Computational analysis of membrane transporters and their substrates

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich-Technischen Fakultät

der Universität des Saarlandes

von

Duy Nguyen

Saarbrücken

Oktober 2017

Tag des Kolloquiums:	02.02.2018
Dekan:	Prof. Dr. rer. nat. Guido Kickelbick
Berichterstatter:	Prof. Dr. Volkhard Helms
	Prof. Dr. Richard Zimmermann
Vorsitz:	Prof. Dr. Karin Römisch
Akad. Mitarbeiter:	Dr. Jessica Hoppstädter

Acknowledgement

I would like to show my appreciation to all of my friends who encouraged and helped me all the time when I was doing this thesis.

I especially appreciate Prof. Dr. Volkhard Helms as a supervisor who cares in remarkable way about his students. He contributed a lot of time for discussion and proofreading.

I also want to give my thank to Prof. Dr. Richard Zimmermann, Prof. Dr. Adolfo Cavalié and Dr. Johanna Dudek for their collaboration in our projects.

My work is funded by Deutsche Forschungsgemeinschaft (DFG) and Graduiertenkolleg (GK) 1276.

I also appreciate Oak Ridge Leadership Computing Facility for the computing time which is essential for the molecular dynamics simulation project.

Last but not least, I would like to give my sincere appreciation to my parents, who are always besides me to support and encourage endlessly for my studying. Words cannot express how lucky and happy I am to have you as my parents.

Abstrakt

In den vergangenen Jahren haben sich rechnerische Technologien sowie die Entwicklung von anspruchsvollen Algorithmen und Software schnell entwickelt. Diese technologischen Fortschritte spielen einen entscheidende Rolle für die bioinformatische Forschung, da die biologischen Daten in Bezug auf Quantität, Qualität und Komplexität exponentiell zunehmen. In dieser Arbeit haben wir in drei Projekten, die auf die Charakterisierung von funktionellen Eigenschaften von Membrantransportsystemen sowie deren Wechselwirkungen mit Substraten und Nicht-Substraten abzielen, Bioinformatik-Werkzeuge/-Techniken entwickelt, umgesetzt und angewendet.

Membrantransporter sind eine sehr wichtige Klasse von integralen Transmembranproteinen, die für den Materialaustausch zwischen Zellen und deren Umgebungen verantwortlich sind. Aufgrund der starken Beziehung mit verschiedenen Krankheiten und abnormen medizinischen Bedingungen wurde und wird die Wechselwirkung von Transportern mit kleinen Arzneimittelmolekülen intensiv untersucht. Im ersten Projekt haben wir eine neuartige Methode für die MdfA-Substratklassifizierung entwickelt. MdfA ist ein Multidrug-Membrantransporter von *E. coli*, der für die Erkennung und den Transport eines breiten Spektrums von Substraten mit nicht verwandten Eigenschaften verantwortlich ist. Im Gegensatz zu anderen herkömmlichen Verfahren, die allgemeine Merkmale wie aus den sequenzen abgeleitete Informationen, molekulare Deskriptoren usw. verwenden, umfasst das neue Verfahren Protein-Ligand-Struktur-Wechselwirkungen und potentielle Energieinformationen, die aus molekulardynamischen Simulationen abgeleitet sind. Allerdings stieß das Verfahren immer noch auf Schwierigkeiten mit dem strukturellen Ähnlichkeitsproblem zwischen Substraten und Nichtsubstraten. Die neue Methode erreichte eine zufriedenstellende Genauigkeit mit 73,12% Klassifizierungsgenauigkeit. Es ist die erste Methode, die Protein-Ligand-Wechselwirkungen bei einem Klassifizierungsproblem für Membrantransport berücksichtigt.

Im nächsten Projekt analysierten wir Proteomikdaten aus Sec61α und TRAP-Stummschaltungsexperimenten, um TRAP-Substrate zu identifizieren und zu charakterisieren. TRAP ist eine assistierende Komponente des Translocon-Komplexes, der für die Protein-Translokation verantwortlich ist. Wir identifizierten erfolgreich einen Satz von TRAP-abhängigen Proteinen aus Massenspektrometrie-Proteomik-Daten. Darüber hinaus zeigte unsere Analyse, dass die Signalpeptide von TRAP-Substraten eine geringe Hydrophobie-Tendenz sowie einen signifikant erhöhten Glycin- und Prolin-Gehalt aufwiesen. Wir schlugen vor, dass TRAP dafür verantwortlich sein

kann, diejenigen Proteine bei der Migration durch den Sec61α-Kanal zu unterstützen, die Signalpeptide mit hohem Glycin-Prolin-Gehalt und geringer Hydrophobizität haben.

Im letzten Projekt haben wir die molekulare Docking-Technik angewendet, um die Bindungsmodi von mehreren Eeyarestatin-Verbindungen (ES1, ES24, ES35 und ES47) mit einem Homologiemodell von humanem Sec61α Protein zu untersuchen. Der Sec61α-Kanal ist nicht nur für die Proteintranslokation verantwortlich, sondern fördert auch Ca^{2+} Leckage. Die Docking-Ergebnisse ergaben, dass sich die energetisch günstigste Bindungsposition von ES1 und ES24 zwischen den H2- und H7-Helices befindet, die die "Türen" des lateralen Tores sind. Daher ist es wahrscheinlich, dass sie die Tor-Funktion behindern können und nach der Bindung den Kanal offen halten. Daher haben wir postuliert, dass ES1 und ES24 die potentiellen "Gate Blocker" sein können, die Ca^{2+} Leckage durch Sec61α fördern. Diese Ergebnisse stimmen mit den Ergebnissen der Calcium-Imaging-Experimente überein, die von unseren Kollegen durchgeführt wurden.

In dieser Arbeit haben wir verschiedene Rechentechniken eingesetzt, um neue mechanistische Einblicke in Transmembran-Transporter zu gewinnen und wichtige Informationen aus der Analyse von Proteomik-Daten zu erhalten. Wir hoffen, dass unsere Arbeit nützliche mikroskopische Details und mögliche Mechanismen für die experimentellen Biologen, die an transmembranen Proteinen arbeiten, zur Verfügung stellt.

Abstract

Recent years have seen fast improvements in computational technologies as well as in the development of sophisticated algorithms and softwares. These technological advances are crucial for bioinformatics research since biological data are exponentially increasing in terms of quantity, quality and complexity. In this thesis, we developed, implemented and applied bioinformatics tools/techniques in three projects that aim at characterising functional properties of membrane transport systems as well as their interactions with substrates and non-substrates.

Membrane transporters are a very important class of integral transmembrane proteins which are responsible for material exchange between cells and their environments. Due to the strong association with various diseases and abnormal medical conditions, the interactions of transporters with small drug molecules are subject of intense studies. In the first project, we developed a novel method for MdfA substrate classification. MdfA is a multidrug membrane transporter of *E. coli*, which is responsible for recognising and transporting a wide spectrum of substrates with unrelated properties. Unlike other conventional methods that utilised general features such as sequence derived information, molecular descriptors, etc. , the new method incorporates protein-ligand structural interactions and potential energy information derived from molecular dynamics simulations. However, the method still encountered difficulties with the structural similarity problem between substrates and non-substrates. The new method achieved a decent performance with 73.12% of classification accuracy. Regardless, this is the first method that considers protein-ligand interactions in a classification problem related to membrane transport.

In the next project, we analysed the proteomics data from Sec61α and TRAP silencing experiments to reveal and characterise TRAP substrates. TRAP is an assisting component of the translocon complex, which is responsible for protein translocation across the membrane of the endoplasmic reticulum. We successfully identified a set of TRAP dependent proteins from mass spectrometry proteomics data. Furthermore, our analysis revealed that the signal peptides of TRAP substrates showed a low hydrophobicity tendency as well as significantly increased glycine and proline content. We propose that TRAP may be responsible for helping those proteins carrying signal peptides with high glycine-proline content and low hydrophobicity to migrate easily through the Sec61α channel.

In the last project, we applied molecular docking to investigate the binding modes of several eeyarestatin compounds (ES1, ES24, ES35 and ES47) to a structural ho-

mology model of human Sec61α protein. The Sec61α channel is not only responsible for protein translocation but also promotes Ca^{2+} leakage. Based on the docking results, we found that the energetically most favourable binding positions of ES1 and ES24 are located in between the H2 and H7 helices, which are the "doors" of the lateral gate. Hence, they are likely to hamper the gate function, keeping it open upon binding. Therefore, we postulated that ES1 and ES24 can be potential "gate blockers" which promote Ca^{2+} leakage via Sec61α. These findings are consistent with the results from calcium imaging experiments which were conducted by our colleagues.

In this thesis, we used various computational techniques to provide new mechanistic insight for transmembrane transporters as well as to reveal important information from the analysis of proteomics data. We hope that our work will provide useful microscopic details and possible mechanisms to the experimental biologists who are working on transmembrane proteins.

Contents

Abbreviations

ER	Endoplasmic Reticulum.
FDR	False Discovery Rate.
LLS	Local Least Squares.
MD	Molecular Dynamics.
MS	Mass Spectrometry.
PBC	Periodic Boundary Conditions.
PLR	Pore-Lining Residues.
PME	particle-mesh Edward.
QSAR	Quantitative Structure-Activity Relationships.
SAM	Significance Analysis of Microarrays.
SP	Signal Peptide.
SRP	Signal Recognition Particle.
TMH	Transmembrane Helix.
TRAP	TRanslocon-Associated Protein.

Chapter 1

Introduction

1.1 Biological membrane

Membranes are basic components of prokaryotic and eukaryotic cells. Biological membranes act as a barrier, separating cellular and sub-cellular compartments/organelles from their surroundings. Cellular membranes are made of three main components: lipids, carbohydrates and membrane proteins [1–3].

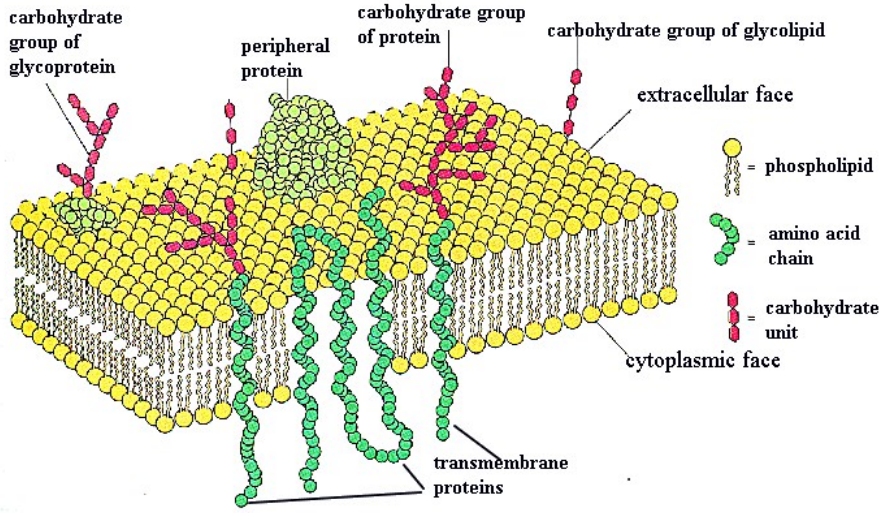


Figure 1.1: A lipid bilayer membrane including peripheral and transmembrane proteins. Figure taken from [2]

Glycerophospholipids (Fig. 1.2), phospholipids and cholesterol are three common types of lipids that are found in biological membranes. Membrane lipids are composed of two major regions: a hydrophilic phosphate head and hydrophobic tails of fatty acids. Because of these two-faced properties, lipid molecules tend to as-

semble into a bilayer geometry (with the hydrophilic heads facing outward and the hydrophobic tails bury deep inside bilayers interior) without any use of energy, or, alternatively, into small vesicles. About 50% of the mass of membranes is composed of lipids. In the plasma membranes of animal cells, cholesterol is accounted for 20% of the lipid amount while it is absent in bacterial membranes or mitochondrial membranes. Cholesterol promotes the stiffness of membranes. The other lipids also take part in cell signalling and and cell recognition.

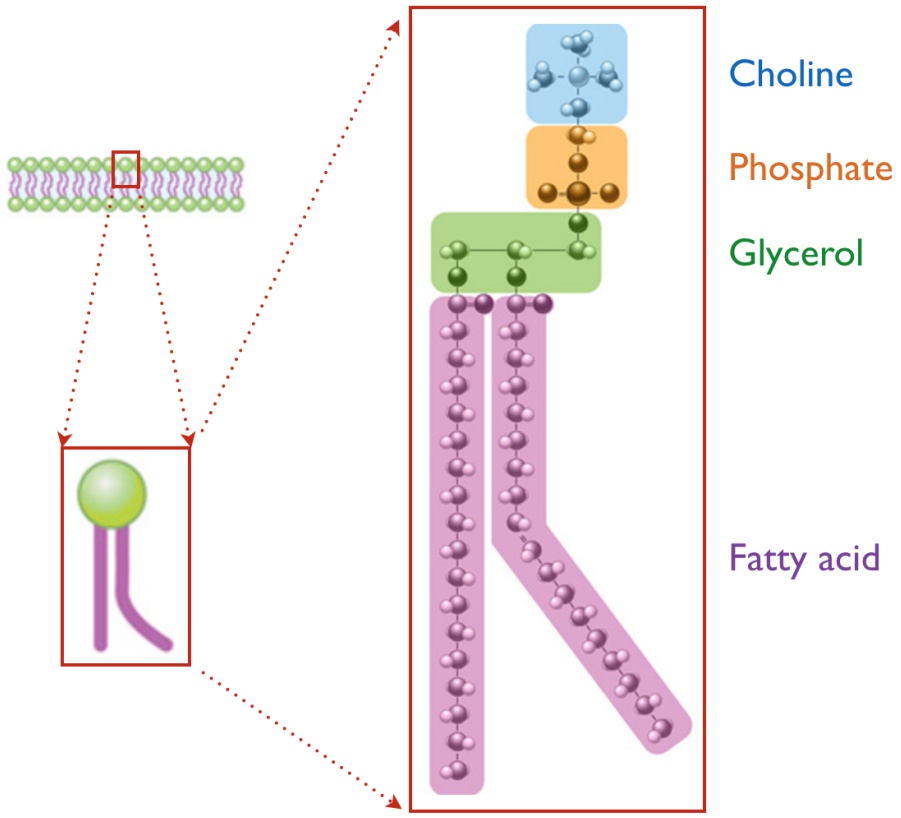


Figure 1.2: Structure of a glycerophospholipid molecule. Figure is adapted from [1]

At physiological condition, cell membranes behave like two dimensional fluids. When the temperature decreases, they become gel-like. The fluidity of the membrane is also affected by fatty acid composition and cholesterol content. Unsaturated fatty acids have bends at the double bonds, hampering the packing of lipid molecules, hence, lowering the transition temperature from gel-like to fluid state (or vice versa). Also, short hydrocarbon chains of lipid molecules have a lower affinity for packing with each other than longer chains, consequently, affecting the transition temperature. The fused-ring structure of cholesterol (Fig. 1.3) interacts with the hydrocarbon chain of neighbouring glycerophospholipids or phospholipids, making the membrane stiffer.

Besides lipid molecules, two other main components of biological membranes are membrane proteins and carbohydrates (i.e. sugars). The functionality and different

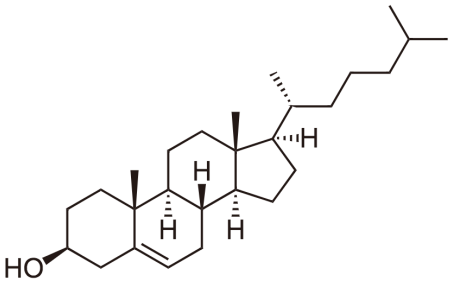


Figure 1.3: Structure of a cholesterol molecule.

types of transmembrane proteins are described in the next section (1.2). Carbohydrates can be attached to either membrane proteins or lipid molecules on the outer side of the membrane (Fig. 1.1). Due to the large number of varieties and combinations of carbohydrates on the cell surface, they are involved in various recognition mechanisms. Therefore, in recent years, more and more studies focus on carbohydrates of cell membrane, experimenting new therapies for cancer and various diseases.

1.2 Transmembrane proteins

1.2.1 Membrane and membrane proteins

Cellular membranes function as barriers and gatekeepers of the cell. They allow some molecules to travel across the lipid bilayer but others cannot. In fact, most of the biological functions/activities which occur at membranes are related to membrane proteins which are embedded in membranes. Membrane proteins play many important roles in the cells of all organisms such as: they transport a wide range of materials (water, ions, metabolites, proteins...) through the membrane, they transmit electrical impulses, they catalyse enzymatic reactions, connect neighbouring cells or extracellular matrix, or keep other proteins to stay in specific locations (anchoring), etc. [4]

There are two types of membrane proteins. The first type includes proteins that span the entire membrane (Fig. 1.1). Those membrane proteins are called *transmembrane* or *integral* proteins. The second type are *peripheral membrane proteins* which are bound at the peripheral surface of the membrane, or bound to other integral membrane proteins. Since this thesis concentrates on membrane transporters, we will focus on transmembrane proteins from now on.

Due to the non-polar, hydrophobic environment of the transmembrane region, transmembrane proteins cannot form any hydrogen bonds or electrostatic interactions with the membrane except of van der Waals interactions. Therefore, amino acid residues of transmembrane proteins can only establish hydrogen bonds among themselves. As a result, there are only two structural options for the membrane-

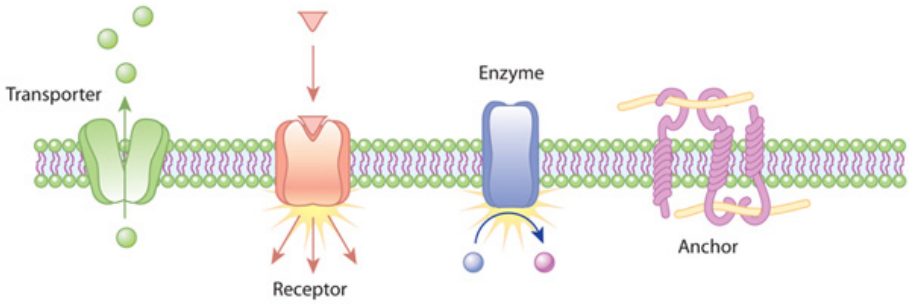


Figure 1.4: Functions of membrane proteins. Figure from [1]

spanning part of transmembrane proteins: *α-helices* and *β-sheets* (Fig. 1.5). Those secondary structures are connected by loops.

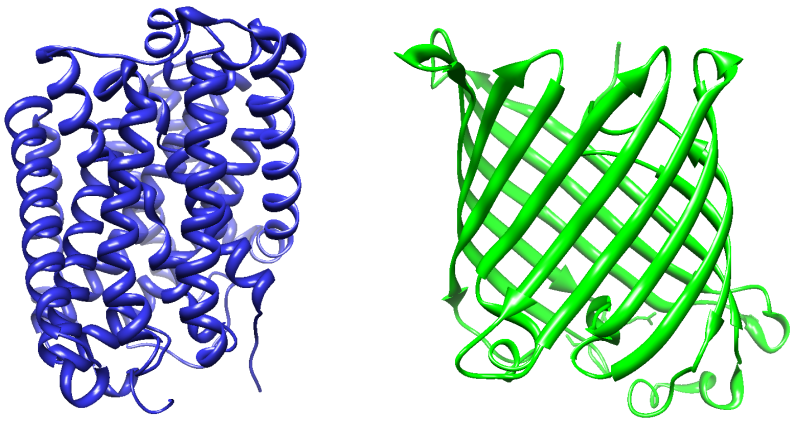


Figure 1.5: Left: α-helical transporter (3GIA). Right: β-barrel transmembrane protein (1QD6)

The β-barrels can only be found in the outer membranes of Gram-negative bacteria and in organelles such as mitochondria and chloroplasts where they allow passive diffusion for small molecules. In contrast, the α-helical bundle transmembrane proteins can be found in all cellular membranes. They are more common and diverse in functionality compared to β-barrels transmembrane proteins. Almost all medically important membrane proteins (enzymes, receptors, channels, transporters...) belong to this group [4].

1.2.2 Transport proteins

Only gas molecules (like oxygen and carbon dioxide) and small hydrophobic molecules can directly travel across phospholipid bilayer membranes by passive diffusion. A net flux is established driven by the concentration gradient across the membrane.

No active transport proteins are needed. Others substances like ions, sugars, amino acids cannot diffuse across the bilayer membrane to fulfil the cell's requirements. Those molecules must be transported across the membrane by a group of integral membrane proteins, transport proteins. Sometimes, passage of some molecules such as water or urea which can also diffuse across the phospholipid bilayer are facilitated by transport proteins. There are three major classes of transport proteins: *ATP-powered pumps* (or *pumps* for simplification), *channels* and *transporters* (Fig. 1.6a).

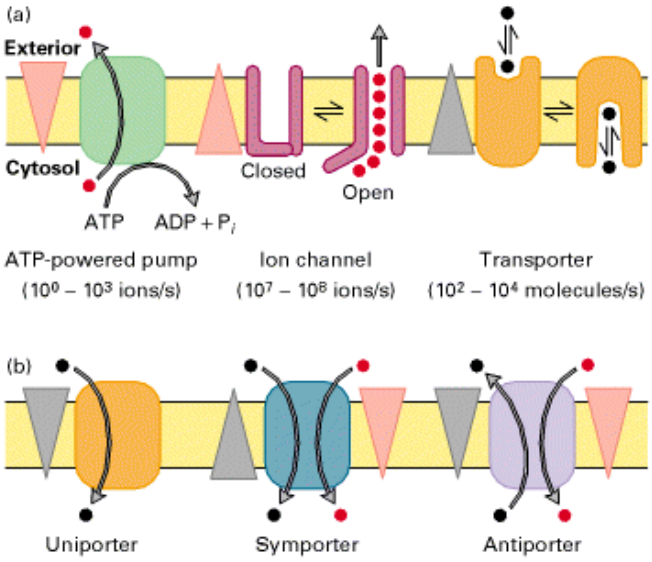


Figure 1.6: (a) The three major types of transport proteins. (b) The three groups of transporters. Gradients are illustrated by triangles pointing toward lower concentration or electrical potential or both. Image from [5]

- ATP-powered pumps are ATPases that utilise energy generated by ATP hydrolysis to drive ions or small molecules across the bilayer membrane against the gradient concentration or electrical potential. This kind of transport is described as *active transport*.

- Channels allow passage of water or come specific ions across the membrane toward lower concentration or along the electrical potential gradient. They form tunnel-like passages which span the entire membrane, allow multiple water or ion molecules to travel across the membrane at the same time.

- Transporters transport a wide range of ions and molecules across cell membranes. Unlike channel proteins, they can only transport one molecule (or a few) at a time. To initiate the transportation, the substrate molecule binds to the transporter. After the substrate is bound, the transporter undergoes a

conformation change which allows the substrate (and only) to be transported across cell membranes. Based on the transport mechanisms, transporter proteins are classified into three classes (Fig. 1.6b):

- *Uniporters* only transport one molecule downhill the concentration or electrical potential gradient.
- *Symporters* can accelerate the movement of 2 type of molecules: one toward lower gradients and the other type against its gradients. Those molecules travel in the same direction.
- The mechanism of *Antiporters* is similar to that of symporters but two types of molecules move in opposite directions. Unlike ATP pumps, symporters and antiporters do not utilise the energy from ATP hydrolysis during the transportation. Symporters and antiporters are also referred as *cotransporters* because they are able to transport two different types of molecules simultaneously.

For simplification, from now on, all transport proteins are referred to as transporters. But what makes transporters so important?

1.2.3 Transporters and Channelopathies

Transporting ions and molecules across the cell membrane is critical for essential biological processes, especially in the human body. Therefore, mutations that disrupt or alter the functionality of transport proteins could deliver devastating effects. Diseases caused by disturbing the function of ion channels or their regulatory proteins are called *channelopathies* [6]. Up until now, many mutations that cause channelopathies have been reported and the number is still increasing [6, 7]. For instance:

- Mutations in sodium channels and potassium channels in the central nervous system lead to epilepsy and migraine [7].
- Disrupting peripheral nerve potassium channels may result in neuromyotonia (a.k.a. Isaac's syndrome) [8].
- Mutations in KCNH2 (human ether-a-go-go related gene) cause type 2 long QT syndrome, a rare genetic disease associated with life-threatening abnormal heartbeat, by disrupting IKr which is an ion channel subunit [9].
- Disturbance in Kir6.2, a major subunit in potassium channels located in pancreatic β cells, by mutations in the KCNJ11 gene lead to hyposecretion of insulin which cause diabetes mellitus [6].
- Disorders in voltage-gated ion channels cause inherited muscle diseases (non-dystrophic myotonias and familial periodic paralyses) [8].

Because of the wide range of diseases which are caused by abnormal functionality of transport proteins, transmembrane transporters have become an important target family for pharmaceutical research.

Another problem regarding transporters that attracted the attention of many scientists is the antibiotics resistance of bacteria due to multidrug efflux pumps [10]. This problem is a headache for many pharmaceutical companies since their drugs are becoming ineffective when bacteria can develop a defense mechanism by emerging mutations in efflux pumps to fight against the new drugs (Fig. 1.7). Moreover, due to the ability of recognising, targeting and extruding a wide range of pharmaceutical drugs out of the cells, some transporters also cause difficulties in tumour chemotherapy and other diseases (e.g. P-glycoprotein).

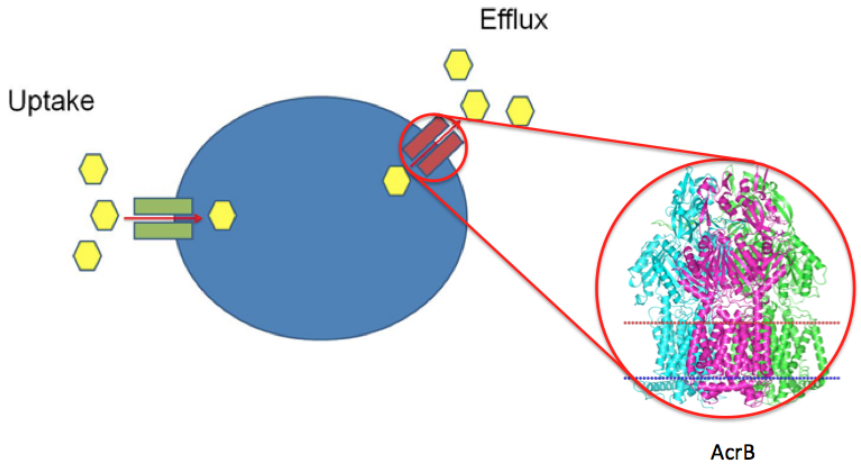


Figure 1.7: Drug uptake and efflux. AcrB is a multidrug efflux pump of *E. coli*. Image from [11]

1.3 Bioinformatics and transmembrane proteins

In the past, the majority of bioinformatics studies of transmembrane proteins has focused on their structural features. For instance, TMX [12] and BTMX [13] assign the burial status of amino acid residues in α-helical and β-barrel transmembrane proteins, respectively. Other bioinformatics tools such as MEMSAT-SVM [14] and TOPCONS [15] determine the topology of transmembrane proteins including the positions of transmembrane helices, re-entrant helices, signal peptides, etc. TMHcon [16] and MEMPACK [17] predict the helix-helix contact map of α-helical transmembrane proteins. Recently, the Pore-Lining Residues (PLR) of transporters can also be predicted by PRIMSIPLR [18] or MEMSAT-SVM. Those methods mentioned above utilised the primary protein sequence to predict the structural features due to the limitation in the number of transmembrane protein 3D structures that

existed until recently. At first, from a set of known transmembrane protein 3D structures, they extracted various features such as evolutionary information, amino acid composition, physicochemical properties, etc. Then, they used machine learning approaches such as Support Vector Machines (SVM), Artificial Neural Networks (ANN) and Hidden Markov Models (HMM) to construct classifiers based on the extracted features.

For the transmembrane proteins for which their 3D structures could be determined, there are various bioinformatics tools/algorithms which were developed for structural traits analysis. For example, TMDET [19] and PPM [20] servers can estimate the position and orientation of the transmembrane proteins inside the biological membrane. Tools such as POCKET [21], LIGSITE [22,23], HOLLOW [24], CAVER [25], PROPORES [26] and PoreWalker [27] can identify pores/pockets/channels inside proteins as well as the PLRs that made up their cavities.

Due to various essential functions of transmembrane proteins as well as their particular distribution in cells (on the membrane which acts as the entrance into the cells), transmembrane proteins are considered as potential drug targets in drug discovery. In fact, the tendency being membrane bound is one of the indicators of drug target likeliness, alongside with hydrophobicity, *in vivo* half-life, non-polar amino acid composition, etc. [28]. Therefore, several substrate/non-substrate classification methods have been developed during the past decades. One of the most popular methods for substrate classification are Quantitative Structure-Activity Relationships (QSAR) models. Initially based on the idea that similar compounds with respect to physicochemical properties possess similar biological effects, QSAR models were often used to validate the affinity of various ligands towards a common protein target. Over the years, QSAR methods have matured and are now widely used. Based on the dimensionality, QSAR methods are categorised as follows [29]:

- 1D-QSAR: only one-value descriptors are considered such as $\mathrm{p}K_a$, $\log P$, etc.
- 2D-QSAR: contains descriptors with structural patterns such as connectivity, 2D pharmacophore, etc.
- 3D-QSAR: 3D structure of ligands is included.
- 4D-QSAR: multiple representations of ligands are considered.
- 5D-QSAR: including 4D-QSAR and the representation of various induced-fit models.
- 6D-QSAR: including 5D-QSAR and multiple solvation schemes.

Besides QSAR models, machine learning approaches such as SVM, random forest, etc. were also applied for substrates classification [30–32]. However, most of the descriptors for classification of mentioned approaches are rather general and do not capture the structural details how proteins interact with their substrates [33].

1.4 Research topics addressed in this thesis

As previously mentioned, most of the substrate classification methods rely on general basic descriptors, neglecting actual interactions between proteins and their ligands. Therefore, in chapter 3, we present a novel substrate classification method which overcomes these issues by integrating Molecular Dynamics (MD) simulations of protein-ligand complexes. The main idea is to observe and extract various potential energy information as well as protein-ligand interactions. Then, by using a machine learning approach, we constructed a substrate classification model based on the information extracted from MD simulations.

In chapter 4, we studied the effects of Sec61α and TRAP silencing. Those are important components of the translocon complex, which is responsible for targeting most of the proteins to their destination in eukaryotic cells. In particular, we mainly investigated the distinct characteristics of the proteins that were affected by TRAP silencing due to the fact that TRAP function and mechanism still remain unknown.

Finally, in chapter 5, we applied a docking algorithm to provide microscopic information, structural features as well as possible mechanisms to explain the Ca^{2+} leakage caused by eeyarestatin compounds binding into the Sec61α cavity.

1.5 Aim of this thesis

In this thesis, we developed and implemented multiple approaches (from MD simulations to differential expression analysis, as well as structural analysis of different protein-ligand complexes) to study transmembrane proteins and their function. Our purpose is to provide useful information for experimental biologists who are working on transmembrane proteins.

