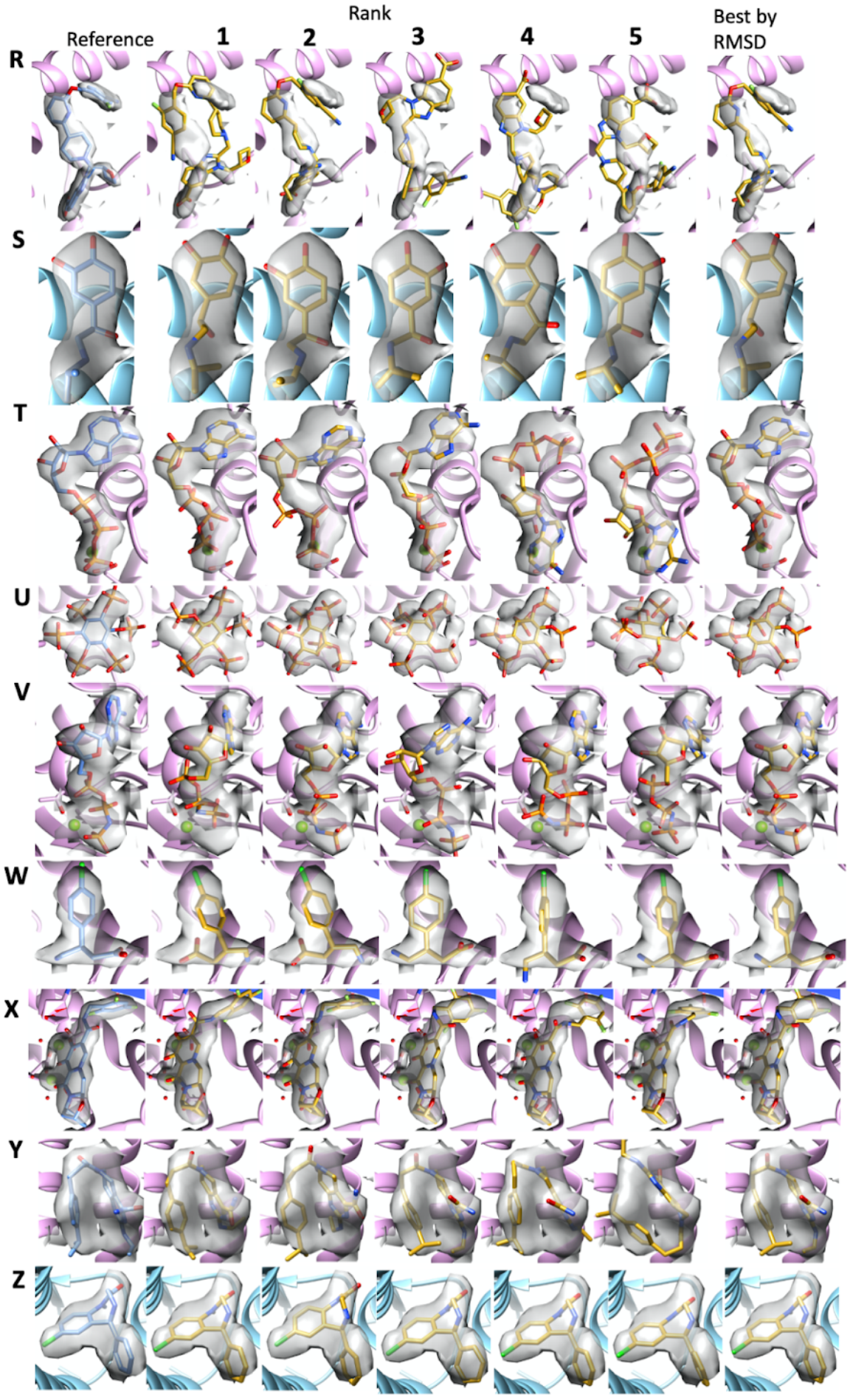

Rank
Reference
1
2
3
4
5
Best by RMSD
R
S
T
U
V
W
X
Y
Z

Figure A9. *A comparison of the deposited ligand conformation (left most, blue) with the top five solutions generated by the GA (yellow) in the high resolution benchmark. The solution that had the best RMSD with the deposited ligand is also shown (right most, yellow).* ***A****, PDB ID: 6X40 Ligand ID: RI5.* ***B****, PDB ID: 6TTE Ligand ID: PTQ.* ***C*** *PDB ID: 6UDP Ligand ID: IMP.* ***D*** *PDB ID: 6TTQ Ligand ID: FBP.* ***E*** *PDB ID: 6A95 Ligand ID: 9SR .* ***F*** *PDB ID: 6OAX Ligand ID: AGS.* ***G*** *PDB ID: 6PEQ Ligand ID: ZK1.* ***H*** *PDB ID: 6UZ8 Ligand ID: R0D.* ***I*** *PDB ID: 6WVR Ligand ID: TA1.* ***J*** *PDB ID: 6UQE Ligand ID: ADP.* ***K*** *PDB ID: 6OO3 Ligand ID: 6EU.* ***L*** *PDB ID: 6PUZ Ligand ID: XXJ.* ***M*** *PDB ID: 6X3T Ligand ID: PFL.* ***N*** *PDB ID: 6KPF Ligand ID: 8D0.* ***O*** *PDB ID: 6QM7 Ligand ID: J6E.* ***P*** *PDB ID: 6TW1 Ligand ID: M4H.* ***Q*** *PDB ID: 6TTI Ligand ID: NXE.* ***R*** *PDB ID: 6X1A Ligand ID: UK4.* ***S*** *PDB ID: 7JJO Ligand ID: 5FW.* ***T*** *PDB ID: 6VFX Ligand ID: ATP.* ***U*** *PDB ID: 6REY Ligand ID: IHP.* ***V*** *PDB ID: 6NYY Ligand ID: ANP.* ***W*** *PDB ID: 7C7Q Ligand ID: 2C0.* ***X*** *PDB ID: 6PUW Ligand ID: KLQ.* ***Y*** *PDB ID: 7CFM Ligand ID: FWX.* ***Z*** *PDB ID: 6X3X Ligand ID: DZP*

