# Biophysical and structural investigation of ABC transporters by compunational pharmacology

Ágota Tóth

# Ph.D. Thesis


## University of Limoges
### ED 615 – Sciences Biologiques et Santé (SBS)
**INSERM UMR1248 – Pharmacologie & Transplantation**


A thesis submitted to University of Limoges
in partial fulfillment of the requirements of the degree of
Doctor of Philosophy
Discipline / Speciality: Computational Chemistry


Presented and defensed by
**Ágota Tóth**


On December 7, 2022

## Biophysical and Structural Investigation of ABC Transporters by Computational Pharmacology

Thesis supervisor: Florent Di Meo, Ph.D.

JURY:

Reporters
Ms. **Isabelle Callebaut**, Ph.D. Senior Researcher, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, UPMC-Sorbonne Université, CNRS, IRD and MNHN, Paris, France
M. **Thomas Stockner**, Ph.D. Associate Professor, Center for Physiology and Pharmacology, Institute of Pharmacology, Medical University of Vienna, Austria

Examiners
Ms. **Elena Papaleo**, Ph.D. Associate Professor, DTU Health Tech, Danish Cancer Society Research Center, Copenhagen, Denmark
M. **Florent Di Meo**, Ph.D. Inserm researcher (CRCN), Inserm U1248 P&T, University of Limoges, France
Ms. **Véronique Blanquet**, Ph.D. Professor, IU1094 Inserm U270 IRD USC1501 INRAE EPIMACT, Institut Ω-Health, Université de Limoges, France
M. **Eric Pinaud**, Ph.D. Senior Researcher, B-Nation, CNRS 7276 Inserm 1262 CRIBL, Institut Ω-Health, Université de Limoges, France

Dedication

*“When you pass through the waters, I will be with you;*
*and when you pass through the rivers, they will not sweep over you.*
*When you walk through the fire, you will not be burned;*
*the flames will not set you ablaze.”*

Isaiah 43:2

## Acknowledgements

First of all, I would like to address my thank to **Florent Di Meo**. I feel lucky to have you as my supervisor. Thank you for enlightening me to theoretical chemistry. I am thankful to you for everything you taught me, your academic and scientific qualities, help, your patience, your sacrifices, and your generosity. It has been a pleasure to work by your side.

I also thank to **all members of the jury**. I am truly honoured that you accepted to be part of this jury. I appreciate your kindness, your expertise and your insightful discussions.

**Isabelle Callebaut**, it is an honour to have you as a jury member. Thank you for accepting to review this work and for your insightful thoughts regarding to ABC transporters.

**Thomas Stockner**, thank you for accepting to review this work. Your expertise in the theoretical methods and degenerate NBSs have brought interesting discussions.

**Elena Papaleo,** I am really grateful to have you as a member of my jury. It is thanks to you that I have found this Ph.D. position.

**Véronique Blanquet** and **Eric Pinaud**, I would like to express my gratitude for expecting being a member of my committee.

I thank all members of the Inserm U1248, especially **Prof. Pierre Marquet** head of the unit, for your help and support during my Ph.D. studies. Thank you for the friendly atmosphere and the programs in and out of the lab. I wish you the best for your personal and professional lives.

I would like to address my gratitude to the **theoretical chemists** in Limoges for your help, insightful discussions, all the knowledge that you taught to me, and the friendly environment during my Ph.D studies. Thanks to you, it was always a pleasure to go to work. I wish you all the best for a bright scientific carrier that you deserve. Special thanks to **Angelika Janaszkiewicz** and **Mehdi Benmameri** for all moments passed together and to tolerate me in the office during these three years. You understand what I mean.

My sincere thanks to all collaborators from other cities and countries who helped me broaden my knowledge and build this research topic. Thank you **Amel Ben Saad** and **Thomas Falguières** for the opportunity to work together on ABCB4. **Louis-Gabriel Bronstein** and **Ali Makky**, thanks for the knowledge about lipids and porphyrin conjugates that our collaborative work gave me. Finally, but not least, I would like to express my gratitude to **Lucie Delemotte** and the **SciLifeLab** for the opportunity to learn about biased methods and techniques to solve the collective variable definition problem. Thank you all for all the insightful and motivating discussions. I wish you the best for your personal and professional lives.

I am also grateful for all the available resources and for the IT support of **Xavier Montagutelli**.

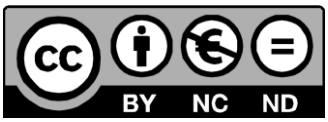
I am grateful to have my friends. Thank you **András Mihálykó** and **Gergő Kulcsár** for keeping the friendship such a long time and for all the distance support. I am thankful to **Anna Csicsics** and **Boglárka Csuka** to listening me when I had a difficult time. I would like to express my gratefulness to my friends from work – **Angelika Janaszkiewicz** and **Veronica Crespi** – who supported me a lot especially during the last month of my Ph.D. when I needed it the most. All of you contributed somehow for this work. I cannot express how grateful I am.

Thank you, **Mathieu Thizy**, and your family for your support during this Covid-19 period that I had to spend far from my family. You helped me a lot not to be totally isolated and to have a calm home office.

I am thankful to have the support of my **godparents**. Finally, but not least my biggest thank is going to **my family** who supported me from the beginning. They helped me to move into another country, they always listen my problems, and encourage me. In particular, I thank my sister, **Eszter Tóth**, who is always ready to answer my English related questions. I am really lucky to have such a great family.

## Rights





## Table of Contents

## List of Figures

## List of Tables

## List of Equations

## Acronyms

| | |
|---|---|
| **ABC** | ATP-Binding Cassette |
| **ABCA** | ATP-Binding Cassette A sub-family |
| **ABCB** | ATP-Binding Cassette B sub-family |
| **ABCC** | ATP-Binding Cassette C sub-family |
| **ABCC1** | First member of ATP-Binding Cassette C sub-family |
| **ABCD** | Administration – Bioavailability – Clearance – Distribution |
| **ABCD** | ATP-Binding Cassette D sub-family |
| **ABCE** | ATP-Binding Cassette E sub-family |
| **ABCF** | ATP-Binding Cassette F sub-family |
| **ABCG** | ATP-Binding Cassette G sub-family |
| **ADP** | Adenosine DiPhosphate |
| **AE** | AutoEncoder |
| **AI** | Artifical Intelligence |
| **APL** | Area Per Lipid |
| **ATP** | Adenosine TriPhosphate |
| **AWH** | Accelerated Weight Histogram |
| **BSEP** | Bile Salt Export Pump |
| **Cryo-EM** | Cryogenic Electron Microscopy |
| **CV** | Collective Variable |
| **DBSCAN** | Density-Based Spatial Clustering of Applications with Noise |
| **DDS** | Drug Delivery System |
| **DHS** | Dihydrosphingosine |
| **FES** | Free Energy Surface |
| **FF** | Force Field |
| **FN** | False Negative |
| **FN** | False Negative |
| **FP** | False Positive |
| **FP** | False Positive |
| **GAFF** | General Amber Force Field |
| **GMM** | Gaussian Mixture Model |
| **IF** | Inward-Facing |
| **InfleCS** | Inflection Core State |
| **ITC** | International Transporter Consortium |
| **(L)ADME(-Tox)** | Liberation – Absorption – Distribution – Metabolism – Elimination - Toxicity |
| **MATE** | Multidrug and Toxin Extrusion Transporter |
| **MD** | Molecular Dynamics |
| **ML** | Machine Learning |
| **MLP** | MultiLayer Perceptron |
| **MM** | Molecular mechanics |
| **MRP** | Multidrug Resistance-associated Protein |
| **MRP1** | First member of Multidrug Resistance-associated Protein |
| **NBD** | Nucleotide Binding Domain |
| **NBS** | Nucleotide Binding Site |
| **NN** | Neural Network |
| **OAT** | Organic Anion Transporter |
| **OATP** | Organic Anion Transporting Polypeptide |

| | |
|---|---|
| **OCT** | Organic Cation Transporter |
| **OF** | Outward-Facing |
| **PBC** | Periodic Boundary Conditions |
| **PC** | Principal Component |
| **PCA** | Principal Component Analysis |
| **PD** | PharmacoDynamics |
| **PHS** | phytosphingosine |
| **PK** | PharmacoKinetics |
| **pmemd** | Particle Mesh Ewald Molecular Dynamics |
| **QM** | Quantum Mechanics |
| **QM** | Quantum Mechanics |
| **RBM** | Restricted Boltzmann Machine |
| **RF** | Random Forest |
| **RMSD** | Root Mean Square Deviation |
| **sander** | Simulated Annealing with NMR-Derived Energy Restraints |
| **SLC** | Solute Carrier |
| **SMD** | Steered Molecular Dynamics |
| **SPH** | Sphingosine |
| **SVM** | Support Vector Machine |
| **TMD** | TransMembrane Domain |
| **TMH** | TransMembrane Helix |
| **TN** | True Negative |
| **TN** | True Negative |
| **TP** | True Positive |
| **TP** | True Positive |
| **UR** | Unlock-Returned |

## Introduction

Over the past decades, the importance of personalized and precision medicine has been growing more attention as shown *e.g.*, by the increased number of drug monitoring therapies. One of the purposes of these approaches is to pay more attention to individual factors. This is even more important when drug cocktails are used, as in post-transplantation therapies. Over the past decades, many pharmacokinetics tools for improved dose individualization and pharmacogenomics have been proposed to better understand inter-individual variability. Many factors have been identified; however, there is still a significant unexplained part which remains misunderstood. Perhaps the combination of low-penetrance or rare variability factors is underestimated. To better understand these processes, several events that are only implicitly included in systemic pharmacology models need to be considered.

Xenobiotics must cross many membranes to reach their target, get metabolized or eliminated. Therefore, membrane crossing events are substantial in pharmacology. Such events affect both drug concentration in several body compartments (pharmacokinetics – PK) as well as their different effects (pharmacodynamics – PD). There is a need for deeper understanding and finer quantification of PK/PD relationships. Drug-membrane crossing is directly involved in local PK, *i.e.*, drug concentration at *e.g.*, the target sites, whether linked with drug therapeutic or adverse effects. Nevertheless, knowledge about membrane crossing mechanisms is still fragmented [1].

Particular attention should be paid to human transporters which drive membrane crossing event. ATP-Binding Cassette (ABC) proteins and solute carrier (SLC) proteins are crucial in drug absorption, distribution, metabolism and elimination (ADME) [2], [3]. Some of them were described as “emerging clinical importance” by the International Transporter Consortium (ITC). For instance, ABCB1 (P-gp), SLCC22A6 (OAT1), SLCO1B1 (OATP1B1), ABCC2 (MRP2), and ABCC4 (MRP4) are involved in drug elimination events [2], [3]. This is indeed particularly important to focus on liver and kidney transporters. These organs are of crucial importance in the ADME model as they govern the modification and filtration of most of the drugs. Consequently, it is important to know more about the functioning of transporters located in these organs since they are directly linked with PK/PD relationship. These transporters have not been all structurally resolved yet, precluding the atomic resolution of their functions. However, one can take advantage from similar resolved transporters (*e.g.*, from the same family) to have an overview and better decipher their activities.

There is currently no single technique capable of providing a comprehensive and dynamic overview of xenobiotic-transporter interactions. Computational chemistry, molecular modelling or *in silico* pharmacology appear relevant to provide an atomic and dynamic picture of these events. Among them, all-atom molecular dynamics (MD) simulations have become a powerful tool to rationalize the structure and functions of transporters, including the role of the surrounding lipid bilayer [4]. Besides, biased MD simulations combined with machine learning techniques were shown to accurately describe transport cycles [5]-[8]. Machine learning can also be considered as advanced analysis technique to *e.g.*, reduce the high-dimensionality of raw outputs from MD simulations [5].

In the present manuscript, computational techniques were used in order to better understand the role and function of ABC membrane transporters. The main focus was paid on ABCC family transporters, more specifically ABCC1 (MRP1) for which its bovine ortholog has been structurally resolved. Structural investigations were carried out using MRP1 as an ABCC

transporter prototype in order to better understand the function of the pharmacologically relevant ABCC2, ABCC3, and ABCC4. The interplay between lipid and protein interactions was checked, as well. Besides, a well-known lipid-transporter, namely ABCB4, was also investigated in the context of joint computational experimental collaboration. Finally, computational techniques were also used to decipher the structure and biophysics of the supramolecular assembly of lipid bilayers incorporating porphyrin conjugates. These systems have been investigated in order to ultimately develop specific drug-release pharmaceutical devices.

The present manuscript is divided into six chapters. First, in Chapter I, the pharmacological background (section I.1), membrane components (section I.2) and different types of transports across them (section I.3), as well as pharmacologically important transporters (section I.4) are detailed. The basic concept of MD simulations is then explained in Chapter II. First a general description is given about the available theoretical methods (section II.1). This is followed by the detailed explanation of molecular mechanics (sections II.2), molecular dynamics (section II.3), parameter definition of force fields (section II.4), biased methods (section II.5), and molecular docking (section II.6). Artificial intelligence as well as machine learning based algorithms and methods are expounded in Chapter III.

Chapter IV is divided into two sections, describing two main studies on ABCC1/MRP1. The first one (section IV.1) gives a wide overview about the dynamics of *b*MRP1 in different states along the transport cycle embedded into different POPC-, POPE-, and cholesterol-based membrane compositions. Section IV.2 focuses on the pre- and post-hydrolysis outward-facing states of *b*MRP1. The results of a study on another ABC transporter, namely ABCB4, is detailed in Chapter V in which particular attention was paid to interactions between transporter and small molecules. Molecular docking study was carried out to assess the influence of CFTR correctors on the bile secretion mediator ABCB4 transporter. Finally, Chapter VI. describes joint computational and experimental investigations of phospholipid-porphyrin conjugates used for *e.g.*, photo-activatable drug delivery systems.

## References


[1] Mateus, A. *et al.* Prediction of intracellular exposure bridges the gap between target- and cell-based drug discovery. *Proc Natl Acad Sci* **114**, (2017).

[2] The International Transporter Consortium *et al.*, 'Membrane transporters in drug development', *Nat Rev Drug Discov*, vol. 9, no. 3, pp. 215–236, Mar. 2010, doi: 10.1038/nrd3028.

[3] M. J. Zamek-Gliszczynski *et al.*, 'Transporters in Drug Development: 2018 ITC Recommendations for Transporters of Emerging Clinical Importance', *Clin Pharmacol The*, vol. 104, no. 5, pp. 890–899, 2018, doi: https://doi.org/10.1002/cpt.1112.

[4] T. Stockner, R. Gradisch, and L. Schmitt, 'The role of the degenerate nucleotide binding site in type I ABC exporters', *FEBS Lett*, vol. 594, no. 23, pp. 3815–3838, Dec. 2020, doi: 10.1002/1873-3468.13997.

[5] O. Fleetwood, M. A. Kasimova, A. M. Westerlund, and L. Delemotte, 'Molecular Insights from Conformational Ensembles via Machine Learning', *Biophys J*, vol. 118, no. 3, pp. 765–780, Feb. 2020, doi: 10.1016/j.bpj.2019.12.016.

[6] F. Noé, G. De Fabritiis, and C. Clementi, ‘Machine learning for protein folding and dynamics’, *Curr Opin Struct Biol*, vol. 60, pp. 77–84, Feb. 2020, doi: 10.1016/j.sbi.2019.12.005.

[7] A. Mardt, L. Pasquali, H. Wu, and F. Noé, ‘VAMPnets for deep learning of molecular kinetics’, *Nat Commun*, vol. 9, no. 1, p. 5, Dec. 2018, doi: 10.1038/s41467-017-02388-1.

[8] D. Mitrovic, S. E. McComas, C. Alleva, M. Bonaccorsi, D. Drew, and L. Delemotte, ‘Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning’. bioRxiv, p. 2022.09.24.509294, Sep. 26, 2022. doi: 10.1101/2022.09.24.509294.

# Chapter I. Pharmacology and drug transport across cell membranes

## I.1. Drug membrane crossing events in pharmacology

Many different types of drugs are taken during our life. In some cases, drug usage is occasional. In case of some diseases, treatments take a long time. For instance, after solid organ transplantation, the treatment even lasts lifelong. The active substance often crosses several membranes until it reaches its target. Therefore, xenobiotic membrane crossing events, that affect both the concentration of the drugs (pharmacokinetics) and the effects of the drugs (pharmacodynamics), are of particular importance in pharmacological research. Impaired transport processes are key in various diseases. One striking and intensively investigated event is drug resistance in cancer cells [1]–[3]. However, in this thesis, the context is the treatment of transplanted people as there was no improvement in the post-transplantation treatment in the last decades. To improve the post-transplantation treatment, among other factors (*e.g.*, patient education), it is important to gain more knowledge about the journey of a drug [1], [4], [5].

Drugs can be taken in numerous ways (Table 1). To be efficient, the active agent must reach its target at a sufficient concentration. Except for local treatments, xenobiotics crossing membranes are distributed by the blood flow, called systemic effect. The LADME-Tox (Liberation, Absorption, Distribution, Metabolism, Excretion, and Toxicity) model is often used to describe the journey of xenobiotics in the body where various events occur [1], [4]–[6].

**Table 1. Drug dosage options [7].**

<table>
<tr><th></th><th colspan="2">Enteral</th><th>Parenteral</th></tr>
<tr><td rowspan="7">Invasive</td><td colspan="2" rowspan="7"></td><td>Intravenous</td></tr>
<tr><td>Intramuscular</td></tr>
<tr><td>Subcutaneous</td></tr>
<tr><td>Intraarterial</td></tr>
<tr><td>Intrathecal</td></tr>
<tr><td>Intracardiac</td></tr>
<tr><td>Intraperitoneal</td></tr>
<tr><td rowspan="10">Non-invasive</td><td rowspan="2">Oral</td><td>Sublingual</td><td>Transdermal</td></tr>
<tr><td>Buccal</td><td>Inhalation</td></tr>
<tr><td colspan="2" rowspan="8">Peroral</td><td>Ocular</td></tr>
<tr><td>Nasal</td></tr>
<tr><td>Oticular</td></tr>
<tr><td>Dermal</td></tr>
<tr><td>Vaginal</td></tr>
<tr><td>Rectal</td></tr>
<tr><td>Urethral</td></tr>
</table>

### I.1.1. The LADME-Tox model

**Liberation**, *i.e.*, the release of an active agent from a drug delivery device, is the first step in drug transport. The place of liberation determines the path of the drug in the body and the next steps of the drug's journey. As mentioned above, the effect may be local, for example, if the uptake is *via* skin, eyes, or ears. However, in many cases, the molecules are absorbed to reach blood circulation [1], [4], [5].

**Absorption**, *i.e.*, the uptake of xenobiotics into the blood circulation, can happen passively and through membrane transporters (see section I.3). The uptake depends on the environment, the physical barrier that the drug should cross, and the physical and chemical properties of the transported molecule. These attributes determine the membrane crossing mechanism. The different types of transport mechanisms through membranes are described in detail in section I.3. Absorption can be delayed to achieve a targeted effect or to prolong the effect of a drug. It can be accomplished by various drug delivery systems (DDSs) (see section I.2.4) [1], [4], [5].

The absorbed drugs (*i.e.*, drugs that have reached systemic blood circulation) are distributed in the body. This happens fast as the blood flow is rapid. In blood, drugs may bind to human serum albumin which is the primary transport and reservoir protein in systemic circulation. Thus, albumin influence the pharmacokinetic and pharmacodynamic properties of drugs [8]. During **distribution**, drugs cross some barriers to get into the target cells. In this process, similar transport mechanisms play an important role like in case of absorption, but interactions with albumin are important, as well [1], [4], [5], [8].

In the body, molecules may undergo chemical transformations to reduce their toxicity. Such events usually take place in hepatocytes as well as in the intestinal gut wall (first pass metabolism) but to a lesser extent. In the liver, this native detoxifying system involves two main types of enzyme-mediated metabolism reactions. Phase I metabolism is often defined by oxidation, reduction, or hydrolysis. These catabolic reactions set up xenobiotics for phase II reactions. Phase II reactions lead to the conjugation of xenobiotics to *e.g.*, glucuronic acid, sulfonic acid or glutathione. Overall, liver metabolism aims to produce more hydrophilic compounds, favouring subsequent elimination. However, metabolites may be more toxic than the parent xenobiotic. In pharmacology, the most studied detoxification Phase I enzymes are members of the cytochrome P450 family. Sometimes a drug requires to be metabolised to become active. In this case, the originally taken xenobiotic is called prodrug, *i.e.*, the inactive precursor of the active molecule. Those xenobiotics and metabolites that are not eliminated during the metabolism steps can be transferred to the bile and delivered to the intestine from where a part can be reabsorbed thus prolonging the drug's effect [1], [4], [5], [9].

**Excretion** can already happen in the liver for more lipophilic compounds; however, kidneys are the main excretory routes. The rate of excretion is different to a great extent of molecules. If these mechanisms do not work properly, *i.e.*, the drugs are not excreted correctly, xenobiotics may accumulate and cause **toxicity**. It can happen, for example, because of altered drug metabolism in elderly people or patients with renal or liver diseases. Drug-drug and drug-food interactions can reduce drug pharmacokinetics, as well. The therapeutic effect is the result of a complex interaction between the drug and the organism depending on many factors (*e.g.*, biological, chemical, and biochemical), in which individual membrane crossing events are key factors [1], [4], [5], [9].

Liberation and toxicity are not always included in the drug quantifying descriptors, so the ADME acronym often appears.

### I.1.2. The ABCD model in pharmacokinetics

For pharmacokinetics, another model, namely ABCD, has been proposed. ABCD comes from Administration, Bioavailability, Clearance, and Distribution. It is tightly connected to the ADME model; however, ABCD provides an overview from the prescription to the patient response.

The above-mentioned absorption is important in all **administration** pathways (Fig. 1), except for intravenous injection where the drug is directly administrated into the plasma. The most common administration is oral whose speed depends on many factors (*e.g.*, fed or fasted stomach, the acidity of the molecule).

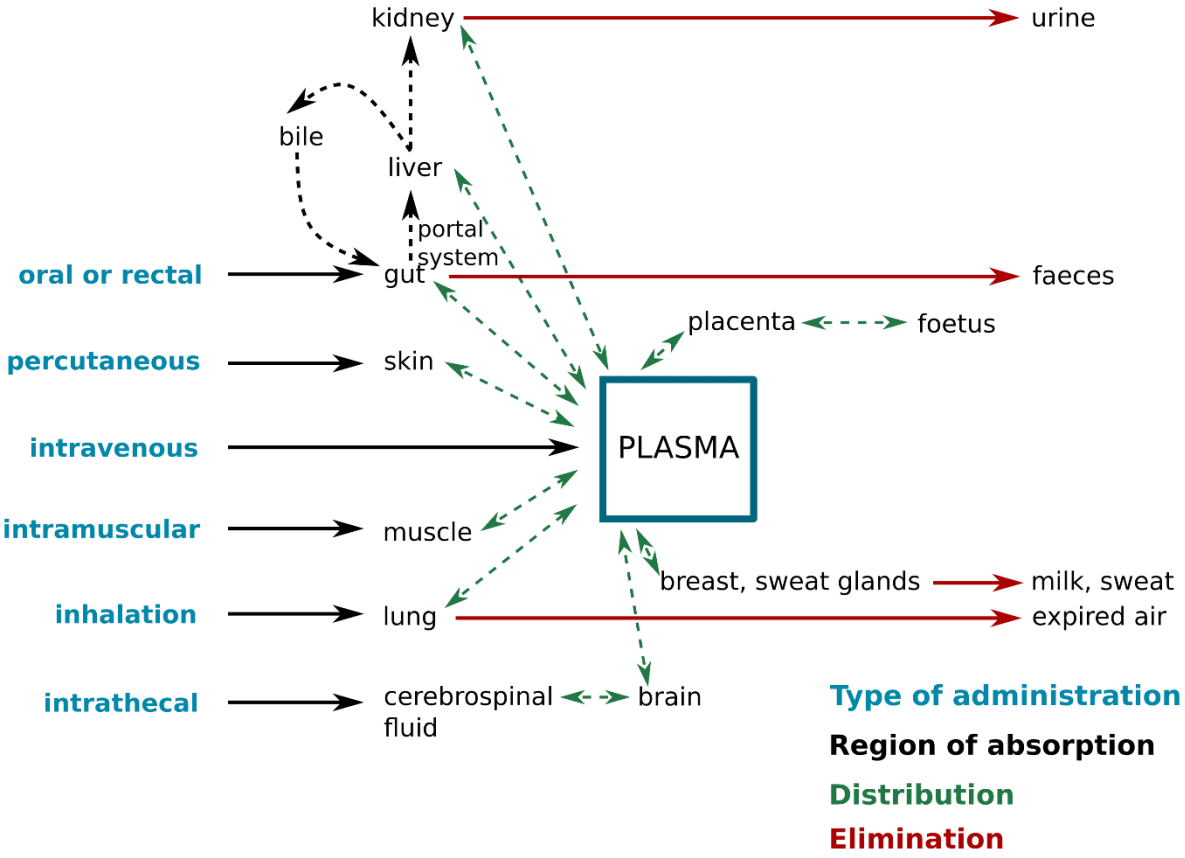



**Figure 1. The routes of drugs in the body depending on the type of administration.**
[Based on Fig. 8.8 of [5] and [4].]