Chapter 2

Theory

2.1 Molecular Dynamics Simulation

Since the very first MD simulation of a biological macromolecule back in 1977 [34], MD simulations have been established as a reliable computational method to study the properties of individual biomolecules and their assemblies by observing their structures, movements, interactions, etc. at the microscopic level. MD simulations can provide the positions and motions of individual particles of an N-body system, as well as 'predict' the interactions between molecules and system properties. Therefore, they can serve as a bridge between theory and experiment: either one conducts a simulation first and tests the predictions of the model by comparing with experimental results; or they act as a complement to existsing experiments, enabling us to discover new knowledge that cannot be found in other ways.

In MD simulations, the position of an individual particle after a short time step is determined by solving the classical equation of motion based on Newton's second law:

$$\vec{F} = m\vec{a} \tag{2.1}$$

where $\vec{F}$ is the force applied on the particle, m is its mass and $\vec{a}$ is its acceleration. The force can be expressed as the gradient of the potential energy U and the acceleration is the second derivative of the position $\vec{r}$. Hence, we have:

$$-\nabla U = m\frac{d^2\vec{r}}{dt^2} \tag{2.2}$$

By solving the classical equation of motion, the position of a particle can be expressed as:

$$\vec{r} = \vec{r}_0 + \vec{v}_0 t + \frac{1}{2}\vec{a}t^2 \tag{2.3}$$

where r_0 and v_0 are initial position and initial velocity, respectively. Additionally, according to Eq. 2.2, a is determined by the potential energy function:

$$\vec{a} = -\frac{1}{m}\nabla U \tag{2.4}$$

The potential energy is a function of the Cartesian positions of all particles in the system. Due to the complex nature of this function, the equations of motion (Eq. 2.2) cannot be solved analytically, and must be solved by numerical methods.

In short, to calculate the trajectory, one only needs the initial positions of the particles and the initial distribution of velocities. The initial positions can be obtained from experimental structures (X-ray, NMR, etc.) or by structural modelling, whereas the initial distribution of velocities are chosen randomly from a Maxwell-Boltzmann or Gaussian distribution with an appropriate magnitude for a given temperature. The initial velocities are generated in such a way that there is initially no overall momentum in the system:

$$\vec{P} = \sum_{i=1}^{N} m_i \vec{v}_i = 0 \tag{2.5}$$

where N is the number of particles in the system.

After a certain number of iteration steps, any occurring center of mass motion is set back to zero so that the simulation system remains at rest.

2.1.1 Potential Energy Function and Molecular Interactions

As described in Eq. 2.2, the force acting on all particles in the system can be derived from the potential energy function U, which is a function of positions, $\vec{R}$, of the N particles in the system. The potential energy is composed of intramolecular interactions, or bonded energy terms and external, or non-bonded energy terms:

$$U_{\text{total}}(\vec{R}) = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{2.6}$$

Bonded Interactions

The bonded interactions are described by three main components which correspond to three types of atom movements: bond-stretching (bond distance), bond angle bending and bond torsion (or dihedral angle). Figure 2.1 illustrates the geometry of the three main components.

$$\begin{aligned} U_{\text{bonded}} &= U_{\text{bonds}} + U_{\text{angles}} + U_{\text{dihedrals}} \\ &= \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi(1 + cos(n\phi - \delta)) \end{aligned} \tag{2.7}$$

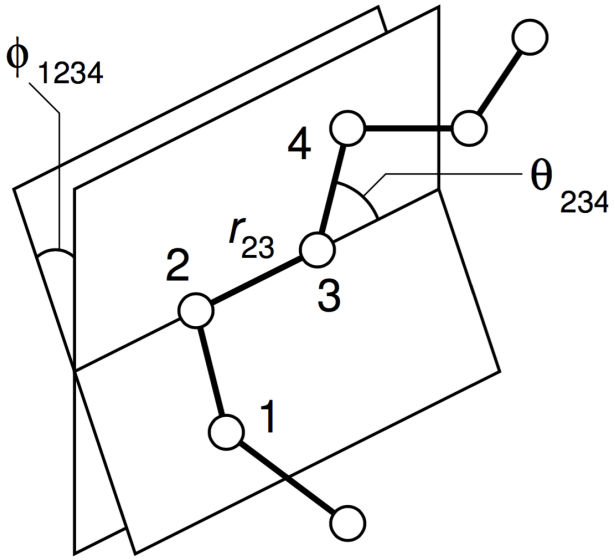


Figure 2.1: Illustration of bond distance r_{23}, bend angle θ_{234} and dihedral angle ϕ_{1234} in a simple molecule (image is adapted from [35]).

In Eq. 2.7, the first term is a harmonic potential which represents the 2-body interaction between two atoms connected by a covalent bond. This term approximates the energy of a bond as its length b deviates from the equilibrium distance b_0. The constant k_b dictates the strength of the bond. Both k_b and b_0 are specific for each pair of bonded atoms.

The second term is also described by a harmonic potential, which represents the additional energy of a shifted bond angle θ from its ideal value θ_0. The force constant k_θ and θ_0 are determined by the atoms which make up the bond angle.

The last term describes the 4-body torsion angle which is the angle between the planes formed by the first three and last three atoms of four consecutively bonded atoms. The term n indicates the periodicity of the torsion angle and δ acts as the phase shift angle.

Consistent sets of force constants, equilibrium values, and other essential parameters for macromolecules MD simulation constitute molecular force fields such as CHARMM [36], AMBER [37], GROMOS [38], etc. Usually, their parameters are determined using quantum mechanics calculation in conjunction with empirical evidence. However, the mentioned force fields only contain parameters for biological macromolecule such as protein, DNA, RNA, etc. For MD simulations of smaller, general molecules (drugs, ligands), one has to manually derive the parameters, which is called "parameterisation", or use any of the suitable force fields for such molecules such as the CHARMM General Force Field (CGenFF) [39] or the Generalized AMBER Force Field (GAFF) [40].

Non-bonded Interactions

The non-bonded potential terms consider interactions between all pair of atoms, excluding pairs of atoms that are already involved in the bonded potential energy terms. Therefore, the computational cost of non-bonded potentials is quite demanding during an MD simulation. In biological macromolecules, the dominant non-bonded interactions are van der Waals and electrostatic interactions. Therefore, the non-bonded potentials are usually comprised of 2 terms: the Lennard-Jones potential and the Coulomb potential:

$$\begin{aligned} U_{\text{non-bonded}} &= U_{\text{LJ}} + U_{\text{elec}} \\ &= \sum_{\text{atoms i,j}} \varepsilon_{ij} \Big[\big(\frac{R_{ij}^{min}}{r_{ij}}\big)^{12} - 2\big(\frac{R_{ij}^{min}}{r_{ij}}\big)^{6} \Big] + \sum_{\text{atoms i,j}} \frac{q_1 q_2}{4\pi\varepsilon_0 r_{ij}} \end{aligned} \tag{2.8}$$

The van der Waals interaction between 2 atoms is the balance between the attractive and the repulsive forces. The attractive force is due to the instantaneous dipole formation caused within an atom or molecule, which induces a dipole in a neighbouring atom or molecule. These instantaneous dipoles arise from the fluctuations in the charge distribution in the electron clouds. As a result, these effects between temporary dipoles on neighbouring atoms give rise to an attractive force. The repulsive force, in the other hand, appears when the distance between two atoms becomes closer, when the electron clouds of two atoms are unfavourably coming into contact. As two atoms or molecules are moving further away from each other, the van der Waals force gradually vanishes.

Based on the characteristics of the van der Waals force, the Lennard-Jones potential is commonly used for representing the van der Waals interactions between pair of atoms by two parameters: the potential well depth ε and the R_{ij}^{min} distance where the potential reaches its minimum. Figure 2.2 describes how the Lennard-Jones potential approximates van der Waals interactions. At close range, the repulsive force shows a steep increase when the electron clouds of the two atoms start to overlap, but it decreases when the distance increases. When the distance reaches the optimal length R_{min}, the potential energy becomes most favourable at that point, i.e. the two atoms are at their equilibrium position. The energy is getting get weaker and gradually vanishes when two atoms are separated by a large distance.

The electrostatic interaction takes into account the atomic charges q of particles. It becomes repulsive when the charges have the same sign and attractive for charges with opposite sign. The electrostatic interaction is represented by Coulomb potential where r_{ij} is the distance between two charged atoms i and j, and $\frac{1}{4\pi\varepsilon_0}$ is the Coulomb's constant.

As mentioned before, the most time-consuming part during MD simulation is the calculation of non-bonded potential energy because one have to take into account every pair of atoms in the whole system. In this naive approach, the complexity of the calculation is $O(N^2)$ since all combinations are evaluated. To reduce the computational cost, other fast evaluation methods have been developed. The common

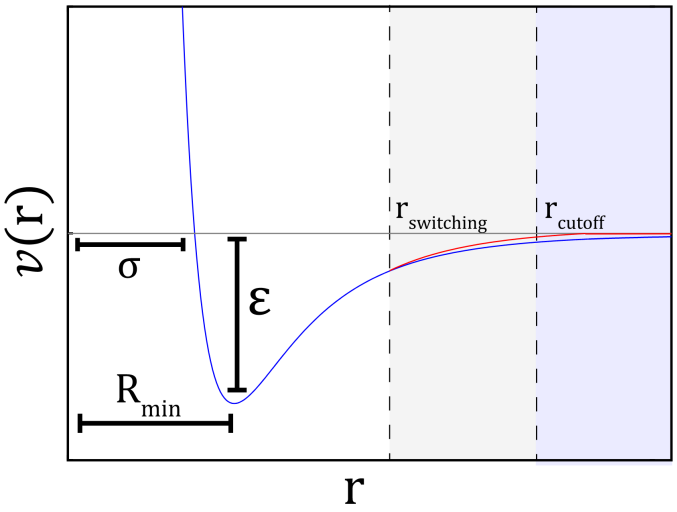


Figure 2.2: The Lennard-Jones 6–12 potential (blue) and the SWITCH cutoff method (red). Image taken from [41].

method to speed up Lennard-Jones potential calculation is cutoff distance, where the interactions between two atoms decay faster (SWITCH) or are completely ignored (TRUNCATION). In the case of the electrostatic potential, the force decays much slower to zero compared to the Lennard-Jones potential. Several studies have shown the importance of long range electrostatic effects in biological molecules [42–44]. Therefore, neglecting long range electrostatic interactions using a cutoff method may severely affect the simulation quality, especially for highly charged systems [45, 46]. The Edward summation has proved to correctly approximate long-range electrostatic interactions for simulations of proteins and enzymes [47]. The particle-mesh Edward (PME) method has also been applied successfully to periodic systems [48]. Although those methods require a larger computational cost than neglecting the long-range electrostatic, they are still remarkably faster than simply summation of the Coulomb potential of all atom pairs in the system.

2.1.2 Integration Algorithms

As mentioned above, because of the complicated nature of potential energy function, the equations of motion can only be solved by numerical methods. Many numerical algorithms have been developed for solving the equations of motion such as Verlet [49], Velocity Verlet [50], Leap-frog [51], and Beeman's algorithm [52, 53]. All of the integration algorithms were developed around the Taylor series expansion, assuming that the positions, velocities and accelerations after a short time step can be approximated from their current values. For brevity, we will omit the vector symbols for coordinates, velocities and accelerations in the following:

$$\begin{aligned} r(t+\delta t) &= r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 + \dots \\ v(t+\delta t) &= v(t) + a(t)\delta t + \frac{1}{2}b(t)\delta t^2 + \dots \\ a(t+\delta t) &= a(t) + b(t)\delta t + \dots \end{aligned} \tag{2.9}$$

Verlet algorithm

This is the very first algorithm developed for solving the integration problem of equations of motion. Verlet considers the summation of the Taylor expansion in both directions of time (forward and reverse time step). Note that this is perfectly fine for a system governed by deterministic dynamics. In the end, all of the odd order derivations cancel out since they have opposite sign:

$$\begin{aligned} r(t+\delta t) &= r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \\ r(t-\delta t) &= r(t) - v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \end{aligned} \tag{2.10}$$

Summing those two equations, we have:

$$\begin{aligned} r(t+\delta t) + r(t-\delta t) &= 2r(t) + a(t)\delta t^2 \\ r(t+\delta t) &= 2r(t) - r(t-\delta t) + a(t)\delta t^2 \end{aligned} \tag{2.11}$$

In short, the Verlet algorithm calculates the next position at $t + \delta t$ by using the current position and acceleration at time t without using the velocity term. It is pretty straightforward and simple but has moderate accuracy.

The Velocity Verlet and Leap-frog algorithm are quite similar to Verlet algorithm. In both algorithms, the position calculation makes use of the velocity term besides the position and acceleration terms and the velocity of the next phase is also calculated.

Velocity Verlet algorithm

$$\begin{aligned} r(t+\delta t) &= r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \\ v(t+\delta t) &= v(t) + \frac{1}{2}\big[a(t) + a(t+\delta t)\big]\delta t \end{aligned} \tag{2.12}$$

Leap-frog algorithm

$$\begin{aligned} r(t+\delta t) &= r(t) + v\Big(t + \frac{1}{2}\delta t\Big)\delta t \\ v\Big(t + \frac{1}{2}\delta t\Big) &= v\Big(t - \frac{1}{2}\delta t\Big) + a(t)\delta t \end{aligned} \tag{2.13}$$

The velocity calculation is a bit different in the Leap-frog algorithm. As one can see in Eq. 2.13, the velocity is calculated at time $t + \frac{1}{2}\delta t$ instead of $t + \delta t$, and it is used to calculate the position in time $t + \delta t$. This expains where the name of the algorithm comes from: the velocities *leap* over the positions by $\frac{1}{2}\delta t$ and then the positions *leap* over the velocities by the same amount of time. As a result, the velocities and the positions have never been calculated at the same time. However, the velocity at a given time can be approximated by:

$$v(t) = \frac{1}{2}\left[v\Big(t - \frac{1}{2}\delta t\Big) + v\Big(t + \frac{1}{2}\delta t\Big)\right] \tag{2.14}$$

Beeman's algorithm

$$\begin{aligned} r(t+\delta t) &= r(t) + v(t)\delta t + \frac{2}{3}a(t)\delta t^2 - \frac{1}{6}a(t-\delta t)\delta t^2 \\ v(t+\delta t) &= v(t) + v(t)\delta t + \frac{1}{3}a(t)\delta t + \frac{5}{6}a(t)\delta t - \frac{1}{6}a(t-\delta t)\delta t \end{aligned} \tag{2.15}$$

The Beeman's algorithm is also a relative of the Verlet algorithm. The advantage of the Beeman's algorithm is the higher precision due to the better expressions for the position and velocity approximation. However, because of the more complex expression, the Beeman's algorithm requires more computing time compared to the Verlet algorithm.

2.1.3 Neighbour Lists

As mentioned in the previous section, the most expensive task during an MD simulation is the evaluation of the non-bonded interactions, especially when one has to evaluate all possible atom pairs for the calculations. Several methods have been developed to reduce the computational cost such as cutoff methods or Edward summation. Let us consider the TRUNCATION cutoff approach, where only short range interaction potentials are evaluated to avoid the expensive calculations. Still, the time to examine and search for the pairs which satisfy the condition $r_{ij} \leqslant r_{cut}$ is quite time consuming. Loup Verlet suggested to construct lists of close pairs of atoms, which are called the neighbour lists, to speed up the calculation process [49]. The construction of neighbour lists is depicted in Fig. 2.3. Basically, the neighbour list of a particular atom i contains all atoms j whose distance from i is smaller than a predefined distance r_{list} with the condition that $r_{list} > r_{cut}$. During the non-bonded interaction calculation step, only the atoms inside the neighbour list are taken into account. From time to time, the lists are updated before atoms from outside r_{list} (black dots in Fig. 2.3) come close to and interact with central atoms (white dots in Fig. 2.3).

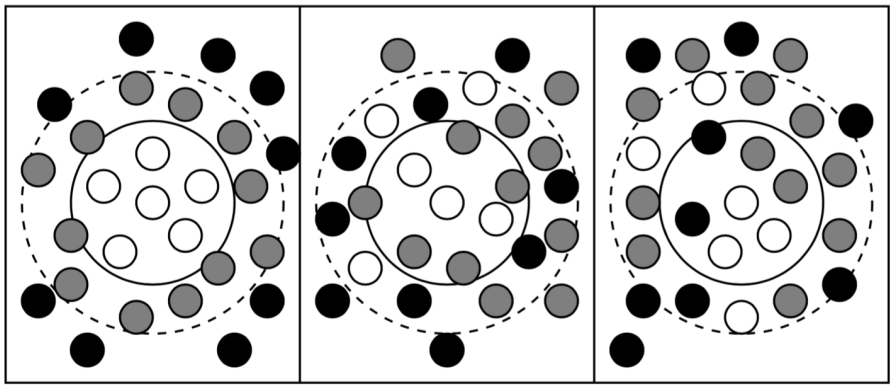


Figure 2.3: The construction of the Verlet neighbour list. From left to right: initial, later and "too late" state. The solid circle and dashed circle depict the potential cutoff range (r_{cut}) and the list range (r_{list}), respectively. Image taken from [35].

2.1.4 Periodic Boundary Conditions

When performing an MD simulation with a relatively small number of particles and the surface effects are not of interest, the Periodic Boundary Conditions (PBC) are applied so that the particles are simulated as if they were in a bulk solution. For example, let us consider a simulation box of five particles in Fig. 2.4. The central box containing original particles (grey circles) is surrounded by eight replicas of itself by simple translations. During MD simulation, the forces calculated on an original particle take into account the particles inside the original box as well as the particles in the replicated boxes. Usually, the size of the box is defined in such a way that a particle cannot interact with its images in the replicated boxes. As soon as a particle moves out of the original simulation box, an image particle moves in to replace it.

2.1.5 Setting up a Molecular Dynamics Simulation

As described earlier, in MD simulations, the time dependent positions of the particles in the molecular system are obtained by solving the equations of motion, using approximation methods and the potential energy function. In the end, the result of an MD simulation is a time series of conformations, which is called a trajectory, of the molecular system, as well as the time series of many different energy terms which were calculated during the MD simulation. To be able to mimic experimental conditions, most molecular MD simulations are performed in the NPT ensemble, which means at a constant number of particles (N), pressure (P) and temperature (T). A typical MD simulation workflow consists of the following steps (Fig. 2.5):

- Initialisation: the first step of ann MD simulation begins with choosing a starting point of the system. Usually, in the simulation of biomolecules, the initial structure of interest is obtained from the Protein Data Bank (http://www.rcsb.org/pdb/).

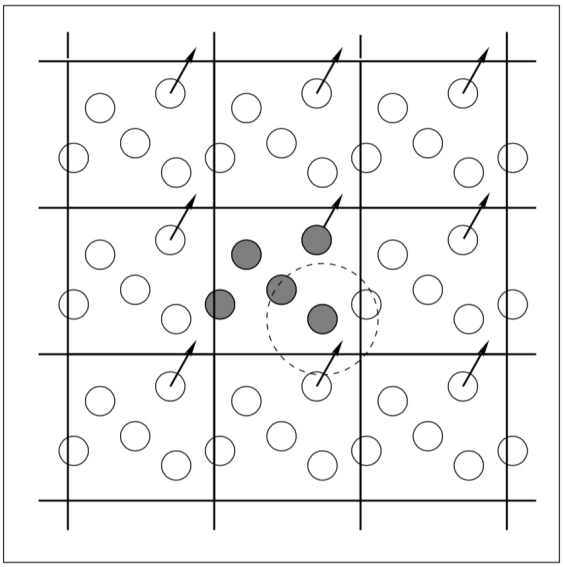


Figure 2.4: Periodic boundary conditions. Image taken from [35].

Sometimes, theoretical models for the structures such as homology models can also be used as a starting point for an MD simulation. However, the models must be chosen with great care since the quality of the models does affect the quality of the simulation. When the initial structures has been chosen, the biomolecules should be put into their innate environment: a water box for soluble proteins or insertion into lipid bilayer for membrane proteins, etc. Finally, the water and the ion molecules are added to neutralise the whole system at the physiological salt concentration.

- Minimisation: before any MD simulation is started, it is advisable to perform an energy minimisation to minimise the whole system. The goal of this process is to remove any strong clashes/interactions that could distort the system, resulting in an unstable simulation.

- Assign initial velocities: first, initial velocities for each atom are assigned at low temperature and the whole system is propagated in time by simulation. Once the whole system is equilibrated, the temperature is slightly increased and new velocities are reassigned and the whole process is repeated until the desired temperature has been reached.

- Equilibration: once the system has reached the desired temperature, the simulation is continued in the NPT ensemble. During this stage, several properties of the system are monitored: the pressure, the temperature, the energy, the structure, etc. The goal of this equilibration step is to run the simulation until those properties are stable with respect to time.

- Production: once the system has been equilibrated, which means the important

properties have become stable over time, one can begin to run the simulation in production phase for the desired time length.

- Analysis: during MD simulation, the positions, velocities as well as various energy terms are saved. Once the simulation is finished, one can use this information to visualise and study the energetic and structural changes at the atomic level during the simulation time, and ultimately, answer the scientific question of the whole MD study.

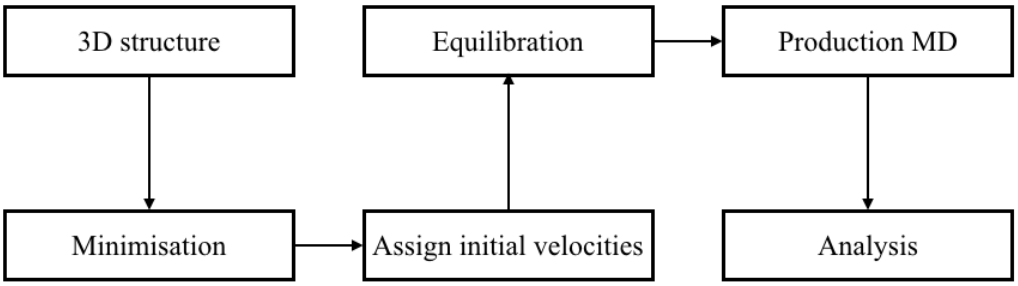


Figure 2.5: A typical MD simulation workflow.

2.2 Machine Learning by Random Forest

The Random Forest (RF) method was introduced by Breiman in 2001 [54] and soon became one of the most popular machine learning methods, especially in bioinformatics studies. The random forest is an ensemble learning method. The main principle of an ensemble method is the combination of many "weak" classifiers to form a "stronger" classifier. In particular, the random forest is a large collection of decision trees which are uncorrelated to each other. The random forest implementation can be summarised as follows [55]:

1. For $b = 1$ to B:
   (a) Draw a bootstrap sample Z^* of size N from the training data
   (b) Grow a decision tree T_b to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size n_{min} is reached.
      i. Randomly select m variables from all predictor variables
      ii. Pick the best variable/split-point among the m
      iii. Split the node into 2 daughter nodes
2. Output the ensemble of trees (the random forest) $\{T_b\}_1^B$

For the classification problem, the resulting class will be the class that has the majority vote from all the decision trees in the forest. For the regression problem, the end result is the average of all decision trees' results.

In various studies, the random forest approach was found to yield a pretty good classification performance. Moreover, due to the small number of parameters, the tuning process of a random forest is much simpler and cost effective than that of other popular machine learning methods such as neural network (NN) or support vector machine (SVM). In fact, there are only two parameters that need to be optimised in the tuning step for a random forest model: the number of decision trees and the number of variables used to build each tree.

Another advantage of random forest over NN and SVM is the interpretability of predictor variables. In data mining and machine learning applications, the impact of various predictor variables onto the model performance is rarely comparable. In most cases, only a small number of variables have significant influence on the response. Therefore, it is quite useful to assess the importance of each input variable in the predictive model. In random forest, the Gini Index is commonly used to evaluate the contribution of predictor variables. The Gini Index at a particular node is defined as:

$$G = \sum_{i=1}^{n_c} p_i(1 - p_i) \tag{2.16}$$

where n_c is the number of classes and p_i is the ratio of class i in the dataset. The importance is then calculated as:

$$I = G_{parent} - G_{split1} - G_{split2} \tag{2.17}$$

The overall variable importance is the average of all splits in the forest.

A common problem that many machine learning methods have to deal with is the imbalance of datasets where the number of observations in a particular class is heavily dominant in the entire dataset. This circumstance could significantly affect the performance of the prediction model and one has to devise various strategies to overcome this problem such as collecting more data, data sampling, etc. However, Dittman *et al.* has shown that the random forest approach is quite robust against the imbalanced data problem, at least for bioinformatics data [56, 57].

2.3 Proteomics Data Analysis

Due to the advances in molecular biotechnology, biological data have become easier and cheaper to generate with greater volumes and various types. As a result, omics data analysis techniques need to be improved and need to become more accurate to adapt and accommodate the exponential growth of the data. In genomics,

transcriptomics or proteomics studies, one of the most common tasks is evaluating the expression levels of genes, RNAs or proteins in different conditions, which is often referred to as *differential expression analysis*. However, before analysing the expression levels of genes/proteins, the pre-processing data procedure is also quite important for the accuracy of the analysis. Usually, the most important steps in pre-processing data are data normalisation and imputation of missing data.

2.3.1 Data Normalisation

Normalisation is an essential procedure in the analysis of omics data. The main goal of normalisation is to remove the non-biological variation between samples (or between different arrays of a microarray experiment). In other words, normalisation helps canceling the technical variation while keeping the biological variation untouched. Many normalisation methods have been developed, however, which method is the most suitable and gives the best result depends on various factors such as the type of data, the design of the experiment, the assumptions made about the data, etc. Therefore, to determine the best method, one has to try several methods and visually inspect the results with the controls. The common normalisation methods are:

Scale normalisation

This is probably the simplest normalisation method. Basically, the method simply adjusts the scale of the data, for example, by setting the range of the data from 0 to 1 using the following formula:

$$x_{normalised} = \frac{x - \bar{x}}{x_{max} - x_{min}} \tag{2.18}$$

where x_{min} and x_{max} are the minimum and maximum values in the dataset; or by setting the median to 0:

$$x_{normalised} = x - x_{median} \tag{2.19}$$

Quantile normalisation

The objective of quantile normalisation is making two or more distributions of expression values to be identical with respect to statistical properties. The idea comes from the quantile-quantile plot technique, which shows that the two distributions are identical if the plot forms a straight diagonal line. The quantile normalisation could be carried out with or without a reference distribution.

For quantile normalisation with a reference distribution, given the X matrix with $p \times n$ dimensions that need to be normalised and the reference vector X_{ref} with p elements, the procedure consists of the following steps:

1. Sort each column in X to give X' and X_{ref} to give X'_{ref}.

2. Get $X_{normalised}$ by setting the highest ordered entry in each column of X' as the value of the highest ordered entry in the reference vector X'_{ref}, then move on with the next highest ordered entry and so on, until the $X_{normalised}$ is a perturbation of the reference distribution X_{ref}.

For quantile normalisation without a reference distribution, the procedure is slightly modified [58]:

1. Sort each column in X to give X'.

2. Calculate the mean in each row of X' and assign the mean to each element in the row to get X'_{mean}.

3. Get $X_{normalised}$ by rearranging each column of X'_{mean} to have the same order as the original X.

Loess (or Lowess) normalisation

Firstly introduced by Dudoit *et al.* [59], the main idea of this method is based on the M/A plot (Fig. 2.6) and Loess local weighted regression method [60]. M and A stand for the difference and the average of the log expression values, respectively. The normalisation is applied for 2 samples (or arrays) at a time. The normalised data points should scatter around the $M = 0$ axis (Fig. 2.6, right image). The normalisation procedure for a matrix X with $p \times n$ dimension can be summarised as follows:

1. For any two columns i, j with expression values x_{ki} and x_{kj}, where k stands for the kth of p rows (which represent genes/probes/proteins in the experiment), calculate the difference $M_k = log_2(x_{ki}/x_{kj})$ and the average $A_k = \frac{1}{2}log_2(x_{ki}x_{kj})$.

2. Fit a loess curve to the M/A plot using the loess regression method to obtain the normalised curve $\hat{M}_k$.

3. Calculate the normalisation adjustment by $M'_k = M_k - \hat{M}_k$ and the normalised expression values can be obtained by $x'_{ki} = 2^{A_k + \frac{M'_k}{2}}$ and $x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$

4. Repeat the whole procedure with all distinct pairwise combinations of columns in the matrix X.

Since this method iterates over all distinct pairs of samples, therefore, it is quite time consuming compared to other methods.

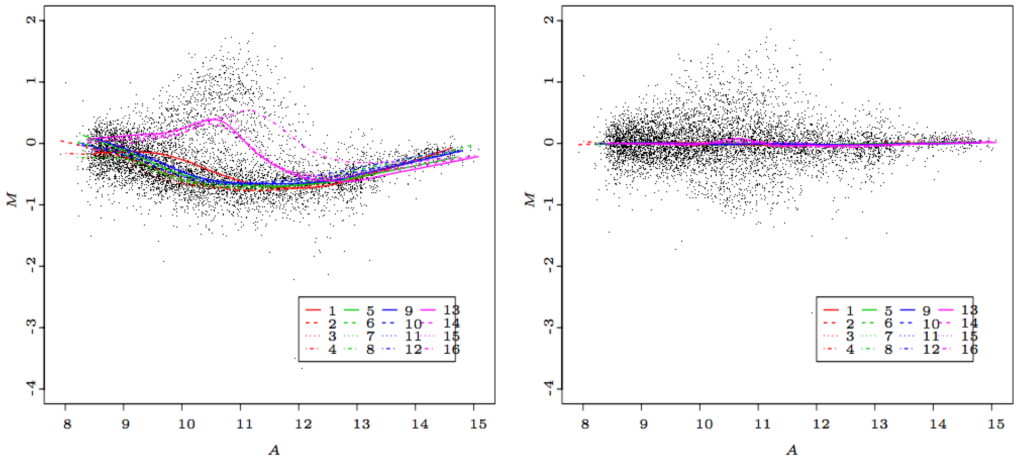


Figure 2.6: Loess normalisation on an cDNA experiment. The coloured lines represent loess curves from different samples. The left image and the right image represent the dataset before and after normalisation. Image taken from [59].

Housekeeping genes

Housekeeping genes are genes that are essential for basic cellular functions. Under normal conditions, they are expressed at a relatively constant rates in all cells. Based on their unique properties, housekeeping genes can serve as "references" to scale other genes. To increase the accuracy and reliability, multiple housekeeping genes are used to normalise the data. However, the expression of some housekeeping genes may vary due to different experimental conditions. Therefore, housekeeping genes should be chosen with care.

2.3.2 Data Imputation

Missing observations are frequently encountered in biological data. In proteomics data from Mass Spectrometry (MS) experiments, there are several reasons which may cause missing data: the peptide is actually present, but it's not detected or falsely identified; or the peptide abundance level is below the detection limit of the instrument; or the peptide is not present at all. The easiest solution for this case is simply ignoring the missing data. However, this could also mean discarding a portion of potential data, and as a result, could significantly affect the result of downstream analysis. Many imputation methods have been developed to tackle this problem, but due to the complicated nature of the missing data, different imputation methods should be used depending on which mechanism that led to missing data. In the scope of this thesis, we applied two imputation methods to handle two different cases of missing data:

Completely missing

The "completely missing" case is the situation when the protein/gene entries do not have any valid data at all. For such cases, low values of expression level are generated based on the current data distribution. The main idea behind this method is to impute missing proteomics data which have expression below the detection limit of the instruments. This method is integrated in Perseus, a proteomics analysis software which is developed and maintained by Tyanova *et al.* [61]. However, visual inspection must be carried out after imputation to accordingly adjust the imputed distribution parameters until desirable results are obtained. Fig. 2.7 shows the imputation with different parameters. The left plot depicts the imputation where the mean is set equal to the overall distribution's mean. Therefore, this strategy does not mimic case when missing data caused by detection limit. The right plot depicts imputation with extremely low values which yields an undesirable bi-modal distribution. The central plot comes from the correctly imputed data with the assumption of low abundant proteins/genes giving rise to missing values.