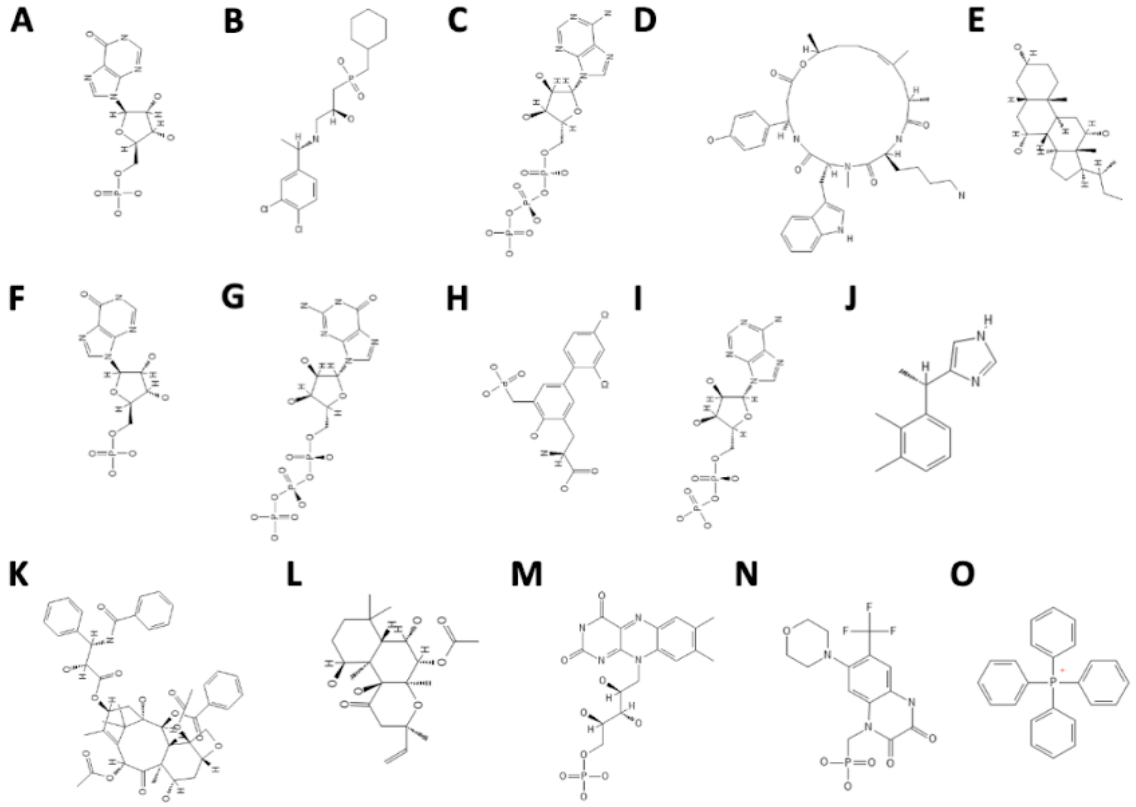
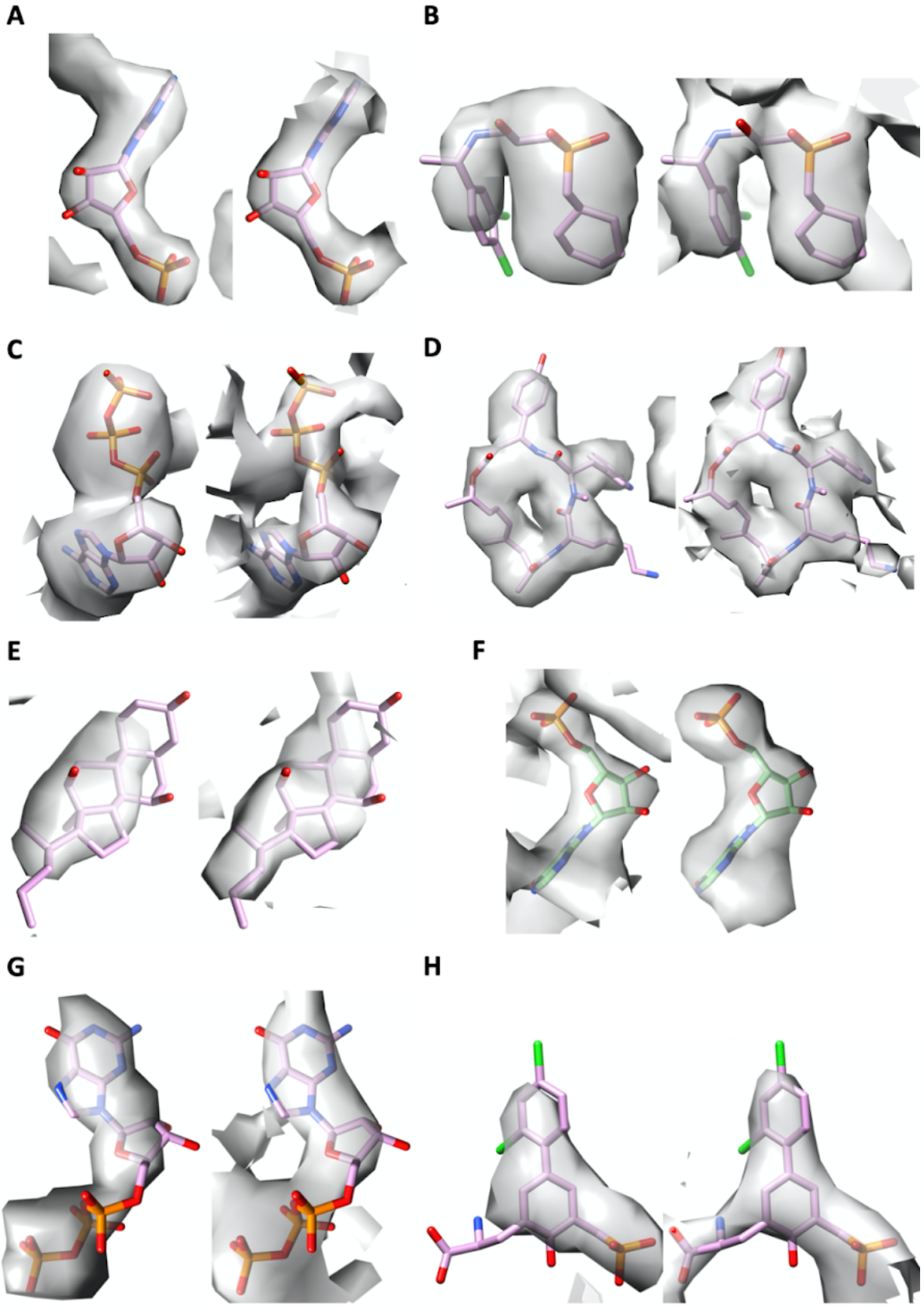


Figure A10. *The chemical schematics of ligands from the low resolution benchmark.* ***A****, PDB_ID: 6U8S Ligand ID: IMP.* ***B****, PDB_ID: 7CUM Ligand ID: 2BV.* ***C****, PDB_ID: 6IP2 Ligand ID: ATP.* ***D****, PDB_ID: 6T24 Ligand ID: 9ZK.* ***E****, PDB_ID: 6N57 Ligand ID: 1N7.* ***F****, PDB_ID: 6U8R Ligand ID: IMP.* ***G****, PDB_ID: 5W3J Ligand ID: GTP.* ***H****, PDB_ID: 6WHV Ligand ID: QGP.* ***I****, PDB_ID: 5OAF Ligand ID: ADP.* ***J****, PDB_ID: 6K42 Ligand ID: CZX.* ***K****, PDB_ID: 6RZB Ligand ID: TA1.* ***L****, PDB_ID: 6R4O Ligand ID: FOK.* ***M****, PDB_ID: 6ZIY Ligand ID: FMN.* ***N****, PDB_ID: 5WEL Ligand ID: ZK1.* ***O****, PDB_ID: 7CKQ Ligand ID: P4P.*