The fraction of an administered drug that enters the systemic circulation is called **bioavailability**. For orally administrated drugs, bioavailability includes the first pass metabolism, *i.e.*, crossing the intestinal gut wall. Bioavailability can rise if this first pass metabolism is reduced, for example, by an alternative, targeted delivery system (see section I.2.4).

After metabolism, a part of the active drug can be excreted. The rate of elimination, *i.e.*, the amount of active substance eliminated per unit of time is called **clearance**. In some cases, the excreted drug ratio concerning the available fraction is large, leading to a lower bioavailability.

The administrated but not excreted xenobiotics are distributed. Thanks to the **distribution,** drugs get close to their targets where they can accomplish their therapeutic effect [1], [4], [5]. As mentioned above, in this process, membrane crossing events are crucial. In the next sections, insights about membranes, membrane crossing mechanisms, and pharmacologically relevant membrane protein families are provided.

## I.2. The structure of a cell membrane

### I.2.1. Membrane construction

Membranes are formed by lipid molecules. Most of the lipids contain a polar (hydrophilic) head group to which two non-polar (hydrophobic) fatty acid tails are bound. More about lipid structures is described in section I.2.3. Lipids can be organized into a thin (4-7 nm) bilayer, building a membrane. The fatty acid tails are oriented to each other creating a two-layers membrane arrangement where the hydrophilic heads form the interface with water molecules (Fig. 2). Different types of biological membranes can be found at diverse places in the human body even though they share a common structure. All types of biological membranes are thin, dynamic, fluid, and composed of lipids and proteins (Fig. 2). They have a separating function. The plasma membrane, a.k.a. cell membrane, segregates the extra- and intracellular space, *i.e.*, it borders the cell from its environment. This separation is essential for Life as this maintains the crucial distinct contents of different cells. Intracellular membranes isolate the contents of cell organelles (*e.g.*, mitochondria, endoplasmic reticulum, or Golgi apparat) to the cytoplasmic compartment. Connections between cell membranes but also between organelles within a cell are made by complicated signalling systems in which proteins are the sensors. Membrane composition alters dynamically according to *e.g.*, environmental modification, but it is under homeostatic control, *i.e.*, it maintains relative stable. The homeostasis of lipids is important as changes can be connected to various diseases which are often connected to a gene mutation that results in an altered lipid composition [10]. Nevertheless, our knowledge about the mechanisms occurring in membranes is still limited. This makes membrane protein, lipid, and signalling system studies substantial [1], [10], [11]. In the context of pharmacology, focus will be on the plasma membrane given its central role in all key steps of drug pharmacokinetics.

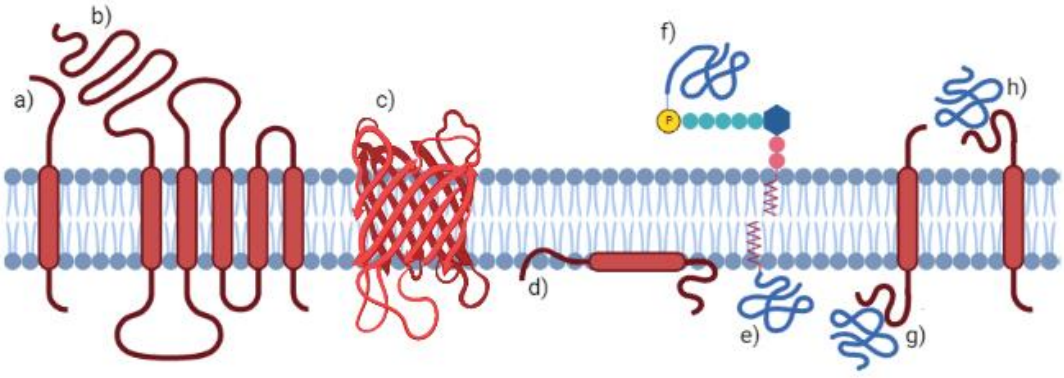



**Figure 2. Schematic lipid bilayer with proteins.**

Transmembrane proteins can cross the membrane by **a)** a single or **b)** multiple $\alpha$-helix, or **c)** $\beta$-barrel. Membrane-associated proteins can anchor by **d)** amphiphilic $\alpha$-helix, **e)** lipid chains, **f)** covalent oligosaccharide linkage, or they can bind **g-h)** to other membrane proteins. [Inspired by Figure 10-17 [1], created with BioRender.com].

### I.2.2. Proteins

There are many different types of proteins in the plasma membrane. These proteins are essential for the proper functioning of a cell as they play a central role in interactions between a cell and its environment. Membrane proteins are responsible for the various functional characteristics of the different cell membranes. The fraction (25-75%) of membrane proteins and their types are highly variable depending on cells and organelles. Membrane proteins can be integrated into the bilayer (transmembrane proteins) as a single or multiple $\alpha$-helix or as a $\beta$-barrel. Most of the proteins span the membrane. Sometimes proteins are just anchored to the membrane surface by an amphiphilic $\alpha$-helix, lipid chains (in the inner leaflet), or a covalent oligosaccharide linkage (in the outer leaflet). Furthermore, proteins can bind to each other, including membrane or cytosol proteins creating a network (Fig. 2). Membrane transporters can form complexes and mediate complicated biological processes, so they are playing an important role in pharmacology, as well [1], [11]. The role of proteins in different types of transport processes across a cell membrane is explained in section I.3. Two families of membrane transporters, that are important in drug influx and efflux, are discussed at the end of this chapter (see section I.4).

### I.2.3. Lipids

#### I.2.3.1. Lipids in general

Lipids are composed of a polar head group and two non-polar fatty acid tails. The polar head can be split into three parts (Fig. 3a): base, phosphate and alcohol, *i.e.*, glycerol for a typical phospholipid. The alcohol is linked to the fatty acid tails and the phosphate, while the phosphate is further linked to a small, polar group (*e.g.*, choline) called base in general (Fig. 3a). The length of fatty acid chains can differ by the number of carbon atoms and double bonds. Saturated chains (*e.g.*, palmitic acid) do not contain any double bonds (Fig. 4). Fatty acid chains that have one or more double bonds are called monounsaturated (*e.g.*, oleic acid) and polyunsaturated, respectively. In presence of a double bond, the cis- or Z-configuration leads to a kink resulting in a shorter acyl chain length as compared to a saturated chain containing the same number of carbon atoms. Fatty acid chains connected to the polar head are often different in a lipid. One is called sn-1 chain, while the other one sn-2 chain following the stereospecific numbering of the head group glycerol (Fig. 3a). In mammalians, many of the fatty acid chains are unsaturated, often polyunsaturated [1], [10]–[12].

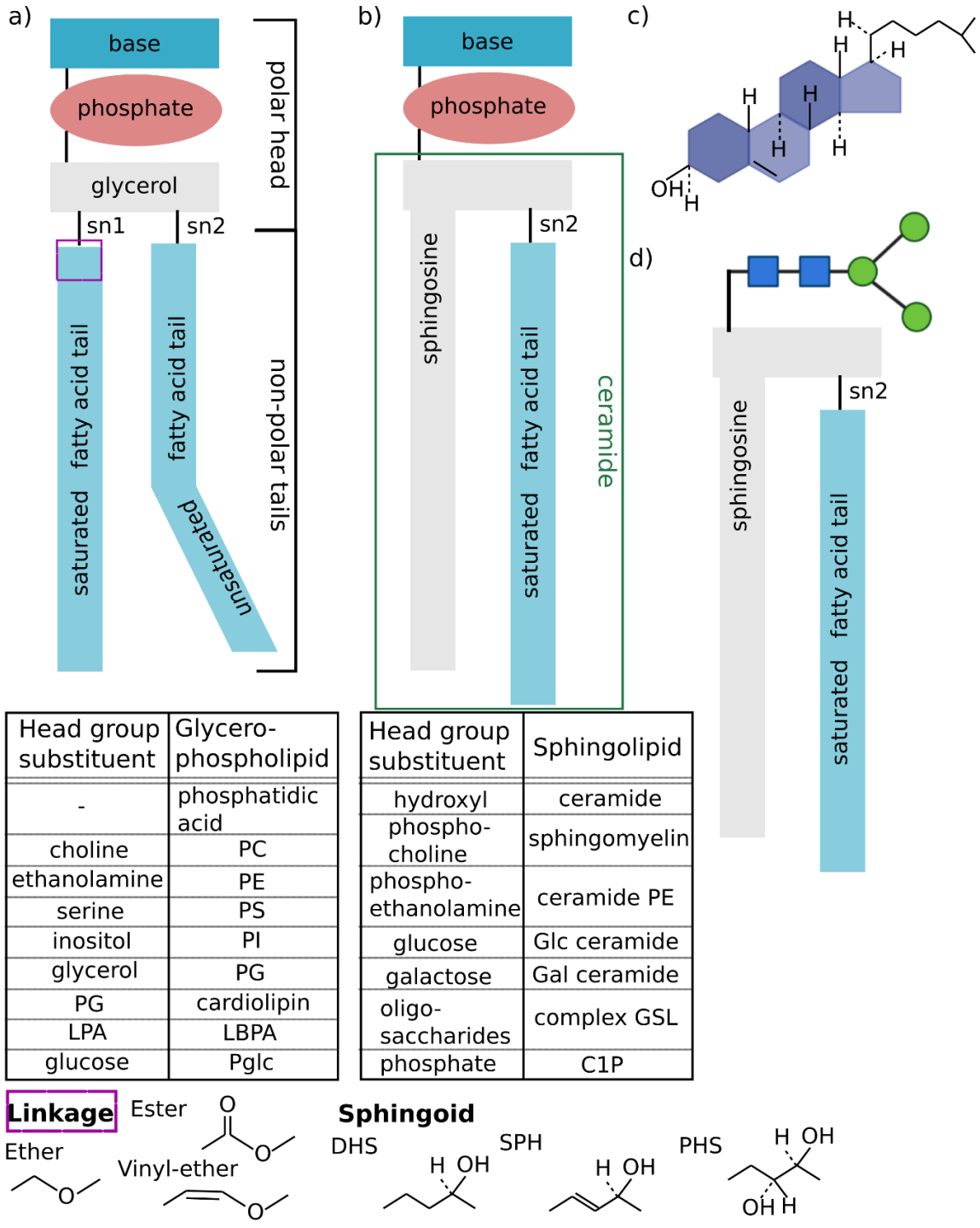



**Figure 3. Different types of lipids.**

**a)** Schematic representation of glycerophospholipids and the possible sn-1 linkages and head groups. **b)** Schematic representation of sphingolipids and the possible sphingoid base types and head groups. **c)** Schematic representation of cholesterol, and **d)** glycolipids. [Inspired by [1], [10], [11].]

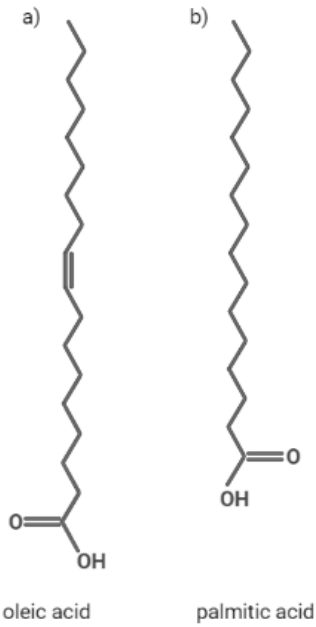



**Figure 4. Different types of side chains.**
**a)** Unsaturated oleic acid and **b)** saturated palmitic acid. [Created with BioRender.com.]

Plasma membranes contain many various lipids; however, most of them are phospholipids, sphingolipids, and sterols, mostly cholesterols in mammalians. Glycolipids often have an important fraction, as well. These lipids will be discussed in detail in the upcoming sections. The ratio of the different types of lipids and proteins is associated with various cell-specific compositions (Table 2) and functions [1], [11]–[13]. It is important to know more about the membrane compositions, membrane proteins and the processes taking place in the membrane. Nowadays, more attention has been paid to lipid composition and its impact on membrane proteins and transport processes both in experimental and theoretical studies [10].

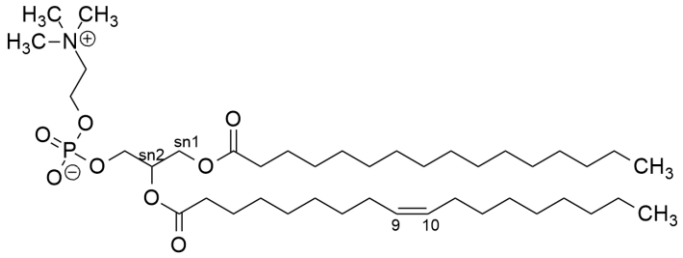
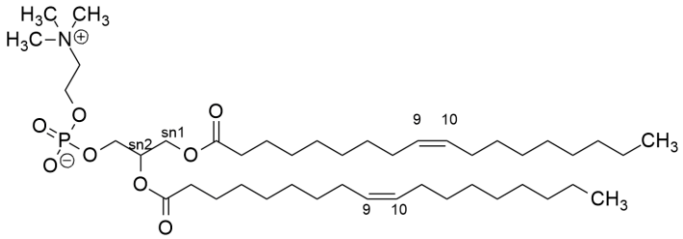
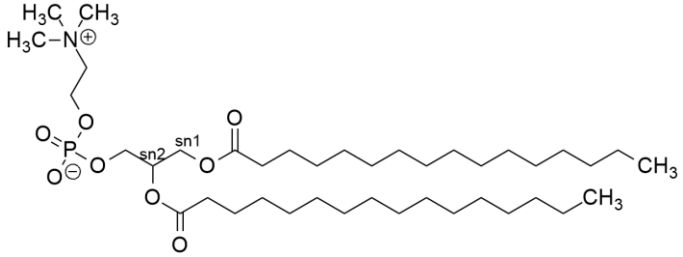
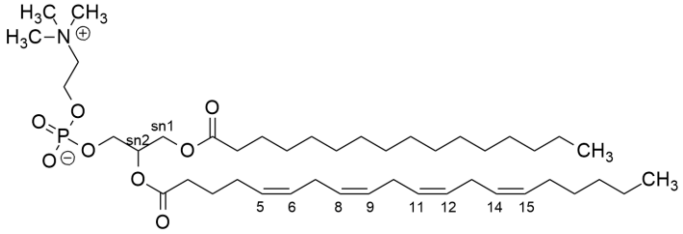
**Table 2. Approximate Lipid Compositions of Different Cell Membranes**.
[Table10-1 from [1].]

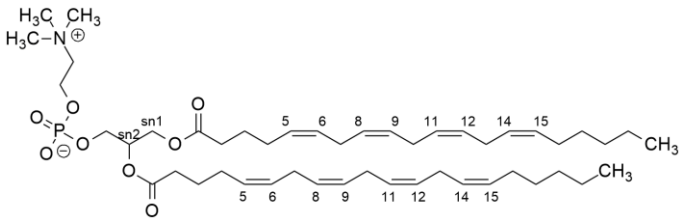
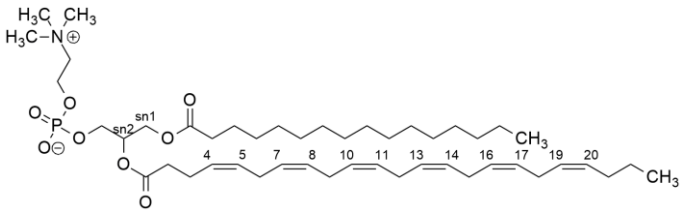
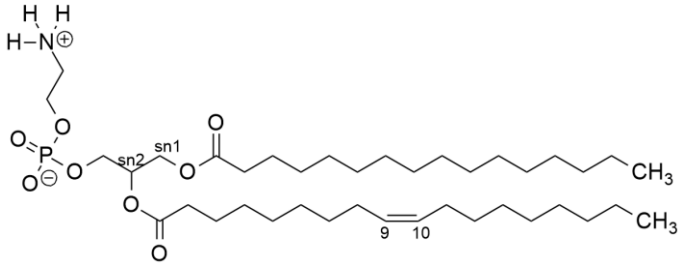
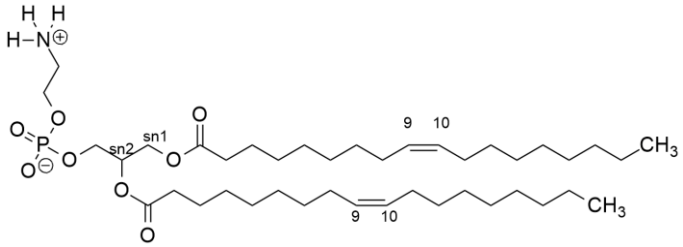
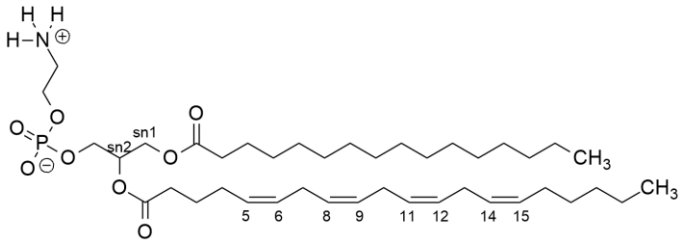
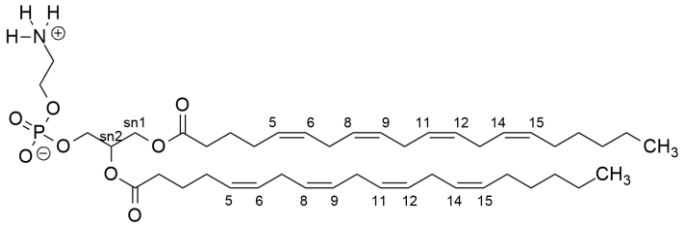
| Lipid | Percentage of total lipid by weight | | | | | |
|---|---|---|---|---|---|---|
| | Liver cell plasma membrane | Red blood cell plasma membrane | Myelin | Mitochondrion (inner and outer membranes) | Endoplasmic reticulum | *E. coli* bacterium |
| Cholesterol | 17 | 23 | 22 | 3 | 6 | 0 |
| Phosphatidylethanolamine | 7 | 18 | 15 | 28 | 17 | 70 |
| Phosphatidylserine | 4 | 7 | 9 | 2 | 5 | trace |
| Phosphatidylcholine | 24 | 17 | 10 | 44 | 40 | 0 |
| Sphingomyelin | 19 | 18 | 8 | 0 | 5 | 0 |
| Glycolipids | 7 | 3 | 28 | trace | trace | 0 |
| Others | 22 | 14 | 8 | 23 | 27 | 30 |

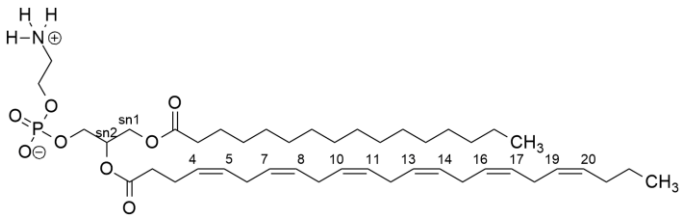
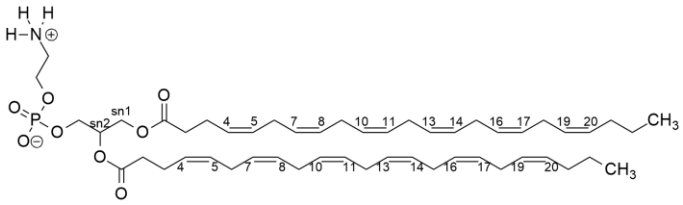
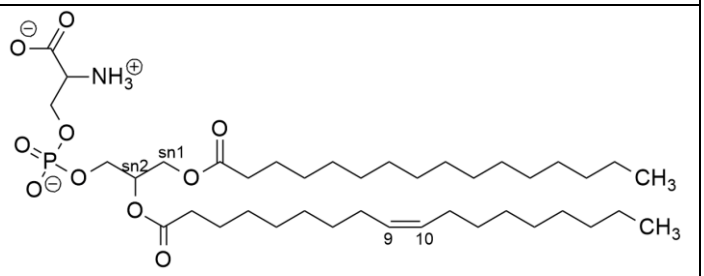
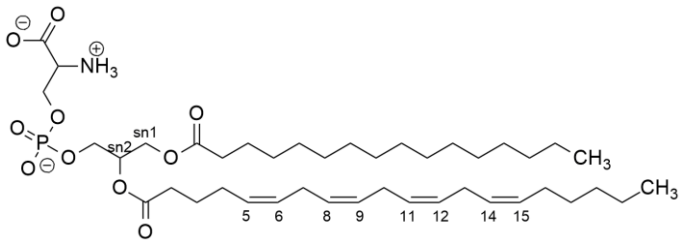
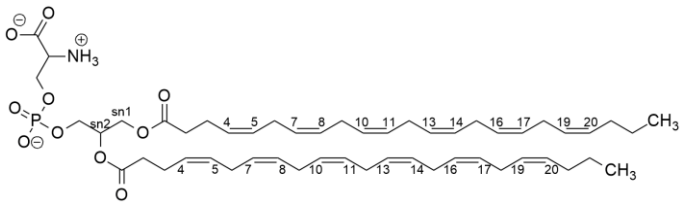
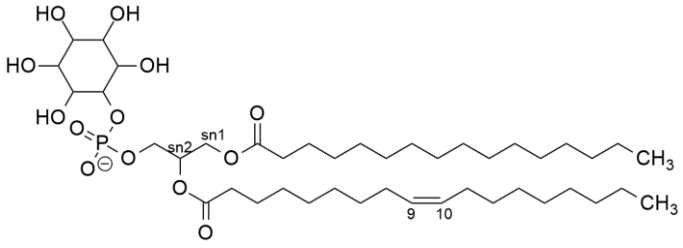
### I.2.3.2. Glycerophospholipids

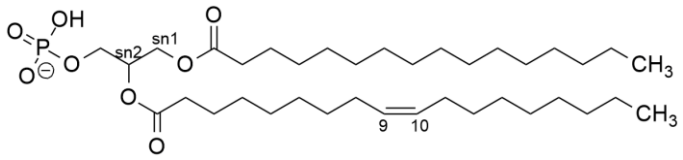
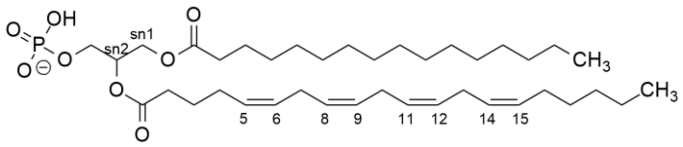
Glycerophospholipids are the major components of cell membranes. Their polar head consists of a base (*e.g.*, choline), a phosphate, and a glycerol. The combination of the two fatty acid chains results in the chemical diversity of glycerophospholipids. The sn-1 chain is either saturated or monounsaturated, while the sn-2 chain is usually mono- or polyunsaturated. There is a difference in side chain lengths in the distinct glycerophospholipids, but their average length is around 3.0-3.5 nm. The linkage of the sn-1 chain can be ester, ether, or vinyl-ether. Various types of head group bases induce a wider glycerophospholipid family (Fig. 4a). The most common phospholipids are listed in Table 3. The type of the headgroup influences the charge of the lipid that can have an impact on the transport processes across the cell membrane, for example [1], [10], [11].

**Table 3. Common phospholipids in an averaged cell membrane [12].**
Lipids were drawn by the ChemDraw® software.

| Abbrevia-tion | Chemical name | sn1 chain / sn2 chain | 2D structures |
|---|---|---|---|
| POPC | 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine | C16:0 / C18:1 |  |
| DOPC | 1,2-dioleoyl-*sn*-glycero-3-phosphatidylcholine | C18:1 / C18:1 |  |
| DPPC | 1,2-dipalmitoyl-*sn*-glycero-3-phosphocholine | C16:0 / C16:0 |  |
| PAPC | 1-palmitoyl-2-arachidonoyl-sn-glycero-3-phosphocholine | C16:0 / C20:4 |  |

| | | | |
|---|---|---|---|
| DAPC | 1,2-diarachidoyl-sn-glycero-3-phosphocholine | C20:4<br>C20:4 |  |
| PUPC | 1-palmitoyl-2-docosahexaenoyl-sn-glycero-3-phosphocholine | C16:0<br>C22:6 |  |
| POPE | 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine | C16:0<br>C18:1 |  |
| DOPE | 1,2-dioleoyl-*sn*-glycero-phosphoethanolamine | C18:1<br>C18:1 |  |
| PAPE | 1-palmitoyl-2-arachidonoyl-sn-glycero-3-phosphoethanolamine | C16:0<br>C20:4 |  |
| DAPE | 1,2-diarachidoyl-sn-glycero-3-phosphoethanolamine | C20:4<br>C20:4 |  |

| | | | |
|---|---|---|---|
| PUPE | 1-palmitoyl-2-docosahexaenoyl- sn-glycero-3-phosphoethanolamine | C16:0<br>C22:6 |  |
| DUPE | 1,2-didocosahexaenoyl-sn-glycero-3-phosphoethanolamine | C22:6<br>C22:6 |  |
| POPS | 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phospho-L-serine | C16:0<br>C18:1 |  |
| PAPS | 1-palmitoyl-2-arachidonoyl-*sn*-glycero-3-phospho-L-serine | C16:0<br>C20:4 |  |
| DUPS | 1,2-didocosahexaenoyl-*sn*-glycero-3-phospho-L-serine | C22:6<br>C22:6 |  |
| POPI | 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphoinositol | C16:0<br>C18:1 |  |

| | | | |
|---|---|---|---|
| POPA | 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphate | C16:0<br>C18:1 |  |
| PAPA | 1-palmitoyl-2-arachidonoyl-sn-glycero-3-phosphate | C16:0<br>C20:4 |  |

### I.2.3.3. Sphingolipids

Sphingolipids have a sphingosine group which is similar to the sn-1 chain and glycerol backbone in glycerophospholipids (Fig. 3a-b). This sphingosine has a long acyl chain with an amino- and two hydroxyl groups (*e.g.*, dihydrosphingosine (DHS), sphingosine (SPH), and phytosphingosine (PHS), see Fig. 3b) [14]–[16]. Sphingolipids vary in length, headgroup, and the number of double bonds and hydroxyl groups of their fatty acid chain. In sphingolipids, the fatty acid chains are more often saturated and longer than in phospholipids. Ceramides are precursors of complex sphingolipids, such as sphingomyelin and glycosphingolipids. Ceramides are composed of a sphingosine (serine and fatty acid) to which another fatty acid chain is bound (Fig. 3b). Sphingomyelin is formed by binding a phosphocholine head group to the ceramide (Fig. 3b) [1], [10], [11].