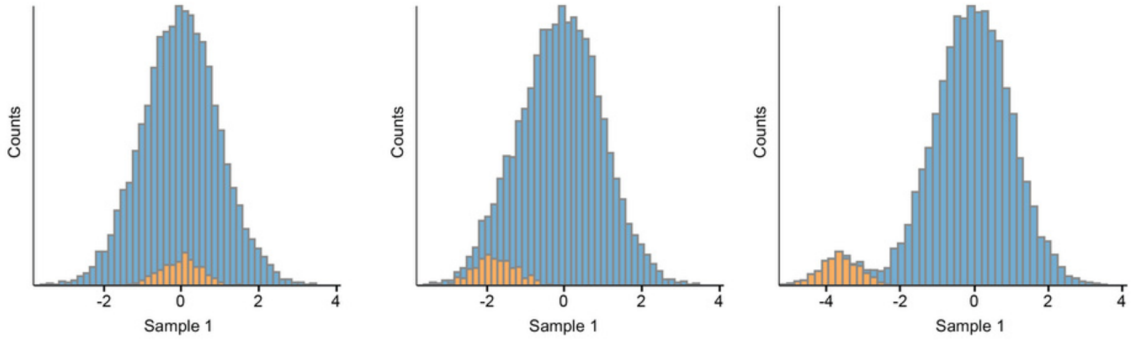


Figure 2.7: Imputation method based on given data distribution. The orange bars represents the imputed data while the blue bars depicts the overall distribution. Three plots come from three imputations with different parameters. Image taken from [61].

Partially missing

In contrast to the "completely missing", the case of "partially missing" data contains at least one valid data point. The safe solution for this case is to impute the new data based on the existing data. The Local Least Squares (LLS) imputation method, which was introduced by Kim *et al.* [62], was used to deal with this problem. Assuming we need to impute missing data in a $m \times n$ DNA microarray experiment, the LLS method consists of 2 steps:

1. Select k genes that are correlated to the missing data by Pearson correlation or ℓ^2-norm. The selected genes are called k-nearest neighbour genes.

2. Apply the local least squares for imputation:

   (a) Based on the k-nearest neighbour genes, derive the matrix $A \in \mathbb{R}^{k\times(n-1)}$ and the two vectors $b \in \mathbb{R}^{k\times 1}$ and $w \in \mathbb{R}^{(n-1)\times 1}$

   (b) The least squares problems for imputation is formulated as:

$$\min_x \|A^T x - w\|_2 \tag{2.20}$$

   (c) The missing value α is determined as follows:

$$\alpha = b^T x = b^T (A^T)^\dagger w \tag{2.21}$$

   where $(A^T)^\dagger$ is the pseudoinverse of A^T.

In their study, Kim showed that the LLS method outperforms its predecessor, the KNNimpute method. Furthermore, the LLS method is also robust and performs pretty well in multiple situations [63].

2.3.3 Significance Analysis of Microarrays – SAM

Modern omics technologies such as microarray or MS can identify and measure the expression of thousands of genes and proteins in a short amount of time and in different experimental conditions. As a consequence, fast and accurate methods are needed to compare and determine the changes or differences in expression between different biological conditions. One of the most popular, well-respected method in differential expression analysis is Significance Analysis of Microarrays (SAM), which was introduced by Tusher *et al.* in 2001 [64]. Initially developed for microarray experiments, SAM identifies genes which are significantly different in term of expression by using a modified t test in conjunction with the False Discovery Rate (FDR) technique. Basically, each gene i is assigned a "relative difference score" $d(i)$ and then compared with a threshold to determine whether it is significantly different or not. The formula for $d(i)$ is:

$$d(i) = \frac{\bar{x}_I(i) - \bar{x}_U(i)}{s(i) + s_0} \tag{2.22}$$

where $\bar{x}_I(i)$ and $\bar{x}_U(i)$ are the average expression level of gene i in conditions I and U, respectively. $s(i)$ is the standard deviation of experiment replicates:

$$s(i) = \sqrt{\frac{1/n_1 + 1/n_2}{n_1 + n_2 - 2}\left\{ \sum_m \Big[x_m(i) - \bar{x}_I(i)\Big]^2 + \sum_n \Big[x_n(i) - \bar{x}_U(i)\Big]^2 \right\}} \tag{2.23}$$

where $\sum_m$ and $\sum_n$ are the summations of expression level in conditions I and U, respectively. n_1 and n_2 are the number of replicates in conditions I and U.

By observing Eq. 2.22, one can see that the $d(i)$ calculation is quite similar to the t test formula except the s_0 term. The s_0 is a small positive constant, which is inserted into the formula to make the variance of $d(i)$ is independent of gene expression, hence, the value of s_0 is chosen in such a way that it minimise the variation of $d(i)$.

Additionally, to tackle the problem of small sample size, instead of performing more experiments, FDR method was applied by permutations of samples and calculate $d_p(i)$ from the permutations. The "expected relative difference" $d_E(i)$ is defined as the average of $d_p(i)$ over all permutations. If the difference of $d(i)$ and $d_E(i)$ is larger than the predefined threshold Δ, gene i is significantly different in terms of expression level in different experimental conditions.

The authors has proven that SAM outperforms the conventional methods of microarray analysis [64]. Even though SAM was initially developed for microarray analysis, it can also be applied to other types of experimental data, for instance, MS proteomics data. In summary, SAM is a robust method and can accommodate various experimental situations.

2.4 Molecular docking and AutoDock

With the development of sophisticated algorithms and software tools and the increasing performance of computing hardware, computer aided drug design has become more and more feasible and crucial for modern drug discovery. In the field of computer aided drug design as well as structural bioinformatics, one of the most popular and well-developed methods is molecular docking. The main goal of docking is finding and evaluating binding conformations between two molecules, usually, a receptor and its ligand. Also, docking can approximate the strength of the binding, the so-called binding affinity, by a scoring function which mimics the potential energy calculations. In this section, we use AutoDock [65–67], a popular docking software, to explain the ideas and algorithm of receptor-ligand docking.

The molecular docking method consists of two main components: the scoring function, which estimates the binding affinity of receptor-ligand binding positions (or *poses*), and the search algorithm.

Scoring function

The scoring function of AutoDock is quite similar to the potential energy function (section 2.1.1) and is composed of 5 terms:

$$\Delta G = W_{vdw} \sum_{i,j} \Big(\frac{A_{ij}}{r_{ij}^{12} - \frac{B_{ij}}{r_{ij}^{6}}} \Big) + W_{hbond} \sum_{i,j} E(t) \Big(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \Big) + W_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{sol} \sum_{i,j} \big(S_i V_j + S_j V_i \big) e^{-r_{ij}^2 / 2\sigma^2} + W_{conf} N_{tors} \tag{2.24}$$

The first term adopts the 12-6 Lennard-Jones potential to represent the van der Waals interaction, while the 12-10 potential mimics the hydrogen bonding with $E(t)$ as a function of the angle t between atoms that form the hydrogen bond. The third term represents the Coulombic potential for electrostatic interactions. The fourth term is the desolvation energy where S stands for solvent accessible surface area of the ligand and V is the volume surrounded by receptor atoms (or the other way around). The last term is the estimation of entropy loss upon ligand binding, which is proportional to the number of rotatable bonds of the ligands. W_{vdw}, W_{hbond}, W_{elec}, W_{sol} and W_{conf} are the weights of each term. To obtain the optimal weight for each term, the authors of AutoDock performed empirical fitting of the scoring function against the experimental binding affinities from receptor-ligand complexes.

There are other scoring functions which are developed by different techniques or concepts such as machine learning based scoring function, knowledge based scoring function and consensus scoring function. Machine learning based scoring functions utilise the nonlinear mapping capability of machine learning methods to map the receptor-ligand interactions to docking scores. In the knowledge based scoring function approach, the occurrence frequency of contacting atoms is observed and converted into a Boltzmann weighted potential by the so-called Boltzmann inversion. However, no single perfect scoring function has been developed yet. Therefore, the main idea of the consensus scoring method is to combine several scoring functions by a voting regime. It thus provides a good agreement of different scoring functions [68–70].

Search algorithm

The Lamarkian genetic algorithm was adapted as a conformational search algorithm in AutoDock. The algorithm imitates the process of population evolution under selection based on Lamarkian ideology, which means the traits acquired during an individual's lifetime can be inherited by its offsprings. A binding pose represents an individual in a population. During a generation, mutations may occur and, as a result, new variants (offsprings) are generated. Only the fittest individual may survive under the evolutional selection. In the docking method, the scoring function plays a role as an external pressure to select the best receptor-ligand binding poses while discarding the worst fitting poses. After several generations with mutations and selection pressure, the population will converge and reach a stationary state with better fitness compared to the original population. The optimal binding poses

are likely present in the converged population. Aside from the Lamarkian genetic algorithm, other searching algorithms have been implemented in docking applications such as Monte Carlo simulation, simulated annealing, ant colony optimisation [71], etc.

Additionally, to shorten the calculating time, AutoDock also precomputes the interaction energies of the receptor with different predefined atom types within a user-defined docking grid box. Therefore, during the calculations of receptor-ligand binding affinity, the estimated binding free energy can be quickly computed by summing up the precomputed values.

Currently, AutoDock only allows partial flexibility of the receptor by treating a list of user-defined residues as flexible ligands during docking, hence, limits the searching result. However, the fully flexible scheme (flexible-ligand-flexible-receptor) is definitely computationally expensive.

Chapter 3

Substrates/Non-substrates classification for the MdfA multidrug transporter

3.1 Background and Motivation

Membrane transporters are a very important class of integral transmembrane proteins that facilitate the exchange of materials between cells and their environments. They are typically grouped into channels/pores, electrochemical potential-driven transporters, primary active transporters, group translocators and transport electron carriers. Each of them plays a vital role for the cell and organism. Overall, approximately 10% of the human genome is related to transporting functions [72]. Due to the strong connection with diseases and abnormal medical conditions, the interaction of transporters with small molecules/drugs has caught a lot of attention and many studies have been carried out to investigate these relationships.

Nowadays, multidrug resistance has become a serious threat to public health. Multidrug transporters such as ABC proteins, AcrB, EmrD, MdfA, etc. play a vital role in the drug defense mechanism of bacteria by expelling drugs out of the cells. They can recognise, attach and eliminate a wide range of chemically unrelated drugs by pushing them out of the cells. Additionally, due to the ability to transport a wide range of chemotherapeutic agents, multidrug transporters also create difficulties in chemotherapeutic treatments. An example for this is the P-glycoprotein 1 (P-gp), an ABC transporter that pumps many foreign substances out of the cells, including cancer chemotherapeutic agents. Consequently, over-expression of P-gp in cancer cells limits the efficacy of anticancer drugs.

Therefore, computer-assisted identification/classification of transporter substrates will make an important contribution to understanding their roles in cells and how they interact with small molecules. Also, it will facilitate selecting better drug can-

didates in the early phase of drug discovery. However, most of the existing substrate classification methods for membrane transporters rely on simple sequence-derived information such as amino acid composition, physicochemical properties, and sequence conservation during evolution [33, 73, 74]. Those descriptors are quite general and do not capture the structural details how membrane proteins interact with their substrates in the internal translocation pores. Such information was simply not available until very recently. Fortunately, this situation has now changed when more and more crystallographic structures of multidrug transporters have been determined. These exciting developments now open up the possibility of integrating structural and dynamic data derived from molecular dynamics simulations of these systems into transporter substrate classification.

3.2 Materials and Methods

The workflow of the substrates classification study is summarised in Fig. 3.1. Basically, the project contains the following steps:

1. Data collection and preparation: in this very first step, the protein structure and its associated ligands were collected and the whole molecular system was prepared.
2. Docking: this step determines the initial binding positions of protein-ligand complexes.
3. System equilibration and relaxation: the main goal of this step is to equilibrate the whole system, keeping it stable and relaxed before the main MD simulation.
4. MD production: the main MD simulation.
5. Analysis: the results were retrieved and analysed.

3.2.1 MdfA structure and collection of ligands

In this chapter, we selected as subject of a substrate classification study MdfA, a well-studied multidrug/proton antiporter of *E. coli*. MdfA contains 12 transmembrane helices, which are divided into 2 pseudo-symmetrical domains, each consisting of 6 helices (Fig. 3.2). Various studies have shown that MdfA can recognise and transport a wide spectrum of substrates with unrelated properties: basic, neutral, lipophilic, hydrophilic, aromatic, zwitterionic, etc. [75,76]. Although many studies were carried out, the transport mechanism of MdfA is still unclear. However, Bibi *et al.* showed that E26 and D34 residues play critical roles in the transport mechanism [77, 78]. Recent findings on the very first X-ray structures of MdfA also emphasised the importance of E26 and D34 as MdfA-ligand binding positions [79].

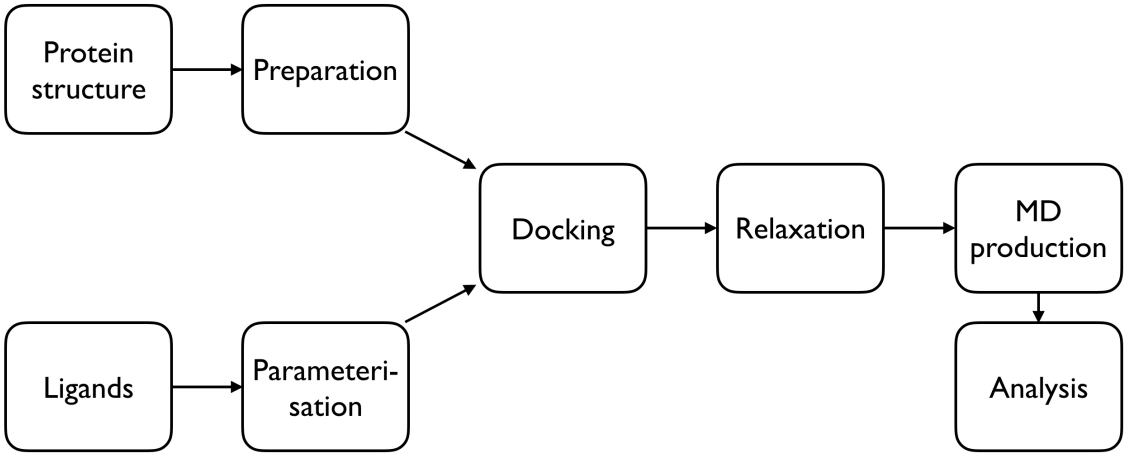


Figure 3.1: The summarised workflow of the substrate classification study.

As it is biochemically well-characterised and of relatively small size (410 amino acids), MdfA is quite suitable for a substrate classification study using MD simulations. One of the very first crystal structures of MdfA from Heng's study [79] (PDB ID: 4ZP0) was chosen as the initial structure for the MD simulations. However, the 4ZP0 structure contains the Q131R engineered mutation. Therefore, the structure was subjected into the Swiss-Pdb Viewer [80] to revert the mutated residue to the original.

Information on known ligands of MdfA was retrieved from various studies [75–78, 81, 82] and their molecular structures were downloaded from CHEMBL [83] and ZINC15 [84] databases. The detailed information on this ligand collection is listed in Table 3.1. Interestingly, there are 2 cases where substrate and non-substrate molecules are highly similar with respect to chemical structure: ethidium bromide vs. propidium iodide; and tetracycline vs. CHEMBL339030, or CHEMBL125158, CHEMBL332172 which are derivatives of tetracycline (Fig. 3.3).

In the scope of this thesis, due to limitations of available computing time, only 16 out of 30 ligands were selected for MD simulations. The 16 selected ligands are composed of 10 substrates, which can be transported by MdfA, and 6 non-substrates, which cannot be transported even though they may bind in the central cavity of MdfA. We also included 4 interesting ligands which are mentioned before: ethidium bromide, propidium iodide, tetracycline and CHEMBL339030. The selection was performed in such a way that the ligands vary with respect to molecular structures and chemical properties. The similarity between two molecules are estimated by the coefficient:

$$\text{Tanimoto coefficient} = \frac{\sum_{j=1}^{k}(a_j \times b_j)}{\sum_{j=1}^{k} a_j^2 + \sum_{j=1}^{k} b_j^2 - \sum_{j=1}^{k}(a_j \times b_j)} \tag{3.1}$$

where a and b are two drug molecules with k dimensions (chemical fingerprint).

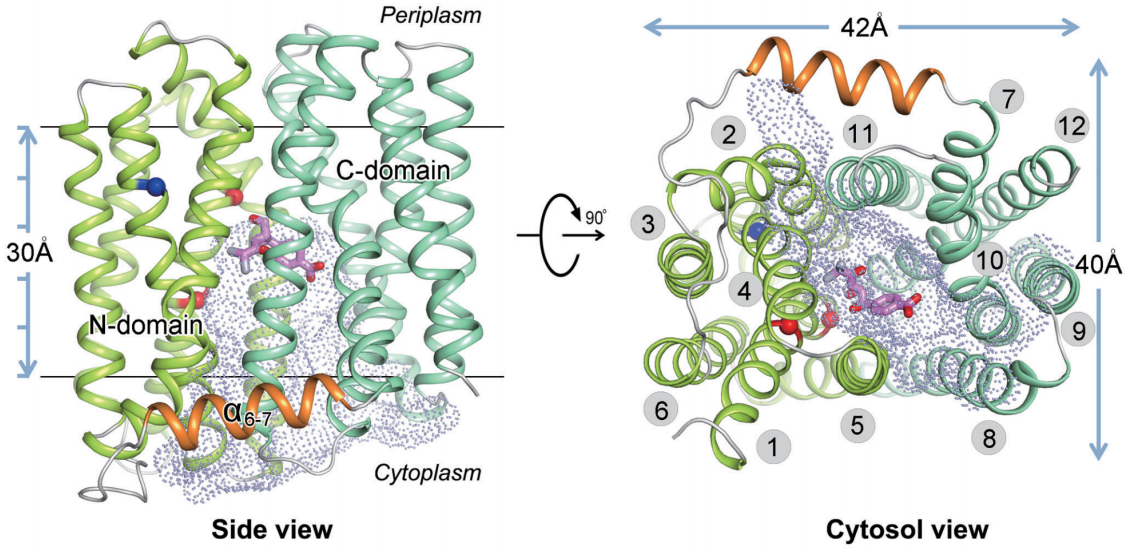


Figure 3.2: X-ray structure of MdfA in the sideview and topview from the cytosolic side. The magenta stick model represent chloramphenicol, one of the most well-known substrate of MdfA. The red spheres represent E26 and D34 residues. The dot-area depicts the inward-facing cavity. Transmembrane helices are numbered in the right image. Image taken from [79].

Table 3.2 shows the Tanimoto coefficient of chemical similarity of the ligands in the selected set. The "Tanimoto avg." is the overall average of the Tanimoto coefficient of a single ligand against the others in the selected set. Taking amp as an example, the Tanimono coefficients of amp with 15 other ligands in the selection are computed, and "Tanimoto avg." is obtained by averaging those 15 Tanimoto coefficients. The Tanimoto coefficient, ranging from 0 (no similarity) to 1 (identical), represents the similarity between two molecules in terms of chemical structures. From the Table 3.2, we can see that the ligands in the selection are quite varied in terms of chemical structures since the averaged Tanimoto coefficients fluctuate around 0.2, which high dissimilarity among the selected ligands.

In summary, for the substrate classification study, we have collected: an MdfA 3D structure (4ZP0); 30 ligands, 16 out of 30 ligands were further selected for MD simulations. The selection, containing 10 substrates and 6 non-substrates, was ensured to vary with respect to chemical structure.

3.2.2 System preparation and equilibration

To accurately simulate the MdfA transporter in its innate environment, the protein was inserted into a pre-equilibrated POPC lipid bilayer membrane model, generated by Visual Molecular Dynamics (VMD) software [85]. All lipid molecules which were located within 0.6 Å from the protein structure after the set-up stage were removed.

Table 3.1: Ligand collection of MdfA multidrug transporter. The charge information was retrieved from the ZINC15 database at pH 7.

Full name	**Abbreviation**	**Charge (e)**	**Substrate/Non-substrate**
Ampicillin*	amp	-1	substrate
Benzalkonium	bzk	1	substrate
Chloramphenicol*	cam	0	substrate
Carbonyl cyanide m-chlorophenylhydrazone	ccc	0	substrate
Chlorhexidine*	chx	2	substrate
Ciprofloxacin	cip	0	substrate
Daunomycin*	dau	1	substrate
Dequalinium	deq	2	substrate
Deoxycholic acid*	dxc	-1	substrate
Ethidium bromide*	ebr	1	substrate
Erythromycin	ery	1	substrate
Kanamycin	kan	4	substrate
Neomycin	neo	4	substrate
Norfloxacin*	nor	0	substrate
Pentamidine*	pen	2	substrate
Puromycin	pur	1	substrate
Pyronin	pyr	1	substrate
Rhodamine 6g	rho	0	substrate
Rifampicin	rif	0	substrate
Thiamphenicol	tpc	0	substrate
Tetraphenylphosphonium*	tpp	1	substrate
Tetracycline*	ttc	0	substrate
4',6-diamidino-2-phenylindole*	dap	2	non-substrate
Diminazene*	dmn	2	non-substrate
Methyl viologen*	mev	2	non-substrate
Nalidixic acid*	nal	-1	non-substrate
Propidium iodide*	pio	2	non-substrate
CHEMBL339030*	tc1	0	non-substrate
CHEMBL125158	tc2	0	non-substrate
CHEMBL332172	tc3	0	non-substrate

(*) Selected ligands for MD simulations.

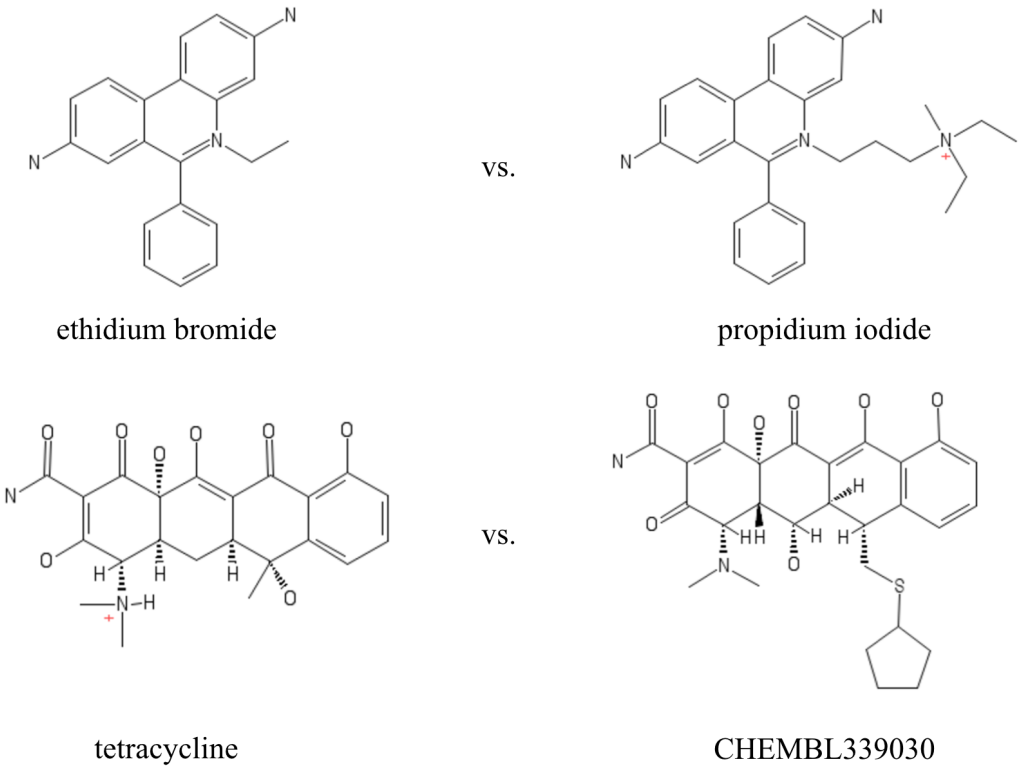


Figure 3.3: Substrates and non-substrates that have similar structures.

Table 3.2: Tanimoto average score of selected ligands.

Ligand	**Tanimoto avg.**	**Ligand**	**Tanimoto avg.**
amp	0.1506	mev	0.1765
cam	0.1910	nal	0.2227
chx	0.1424	nor	0.2216
dap	0.2228	pen	0.1709
dau	0.1430	pio	0.2485
dmn	0.1708	tc1	0.1606
dxc	0.0602	tpp	0.0635
ebr	0.2494	ttc	0.1725

Afterwards, the protein-membrane system was solvated and any water molecules that positioned inside the POPC membrane were removed. Finally, sodium and chloride ions were added to neutralise and maintain the whole system at the physiological salt concentration (0.15 M NaCl). Overall, the complete system has dimen-

sion of 90 Å × 90 Å × 90 Å with 57 764 atoms. This is large enough to satisfy the periodic boundary conditions.

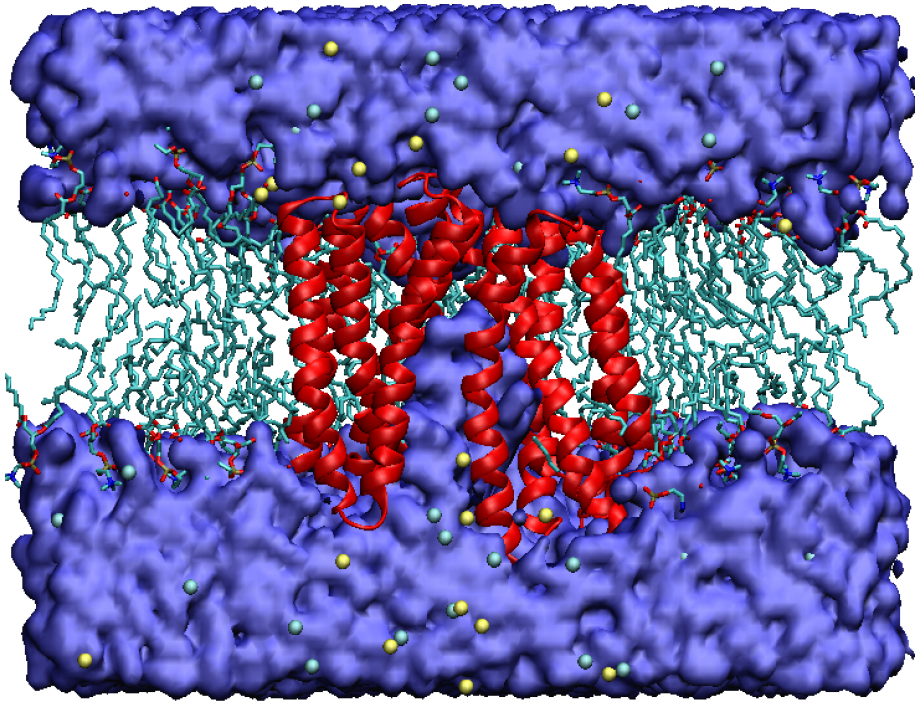


Figure 3.4: Embedded MdfA (red ribbon) in the POPC lipid bilayer membrane (cyan tails) with water (blue mass) and ions (yellow beads - sodium, cyan beads - chloride).

After the complete system (protein, membrane, water and ions) was assembled, it was subjected to a short MD simulation to "melt" the added lipids and system equilibration. Since the membrane patch was generated by VMD, it has not been equilibrated. Therefore, it is essential to perform a short MD simulation in which everything (protein, water, ions, lipid headgroups) except the lipid tails is fixed to obtain the disordered, fluid-like model for the lipid bilayer. The "melting" process included 10 000 steps of minimisation, followed by an MD simulation in the NPT ensemble over 1 ns with 1 fs timestep. After the "melting" run was completed, the system underwent a further equilibration over 2 ns with harmonically restrained protein, followed by a 10 ns simulation without any restraint at all (free system). The resulting system is depicted in Fig. 3.4. In this thesis, we adopted the NAMD package [86] and CHARMM36m force field [87] for MD simulations. Temperature and pressure during simulation were controlled by Langevin dynamics. Further details on simulation parameters are given in the theory section.

3.2.3 Ligand parameterisation

As briefly mentioned in section 2.1.1, force fields such as CHARMM, AMBER, GROMOS, etc. are not suitable for simulating small molecules (ligands, drugs, etc.) because their parameters only describe a relatively limited set of macromolecules such as protein, DNA, lipids, etc. Therefore, force fields for small, general, drug-like molecules have been specifically developed such as CGenFF and GAFF. However, even those force fields cannot possibly cover every existing small molecule. They only provide the parameters for typical elements of many chemically/biologically relevant small molecules. Hence, to address these limitations, one has to refine or even re-parameterise the parameter set. Since different force fields were developed based on different philosophies, and we have used CHARMM36m force field to simulate the protein-membrane system, thus for consistency, CGenFF was adopted for parameterising the MdfA ligands.

Parameters for a new molecule are evaluated and generated by CGenFF based on a set of parameters from pre-determined, chemically/biologically relevant molecules. Every new parameter is associated with a penalty score based on the similarity between the input molecule and the pre-determined molecules. In CGenFF, a penalty score > 50 indicates a poor analogy (Fig. 3.5), hence, the input molecule need to be manually re-parameterised. The Force Field Toolkit (ffTK) [88] was designed and implemented for the ease of developing CHARMM compatible force field parameters. Basically, the parameters that need to be optimised for a small molecule are: partial atomic charges, bonds, angles and dihedrals.

```
RESI tpp            0.000 ! param penalty= 259.000 ;
GROUP             ! CHARGE   CH_PENALTY
ATOM P1     PG0     1.952 !  360.525! typer WARNING:
phosphorus not explicitly supported
ATOM C4     CG2R61 -0.410 !  180.737   <--
ATOM C3     CG2R61 -0.249 !   19.992
ATOM C2     CG2R61 -0.115 !    6.325
ATOM C1     CG2R61 -0.115 !    0.000
ATOM C6     CG2R61 -0.115 !    6.325
ATOM C5     CG2R61 -0.249 !   19.992
ATOM C7     CG2R61 -0.410 !  180.737   <--
ATOM C8     CG2R61 -0.249 !   19.992
ATOM C9     CG2R61 -0.115 !    6.325
ATOM C10    CG2R61 -0.115 !    0.000
```

Figure 3.5: A sample parameter file generated by CGenFF. The red arrows indicate the parameters with high penalty (penalty = 180.737).

According to CHARMM philosophy [89], the partial charge of an atom is determined by its interactions with water molecules. The interaction profile was obtained from quantum mechanical (QM) calculations performed by me using the Gaussian software [90]. Subsequently, ffTK evaluates and computes the atomic charges using molecular mechanics (MM) theory until they reasonably fit the QM interaction pro-

file. This two-step procedure is also applied for parameterising bonds, angles and dihedral, where target data is obtained by QM calculations, and followed by parameter validation/optimisation until they match the QM target data at an acceptable threshold (Fig. 3.6).