A
B
C
D
E
F
G
H

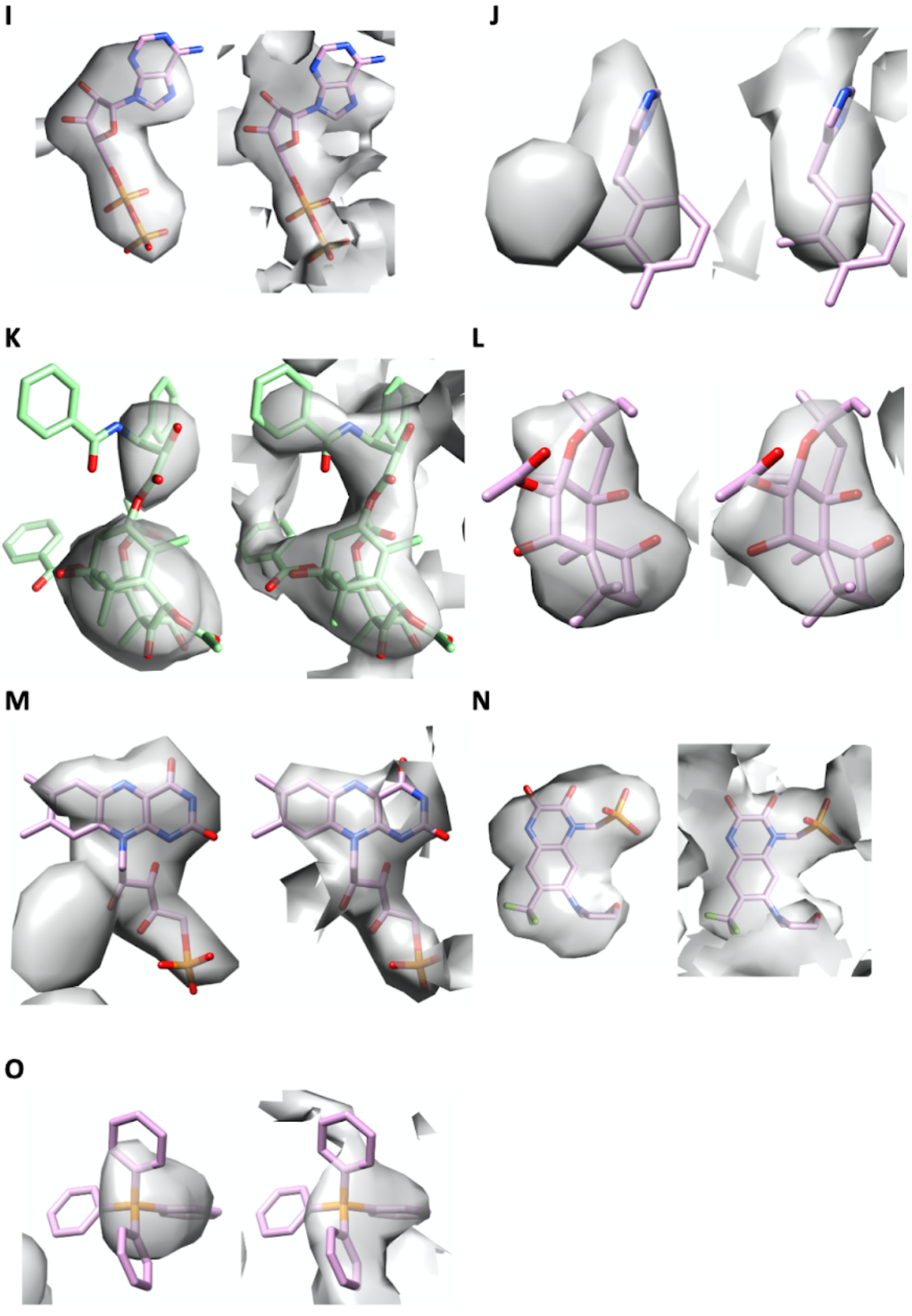
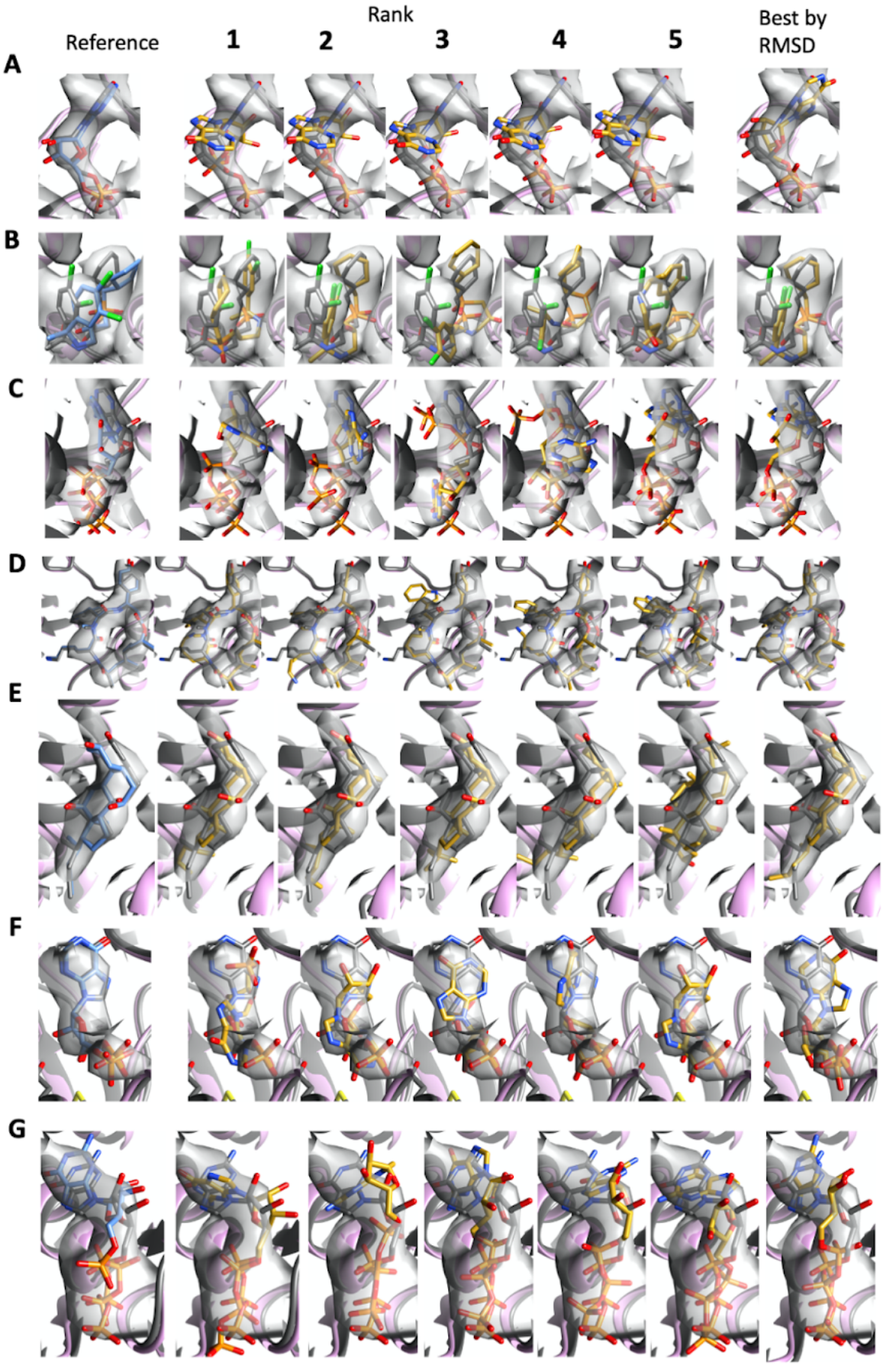

I
J
K
L
M
N
O

Figure A11. *The calculated density difference maps (left) with full maps (right) for each case in the high resolution benchmark.* ***A****, PDB_ID: 6U8S Ligand ID: IMP.* ***B****, PDB_ID: 7CUM Ligand ID: 2BV.* ***C****, PDB_ID: 6IP2 Ligand ID: ATP.* ***D****, PDB_ID: 6T24 Ligand ID: 9ZK.* ***E****, PDB_ID: 6N57 Ligand ID: 1N7.* ***F****, PDB_ID: 6U8R Ligand ID: IMP.* ***G****, PDB_ID: 5W3J Ligand ID: GTP.* ***H****, PDB_ID: 6WHV Ligand ID: QGP.* ***I****, PDB_ID: 5OAF Ligand ID: ADP.* ***J****, PDB_ID: 6K42 Ligand ID: CZX.* ***K****, PDB_ID: 6RZB Ligand ID: TA1.* ***L****, PDB_ID: 6R4O Ligand ID: FOK.* ***M****, PDB_ID: 6ZIY Ligand ID: FMN.* ***N****, PDB_ID: 5WEL Ligand ID: ZK1.* ***O****, PDB_ID: 7CKQ Ligand ID: P4P.*