The attachment of oligosaccharide chains to a ceramide results in a glycosphingolipid (Fig. 3d). The oligosaccharide chains can vary leading to a broad range of functions depending on the bond locations in glycosphingolipids [1], [10], [11]. For instance, the oligosaccharide chains determine the various blood types (A, B, AB, and O) in erythrocytes [11]. Altogether with membrane glycoproteins, oligosaccharide chains form the so-called glycocalyx or cell coat. It has a role in cell-recognition processes (*e.g.*, in cell-cell adhesion) through which they sometimes are involved in the entry of certain bacterial toxins or viruses [1], [10], [11].

### I.2.3.4. Cholesterol

Cholesterol is the primary mammalian sterol. The plasma membrane can contain up to one cholesterol for one phospholipid. The exact ratio depends on the cell type and organelles. The structure of cholesterol differs from the structure of lipids. It has a polar hydroxyl group at the C-3 position, followed by a rigid steroid ring bearing a short non-polar aliphatic tail (Fig. 3c). The hydroxyl group is oriented toward the water phase as the polar head of glycerophospholipids and sphingolipids. Due to either its biophysical properties on the membrane structure or interactions with membrane proteins, many biological processes require the presence of cholesterol. Furthermore, it often acts as a functional partner with sphingolipids through H-bonds (*e.g.*, stabilizing nanodomains) [1], [10], [11].

### I.2.4. Synthetic phospholipids and drug delivery systems

Some other amphiphilic molecules, that contain both hydrophilic and hydrophobic blocks, can self-assemble into many different types of structures, such as micelles, vesicles, planar bilayers, nanotubes, or nanofibers. Therefore, these types of molecules are in the focus of pharmaceutical and industrial applications. Phospholipid-porphyrin conjugates are versatile in this field. These phospholipid-porphyrin conjugates are derived from phosphatidylcholine in which the sn-1 and -2 chains can be substituted by a porphyrin. Binding between phosphatidylcholine and porphyrin moieties is ensured by a chemical linker for which length can differ, leading to different self-assemble membrane structures (see Chapter VI.) [17].

For example, phospholipid-porphyrin conjugates can be used as photoactivatable DDSs. DDSs (formulations or devices) aim to control the rate, time, and place of the release of drugs. This field has greatly developed over the last decades. Targeted DDSs specifically reach their target cells, organs, or tissues to *e.g.*, reduce adverse effects, and/or improve the therapeutic effect. To achieve this, drugs should be packed into delivery vehicles, such as micelles, liposomes, or nanoparticles formed by phospholipid-porphyrin conjugates, for example [6], [7], [17], [18].

### I.2.5. Membrane structures

#### I.2.5.1. Asymmetry and lipid dynamics

Membrane asymmetry is necessary for the life of cells. Lipid bilayers are two-dimensional fluids in which lipid molecules can freely move even though most of the lipids rarely relocate from one leaflet to the other. However, spontaneous flip-flop events were described to occur faster in case of cholesterol than other lipids [1], [11]. Special membrane proteins (namely, flippases, floppases and scramblases) facilitate the flip-flop of lipids (Fig. 5). Flippases move lipids (mostly phosphatidylserine and phosphatidylethanolamine) from the outer leaflet to the inner leaflet while floppases carry lipids (cholesterol, sphingolipid, and phosphatidylcholine) in the opposite direction. Both flippases and floppases are highly specific and ATP-dependent; thus, they can carry lipids against the concentration gradient. Scramblases are usually less substrate specific and ATP-independent; thus, they facilitate lipid translocation along the concentration gradient in both directions (to the outer or to the inner leaflet). Flippase and floppase activities are responsible for the asymmetry of eukaryotic membranes (see table 4 for a proposed realistic membrane composition). In general, the outer leaflet consists of more saturated lipids while the inner leaflet contains more unsaturated and charged lipids. The asymmetry is striking for glycolipids as they are always in the outer leaflet. The presence of certain lipid compositions on the surface is a specific signal. One example is the appearance of phosphatidylserine on the external cell surface which is a signal for apoptosis (*i.e.*, programmed cell death). In this context, scramblases guide the fast presentation of phosphatidylserine lipids on the surface [1], [11], [19].

Besides, asymmetry in the two leaflets of a mammalian membrane, *i.e.*, segregation of the lipids and proteins is observable in each leaflet. Contrary to translocation, lipid molecules diffuse rapidly laterally and their rotation around their long axis is fast, as well. Proteins and lipids, that work together to accomplish their function, often form specialized domains that are frequently enriched in *e.g.*, sphingolipids and cholesterol. Large-scale segregation in living cell membranes is rare, the segregation is highly dynamic thanks to the lateral diffusion [1].

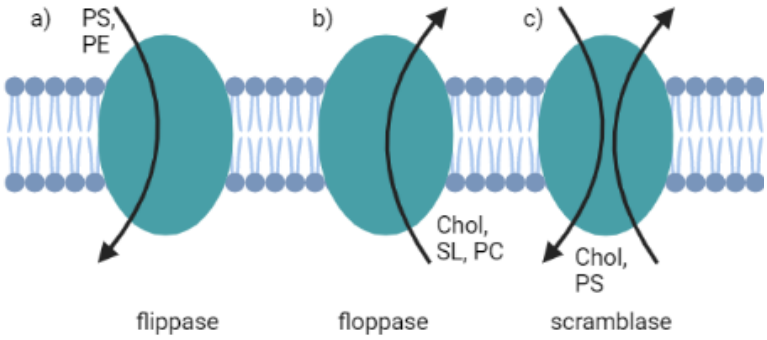



**Figure 5. Flip-flop mechanisms.**
**a)** Flippases, **b)** floppases, and **c)** scramblases. [Based on [19], created with BioRender.com.]

**Table 4. Realistic (asymmetric) mammalian plasma membrane composition used in our group.**
[Based on the work of Ingólfsson *et al.* [Table S2 [12].]

| | | Inner leaflet count | Inner leaflet mole fraction | Outer leaflet count | Outer leaflet mole fraction | Overall mole fraction |
|---|---|---|---|---|---|---|
| Phosphatidyl-choline (PC) | POPC | 550 | 0.059 | 1205 | 0.122 | 0.091 |
| | DOPC | 49 | 0.005 | 106 | 0.011 | 0.008 |
| | PAPC | 129 | 0.014 | 283 | 0.029 | 0.021 |
| | DAPC | 16 | 0.002 | 35 | 0.004 | 0.003 |
| | PUPC | 32 | 0.003 | 71 | 0.007 | 0.005 |
| Phosphatidyl-ethanolamine (PE) | POPE | 569 | 0.061 | 135 | 0.014 | 0.037 |
| | DOPE | 190 | 0.020 | 44 | 0.004 | 0.012 |
| | PAPE | 522 | 0.056 | 124 | 0.013 | 0.034 |
| | DAPE | 332 | 0.036 | 78 | 0.008 | 0.021 |
| | PUPE | 190 | 0.020 | 44 | 0.004 | 0.012 |
| | DUPE | 95 | 0.010 | 22 | 0.002 | 0.006 |
| Phospha-tidylserine (PS) | POPS | 200 | 0.021 | 0 | 0.000 | 0.010 |
| | PAPS | 461 | 0.049 | 0 | 0.000 | 0.024 |
| | DUPS | 20 | 0.002 | 0 | 0.000 | 0.001 |
| Phospha-tidic acid (PA) | POPA | 46 | 0.005 | 0 | 0.000 | 0.002 |
| | PAPA | 39 | 0.004 | 0 | 0.000 | 0.002 |
| | **Total** | **3440** | **0.367** | **2147** | **0.218** | **0.290399709** |

Lateral diffusion can be calculated from the mean-squared displacement (MSD) (Eq. 1). The diffusion in a liquid is a random motion of particles in a medium, called Brownian motion. To a given extent, it can be approximated by the Einstein relation (Eq. 2), as well. Einstein relation describes the connection between fluctuations caused by diffusion and dissipation expressed by a finite mobility [1], [20]. Lateral diffusion is typical not only for the lipids but for the membrane proteins, as well.

$$\boldsymbol{D_l = \frac{1}{2n} \lim_{t\to\infty} MSD(t)}$$

$$MSD = \langle x(t) - x(0) \rangle^2 = \frac{1}{N} \sum_{n=1}^{N} \big(x_n(t) - x_n(0)\big),$$

$x_n(0)$ – reference point for particle $n$

**Equation 1. Lateral diffusion coefficient at a given time for all particles (N) [21].**

$$\boldsymbol{D = k_B T \mu}$$

$\mu$ – mobility, $k_B T$ – thermal energy

**Equation 2. Einstein relation for the diffusion coefficient [20].**

### I.2.5.2. Thickness

The thickness of a membrane strongly depends on the lipid composition. A typical total thickness is around 5 nm. However, pure membranes including only unsaturated chains are slightly thinner than those built by saturated tails containing the same number of carbon atoms. The thickness is strongly connected to the lipid order, as well. Ordered (gel) lipids are straight packed and thicker than disordered (fluid) lipids. In other words, structural parameters that influence the fluidity (see next section) have a strong impact on membrane thickness, as well. Importantly, membrane thickness values strongly depend on how it is measured or calculated [1], [11], [22], [23]; therefore, several thickness definitions exist.

- $D_B$ is the overall bilayer thickness, also called Luzzati thickness, *i.e.*, the distance between the locations where the water density drops half its bulk value.
- $D_{HH}$ is the head-to-head distance, *i.e.*, the distance between the two peaks in the electron density profile. It can be approximated by $D_{PP}$.
- $D_{PP}$ is the phosphate-phosphate distance, *i.e.*, the distance between the maximal phosphate group densities.
- $2D_C$ is the hydrocarbon chain thickness.

When calculated from simulations, all these thicknesses can be directly or indirectly compared to experimentally observed values [22], [23].

### I.2.5.3. Fluidity, order of the membrane, lipid phase

Lipid membranes can be featured by a characteristic temperature at which the liquid-to-gel phase transition occurs. The so-called phase transition temperature varies upon the lipid composition of the membrane. Indeed, at a given temperature, saturated lipids favour more compact, more ordered and less fluid membrane whereas unsaturated lipids have the opposite

impact. The fluidity is linked to the lipid order parameter that can be calculated either from NMR experiments or from simulations (by Eq. 3) [1], [11], [22].

$$S_{CD} = \frac{1}{2}\langle 3cos^2\theta - 1\rangle \in [-0.5,1]$$

$\theta$ – angle between a C-H bond of the lipid tail and the normal vector of the membrane (Fig. 6)

**Equation 3. Lipid order.**

$S = 1$ means lipid chains are perfectly aligned to the membrane normal,
$S = -0.5$ means anti-alignment, $S = 0$ can mean perfect order and disorder, as well [22].

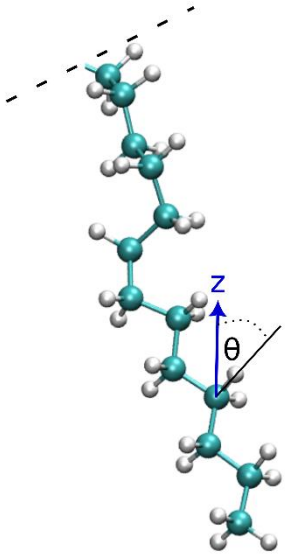



**Figure 6. Definition of the $\theta$ angle in $S_{CD}$ calculation.**
Z is the normal of the lipid bilayer.

The order of a membrane and its fluidity are strongly connected (Fig. 7). The liquid-disordered ($L_\alpha$), also called liquid-crystalline phase, is fluid and the lipids have high lateral mobility. The gel phase ($L_\beta$) is highly ordered and the lipids have much less lateral mobility. The environment and the proportion of cholesterol also affect the fluidity of a lipid bilayer. The effect of the cholesterol depends on the original order and phase of the lipid bilayer. Cholesterol decreases the lipid order in the gel phase while increases it in the fluid phase. Cholesterol can shift the liquid-disordered and the gel phase into an intermediate phase, called liquid-ordered ($L_o$) [24], [25].

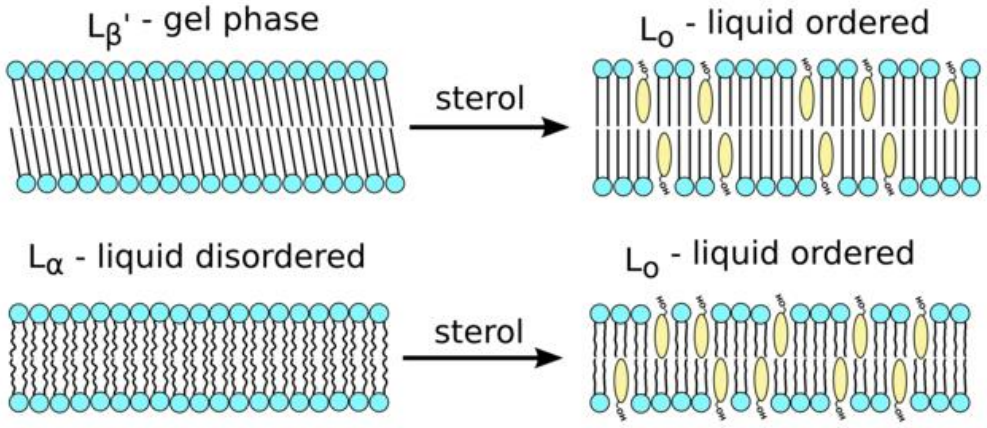



**Figure 7. Different types of lipid phases.**
[Figure 3.1.3. from [26].]

### I.2.5.4. Curvature, lipid shape and lipid packing

As many membrane properties, membrane curvature depends on the saturation of the lipid tails and the head group of the lipids. Membrane curvature determines the shape of the whole compartment such as cells. Structural lipid shapes are of particular importance: cone-shaped lipids form micelles, while cylinder-shaped lipids, such as phospholipids, form bilayers. Lipids with a large or small head group on the surface induce positive or negative curvature, respectively. Positive and negative curvatures result in micelles and inverted micelles, respectively (Fig. 8). Membrane asymmetry (section I.2.6.1) plays an important role in local curvature of a cell membrane, as the cone- and cylinder-shaped lipids are not equally distributed. Furthermore, thanks to the diffusion (section I.3.3.2-3), this local curvature may be dynamic if no proteins are involved in maintaining the supramolecular assembly [1], [17], [27].

The shape of a lipid indicates its proper packing. Thus, for pure lipid bilayers, the curvature can be predicted from the packing parameter which is the quotient of the molecular volume of the hydrophobic part of the lipids ($v$) and the product of the cross-section surface area of the head group ($a$) and the length of the fatty acid chain in their all trans conformation ($l$) [17], [27].

$$\boldsymbol{P = \frac{v}{al}}$$

**Equation 4. Packing parameter [17], [27].**

The structure of a membrane can be predicted:

- $P < \frac{1}{3}$ for micelles
- $P = 1$ for bilayers
- $P > 1$ for inverted micelles [1], [27]

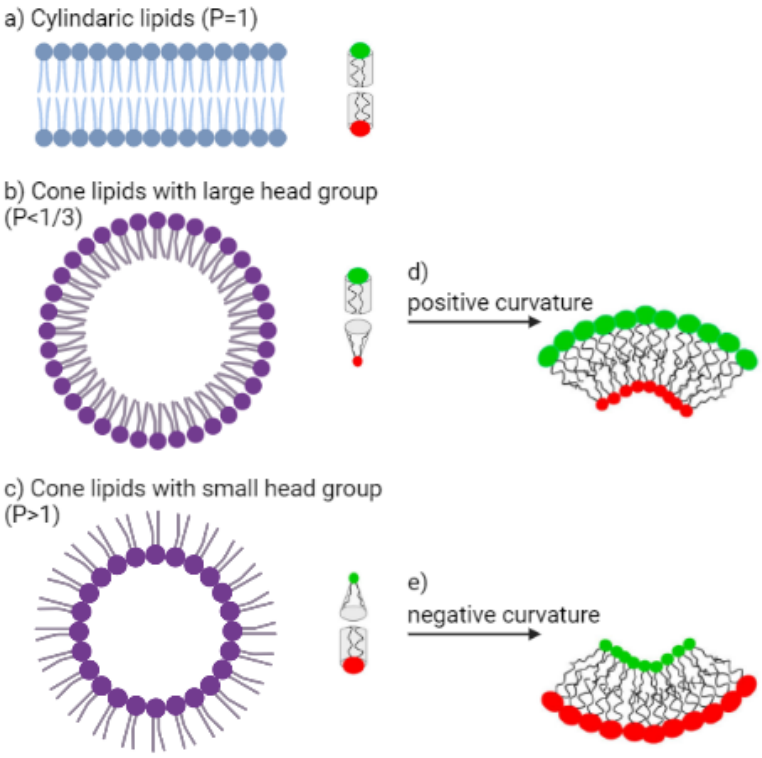



**Figure 8. Lipid bilayer curvatures and structures.**
**a)** Bilayer, **b)** micelles, **c)** inverted micelles, **d)** positive and **e)** negative curvature.
[Based on Figure 1 from [27], created with BioRender.com.]

### I.2.5.5. Area per lipid

Another parameter widely used in membrane biophysics is the area per lipid (APL) which is the average cross-section area occupied by a lipid polar head in the bilayer. It also depends on the lipid composition, and the saturation, so it is tightly connected to the order parameter and fluidity. APL of a pure bilayer can be easily calculated from simulations as the x-y area of the model divided by the number of lipids in one leaflet. It assumes that the number of lipids is symmetric in the two leaflets [22]. Realistic mammalian membranes are asymmetric and include proteins, leading to a more challenging assessment of APL. A better approximation can be the quotient of the volume of the bilayer and the overall bilayer thickness. However, the proper volume calculation can be complicated in presence of proteins [28].

$$\boldsymbol{A_L = \frac{L_x L_y}{n} = \frac{V_{bilayer}}{D_B}}$$

$n = \frac{No.of\ lipids}{2}$, $L_x L_y$ – x and y length of the bilayer
$V_{bilayer}$ – volume of the bilayer, $D_B$ – overall bilayer thickness

**Equation 5. Area per lipid of a pure bilayer [22].**

## I.3. Different types of transport across a cell membrane

### I.3.1. Overview

Cell membranes are more or less permeable for molecules. The transport mechanism of molecules strongly depends on their physical and chemical properties. There exist spontaneous and protein mediated transports. Depending on the transport mechanism, the latter can be divided into three sub-categories: facilitated diffusion, primary and secondary (or even tertiary) active transport [1], [11].

Hydrophobic molecules (*e.g.*, $O_2$, $CO_2$, $N_2$, and steroid hormones) can passively cross the membrane as well as small uncharged polar molecules (*e.g.*, $H_2O$, urea, glycerol, and $NH_3$) to a lesser extent. Spontaneous membrane crossing for large uncharged polar molecules (*e.g.*, glucose and sucrose) is kinetically more challenging and almost unlikely. Finally, ions (*e.g.*, $Na^+$, $HCO_3^-$, $K^+$, $Ca^{2+}$, $Cl^-$, and $Mg^{2+}$) cannot cross the membrane spontaneously. Membrane proteins help the transport of both large uncharged polar molecules and ions [1], [11].

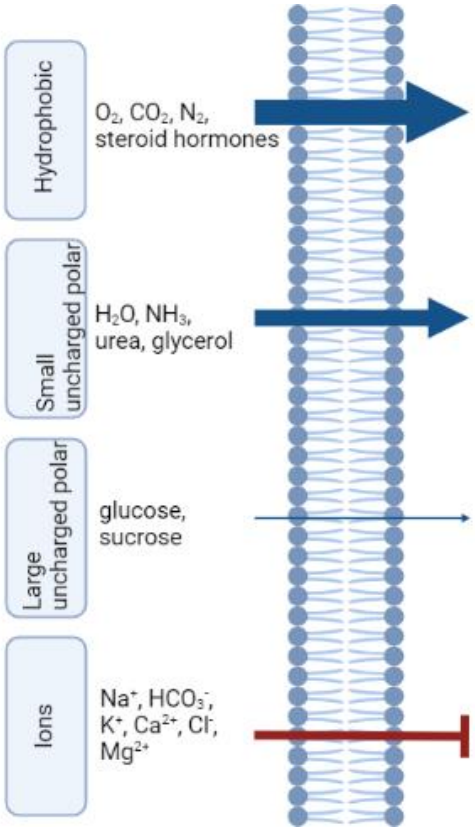



**Figure 9. Plasma membrane permeability of different types of molecules.**
[Based on Figure 11-1 from [1], Created with BioRender.com.]

### I.3.2. Passive permeation

Passive permeation, earlier called diffusion, follows the concentration gradient; thus, the membrane crossing is passive. The rate of a small molecule's diffusion in a solution can be described by the Fick's law (Eq. 6), *i.e.*, the molecular change of gas molecules over time. In case of passive membrane permeation, the rate is directly proportional to the surface area and the concentration gradient along the membrane, while inversely proportional to the membrane

thickness. Furthermore, it depends on the diffusion coefficient that is different according to the size, physical, and chemical properties of the diffused molecule [1], [11].

$$\boldsymbol{V_x}\left[\frac{mol}{t}\right] = \frac{\Delta m}{\Delta t} = \frac{\boldsymbol{DA\Delta C}}{\boldsymbol{\Delta x}}$$

$\Delta t$ – time difference, $\Delta m$ –mole difference,
$V_x$ – speed of the diffusion, $D$ – diffusion constant, $\Delta C$ – concentration difference,
$A$ – surface, $\Delta x$ – membrane thickness

**Equation 6. Fick's law [1], [11].**

However, a membrane is not a homogeneous solvent. Therefore, inhomogeneous solubility-diffusion models were developed, by dividing a membrane into several "slices" with different biophysical properties. Nevertheless, these models are still limited, *e.g.*, for small molecules. Recently, more advanced approaches based on *e.g.*, Markov state models were proposed to describe the kinetics of drug passive permeation events. All models exhibit limitations as the rate limiting step depends on the type of the diffused molecule. However, advanced models can be used to build a structure-kinetic relationship of drug membrane permeation. This allows the assessment of the permeability coefficient ($P$) of a molecule in a given membrane. It depends on the membrane composition and the diffused molecule. It is experimentally measurable, as well (Fig. 9). However, both computer and experimental approaches suffer from user-dependent biases and model limitations precluding their use for systematic large screening [1], [11], [29].

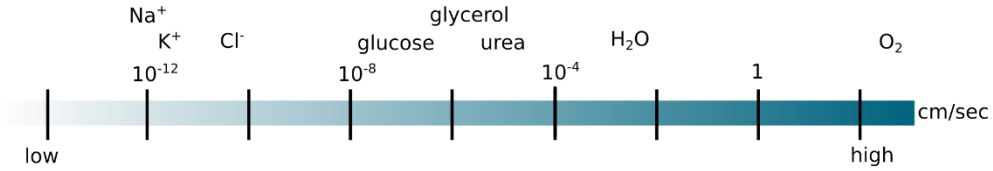



**Figure 10. The permeability coefficient of various molecules.**
[Adapted from Figure11-2 of [1].]

### I.3.3. Facilitated diffusion

Passive permeation was suggested to be either an unlikely or minor event for most of the xenobiotics used nowadays. Molecules that cannot cross the membrane spontaneously can undergo facilitated diffusion. The transport still happens following the concentration gradient; however, its rate is accelerated thanks to carriers (*e.g.*, glucose transporters) or channels (*e.g.*, aquaporins). Nevertheless, transport is saturable (Fig. 11) for carriers. Depending on the concentration gradient, carriers always open up only for one side of a cell membrane at a time, alternating between two states around $10^3$ times in a second. Contrary to carriers, an open channel can be open on both sides of the membrane at the same time. Without the alternation mechanism, they facilitate the transport across a membrane faster than carriers (around $10^{6\text{-}7}$ times in a second). All channels and carriers demonstrate high specificity for their substrate. They can carry only one molecule at the same time (uniporter) [1], [11].

The maximum rate of carrier-mediated transport is connected to their affinity ($K_M$ value in Fig. 11) for their substrates. For example, GLUTs translocate glucose and other

monosaccharides with distinct affinities. Glucose affinities also vary for the different types of GLUTs. Interestingly, monosaccharides can exhibit larger affinities for GLUTs without triggering the opening, leading to the inhibition of the transporter [1], [11].