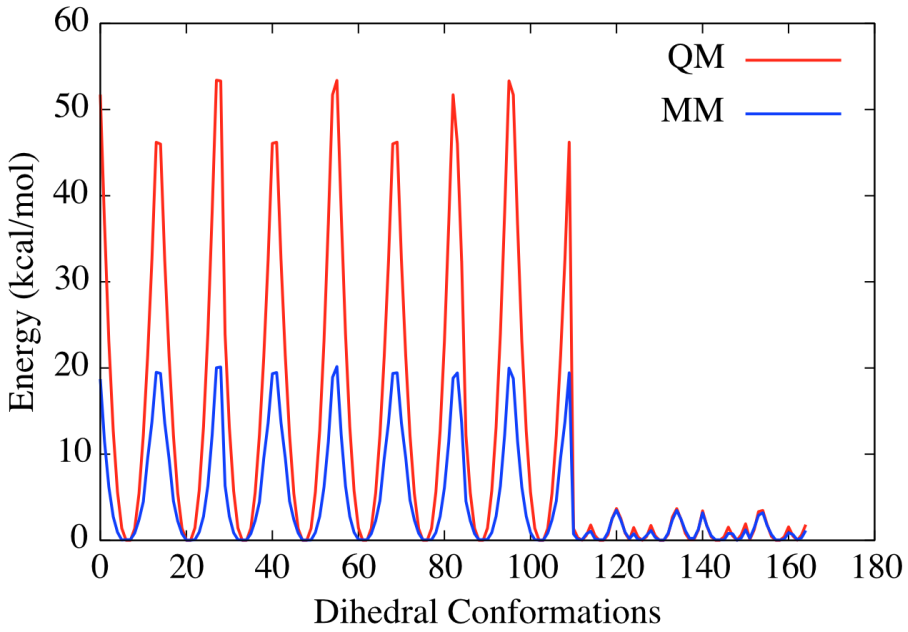


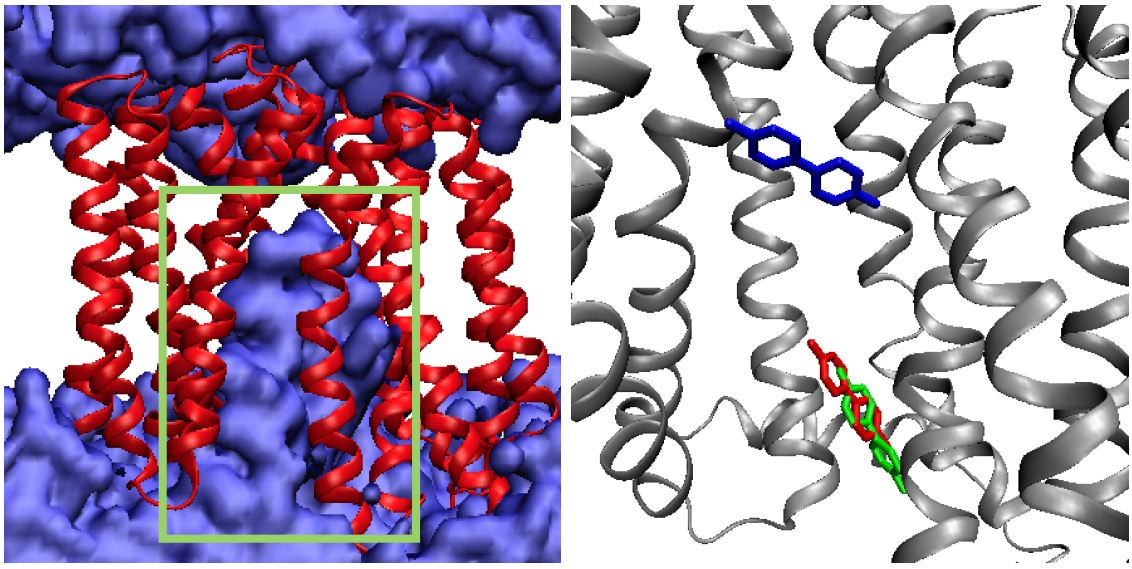
Figure 3.6: Dihedral optimisation by matching MM profile (blue) to QM profile (red).

After the parameterisation for all 16 selected ligands is completed, they are ready for MD simulation. But before the actual MD simulations, we need to place the ligands inside the MdfA central cavity.

3.2.4 Docking and relaxation

We adopted the AutoDock package to determine the initial binding positions of the ligands. To ensure that the ligands are positioned inside MdfA, a docking gridbox (Fig. 3.7a) was defined in such a way that it covers the internal space inside the MdfA structure, hence it only allows the ligands to be placed inside the cavity. The docking was carried out using the default parameters: 2.5×10^6 energy evaluations, 27×10^3 generations.

After the docking was finished, all the resulting poses were clustered using AutoDock based on their similarity, which means ligands that occupy the same space and have a similar orientation are assigned to the same group (cluster). Afterwards, we chose the most favourable pose (lowest estimated binding free energy) in each cluster to be the initial conformation for an MD simulation.



(a) (b)

Figure 3.7: (a) Docking gridbox (green) for MdfA ligands. (b) Different docking poses (red, blue and green) of methyl viologen inside the central cavity of MdfA (gray).

When a ligand is inserted into the protein-membrane system, a short equilibration step (1 ns with 0.5 fs timestep) is required to maintain the stability of the system. Also, water molecules which overlap with the ligand should be removed.

3.2.5 Production MD simulation

After all these preparation steps, we have:

- an equilibrated protein-membrane system
- ligand force field parameters for MD simulations
- initital binding positions of ligands inside MdfA cavity

Since we had assembled all necessary components, the production MD simulations could be prepared and carried out. All production MD simulations were performed in the NPT ensemble over a length of 100 ns using a 2 fs timestep. The trajectory of the whole system as well as many energy terms were recorded. Additionally, for the analysis of energy change upon ligand binding, the protein in its apo form (protein-membrane only) and isolated ligands in a water box were also simulated (100 ns, 2 fs timestep). All production MD simulations were carried out on Titan Cray XK7 of the Oak Ridge Leadership Computing Facility.

3.2.6 MD result analysis

After the production MD simulations were finished, all log files which contain system trajectory and energy terms (Fig. 3.8) were retrieved by `Perl` scripts for further analysis. Information was retrieved for the last 20 ns of the production MD simulations since at the last 20 ns, the ligands had spent some time to accommodate and interact with the protein, in particular, with the PLRs. Therefore, at first, the extracted MD information such as energies, contacts, etc. were the average of the last 20 ns. Because we were investigating protein-ligand interactions, the intra- and intermolecular energies in various states (protein-ligand complex, protein in apo form and isolated ligands) were considered: bond, angle, dihedral, improper, Van der Waals and electrostatic. We also calculated the changes in terms of bonded (bond, angle, dihedral, improper) and non-bonded (VdW, electrostatic) energies of the ligands upon binding to the protein and the protein upon binding of the ligands.

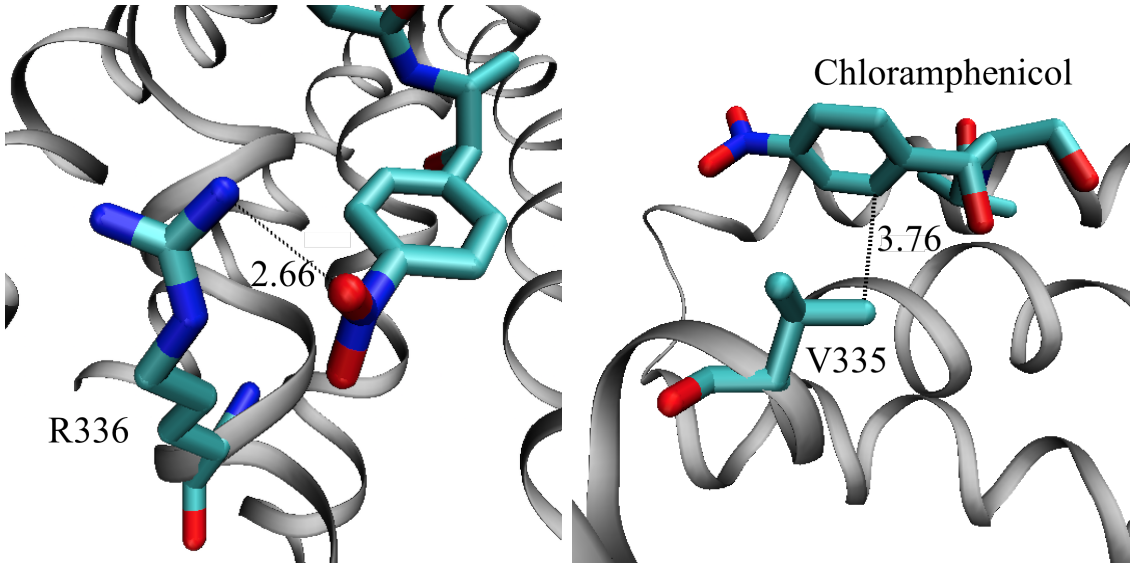
```
ETITLE:      TS           BOND          ANGLE          DIHED          IMPRP
ELECT            VDW       BOUNDARY          MISC        KINETIC
TOTAL          TEMP      POTENTIAL         TOTAL3        TEMPAVG
PRESSURE      GPRESSURE          VOLUME       PRESSAVG      GPRESSAVG

ENERGY: 1033000      3534.8495     16402.5419     11694.8577       297.4074
-146111.2574      1280.4303         0.0000         0.0000     39170.5750
-73730.5957       311.2904   -112901.1707    -73451.9630       311.6126
-125.1031       -83.8856    557139.2523        16.9373        19.7566

WRITING EXTENDED SYSTEM TO RESTART FILE AT STEP 1033000
PRESSURE: 1033100 79.5651 -122.048 79.3791 -122.048 441.117 -259.405 79.378
-259.403 -10.2699
```

Figure 3.8: A sample from a log file which contains records of various energy terms (output from NAMD).

Additionally, we also monitored the contacts between protein and ligands which can be represented by hydrogen bonds and hydrophobic interactions between the Pore-Lining Residues (PLRs) of MdfA and the ligands. The pore-lining residues, which are residues that form the cavity inside MdfA, were determined by PoreWalker [27]. The interaction information was extracted from the trajectory files. Using `tcl` scripting in VMD, hydrogen bond and hydrophobic interactions were defined as follows:

- Hydrogen bond: is defined if hydrogen bond donor and hydrogen bond acceptor between MdfA and ligand are within 3 Å [91] (Fig. 3.9a).
- Hydrophobic interaction: if hydrophobic atoms (C and S) between MdfA and ligand are within 3.9 Å [91] (Fig. 3.9b).

Finally, by utilising this information that we derived from MD simulations (energy terms, interactions), docking result (estimated binding energy), as well as other



(a) (b)

Figure 3.9: Examples for (a) hydrogen bond and (b) hydrophobic interaction between chloramphenicol and MdfA.

molecular descriptors (PaDEL [92]) as predictive features, we trained a substrate/non-substrate classification model using the random forest approach. The classification model was implemented and optimised with the `randomForest` R package [93].

3.2.7 Training and testing scheme for classification model

Similarly to other machine learning methods, one has to come up with an appropriate training and testing scheme to validate the performance of the prediction model and overcome common "obstacles" such as imbalance data or overfitting. Even though the random forest model is unlikely to be susceptible to imbalanced data (see section 2.2), one should pay attention to the overfitting problem while training classification model.

Therefore, we adopted a modified Leave-One-Out Cross-Validation (LOOCV) scheme. Normally, in the LOOCV approach, for data with n observations, only one observation is set aside to be the test set (unknown data), the rest $(n-1)$ will become the training set to build the prediction model. The performance of the model will be validated by the single test observation. Iteratively, in the next round, another observation will be left out to be the test set until all observations in the data have become the test set once. The overall performance is the average performance of all models over the iteration.

In our situation, let us assume that we have n MD simulations (observations) from n protein-ligand poses, which represent the docking results of 16 selected ligands (clustered docking poses). Apparently, one ligand could have more than 1

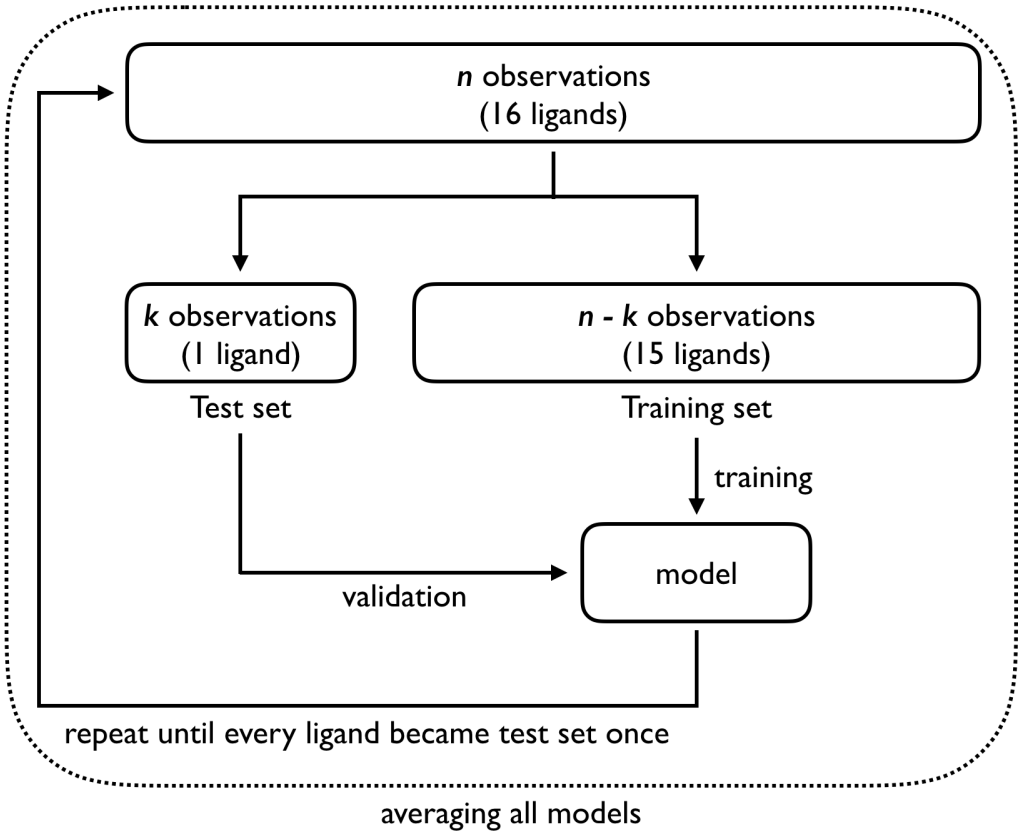


Figure 3.10: Training and testing scheme for classification model.

binding pose. Thus, if we only exclude one observation, the training set still contains information of the left out data due to the other relevant binding poses of the same ligand. Therefore, instead of following the conventional LOOCV, we decided to put aside all binding poses of a particular ligand to be the test set, and the rest will become the training set. The modified LOOCV is illustrated in Fig. 3.10.

The performance of the prediction will be assessed by the common measurement of accuracy:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.2}$$

where TP, TN, FP and FN stand for the number of true positives, true negatives, false positives and false negative, respectively.

3.3 Results and Discussion

3.3.1 Docking result and feature extraction

The docking runs gave 94 binding positions for the 16 selected ligands. The detailed number of binding positions for 16 ligands are given in Table 3.3.

Table 3.3: Number of binding positions for 16 selected ligands inside the MdfA cavity.

Substrates	**No. of poses**	**Non-Substrates**	**No. of poses**
amp	4	dap	6
cam	7	nal	4
chx	9	mev	3
dau	8	dmn	9
dxc	4	pio	7
ebr	3	tc1	8
nor	3		
pen	10		
tpp	2		
ttc	7		

Total: 94 (57 for substrates and 37 for non-substrates)

Consequently, there are 94 protein-ligand complexes for MD simulations, in addition to the protein in apo form and 16 isolated ligands. In total, we have conducted 11 100 ns (or 11.1 μs) of MD simulations. The isolated ligand simulations are much less compute intensive due to the obviously small size compare to the full-size system (protein-membrane-ligand).

The PoreWalker software identified 74 PLRs inside MdfA and we have collected 2756 molecular descriptors from PaDEL. Including various energy terms that we extracted from the MD simulations and docking runs, in total , we obtained 2929 features for the substrate classification model. The feature set includes:

- 8 energy terms from protein-ligand complex simulations with respect to bond, angle, dihedral, improper, VdW, electrostatic and their combinations (*bonded* = *bond* + *angle* + *dihedral* + *improper* and *non-bonded* = *VdW* + *electrostatic*).
- 8 energy terms which describe the changes of a ligand upon binding: $\Delta E_{\text{lig}} = E_{\text{complex}} - E_{\text{isolated\_ligand}}$

- 8 energy terms which describe the changes of the protein upon ligand binding: $\Delta E_{\text{pro}} = E_{\text{complex}} - E_{\text{pro\_apo}}$
- 148 features of hydrogen bonds and hydrophobic interactions of 74 PLRs
- 2756 molecular descriptors from the PaDEL software
- Docking estimated binding energy

During the pre-process of implementing the classification model, we removed those features which have more than 50% invalid data or are correlated with other features.

3.3.2 MdfA substrate classification model

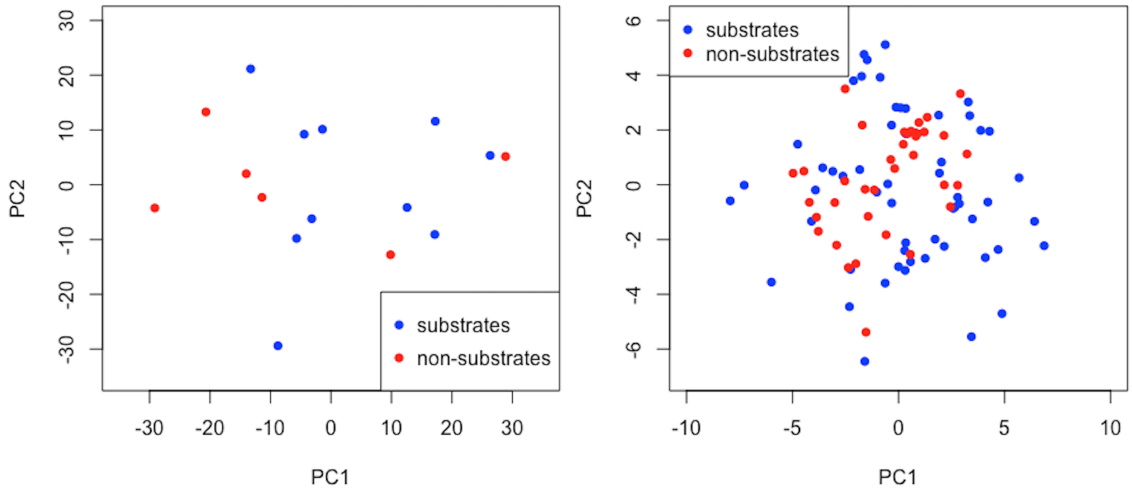
Due to the extremely large number of features (2929) but a relatively small dataset (94), the features were carefully assessed to optimise the performance of the classifier. In fact, a "naive" model was initially built based on the whole feature set but its accuracy was only slightly better than random (overall $ACC \approx 56.2\%$).

PaDEL molecular descriptors

Although the majority of features are molecular descriptors from PaDEL($\approx 94\%$), they were simply extracted from the 2D and 3D information of ligands, not considering any interactions between protein and ligands. Hence, the molecular descriptors are somewhat redundant due to the fact that all binding poses of the same ligand have identical molecular descriptors. In fact, there are only 16 distinct sets of molecular descriptors for 16 selected ligands. A Principal Component Analysis (PCA) was carried out on 16 distinct molecular descriptor set and showed that, with such a small dataset (16 data points), there is no clear separation between substrates and non-substrates with respect to PaDEL descriptors (Fig. 3.11a). Additionally, due to the massive amount of molecular descriptors, they may hinder the contribution of other features because only a portion of the feature set will be randomly selected for tree building in the random forest method. Hence, it is beneficial to exclude molecular descriptors from model training.

Protein-ligand interactions and ligand trajectories

Even after removing the PaDEL molecular descriptors, the prediction performance did not improve much. In fact, it was practically similar to the naive model with $ACC \approx 56.6\%$. Therefore, the same PCA analysis was also done for hydrogen bonds and hydrophobic interactions (Fig. 3.11b). Surprisingly, from the PCA result, it is unfeasible to differentiate substrates and non-substrates based on MdfA-ligand interactions in the last 20 ns of the MD simulation. Additionally, in the first principal component, the top contributors are among the hydrophobic interactions and the



(a) (b)

Figure 3.11: PCA analysis of (a) PaDEL molecular descriptors; and (b) hydrogen bonds, hydrophobic interactions between ligands and MdfA.

contribution of the two known important PLRs, E26 and D34, are not substantial at all (Fig. 3.12).

Therefore, to investigate the sources of this observation, we visually inspected the trajectories of all 16 selected ligands during the MD simulations (Fig. 3.13 to 3.16). As one can see in the trajectories, in most of the cases, the ligands were moving forward to and finally populated the same central area of the cavity. These events could somehow explain why Mdfa-ligand interactions are indistinguishable between substrates and non-substrates in the last 20 ns of the MD simulations.

Being aware of the problems that could potentially hinder the classification performance, instead of using the information from the last 20 ns of the MD simulations, we extracted the energies and interaction information from the first 20 ns and started re-training and re-testing the classification model. The model's performance now improved to 73.1% in accuracy, which means nearly 20% improvement compared to the previous model. The details of the model's performance are given in Table 3.4.

The model was further analysed and the variable importance revealed that, unexpectedly, the bonded energy terms were the key players in the substrates/non-substrates classification (Tab. 3.5). In the beginning, we speculated that the difference between substrates and non-substrates may come from the protein-ligand interactions, which means the non-bonded energies such as VdW and electrostatic should be the most important contributors to the classification. However, the analysis of the prediction model has forsaken that theory. Instead, from the obtained result, we could infer that the protein conformation needs to be rearranged in response to accomodating various ligands. And since the protein conformation adjusts differently according to the nature of the ligand, whether it is substrate or non-substrate, which in turn, leads to different bonded energies between protein-substrate and protein-

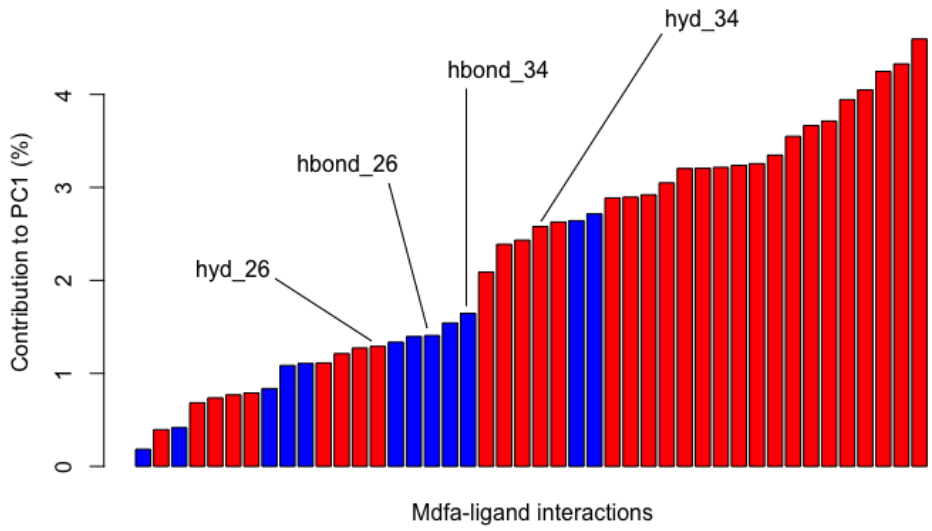


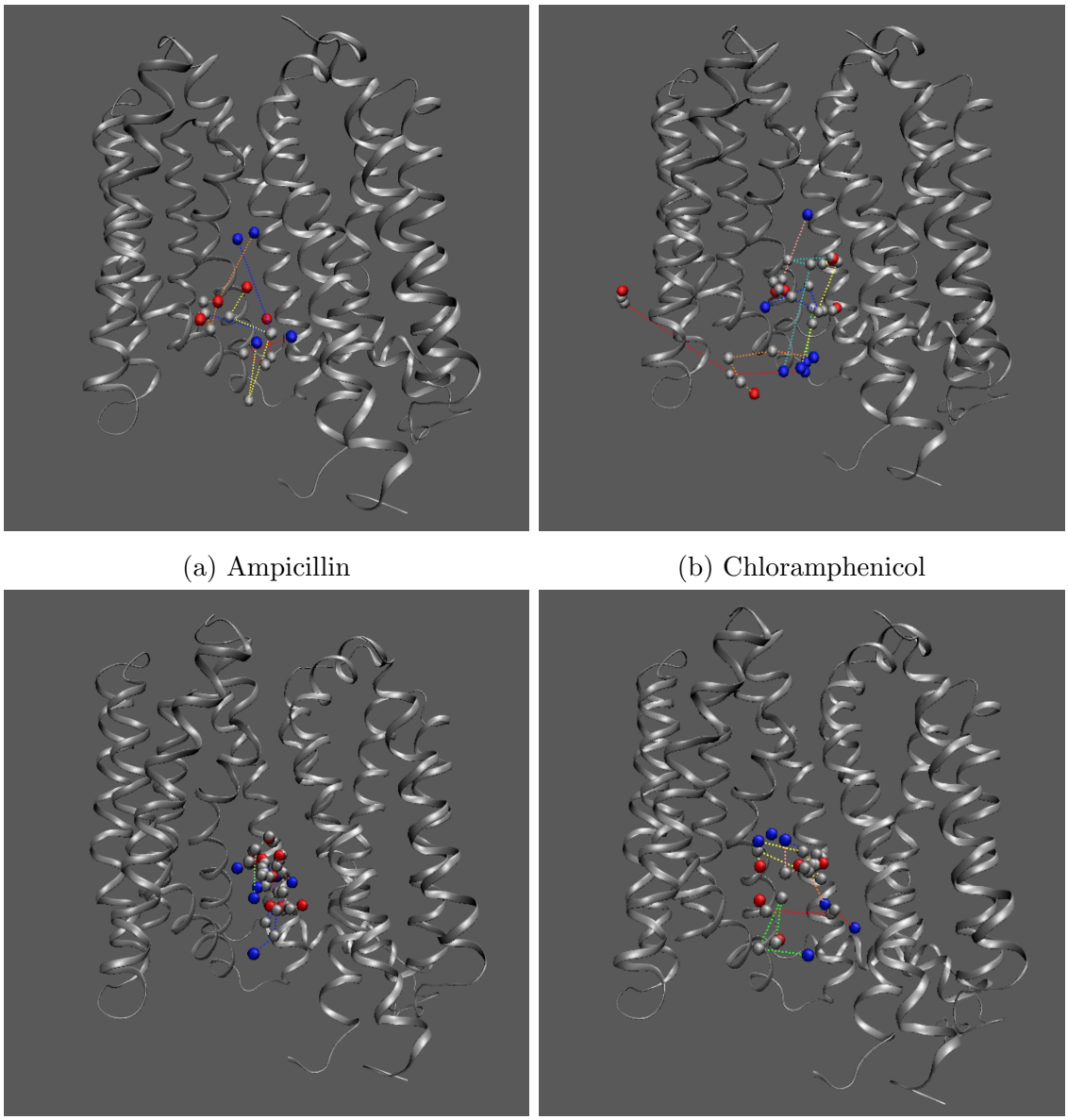
Figure 3.12: The contribution of hydrogen bonds (blue bars) and hydrophobic interactions (blue bars) in PC1. "hyd_26" and "hbond_26" represent hydrophobicity interaction and hydrogen bond contribution at residue E26. The same notation is applied for residue D34.

Table 3.4: Performance of classification models (in accuracy percentage) using the training and testing regime described in section 3.2.7.

Test ligand	**ACC**	**Test ligand**	**ACC**
amp	100.00	mev	100.00
cam	85.71	nal	100.00
chx	100.00	nor	100.00
dap	66.67	pen	30.00
dau	87.50	pio	0.00
dmn	100.00	tc1	0.00
dxc	100.00	tpp	0.00
ebr	100.00	ttc	100.00

Averaged ACC: 73.12%

nonsubstrate complexes.



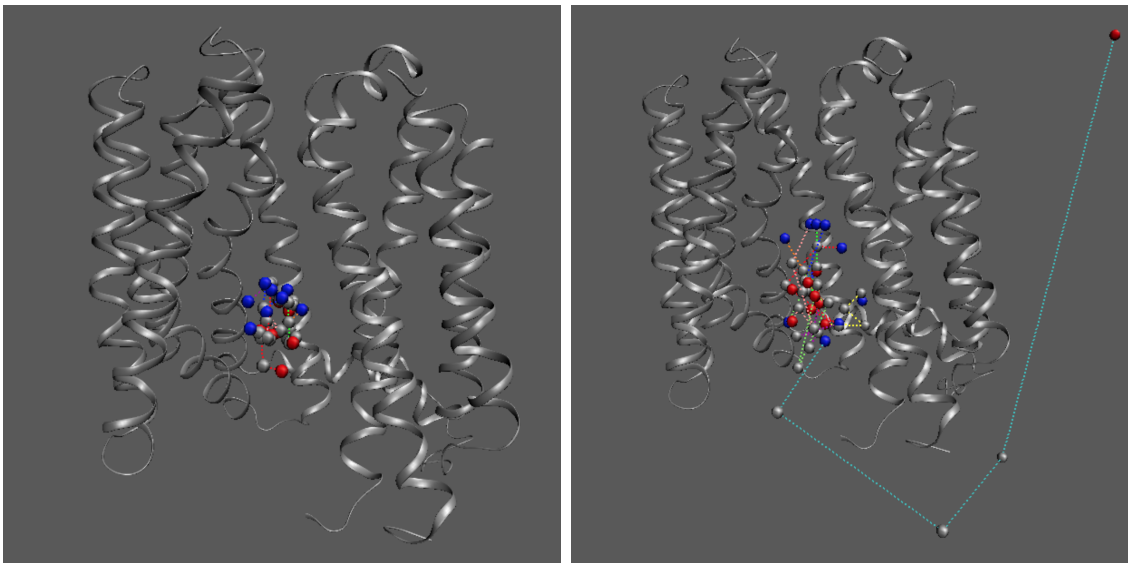
(a) Ampicillin

(b) Chloramphenicol

(c) Chlorhexidine

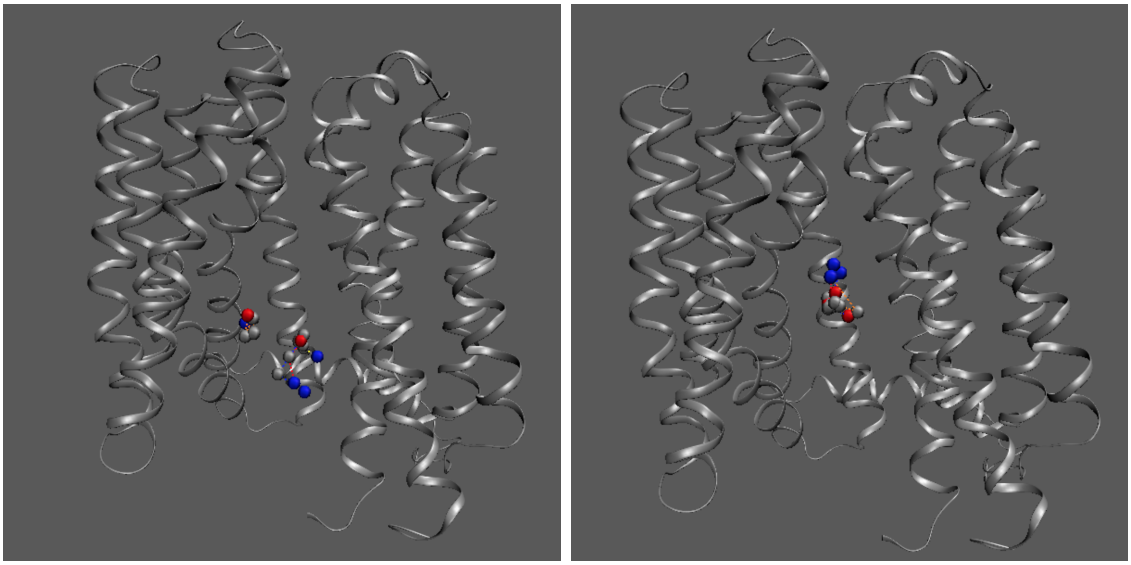
(d) 4,6-diamidino-2-phenylindole

Figure 3.13: Trajectories during 100 ns MD simulation of (a) ampicillin, (b) chloramphenicol, (c) chlorhexidine and (d) 4,6-diamidino-2-phenylindole. The beads represent the center of mass of the ligands. The blue beads indicate the starting positions, the red beads indicate the final positions and the grey beads indicate the intermediate positions. The coloured dashed lines depict the trajectories of different initial binding positions. Lipid molecules and waters are not shown for clarity.