Rank
Reference
1
2
3
4
5
Best by RMSD
A
B
C
D
E
F
G

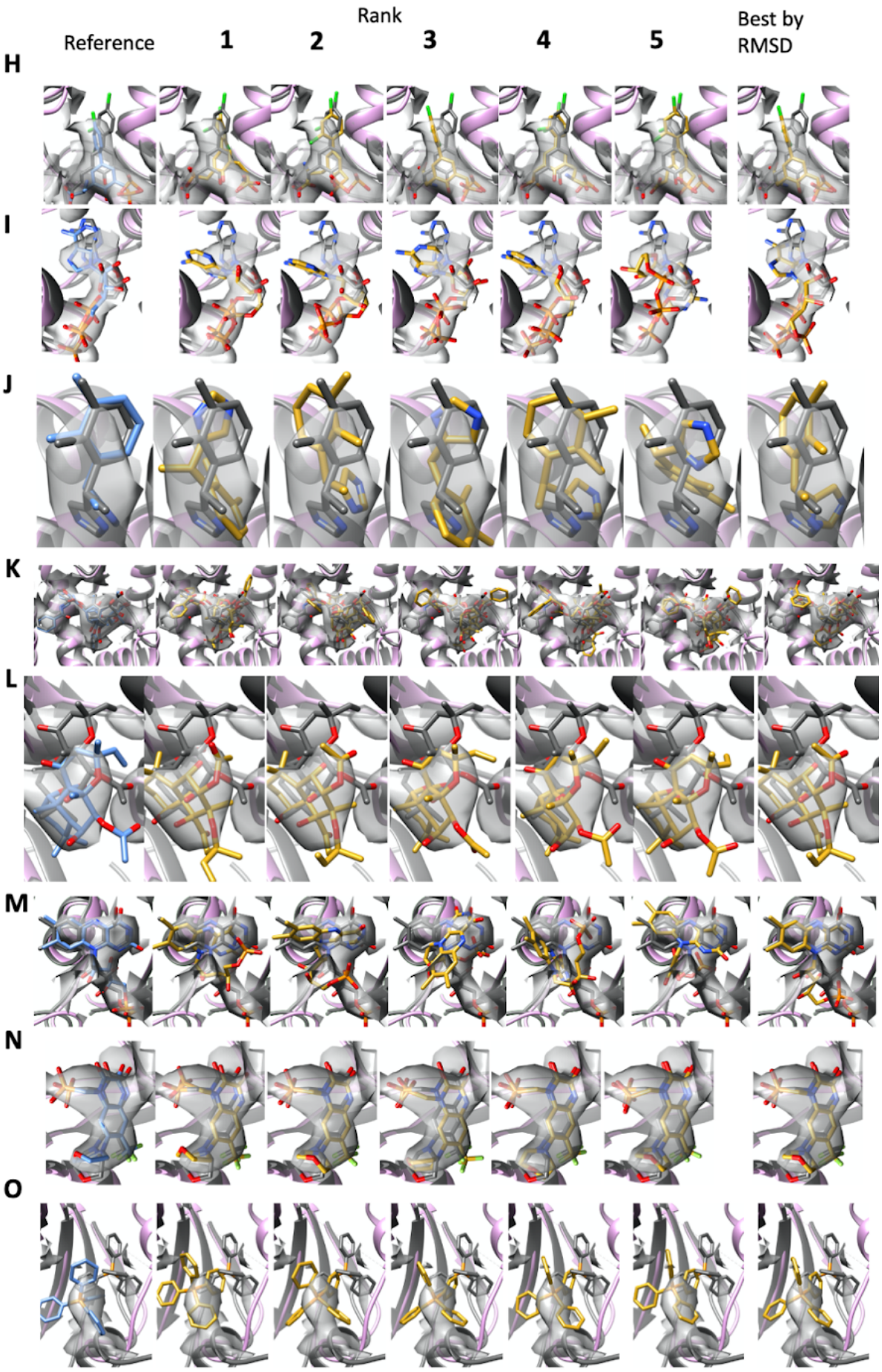

Rank
Reference
1
2
3
4
5
Best by RMSD
H
I
J
K
L
M
N
O

Figure A12 *A comparison of the deposited ligand conformation (left most, blue) with the top five solutions generated by the GA (yellow) in the lower resolution benchmark. The solution that had the best RMSD with the deposited ligand is also shown (right most, yellow). High resolutions control structures are structurally aligned in each image (dark grey).* ***A****, PDB_ID: 6U8S Ligand ID: IMP.* ***B****, PDB_ID: 7CUM Ligand ID: 2BV.* ***C****, PDB_ID: 6IP2 Ligand ID: ATP.* ***D****, PDB_ID: 6T24 Ligand ID: 9ZK.* ***E****, PDB_ID: 6N57 Ligand ID: 1N7.* ***F****, PDB_ID: 6U8R Ligand ID: IMP.* ***G****, PDB_ID: 5W3J Ligand ID: GTP.* ***H****, PDB_ID: 6WHV Ligand ID: QGP.* ***I****, PDB_ID: 5OAF Ligand ID: ADP.* ***J****, PDB_ID: 6K42 Ligand ID: CZX.* ***K****, PDB_ID: 6RZB Ligand ID: TA1.* ***L****, PDB_ID: 6R4O Ligand ID: FOK.* ***M****, PDB_ID: 6ZIY Ligand ID: FMN.* ***N****, PDB_ID: 5WEL Ligand ID: ZK1.* ***O****, PDB_ID: 7CKQ Ligand ID: P4P.*

Table A1 A table of the CCC, strain energies, and fitting scores of the reference ligand, top 5 conformations and solutions with the closest RMSD to the reference for each case in the high resolution benchmark.

PDB	Ligand ID	Rank	Fitting Score	CCC	Strain energy (TEU)	RMSD
6X40	RI5	Reference	-	0.130	0.650	0.0
		1	2301.46	0.125	1.056	3.92
		2	2286.0	0.128	1.04	0.94
		3	2276.87	0.125	1.74	3.82
		4	2257.57	0.129	1.47	3.86
		5	2194.60	0.156	1.23	0.925
		Best	2194.60	0.156	1.23	0.925
6TTE	PTQ	Reference	-	0.149	5.71	0.0
		1	904.73	0.134	6.18	1.32
		2	849.77	0.121	11.24	3.25
		3	818.10	0.119	11.16	3.13
		4	777.80	0.111	12.13	2.58
		5	775.60	0.106	15.18	5.48
		Best	904.73	0.134	6.18	1.32
6UDP	IMP	Reference	-	0.207	4.86	0.0
		1	1887.48	0.210	3.73	0.70
		2	1866.54	0.212	6.84	0.77
		3	1782.75	0.207	3.85	0.76
		4	1715.72	0.187	7.90	1.36
		5	1694.66	0.185	6.24	5.64
		Best	1887.48	0.210	3.73	0.704

6TTQ	FBP	Reference	-	0.124	1.95	0.0
		1	1296.55	0.122	3.28	0.47
		2	1295.87	0.116	3.78	0.67
		3	1295.45	0.105	2.50	0.79
		4	1290.01	0.106	3.69	0.68
		5	1272.51	0.118	3.82	0.71
		Best	1296.55	0.122	3.28	0.47
6A95	9SR	Reference	-	0.124	1.95	0.0
		1	1296.55	0.122	3.28	0.47
		2	1295.87	0.116	3.78	0.67
		3	1295.45	0.105	2.50	0.79
		4	1290.01	0.106	3.69	0.68
		5	1272.51	0.118	3.82	0.71
		Best	1296.55	0.122	3.28	0.47
6OAX	AGS	Reference	-	0.185	3.79	0.0
		1	1125.64	0.182	4.30	5.44
		2	1118.65	0.185	2.40	5.55
		3	1118.24	0.183	5.83	0.79
		4	1113.28	0.204	5.12	1.20
		5	1112.02	0.187	3.18	5.50
		Best	1118.24	0.183	5.83	0.79
6PEQ	ZK1	Reference	-	0.185	3.79	0.0
		1	1125.64	0.181	4.30	5.44
		2	1118.65	0.185	2.40	5.55
		3	1118.24	0.183	5.83	0.79
		4	1113.2	0.204	5.12	1.20

		5	1112.02	0.187	3.18	5.5
		Best	1118.2	0.183	5.83	0.79
6UZ8	R0D	Reference	-	0.111	1.81	0.0
		1	822.67	0.076	3.02	7.00
		2	817.39	0.086	4.89	7.08
		3	813.46	0.055	4.11	6.83
		4	807.41	0.063	4.69	6.59
		5	802.25	0.061	5.92	7.09
		Best	768.07	0.110	2.22	0.66
6WVR	TA1	Reference	-	0.162	12.05	0.0
		1	1384.04	0.133	46.42	2.82
		2	1384.04	0.134	48.05	2.83
		3	1262.26	0.116	39.08	6.14
		4	1257.97	0.124	34.04	2.40
		5	1244.80	0.090	46.56	3.93
		Best	1174.06	0.133	41.12	1.94
6UQE	ADP	Reference	-	0.403	5.35	0.0
		1	2302.45	0.398	8.27	0.87
		2	2231.58	0.400	10.24	1.30
		3	2127.68	0.394	4.09	2.46
		4	2078.69	0.372	7.03	2.45
		5	2069.42	0.385	5.51	1.68
		Best	2302.45	0.398	8.27	0.87
6OO3	6EU	Reference	-	0.112	18.46	0.0
		1	1187.77	0.082	16.91	1.42

		2	1125.38	0.073	19.27	1.47
		3	1084.09	0.065	17.18	2.26
		4	1064.64	0.057	24.62	5.50
		5	1054.18	0.052	26.12	4.31
		Best	1187.77	0.082	16.91	1.42
6PUZ	XXJ	Reference	-	0.178	7.43	0.0
		1	1040.01	0.140	18.56	4.09
		2	1027.96	0.136	17.93	8.63
		3	1027.65	0.145	17.54	1.86
		4	1017.87	0.125	16.55	8.81
		5	1016.58	0.123	13.26	3.23
		Best	1027.65	0.145	17.54	1.86
6X3T	PFL	Reference	-	0.091	0.618	0.0
		1	784.25	0.062	5.15	4.53
		2	781.68	0.082	1.44	0.72
		3	771.69	0.073	1.00	2.21
		4	767.94	0.046	5.60	2.27
		5	764.13	0.066	1.28	0.95
		Best	781.68	0.082	1.44	0.72
6KPF	8D0	Reference	-	0.117	13.24	0.0
		1	1370.67	0.129	15.91	1.52
		2	1331.40	0.116	18.67	7.00
		3	1330.19	0.111	12.06	1.18
		4	1305.65	0.117	14.96	2.12
		5	1286.58	0.079	18.17	2.21
		Best	1250.4	0.124	23.63	1.03