All cells in a human body have one or more distinct GLUT transporters whose affinities depend on the role of glucose in the cell. High affinity means that the change in glucose level in plasma does not influence the glucose uptake of the cells (*e.g.*, liver, kidneys). Other cells (*e.g.*, brain and placenta) have lower affinity and thus are more sensitive to glucose concentration. In this case, cell glucose uptake is limited in order to take the necessary glucose concentration for proper functioning [1], [11].

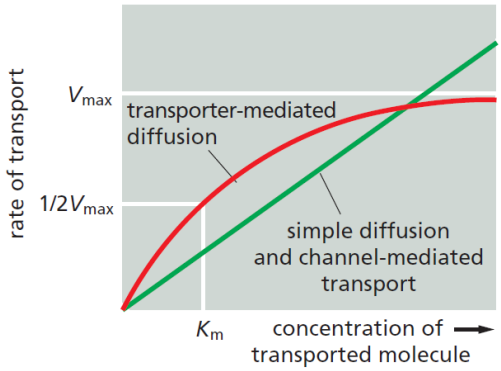



**Figure 11. Kinetics of diffusion.**
[Figure 11-6 from [1].]

### I.3.4. Primary active transport

Active transport requires energy for its function, as substrate transport occurs against the concentration gradient. ATP-hydrolysis drives the transport of ATP-driven pumps (such as $Na^+/K^+$-ATPase, $Ca^{2+}$-ATPase, $H^+/K^+$-ATPase, and $H^+$-ATPase), and ATP-binding cassette (ABC) transporters (details in section I.4). A typical example of primary active transport is the activity of $Na^+/K^+$-ATPase. During its transport cycle, $Na^+/K^+$-ATPase transports out three $Na^+$ cations against two $K^+$ cations which are imported into the cytoplasmic compartment. The transport cycle of $Na^+/K^+$-ATPase can be specifically inhibited by heart glycosides (organic steroid derivatives). $Na^+/K^+$-ATPase maintains the high $K^+$ intracellular concentration in many cells such as erythrocytes or kidney tubular proximal cells. It also participates to maintain the high extracellular $Na^+$ concentration. $Na^+/K^+$-ATPase can be found in all plasma membranes given its central role in maintaining the resting membrane potential. There exist other ATP-dependent exchangers of which membrane expression depends on the cell type. For example, the $H^+/K^+$-ATPases and the $H^+$-ATPases are observed in kidneys, playing an important role in body homeostasis [1], [11].

### I.3.5. Secondary (tertiary) active transport

In case of secondary active transport, the electrochemical gradient of a compound provides the energy required for the transport of another molecule. For instance, the $Na^+$ gradient maintained by the $Na^+/K^+$-ATPase can cover the energy required to ion coupled co- and anti-transporters. In case of co-transport (*e.g.*, $Na^+$-glucose), the two transported molecules are translocated in the same direction, *e.g.*, from the cytoplasm to the extracellular compartment in case of SGLT1. Contrary, the direction of the transport is different for the two transported molecules in antiporters. OAT1 is an example of tertiary active transport (Fig. 12)

as it uses the gradient created by $Na^+$/dicarboxylate cotransporter which uses itself the $Na^+$ gradient from $Na^+/K^+$-ATPase. In this cascade, the $Na^+$/dicarboxylate is a secondary active transporter [11], [30]–[32].

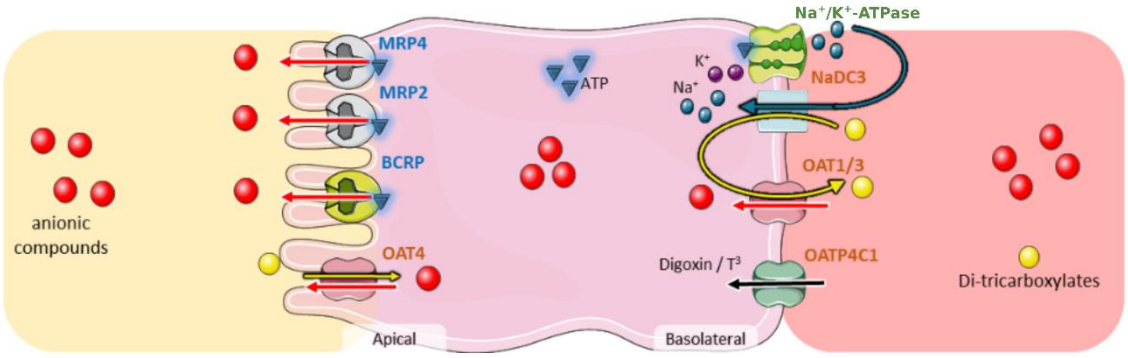



**Figure 12. Secondary (NaDC3) and tertiary (OAT1/3) active transport mechanisms.**
[Slightly adapted Figure 1A. from [31].]

## I.4. Transporters in pharmacology

### I.4.1. Overview

Membrane proteins are important in various aspects along the drug's journey in the body. Transporters facilitate the uptake or elimination of a broad range of substrates. Their primary role is to transport nutrients and endogenous substrates, but they can also protect the body from toxins. Transporters can translocate a wide range of molecules, including xenobiotics that are structurally similar to their physiological substrates. Besides, drug transporters are mainly expressed in tissues with barrier functions *e.g.*, liver and kidneys. These tissues often adopt an epithelial architecture in which cells are often polarized, as they are at the interface between distinct fluids (*e.g.*, bile/blood or urine/blood, respectively for hepatocytes or tubular proximal cells). Drug transporters can be grouped depending on the direction of the xenobiotic transport. Influx transporters are involved in the uptake of the molecules into cells, while efflux transporters are responsible for their elimination. Consequently, often, cellular expressions of the different types of drug transport differ between polarized cell membranes. Drug membrane transporters belong to either the Solute Carrier (SLC) or ATP-Binding Cassette (ABC) superfamilies. Members of these superfamilies have been reported as pharmacologically important by the International Transporter Consortium (ITC) as these proteins play a central role in drug disposition [33], [34]. Membrane transporters located in liver and kidney are crucial since these organs are involved in most of the drug elimination events (*ca.* 95%) [9].

### I.4.2. Solute carrier superfamily

The SLC superfamily is ubiquitous, it includes more than 400 members divided into more than 65 families for which transport cycles can occur under different mechanisms. Some of them benefit from the electrochemical potential difference of their substrates (facilitated transport), while others transport substrates against an electrochemical gradient using an ion gradient produced by a primary active transporter (secondary active transport). SLC transporters are mostly responsible for the uptake of drugs from the blood flow even though few members were described to extrude their substrates. Below the role of SLC transporters relevant in drug disposition namely, organic cation and anion transporters (OCTs, OATs), multidrug and toxin

extrusion transporters (MATEs), and organic anion transporting polypeptides (OATPs), is described [5], [9], [35].

OCTs and OATs belong to the SLC22 family. They are composed of 12 α-helical transmembrane helices (TMHs) divided into two bundles connected by a large intracellular loop between transmembrane domains (TMDs) 6 and 7. OCTs are uniporters, and they mediate the uptake of small positively charged substrates including both endogenous compounds and xenobiotics. Their transport mechanism occurs by facilitated transport *i.e.*, governed by the concentration gradient. Their expression level can be associated with diseases. For example, *h*OCT1 is overexpressed in chronic myeloid leukaemia cell lines, and its expression is reduced in liver injuries [5], [9], [35].

OATs are antiporters; thus, they carry small anions, both endogenous compounds and xenobiotics. Their transport mechanism occurs under tertiary active transport: substrate uptake is ensured by the anti-transport of endogenous dianions (*e.g.*, α-ketoglutarate). High intracellular concentrations of dianions are governed by the function of NADC3 which relies on $Na^+/K^+$-ATPase (Fig. 12). OAT1 and OAT3 are particularly important in drug disposition. They are mostly expressed in kidney cells participating in drug cellular uptake from blood prior to their elimination into the urine. Interestingly, their substrate specificity is wide and partially overlapping [5], [9], [30], [32], [35].

OATPs belong to the SLCO or SLC21 family. They transport endogenous and exogenous amphipathic molecules. Their transport mechanism remains unclear. Most of the OATPs are ubiquitously expressed; however, some of them have tissue-specific expression. For example, OATP1B1 and OATP1B3 are liver-specific, while OATP4C1 is kidney-specific. OATPs also show disease specific changes in their expression level, *e.g.*, OATP1B3 is decreased in hepatocellular carcinoma [5], [9], [35].

MATEs are also important in cationic xenobiotic transport as OCTs. Contrary to most of the SLCs, MATE is responsible for drug efflux. It transports a wide range of substrates using proton co-transport. Many of its substrates overlap with OCT substrates. MATE is principally expressed in the liver and kidneys; thus, it is involved in clearing the plasma [5], [9], [35].

### I.4.3. ATP-binding cassette proteins

#### I.4.3.1. A general description

ABC transporters use the energy from ATP-hydrolysis during their transport cycle. ABC proteins transport endogenous substrates, but some of them are important in drug transport, as well. ABC proteins can be found in both prokaryotes and eukaryotes. Among the prokaryotic ABC transporters, both efflux and influx proteins exist. Mammalian ABC transporters, except ABCA4 and ABCD4, are efflux proteins.

ABC transporters share structural similarities. They are made of at least two transmembrane domains (TMDs) and two nucleotide binding domains (NBDs). Eukaryotic ABC proteins are divided into seven sub-families (ABCA-ABCG) from which the members of the ABCE and ABCF sub-families lack transmembrane domains. Therefore, they cannot act as a transporter or channel. ABCE and ABCF proteins are bound to their target molecules or other proteins, and they play a role in ribosome function and regulation [36]–[40]. The other five sub-families (ABCA-ABCD, ABCG) are transporters or involve in substrate membrane crossing events (Fig. 13a). Members of ABCB, ABCC and ABCG families play a role in drug transport, and some of them were reported as emerging clinical importance by the ITC [2], [5], [9], [33], [34].

During many years, ABC proteins were classified into type I and type II. Recently, a new classification was proposed based on the TMD sequence homology and structures. This recent classification results in seven folds/types for which different physiological functions are observable (Table 5). In bacteria, all types of ABC proteins are presented, while in eukaryotes, only type IV-V proteins (mainly exporters) can be found [38]. The focus of this thesis is on eukaryotic ABC transporters (type IV and V), as they are involved in many physiological and pharmacological processes and their dysfunction cause various diseases.

**Table 5. New classification of ABC transporters.**
[Based on Table 1 [38].]

| Folds | Function | Examples | |
|---|---|---|---|
| | | Prokaryotic | Eukaryotic |
| Type I | Importer | $MalFGK_2$ | - |
| Type II | Importer | $BtuC_2D_2$ | - |
| Type III | Importer | Folate importer | - |
| Type IV | Exporter and importer | Sav1866, MsbA, TmrAB | P-gp, MRP1, CFTR |
| Type V | Exporter and importer | Wzm-Wzt | ABCG2 |
| Type VI | Extractor | $LptB_2FG$ | - |
| Type VII | Mechanotransmitter | MacB | - |

### I.4.3.2. Eukaryotic ABC transporters

Eukaryotic ABC transporters contain six transmembrane helices (TMHs) in each TMDs. TMDs and NBDs can be coded as a full transporter (*e.g.*, type IV – ABCB, ABCC, ABCD) or by subunits (type V – ABCA, ABCG) resulting in either homo- or heterodimer (Fig. 13a-b). The TMHs are longer in full (type IV) transporters, and they cross each other, while dimers (type V transporters) are symmetrical (Fig. 13b). NBDs are evolutionarily conserved in both types. ABC transporter can jointly bind two ATP molecules and two magnesium in two nucleotide binding sites (NBSs) which are located at the interface of the two NBDs. ATP molecules and magnesium cations are bound to conserved motifs (Fig. 15b). These conserved motifs are also involved in ATP-hydrolysis that provides the energy for the transport cycle. Each NBS is made of the Walker A and B motifs, A-, Q- and H-loops of one NBD and the signature motif and X-loop of the other NBD (Fig. 15b). An aromatic residue (often a tyrosine) in the A-loop binds the ATP by forming $\pi$–stacking interactions between aromatic rings. The D-loops are responsible for the stabilization of the ATP-binding and the binding of the two NBDs together. The glutamic acid in Walker B, the glutamine in Q-loop, the H-loop, and the signature motif interact with the ATP. Moreover, these motifs may have an essential role in ATP-hydrolysis. The Q-loop was also suggested to participate in the signal transmission from NBDs to TMDs. The intracellular helices between the TMHs adopting a “ball-and-socket” conformation with NBDs are called coupling helices. They interact with the NBDs helping the information flow during the substrate translocation as well as ATP-hydrolysis [2], [38]–[40].

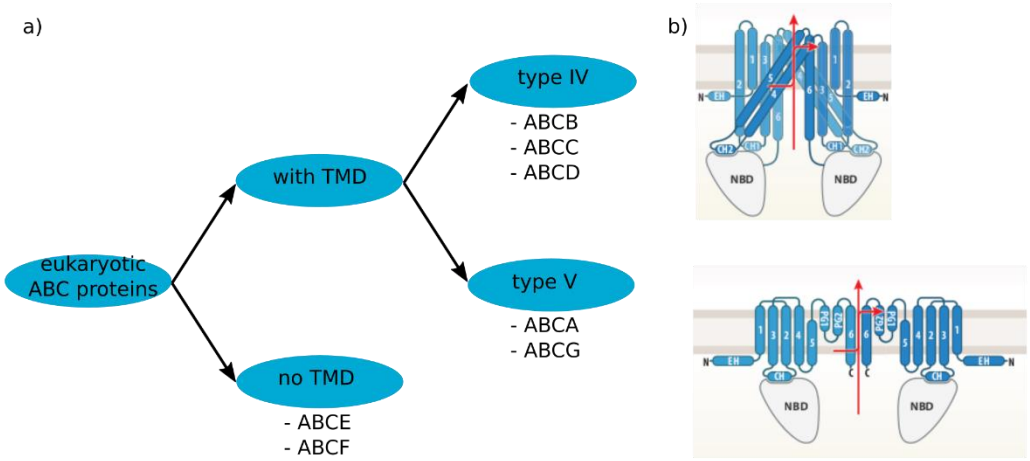



**Figure 13. Eukaryotic ABC proteins.**

**a)** Classification of the eukaryotic ABC proteins. **b)** Schematic TMD topology of type IV and type V transporters are taken from Figure 2 of [38].

The transport cycle of mammalian ABC transporters implies alternation between two (inward-facing – IF and outward-facing – OF) conformations. As most of the eukaryotic ABC proteins are exporters (exceptions are listed in Table 6), they take the drug from the cell in the IF conformation, and release it outside of the cell in the OF conformation (Fig. 15a). Conformational transition is expected to occur via occluded states where the NBDs and the extracellular gate are also closed. The IF-to-OF conformational change is expected to be facilitated by substrate-binding. ATP-hydrolysis, or as it was proposed recently, the phosphate release triggers the conformational changes necessary to go back from OF-to-IF conformation (Fig. 14) [38]–[41].

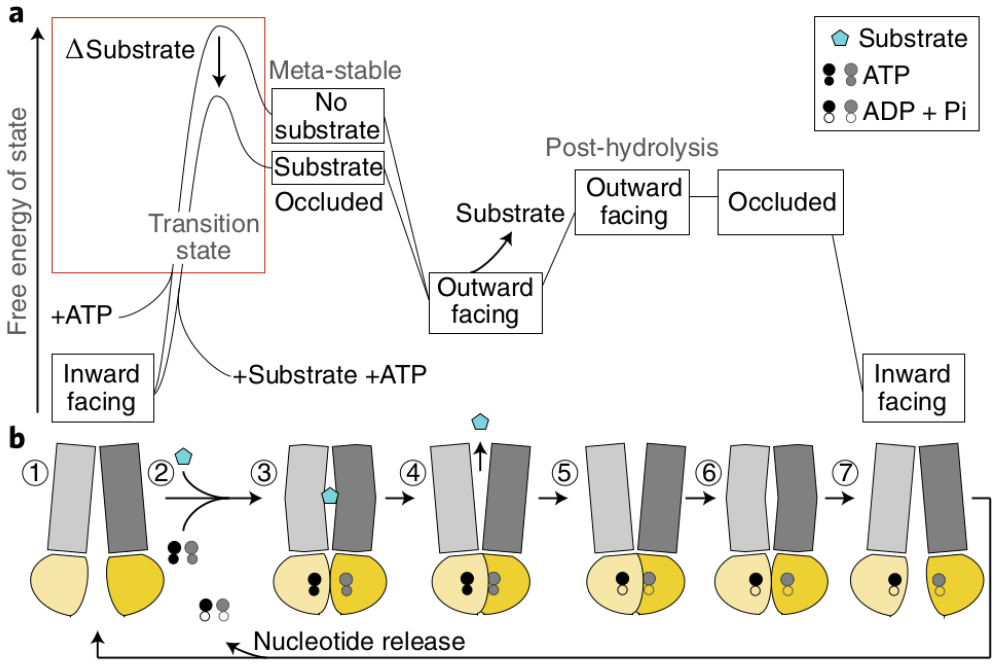



**Figure 14. The thermodynamic model of an ABC transporter.**

[Fig. 3 from [40].]

It was also recently suggested that this transition happens via asymmetric unlock-returned (UR) turnover conformations, where the NBDs open only partially. This partial opening of the NBDs is followed by the opening of the intracellular gate. The proper opening of the NBDs and the ADP-to-ATP exchange is the last step, resetting the transporter for a new transport cycle [38], [41]. However, there are many unanswered questions. For example, it is not clear (i) if the substrate- or the ATP-binding happens first, and (ii) if the NBDs keep interacting along the transport cycle. Constant weak interactions at the NBD interface would mean that the IF wide open conformation may not exist under physiological conditions. In other words, resolved wide open structures would be an artefact from experimental conditions, *e.g.*, when using detergent-purified structures [2], [39]–[42]. Thanks to Cryo-electron microscopy (cryo-EM), it is now possible to catch structures during changes [41]. However, our knowledge about the detailed atomistic mechanism is still fragmented.

NBDs have drawn particular attention. In many ABC proteins, deviations from the consensus sequence were observed for all ABC families but ABCD. Sequence deviation in NBSs results in the preference of ATP-hydrolysis in one NBS, namely in NBS2, which is called canonical site. The slower ATP-hydrolysis in NBS1 also called degenerate site can be explained by differences in the sequence. NBS2 (canonical site) contains important residues, such as a glutamic acid, a tyrosine and a glycine in the Walker B motif, A-loop and signature sequence, respectively. When these residues are mutated, kinetics are modified leading to the existence of a degenerate NBS1. For instance, mutation of the signature sequence is present in all degenerated ABC proteins, while the other mutations are more specific [39].

**Table 6. Non exporter eukaryotic ABC proteins [38], [39].**

<table>
<tr><td colspan="3">Channels</td><td colspan="2">Transporters</td></tr>
<tr><td>ABCC7</td><td>ABCC8</td><td>ABCC9</td><td>ABCA4</td><td>ABCD4</td></tr>
<tr><td>CFTR –<br>Cystic Fibrosis<br>Transmembrane-<br>conductance<br>Regulator</td><td>SUR1</td><td>SUR2</td><td>ABCR</td><td></td></tr>
<tr><td>chloride channel</td><td colspan="2">sulfonylurea receptors regulating the<br>Kir6.1 and Kir6.2 potassium channels</td><td>retina-specific<br>importer</td><td>lysosomal B12<br>vitamin<br>importer</td></tr>
</table>

### I.4.3.3. ABCC family

In many ABCC proteins (*e.g.*, ABCC1 and ABCC4), the glycine of the signature sequence is replaced by valine. The swap of the C-terminal glutamic acid of the Walker B motif into aspartic acid results in less proper interaction with the ATP. It slows down the ATP-hydrolysis rate about 3-fold. In some ABC proteins (*e.g.*, ABCC7 but also in the ABCB11 a.k.a. BSEP member), the C-terminal glutamic acid is replaced by another amino acid (*e.g.*, methionine or serin). In ABCC proteins, the A-loop tyrosine is mostly replaced by tryptophan in NBS1 which may lead to stronger $\pi$–stacking due to the larger overlap between molecular orbitals. Moreover, the interface between NBDs and TMDs may also differ in ABCC proteins. For example, in the NBD1 of ABCC1, there is a 10-13 residue long deletion which prevents the formation of a proper “ball-and-socket" arrangement which may be involved in lower signal transduction between NBD and TMHs. All these changes influence the ATP-binding environment, so most probably the ATP-hydrolysis rate, as well [2], [39], [43].

Most of the ABCC proteins but also two of the ABCB proteins (transporter associated with antigen processing – TAP – proteins) have an extra N-terminal TMD (Table 7). The so-called $TMD_0$ is expected to be involved in protein trafficking (ABCC1), but it may also be responsible for the connection with other functional partner proteins (as in ABCC8, and ABC9). However, it has been suggested that $TMD_0$ is not involved in the substrate transport in case of ABCC1. For the other $TMD_0$ containing proteins, the role of $TMD_0$ is still elusive [2], [44]. Proteins containing this extra domain are called "long", while proteins without $TMD_0$ are called "short" [2], [38], [43], [44].

The members of the ABCC sub-family have a small helix at the N-terminal side of TMD1. The so-called $L_0$ is present independently of the presence of $TMD_0$. $L_0$ was shown to play an important role in the proper folding and function of the protein. $L_0$ is conserved in all members of the ABCC family even though its size may be different. In ABCC1, the fold of $L_0$ residues is unusual: it folds to the membrane-cytosol interface [2], [38], [43], [44]. This remarkable fold may be noticeable in other ABCC proteins.

**Table 7. "Short" (without $TMD_0$) and "long" (with TMD0) MRPs and other TMD0 containing proteins [2].**

| "Short" MRPs | | "Long" MRPs | | Other proteins with TMD0 | |
|---|---|---|---|---|---|
| ABCC4 | MRP4 | ABCC1 | MRP1 | ABCC8 | SUR1 |
| ABCC5 | MRP5 | ABCC2 | MRP2 | ABCC9 | SUR2 |
| ABCC11 | MRP8 | ABCC3 | MRP3 | ABCB2 | TAP1 |
| ABCC12 | MRP9 | ABCC6 | MRP6 | ABCB3 | TAP2 |
| | | ABCC10 | MRP7 | ABCB9 | TAP-like |

### I.4.3.4. Multidrug resistance-associated proteins

Nine of the ABCC family (Table 7) compose a group of proteins, called multidrug resistance-associated proteins (MRPs), as they were originally proposed to play a role in drug resistance in tumour cells. Four MRPs (MRP2, MRP3, MRP4, MRP6) are present in liver and/or kidney cells, and three of them are involved in many pharmacological processes. For instance, MRP2 can be found in the canalicular membrane but also in the apical membrane of hepatocytes. It is also localized to the apical membrane of other polarized cells, such as renal proximal tubule epithelia, or intestinal epithelia. MRP2 transports a wide range of metabolized molecules. Moreover, the improper function of MRP2 can result in inefficient bilirubin elimination, the so-called Dubin and Johnson disease. MRP3 can be mainly found in the basolateral membrane of many cells. However, in hepatocytes, its expression is low. The dysfunction of MRP3 is not connected to any disease, but it is upregulated in cholestasis as it may be the "back-up system" of MRP2. MRP4 can be found in numerous tissues, among them in both membranes of liver hepatocytes, as well as in kidney proximal tubular cells. It also exports a large variety of substrates. It was suggested to act as a back-up partner of MRP2 [2].

### I.4.3.5. ABCC1/MRP1

MRP1 is ubiquitous; however, it is not expressed in liver cells. MRP1 shows appropriate (around 40-60%) sequence similarity to other unresolved MRPs located in the liver. The highest sequence similarity has been observed to MRP3 (58%). Regarding to the expression and substrate specificity, MRP4 is the most similar to MRP1. MRP1 seems to be a good candidate as a prototype for MRPs located in the liver and kidneys, as until the release of

Alpha Fold 2 in 2021, the availability of the resolved structures was limited [45]. Therefore, a common technique to study the dynamical behaviour of a non-resolved but pharmacologically important MRP was homology modelling or protein threading techniques as exemplified by the model of MRP4 proposed in 2018 [46]. MRP1 was used as a prototype for pharmacologically relevant MRPs, such as MRP2 and MRP4.

Cryo-EM structures of bovine MRP1 (*b*MRP1) are available in different conformations, namely in $IF_{apo}$ (PDBID: 5UJ9), IF leukotriene C4 ($LTC_4$) substrate-bound (PDBID: 5UJA), OF ATP-bound (PDBID: 6BHU) and OF ATP-ADP-bound (PDBID: 6UY0) [43], [44], [47]. Moreover, *b*MRP1 has high (91%) sequence similarity to human MRP1 (*h*MRP1) [2], [38], [44].

All the above mentioned cryo-EM structures were resolved with relatively high resolution (3.14 Å – 3.49 Å) obtained using a detergent-purified experimental setup. The resolution for the $TMD_0$ is however lower [44]. The cryo-EM resolved OF structures suggested the importance of cholesterol in the membrane, as protein structures were co-resolved with three tightly bound cholesterols. It was recently acknowledged that lipids might play an important role in allosteric effects by facilitating the information flow, as observed in ABCB1/P-gp [48]–[50]. Changes on one side of the protein, *e.g.*, binding of a ligand, influence a distant site of the protein. Allostery seems to be important in the proper functioning of many membrane proteins [51].