(a) Daunomycin

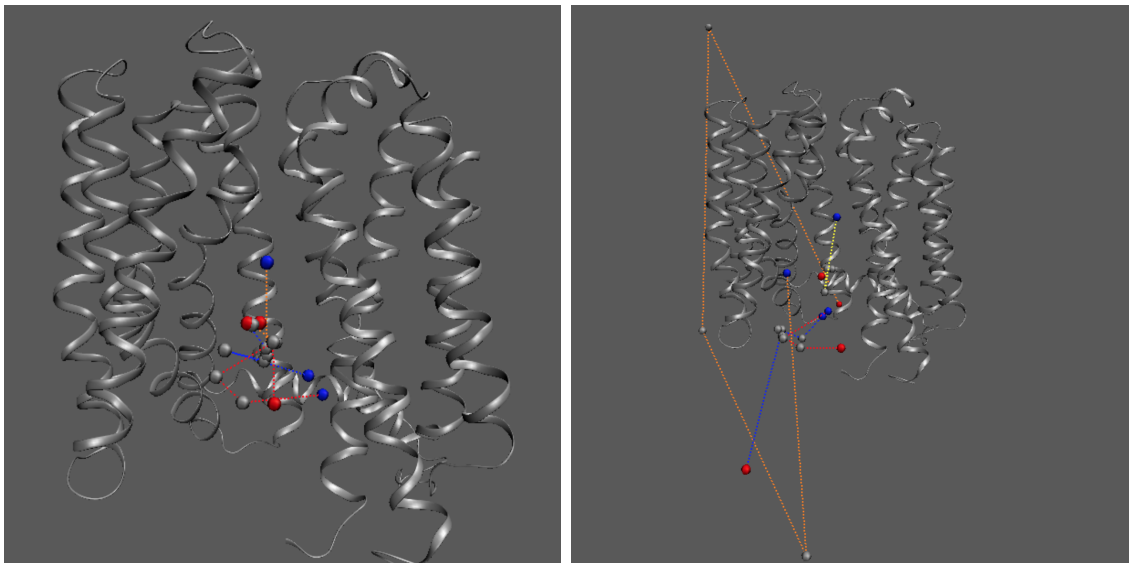
(b) Diminazene



(c) Deoxycholic acid

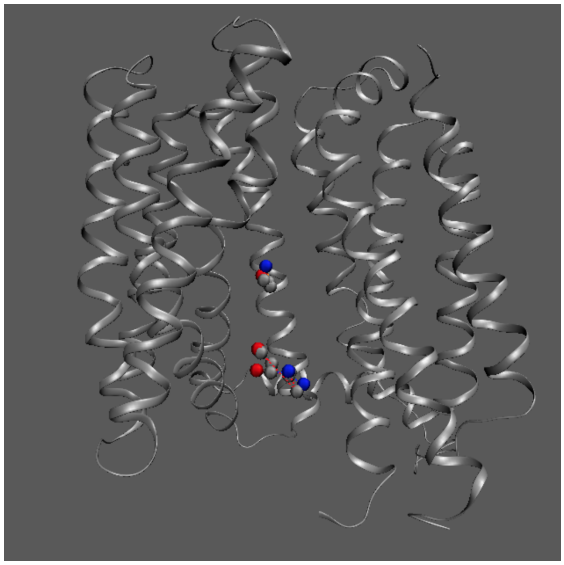
(d) Ethidium bromide

Figure 3.14: Trajectories during 100 ns MD simulation of (a) daunomycin, (b) diminazene, (c) deoxycholic acid and (d) ethidium bromide. The colouring scheme is similar to Fig. 3.13.

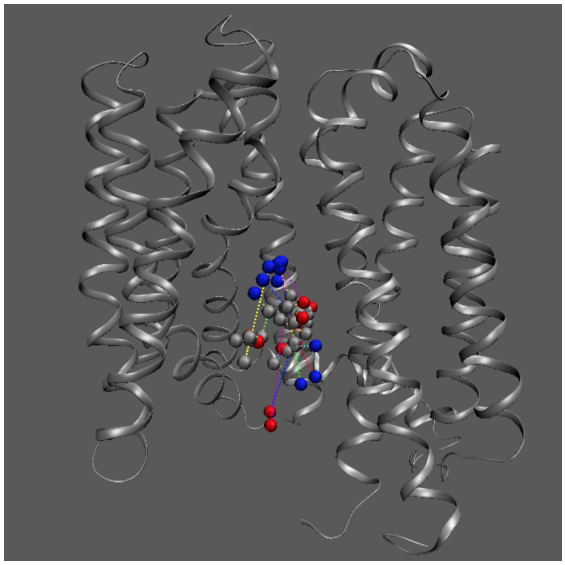


(a) Methyl viologen

(b) Nalidixic acid

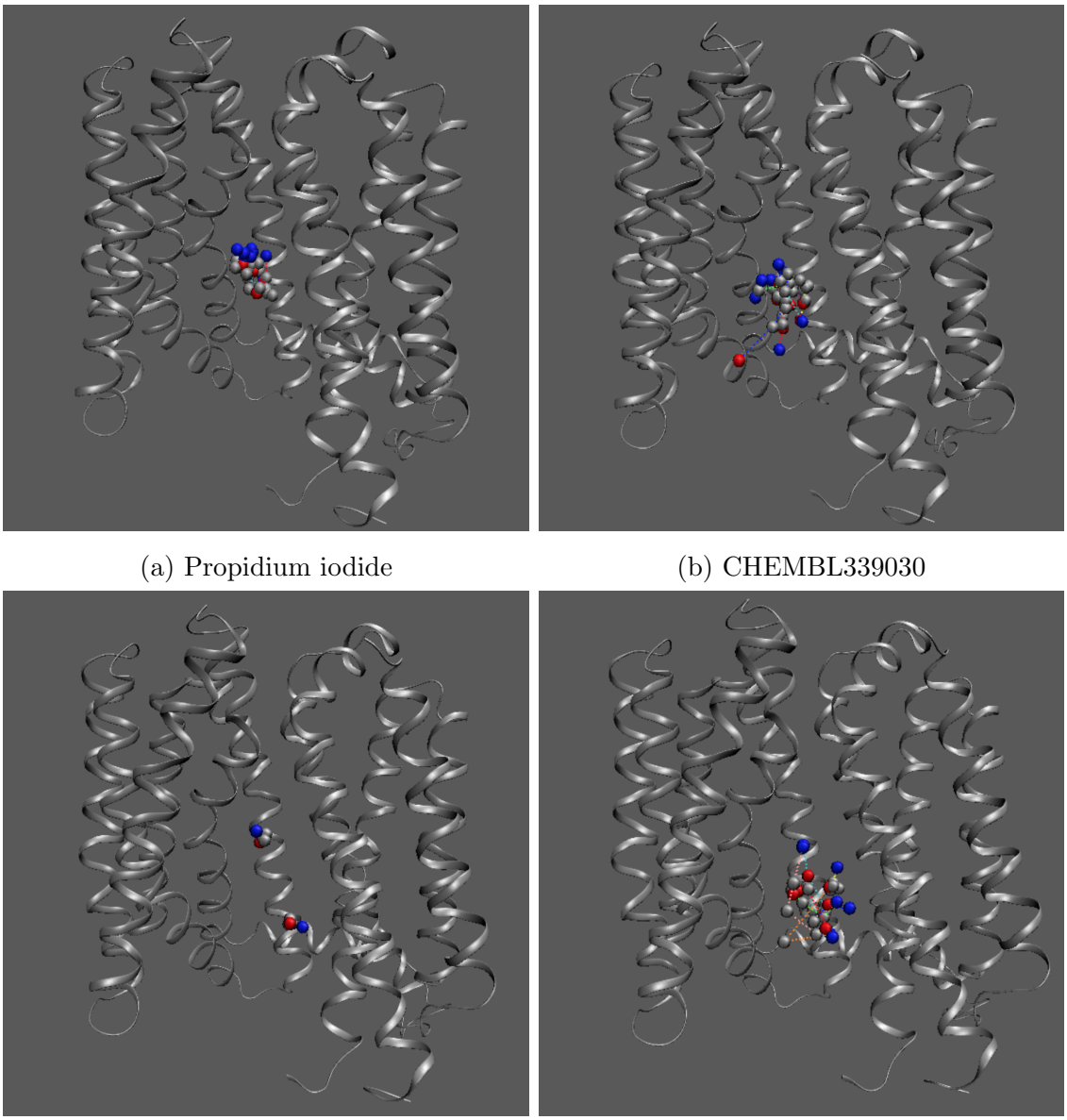


(c) Norfloxacin



(d) Pentamidine

Figure 3.15: Trajectories during 100 ns MD simulation of (a) methyl viologen, (b) nalidixic acid, (c) norfloxacin and (d) pentamidine. The colouring scheme is similar to Fig. 3.13.



(a) Propidium iodide

(b) CHEMBL339030

(c) Tetraphenylphosphonium

(d) Tetracycline

Figure 3.16: Trajectories during 100 ns MD simulation of (a) propidium iodide, (b) CHEMBL339030, (c) tetraphenylphosphonium and (d) tetracycline. The colouring scheme is similar to Fig. 3.13.

Table 3.5: Some of the top important features from the random forest classification model.

Feature	Gini Index	Feature	Gini Index
$E_{\text{complex\_bonded}}$	7.17	$E_{\text{complex\_bond}}$	0.98
$E_{\text{complex\_angle}}$	3.76	$E_{\text{complex\_nonbonded}}$	0.83
$E_{\text{complex\_dihedral}}$	3.62	$E_{\text{complex\_VdW}}$	0.77
$\Delta E_{\text{ligand\_electrostatic}}$	1.87	hyd_127	0.77
Estimated E_{binding}	1.79	hyd_353	0.76
$\Delta E_{\text{ligand\_nonbonded}}$	1.70	$E_{\text{complex\_electrostatic}}$	0.74

$E_{\text{bonded}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}} + E_{\text{improper}}$

$E_{\text{nonbonded}} = E_{\text{VdW}} + E_{\text{electrostatic}}$

Problematic cases

Although the model has been significantly improved, it still performed poorly for several ligands (Tab. 3.4). Interestingly, both problematic cases that we mentioned before (Fig. 3.3) all suffered poor performance. We speculate that those two problematic pairs adversely affect the classification performance due to the structural similarity between substrates and non-substrates. Therefore, we carried out two tests to verify that theory:

1. Training and testing the classifiers without propidium iodide and CHEMBL339030 (both are non-substrates)
2. Training and testing the classifiers without ethidium bromide and tetracycline (both are substrates)

While the performance was significantly enhanced with 89.29% accuracy in the first test (Tab. 3.6), the model in the second test still suffered from a mediocre accuracy at 55.99%. These tests showed that the structural similarity does affect the classifier performance. Moreover, from the result of these tests, we can also infer that the two problematic non-substrates, propidium iodide and CHEMBL339030, do behave similarly to other substrates inside the MdfA cavity, at least in the first 20 ns, which is quite troublesome for substrates/non-substrates classification even with the aid of MD simulations.

Summary

In this project, we developed a novel method for MdfA substrate classification. Unlike other conventional classification methods which utilise general features (sequence derived information, molecular descriptors, etc.), the new method incorporates protein-ligand structural interactions and potential energy information derived

Table 3.6: Performance of classification models in the first test - without propidium iodide (pio) and CHEMBL339030 (tc1).

Test ligand	**ACC**	**Test ligand**	**ACC**
amp	100.00	mev	100.00
cam	100.00	nal	100.00
chx	100.00	nor	100.00
dap	66.67	pen	50.00
dau	100.00	pio	NA
dmn	33.33	tc1	NA
dxc	100.00	tpp	100.00
ebr	100.00	ttc	100.00

Averaged ACC: 89.29%

from MD simulations of different protein-ligand complexes. Although the method encountered difficulties with the structural similarities between substrates and non-substrates, it still reached a decent performance with 73.12% accuracy. However, due to the limit of computational facility, the project was conducted with a relatively small dataset (16 selected ligands) and basic MD simulation setup (only 1 MD run per binding pose). Regardless, this is the first method that considers protein-ligand interactions into a classification problem, hence, paving the way for further developments in drug discovery and other applications.

Chapter 4

Gene silencing combined with quantitative proteomics reveals client spectrum of TRAP complex

This work is submitted as "Nguyen D, Stutz R, Schorr S, Lang S, Pfeffer S, Freeze HH, Förster F, Helms V, Dudek J, Zimmermann R. **Proteomics explains client specificity of the translocon-associated protein in ER protein import**". My contribution in this work was to conduct all proteomics data analysis.

4.1 Background and Motivation

In mammalian cells, the Endoplasmic Reticulum (ER) is involved in protein synthesis, protein folding and acts as a gateway into endocytic and exocytic pathways for the majority of soluble proteins [94, 95]. Typically, proteins are transported into the ER membrane by the ER protein translocon in a co-translational mode. The translocon is composed of Sec61 complex and additional components (Fig. 4.1) which are involved in nascent precursor polypeptides processing and translocation. When the nascent precursor polypeptide emerges from the ribosomes, the Signal Recognition Particle (SRP) recognise Signal Peptide (SP) and Transmembrane Helix (TMH) of the nascent chain. Afterwards, the whole complex, which comprises of ribosome, nascent chain and SRP, is guided to the ER membrane by an SRP receptor (SR) [96, 97]. Then the precursor polypeptide is inserted into the Sec61 complex [98–100], which either happens spontaneously or with the aid of other assisted components [101–104], e.g. the TRanslocon-Associated Protein (TRAP) complex [103–111].

The TRAP complex, originally termed the Signal-Sequence Receptor (SSR), was revealed to be associated with nascent chain [113,114] and Sec61 [106,107]. In an *in vitro* study, TRAP complex stimulates the translocation of many proteins, but not

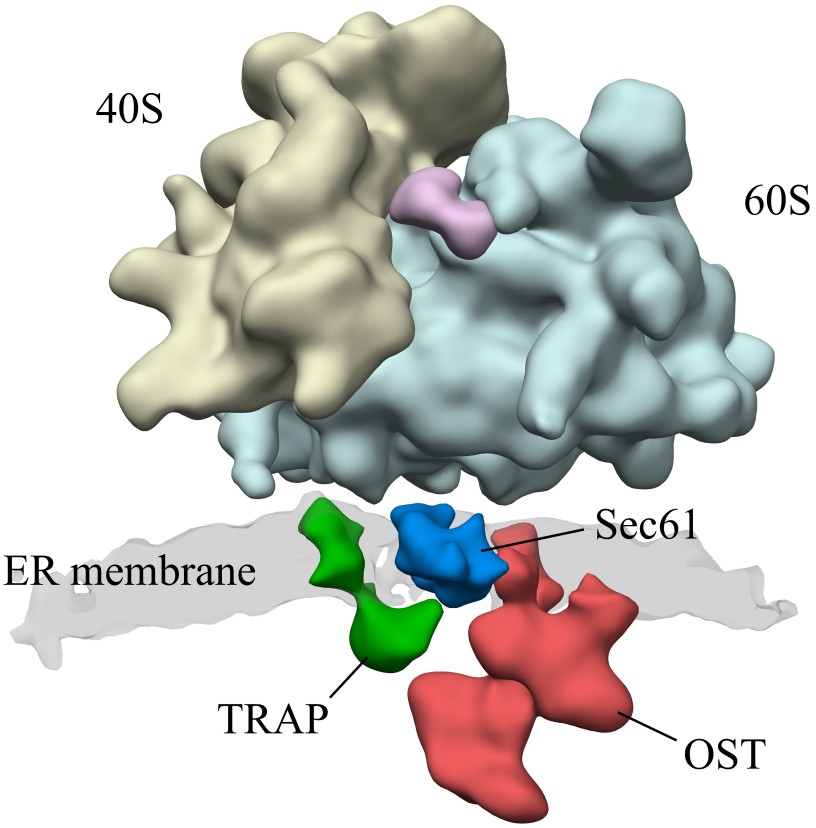


Figure 4.1: Subtomogram average of ribosome/translocon/nascent chain (magenta density) complex. Image is adopted from [112].

all, in a manner that is influenced by their SP [104]. A recent study also showed that TRAP may affect protein topology [111]. Additionally, mutations in human SSR3 and SSR4 (two subunits of TRAP complex) cause the loss of TRAP and Congenital Disorder of Glycosylation (CDG), which suggests that TRAP may be involved in protein N-glycosylation [115, 116].

However, in the mentioned studies, the experiments were designed in cell-free conditions with a small set of synthesised precursor proteins [104], or are biased towards a model precursor. They do not clearly clarify the precursor polypeptides' properties that make them dependent on TRAP complex under normal physiological conditions. In this work, we combined siRNA-mediated gene silencing in HeLa cells with label-free mass spectrometry-based (MS) proteomics analysis and differential expression analysis to properly identify and characterise the TRAP dependent precursor polypeptides in human cells.

The summarised workflow is represented in Figure 4.2. In summary, the workflow is composed of the following steps:

1. TRAP and Sec61α silencing by siRNA-mediated method. Two different siRNAs were used for each target.

2. Perform label-free MS proteomics analysis to obtain the protein abundance profile in HeLa cells in both control and siRNA-mediated samples.
3. Perform differential expression analysis to identify TRAP clients.
4. Characterise TRAP clients.

In this work, my contribution was to perform all analysis in steps 3 and 4.

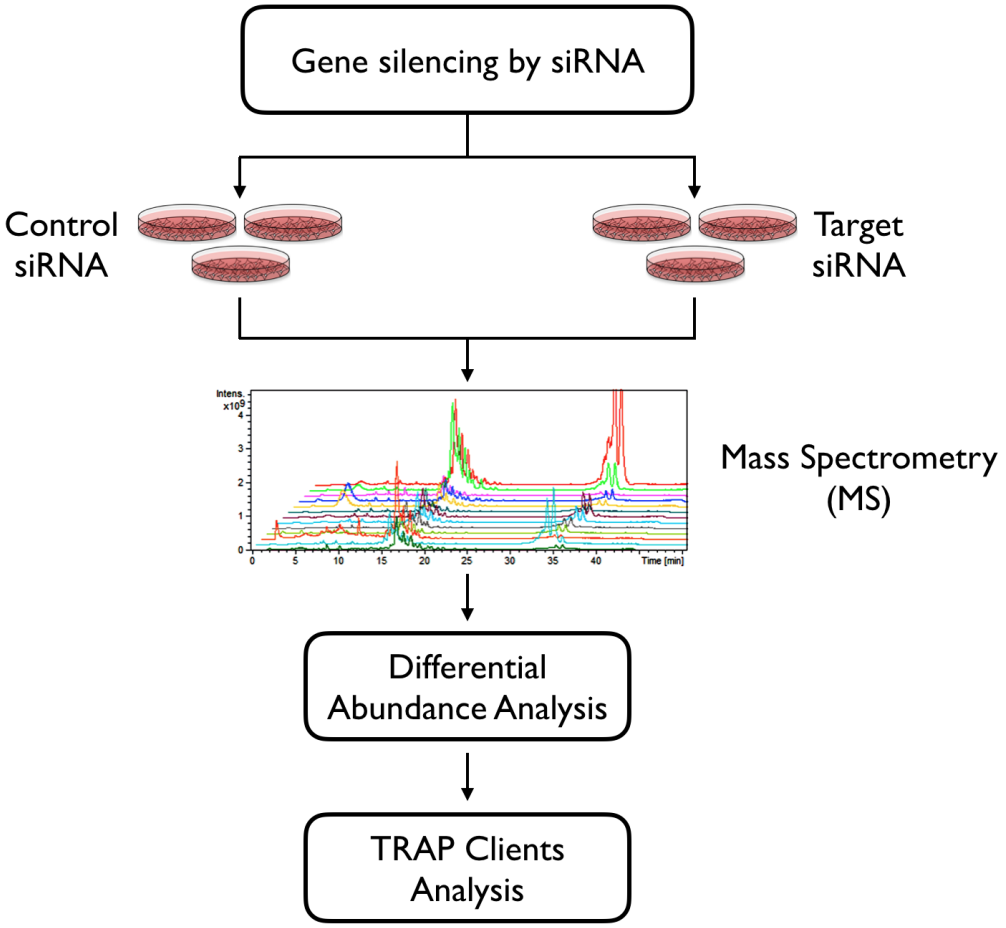


Figure 4.2: The workflow to characterise TRAP clients.

4.2 Materials and Methods

4.2.1 Differential expression analysis

Mass Spectrometry proteomics data pre-processing

The MS proteomics data was prepared and processed by Dr. Nagarjuna Nagaraj (Max-Planck Institue of Biochemistry, Biochemistry core facility, Martinsried, Ger-

many) using the MaxQuant package [117]. The resulting data contains Label-Free Quantification (LFQ) intensities for the whole proteome in the cells. For the ease of calculations, all protein intensities were transformed into $\log_2$ values. A siRNA-mediated gene silencing dataset consists of 3 replicates for each condition: 1 control and 2 different siRNA-treated samples (Fig. 4.3). Two proteins were targeted in the silencing experiments: Sec61α and TRAP. Two independent (but identical) silencing experiments were carried out on Sec61α while there were three independent experiments for TRAP. In total, we have:

- Sec61α silencing experiment:
  - 6 control replicates
  - 12 replicates of 2 different siRNAs
- TRAP silencing experiment:
  - 9 control replicates
  - 18 replicates of 2 different siRNAs

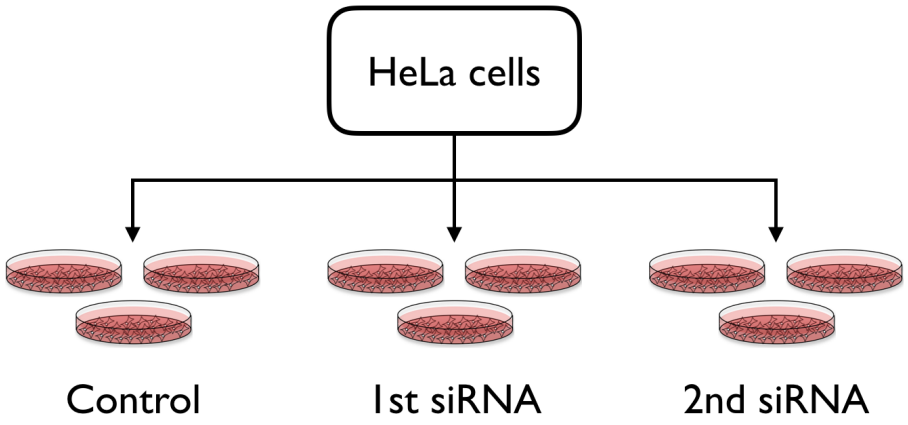


Figure 4.3: The design of a siRNA gene silencing experiment.

Due to the fact that the silencing experiments of the same target protein were conducted independently and at different points of time, even though the experiment setup was identical, the proteomics data from those experiments showed clear differences (Fig. 4.4) due to cell ageing and batch effects. Therefore, the normalisation method should be chosen with care to minimise those effects. In this work we adopted a gene-based quantile normalisation. Compared to the traditional quantile normalisation (see section 2.3.1), instead of normalising proteomics data across all experiments to make them statistically identical, this modified quantile normalisation was executed in a manner that the distributions of a gene across all experiments are statistically identical. This modified method can remove the batch effects that

only affect specific subsets of genes and/or affect different genes in different ways, which cannot be achieved by traditional normalisation methods [118].

Similar to other proteomics data analysis methods, MS also encounters the problem of missing data. In this work, we adopted the data imputation strategy as described in section 2.3.2. However, the proteins that have more than 50% missing data in the control samples will be removed.

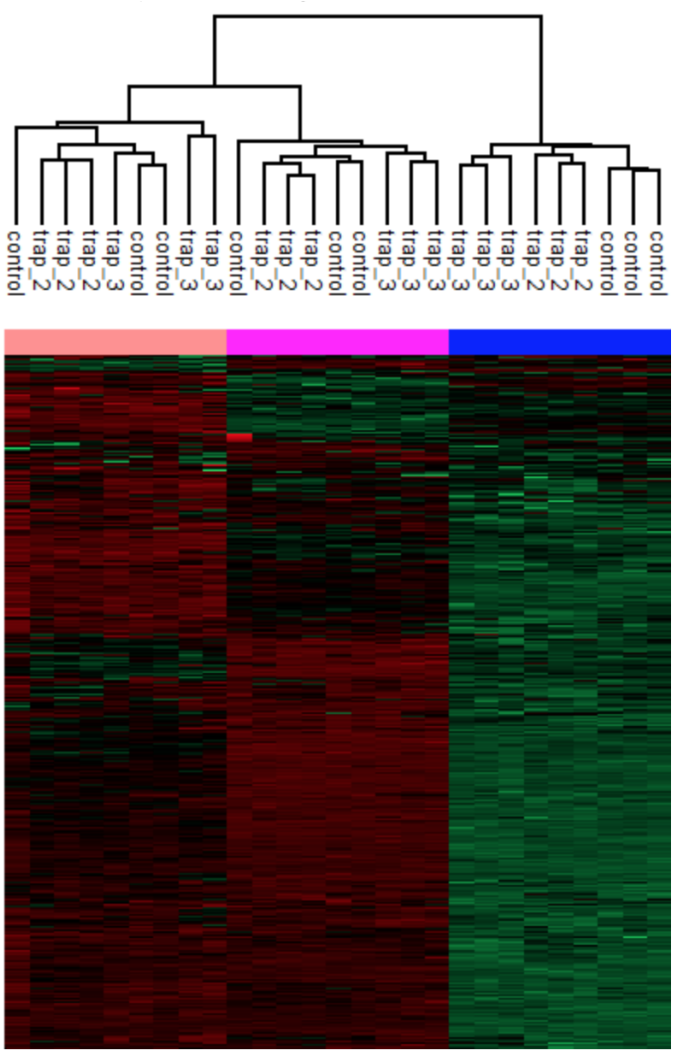


Figure 4.4: Hierarchical clustering heat map of 3 unnormalised independent TRAP silencing experiments. *"control"* columns represent cells in control condition while *"trap_2"* and *"trap_3"* represent cells in two different siRNA-mediated conditions. The three coloured bars indicate three independent experiments: blue - 1st, magenta - 2nd, pink - 3rd.

Differential expression analysis

The differential expression analysis was carried out using the Significance Analysis of Microarrays (SAM) method. Even though the SAM method was originally developed for microarray analysis, it is still applicable for MS proteomics data. By combining modified t-test with False Discovery Rate (FDR) method, SAM has overcome the problem of small sample size in the majority of proteomics data (see section 2.3.3 for more details). In this work, a protein is defined as significantly different (either negatively or positively) if its FDR in comparison against control samples is lower or equal to 5% by SAM test. Since we conducted the silencing experiments with two different siRNAs, only the overlapped significantly affected proteins were considered, e.g. a protein is considered as negatively affected (down-regulated) by TRAP silencing if it is negatively affected in both siRNA conditions.

All the pre-processing steps (normalisation, imputation) and differential expression analysis were done in R with the following packages: `preprocessCore` [119] for normalisation, `pcaMethods` and `impute` for data imputation [120], `samr` [121] for differential expression analysis.

4.2.2 Downstream analysis

Sub-cellular localisation for significantly affected proteins

To verify the effectiveness of silencing experiments and applicability of MS method, we identified the sub-cellular location of all affected proteins. Since Sec61α and TRAP are critical components in the translocon complex that guides proteins into endocytic and exocytic pathways, we expect that those proteins that are associated with ER, golgi, membrane, endosome, secretory pathways, etc. should be affected by the silencing. The sub-cellular localisation was carried out using GO slim annotations [122–124].

TRAP clients characterisation

All Sec61α and TRAP clients/substrates (negatively affected proteins caused by Sec61α/TRAP silencing) which were identified in the previous steps were further analysed for the characterisation. Specifically, we analysed the hydrophobicity and amino acid content of their N-terminal SP and TMH. To obtain the SP and TMH information of affected proteins, the associated curated entries from UniProtKB [125] were downloaded and extracted. Afterwards, the hydrophobicity score and amino acid composition of SP and TMH were derived:

- Hydrophobicity score: is the average score of the accumulated hydrophobicity values across all residues in the SP/TMH sequence. The hydrophobicity value is based on the Kyte-Doolittle hydrophobicity scale [126].

- Amino acid composition: is the contribution of amino acid types to the content of the SP/TMH sequence:

$$\mathrm{AAC}(i) = \frac{\text{Number of amino acids of type } i}{\text{Total length of SP/TMH}} \tag{4.1}$$

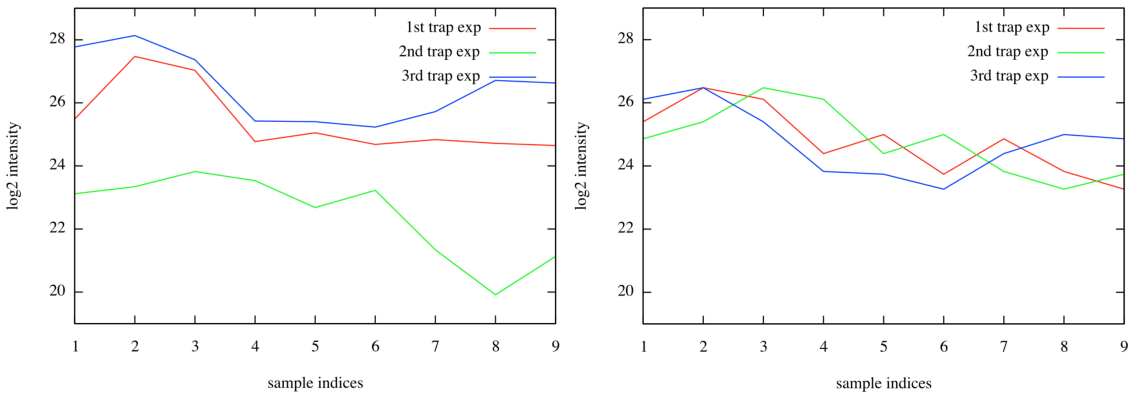
Additionally, the N-glycosylation site information of the substrates was also retrieved from UniProtKB entries.