6QM7	J6E	Reference	-	0.084	6.25	0.0
		1	793.11	0.082	7.27	0.641
		2	777.68	0.074	11.16	0.881
		3	738.98	0.069	7.01	1.89
		4	682.81	0.042	14.9	2.51
		5	682.4	0.059	8.89	8.25
		Best	793.11	0.082	7.27	0.641
6TW1	M4H	Reference	-	0.179	3.91	0.0
		1	886.40	0.218	5.02	1.99
		2	884.95	0.212	3.13	1.15
		3	871.6	0.180	5.67	1.58
		4	864.72	0.207	4.86	2.05
		5	849.55	0.209	6.97	3.26
		Best	884.95	0.212	3.13	1.15
6TTI	NXE	Reference	-	0.281	1.12	0.0
		1	1938.59	0.280	7.55	4.62
		2	1870.21	0.251	3.93	4.61
		3	1669.67	0.242	6.21	1.46
		4	1615.85	0.318	6.71	5.16
		5	1615.17	0.204	3.30	3.92
		Best	1669.67	0.242	6.21	1.46
6X1A	UK4	Reference	-	0.168	12.77	0.0
		1	965.34	0.062	12.86	3.96
		2	955.34	0.089	18.17	2.46
		3	934.96	0.061	13.46	9.32

		4	921.89	0.052	14.57	9.23
		5	913.28	0.048	10.87	8.19
		Best	955.34	0.089	18.17	2.46
7JJO	5FW	Reference	-	0.290	5.10	0.0
		1	1830.03	0.306	6.15	1.02
		2	1820.91	0.308	8.82	0.78
		3	1702.36	0.302	5.46	1.51
		4	1672.04	0.296	2.72	1.65
		5	1660.18	0.274	4.15	1.40
		Best	1820.91	0.308	8.82	0.78
6VFX	ATP	Reference	-	0.182	7.72	0.0
		1	1311.00	0.162	5.73	1.04
		2	1242.66	0.128	9.99	1.52
		3	1240.25	0.164	8.25	1.33
		4	1234.93	0.124	11.79	7.48
		5	1212.43	0.137	4.93	7.64
		Best	1311.00	0.162	5.73	1.04
6REY	IHP	Reference	-	0.131	7.34	0.0
		1	1131.91	0.098	8.39	1.94
		2	1123.76	0.114	10.57	1.70
		3	1102.87	0.090	8.63	2.63
		4	1099.60	0.091	2.52	2.02
		5	1098.12	0.101	6.99	1.79
		Best	1123.76	0.114	10.57	1.70
6NYY	ANP	Reference	-	0.131	7.34	0.0

		1	1131.91	0.098	8.39	1.94
		2	1123.76	0.114	10.57	1.70
		3	1102.87	0.090	8.63	2.63
		4	1099.6	0.091	2.52	2.02
		5	1098.12	0.101	6.99	1.79
		Best	1123.76	0.114	10.57	1.70
7C7Q	2C0	Reference	-	0.056	5.20	0.0
		1	1277.70	0.055	9.25	2.31
		2	1273.90	0.048	6.95	2.32
		3	1273.13	0.073	5.14	0.73
		4	1268.27	0.061	5.70	1.11
		5	1264.86	0.057	3.27	0.65
		Best	1264.86	0.057	3.27	0.65
6PUW	KLQ	Reference	-	0.241	2.16	0.0
		1	1478.84	0.219	10.90	1.26
		2	1473.19	0.233	9.56	1.00
		3	1419.65	0.218	8.93	0.95
		4	1381.59	0.240	5.63	1.08
		5	1341.79	0.231	9.74	1.12
		Best	1419.65	0.218	8.93	0.95
7CFM	FWX	Reference	-	0.135	10.74	0.0
		1	1267.11	0.117	16.18	1.54
		2	1225.47	0.120	14.74	1.16
		3	1200.88	0.112	15.66	1.34
		4	1194.64	0.092	11.00	5.21
		5	1185.05	0.103	12.86	1.33

		Best	1225.47	0.120	14.74	1.16
6X3X	DZP	Reference	-	0.095	0.966	0.0
		1	1322.29	0.103	0.85	0.59
		2	1310.19	0.103	0.17	1.17
		3	1300.58	0.107	0.61	0.90
		4	1291.29	0.105	1.08	0.57
		5	1283.06	0.091	1.79	1.17
		Best	1291.29	0.105	1.08	0.57

Table A2. A table fo the CCC, strain energies, and fitting scores of the reference ligand, top 5 conformations and solutions with the closest RMSD to the reference for each case in the low resolution benchmark. High resolution control structures are also indicated.

PDB	Ligand ID	Control PDB	Rank	Fitting Score	CCC	Strain energy (TEU)	RMSD
6U8S	IMP	6UDP	Reference	-	0.104	6.14	0.0
			1	1022.76	0.012	7.59	4.30
			2	1019.95	0.026	7.43	4.25
			3	1007.95	0.014	6.22	3.93
			4	1002.15	0.017	3.24	3.87
			5	1000.41	0.025	6.19	4.00
			Best	912.5	0.075	7.22	1.42
7CUM	2BV	7C7S	Reference	-	0.126	5.42	0.0
			1	1017.55	0.134	10.61	5.49
			2	994.13	0.158	8.85	2.57
			3	976.41	0.152	8.86	3.58
			4	972.00	0.150	8.46	3.51
			5	970.13	0.153	21.63	4.93
			Best	994.13	0.158	8.85	2.57

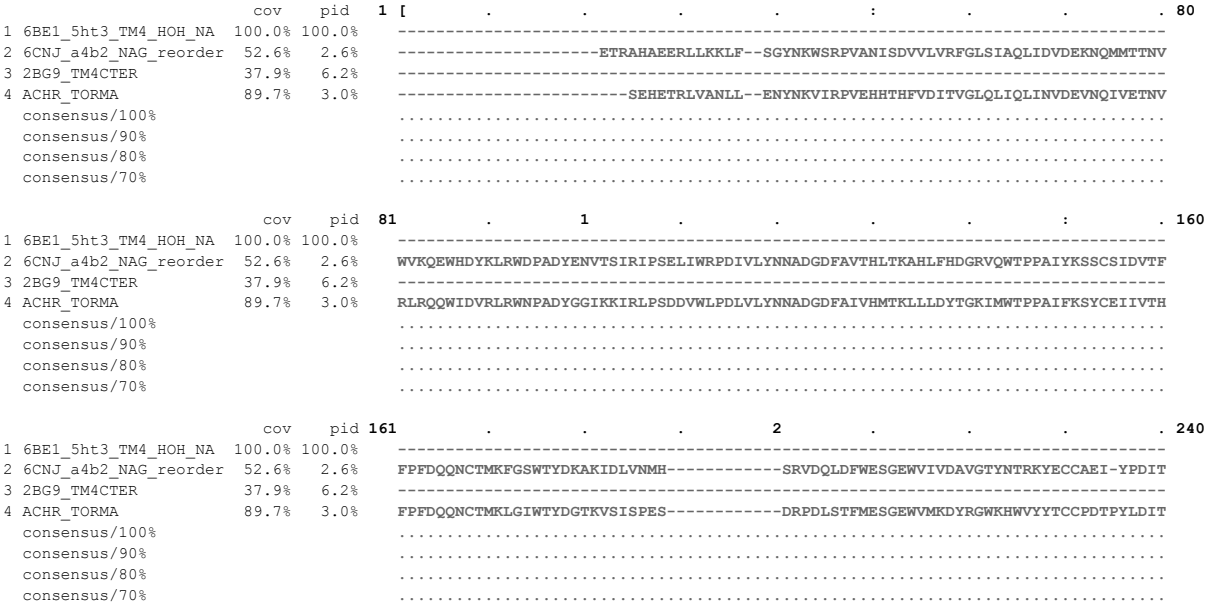
6IP2	ATP	2ZAN	Reference	-	0.108	5.69	0.0
			1	861.07	0.097	11.26	3.98
			2	856.34	0.104	11.62	4.11
			3	854.11	0.066	9.62	6.47
			4	853.26	0.064	6.84	6.39
			5	847.85	0.099	8.64	1.76
			Best	847.85	0.099	8.64	1.76
6T24	9ZK	6T23	Reference	-	0.067	8.31	0.0
			1	812.40	0.057	21.21	1.35
			2	761.72	0.056	22.84	1.53
			3	736.23	0.050	19.32	1.48
			4	727.81	0.051	18.13	1.85
			5	722.28	0.038	20.8	2.09
			Best	812.40	0.057	21.21	1.35
6N57	1N7	6PST	Reference	-	0.108	5.65	0.0
			1	752.27	0.110	6.415	1.44
			2	747.70	0.103	9.02	1.32
			3	742.34	0.105	9.68	1.42
			4	735.7	0.106	10.52	1.63
			5	734.4	0.083	7.43	6.37
			Best	733.76	0.107	6.71	1.30
6U8R	IMP	6UDP	Reference	-	0.082	0.71	0.0
			1	956.48	0.032	6.58	5.58
			2	954.70	0.024	4.54	5.19
			3	952.34	0.013	5.82	3.92