The resolution of TMHs enables the examination of substrate-binding. ABCC1 bound to $LTC_4$ was resolved exhibiting the substrate structure 10 Å deep from the cytoplasm. Substrate binding was assigned to a strong H-bond network as well as van der Waals interactions with the TMDs, leading to two distinct binding pockets. The P-pocket contains positively charged residues, and it is connected to the glutathione moiety of $LTC_4$. The H-pocket is hydrophobic, formed by residues only from one side of the binding site. It interacts with the lipid tail of $LTC_4$ (Fig. 15a). Many of the residues, forming H-bonds with $LTC_4$, were shown to be important in the transport of other substrates, as well. Binding of a substrate brings the TMDs closer to each other which in turn is expected to favour NBD-NBD interactions. This supports the already mentioned theory that substrate-binding facilitates the IF-to-OF large-scale conformational changes. The binding pocket no longer exists in the OF conformation, residues forming the two pockets pulling apart. The positional changes of important binding residues help to move the substrate towards the extracellular space by decreasing non-covalent interactions with the substrate. The substrate is expected to be released before ATP-hydrolysis. OF-to-IF transition, *i.e.*, the NBD dissociation was suggested to be the rate-limiting step in the transport cycle [43], [44], [47].

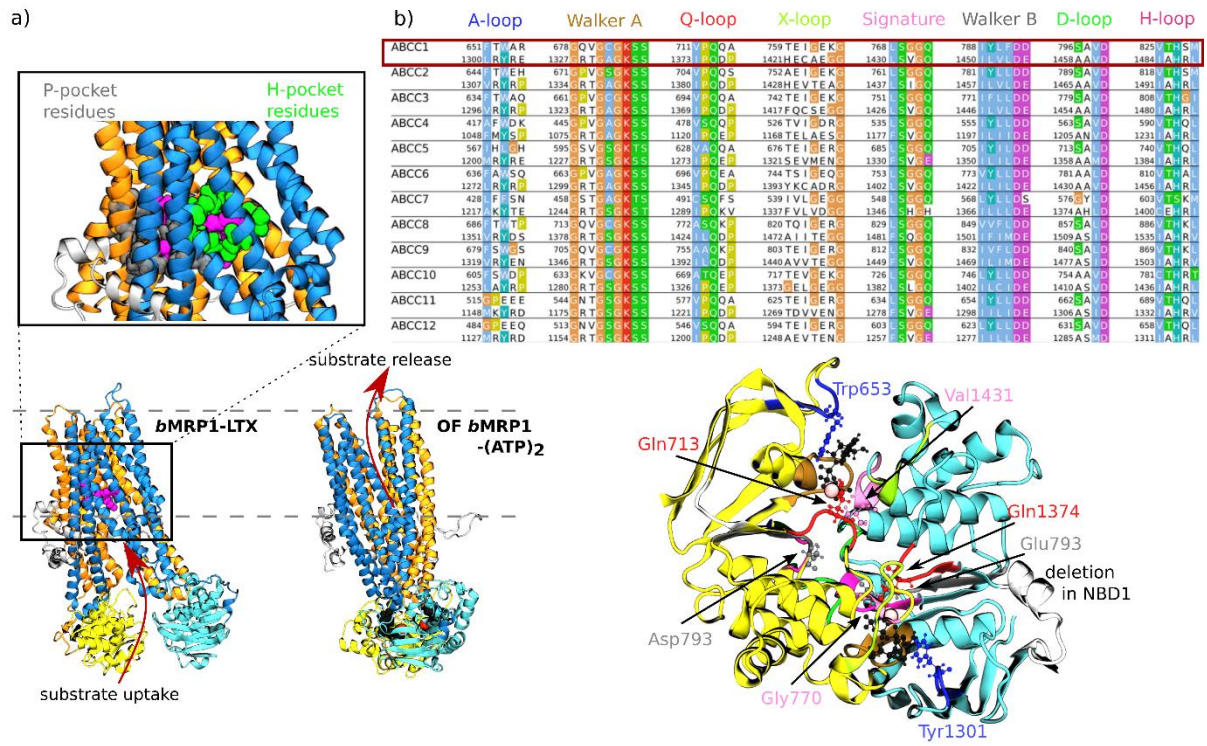



**Figure 15. ABCC family.**

**a)** Inward- and outward-facing conformations of MRP1. $L_0$ is coloured grey, TMD1 orange, NBD1 yellow, TMD2 blue, and NBD2 cyan. $LTC_4$ (LTX) and ATP-$Mg^{2+}$ are shown in magenta and black-red, respectively. The P- and H-pockets are shown in the small picture. **b)** Conserved motifs in the NBDs are shown as sequence (top) Fig. 2 from [39] and on MRP1 (bottom). A-loop, Walker A, Q-loop, X-loop, Signature sequence, Walker B, D-loop, and H-loop are coloured differently. Important residues in these motives have CPK representation (bottom).

## References


[1] B. Alberts, *Molecular biology of the cell*, Sixth edition. New York, NY: Garland Science, Taylor and Francis Group, 2015.

[2] T. Kroll, M. Prescher, S. H. J. Smits, and L. Schmitt, 'Structure and Function of Hepatobiliary ATP Binding Cassette Transporters', *Chem Rev*, vol. 121, no. 9, pp. 5240–5288, May 2021, doi: 10.1021/acs.chemrev.0c00659.

[3] H. Rezayatmand, M. Razmkhah, and I. Razeghian-Jahromi, 'Drug resistance in cancer therapy: the Pandora's Box of cancer stem cells', *Stem Cell Res Ther*, vol. 13, no. 1, p. 181, Dec. 2022, doi: 10.1186/s13287-022-02856-6.

[4] M. P. Doogue and T. M. Polasek, 'The ABCD of clinical pharmacokinetics', *Ther Adv Drug Saf*, vol. 4, no. 1, pp. 5–7, Feb. 2013, doi: 10.1177/2042098612469335.

[5] H. P. Rang and M. M. Dale, Eds., *Rang and Dale's pharmacology*, 8. ed. Amsterdam Heidelberg: Elsevier, Churchill Livingstone, 2016.

[6] K. K. Jain, *Drug Delivery System*, Softcover reprint of the original 2nd ed. 2014 édition. Humana, 2016.

[7] I. Antal, 'Hatóanyag-felszabadító rendszerek előadás vázlat és segédlet'. 2016.

[8] S. Tayyab and S. R. Feroz, 'Chapter Nine - Serum albumin: clinical significance of drug binding and development as drug delivery vehicle', in *Advances in Protein Chemistry and Structural Biology*, vol. 123, R. Donev, Ed. Academic Press, 2021, pp. 193–218. doi: https://doi.org/10.1016/bs.apcsb.2020.08.003.

[9] G. You and M. E. Morris, Eds., *Drug transporters: molecular characterization and role in drug disposition*, Second edition. Hoboken, New Jersey: John Wiley & Sons, Inc, 2014.

[10] T. Harayama and H. Riezman, 'Understanding the diversity of membrane lipid composition', *Nat Rev Mo. Cell Biol*, vol. 19, no. 5, pp. 281–296, May 2018, doi: 10.1038/nrm.2017.138.

[11] Fonyó A. and Ligeti E., *Az orvosi élettan tankönyve*, Fourth edition. Budapest: Medicina Könyvkiadó Zrt., 2008.

[12] H. I. Ingólfsson *et al.*, 'Lipid Organization of the Plasma Membrane', *J Am Chem Soc*, vol. 136, no. 41, pp. 14554–14559, Oct. 2014, doi: 10.1021/ja507832e.

[13] M. Doktorova, J. L. Symons, and I. Levental, 'Structural and functional consequences of reversible lipid asymmetry in living membranes', *Nat Chem Biol*, vol. 16, no. 12, pp. 1321–1330, Dec. 2020, doi: 10.1038/s41589-020-00688-0.

[14] H. Fyrst, D. R. Herr, G. L. Harris, and J. D. Saba, 'Characterization of free endogenous C14 and C16 sphingoid bases from Drosophila melanogaster', *J Lipid Res*, vol. 45, no. 1, pp. 54–62, Jan. 2004, doi: 10.1194/jlr.M300005-JLR200.

[15] S. M. Mandala *et al.*, 'Sphingoid base 1-phosphate phosphatase: A key regulator of sphingolipid metabolism and stress response', *Proc Natl Acad Sc. USA*, vol. 95, no. 1, pp. 150–155, Jan. 1998, doi: 10.1073/pnas.95.1.150.

[16] M. J. Hernández-Corbacho, M. F. Salama, D. Canals, C. E. Senkal, and L. M. Obeid, 'Sphingolipids in mitochondria', *Biochim Biophys Acta Mol Cell Biol Lipids*, vol. 1862, no. 1, pp. 56–68, Jan. 2017, doi: 10.1016/j.bbalip.2016.09.019.

[17] L.-G. Bronstein, Á. Tóth, P. Cressey, V. Rosilio, F. Di Meo, and A. Makky, 'Phospholipid–porphyrin conjugates: deciphering the driving forces behind their supramolecular assemblies', *Nanoscale*, vol. 14, no. 19, pp. 7387–7407, 2022, doi: 10.1039/D2NR01158A.

[18] 'Drug Delivery Systems'. https://www.nibib.nih.gov/science-education/science-topics/drug-delivery-systems-getting-drugs-their-targets-controlled-manner (accessed Sep. 07, 2022).

[19] M. R. Clark, 'Flippin' lipids', *Nat Immunol*, vol. 12, no. 5, pp. 373–375, May 2011, doi: 10.1038/ni.2024.

[20] V. Blickle, T. Speck, C. Lutz, U. Seifert, and C. Bechinger, 'Einstein Relation Generalized to Nonequilibrium', *Phys Rev Lett*, vol. 98, no. 21, p. 210601, May 2007, doi: 10.1103/PhysRevLett.98.210601.

[21] 'Mean Square Displacement — GROMACS 2022.3 documentation'. https://manual.gromacs.org/current/reference-manual/analysis/mean-square-displacement.html (accessed Sep. 07, 2022).

[22] D. J. Smith, J. B. Klauda, and A. J. Sodt, 'Simulation Best Practices for Lipid Membranes [Article v1.0]', *LiveCoMS*, vol. 1, no. 1, 2019, doi: 10.33011/livecoms.1.1.5966.

[23] C. T. Boughter, V. Monje-Galvan, W. Im, and J. B. Klauda, 'Influence of Cholesterol on Phospholipid Bilayer Structure and Dynamics', *J Phys Chem B*, vol. 120, no. 45, pp. 11761–11772, Nov. 2016, doi: 10.1021/acs.jpcb.6b08574.

[24] M. R. Krause and S. L. Regen, 'The Structural Role of Cholesterol in Cell Membranes: From Condensed Bilayers to Lipid Rafts', *Acc Chem Res*, vol. 47, no. 12, pp. 3512–3521, Dec. 2014, doi: 10.1021/ar500260t.

[25] G. W. Feigenson, 'Phase diagrams and lipid domains in multicomponent lipid bilayer mixtures', *Biochim Biophys Acta Biomembr*, vol. 1788, no. 1, pp. 47–52, Jan. 2009, doi: 10.1016/j.bbamem.2008.08.014.

[26] ‘3.1: Membrane Phase Transitions’, *Physics LibreTexts*, Apr. 09, 2015. https://phys.libretexts.org/Courses/University_of_California_Davis/UCD%3A_Biophysics_241_-_Membrane_Biology/03%3A_Membrane_Phases_and_Morphologies/3.01%3A_Membrane_Phase_Transitions (accessed Sep. 05, 2022).

[27] N. Yandrapalli, D. Muriaux, and C. Favard, ‘Lipid domains in HIV-1 assembly’, *Front Microbiol*, vol. 5, May 2014, doi: 10.3389/fmicb.2014.00220.

[28] N. Kučerka, M.-P. Nieh, and J. Katsaras, ‘Fluid phase lipid areas and bilayer thicknesses of commonly used phosphatidylcholines as a function of temperature’, *Biochim Biophys Acta Biomembr*, vol. 1808, no. 11, pp. 2761–2771, Nov. 2011, doi: 10.1016/j.bbamem.2011.07.022.

[29] C. J. Dickson, V. Hornak, R. A. Pearlstein, and J. S. Duca, ‘Structure–Kinetic Relationships of Passive Membrane Permeation from Multiscale Modeling’, *J Am Chem Soc*, vol. 139, no. 1, pp. 442–452, Jan. 2017, doi: 10.1021/jacs.6b11215.

[30] A. Janaszkiewicz *et al.*, ‘Insights into the structure and function of the human organic anion transporter 1 in lipid bilayer membranes’, *Sci Rep*, vol. 12, no. 1, p. 7057, Dec. 2022, doi: 10.1038/s41598-022-10755-2.

[31] Q. Faucher, H. Alarcan, P. Marquet, and C. Barin-Le Guellec, ‘Effects of Ischemia-Reperfusion on Tubular Cell Membrane Transporters and Consequences in Kidney Transplantation.’, *J Clin Med*, vol. 9, no. 8, Aug. 2020, doi: 10.3390/jcm9082610.

[32] A. Janaszkiewicz *et al.*, ‘Substrate binding and lipid-mediated allostery in the human organic anion transporter 1 at the atomic-scale’. bioRxiv, p. 2022.07.14.500056, Jul. 15, 2022. doi: 10.1101/2022.07.14.500056.

[33] The International Transporter Consortium *et al.*, ‘Membrane transporters in drug development’, *Nat Rev Drug Discov*, vol. 9, no. 3, pp. 215–236, Mar. 2010, doi: 10.1038/nrd3028.

[34] M. J. Zamek-Gliszczynski *et al.*, ‘Transporters in Drug Development: 2018 ITC Recommendations for Transporters of Emerging Clinical Importance’, *Clin Pharmacol Ther*, vol. 104, no. 5, pp. 890–899, 2018, doi: https://doi.org/10.1002/cpt.1112.

[35] M. K. DeGorter, C. Q. Xia, J. J. Yang, and R. B. Kim, ‘Drug Transporters in Drug Efficacy and Toxicity’, *Annu Rev Pharmacol Toxicol*, vol. 52, no. 1, pp. 249–273, Feb. 2012, doi: 10.1146/annurev-pharmtox-010611-134529.

[36] V. Vasiliou, K. Vasiliou, and D. W. Nebert, ‘Human ATP-binding cassette (ABC) transporter family’, *Hum Genomics*, vol. 3, no. 3, Art. no. 3, Dec. 2009, doi: 10.1186/1479-7364-3-3-281.

[37] C. Crowe-McAuliffe *et al.*, ‘Structural basis of ABCF-mediated resistance to pleuromutilin, lincosamide, and streptogramin A antibiotics in Gram-positive pathogens’, *Nat Commun*, vol. 12, no. 1, p. 3577, Dec. 2021, doi: 10.1038/s41467-021-23753-1.

[38] C. Thomas and R. Tampé, ‘Structural and Mechanistic Principles of ABC Transporters’, *Annu Rev Biochem*, vol. 89, no. 1, pp. 605–636, Jun. 2020, doi: 10.1146/annurev-biochem-011520-105201.

[39] T. Stockner, R. Gradisch, and L. Schmitt, ‘The role of the degenerate nucleotide binding site in type I ABC exporters’, *FEBS Lett*, vol. 594, no. 23, pp. 3815–3838, Dec. 2020, doi: 10.1002/1873-3468.13997.

[40] S. Srikant and R. Gaudet, ‘Mechanics and pharmacology of substrate selection and transport by eukaryotic ABC exporters’, *Nat Struct Mol Biol*, vol. 26, no. 9, Art. no. 9, Sep. 2019, doi: 10.1038/s41594-019-0280-4.

[41] S. Hofmann *et al.*, ‘Conformation space of a heterodimeric ABC exporter under turnover conditions’, *Nature*, vol. 571, no. 7766, Art. no. 7766, Jul. 2019, doi: 10.1038/s41586-019-1391-0.

[42] H. Göddeke *et al.*, ‘Atomistic Mechanism of Large-Scale Conformational Transition in a Heterodimeric ABC Exporter’, *J Am Chem Soc*, vol. 140, no. 13, pp. 4543–4551, Apr. 2018, doi: 10.1021/jacs.7b12944.

[43] Z. L. Johnson and J. Chen, ‘ATP Binding Enables Substrate Release from Multidrug Resistance Protein 1’, *Cell*, vol. 172, no. 1–2, pp. 81-89.e10, Jan. 2018, doi: 10.1016/j.cell.2017.12.005.

[44] Z. L. Johnson and J. Chen, ‘Structural Basis of Substrate Recognition by the Multidrug Resistance Protein MRP1’, *Cell*, vol. 168, no. 6, pp. 1075-1085.e9, Mar. 2017, doi: 10.1016/j.cell.2017.01.041.

[45] M. Akdel *et al.*, ‘A structural biology community assessment of AlphaFold 2 applications’, Biophysics, preprint, Sep. 2021. doi: 10.1101/2021.09.26.461876.

[46] B. Chantemargue *et al.*, ‘Structural patterns of the human ABCC4/MRP4 exporter in lipid bilayers rationalize clinically observed polymorphisms’, *Pharmacol Res*, vol. 133, pp. 318–327, 2018, doi: 10.1016/j.phrs.2018.02.029.

[47] L. Wang, Z. L. Johnson, M. R. Wasserman, J. Levring, J. Chen, and S. Liu, ‘Characterization of the kinetic cycle of an ABC transporter by single-molecule and cryo-EM analyses’, *eLife*, vol. 9, p. e56451, May 2020, doi: 10.7554/eLife.56451.

[48] K. Immadisetty, J. Hettige, and M. Moradi, ‘Lipid-Dependent Alternating Access Mechanism of a Bacterial Multidrug ABC Exporter’, *ACS Cent Sci*, vol. 5, no. 1, pp. 43–56, Jan. 2019, doi: 10.1021/acscentsci.8b00480.

[49] K. Kapoor, S. Pant, and E. Tajkhorshid, ‘Active participation of membrane lipids in inhibition of multidrug transporter P-glycoprotein’, *Chem Sci*, vol. 12, no. 18, pp. 6293–6306, 2021, doi: 10.1039/D0SC06288J.

[50] X. Cong, Y. Liu, W. Liu, X. Liang, and A. Laganowsky, ‘Allosteric modulation of protein-protein interactions by individual lipid binding events’, *Nat Commun*, vol. 8, no. 1, p. 2203, Dec. 2017, doi: 10.1038/s41467-017-02397-0.

[51] Z. Cournia and A. Chatzigoulas, ‘Allostery in membrane proteins’, *Cur Opin Struct Bio*, vol. 62, pp. 197–204, Jun. 2020, doi: 10.1016/j.sbi.2020.03.006.

# Chapter II. Computational chemistry

## II.1. General description of computational chemistry

Many proteins are involved in drug disposition (shown in Chapter 1). For example, MRP2 and MRP4 were highlighted to be important in pharmacological processes in liver and kidney cells [1], [2]. It is essential to know more about the dynamics and substrate transport mechanisms of pharmacologically relevant proteins. Computational chemistry can help to model proteins in the membrane and to give some hints about the dynamics of proteins and the influence of the surrounding lipids [3]–[5]. Computational chemistry is a field in which the combination of mathematical methods and fundamental laws of physics are applied to study chemical and/or biological models. Historically, different computational approaches exist depending on the object of interest: electrons or nuclei respectively for quantum mechanical and molecular mechanics methods.

**Quantum mechanics** (QM) describes molecules as collections of nuclei and electrons focusing on the behaviour of the latter. It is computationally expensive as it approximates the wave function described in the Schrödinger equation [6]. QM in detail is not part of this thesis. As QM is computationally expensive for large molecules, **molecular mechanics** (MM) is more commonly used to model molecules for which electronic behaviours are not considered. MM does not explicitly consider electrons since the nucleus and its electrons are considered as one atom. In molecular mechanics, molecules are described by the "ball and spring" model. Atoms are represented as different sized balls, while the bonds are springs with different lengths [6], [7]. More about MM can be found in section II.2. Both QM and MM can be used in the so-called hybrid QM/MM methods to take advantage of these techniques to calculate the free energy landscape of enzymatic reactions, for example [6]. Although, QM/MM is out of the topic of this thesis, as well.

**Molecular dynamics** (MD) model ensembles of different molecules, such as lipids and/or proteins. MD methods are based on MM for which dynamics is included by iteratively solving the Newton's equations of motions providing a numerical solution. MD simulations sample the distribution of choice [3], [8]. MD is described in detail in section II.3. However, the timescale of basic molecular dynamics simulations is often too short to catch biologically important processes. In many cases, the sampling must be either accelerated or enhanced. This can be achieved by using one of the **enhanced sampling methods** described in section II.5.

## II.2. Molecular mechanics

### II.2.1. Newton's classical physics as the essentials of molecular mechanics methods

MM is used to model molecules based on Newton's laws of motion (a.k.a. classical mechanics). According to Newton's first law, a body will stay in a resting state or move with constant velocity ($v$) if the sum of all external forces equals zero. In accordance with the second law, an external force ($\vec{F}$) produces an acceleration ($\vec{a}$) on a body that is inversely proportional to the mass ($m$) of the body (Eq. 7) [3], [7], [9].

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

**Equation 7. Newton's second law**

The third law states that if a body "A" applies a force onto body "B", body "B" applies equal but opposite force on body "A" ($\overrightarrow{F_{AB}} = -\overrightarrow{F_{BA}}$). This indicates that two objects can expend either attractive or repulsive forces. If well parameterized, Newton's laws give an appropriate estimation for molecular systems [3].

From now on, this section contains how molecular systems can be described in the framework of classical mechanical physics [3], [7], [9]:

- The atoms ($N$) of a system are described by Cartesian coordinates: $\vec{X} \equiv (x_1, y_1, z_1, \dots, x_N, y_N, z_N)$ where 1,2, …, $N$ refer to particles.
- The motion of these particles at any time ($t$) can be described by Cartesian position vectors: $\vec{r}(t) = \big(x(t), y(t), z(t)\big)$.
- The velocity at a given time is the derivative of the position vector: $\vec{v}(t) = \frac{d\vec{r}}{dt} = \dot{\vec{r}}$.
- While the acceleration is the second derivative of the position (or the derivative of the velocity): $\vec{a}(t) = \frac{d^2\vec{r}}{dt^2} = \ddot{\vec{r}} = \frac{d\vec{v}}{dt}$.
- Newton's second law can thus be written as $\vec{F} = m\frac{d^2\vec{r}}{dt^2}$. It is a second order differential equation. To get an explicit solution, starting positions ($r(0)$) and velocities ($v(0)$) should be defined[1].
- According to Newton's first law, the position at time ($t$) is $\vec{r}(t) = \vec{r}(0) + \vec{v}t$. This is a trajectory given by a Taylor expansion[2] as $\vec{v}(t) = \dot{\vec{r}} \rightarrow \vec{r}(t) = \vec{r}(0) + \dot{\vec{r}}t$.
- In the modelled systems, many particles ($N$) can be found. The force of a particle ($i$) depends on all other particles and the velocity of itself: $\vec{F}_i = \vec{F}_i\big(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N, \dot{\vec{r}}_i\big) = m\ddot{\vec{r}}$.
- Often, the momenta of Cartesian coordinates ($\vec{p}$) is used: $\vec{p}_i = m_i\vec{v}_i = m_i\dot{\vec{r}}_i$. This makes possible to change the space for a 6N-dimensional space called phase space where the state of a particle can be clearly specified by six quantities ($x, y, z, p_x, p_y, p_z$). The force can be written as $\vec{F}_i = m_i\vec{a}_i = m_i\frac{d\vec{v}_i}{dt} = \frac{d\vec{p}_i}{dt}$ in this phase space.
- Generalized coordinates ($q$) can be used to describe the whole system; therefore, they are often applied in theoretical chemistry. Generalized coordinates are another set of 3N coordinates ($\vec{q}_1, \vec{q}_2, \dots, \vec{q}_{3N}$), in a 3D system, and they derive from the original Cartesian coordinates ($\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N$) by a coordinate transformation: $q_\alpha = f_\alpha(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N), where\ \alpha = 1, \dots, 3N$.
  (It has a unique inverse: $r_i = g_i(\vec{q}_1, \vec{q}_2, \dots, \vec{q}_{3N}), where\ i = 1, \dots, N$.)
- Taking a generalized formulation, the force can be written as $F_i(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N) = \frac{-\partial}{\partial r_i}U(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$, where $U(\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$ is the potential energy function (see its calculation in section II.2.2). This results in the following form of Newton's second law:

[1] A second order differential equation can have infinite solution from which one is calculated where $f(0) = a, f'(0) = b. a, b$ are given numbers [10]

[2] Taylor series expansion of a function ($f$) around $x_0$ is
$f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$ [10]

$\frac{-\partial}{\partial r_i} U(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N) = m\frac{d^2\vec{r}}{dt^2} = \frac{d\vec{p}_i}{dt}$. It is a good approximation as nuclei are heavy enough to behave similarly as classical particles.

- The kinetic energy of the momenta of Cartesian coordinates of given velocities can be calculated as $K(\vec{p}) = K(\dot{\vec{r}}_1, \ldots, \dot{\vec{r}}_N) = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\vec{r}}_i^2$.
- The total energy ($E$) is given as the sum of the kinetic and potential energies: $E = \frac{1}{2}\sum_{i=1}^{N} m_i \dot{\vec{r}}_i^2 + U(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N)$.