Furthermore, to investigate the distinct properties of TRAP clients, we also compared the SP of TRAP clients with their homologs in *S. cerevisiae* due to the fact that yeast does not possess TRAP complex. TRAP clients were processed by the BLAST package [127] against 7904 yeast protein sequences from UniProtKB to identify the homologs and their SP.

4.3 Results and Discussion

4.3.1 Data normalisation method

Fig. 4.5 shows the intensity distributions of SSR2, a subunit of TRAP complex, before and after gene-based quantile normalisation.

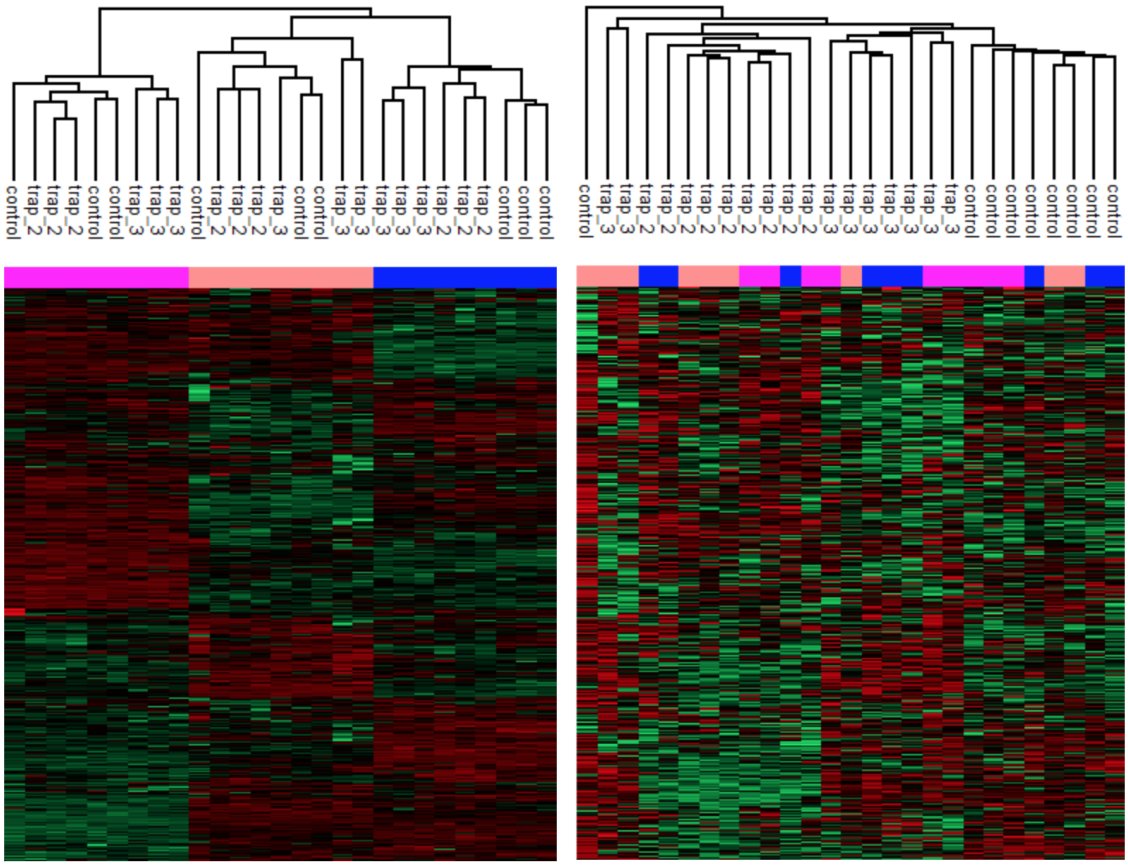


(a) Before normalisation (b) After normalisation

Figure 4.5: The SSR2 intensity profile across all experiments (red - 1st experiment, green - 2nd experiment, blue - 3rd experiment) before and after gene-based quantile normalisation. The horizontal axis indicates sample conditions: 1 to 3 - control, 4 to 6 - 1st siRNA, 7 to 9 - 2nd siRNA.

Based on the visual inspection of the intensity distributions and the clustering heat map (Fig. 4.6b), the gene-based quantile normalisation produces comparable intensity distributions across all experiments even though they were conducted at different points of time. Moreover, the comparison of hierarchical clustering heat

maps indicates that the gene-based quantile normalisation outperforms the traditional quantile normalisation in terms of batch effects correction for the special data sets available to us. As one can see in Fig. 4.6, the gene-based normalisation gives better clustering result with clear separation between different experimental conditions (control, 1st siRNA, 2nd siRNA) while the traditional normalisation still suffers heavily from batch effects. In fact, there is visually no improvement when comparing the traditional quantile normalisation heat map against unnormalised data (Fig. 4.4).



(a) Traditional quantile normalisation (b) Gene-based quantile normalisation

Figure 4.6: Comparison between (a) traditional quantile normalisation and (b) gene-based quantile normalisation. The labels are identical to Fig. 4.4

4.3.2 Sec61α silencing experiments: experimental strategy for substrates identification

The Sec61 complexes in HeLa cells were knocked-out by the introduction of 2 different siRNAs that target Sec61α. A previous study showed that the silencing effects

reduces more than 90% of Sec61 complexes without affecting cell growth, cell viability or cell morphology [102].

About 7000 distinct proteins were identified in Sec61α silencing experiments by MS, roughly 50% of the human proteome. After removal of invalid data, the unavailability of GO annotations and sample data, 5129 proteins that were detected in all experiments were further analysed.

In the comparison of siRNA-treated samples against control samples, we have identified 824 proteins significantly affected by Sec61α depletion: 482 were down-regulated and 342 were up-regulated. Being the target of the silencing experiments, Sec61α was clearly degraded (Fig. 4.7), as expected. In addition, the volcano plot shows that, along with Sec61α, two other subunits of the Sec61 complex, Sec61β and Sec61γ were negatively affected due to the depletion of Sec61α. Also, two subunits of the SRP receptor (SRPRA and SRPRB), which were revealed as compensatory components in a previous Sec61α silencing study [102], were among the positively affected proteins.

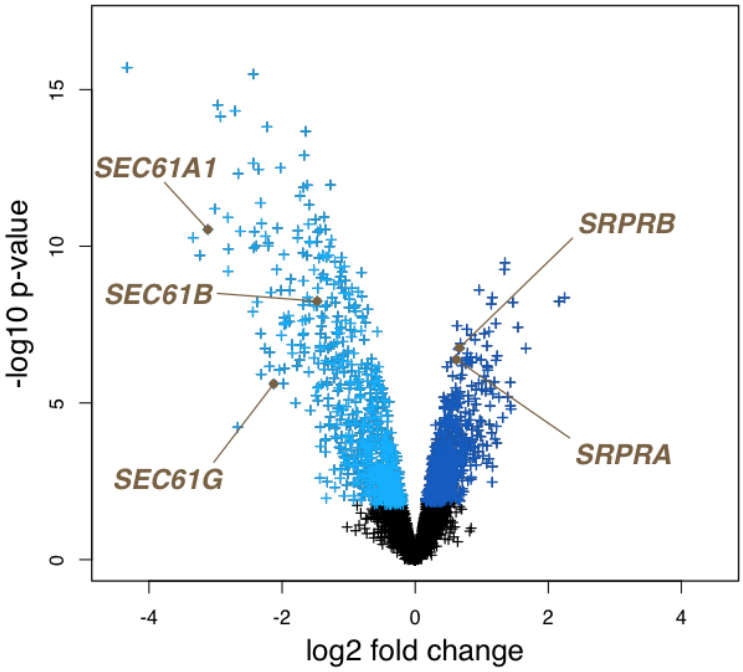
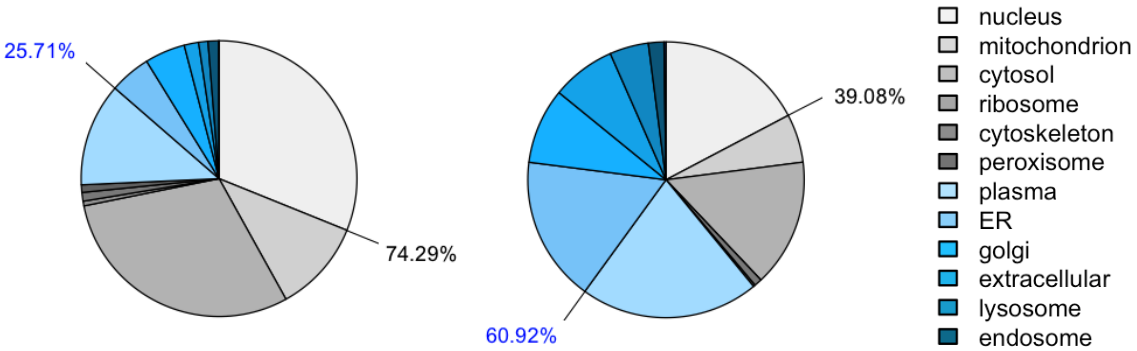


Figure 4.7: Volcano plot of Sec61α silencing experiment. The points represent the whole quantified proteins while the blue points indicate the significantly affected proteins: the light blue points on the left and the dark blue points on the right represent the negatively affected proteins and the positively affected proteins, respectively.

In Fig. 4.8, the sub-cellular localisation analysis has shown that 60.92% of the negatively affected proteins were belong to organelles of the endocytic and exocytic pathways (plasma membrane, ER, golgi, extracellular region, lysosome, endosome and vacuole). This is a more than two-fold enrichment compared to the whole

quantified proteome (25.71%).



(a) Total quantified proteins (b) Negatively affected proteins

Figure 4.8: Sub-cellular localisation of (a) the whole quantified proteome and (b) the negatively affected proteins of Sec61α silencing experiments. The blue parts indicates the organelles of endocytic and exocytic pathways.

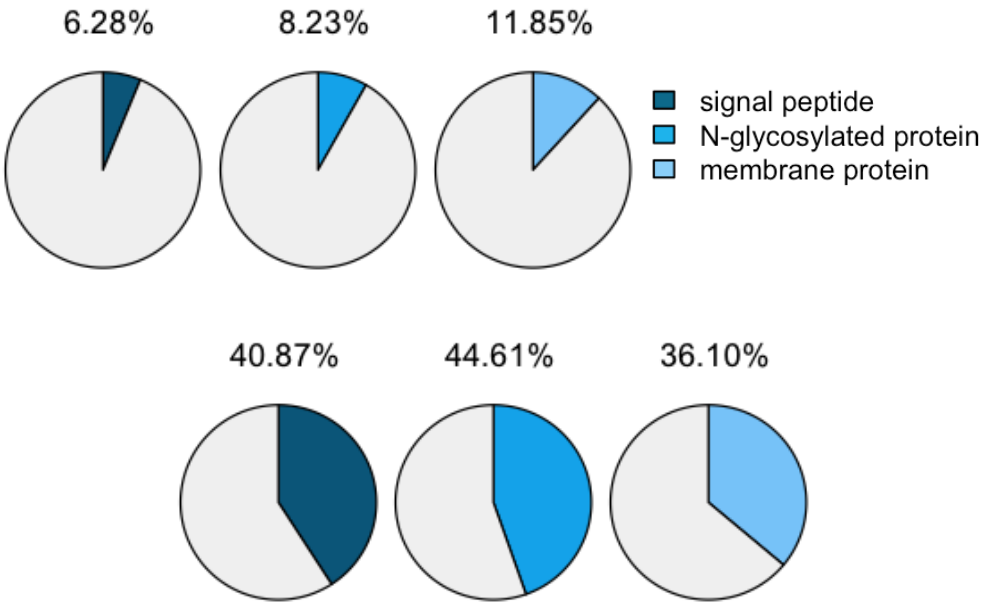


Figure 4.9: Contribution of proteins containing SP, N-glycosylated sites and membrane proteins in (upper row) the whole quantified proteome and (lower row) the negatively affected proteins of Sec61α silencing experiments.

Additionally, we also detected a significant enrichment of proteins containing SP, N-glycosylated sites and membrane proteins among the negatively affected proteins

(Fig. 4.9). Those observations suggest that the down-regulated proteins are Sec61α substrates and the up-regulated proteins could potentially serve as compensatory components due to the silencing effects.

In short, we have successfully identified and analysed the substrate spectrum of the Sec61 complex from MS proteomics data, paving the way for subsequent analysis of specific substrates of other transport components, e.g. the TRAP complex.

4.3.3 TRAP silencing experiments: characterisation of TRAP clients

Differential expression analysis

Similar silencing experiments were conducted for the TRAP complex with two different siRNAs, targeting the TRAPβ (or SSR2) subunit. In the previous study with the same experiment design, the silencing depleted 90% of TRAP complexes, without any significant effects on cell growth, cell viability or cell morphology [112].

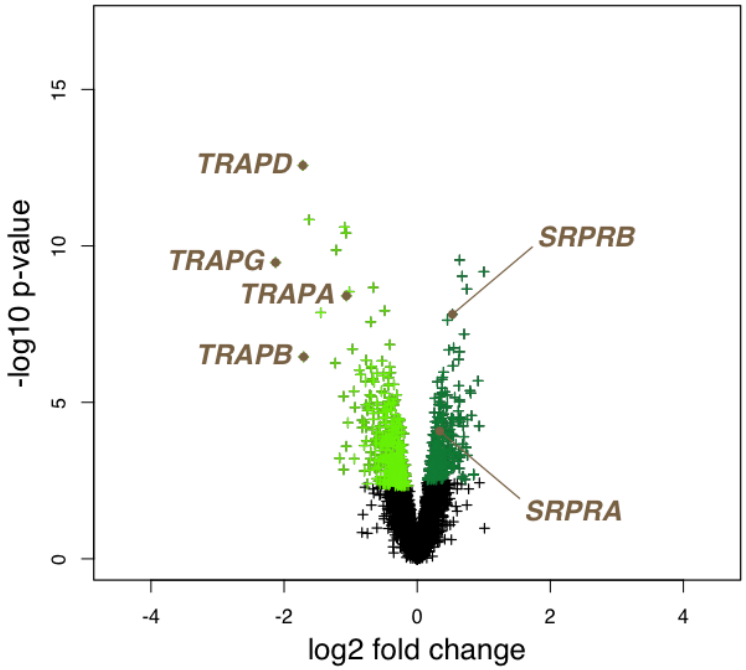
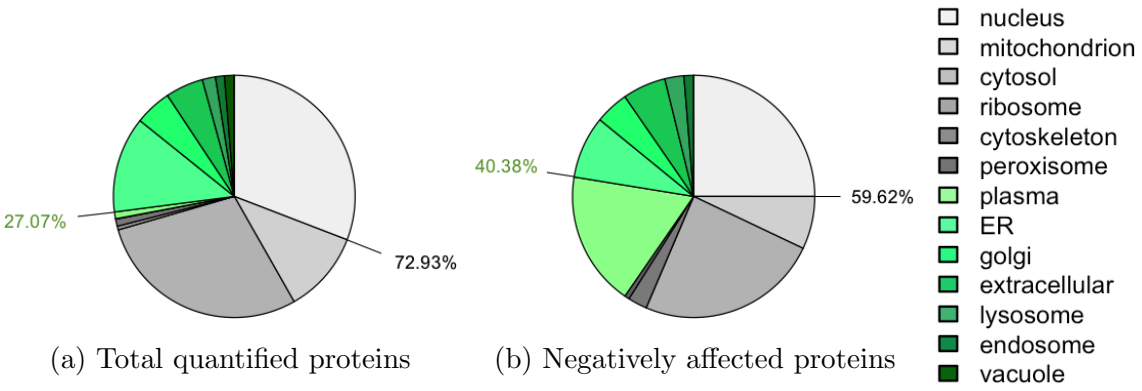


Figure 4.10: Volcano plot of TRAP silencing experiment. The points represent all quantified proteins while the green points indicate the significantly affected proteins: the light green points on the left and the dark green points on the right represent the negatively affected proteins and the positively affected proteins, respectively.

Approximately 8500 different proteins were quantified by MS in TRAP silencing experiments. Among those, 5911 proteins were detected in all experiments for



(a) Total quantified proteins (b) Negatively affected proteins

Figure 4.11: Sub-cellular localisation of (a) the full quantified proteome and (b) the negatively affected proteins of TRAP silencing experiments. The green parts indicate the organelles of endocytic and exocytic pathways.

further analysis. Out of 5911, 257 were identified as significantly affected proteins: 180 negatively affected proteins, including the target protein TRAPβ; and 77 positively affected proteins. Besides, other TRAP subunits such as TRAPα, TRAPγ and TRAPδ were negatively affected by TRAPβ depletion (Fig. 4.10). The positively affected proteins also included SRP receptor subunits. Sub-cellular localisation analysis verified that 40.38% of the negatively affected proteins belong to organelles of endocytic and exocytic pathways (Fig. 4.11). This is an about 1.5 fold enrichment compared the full quantified proteome. The GO annotations of the negatively affected proteins also revealed a significant enrichment of proteins containing sp, N-glycosylated sites and membrane proteins (Fig. 4.12). Compared to Sec61α silencing experiments, these number of substrates are expected since the TRAP complex is a precursor-specific auxiliary transport component to the Sec61 complex.

TRAP clients characterisation

The hydophobicity analysis of Sec61 substrates showed that their SP were less hydrophobic compared to the overall hydrophobicity of all human SP (Fig. 4.13a), indicating that Sec61 has a higher affinity to nascent chains with higher hydrophobic SP. Regarding TRAP substrates, their SP tend to have lower hydrophobicity compared to the average. Interestingly, the SP of TRAP substrates showed a significantly higher content of glycine and proline (GP) (Fig. 4.13b) than all human SP and all human proteins (Fig. 4.15a). Visual inspection of the TMH of proteins that do not have cleavable SP showed a lower hydrophobicity tendency and higher GP content although they are not statistically significant (Fig. 4.14).

To further verify the unusually high GP content in the SP of TRAP substrates, we extracted SP sequences of TRAP substrates' homologs from *S. cerevisiae* and

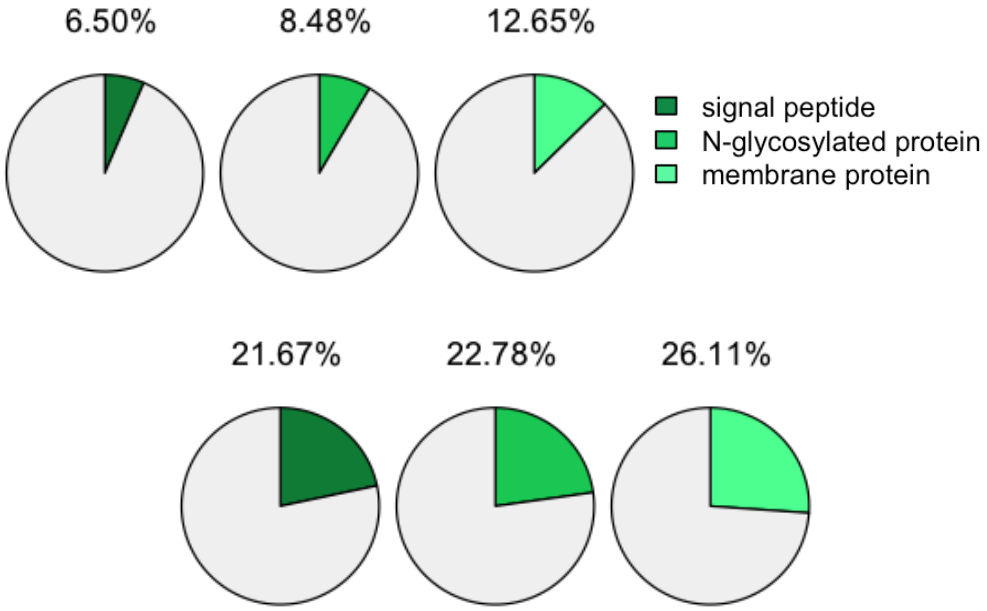
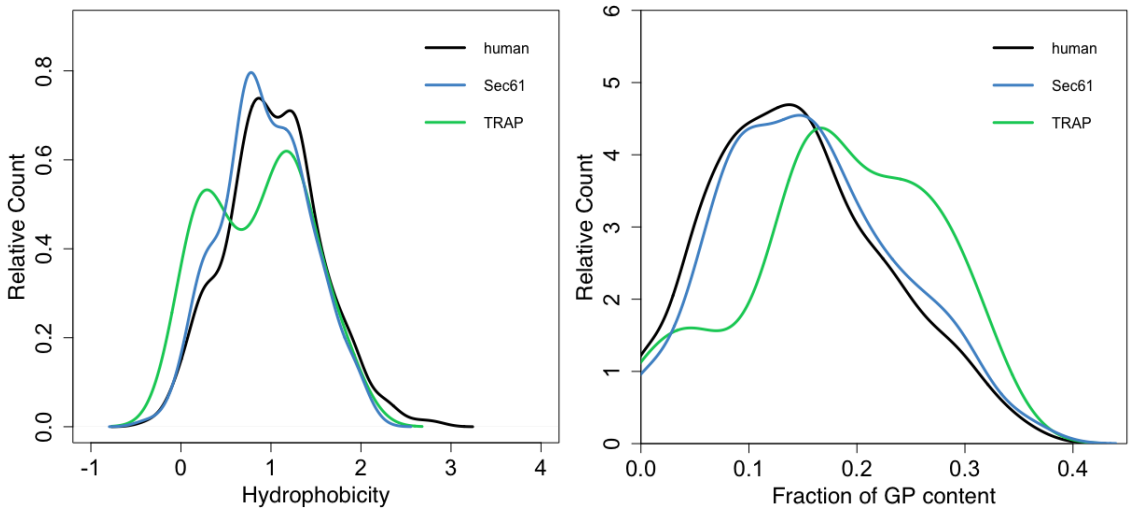
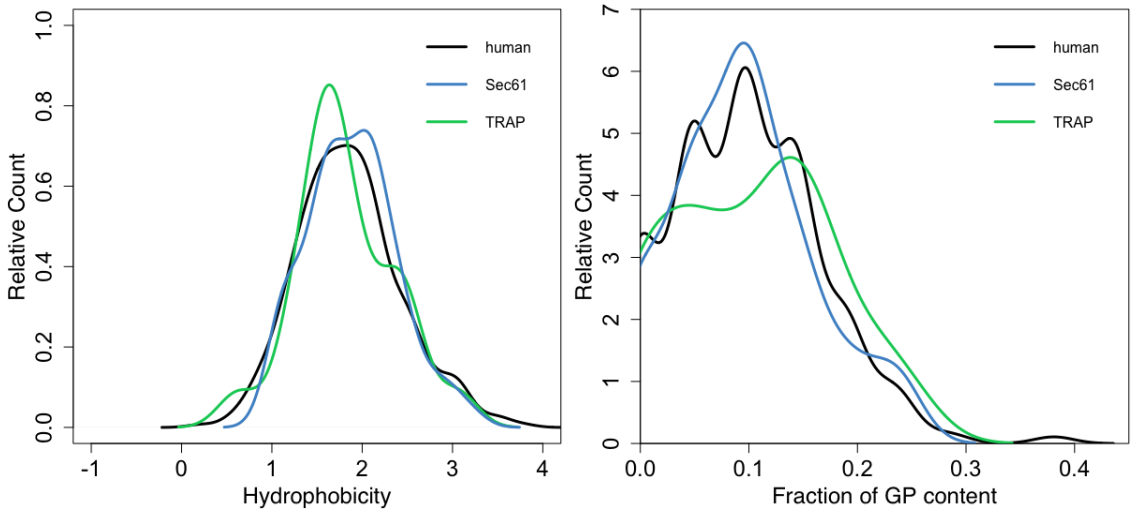


Figure 4.12: Contribution of proteins containing SP, N-glycosylated sites and membrane proteins in (upper row) the full quantified proteome and (lower row) the negatively affected proteins of TRAP silencing experiments, respectively.



(a) (b)

Figure 4.13: Distribution of (a) hydrophobicity score and (b) GP content of SP which belong to the full human proteome (black), Sec61 substrates (blue) and TRAP substrates (green).



(a) (b)

Figure 4.14: Distribution of (a) hydrophobicity score and (b) GP content of TMH (of membrane proteins that do not have cleavable SP) which belong to the full human proteome (black), Sec61 substrates (blue) and TRAP substrates (green).

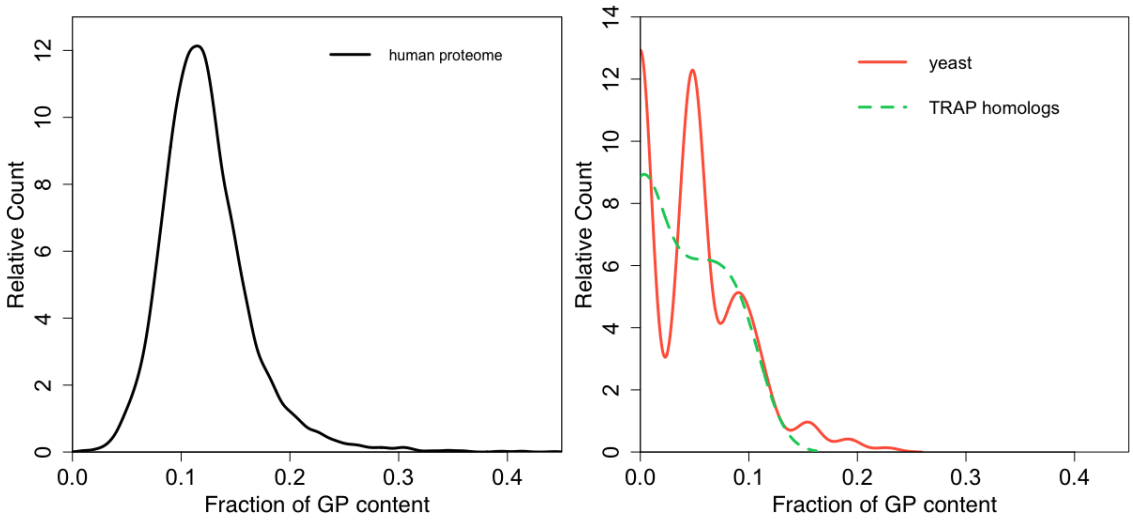
analysed the SP of homologs in a similar fashion. Out of almost 8000 yeast protein sequences from UniProtKB, over 800 contain cleavable SP. However, only 7 homologs contain SP. By comparison, the SP set in yeast homologs showed a lower GP content than human SP (Fig. 4.15b). Since yeast does not have TRAP complex, these findings support the relevance of high GP content of TRAP substrates in human cells.

4.3.4 Discussion

By applying unbiased experiments in living human cells, our investigation on putative TRAP substrates has revealed that they are enriched in the endocytic and exocytic pathways as expected. Additionally, they showed a low hydrophobicity tendency and significantly high GP content in their cleavable SP. In particular, the prion protein (PrP), one of the TRAP dependent protein [104], fits these observations.

Since the high GP content could potentially impede the insertion of SP into the channel, we speculated that TRAP activity may be necessary for those SP with high GP content. The unusual high GP content in SP could come from mutations over the course of evolution and TRAP may play a role to compensate this high accumulation of GP content. On the contrary, due to the absence of the TRAP complex in yeast, SP with high GP content could have been eliminated due to selection pressure.

Due to the high GP content and low hydrophobicity of the SP of TRAP substrates, they may spend a longer period of time on the cytosolic surface of the Sec61 channel. Therefore, we further propose that the TRAP complex may stabilise SP



(a) (b)

Figure 4.15: (a) GP content distribution of the whole human proteome; (b) GP content distributions of SP which belong to the full *S. cerevisiae* proteome (continuous red line) and to homologs of TRAP clients (green dashed line).

on the cytosolic site for easier translocation through the Sec61 channel.

Chapter 5

Enhancement of Sec61-mediated Ca^{2+} leakage from endoplasmic reticulum by eeyarestatin compounds

This work is in preparation for submission as "Gamayun I, Klein MC, Lee PH, Nguyen D, Flitsch SL, Whitehead R, Swanton E, High S, Helms V, Zimmermann R, Cavalié A. **Enhancement of Sec61-mediated Ca^{2+} leakage from endoplasmic reticulum by eeyarestatin compounds**". My contribution in this work is conducting all docking experiments. Only my computational results are presented in the following. They are accompanied by an extensive experimental part that is not shown here.

5.1 Background and Motivation

Calcium ions (Ca^{2+}) play important roles in many physiological and biological processes of the cell such as signal transduction, cofactor of enzymes, bone formation, etc. Therefore, calcium levels are tightly regulated, especially in mammals, by Ca^{2+} channels/transporters which can allow calcium entry or removal from the cell or cellular compartments. Sec61 complexes are not only vital components in protein biogenesis, they also operate as permeable Ca^{2+} ion channels in ER membrane [128]. Under normal conditions, the Ca^{2+} concentration is maintained at a low level (0.05–0.1 μM) by the control of two types of pumps: SERCAs (Sarcoplasmic Endoplasmic Reticulum Calcium ATPases) and PMCAs (Plasma Membrane Calcium ATPases). Meanwhile, the Ca^{2+} concentration inside the ER is high (100–800 μM). However, due to a constant Ca^{2+} leakage from the ER, this distribution is always under pres-

sure. This leakage is supposed to relieve the stress for ER when the Ca^{2+} concentration is too high, hence, inducing calcium signal transduction [129]. In Sec61α, the lateral gate, which is composed of transmembrane helices 2 (position 77 to 96), 3 (118 to 138), 7 (289 to 309) and 8 (355 to 375), is believed to be responsible for the insertion of transmembrane domains into the ER membrane [130]. Therefore, it is possible that the open lateral gate is also responsible for Ca^{2+} leakage.

Eeyarestatin 1 (ES1) has been revealed as a potent inhibitor of ER Associated protein Degradation (ERAD) and of protein translocation through ER by targeting the Sec61 complex [131]. However, the inhibiting mechanism of ES1 is still unclear. Additionally, because of the connection between Sec61 and Ca^{2+} leakage, the effect of ES1 on the ER calcium homeostasis is also in question. Several Ca^{2+} imaging experiments have been conducted by our colleague, Igor Gamayun, to address this question. By observing the changes of Ca^{2+} concentration inside ER and cytosol under the effect of ES1 and its analogues (ES24, ES35 and ES47), the experiments revealed that ES1 and ES24 weaken the ER calcium homeostasis by promoting Ca^{2+} leakage from the ER. On the other hand, ES35 did not show any effect while ES47 slightly inhibited the Ca^{2+} leakage. To be able to understand these effects at the molecular level, we carried out several docking experiments between a human Sec61α structure and eeyarestatin compounds. Additionally, in a study about inhibition effects of eeyarestatin 1 against p97/VCP and ERAD, Wang *et al.* showed that the nitrofuran-containing (NFC) group (Fig. 5.4) is responsible for the inhibitory effects [132]. Therefore, NFC was also included in the docking experiments.