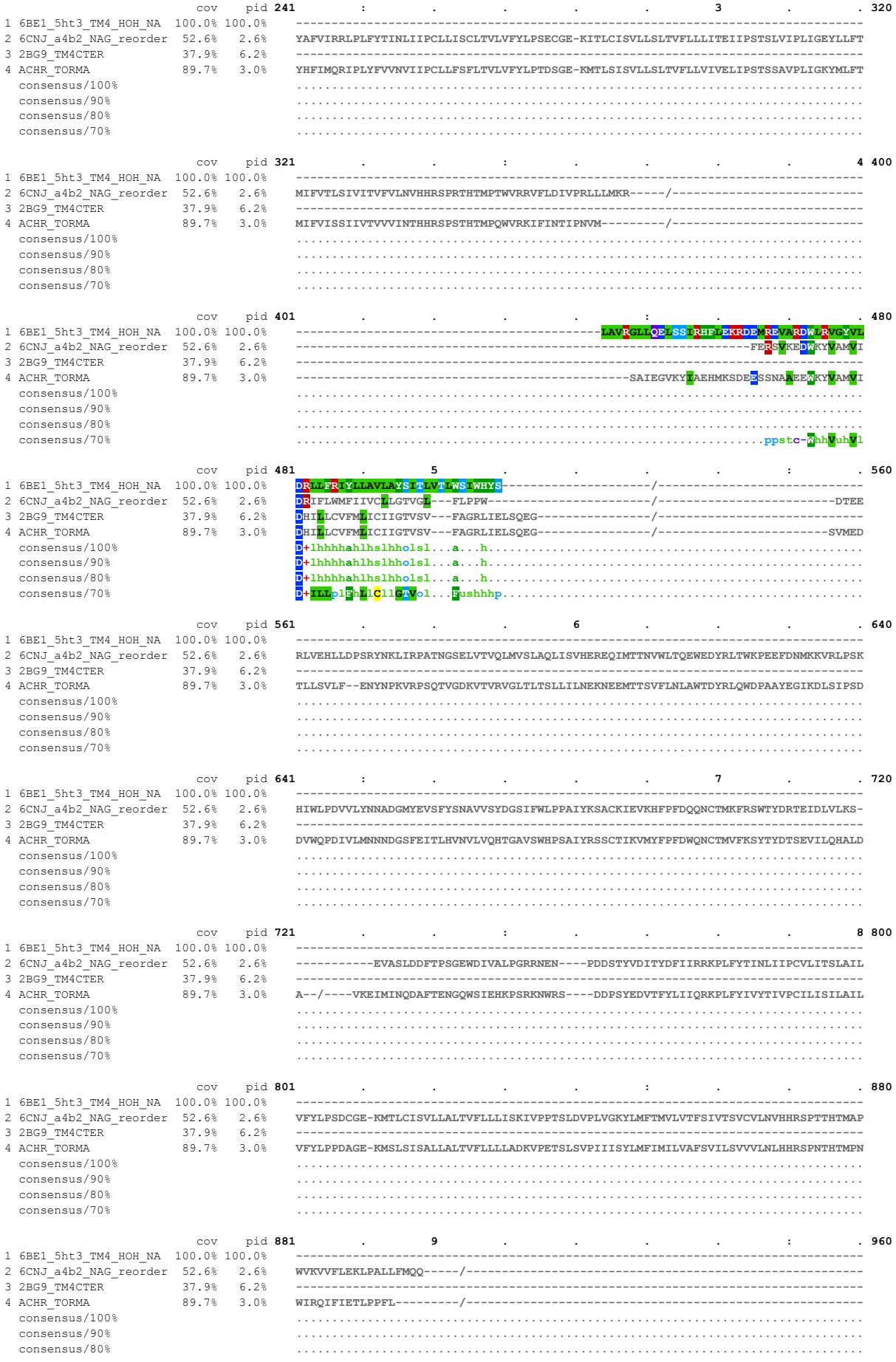
			4	951.37	0.034	0.12	2.41
			5	951.34	0.020	4.23	5.97
			Best	910.92	0.060	3.89	1.63
5W3J	GTP	4U3J	Reference	-	0.285	6.03	0.0
			1	2568.32	0.325	7.95	2.64
			2	2525.3	0.221	8.79	2.75
			3	2519.77	0.321	9.57	2.69
			4	2519.3	0.178	9.57	2.41
			5	2506.17	0.295	7.22	3.27
			Best	2473.40	0.298	10.35	1.66
6WVH	QGP	6USV	Reference	-	0.126	8.27	0.0
			1	950.22	0.002	9.28	4.81
			2	938.01	0.067	9.69	2.12
			3	934.78	0.094	9.30	1.31
			4	934.60	0.034	7.16	4.85
			5	930.43	0.070	7.78	2.33
			Best	934.78	0.094	9.30	1.31
5OAF	ADP	2C9O	Reference	-	0.100	1.55	0.0
			1	970.51	0.086	7.98	2.25
			2	970.34	0.057	9.10	3.04
			3	967.8	0.065	6.31	2.69
			4	965.36	0.083	5.17	1.67
			5	963.63	0.048	7.73	5.55
			Best	908.02	0.0984	10.25	1.49
6K42	CZX	6k41	Reference	-	0.136	1.66	0.0