### II.2.2. Total potential energy

As mentioned in the introduction of this chapter, molecules are described by the "ball and spring" model in molecular mechanics. Between the atoms, different forces are present to keep them in a proper conformation of a molecule. Interactions between atoms can be either bonded (*i.e.*, the stretching of the bond, the bending of an angle and the bond torsion) or non-bonded (*i.e.*, electrostatic and van der Waals interactions). These interactions characterize the interplay in molecules, proteins, lipids, *etc*. All necessary information to describe a biological system is defined in force fields (FFs). Many different FFs are available (more in section II.4). The total potential energy of an FF ($U_{total} = U(\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N)$) can be approached by the sum of the energies resulted by the bonded and non-bonded interactions, and sometimes, by the coupling between bonded interactions ($U_{cross}$) (Eq. 7) [3], [7], [9]. In the next sections, these components of the potential energy function are detailed.

$$\boldsymbol{U_{total} = U_{bonded} + U_{non-bonded} + U_{cross}}$$
$$U_{total} = \left(U_{streching} + U_{bending} + U_{torsion}\right) + \left(U_{electrostatic} + U_{vanderWaals}\right) + U_{cross}$$

**Equation 8. An approximation of the total potential energy of a system.**

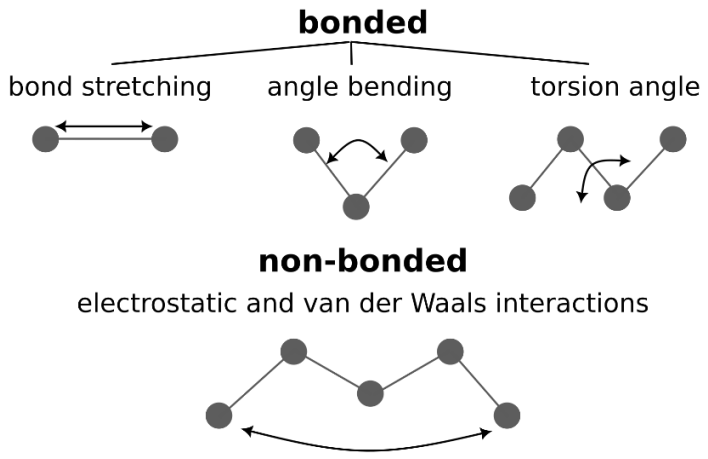



**Figure 16. Bonded and non-bonded interactions.**
[Based on Figure 2.1 [7].]

### II.2.2.1. Bond stretching

The potential energy of bond stretching ($U_{str}$) is a spring that is defined around an equilibrium (mathematically described as the minimum energy point) bond length ($r_{eq}$). In the simplest case, a harmonic potential and second order Taylor expansion are sufficient to model the potential energy of the bond stretching (Eq. 9) [7], [9].

$$\boldsymbol{U_{str}(r)} = U(r_{eq}) + \frac{dU(r_{eq})}{dr}(r - r_{eq}) + \frac{1}{2}\frac{d^2U(r_{eq})}{dr^2}(r - r_{eq})^2 = \frac{\boldsymbol{k}}{\boldsymbol{2}}(\Delta \boldsymbol{r})^2$$

**Equation 9. Bond stretching potential.**

$U_{str}(r)$ can be written simply since 1) $U(r_{eq})$, the zero point for the energy scale is normally defined to zero; 2) $\frac{dU(r_{eq})}{dr}$ is zero as the first derivative in minimum is zero, and 3) the second derivative is the force constant following the Hooke's law[3] [7], [9].

In some systems, *e.g.*, when the vibrational frequencies should be included in the FF, this approximation results in significantly different values compared to experiments. The FF can be improved by including higher order terms in the Taylor expansion. However, it has a computational price as more parameters should be assigned. Moreover, *e.g.*, the cubic harmonic constant is negative; thus, if the higher terms end at the cubic term the molecule can "explode".

The accuracy can be improved by using the Morse potential: $E_{str(Morse)}(\Delta r) = D\left(1 - e^{-\alpha \Delta r}\right)^2$, where $\alpha = \sqrt{\frac{k}{2D}}$ is the fitting constant and D is the dissociation energy. However, the computational cost of the Morse potential is significantly higher than the harmonic approximation, whereas the latter is reasonably accurate on the possible distances for most modelled systems [7], [9].

### II.2.2.2. Angle bending

The potential energy of angle bending is also estimated by a polynomial (Taylor) expansion around an equilibrium angle ($\theta_{eq}$) (Eq. 10). As for the bond stretching, the Taylor series usually terminates at second order since this approximation is already sufficient. Higher order terms can be included to improve the accuracy or to reproduce vibrational frequencies. Angles of the di- or trivalent central atoms are special, and the Taylor expansion should contain at least three order terms to handle boundaries for the linearity of the harmonic force constant. Special atom types are often defined for di- or trivalent central atoms (see section II.2.1.4) [7], [9].

$$\boldsymbol{U_{bend}(\theta_{eq})} = \frac{\boldsymbol{k}}{\boldsymbol{2}}(\boldsymbol{\Delta\theta})^2$$

**Equation 10. Angle bending potential.**

[3] Hooke's law: $U = \frac{1}{2}k(\Delta r)^2$ [9]

### II.2.2.3. Torsion angle

Torsion angle can be defined for four (A, B, C, D) atoms in a row as it is the angle between the projection of two bonds, AB and CD, into the BC bond bisecting plane. By definition, this angle is periodic. It is defined either in the range $[0, 2\pi]$ or $[-\pi, \pi]$. It would be obvious to take the positive angles, but for many torsion angles, $\pi$ corresponds to the minimum energy. As the angle is periodic, it is logical to use a periodic function as expansion, such as the Fourier series (Eq. 11). Thus, $\omega = \pi$ results in the minimum energy as the sum is zero, as $cos(\pi) = -1$. Contrary to the bond stretching and angle bending potentials, the rotational cost around a single bond is often low; therefore, a bigger deviation from the equilibrium (minimum) is more possible. This is well represented by the periodic angle definition and the application of Fourier series. The first three terms of the Fourier series do not result in a perfect fitting, but in most cases, it gives a good estimation from the chemical point of view [7], [9].

$$\boldsymbol{U_{tors}(\omega) = \frac{1}{2}\sum_{j} V_j\left[1 + (-1)^{j+1} cos\left(j\omega + \psi_j\right)\right]}$$

$\psi$ – phase angle, the difference from a reference
$V_j$ – signed amplitudes, $j$ – set of periodicities

**Equation 11. Torsion angle potential.**

Signed amplitudes ($V_j$) and set of periodicities ($j$) are specific to a given torsion angle (*i.e.*, set of A, B, C, D atoms) [7], [9].

### II.2.2.4. Out-of-plane bending

A special four-atom system is when three of the atoms are connected to a central atom that is out of the plane. In this case, there is a significant energy penalty; therefore, a special potential term is often added to the equation. The deviation from the plane can be defined either by an angle ($\theta$) or a distance ($r$) (Fig. 17). The angle ($\theta$) is defined between the central atom plane and the plane made by the other three atoms. The distance ($r$) is defined as the distance between the central atom and its projection onto the plane made by the other three atoms. Torsion angles defined like this are called “improper torsions” [7], [9].

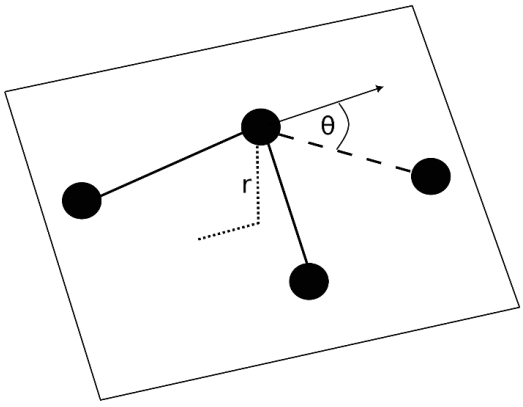



**Figure 17. Out-of-plane bending angle ($\theta$) and distance ($r$) definition.**
[Based on Figure 2.5 of [9].]

### II.2.2.5. Electrostatic interactions

Electrostatic interactions result from the internal implicit distribution of electrons. So that a molecule can have “positive” and “negative” regions. Electrostatic interactions are part of non-bonded interactions that are weaker than bonded interactions. Electrostatic interactions are calculated by the first order Coulomb potential between the partial atomic charges [7], [9].

$$U_{elec}(r_{j,i}) = \sum_{i} \sum_{j>i} \frac{q_i q_j}{\varepsilon r_{i,j}}$$

$q_i$, $q_j$ – partial atomic point charges, $\varepsilon$ – dielectric constant

**Equation 12. Electrostatic potential.**

### II.2.2.6. Lennard-Jones potential

The Lennard-Jones (LJ) potential describes the repulsion and attraction between two non-connected atoms. It is zero at large distances and repulsive at very short distances. At intermediate distances, a small attraction is observable due to the so-called van der Waals interactions which are often described by a dipole-dipole interaction model between two atoms (A, B). A popular function to model these interactions is the 12-6 LJ potential (Eq. 13) where the repulsion has $r^{-12}$ and the attraction $r^{-6}$ dependence. In a system, this equation should be calculated for all pairs of atoms. However, since this term is often null at a large distance, distance cut-off is used to decrease the computational time [7], [9].

$$U_{LJ}(r_{AB}) = 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right]$$

$\sigma_{AB}$ – minimum distance, $\varepsilon$ – depth of this min. distance, $r_{AB}$ – actual distance
Power twelve and six describe the repulsive and attraction terms, respectively.

**Equation 13. Van der Waals potential.**

Special attention should be paid to H-bonds, as an H-bond is a big attraction between a hydrogen and another negative heteroatom (*e.g.*, O, N). A specific potential can then be used for H-bonds [7], [9].

$$U_{E-bond}(r_{AB}) = 4\varepsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{10}\right]$$

**Equation 14. Van der Waals potential for H-bonds.**

### II.2.2.7. Cross potential energy

The above-mentioned stretching, bending and torsion are coupled, *i.e.*, if one is changing it induces the change of another one. For instance, if the angle becomes smaller in a water molecule, the bond length becomes longer as the repulsion between the two H atoms is increased. In FFs, different cross terms can be defined and added to the total potential energy to include this coupling that can improve the FF definition [7], [9].

The possible cross terms are usually written as first-order Taylor series [7]:

- Stretch – bend for A, B, and C atoms:
  $U_{stretch-bend} = k^{ABC}\left(\theta^{ABC} - \theta_0^{ABC}\right)\left[\left(r^{AB} - r_0^{AB}\right) - \left(r^{BC} - r_0^{BC}\right)\right]$

  This is the most important term as the bond stretching and the angle bending are rather coupled.
- Stretch – stretch for A, B, and C atoms:
  $U_{stretch-stretch} = k^{ABC}\left(r^{AB} - r_0^{AB}\right)\left(r^{BC} - r_0^{BC}\right)$
- Bend -bend for A, B, C, and D atoms:
  $U_{bend-bend} = k^{ABCD}\left(\theta^{ABC} - \theta_0^{ABC}\right)\left(\theta^{BCD} - \theta_0^{BCD}\right)$
- Stretch – torsion for A, B, C, and D atoms:
  $U_{stretch-torsion} = k^{ABCD}\left(r^{AB} - r_0^{AB}\right)cos(n\omega^{ABCD})$
- Bend – torsion for A, B, C, and D atoms:
  $U_{bend-torsion} = k^{ABCD}\left(\theta^{ABC} - \theta_0^{ABC}\right)cos(n\omega^{ABCD})$
- Bend – torsion – bend for A, B, C, and D atoms:
  $U_{bend-torsion-bend} = k^{ABCD}\left(\theta^{ABC} - \theta_0^{ABC}\right)\left(\theta^{BCD} - \theta_0^{BCD}\right)cos(n\omega^{ABCD})$

## II.3. Molecular dynamics

### II.3.1. An introduction to molecular dynamics simulations

MD techniques provide numerical solutions for the classical equations of motion and generate a trajectory to extract macroscopic observables. The widely used MD techniques can provide an atomistic description of molecular systems; therefore, they can describe the transport processes of different proteins as well as the influence of the membrane composition [3]–[5].

**Basic concepts of MD** [7]–[9], [11]**:**

- The state of the system is described by the positions ($x$) and momenta ($p$) of all particles:
  $X' = \left(x_1, y_1, z_1, p_{x1}, p_{y1}, p_{z1}, \dots\right) = (q, p)$
  $q = (x_1, y_1, z_1, x_2, y_2, z_2, \dots), p = \left(p_{x1}, p_{y1}, p_{z1}, p_{x2}, p_{y2}, p_{z2}, \dots\right)$
- The above-mentioned Newton law-based equation $\frac{-\partial U}{\partial q} = \frac{dp}{dt}$ can be solved using the Lagrange function or the Hamilton function (Eq. 15). The Lagrange calculation is a single second-order differential equation, and it can be applied only for spatial coordinates and velocities. The Hamilton calculation is a set of first-ordered differential equations and it is good for any set of non-redundant variables. This calculation is more general.

$$\textbf{a)}\ L = T - V, \frac{d}{dt}\frac{\partial L}{\partial p} - \frac{\partial L}{\partial q} = 0$$
$$\textbf{b)}\ H = T + V, \frac{\partial H}{\partial q} + \frac{\partial p}{dt} = 0, \frac{\partial H}{\partial p} - \frac{\partial q}{dt} = 0$$

**Equation 15. Calculation of Newton's law-based equations in MD.**
**a)** Lagrange and **b)** Hamilton functions.

- The time step is an important parameter. It can be increased without decreasing the numerical stability if the degrees of freedom of the highest frequencies are removed. To consider the vibration of hydrogens (fastest movement in molecular systems), the time step should be smaller, resulting in a higher computational cost. For most MD simulations, bonds involving hydrogens are often restrained. A commonly applied algorithm to get rid of these degrees of freedom is the SHAKE which changes the coordinates to fulfil the distance constraints considering a relative tolerance. Another algorithm is the LINCS which resets the lengths after an unconstrained update. The second one is more stable, but it can only be used with bond constraints and isolated angle constraints.
- Most MD systems have an absolute temperature ($T$), defined at the beginning. Although, sometimes the so-called inverted temperature ($\frac{1}{k_B T}$, where $k_B$ is the Boltzmann constant) is used.
- Extended (auxiliary) variable ($\lambda$) can be defined, as well. $\lambda$ can be fix or change according to a pre-defined plane or dynamically. Its principle depends on the type of the MD simulation (see *e.g.*, in section II.5.4).
- The system is placed in a cubic box; therefore, it is common to use periodic boundary conditions (PBCs) to avoid the boiling of the solvent molecules into the "empty" space and to minimize surface effects. PBC means that the cubic box, in which the system is modelled, is duplicated in all directions. If a particle leaves the box on one side, its image enters into the box next to that site. An evident question is how much bigger the box should be than the modelled system. The size of the box should be defined according to the largest cutoff applied (typically 8-12 Å).

### II.3.2. Different ensembles

Different ensembles can be considered for MD simulations, depending on the thermodynamic conditions to be modelled. Simulating isolated systems following the classical laws results in an NVE (constant number of particles, volume, and energy) ensemble also known as microcanonical ensemble. Closed but not heat isolated systems can be approximate as simulations performed in NVT (constant number of particles, volume, and temperature) ensemble (or canonical ensemble). Finally, NPT (constant number of particles, pressure, and temperature) ensemble model isothermal and isobaric conditions which are often the most relevant for biological systems [7]–[9], [12]. To do so, temperature and/or pressure should be adaptable.

#### II.3.2.1. Temperature coupling (thermostats)

Since temperature and total kinetic energy are closely related[4], the velocities are scaled while the temperature remains constant. However, the resulted trajectory can be not totally reliable as it is no longer Newtonian. Several temperature coupling methods exist depending on the FF parametrization or the used code. Commonly used thermostats are the Berendsen [13], Nosé–Hoover [14], and Langevin [12]. Berendsen coupling applies a scaling between a surrounding constant temperature ($T_0$) thermal bath and the system by adding a dissipative Langevin force. In case of the Nosé-Hoover thermostat, added independent variables control

[4] $\langle E_{kin} \rangle = \frac{1}{2}(3N_{atom} - N_{constraint})kT$

the temperature [7], [15]. Finally, the Langevin thermostat is a stochastic approach. It thermalizes each degree of freedom independently, and it uses a randomly added force simulating the random kicks between particles [12]. Temperature in simulations can usually be monitored by ensuring that velocities of particles adopt the Maxwell-Boltzmann distribution. All these coupling methods can be used to keep a chosen temperature or to slowly heat the system. For constant pressure simulations, temperature should usually be already equilibrated using constant volume (NVT) simulations [16].

#### II.3.2.2. Pressure coupling (barostats)

Pressure can be coupled similarly to temperature using Berendsen barostat. A constant "pressure bath" is used and a Langevin force is added to the speed. On the other hand, Monte Carlo barostat samples volume fluctuation and it does not compute the virial[5] so pressure is not available at runtime [16]–[18].

### II.3.3. Solvent/water models

Biological systems almost systematically include water molecules. Therefore, different models exist to mimic water molecules, but there also exist optimized forcefields for other solvents (*e.g.*, methanol, dichloromethane). Solvent models can be divided into two main families: explicit and implicit solvent. In explicit water models, water molecules are present in the simulation as particles while in implicit water models, water is modelled as a homogeneous continuum, for which only electrostatic contribution is included. For water molecules, explicit models are used in MD simulations since implicit models cannot account solvent-solute H-bond network, which is known to play a central role in biology. Explicit models can be divided into two groups: 3- and 4-point models. In 3-point models, water is modelled by three points corresponding to its three atoms. While in 4-point models, the delocalized charge of water's O atom is also included [9]. 5-point water models exist, as well (*e.g.*, TIP5P [19]). However, the larger the number of points by a water molecule, the larger the computational cost. This is even more important as biological systems are mostly made of water molecules [9].

**Table 8. Commonly used water models.**

| 3-point models | | 4-point models | |
|---|---|---|---|
| SPC | Simple Point Charge | OPC | Optimal Point Charges |
| SPC/E | Extended Simple Point Charge | TIP4P | Transferable Intermolecular Potentials 4 Point |
| TIP3P | Transferable Intermolecular Potentials 3 Point | | |

### II.3.4. Running MD simulations

Prior to an MD production run, the modelled molecule should be (i) minimized, (ii) thermalized, and (iii) equilibrated, such as the system is slowly heated to the targeted temperature, and it can relax and reach a preliminary realistic configuration. After these preparation steps, production can be started, where the new velocities and coordinates of the system are calculated by applying either the leap-frog or velocity Verlet algorithm [11].

[5] $\Xi = \frac{-1}{2}\sum_{i<j}^{N} r_{ij}^{n} \otimes F_{ij}$

Leap-frog does not update velocities and positions at the same time, but with a half time step shift (Eq. 16) [20].

$$v\left(t+\frac{1}{2}\Delta t\right) = v\left(t-\frac{1}{2}\Delta t\right) + \frac{\Delta t}{m}F(t)$$
$$r(t+\Delta t) = r(t) + \Delta t v\left(t+\frac{1}{2}\Delta t\right)$$

**Equation 16. The update of velocities ($v$) and positions ($r$) in leap-frog algorithm.**

The velocity Verlet integrator is more precise. It is compatible with maintained (coupled) temperature or pressure simulations (sections II.3.1.1 and II.3.1.2). In this case, positions and velocities are updated at the same time (Eq. 17). For a standard MD output, where there is no position or velocity coupling, the two integrators are giving the same results [11].

$$v(t+\Delta t) = v(t) + \frac{\Delta t}{2m}[F(t)+F(t+\Delta t)]$$
$$r(t+\Delta t) = r(t) + \Delta t v + \frac{\Delta t^2}{2m}F(t)$$

**Equation 17. The update of velocities ($v$) and positions ($r$) in velocity Verlet algorithm.**

Both algorithms are implemented in Amber and Gromacs, in two commonly used software for MD [11], [16]. Moreover, different routines exist to run these algorithms on CPU and GPU.

In Amber, two routines are implemented to run MD simulations and preparatory steps: 1) sander (Simulated Annealing with NMR-Derived Energy Restraints) and 2) pmemd (Particle Mesh Ewald Molecular Dynamics). More features have been made available in both routines as pmemd can provide better performance. However, there are still some features that are only supported in sander and *vice versa*. Sander is using the simplest equation to calculate the potential energy (Eq. 18). Pmemd aims to improve the performance by treating the long-ranged component of the non-bonded electrostatic interactions. This is the primary engine for molecular dynamics in Amber. Pmemd supports the use of GPUs, even parallelly. The format of the MD-parameter input file (*in/mdin*) is the same for both routines and it contains all information about the run, such as temperature, MD length, timestep, *etc*. Both routines create a trajectory, often stored in a series of binary trajectory files (*netcdf* format). The trajectory could be stored in one file, but for security reasons, it is better to use separated files. To save space, trajectory files usually do not contain the velocities. However, velocities are stored in checkpoint files that are saved less often. These files make it possible to continue or restart a simulation from the stored point [16].

To run an MD simulation in Gromacs, everything should be defined in an input (*mdp*) file. Moreover, a run input (*tpr*) file is necessary which combines the information from a topology (*top*), a structure (*gro*), an MD-parameter (*mdp*), and optionally an index (*ndx*) file. This can be created by the *gmx grompp* command. Simulation, energy minimization, and normal mode analysis can be run by the *gmx mdrun* command. The two most important routines that can be used in parallelization are the domain decomposition and the particle mesh Ewald. In domain decomposition, the components of the non-bonded interactions are broken into domains following the principle that most interactions are local. Particle mesh Ewald is the same as in Amber. Both domain decomposition and particle mesh Ewald create a trajectory (*trr*) file storing

the positions, velocities, and optionally forces. Checkpoint files are created to allow the restart of simulations starting from a given point [11].

### II.4. Parameter definition, the force field

#### II.4.1. Basic concepts

The use of FFs is central for molecular simulations. Potentials, described in sections II.2.2.1-II.2.2.7, require a careful definition of each parameter. Technically, FFs often have optimized parameters available in the code *e.g.*, for proteins, lipids, water, ions and small molecules. A key point is the use of compatible forcefields to each other. Moreover, computational accuracy can be increased by combining specific FFs which were tested and validated as more efficient together (*e.g.*, ff19SB [21] with OPC).

Nowadays, various FFs are available. Some of the FFs are better fitted for large systems, others for small/medium sized molecules. Some FFs are better for ordered proteins; others are as much as possible optimized for intrinsically disordered proteins. Newer FFs aim to be efficient for both ordered and disordered proteins and to fit well for all system sizes. FFs differ in the functional form, *i.e.*, the parameter fitting algorithm of the above-described energy potentials (section II.2.1).

The most time-consuming term in the FF energy calculation is the non-bonded potential. Thus, many FFs apply a cutoff distance (*e.g.*, 10 Å) for van der Waals interaction calculations. To avoid the distance calculation in every step, distances are calculated on the starting structure and saved into a list which is updated *e.g.*, every 20 steps during the simulations. This can save some calculation time. However, often an optimized geometry is in the focus of interest; thus, for large molecules, the cutoff does not result in time gain because of a complex geometry optimization. The use of the same cutoff is more problematic in case of Coulomb interactions as its relationship with the distance is $r^{-3}$ (dipole-dipole) or $r^{-1}$ (charge-charge) contrary to $r^{-6}$ for van der Waals. Another problem is how to treat the electrostatic energies out of the cutoff. If they are set to zero, the energy function will no longer be continuous. Furthermore, it would neglect a significant contribution from long-range electrostatic interactions. The best is to use two cutoffs and a switching function that is smoothly reducing the potential to zero. Modern FFs use fast multipole or Ewald sum methods to achieve this smoothing as well as to include long-range interaction [7], [9]. The Ewald sum concept is to split the electrostatic interactions into near and far contributions [7].

FF definitions are often optimized for MD programs as often an MD software and an FF have been developed parallelly. However, with some limitations, it is possible to use other FFs in an MD program. In the next two sections, parameter definition approaches used in two extensively used molecular dynamics programs (namely Amber, and Gromacs) are described. Then, an overview about a machine learning (ML)-based force field approach is proposed. Finally, a technical example of how to parameterize missing terms for small molecules is explained.

#### II.4.2. Parameter definition in Amber

Parameters should be defined according to the potential energy calculation. In the simplest case, Amber uses the following Hamiltonian for the potential energy calculation [16].

$$U_{tot} = \sum_{bonds} k_b(r - r_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} V_n[1 + cos(n\phi - \gamma)] + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}}\right]$$

$k_b, k_\theta$ – force constants, $r_0$ – equilibrium bond length, $\theta_0$ – equilibrium bond angle, $V_n$ – torsion barrier term, $\gamma$ – phase, $n$ – periodicity, $A_{ij}, B_{ij}$ – specific parameters, $q_i, q_j$ – partial charges

**Equation 18. Simplest potential energy calculation in Amber.**

Many parameters, such as, $k_b$, $r_0$, $k_\theta$, $\theta_0$, $V_n$, $\gamma$, $A_{ij}$, and $B_{ij}$ are defined in parameter files. For the torsion term, a so-called divider integer which splits the torsion term into individual contributions is defined, as well. Special flags SCNB and SCEE handle 1-4 non-bonded interactions that are interactions between two atoms connected via three consecutive bonds but for which non-covalent terms should be fitted. Out-of-plane is handled as a special improper torsion angle. The last part of the potential energy calculation describes the non-bonded terms where $A_{ij} = \varepsilon R_{(i,j)min}^{12}$ and $B_{ij} = 2\varepsilon R_{(i,j)min}^{6}$ are specific parameters for atom types $i$ and $j$ to describe the van der Waals interactions. In Amber $\varepsilon$ and $R_{(i,j)min}$ values are defined in parameter files for a given atom type. Partial charges ($q_i, q_j$) do not appear directly in parameter files. They are pre-calculated for protein and nucleic acid atoms and stored in fragment libraries [16]. Recommended force field files for Amber family are listed in Table 9.