5.2 Materials and Methods

5.2.1 Preparation of 3D structures

Due to the unavailability X-ray structure of human Sec61α in the open state, homology modelling was used to construct a comparative human Sec61α structure. The materials for homology modelling consisted of:

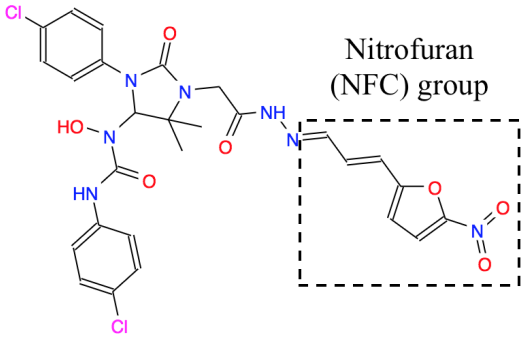
- the human Sec61α protein sequence
- the template 3D structure of a homologous protein

The sequence of the human Sec61α protein, which contains 476 amino acids, was retrieved from the UniProtKB database (UniProt ID: P61619). As template structure, the crystal structure 3JC2 [99] of canine Sec61α in an open conformation was selected since both human and canine sequences share 99.8% sequence identity.

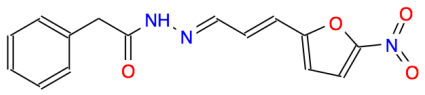
The 3JC2 structure shows Sec61α in an open conformation but is lacking structural information for the helical plug region (residue 63 to 69). Since in the open conformation, the plug is believed to be out of the way of the translocation pathway, we assumed that it does not form relevant interactions with the substrates.

Homology modelling was carried out using the MODELLER 9.17 package [133]. The plug region was modelled and optimised as a loop structure. Subsequently, the homology model was subjected to energy minimisation, using the NAMD package, to relax side chain atoms. Afterwards, the resulting model was validated by the ProSa-web server [134], along with other reference structures.

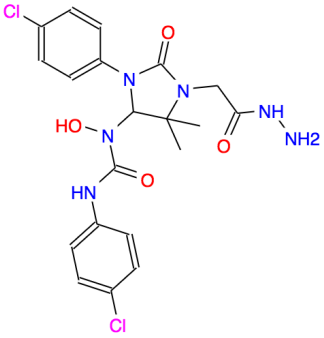
The 3D structures of eeyarestatin compounds (Fig. 5.1) were generated by the Open Babel package [135].



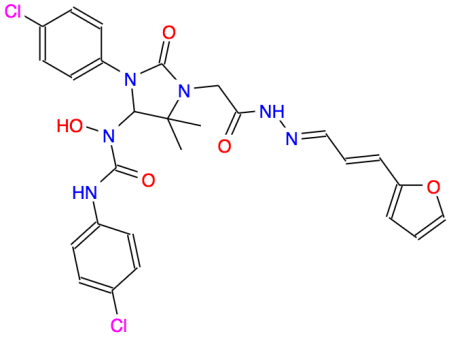
(a) Eeyarestatin 1



(b) Eeyarestatin 24



(c) Eeyarestatin 35



(d) Eeyarestatin 47

Figure 5.1: Chemical structure of eeyarestatin compounds.

5.2.2 Docking protocols

The protonation states of protein residues and partial charges of ligands were assigned by the `prepare_receptor` and `prepare_ligand` modules of the AutoDock4 package [67]. Because the binding positions of eeyarestatin inside Sec61α are still unknown, for each compound, the docking calculations were performed in two consecutive steps:

1. In the first docking step, we adopted a relatively large grid box (Fig. 5.2, black box), covering the entire cavity of Sec61, to scan for energetically favourable

conformations of the ligand inside the Sec61 pocket. The Lamarckian genetic algorithm was used for the optimisation of the ligand conformations and orientations (2.5×10^6 energy evaluations and 27×10^3 generations).

2. In the second docking step, the size of the grid box (Fig. 5.2, red box) was scaled down based on the population of the most stable binding positions of the ligand in the first run. In the second, finer run, more stringent parameters (100×10^6 energy evaluations and 0.5×10^6 generations) were used.

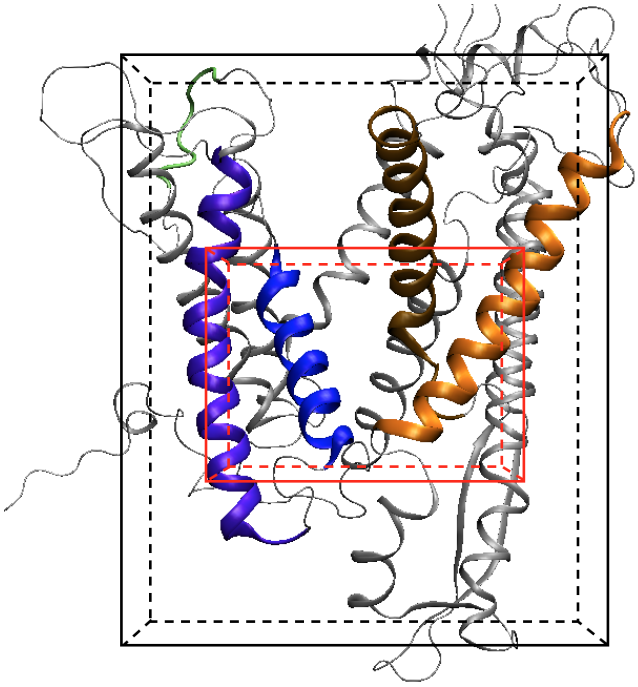


Figure 5.2: Docking grid boxes, for coarse docking (black) and fine docking (red)

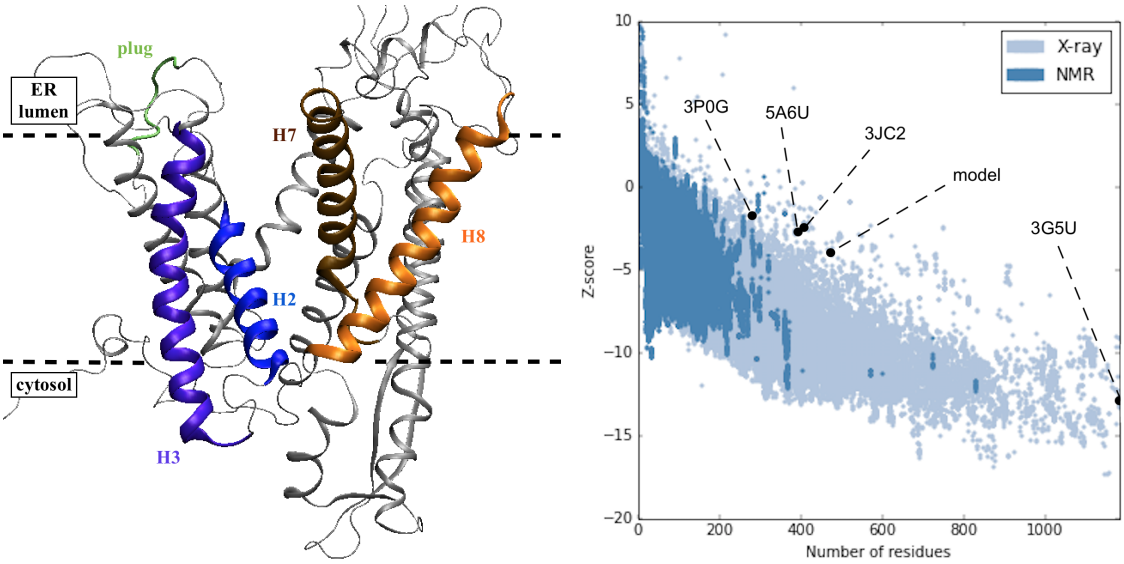
When the docking calculations finished, the results were visually inspected and Sec61α-ligand interactions were visualised by LigPlot$^+$ program [91].

5.3 Results and Discussion

5.3.1 Homology model

The resulting homology model of human Sec61α is depicted in Fig. 5.3a. At first, we wondered how well the MODELLER protocol works for a transmembrane protein. The analysis of the homology model and other reference structures (two canine Sec61α structures, a GPCR and an ABC transporter) from ProSa-web showed that the Z-score of the homology model is within the range of scores typically found in experimentally determined structures of protein chains (Fig. 5.3b). In other words,

the structural integrity of the homology model is comparable to the experimental 3D structures, thus, the model appears of suitable quality for further docking experiments.



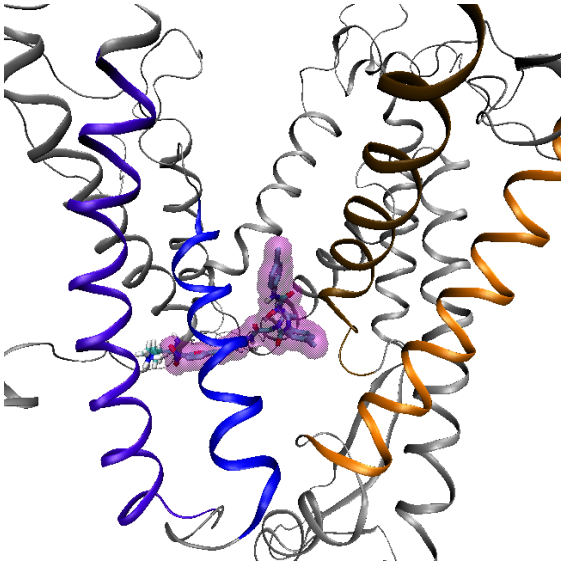
(a) Human Sec61α homology model (b) Model validation

Figure 5.3: (a) Side view of the hSec61α model seen from the lateral gate that is formed by the coloured transmembrane helices 2, 3, 7 and 8. The helical plug of Sec61 is coloured green. The dashed line depicts the putative position of the ER membrane. (b) Model validation using the ProSa-web server. The blue and light-blue points represent the Z-scores of experimental protein structures (mostly soluble proteins) while the black dots depict the Z-scores of 5A6U, 3JC2, 3P0G (GPCR), 3G5U (ABC transporter) and the human Sec61α model.

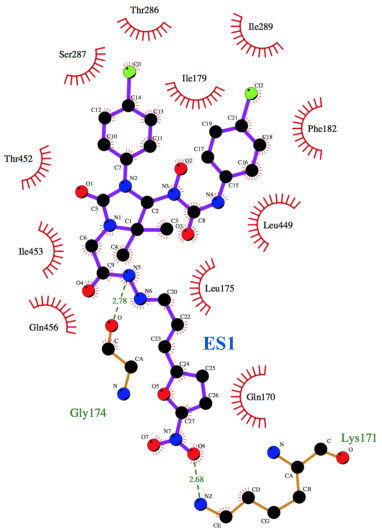
5.3.2 Docking results

Eeyarestatin 1 (ES1)

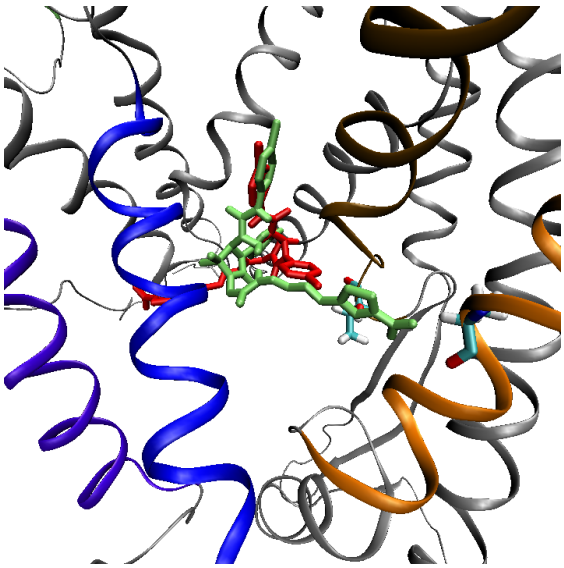
The binding pose with highest predicted affinity (estimated $\Delta G = -9.7\ kcal/mol$) of ES1 inside the Sec61α cavity is shown in Fig. 5.4. In this conformation, ES1 forms 2 hydrogen bonds with K171 and G174 (the region at the end of H4 and loop connecting H4–H5) and is located inside the pocket facing towards the cytosol. In this conformation, ES1 does not seem to have any interaction/hindrance to the gate helices. However, in a slightly less favorable docking position ($\Delta G = -9.62\ kcal/mol$), the nitrofuran group of ES1 does interact with H7 and H8 at T286 and S376, respectively, and is slightly shifted towards the area between H2 and H7 (Fig. 5.4c), which could hamper the gate closure, hence, promoting the Ca^{2+} leakage.



(a) ES1 binding position (side view)



(b) ES1–Sec61α interactions



(c) Different poses of ES1

Figure 5.4: The predicted binding position of ES1 with the best score inside the human Sec61α homology model. The lateral gate helices and the plug are coloured as in Fig. 5.3a. The surface of ES1 is marked in magenta. All residues (K171, G174) which form hydrogen bonds with ES1 and ES1 itself are illustrated in Licorice model. (a) The binding pose in side view perspective towards the lateral gate. (b) ES1–Sec61α interactions illustrated by the LigPlot$^+$ program. (c) Different docking positions of ES1: the red pose is equivalent to that shown in (a). It has $\Delta G = -9.7\ kcal/mol$. The green pose has $\Delta G = -9.62\ kcal/mol$.

Eeyarestatin 24 (ES24)

The most favourable docking position of ES24 has an estimated binding energy of $-9.04\ kcal/mol$, and forms 3 hydrogen bonds with K282, T286 and N288 at the N-terminal end of H7 (Fig. 5.5b). Compared to ES1, the aromatic group of ES24 is placed in the same area as the nitrofuran group of ES1, and is surrounded by H2, H7 and H8, which could potentially hamper the closure of the lateral gate. The aromatic group stays in similar positions in two other less favourable docking poses of ES24 (Fig. 5.5c).

Eeyarestatin 35 (ES35)

The best pose of ES35 (with $\Delta G = -7.91\ kcal/mol$) forms 4 hydrogen bonds with Y279, K282, Y285 and T286, in the loop region upstream of H7 (Fig. 5.6a). This binding position is further away from the gate compared to ES1 and ES24. However, alternative docking positions of ES35 (Fig. 5.6c) occupy the area between H2 and H7, which could hinder the function of the gate. Especially, the binding position with $\Delta G = -7.78\ kcal/mol$, forms 2 hydrogen bonds with V85 and I289, and may have the potential to keep the gate open.

Eeyarestatin 47 (ES47)

The best pose of ES47 ($\Delta G = -9.67\ kcal/mol$) forms 2 hydrogen bonds with K282, Q456 and does not seem to interfere with the gate activity (Fig. 5.7a). Another less favourable docking pose ($\Delta G = -8.12\ kcal/mol$) is slightly shifted towards the area between H2 and H7 but does not form any hydrogen bonds with Sec61α.

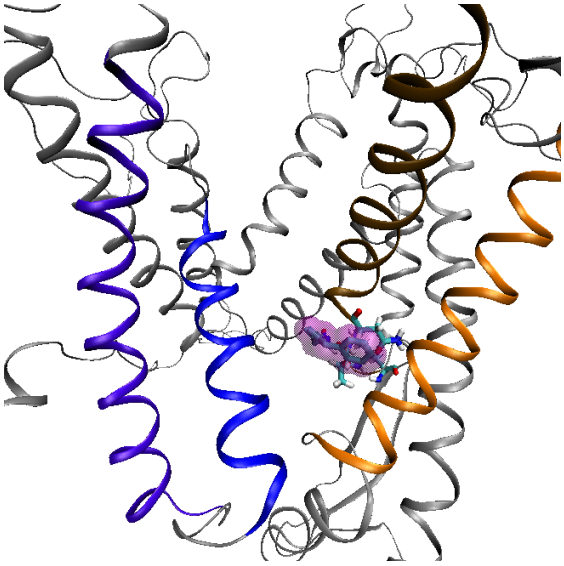
Nitrofuran–containing group (NFC)

Of all the docking poses of NFC, only one pose is located in the area between the lateral gates helices. This is in fact the most favourable binding position with $\Delta G = -5.69\ kcal/mol$ (Fig. 5.8c). In this case, NFC occupies a very similar area as E24 (Fig. 5.8d).

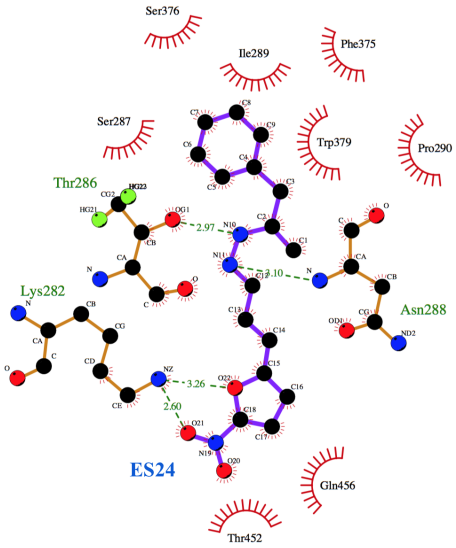
5.3.3 Interpretation

Overall, the molecular docking can be reasonably well correlated with the experimental findings on calcium leakage via Sec61α protein (ES1 and ES24 promote the leakage whereas E35 and E47 do not):

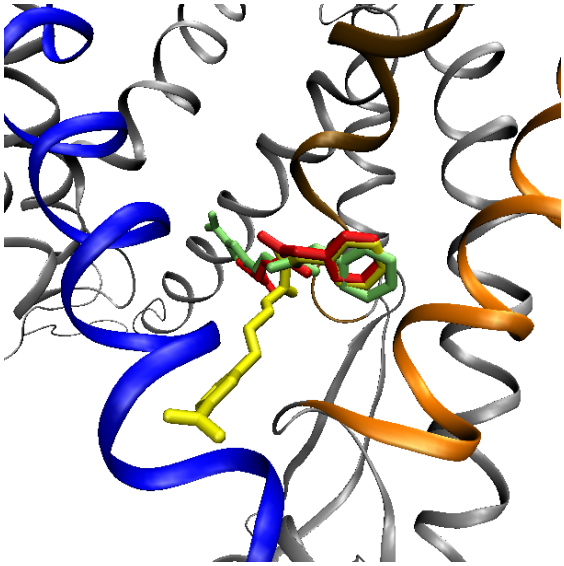
- ES1: although the most favourable binding pose does not quite hamper the gate mechanism, an alternative, slightly less favourable pose (only $0.08\ kcal/mol$ difference) sits between the gate helices, hence, restricts the gate movement. Surprisingly, the nitrofuran group in this conformation occupies a similar area as ES24 and NFC (Fig. 5.9a).



(a) ES24 binding position (side view)

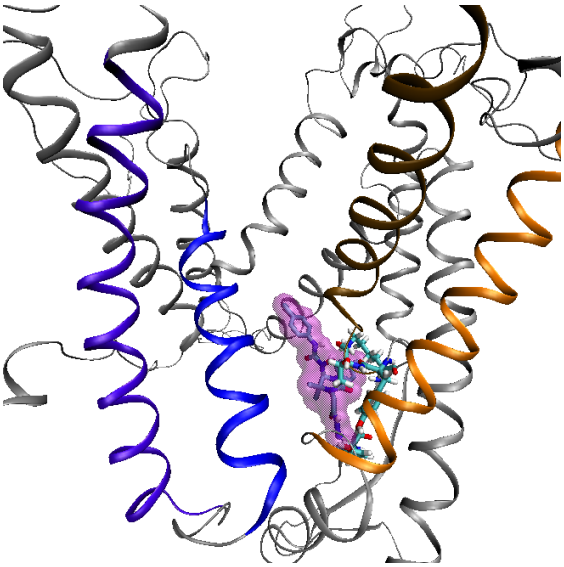


(b) ES24–Sec61α interactions

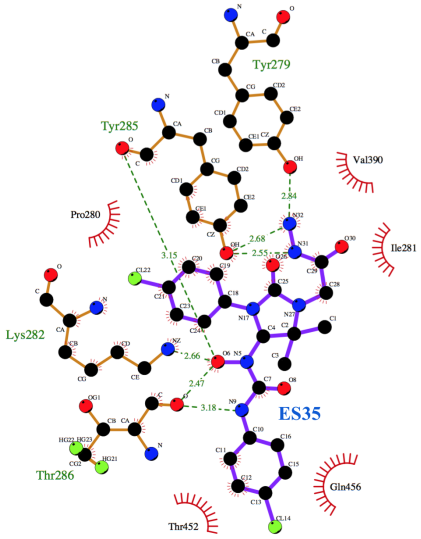


(c) Different poses of ES24

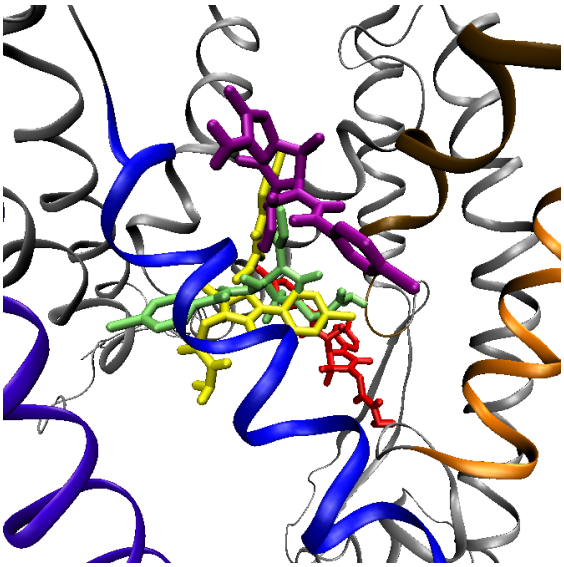
Figure 5.5: The predicted binding position of ES24 with the best score inside the human Sec61α homology model. The lateral gate helices and the plug are coloured as in Fig. 5.3a. The surface of ES24 is marked in magenta. All residues (K282, T286, N288) which form hydrogen bonds with ES24 and ES24 itself are illustrated in Licorice model. (a) The binding pose in side view perspective towards the lateral gate. (b) ES24–Sec61α interactions illustrated by the LigPlot$^+$ program. (c) Different docking positions of ES24: red - $\Delta G = -9.04\ kcal/mol$, equivalent to the pose shown in (a); green - $\Delta G = -8.29\ kcal/mol$; yellow - $\Delta G = -7.55\ kcal/mol$.



(a) ES35 binding position (side view)

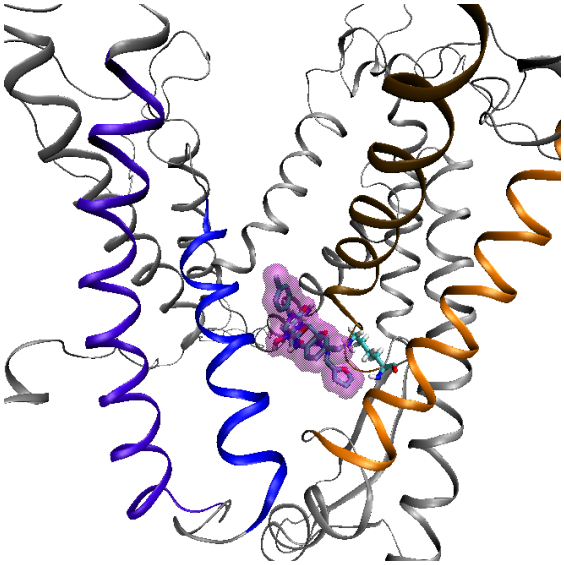


(b) ES35–Sec61α interactions

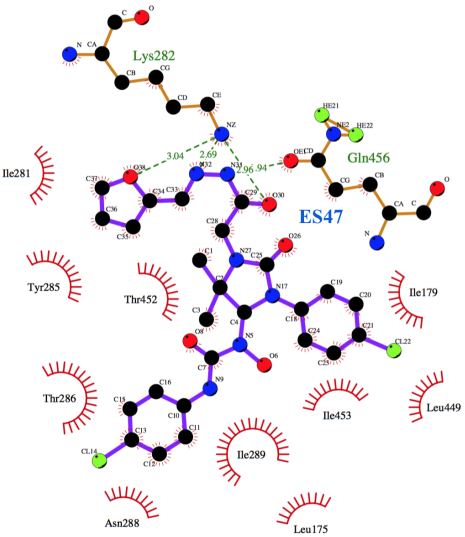


(c) Different poses of ES35

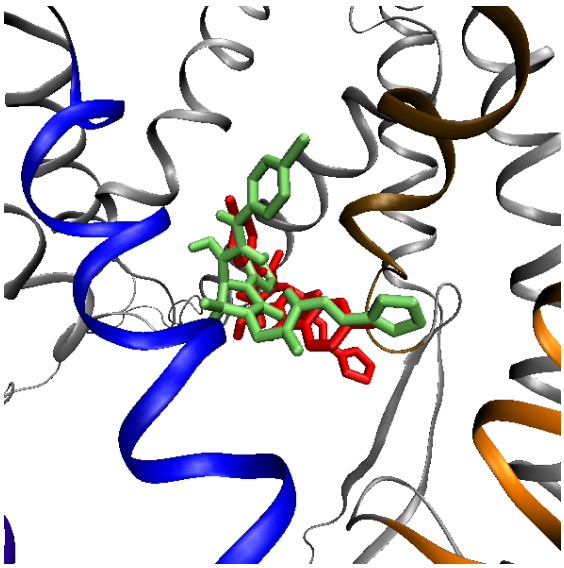
Figure 5.6: The predicted binding position of ES35 with the best score inside the human Sec61α homology model. The lateral gate helices and the plug are coloured as in Fig. 5.3a. The surface of ES35 is marked in magenta. All residues (Y279, K282, Y285, T286) which form hydrogen bonds with ES35 and ES35 itself are illustrated in Licorice model. (a) The binding pose in side view perspective towards the lateral gate. (b) ES35–Sec61α interactions illustrated by the LigPlot$^+$ program. (c) Different docking positions of ES35: red - $\Delta G = -7.91\ kcal/mol$, equivalent to the pose shown in (a); green - $\Delta G = -7.78\ kcal/mol$; yellow - $\Delta G = -7.24\ kcal/mol$; purple - $\Delta G = -6.78\ kcal/mol$.



(a) ES47 binding position (side view)

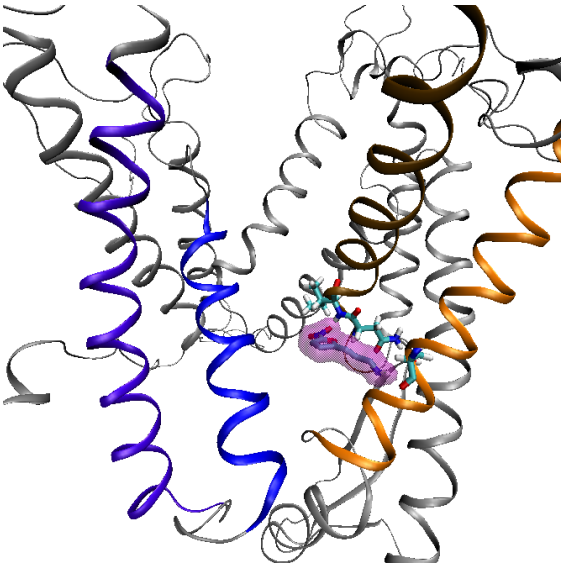


(b) ES47–Sec61α interactions

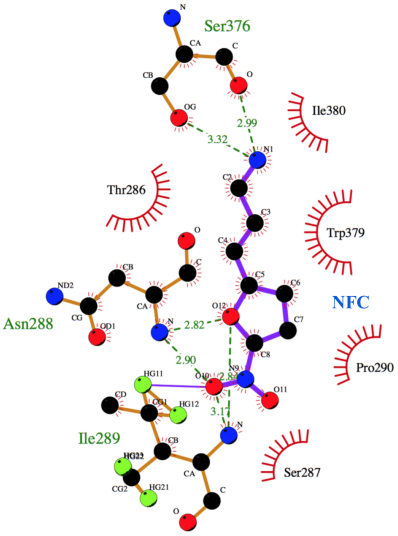


(c) Different poses of ES47

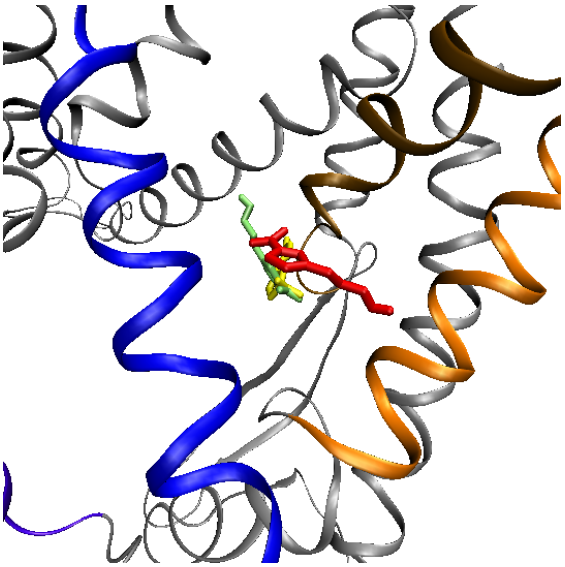
Figure 5.7: The predicted binding position of ES47 with the best score inside the human Sec61α homology model. The lateral gate helices and the plug are coloured as in Fig. 5.3a. The surface of ES47 is marked in magenta. All residues (K282, Q456) which form hydrogen bonds with ES47 and ES47 itself are illustrated in Licorice model. (a) The binding pose in side view perspective towards the lateral gate. (b) ES47–Sec61α interactions illustrated by the LigPlot$^+$ program. (c) Different docking positions of ES47: red - $\Delta G = -9.67\ kcal/mol$, equivalent to the pose shown in (a); green - $\Delta G = -8.12\ kcal/mol$.