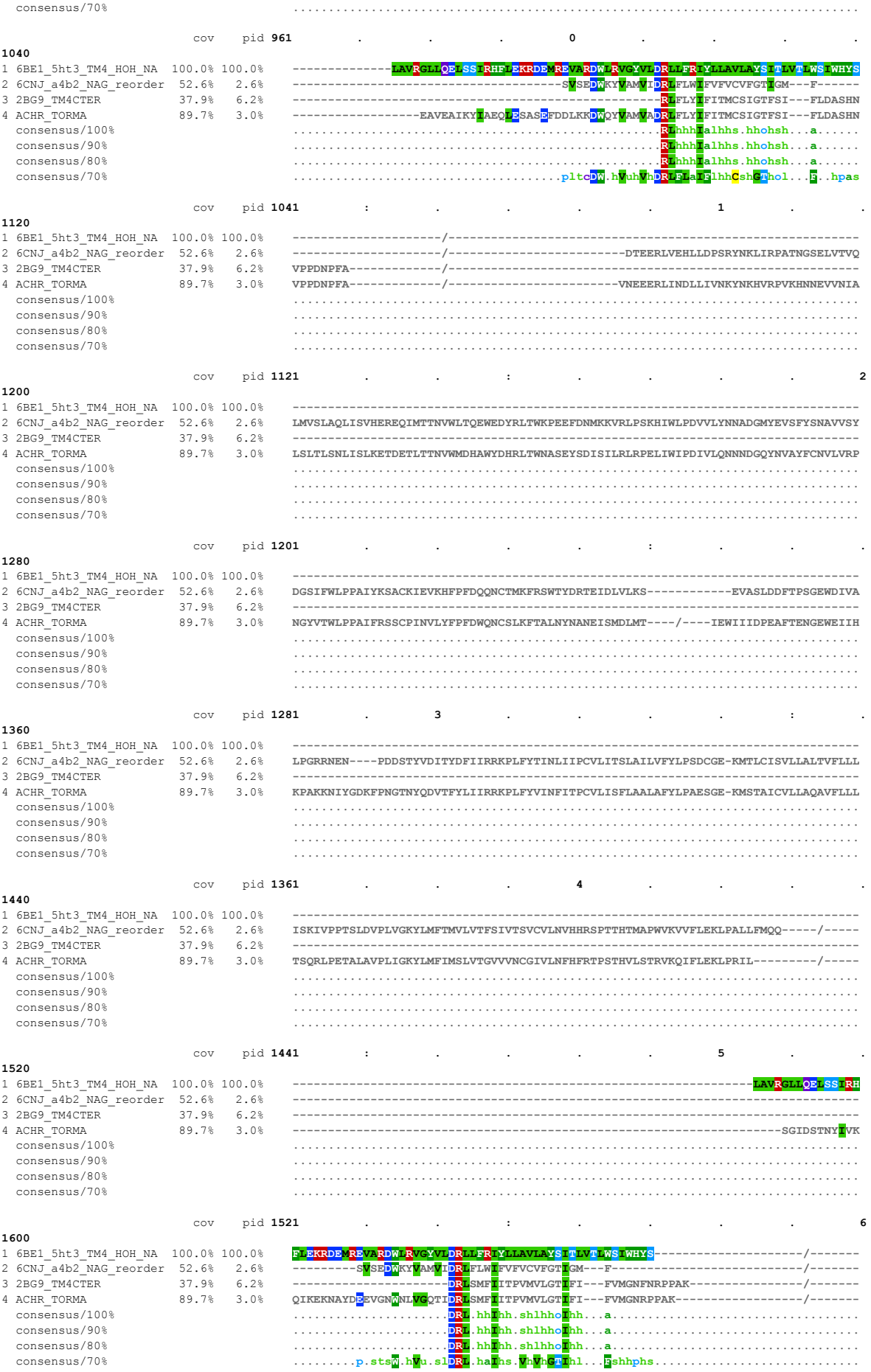
			1	1345.80	0.119	1.44	4.64
			2	1304.45	0.092	0.71	1.53
			3	1303.72	0.125	0.86	5.12
			4	1291.64	0.136	1.20	1.81
			5	1288.51	0.058	1.73	4.77
			Best	1242.38	0.140	0.70	1.42
6RZB	TA1	6WVR	Reference	-	0.255	10.52	0.0
			1	1390.30	0.229	38.24	5.02
			2	1359.60	0.243	39.41	5.42
			3	1350.22	0.236	43.70	5.30
			4	1338.47	0.223	30.63	2.84
			5	1337.99	0.225	40.12	6.00
			Best	1307.92	0.268	43.03	2.47
6R4O	FOK	1CUL	Reference	-	0.165	6.97	0.0
			1	912.48	0.183	10.42	4.39
			2	905.55	0.185	10.39	4.38
			3	898.86	0.170	3.80	0.77
			4	894.65	0.174	7.07	0.96
			5	891.63	0.154	6.86	0.80
			Best	898.8	0.170	3.80	0.77
6ZIY	FMN	3I9V	Reference	-	0.075	11.53	0.0
			1	1066.60	0.069	5.84	2.75
			2	1065.01	0.070	14.7	2.89
			3	1060.74	0.079	15.43	4.08
			4	1060.71	0.073	9.93	4.80
			5	1047.42	0.074	13.31	2.60

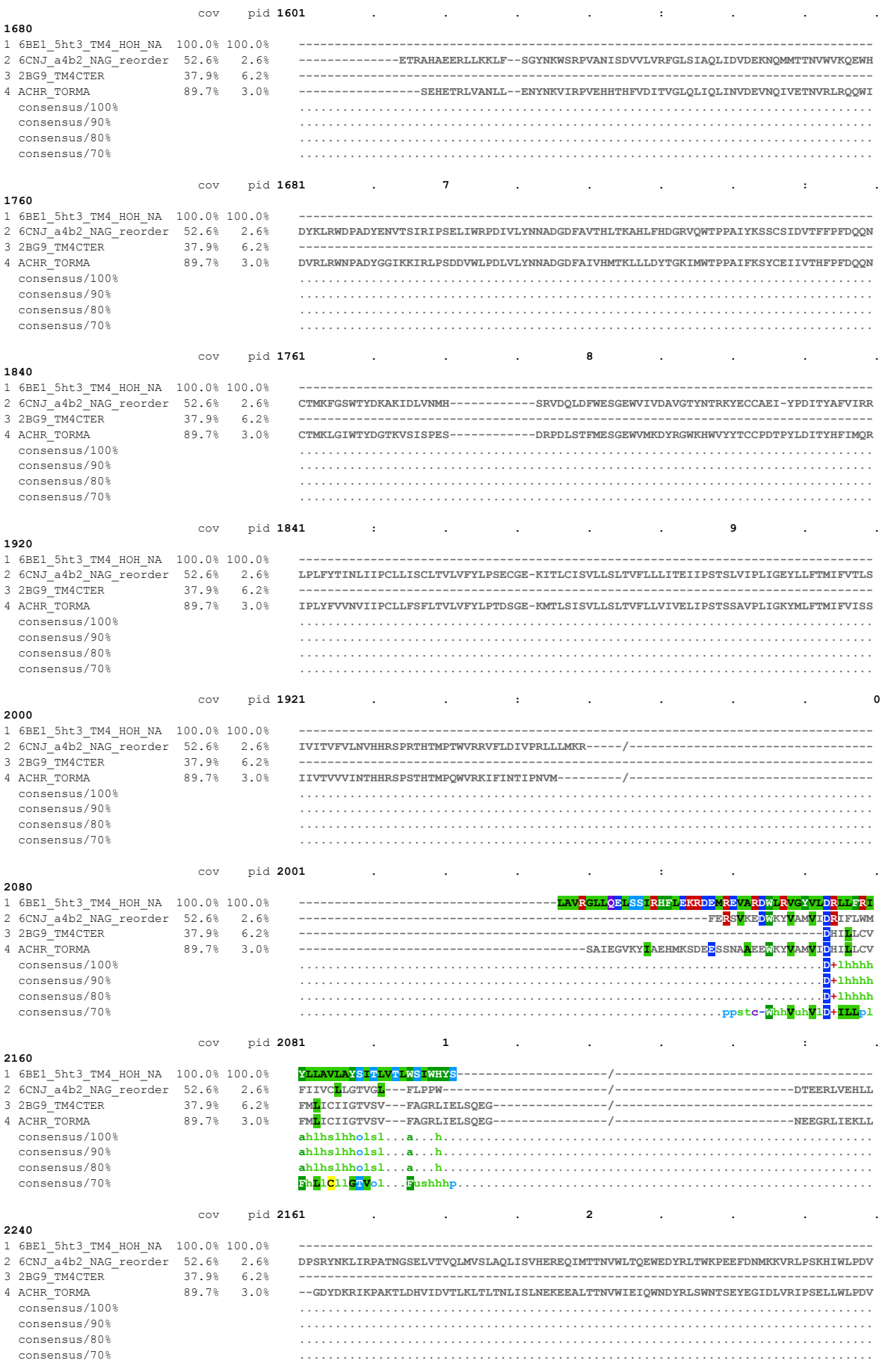
			Best	992.86	0.076	12.10	1.74
5WEL	ZK1	6FQK	Reference	-	0.136	4.35	0.0
			1	730.83	0.138	4.33	0.99
			2	730.09	0.140	4.32	0.97
			3	725.34	0.133	5.26	1.55
			4	718.01	0.139	5.12	1.18
			5	717.57	0.138	5.63	1.22
			Best	730.09	0.140	4.32	0.97
7CKQ	P4P	2BOW	Reference	-	0.051	0.54	0.0
			1	420.34	0.031	0.49	5.48
			2	417.43	0.047	0.48	4.99
			3	417.26	0.044	0.48	5.15
			4	416.70	0.036	0.50	4.52
			5	415.74	0.020	0.52	5.05
			Best	409.03	0.036	0.49	1.47