Amber routines do not read these parameter files directly. Parameters for a system are stored in a special, but ASCII, topology file (*prmtop*). Even if it is possible to modify this topology file using any text editor, it is not recommended as its format should match the system reading routine. Amber tools, *e.g.*, tLeaP or ParmEd can be used to generate and safely modify parameter files. Amber force fields were developed for the Amber software but it is possible to use them in other software [11], [16].

**Table 9.Recommended force fields in Amber.**

| | | | |
|---|---|---|---|
| Protein | FF14SB or FF19SB [21], [22] | Ions | Monovalent Ion parameters from Joung & Cheatham [23], [24] |
| Lipid | Lipid17 or Lipid21 [25], [26] | Substrates | GAFF2 [27] |
| Water | TIP3P or OPC [28]–[31] | Carbohydrate | GLYCAM [32] |
| Nucleotides | modified DNA.OL15 [33], [34], modified RNA.OL3 [34], [35] | ATP and other cofactors, ions, lipids, Carbohydrates, *etc.* | AMBER parameter database from [36] |

### II.4.3. Parameter definition in Gromacs

The potential energy function in Gromacs (Eq. 19) can be divided into three parts: 1) non-bonded, 2) bonded, and 3) restraints. There is a choice of function for both non-bonded

and dihedral interactions. Gromacs uses some extra terms during the potential function calculation compared to Amber [11], [16].

**Additional terms in Gromacs** [11], [16]**:**

- The Urey-Brandley term represents bond-angle vibration between atom triplets ($u = r_{ijk}$ and $u = \theta_{ijk}$ for bond and angle, respectively).
- The simplest improper dihedral is a harmonic potential, and it keeps the planar groups planar.
- The CMAP term is responsible for the torsional correction.
- Many different restrains are available for positions, angles, and distances. These are available in the manual, and they will not be explained here in detail.

$$U_{tot} = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedrals} k_\phi[1 + cos(n\phi - \phi_s)] + \sum_{Urey-Brandley} k_u(u - u_0)^2 + \sum_{impropers} k(\xi - \xi_0)^2 + \sum_{\phi,\psi} U_{CMAP} + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \varepsilon \left[ \left( \frac{R_{(i,j)min}}{r_{ij}} \right)^{12} - \left( \frac{R_{(i,j)min}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\varepsilon r_{ij}} + \sum U_{restrains}$$

$k_b, k_\theta, k_\phi, k_u, k$ – force constants, $b_0$ – equilibrium bond length, $\theta_0$ – equilibrium bond angle, $\phi_s$ – phase, $n$ – periodicity, $u_0$ – equilibrium bond-angle vibration, $\xi_0$ – equilibrium improper dihedral, $R_{(ij)}$, – specific parameters, $q_i, q_j$ – partial charges

**Equation 19. Simplest potential energy calculation in Gromacs [11], [16].**

A topology (*top*) file contains all connections in the system. It is an ASCII file but during the run, the Gromacs pre-processor (*gmx grompp*) creates a binary topology file. Gromacs also creates a specific structure file (*gro*) that can store the velocities, as well. The topology file calls (includes) force field files (*itp*). During the MD simulations, parameters are often included from the *toppar* folder that can be created by CHARMM-GUI [37] or "manually" by copying the necessary files. There is a basic *forcefield.itp* file that determines the atom types and all parameters for the potential energy calculation (Eq. 19). In addition, other *itp* files contain information for the protein, lipids and all molecules included in the system. So, force field files have a different structure compared to Amber [11]. CHARMM force field is often used in the Gromacs software but it is possible to use other FFs, *e.g.*, from the Amber family.

### II.4.4. Machine learning-based force fields

ML technics are involved in many fields; thus, in computational chemistry, as well. One promising approach is the ML-based force field generation. Other applications can be found in Chapter III. The idea of an ML-based FF is to treat large biological systems accurately but at a reasonable time and with acceptable computational resources. It can be achieved by learning functional relationships between inputs and outputs. Thus, suitable reference data with energies and forces from *ab initio* calculations are needed. This is currently a bottleneck, but if enough data are available, an ML-based FF is adaptive. It can learn new configurational environments at any time when data are made available [4], [38]. ML-based FFs are not yet

commonly used; however, this is likely the future as more and more data are available for the learning process and these FFs will provide a more accurate parameter set.

### II.4.5. Parameterization

In the Amber family, parameters for a broad range of atom types are included in the General Amber Force Field (GAFF). However, for some bond, angle, torsion, *etc*. parameters may be not available in the applied force field. In this case, parameters should be generated. This parameterization is often necessary for small molecules such as ligands (*e.g.*, substrates, inhibitors) as they frequently contain special chemical moieties. As electrons are not explicitly treated (the nucleus and its electrons are considered as one particle), only atom types, the connection between them and a starting geometry should be provided during parameterization. Parameters can be generated *e.g.*, using the Antechamber suite that is based on the General Amber Force Field. Often, missing parameters can be extrapolated from QM-based calculations, *e.g.*, for partial atomic charge procedures [7], [9], [16].

## II.5. Enhanced sampling methods and free energy assessment techniques

### II.5.1. A need for enhanced sampling methods

Often, biological events (up to seconds) are far beyond the reach of the current timescale of conventional MD simulations (few μs). There is thus a need to speed up the sampling. An “old” option in many FFs is to use “united atoms”. This means that one molecule is used to model a group of atoms, *e.g.*, $CH_2$. However, often, this is nowadays not enough to catch biologically important processes *e.g.*, transport cycle.

This means that conventional all-atom or united-atom MD simulations cannot explore the full configuration space; thus, the system is quasi-nonergodic. In many cases, the sampling should be accelerated to reach the timescale of biological processes. Enhanced sampling techniques, that are based on statistical and physical ideas and principles, aim to achieve the ergodic dynamics, *i.e.*, an infinitely long trajectory would cover the whole configurational space. Some of these enhanced sampling methods permit only semi-quantitative estimation (exploration), but many of them also enable probability distribution and free energy estimation [8].

Plenty of different enhanced sampling methods are available; therefore, it is difficult to choose the most appropriate one for a given problem. An enhanced sampling classification has been recently introduced by Jérôm Hénin e*t al.* [8]. Moreover, a decision tree (Fig. 18) helping to find the most suitable method has been proposed, as well [8]. Even though other classifications are possible and different approaches can be used to solve a given problem, this initiative might significantly help the MD community (Fig. 18).

Some methods are based on strategies from probability theory, others are based on physical properties. In many techniques, so-called collective variables (CVs) are used to accelerate the sampling by focusing on a given “motion” of interest. This is often achieved by introducing an additional potential acting on selected degrees of freedom (CV). In other words, the system will sample along the expected direction and map the conformational space and the free energy landscape according to the CV.

A good CV should be

- continuously defined and differentiable,
- low-dimensional, and
- capable of separating target states of interest [39].

The definition of CVs is likely one of the most challenging steps in CV-based biased MD methods. Indeed, applying an extra force not on the slowest motions can easily lead to false results. Many different CVs (*e.g.*, root mean square deviation – RMSD, distances, angles, or dihedral angles) are defined in the available MD programs. There is a possibility to define a unique CV, as well. However, nowadays the best methods to find the proper CVs are (partially) ML-dependent. The basics of ML and useful algorithms in CV definition are described in Chapter III.

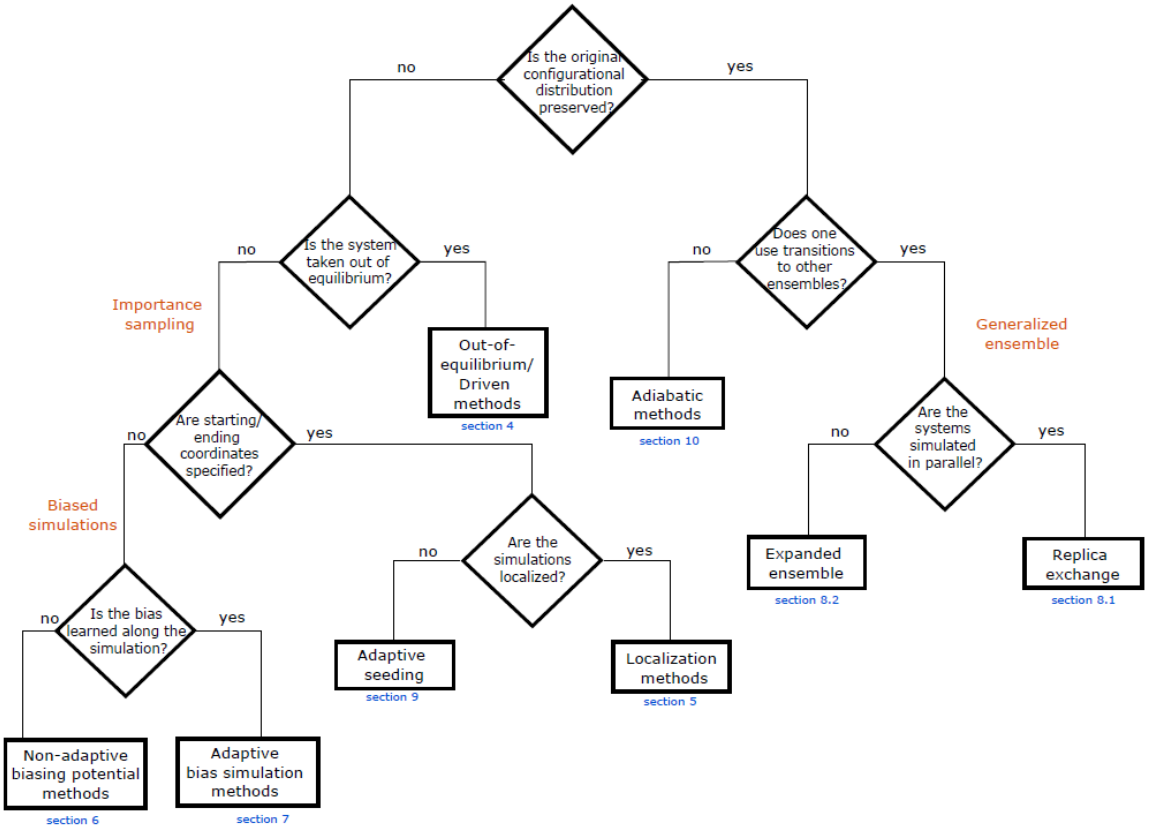



**Figure 18. Enhanced sampling classification based on a decision tree.**
[Figure 1 from [8].]

In the next sections, the focus will be on biased methods that were considered for the investigation of the transport cycle of MRP1 and the mapping of the free energy surface. For sake of time, no results will be provided in the present manuscript as this has become the follow-up work of the present thesis.

### II.5.2. Metadynamics

A well-known CV-based biased MD method is metadynamics, where the valleys of the free energy surface are filled by adding Gaussian hill-like potential along the simulation. Assuming the convergence at the end of the MD simulations, the effective potential becomes flat and the system can freely diffuse. Since information about the Gaussian function deposits is saved during the simulation, the reconstruction of the free energy surface is possible by summing all opposite gaussian functions.

Visually, we can picture this approach by a pedestrian (biased simulation), who fell in a well (free energy local minimum) during the night. He cannot climb and see the shallowest point, as the simulation does not know neither in which direction is the lowest energy pathway. But the pedestrian adds some sand (Gaussian hills). He can slowly fill the pool and find a way out (metadynamics) [40], [41].

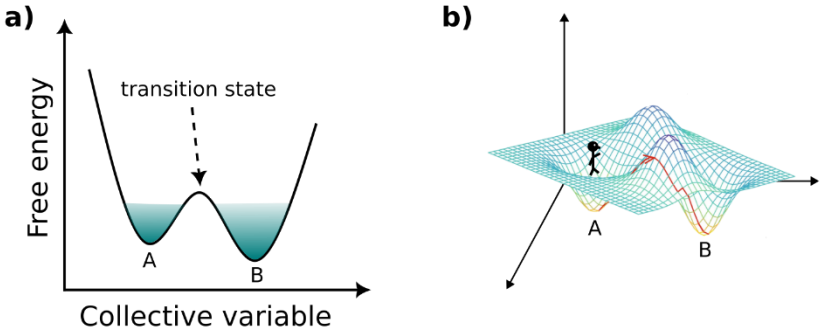



**Figure 19. Illustration of biased techniques.**
**a)** FES according to a reaction coordinate. **b)** 3D free energy surface with a random walker showing the Gaussian hills filling in metadynamics.

### II.5.3. High temperature

An alternative to speed up the sampling without biasing a given CV is to apply a higher temperature (*e.g.*, 375 K instead of the standard 310 K) during the simulations. This has been successfully achieved by Goddeke *et al.* on a bacterial thermophilic ABC transporter, for example [42]. A similar approach is the simulated annealing where the system is consecutively heated up to high temperature several times and then smoothly cooled down in a controlled way [11].

Both methods assume that systems are more dynamic at high temperatures as the metastability of the system is reduced. Therefore, the possibility to capture conformational changes is more likely by crossing barriers of several $kT$ which occur on the millisecond timescale [42], [43]. However, force field parameters might not be suitable for high temperatures. Moreover, the success of the sampling can also depend on the simulated protein.

### II.5.4. Steered molecular dynamics

In steered molecular dynamics (SMD) simulations, an external force is applied to force a molecular motion while analysing the system responses. The force pulls or pushes the system along certain degrees of freedom defined by a CV set. Otherwise, a scaling factor ($\lambda$) is applied

to control some selected potentials during the total potential energy calculation. With this method, the system is biased into an already defined direction; thus, *a priori* knowledge about the system is required. The applied force can be small, in which case, this method is a localization method since the sampling is restrained in a small region of the configurational space. However, usually a relatively big force can be applied to force the transition (*e.g.,* ligand unbinding) that may also result in improbable states. Moreover, only one pathway is sampled if replicas are not used. Whereas, to get the free energy surface (FES), SMD should be run several times starting from different points, increasing the computational cost. However, preliminary SMD simulations can be carried out to provide an initial path that can be refined using more advanced techniques (see *e.g.,* section II.5.6). The applied strong force leads to significant deviations from equilibrium, so SMD is in the out-of-equilibrium category in the above-mentioned classification [8], [16], [43]–[45].

### II.5.5. Accelerated Weight Histogram method

The Accelerated Weight Histogram (AWH) method is a multiple state reweighting method that belongs to the expanded ensemble group in the classification proposed by Jérôm Hénin e*t al.* [8]. It is a relatively new and efficient sampling method in which the system is adaptively biased along the simulation using the framework of the probability weight histogram. One replica, called walker, moves between states along a CV set. Many walkers can be run parallelly to sample faster the configurational space. However, the number of walkers scales with the sampling speed only to a certain extent [8], [41], [43], [46].

The system is described by configurations ($x$) and a system parameter ($\lambda$). The latter is updated according to a given probability distribution. Compared to other biased methods, the target distribution ($\rho$) can be chosen freely, and it is directly controlled. A flat distribution often works well. The initial convergence is fast and exponential, while later it is slower, and it is in the order of the root square of the sample size. The main advantage is that there is only one convergence parameter which is uncritical, and the bias is automatically updated. The walkers share the same bias [8], [41], [43], [46].

Using an initial and arbitrary bias, the real distribution is first estimated then the bias is updated. This process is repeated iteratively until convergence taking advantage of bias history as well as new samples (Fig. 20). Different types of bias can be applied such as harmonic umbrella potential or Boltzmann inversions of a convolution of Gaussians (which is the default, as it is smooth). Therefore, the free energy should be updated iteratively (Eq. 20) following an update size that is constant at the beginning and following Eq. 21 when the update weight is constant. The only tuned parameter is the initial update size set by the initial error and the diffusion. If they are not set correctly, it can cause slow convergence (large parameters) or the pull apart of the system (small parameters) [8], [41], [43], [46].

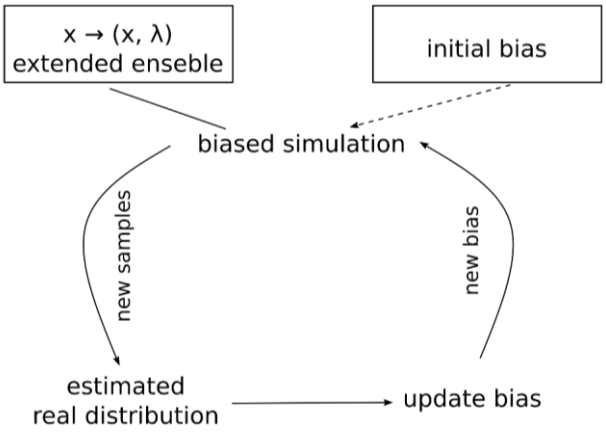



**Figure 20. Schematic AWH method [41], [46].**

$$\Delta F_n(\lambda) = -ln\left(1 + \frac{\Delta W_n(\lambda)}{N_n \pi(\lambda)}\right)\frac{1}{N_n}$$

$W_n$ – collected samples, $N_n$ – prior number of samples
$\pi$ – target distribution (flat), $\lambda$ – extended parameter *e.g.*, temperature or pressure

**Equation 20. Free energy update.**

$$\frac{\mathbf{1}}{\boldsymbol{N_0}} = \frac{\mathbf{1}}{\boldsymbol{N_0(\epsilon_0, D)}}\,\boldsymbol{D\epsilon_0^2}$$

$\epsilon_0$ – initial error in the free energy, $D$ – diffusion along the reaction coordinate
$N_0$ – initial effective number of samples

**Equation 21. Update size.**

### II.5.6. String method with swarms of trajectories

Even though a transition path in a high dimensional CV space has been already defined from biased techniques methods (*e.g.*, steered MD, metadynamics), it is possible that conformational transition was not properly sampled to quantitatively map the FES. Also, the sampled path can differ in the minimum energy path, meaning that the calculated path is not the optimal one. Fortunately, the use of minimum energy path techniques such as the string method with swarms of trajectories can be particularly relevant.

According to the above-mentioned classification, the string method with swarms of trajectories belongs to the adaptive seeding methods. More precisely, it is a transition path-finding method that belongs to the weighted ensemble methodologies. This method requires an *a priori* defined dominant transition pathway which will be refined and minimized using the average dynamic drift of a couple of CVs. The goal is to find the minimum free energy pathway between two main states using milestone conformations [8], [16], [47]–[50].

To achieve it, the dominant transition pathway provides starting points for a myriad of short and unbiased trajectories, called swarms (Fig. 21). The dynamic drift of a couple of CVs is estimated on-the-fly which adapts the original pathway in the direction of the minimum free energy pathway until convergence. Meanwhile, the convergence into the same basin should

be avoided; therefore, a bias is applied on the distances between path images [8], [16], [47]–[50].

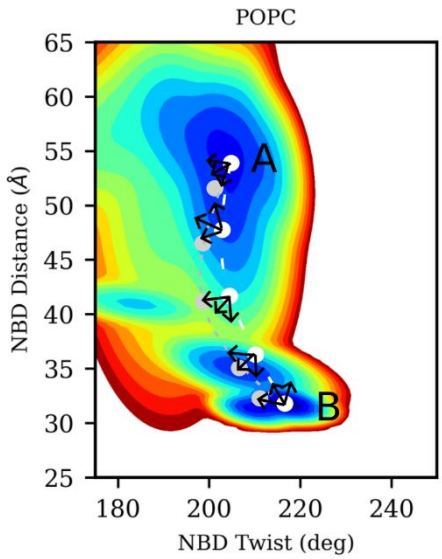



**Figure 21. The principle of string method with swarms of trajectories illustrated on an energy surface of *b*MRP1-(ATP)$_2$ structure in POPC membrane.**

This method is robust, and the calculation time does not scale with the number of CVs. However, it is biased by the original path, and thereby, it cannot find multiple pathways. Due to the many short simulations in parallel and the number of replicas for each, this method requires large computational resources [8], [16], [47]–[50].

### II.6. Molecular docking

Historically, molecular docking calculations aim to find the optimal orientation between two molecules. Nowadays, even though, molecular docking can be used between a wide range of pairs of molecules, it is commonly used to sample the orientation of small molecules in large systems such as proteins, DNA or polymers. To do so, a very simplistic FF-based model for non-covalent interactions is considered and minimized by, often, randomly moving small molecules [7], [9].

In the field of pharmaceutical sciences, molecular docking has been extensively used in drug development in order to find key non-covalent interactions between candidates and targets of biological interest by predicting the proper orientations. Using virtual screening, the number of molecules to be tested experimentally can be decreased. The docking procedure can reduce the time and the cost of drug development by highlighting some relevant drug candidates [7], [9].

If no information about the target binding site is available, molecular docking can be used to map plausible binding sites on the whole protein. The so-called blind docking; however, must be considered carefully since the searching space is too big for an accurate binding prediction. If the binding sites are known from previous blind docking calculations or resolved structure(s), the search area must be narrowed down. This is called refined docking [7], [9], [51], [52].

Docking a molecule has many degrees of freedom – three translational, three rotational and the conformations of small molecules (*e.g.*, dihedral angles) can vary according to its chemical

environment. This is also true *e.g.*, for the target protein residues that can rotate, as well. To reduce the complexity and the computational time, most of the protein residues are rigid and only a few selected dihedral angles of the binding site residues are allowed to rotate. Using this docking approach, the number of molecules that can be docked is limited. Even if the docking procedure is automated, it is not possible to dock billions of molecules with high accuracy within an acceptable time. Many poses can be generated; however, the hit rate is often low, as some of the poses are not relevant based on distances, others based on an energy rank. However, this approach must be very carefully considered since its chemical accuracy remains unsatisfactory. Such calculations require either strong experimental evidences or further computational chemistry techniques [7], [9], [51], [52].

Nevertheless, ML-based molecular docking methods have been recently developed. They aim to improve the hit discovery rate and speed up the search. They make possible the fast docking of many millions-to-billions molecules, and the reduction of the enormous possible molecules for a smaller set [51], [52].

**The basic steps of deep docking** [51]**:**

- To use a big database such as ZINC15.
- To calculate a descriptor *e.g.*, molecular fingerprint.
- To dock a part of the dataset using a conventional docking protocol, described above.
- To relate docking scores to descriptors and train an ML-based model, preferably a deep learning approach.
- To use the ML model to predict docking poses for the rest of the database. A part of the predicted virtual hits is used to augment the training dataset so that it is starting from the conventional docking iteratively, until a pre-defined iteration or convergence.

At the end of these iterative steps, the most relevant hits (that can be still a lot) will be used for further investigation, *e.g.*, for conventional docking methods from which finally the best candidates can be considered, after strong external validation, for more expensive and time-consuming experimental studies [7], [9], [51], [52].

## References


[1] The International Transporter Consortium *et al.*, ‘Membrane transporters in drug development’, *Nat Rev Drug Discov*, vol. 9, no. 3, pp. 215–236, Mar. 2010, doi: 10.1038/nrd3028.

[2] M. J. Zamek-Gliszczynski *et al.*, ‘Transporters in Drug Development: 2018 ITC Recommendations for Transporters of Emerging Clinical Importance’, *Clin Pharmacol Ther*, vol. 104, no. 5, pp. 890–899, Nov. 2018, doi: 10.1002/cpt.1112.

[3] M. E. Tuckerman, *Statistical mechanics: theory and molecular simulation*. Oxford ; New York: Oxford University Press, 2010.

[4] O. T. Unke *et al.*, ‘Machine Learning Force Fields’, *Chem Rev*, vol. 121, no. 16, pp. 10142–10186, Aug. 2021, doi: 10.1021/acs.chemrev.0c01111.

[5] R. Salomon-Ferrer, D. A. Case, and R. C. Walker, ‘An overview of the Amber biomolecular simulation package: Amber biomolecular simulation package’, *WIREs Comput Mol Sci*, vol. 3, no. 2, pp. 198–210, Mar. 2013, doi: 10.1002/wcms.1121.

[6] O. A. Arodola and M. E. Soliman, ‘Quantum mechanics implementation in drug-design workflows: does it really help?’, *Drug Des Devel Ther*, vol. Volume 11, pp. 2551–2564, Aug. 2017, doi: 10.2147/DDDT.S126344.

[7] F. Jensen, *Introduction to computational chemistry*, 2nd ed. Chichester, England ; Hoboken, NJ: John Wiley & Sons, 2007.

[8] J. Hénin, T. Lelièvre, M. R. Shirts, O. Valsson, and L. Delemotte, ‘Enhanced sampling methods for molecular dynamics simulations’. arXiv, Feb. 08, 2022. Accessed: Aug. 23, 2022. [Online]. Available: http://arxiv.org/abs/2202.04164

[9] C. J. Cramer, *Essentials of computational chemistry: theories and models*, 2nd ed. Chichester, West Sussex, England ; Hoboken, NJ: Wiley, 2004.