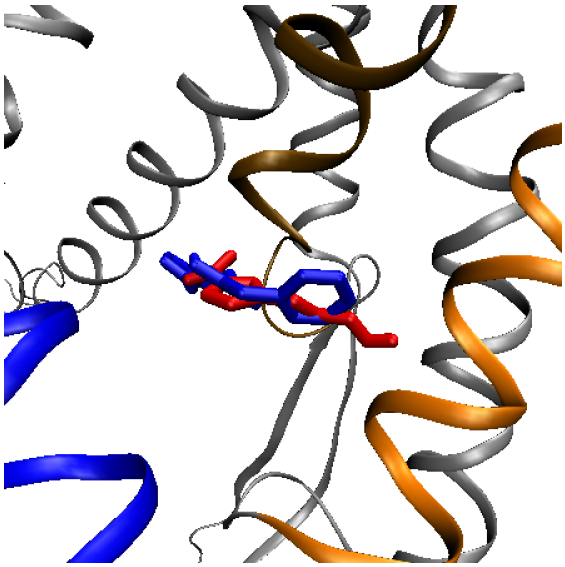
(a) NFC binding position (side view)



(b) NFC–Sec61α interactions



(c) Different poses of NFC



(d) NFC and E24

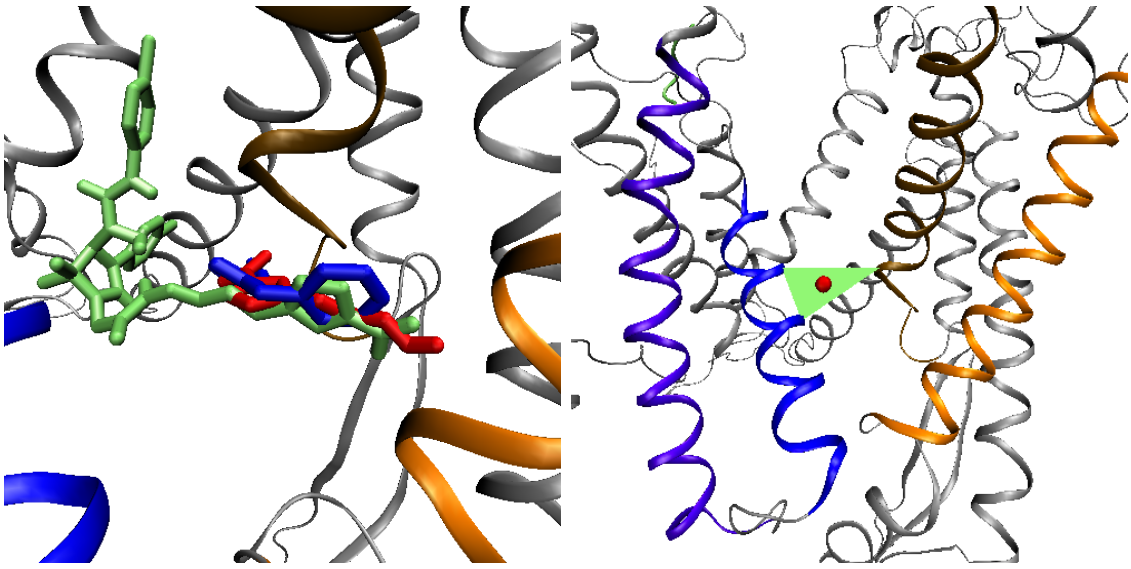
Figure 5.8: The predicted binding position of NFC with the best score inside the human Sec61α homology model. The lateral gate helices and the plug are coloured as in Fig. 5.3a. The surface of NFC is marked in magenta. All residues (N288, I289, S376) which form hydrogen bonds with NFC and NFC itself are illustrated in Licorice model. (a) The binding pose in side view perspective towards the lateral gate. (b) NFC–Sec61α interactions illustrated by the LigPlot$^+$ program. (c) Different docking positions of NFC: red - $\Delta G = -5.69\ kcal/mol$, equivalent to the pose shown in (a); green - $\Delta G = -5.45\ kcal/mol$; yellow - $\Delta G = -5.44\ kcal/mol$. (d) Docking poses of NFC (red) and ES24 (blue)

- ES24: most of the top docking poses of ES24 stay in the area between the lateral gate helices.

- ES35: the best docking pose is consistent with leakage data. The other, less favourable conformations seem to suggest that ES35 supports the calcium leakage by hindering the gate movement. But one has to note that the estimated binding energy of ES35 is quite low compared to the other compounds.

- ES47: the docking result suggests that ES47 should not hamper lateral gate movement: the best docking pose does not seem to affect the gate. The alternative docking pose is located between the gate helices but this conformation is not stabilised by hydrogen bonds between ES47 and protein.

Additionally, we carried out a distance analysis of the best docking position of the compounds to the shortest distance between H2 and H7 (Fig. 5.9b and Tab. 5.1). This analysis showed that E35 and E47 are located further away from the gate compared to the rest, hence, the inhibitory effect is possibly decreased, which is consistent with the result of Ca^{2+} imaging experiments.



(a) (b)

Figure 5.9: (a) Best docking poses of ES1 (green), ES24 (blue) and NFC (red). (b) The shortest distance between H2 and H7 is illustrated by the green triangle (composed of the Cα of T86, L89 and P290), the red sphere illustrates the center of mass of the triangle.

Table 5.1: Distance of the compounds to the red sphere marking the shortest distance between H2 and H7 (Fig. 5.9b). The distances from the red sphere to the center of mass and to the closest atom of the compounds are evaluated.

Compound	**Center of mass (Å)**	**Closest atom (Å)**
ES1	9.98	6.23
ES24	9.87	6.08
ES35	15.06	11.74
ES47	10.92	7.41
NFC	8.65	6.80

5.4 Summary

In this project, we applied molecular docking using AutoDock to investigate the binding modes of several eeyarestatin compounds inside the homology model of human Sec61α. From the docking results, we postulate that ES1 and ES24 can potentially block the lateral gate function since docking predicted that they bind between the H2 and H7 helices of the gate. As a consequence, they keep the gate open as long as they bind in that position, hence, promoting Ca^{2+} leakage via Sec61α. On the other hand, ES35 and ES47 were located further away (compared to ES1 and ES24) from the gate. Therefore, they likely do not block the gate and they do not stimulate Ca^{2+} leakage. These findings from the docking results are consistent with the results from the calcium imaging experiments which were conducted by our colleagues. In short, the results from the docking experiments provided new mechanistic insight how the eeyarestatin compounds may bind to the human Sec61α protein.

Chapter 6

Conclusions

This thesis presents results from three different projects that aim at characterising functional properties of membrane transport systems and of their interaction with substrates and inhibitors. Chapter 3 on the MdfA transporter addresses an experimentally well characterised transporter. Chapter 4 and 5 address the human Sec61 translocation and accessory proteins (chapter 4). In all cases, the computational techniques provided new mechanistic insight (chapter 3 and 5) or were of integral importance for analysis of the primary data (chapter 4).

In chapter 3, we presented a novel method for MdfA substrate classification. The method incorporates protein-ligand interactions as well as various potential energy terms from multiple MD simulations. Overall, the method achieved a decent performance with 73.12% in accuracy although the approach still has unsolved issues with problematic cases, which is due to the similarity in chemical structure among substrates and non-substrates. However, this is a challenging problem since structurally similar compounds behave indistinguishably inside the Sec61α pocket, at least on the timescale of 100 ns. The presented approach is the first method in substrate classification that integrates the structural interactions and potential energies of protein-ligand complexes by running multiple MD simulations. Therefore, this method could potentially promote further developments in drug discovery and other applications.

In chapter 4, we successfully identified Sec61α and TRAP dependent proteins from MS proteomics data. Furthermore, we also characterised TRAP clients to help unravel TRAP function. Through our studies, we discovered that TRAP clients SP have a tendency of low hydrophobicity and higher-than-average GP content. Based on those observations, we have proposed that TRAP may be responsible for helping the proteins that have SP with high GP content and less hydrophobicity to migrate easily through the Sec61α channel.

In chapter 5, we suggested several binding modes of various eeyarestatin compounds into a homology model of human Sec61α. By inspecting, analysing and comparing binding positions of eeyarestatin compounds, we proposed that ES1 and ES24 are likely to hamper the function of the lateral gate by sitting in between H2

and H7, which are the "doors" of the lateral gate, hence, promoting Ca^{2+} leakage by Sec61α. This observation is consistent with the findings from the calcium imaging experiments which were conducted by our colleagues.

List of Figures

List of Tables

Bibliography

[1] Cell Membranes. Online. http://www.nature.com/scitable/topicpage/cell-membranes-14052567.

[2] Membrane Structure and Function. Online. http://www.cytochemistry.net/cell-biology/membrane_intro.htm.

[3] BS Brown. *Biochemistry and Molecular Biology Education*, chapter Biological membranes. Wiley, 1997.

[4] G. von Heijne. The membrane protein universe: what's out there and why bother? *J Intern Med*, 261:543–557, 2007.

[5] Harvey Lodish, Arnold Berk, S Lawrence Zipursky, Paul Matsudaira, David Baltimore, and James Darnell. *Molecular Cell Biology*, chapter 15. W. H. Freeman, fourth edition, 2000.

[6] Robert S. Kass. The channelopathies: novel insights into molecular and genetic mechanisms of human disease. *J Clin Invest.*, 115:1986–1989, 2005.

[7] Dimitri M. Kullmann and Stephen G. Waxman. Neurological channelopathies: new insights into disease mechanisms and ion channel function. *J Physiol*, 588:1823–1827, 2010.

[8] Michael R Rose. Neurological channelopathies: Dysfunctional ion channels may cause many neurological diseases. *BMJ*, 316:1104, 1998.

[9] Jill V. Hunter and Arthur J. Moss. Seizures and arrhythmias: Differing phenotypes of a common channelopathy? *Neurology*, 72:208–209, 2009.

[10] Piddock LJ. Multidrug-resistance efflux pumps - not just for resistance. *Nat Rev Microbiol.*, 4:629–36, 2006.

[11] Drug Transporter Interactions. Online. http://landing.quotientbioresearch.com/blog/bid/80654/Drug-Transporter-Interactions-Efflux-Testing-Strategy.

[12] Park Y, Hayat S, and Helms V. Prediction of the burial status of transmembrane residues of helical membrane proteins. *BMC Bioinformatics*, 8:302, 2007.

[13] Hayat S, Walter P, Park Y, and Helms V. Prediction of the exposure of transmembrane beta barrel residues from protein sequence. *Journal of Bioinformatics and Computational Biology*, 1:43–65, 2011.

[14] Nugent T and Jones DT. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, 10(159), 2009. doi:10.1186/1471-2105-10-159.

[15] Tsirigos KD, Peters C, Shu N, Käll L, and Elofsson A. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res*, 43(W1):W401–7, 2015.

[16] Fuchs A, Kirschner A, and Frishman D. Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. *Proteins*, 74:857–871, 2009.

[17] Timothy Nugent and David T. Jones. Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. *PLOS Computational Biology*, 6(3), 2010. doi:10.1371/journal.pcbi.1000714.

[18] Nguyen D, Helms V, and Lee PH. PRIMSIPLR: prediction of inner-membrane situated pore-lining residues for alpha-helical transmembrane proteins. *Proteins*, 82(7):1503–11, 2014.

[19] Tusnády GE, Dosztányi Z, and Simon I. TMDET: web server for detecting transmembrane regions of proteins by using their 3d coordinates. *Bioinformatics*, 21(7):1276–7, 2005.

[20] Lomize MA, Pogozheva ID, Joo H, Mosberg HI, and Lomize AL. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res*, 40(Database issue):D370–6, 2012. doi:10.1093/nar/gkr703.

[21] Levitt DG and Banaszak LJ. POCKET: a computer-graphics method for identifying and displaying protein cavities and their surrounding amino-acids. *J Mol Graphics*, 10:229–234, 1992.

[22] Hendlich M, Rippmann F, and Barnickel G. LIGSITE: automatic and efficient detection of potential small molecule binding sites in proteins. *J Mol Graph Model*, 15:359–363, 1997.

[23] Huang BD and Schroeder M. LIGSITE^{csc}: predicting ligand binding sites using the connolly surface and degree of conservation. *J Mol Graph Model*, 15:359–363, 1997.

[24] Ho BK and Gruswitz F. HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct Biol*, 8(49), 2008. doi:10.1186/1472-6807-8-49.

[25] Petrek M, Otyepka M, Banas P, Kosinova P, and Koca J. CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, 7:316, 2006.

[26] Po-Hsien Lee and Volkhard Helms. Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. *Proteins: Structure, Function, and Bioinformatics*, 80:421–432, 2012.

[27] Marialuisa Pellegrini-Calace, Tim Maiwald, and Janet M. Thornton. PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput Biol.*, 5:e1000440, 2009.

[28] Simon C. Bull and Andrew J. Doig. Properties of protein drug target classes. *PLoS One*, 10(3):e0117955, 2015.

[29] Lill MA. Multi-dimensional QSAR in drug discovery. *Drug Discov Today*, 12(23–24):1013–7, 2007.

[30] Nitish K. Mishra, Junil Chang, and Patrick X. Zhao. Prediction of membrane transport proteins and their substrate specificities using primary sequence information. *PLoS One*, 9(6), 2014.

[31] Levatić J, Curak J, Kralj M, Smuc T, Osmak M, and Supek F. Accurate models for P-gp drug recognition induced from a cancer cell line cytotoxicity screen. *J Med Chem*, 56(14):5691–708, 2013.

[32] Wang YH, Li Y, Yang SL, and Yang L. Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach. *J Chem Inf Model*, 45(3):750–7, 2005.

[33] Terry R. Stouch and Olafur Gudmundsson. Progress in understanding the structure-activity relationships of p-glycoprotein. *Advanced Drug Delivery Reviews*, 54(3):315–328, 4 2002.

[34] J.A. Mc Cammon, B.R. Gelin, and M. Karplus. Dynamics of folded proteins. *Nature*, 267(5612):585–90, 1977.

[35] A.P. Allen. Introduction to molecular dynamics simulation. *Computational Soft Matter: From Synthetic Polymers to Proteins*, 23:1–28, 2004.

[36] B.R. Brooks, R.E. Bruccoleri, B.D. Olafson, S. States, D.J. Swaminathan, and M. Karplus. CHARMM – a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4:187–217, 1983.

[37] S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta, and P. Weiner. A new force-field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

[38] C. Oostenbrink, Alessandra Villa, A.E. Mark, and W.F. Van Gunsteren. A new force-field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.*, 106:765–784, 1984.

[39] K. Vanommeslaeghe, E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A.D. MacKerell. CHARMM General Force Field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comp. Chem.*, 31(4):671–690, 2010.

[40] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, and D.A. Case. Development and testing of a general amber force field. *J. Comp. Chem.*, 25:1157–1174, 2004.

[41] `https://en.wikibooks.org/wiki/Molecular_Simulation`, 2017. Accessed: 13 June 2017.

[42] P.G. Thomas, A.J. Russel, and A.R. Fersht. Tailoring the *p*H dependence of enzyme catalysis using protein engineering. *Nature*, 318:375–376, 1985.

[43] A.J. Russell and A.R. Fersht. Rational modification of enzyme catalysis by engineering surface charge. *Nature*, 328:496–500, 1987.

[44] M.W. Pantoliano, M. Whitlow, J.F. Wood, M.L. Rollence, B.C. Finzel, G.L. Gilliland, T.L. Poulos, and P.N. Bryan. The engineering of binding affinity at metal ion binding sites for the stabilization of proteins: subtilisin as a test case. *Biochemistry*, 27(22):8311–8317, 1988.

[45] H.E. Alper and R.M. Levy. Computer simulations of the dielectric properties of water: Studies of the simple point charge and transferrable intermolecular potential models. *J. Chem. Phys.*, 91(2):1242–1251, 1989.

[46] R.H. Stote and M. Karplus. Zinc binding in proteins and solution: A simple but accurate nonbonded representation. *Proteins: Struct. Func. Gen.*, 23:12–31, 1995.

[47] M.P. Allen and D.J. Tildesley. Computer simulation of liquids. *Oxford University Press*, 1989.

[48] D.M. York, T.A. Darden, and L.G. Pedersen. The effect of long-range electrostatic interactions in simulations of macromolecular crystals: a comparison of the Ewald and truncated list methods. *Journal of Chemical Physics*, 99:8345–8348, 1993.

[49] L. Verlet. Computer "experiments" on classical fluids. I. thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.*, 159:98–103, 1967.

[50] W.C. Swope, H.C. Andersen, P.H. Berens, and K.R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: application to small water clusters. *J. Chem. Phys.*, 76:637–649, 1982.

[51] R.W. Hockney and J.W. Eastwood. *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules.* Adam Hilger, Bristol, 1988.

[52] P. Schofield. Computer simulation studies of the liquid state. *Computer Physics Communications*, 5(1):17–23, 1973.

[53] D. Beeman. Some multistep methods for use in molecular dynamics calculations. *Journal of Computational Physics*, 20(2):130–139, 1976.

[54] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[55] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: data mining, inference, and prediction*, chapter 15. Springer, second edition, 2009.

[56] D.J. Dittman, T.M. Khoshgoftaar, and A. Napolitano. The effect of data sampling when using Random Forest on imbalanced bioinformatics data. IEEE International Conference on Information Reuse and Integration, 2015.

[57] D.J. Dittman, T.M. Khoshgoftaar, and A. Napolitano. Is data sampling required when using random forest for classification on imbalanced bioinformatics data? In T. Bouabana-Tebibel and S. H. Rubin, editors, *Theoretical Information Reuse and Integration*, volume 446 of *Advances in Intelligent Systems and Computing*, pages 157–171. Springer, 2016.

[58] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, 2003.

[59] S. Dudoit, Y.H. Yang, M.J. Callow, and T.P. Speed. Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica*, 12:111–139, 2002.

[60] W.S. Cleveland and S.J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1998.

[61] S. Tyanova, T. Temu, P. Sinitcyn, A. Carlson, M.Y. Hein, T. Geiger, M. Mann, and J. Cox. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, 13:731–740, 2016.

[62] H. Kim, G.H. Golub, and H. Park. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187–198, 2005.

[63] T. Aittokallio. Dealing with missing values in large-scale studies: microarray data imputation and beyond. *Briefings in Bioinformatics*, 2(2):253–264, 2009.

[64] Virginia Goss Tusher, Robert Tibshirani, , and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(18):5116–5121, 2001.

[65] R. Huey, G.M. Morris, A.J. Olson, and D.S. Goodsell. A semiempirical free energy force field with charge-based desolvation. *J Comput Chem*, 28(6):1145–1152, 2007.

[66] G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem*, 19(14):1639–1662, 1998.

[67] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell, and A.J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–2791, 2009.

[68] S.Y. Huang, S.Z. Grinter, and X.Q. Zou. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. *Phys Chem Chem Phys*, 12(40):12899–12908, 2010.

[69] D.B. Kitchen, H. Decornez, J.R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat Rev Drug Discov*, 3(11):935–949, 2004.

[70] E. Yuriev and P.A. Ramsland. Latest developments in molecular docking: 2010-2011 in review. *J Mol Recognit*, 26(5):215–239, 2013.

[71] O. Korb, T. Stützle, and T.E. Exner. An ant colony optimization approach to flexible protein–ligand docking. *Swarm Int*, 1(2):115–134, 2007.

[72] M.A. Hediger, B. Clémençon, R.E. Burrier, and E.A. Bruford. The ABCs of membrane transporters in health and disease (SLC series): introduction. *Mol Aspects Med*, 34(2–3):95–107, 2013.

[73] Ricardo J. Ferreira, Maria-José U. Ferreira, and Daniel J. V. dos Santos. Reversing cancer multidrug resistance: insights into the efflux by abc transports from in silico studies. *WIRES Comput Mol Sci*, 5(1):27–55, 2015.

[74] Vasanthanathan Poongavanam, Norbert Haider, and Gerhard F. Ecker. Fingerprint-based in silico models for the prediction of P-glycoprotein substrates and inhibitors. *Bioorg Med Chem*, 20(18):5388–5395, 2012.

[75] R. Edgar and E. Bibi. MdfA, an *Escherichia coli* multidrug resistance protein with an extraordinarily broad spectrum of drug recognition. *J Bacteriol*, 179(7):2274–2280, 1997.

[76] Haixia Liu, Xiaoqiang Liu, Yinqian Li, and Caiju Hao. Effect of six fluoroquinolones on the expression of four efflux pumps in the multidrug resistant *Escherichia coli* isolates. *World Journal of Microbiology and Biotechnology*, 31(7):1041–1048, 2015.

[77] Julia Adler, Oded Lewinson, and Eitan Bibi. Role of a conserved membrane-embedded acidic residue in the multidrug transporter MdfA. *Biochemistry*, 43(2):518–525, 2004.

[78] N. Fluman, C.M. Ryan, J.P. Whitelegge, and E. Bibi. Dissection of mechanistic principles of a secondary multidrug efflux protein. *Mol Cell*, 47(5):777–787, 2012.

[79] Jie Heng, Yan Zhao, Ming Liu, Yue Liu, Junping Fan, Xianping Wang, Yongfang Zhao, and Xuejun C Zhang. Substrate-bound structure of the *E. coli* multidrug resistance transporter MdfA. *Cell Res*, 25(9):1060–1073, 2015.

[80] N. Guex and M.C. Peitsch. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis*, 18:2714–2723, 1997.

[81] M.L. Nelson, B.H. Park, J.S. Andrews, V.A. Georgian, R.C. Thomas, and S.B. Levy. Inhibition of the tetracycline efflux antiport protein by 13-thio-substituted 5-hydroxy-6-deoxytetracyclines. *J Med Chem*, 36(3):370–377, 1993.

[82] N. Fluman, J. Adler, S.A. Rotenberg, M.H. Brown, and E. Bibi. Export of a single drug molecule in two transport cycles by a multidrug efflux pump. *Nat Commun*, 5, 2014. doi:10.1038/ncomms5615.

[83] A.P. Bento, A. Gaulton, A. Hersey, L.J. Bellis, J. Chambers, M. Davies, F.A. Krüger, Y. Light, L. Mak, S. McGlinchey, M. Nowotka, G. Papadatos, R. Santos, and J.P. Overington. The ChEMBL bioactivity database: an update. *Nucleic Acids Res*, 42:1083–1090, 2014.

[84] Teague Sterling and John J. Irwin. ZINC 15 — ligand discovery for everyone. *J Chem Inf Model*, 55(11):2324–2337, 2015.

[85] Humphrey, Dalke W., A., and K. Schulten. VMD - Visual Molecular Dynamics. *J Molec Graphics*, 14:33–38, 1996.

[86] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kale, and Klaus Schulten. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry*, 26:1781–1802, 2005.

[87] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M Feig, B.L. de Groot, H. Grubmuller, and A.D. MacKerell. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14:71–73, 2016.

[88] C.G. Mayne, J. Saam, K. Schulten, E. Tajkhorshid, and J.C. Gumbart. Rapid parameterization of small molecules using the Force Field Toolkit. *J. Comput. Chem.*, 34:2757–2770, 2013.

[89] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, S. Ha H. Guo, D. Joseph, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, I.W.E. Reiher, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.*, 102:3586–3616, 1998.

[90] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, Ö. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[91] R.A. Laskowski and M.B. Swindells. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *J. Chem. Inf. Model.*, 51:2778–2786, 2011.

[92] Chun Wei Yap. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry*, 32(7):1466–1474, 2011.

[93] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[94] Dudek J, Pfeffer S, Lee PH, Jung M, Cavalié A, Helms V, Förster F, and Zimmermann R. Protein transport into the human endoplasmic reticulum. *J Mol Biol*, 427:1159–75, 2015.

[95] S. Pfeffer, J. Dudek, R. Zimmermann, and F. Förster. Organization of the native ribosome-translocon complex at the mammalian endoplasmic reticulum membrane. *Biochim. Biophys. Acta*, 1860:2122–2129, 2016.

[96] Egea PF, Stroud RM, and Walter P. Targeting proteins to membranes: structure of the signal recognition particle. *Curr Opin Struct Biol*, 15(2):213–20, 2005.

[97] Mario Halic, Michael Blau, Thomas Becker, Thorsten Mielke, Martin R. Pool, Klemens Wild, Irmgard Sinning, and Roland Beckmann. Following the signal sequence from ribosomal tunnel exit to signal recognition particle. *Nature*, 444:507–511, 2006.

[98] Görlich D, Prehn S, Hartmann E, Kalies KU, and Rapoport TA. A mammalian homolog of SEC61p and SECYp is associated with ribosomes and nascent polypeptides during translocation. *Cell*, 71(3):489–503, 1992.

[99] R.M. Voorhees and R.S. Hegde. Structure of the Sec61 channel opened by a signal sequence. *Science*, 351:88–91, 2016.

[100] Stefan Pfeffer, Laura Burbaum, Pia Unverdorben, Markus Pech, Yuxiang Chen, Richard Zimmermann, Roland Beckmann, and Friedrich Förster. Structure of the native Sec61 protein-conducting channel. *Nature Communications*, 6(8403), 2015. doi:10.1038/ncomms9403.

[101] Görlich D and Rapoport TA. Protein translocation into proteoliposomes reconstituted from purified components of the endoplasmic reticulum membrane. *Cell*, 75(4):615–30, 1993.

[102] Lang S, Benedix J, Fedeles SV, Schorr S, Schirra C, Schäuble N, Jalal C, Greiner M, Hassdenteufel S, Tatzelt J, Kreutzer B, Edelmann L, Krause E, Rettig J, Somlo S, Zimmermann R, and Dudek J. Different effects of Sec61α, Sec62 and Sec63 depletion on transport of polypeptides into the endoplasmic reticulum of mammalian cells. *J Cell Sci*, 125:1958–69, 2012.

[103] Migliaccio G, Nicchitta CV, and Blobel G. The signal sequence receptor, unlike the signal recognition particle receptor, is not essential for protein translocation. *J Cell Biol*, 117(1):15–25, 1992.

[104] Fons RD, Bogert BA, and Hegde RS. Substrate-specific function of the translocon-associated protein complex during translocation across the ER membrane. *J Cell Biol*, 160(4):529–39, 2003.

[105] Martin Wiedmann, Teymuras V. Kurzchalia, Enno Hartmann, and Tom A. Rapoport. A signal sequence receptor in the endoplasmic reticulum membrane. *Nature*, 328:830–833, 1987.

[106] Shibatani T, David LL, McCormack AL, Frueh K, and Skach WR. Proteomic analysis of mammalian oligosaccharyltransferase reveals multiple subcomplexes that contain Sec61, TRAP, and two potential new subunits. *Biochemistry*, 44(16):5982–92, 2005.

[107] Dejgaard K, Theberge JF, Heath-Engel H, Chevet E, Tremblay ML, and Thomas DY. Organization of the Sec61 translocon, studied by high resolution native electrophoresis. *J Proteome Res*, 9(4):1763–71, 2010.

[108] Bañó-Polo M, Martínez-Garay CA, Grau B, Martínez-Gil L, and Mingarro I. Membrane insertion and topology of the translocon-associated protein (TRAP) gamma subunit. *Biochim Biophys Acta*, 1859(5):903–909, 2017.

[109] Jean-François Ménétret, Ramanujan S. Hegde, Mike Aguiar, Steven P. Gygi, Eunyong Park, Tom A. Rapoport, and Christopher W. Akey. Single copies of Sec61 and TRAP associate with a non-translating mammalian ribosome. *Structure*, 16(7):1126–1137, 2008.

[110] Pfeffer S, Dudek J, Schaffer M, Ng BG, Albert S, Plitzko JM, Baumeister W, Zimmermann R, Freeze HH, Engel BD, and Förster F. Dissecting the molecular organization of the translocon-associated protein complex. *Nat Commun*, 8(14516), 2017. doi:10.1038/ncomms14516.

[111] Sommer N, Junne T, Kalies KU, Spiess M, and Hartmann E. TRAP assists membrane protein topogenesis at the mammalian ER membrane. *Biochim Biophys Acta*, 1833(12):3104–11, 2013.

[112] Stefan Pfeffer, Johanna Dudek, Marko Gogala, Stefan Schorr, Johannes Linxweiler, Sven Lang, Thomas Becker, Roland Beckmann, Richard Zimmermann, and Friedrich Förster. Structure of the mammalian oligosaccharyltransferase complex in the native ER protein translocon. *Nature Communications*, 5, 2014. doi:10.1038/ncomms4072.

[113] Wiedmann M, Goerlich D, Hartmann E, Kurzchalia TV, and Rapoport TA. Photocrosslinking demonstrates proximity of a 34 kDa membrane protein to

different portions of preprolactin during translocation through the endoplasmic reticulum. *FEBS Lett*, 257(2):263–8, 1989.

[114] Conti BJ, Devaraneni PK, Yang Z, David LL, and Skach WR. Cotranslational stabilization of Sec62/63 within the ER Sec61 translocon is controlled by distinct substrate-driven translocation events. *Mol Cell*, 58(2):269–83, 2015.

[115] Losfeld ME, Ng BG, Kircher M, Buckingham KJ, Turner EH, Eroshkin A, Smith JD, Shendure J, Nickerson DA, Bamshad MJ, and Freeze HH. A new congenital disorder of glycosylation caused by a mutation in SSR4, the signal sequence receptor 4 protein of the TRAP complex. *Hum Mol Genet*, 23(6):1602–5, 2014.

[116] Ng BG, Raymond K, Kircher M, Buckingham KJ, Wood T, Shendure J, Nickerson DA, Bamshad MJ, Wong JT, Monteiro FP, Graham BH, Jackson S, Sparkes R, Scheuerle AE, Cathey S, Kok F, Gibson JB, and Freeze HH. Expanding the molecular and clinical phenotype of SSR4-CDG. *Hum Mutat*, 36(11):1048–51, 2015.

[117] Jürgen Cox and Matthias Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26:1367–1372, 2008.

[118] Jeffrey T. Leek, Robert B. Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*, 11(10):733–9, 2010.

[119] Benjamin Milo Bolstad. *preprocessCore: A collection of pre-processing functions*, 2016. R package version 1.34.0.

[120] Wolfram Stacklies, Henning Redestig, Matthias Scholz, Dirk Walther, and Joachim Selbig. pcamethods – a bioconductor package providing pca methods for incomplete data. *Bioinformatics*, 23:1164–1167, 2007.

[121] R. Tibshirani, G. Chu, Balasubramanian Narasimhan, and Jun Li. *samr: SAM: Significance Analysis of Microarrays*, 2011. R package version 2.0.

[122] `http://www.geneontology.org/page/go-slim-and-subset-guide`, 2017. Accessed: 13 June 2017.

[123] Ashburner *et al.* Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–9, 2000.

[124] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. *Nucl Acids Res*, 43(Database issue):D1049-56, 2015. doi:10.1093/nar/gku1179.

[125] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucl Acids Res*, 45(Database issue): D158D169, 2017. doi:10.1093/nar/gkw1099.

[126] Davor Juretić, Bono Lučić, Damir Zucić, and Nenad Trinajstić. Protein transmembrane structure: recognition and prediction by using hydrophobicity scales through preference functions. *Theoretical and Computational Chemistry*, 5:405–445, 1998.

[127] Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K., and Madden T.L. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(421), 2009. doi:10.1186/1471-2105-10-421.

[128] S. Schorr, M. C. Klein, I. Gamayun, A. Melnyk, M. Jung, N. Schäuble, Q. Wang, B. Hemmis, F. Bochen, M. Greiner, P. Lampel, S. K. Urban, S. Hassdenteufel, J. Dudek, X. Z. Chen, R. Wagner, A. Cavalié, and R. Zimmermann. Co-chaperone specificity in gating of the polypeptide conducting channel in the membrane of the human endoplasmic reticulum. *J Biol Chem*, 290(30):18621–35, 2015.

[129] Clapham DE. Calcium signalling. *Cell*, 131(30):1047–58, 2007.

[130] Reithinger JH, Yim C, Kim S, Lee H, and Kim H. Structural and functional profiling of the lateral gate of the Sec61 translocon. *J Biol Chem*, 289:15845–55, 2014.

[131] Cross BC, McKibbin C, Callan AC, Roboti P, Piacenti M, Rabu C, Wilson CM, Whitehead R, Flitsch SL, Pool MR, High S, and Swanton E. Eeyarestatin I inhibits Sec61-mediated protein translocation at the endoplasmic reticulum. *J. Cell Science*, 122(30):4393–440, 2009.

[132] Qiuyan Wang, Bidhan A. Shinkre, Jin gu Lee, Marc A. Weniger, Yanfen Liu, Weiping Chen, Adrian Wiestner, William C. Trenkle, and Yihong Ye. The ERAD inhibitor Eeyarestatin I is a bifunctional compound with a membrane-binding domain and a p97/VCP inhibitory group. *PLoS One*, 5(11):e15479, 2010.

[133] M.A. Marti-Renom, A.C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Sali. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.*, 29:291–325, 2000.

[134] M. Wiederstein and M.J. Sippl. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research*, 35:W407–10, 2007.

[135] The Open Babel package (2.4.0). `http://openbabel.org/wiki/Main_Page`.