Chapter 6




cov pid 241 : . . . . 3 . . 320
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% YAFVIRRLPLFYTINLIIPCLLISCLTVLVFYLPSECGE-KITLCISVLLSLTVFLLLITEIIPSTSLVIPLIGEYLLFT
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% YHFIMQRIPLYFVVNVIIPCLLFSFLTVLVFYLPTDSGE-KMTLSISVLLSLTVFLLVIVELIPSTSSAVPLIGKYMLFT
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....
cov pid 321 . . : 4 400
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% MIFVTLSIVITVFVLNVHHRSPRTHTMPTWVRRVFLDIVPRLLLMKR-----/----------
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% MIFVISSIIVTVVVINTHHRSPSTHTMPQWVRKIFINTIPNVM---------/----------
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....
cov pid 401 . . . . : . . . 480
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------LAVRGLLQELSSIRHFLEKRDEMREVARDWLRVGYVL
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% ----------FERSVKEDWKYVAMVI
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% ----------SAIEGVKYIAEHMKSDEESSNAAEEWKYVAMVI
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....ppstc-WhhVuhVl
cov pid 481 . 5 . . . . : . 560
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% DRLLFRIYLLAVLAYSITLVTLWSIWHYS----------/----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% DRIFLWMFIIVCLLGTVGL---FLPPW----------/----------DTEE
3 2BG9_TM4CTER 37.9% 6.2% DHILLCVFMLICIIGTVSV---FAGRLIELSQEG----------/----------
4 ACHR_TORMA 89.7% 3.0% DHILLCVFMLICIIGTVSV---FAGRLIELSQEG----------/----------SVMED
consensus/100% D+lhhhhahlhslhholsl...a...h..........
consensus/90% D+lhhhhahlhslhholsl...a...h..........
consensus/80% D+lhhhhahlhslhholsl...a...h..........
consensus/70% D+lLLplFhLlClllGTvol...Fushhhp..........
cov pid 561 . . . 6 . . . 640
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% RLVEHLLDPSRYNKLIRPATNGSELVTVQLMVSLAQLISVHEREQIMTTNVWLTQEWEDYRLTWKPEEFDNMKKVRLPSK
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% TLLSVLF--ENYNPKVRPSQTVGDKVTVRVGLTLTSLLILNEKNEEMTTSVFLNLAWTDYRLQWDPAAYEGIKDLSIPSD
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....
cov pid 641 : . . . . 7 . . 720
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% HIWLPDVVLYNNADGMYEVSFYSNAVVSYDGSIFWLPPAIYKSACKIEVKHFPFDQQNCTMKFRSWTYDRTEIDLVLKS-
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% DVWQPDIVLMNNNDGSFEITLHVNVLVQHTGAVSWHPSAIYRSSCTIKVMYFPFDWQNCTMVFKSYTYDTSEVILQHALD
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....
cov pid 721 . . : 8 800
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% ----------EVASLDDFTPSGEWDIVALPGRRNEN----PDDSTYVDITYDFIIRRKPLFYTINLIIPCVLITSLAIL
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% A--/----VKEIMINQDAFTENGQWSIEHKPSRKNWRS----DDPSYEDVTFYLIIQRKPLFYIVYTIVPCILISILAIL
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....
cov pid 801 . . . . : . . . 880
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% VFYLPSDCGE-KMTLCISVLLALTVFLLLISKIVPPTSLDVPLVGKYLMFTMVLVTFSIVTSVCVLNVHHRSPTTHTMAP
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% VFYLPPDAGE-KMSLSISALLALTVFLLLLADKVPETSLSVPIIISYLMFIMILVAFSVILSVVVLNLHHRSPNTHTMPN
consensus/100% ....
consensus/90% ....
consensus/80% ....
consensus/70% ....
cov pid 881 . 9 . . . . : . 960
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0% ----------
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6% WVKVVFLEKLPALLFMQQ-----/----------
3 2BG9_TM4CTER 37.9% 6.2% ----------
4 ACHR_TORMA 89.7% 3.0% WIRQIFIETLPPFL---------/----------
consensus/100% ....
consensus/90% ....
consensus/80% ....


consensus/70%
cov pid 961 0
1040
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1041 1
1120
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1121 2
1200
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1201
1280
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1281 3
1360
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1361 4
1440
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1441 5
1520
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1521 6
1600
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%


cov pid 1601
1680
1 6BE1_5ht3_TM4_HOH_NA 100.0% 100.0%
2 6CNJ_a4b2_NAG_reorder 52.6% 2.6%
3 2BG9_TM4CTER 37.9% 6.2%
4 ACHR_TORMA 89.7% 3.0%
consensus/100%
consensus/90%
consensus/80%
consensus/70%
cov pid 1681
1760
cov pid 1761
1840
cov pid 1841
1920
cov pid 1921
2000
cov pid 2001
2080
cov pid 2081
2160
cov pid 2161
2240

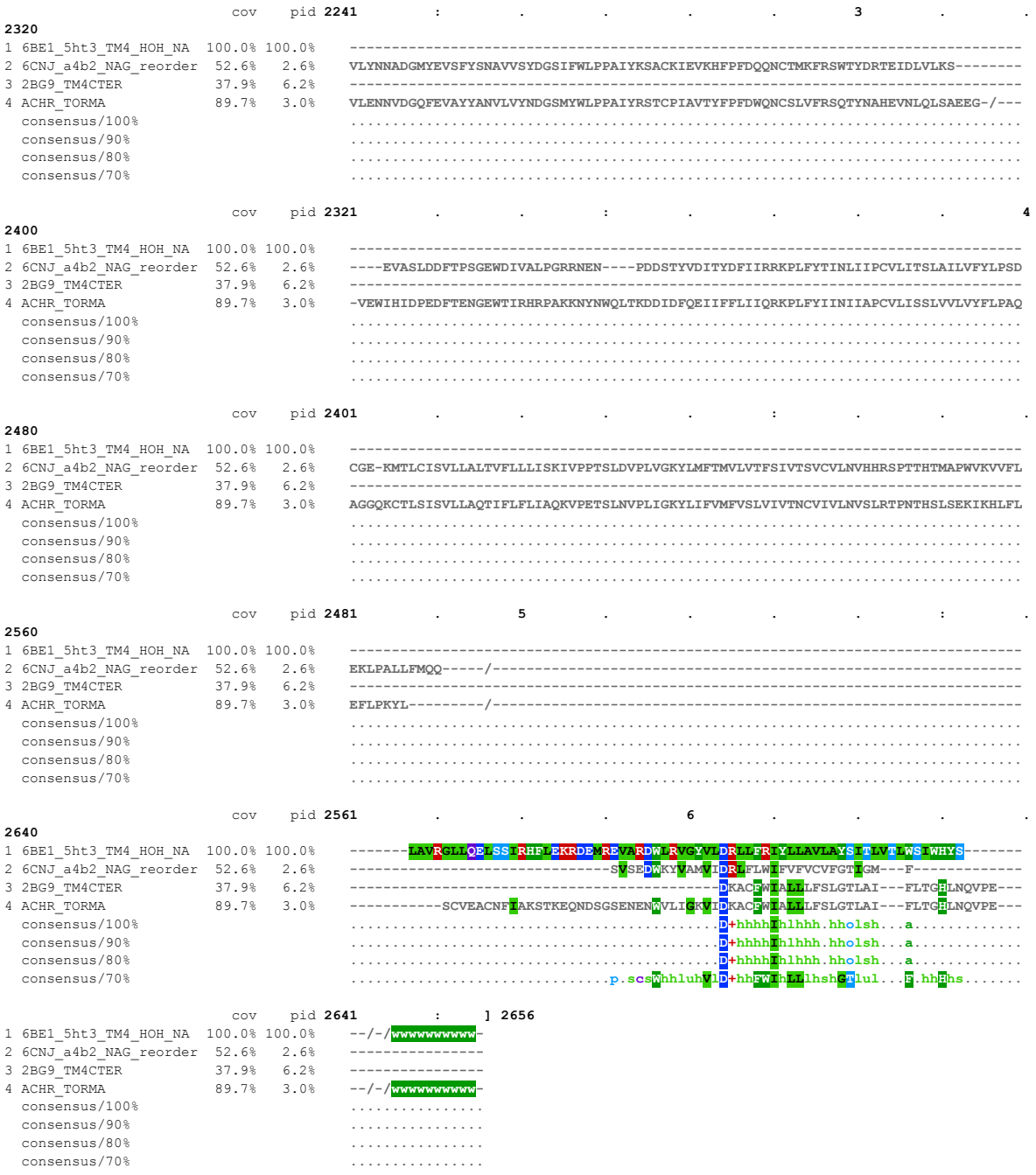


Figure A13. *The initial alignment for homology modelling provided by Dr Joseph Nwecombe.*