[10] Z. Vágó, ‘Matematikai analízis I.’ Pázmány Egyetem eKiadó, 2010.

[11] M. J. Abraham, D. van der Spoel, E. Lindahl, B. Hess, and the GROMACS development team, ‘GROMACS User Manual version 5.0.4’. 2014. [Online]. Available: www.gromacs.org

[12] R. L. Davidchack, R. Handel, and M. V. Tretyakov, ‘Langevin thermostat for rigid body dynamics’, *J Chem Phys*, vol. 130, no. 23, p. 234101, Jun. 2009, doi: 10.1063/1.3149788.

[13] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, ‘Molecular dynamics with coupling to an external bath’, *J Chem Phys*, vol. 81, no. 8, pp. 3684–3690, Oct. 1984, doi: 10.1063/1.448118.

[14] D. J. Evans and B. L. Holian, ‘The Nose–Hoover thermostat’, *J Chem Phys*, vol. 83, no. 8, pp. 4069–4074, Oct. 1985, doi: 10.1063/1.449071.

[15] D. Marenduzzo, ‘Thermostats in Molecular dynamics’. Accessed: Sep. 07, 2022. [Online]. Available: https://www2.ph.ed.ac.uk/~dmarendu/MVP/MVP03.pdf

[16] D. A. Case *et al.*, ‘AMBER 2020 Reference Manual’. University of California, San Francisco, 2020. [Online]. Available: https://ambermd.org/doc12/Amber20.pdf

[17] M. Fernández-Pendás, B. Escribano, T. Radivojević, and E. Akhmatskaya, ‘Constant pressure hybrid Monte Carlo simulations in GROMACS’, *J Mol Model*, vol. 20, no. 12, p. 2487, Nov. 2014, doi: 10.1007/s00894-014-2487-y.

[18] ‘Practical considerations for Molecular Dynamics Controlling Pressure’, 2022 2018. https://computecanada.github.io/molmodsim-md-theory-lesson-novice/08-barostats/index.html (accessed Sep. 07, 2022).

[19] Y. Khalak, B. Baumeier, and M. Karttunen, ‘Improved general-purpose five-point model for water: TIP5P/2018’, *J Chem Phys*, vol. 149, no. 22, p. 224507, Dec. 2018, doi: 10.1063/1.5070137.

[20] R. W. Hockney, S. P. Goel, and J. W. Eastwood, ‘Quiet high-resolution computer models of a plasma’, *J Comput Phys*, vol. 14, no. 2, pp. 148–158, Feb. 1974, doi: 10.1016/0021-9991(74)90010-2.

[21] C. Tian *et al.*, ‘ff19SB: Amino-acid specific protein backbone parameters trained against quantum mechanics energy surfaces in solution’, *J Chem Theory Comput*, vol. 16, no. 1, Art. no. BNL-212345-2019-JAAM, Nov. 2019, doi: 10.1021/acs.jctc.9b00591.

[22] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, ‘ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB’, *J Chem Theory Comput*, vol. 11, no. 8, pp. 3696–3713, Aug. 2015, doi: 10.1021/acs.jctc.5b00255.

[23] I. S. Joung and T. E. Cheatham, ‘Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations’, *J Phys Chem B*, vol. 112, no. 30, pp. 9020–9041, Jul. 2008, doi: 10.1021/jp8001614.

[24] I. S. Joung and T. E. Cheatham, ‘Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters’, *J Phys Chem B*, vol. 113, no. 40, pp. 13279–13290, Oct. 2009, doi: 10.1021/jp902584c.

[25] C. J. Dickson *et al.*, ‘Lipid14: The Amber Lipid Force Field’, *J Chem Theory Comput*, vol. 10, no. 2, pp. 865–879, Feb. 2014, doi: 10.1021/ct4010307.

[26] C. J. Dickson, R. C. Walker, and I. R. Gould, ‘Lipid21: Complex Lipid Membrane Simulations with AMBER’, *J Chem Theory Comput*, vol. 18, no. 3, pp. 1726–1736, 2022, doi: 10.1021/acs.jctc.1c01217.

[27] 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 Comput Chem*, vol. 25, no. 9, pp. 1157–1174, Jul. 2004, doi: 10.1002/jcc.20035.

[28] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, ‘Comparison of simple potential functions for simulating liquid water’, *J Chem Phys*, vol. 79, no. 2, pp. 926–935, Jul. 1983, doi: 10.1063/1.445869.

[29] A. D. MacKerell *et al.*, ‘All-atom empirical potential for molecular modeling and dynamics studies of proteins’, *J Phys Chem B*, vol. 102, no. 18, pp. 3586–3616, Apr. 1998, doi: 10.1021/jp973084f.

[30] D. J. Price and C. L. Brooks, ‘A modified TIP3P water potential for simulation with Ewald summation’, *J Chem Phys*, vol. 121, no. 20, pp. 10096–10103, Nov. 2004, doi: 10.1063/1.1808117.

[31] S. Izadi, R. Anandakrishnan, and A. V. Onufriev, ‘Building Water Models: A Different Approach’, *J Phys Chem Lett*, vol. 5, no. 21, pp. 3863–3871, Nov. 2014, doi: 10.1021/jz501780a.

[32] K. N. Kirschner *et al.*, ‘GLYCAM06: A generalizable biomolecular force field. Carbohydrates: GLYCAM06’, *J Comput Chem*, vol. 29, no. 4, pp. 622–655, Mar. 2008, doi: 10.1002/jcc.20820.

[33] K. L. Meagher, L. T. Redman, and H. A. Carlson, ‘Development of polyphosphate parameters for use with the AMBER force field’, *J Comput Chem*, vol. 24, no. 9, pp. 1016–1025, Jul. 2003, doi: 10.1002/jcc.10262.

[34] R. Galindo-Murillo *et al.*, ‘Assessing the Current State of Amber Force Field Modifications for DNA’, *J Chem Theory Comput*, vol. 12, no. 8, pp. 4114–4127, Aug. 2016, doi: 10.1021/acs.jctc.6b00186.

[35] M. Zgarbová *et al.*, ‘Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles’, *J Chem Theory Comput*, vol. 7, no. 9, pp. 2886–2902, Sep. 2011, doi: 10.1021/ct200162x.

[36] R. Bryce, ‘AMBER parameter database’. http://amber.manchester.ac.uk/ (accessed Sep. 07, 2022).

[37] S. Jo, T. Kim, V. G. Iyer, and W. Im, ‘CHARMM-GUI: A web-based graphical user interface for CHARMM’, *J Comput Chem*, vol. 29, no. 11, pp. 1859–1865, Aug. 2008, doi: 10.1002/jcc.20945.

[38] V. Botu, R. Batra, J. Chapman, and R. Ramprasad, ‘Machine Learning Force Fields: Construction, Validation, and Outlook’, *J Phys Chem C*, vol. 121, no. 1, pp. 511–522, Jan. 2017, doi: 10.1021/acs.jpcc.6b10908.

[39] M. M. Sultan and V. S. Pande, ‘Automated design of collective variables using supervised machine learning’, *J Chem Phys*, vol. 149, no. 9, p. 094106, Sep. 2018, doi: 10.1063/1.5029972.

[40] N. Bešker and F. L. Gervasio, ‘Using Metadynamics and Path Collective Variables to Study Ligand Binding and Induced Conformational Transitions’, in *Computational Drug Discovery and Design*, vol. 819, R. Baron, Ed. New York, NY: Springer New York, 2012, pp. 501–513. doi: 10.1007/978-1-61779-465-0_29.

[41] *BioExcel Webinar #39: Accelerating sampling in GROMACS with the AWH method*, (2019). Accessed: Sep. 07, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=SN-Ompgp-mE

[42] H. Göddeke and L. V. Schäfer, ‘Capturing Substrate Translocation in an ABC Exporter at the Atomic Level’, *J Am Chem Soc*, vol. 142, no. 29, pp. 12791–12801, Jul. 2020, doi: 10.1021/jacs.0c05502.

[43] V. Lindahl, J. Lidmar, and B. Hess, ‘Accelerated weight histogram method for exploring free energy landscapes’, *J Chem Phys*, vol. 141, no. 4, p. 044110, Jul. 2014, doi: 10.1063/1.4890371.

[44] B. Isralewitz, J. Baudry, J. Gullingsrud, D. Kosztin, and K. Schulten, ‘Steered molecular dynamics investigations of protein function’, *J Mol Graph Model*, vol. 19, no. 1, pp. 13–25, Feb. 2001, doi: 10.1016/S1093-3263(00)00133-9.

[45] B. Isralewitz, M. Gao, and K. Schulten, ‘Steered molecular dynamics and mechanical functions of proteins’, *Curr Opin Struct Biol*, vol. 11, no. 2, pp. 224–230, Apr. 2001, doi: 10.1016/S0959-440X(00)00194-9.

[46] *BioExcel Webinar #54: Applying the Accelerated Weight Histogram method to alchemical transformations*, (2021). Accessed: Sep. 07, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=E5nGLcbyqTQ

[47] H. Chen *et al.*, ‘A Companion Guide to the String Method with Swarms of Trajectories: Characterization, Performance, and Pitfalls’, *J Chem Theory Comput*, vol. 18, no. 3, pp. 1406–1422, Mar. 2022, doi: 10.1021/acs.jctc.1c01049.

[48] A. C. Pan, D. Sezer, and B. Roux, ‘Finding Transition Pathways Using the String Method with Swarms of Trajectories’, *J Phys Chem B*, vol. 112, no. 11, pp. 3432–3440, Mar. 2008, doi: 10.1021/jp0777059.

[49] B. Roux, ‘String Method with Swarms-of-Trajectories, Mean Drifts, Lag Time, and Committor’, *J Phys Chem A*, vol. 125, no. 34, pp. 7558–7571, Sep. 2021, doi: 10.1021/acs.jpca.1c04110.

[50] S. E. McComas, D. Mitrovic, C. Alleva, M. Bonaccorsi, D. Drew, and L. Delemotte, ‘Determinants of sugar-induced influx in the mammalian fructose transporter GLUT5’. bioRxiv, Jun. 18, 2022. doi: 10.1101/2022.06.17.495601.

[51] F. Gentile *et al.*, ‘Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery’, *ACS Cent Sci*, vol. 6, no. 6, pp. 939–949, Jun. 2020, doi: 10.1021/acscentsci.0c00229.

[52] K. Crampon, A. Giorkallos, M. Deldossi, S. Baud, and L. A. Steffenel, ‘Machine-learning methods for ligand–protein molecular docking’, *Drug Discov Today*, vol. 27, no. 1, pp. 151–164, Jan. 2022, doi: 10.1016/j.drudis.2021.09.007.

# Chapter III. Artificial intelligence and advanced analysis tools

## III.1. The relevance of artificial intelligence in computational chemistry

Defining important features from biological data is essential to understand biological systems. In other words, finding the most relevant characteristics in a big amount of data can help in understanding biological processes. As mentioned in Chapter II, ML-based techniques require a large amount of available data. The aim is often to find the most relevant features among a myriad of possible paths. ML techniques are increasingly present in theoretical chemistry [3]. In this chapter, a short description about ML (Section III.1), its importance in computational methods (Section III.2), such as solving the aforementioned CV definition problem, and ML-based analysis tools used in present thesis can be found (Section III.3).

## III.2. The basics of machine learning

### III.2.1. Artificial intelligence and machine learning

There is no universal definition of artificial intelligence (AI). According to one of the proposed definitions, AI is “the study of how to make computers do things at which, at the moment, people are better” [4]. The goal is to create computational agents that act intelligently and can perform faster a task which is normally performed by human intelligence. This means that the computer can sense, cooperate, learn, and discover. An intelligent agent acts according to the circumstances, dynamically adapts to its environments and goals, learns from experience, and is aware of its limitations. ML improves the performance of tasks through observation and learning. It can identify patterns from data and use them for prediction about unseen data. ML acquires knowledge, so the learning procedure, called training of the model, is a key component [5], [6].

### III.2.2. Training types

Depending on the type and the amount of available data, learning can be supervised, semi-supervised, unsupervised, or reinforcement. **Supervised learning** is the most powerful as it learns from input and annotated output pairs. It is an iterative process. There is a prediction for a known target and the algorithm tries to minimize the difference between them. Thanks to a learned function during the training, it can well distinguish between tiny different states. Supervised learning can be divided into classification (discrete output) and regression (continuous output). Classification predicts group membership, while regression provides a score on a continuous scale. In case of **unsupervised learning**, the output for the inputs is unknown. Therefore, it can be used to identify trends, and patterns or cluster/classify data. Unsupervised learning can be divided into clustering and dimensionality reduction. Clustering aims to group data according to their similarity. Dimensionality reduction is useful *e.g.*, for CV prediction as the goal is to find the most representative features in the dataset. **Semi-supervised learning** can be used if annotated outputs are available only partially. In case of **reinforcement learning**, a system, which can learn from the environment, is built. This type of learning is less popular in theoretical chemistry [6]–[10].

For all training types, data representation is the most important. Data should be cleaned and prepared. The conversion of data into a more suitable representation for an ML algorithm, called featurization, is often necessary as it increases the accuracy of training. In many cases, it is also important to clean the dataset as it often contains outliers, omissions, and errors. This step can include the normalization of data. For supervised learning, the dataset should be

divided into training, validation, and test sets. The training set is used to teach the model, while the validation set is important to iteratively evaluate and improve an ML model. The test set is used for the final evaluation. The best ratio for these sets is 70:15:15 (training:validation:test). The split of these sets can be achieved by a random split or by using clustered data to result in a diverse split. It is important to have equilibrated samples for each possible output in the training set to avoid over- or underfitting. Overfitting means that the model fits well for all the training points, but it fails to predict any data outside of the training samples as it learns the noise in the dataset. The underfitting is its opposite, the model fails for all input due to a poor fitting of real data [6]–[10].

It is not trivial to find a good fitting; therefore, validation of an ML model is a crucial step. The most known validation technique is the k-fold cross-validation. This means that the training and validation sets are alternating, such as every data is used for training and once also for validation. Different splits happen k-times (Fig. 22). 5-Fold cross-validation is a commonly used method as it has a good balance between computational complexity and validation accuracy. Other but similar methods to create train and validation sets exist, as well [6], [11].

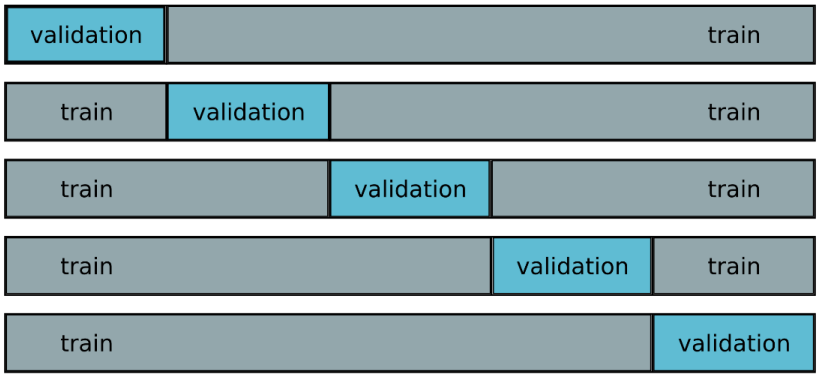



**Figure 22. 5-Fold cross-validation [6].**

During this cross-validation and at the end of the training as well, the performance of the model should be assessed by a metric. The calculation of this metric depends *e.g.,* on the training type, but many of the available metrics are based on ratios of the predicted and actual values (Table 10). True Negative (TN) means that the predicted and the actual output, both are negative (0), while they are both positive (1) in case of True Positive (TP). In these events, the prediction is good. In case of False Positive (FP) and False Negative (FN), the prediction was incorrect positive instead of negative and negative contrary to positive, respectively. A widely used metric is the accuracy (Eq. 22a). However, the calculation of accuracy does not consider the imbalance of a dataset. This can be treated by the balanced accuracy (Eq. 22b) [6], [11].

**Table 10. Connections between predicted and actual values.**

| | | Predicted: *Negative* | Predicted: *Positive* |
|---|---|---|---|
| **Actual** | *Negative* | True Negative (TN) | False Positive (FP) |
| | *Positive* | False Negative (FN) | True Positive (TP) |

$$\mathbf{a})\ Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\mathbf{b})\ Acc_{balanced} = \frac{\frac{TP}{TP + FN} + \frac{TN}{TN + FP}}{2}$$

**Equation 22. Calculation of accuracy.**
**a)** Basic definition and **b)** considering the imbalance of a dataset.

Other metrics, such as precision, recall and F-score, are commonly calculated, as well (Eq. 23). Precision defines how accurate the model is on values predicted as positive. It is useful in cases where the aim is to avoid false positive values, *e.g.*, in spam detection for which it is unfavourable to mark important e-mails as spam. On the other hand, recall calculates how many actual positive values were predicted as positive. It is useful if false negative events must be prevented. For example, predicting a sick patient as healthy is disadvantageous. The F-score is the function of these two metrics (precision and recall); thus, it helps to keep a balance between precision and recall. To assess the performance of the model, significance testing of these metrics is important to see if the model is better than random chance[6], [11].

$$\mathbf{a})\ Precision = \frac{TP}{TP + FP}$$

$$\mathbf{b})\ Recall = \frac{TP}{TP + FN}$$

$$\mathbf{c})\ F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

**Equation 23. Some commonly used metrics.**
**a)** Precision, **b)** recall, **c)** F-score.

In all learning types (supervised, unsupervised, semi-supervised, reinforcement), many different models exist for training. It is not always obvious which model fits better to a given problem. In the next sections, different training models are described.

### III.2.3. Unsupervised training models

The algorithms presented in this section are dimensionality reduction and/or clustering/classifying techniques. The goal of ***dimensionality reduction*** is to result in a new, smaller (reduced) dataset that has fewer dimensions than the original one. At the same time, this new dataset carries the information contained in the original data. However, the new dataset will never totally contain the same information, as unimportant features are not taken

into account. But many models are available that provide a good approximation and keep the high variance of a dataset. The goal of ***clustering*** is to group the data points according to their similarity using a measure (similarity criterion). To run a clustering algorithm, feature (criteria) and similarity metric should be selected. Finally, the model should be validated. Often, the expected group number should be pre-defined. The success of clustering depends on the feature selection and the expected group number. A clustering algorithm can be partitional or hierarchical. In the latter, it is possible to divide a group into sub-groups. It is possible to assign a data point to a class (hard clustering) or calculate the probability of belonging to a group (soft clustering). Clustering is commonly used in molecular dynamics simulations when a criterion has a multipeak density distribution. It can be used to find some patterns in the dataset, but the separation between groups is sometimes unclear. Clustering performs dimensionality reduction, as well [8]. In this section, algorithms/methods to run dimensionality reduction or clustering will be described.

### Dimensionality reduction by Principal Component Analysis

Principal Component Analysis (PCA) performs a linear mapping of input features to a low-dimensional representation, *i.e.*, linearly uncorrelated variables called principal components (PCs). These PCs are orthogonal to each other. PCA is the most often used dimensionality reduction algorithm. Its performance is highly dependent on the input coordinates [3], [6], [8].

A data matrix ($M$) contains the coordinates for all frames; thus, its dimension is $M \times N$, where $M$ is the number of frames, and $N$ is three times (corresponding to (x, y, z) coordinates) the number of atoms. To guarantee the translational invariance, a zero-mean data matrix ($M'$) is created by subtracting the mean of the original matrix. In the next step, the covariance matrix ($C$) of the data is estimated (Eq. 24a). Diagonalization of this covariance matrix (Eq. 24b) provides the eigenvectors ($\boldsymbol{v}_\alpha$) and eigenvalues ($\lambda_\alpha$). Eigenvalue $\lambda_\alpha$ represents the variance of the data along the direction defined by eigenvector $\boldsymbol{v}_\alpha$ [3], [6], [8].

$$\mathbf{a}) \; C = \sum_{i=1}^{n} \frac{(x_i - \tilde{x})(y_i - \tilde{y})}{n-1}$$

$$\mathbf{b}) \; C\boldsymbol{v}_\alpha = \lambda_\alpha \boldsymbol{v}_\alpha, \qquad \lambda_1 > \lambda_2, \ldots > \lambda_n$$

**Equation 24. Covariance matrix (a) and its diagonalization (b).**

The eigenvectors of the most relevant eigenvalues are the principal components on which the projection happens. The first principal component represents the most varying direction, the second principal component corresponds to the second most important motion, *etc.* As the last principal components only represent a small variance in the dataset, they can be omitted, *i.e.*, dimensionality reduction is achieved. It is important to know which PCs are the most relevant. If there is a gap between their variance, that can be an obvious choice. Although, usually, there is no observable gap. In this case, the sum of variances should reach a given fraction which is often set to 0.95-0.98. The dimensionally reduced output (linear combinations of the original dimensions) is produced by a matrix ($P$) containing the eigenvectors for the highest eigenvalues (Eq. 25) [3], [6], [8].

$$\boldsymbol{M'P = T}$$
$$P = [[v_1] \quad [v_2] \quad \cdots]$$

**Equation 25. Dimensionally reduction step in PCA.**

Using PCA, the dimension of a dataset can be greatly reduced. Though, in some cases, the interpretation of the results can be difficult as it is not biologically plausible. Despite some difficulties, PCA is widely and successfully used in theoretical chemistry and biology [3], [6], [8].

### Clustering by k-means

K-means is one of the most known clustering algorithms. "K" starting cluster centres should be chosen, then each data point is assigned to the nearest cluster based on a metric (*e.g.*, Euclidean[6] or Manhattan[7] distances). The centres are redefined based on the centroid of each cluster and the data points are again assigned to the clusters. This is iteratively done via the k-means clustering algorithm, until a threshold or a given number of iterations is reached. In other words, the algorithm learns which data point belongs to which cluster. The definition of k-means does not guarantee that the global minimum is reached. Another drawback of k-means is that it always results in "k" clusters, even if it does not have a biological sense. Moreover, the clusters cannot be too much different in shape. A Gaussian Mixture Model based (see section III.3.1) or density-based clustering gives better results for noisy datasets, for example [6], [8].

### Density-based clustering

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) relays on local density. Distances within a group are smaller than between groups. However, a threshold to eliminate noise should be defined. The choice of a threshold is not always trivial. A. Rodriguez and A. Laio have proposed another but similar algorithm to find density peaks. For each point, a local density and its distance from higher density points are calculated. Points with density maxima will be selected as cluster centres and all the other points will be assigned to the nearest cluster. Points with high distance but low density are called outliers which are single points in a cluster. This algorithm does not require any parameter definition; thus, it does not depend on a good choice of parameters [6], [12].

## III.2.4. Supervised training models

The algorithms presented in this section are classification or regression techniques. The goal of ***classification*** is to predict group membership, *i.e.*, label the output data using a trained decision function (classifier). A classification algorithm can be linear or high dimensional, but both result in a discrete output. The goal of ***regression*** is to analyse the relationship between two or more variables by providing a score on a continuous scale. Regression algorithms are good to describe and test relationships between an output and some input variables. The algorithm learns a regression function that predicts an output from input predictors [6], [13]–

[6] Euclidean distance of $A(x_a, y_a)$ and $B(x_b, y_b)$ is $\sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}$

[7] Manhattan distance of $A(x_a, y_a)$ and $B(x_b, y_b)$ is $|x_b - x_a| + |y_b - y_a|$

[15]. In this section, algorithms/methods to run dimensionality reduction or clustering will be described.

### Classification by Support Vector Machine

A Support Vector Machine (SVM) is a linear classifier that can distinguish between two classes. This algorithm is searching for a hyperplane (Eq. 26) that maximizes the distance (margin) between the support vectors of the class labels (Fig 23). It requires a set of features as input. Feature selection can be done *a priori* (filter method), or it can be part of the classifier (embedded and wrapped methods). In the first case, a dimensionality reduction algorithm is run *a priori*, while in the second case, the SVM can be used as a dimensionality reduction algorithm [6], [13].

$$\boldsymbol{y = 1[w^T X + b > 0]}$$

$y$ – output label, $X$ – high dimensional input vector,
1 – indicator function, $w$ – learnt vector of coefficients
$b$ – scalar bias

**Equation 26. The equation of a hyperplane for two features.**

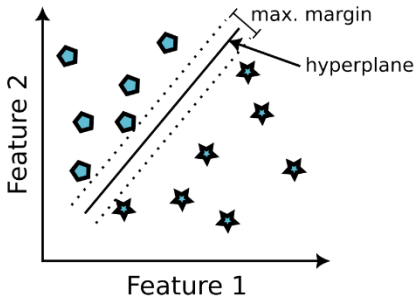



**Figure 23. 2-Dimensional illustration of SVM [6].**

The hyperplane is often not curved for three or more features either. However, for more difficult classification problems, a curved hyperplane is necessary. In this case, a kernel method is required to transform the support vectors to a higher-dimensional input space [6], [13].

### Logistic regression

Logistic regression is an advanced version of linear regression, and it is similar to multiple linear regression but with a binomial response. Logistic regression gives the probability of an outcome based on the inputs by the natural logarithm of an odds ratio (Eq. 27). It maximizes the probability of being in one class or another [13]–[15].

$$\boldsymbol{ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m}$$

$\pi$ – probability (output is the outcome of interest | $X_1$=$x_1$,$X_2$=$x_2$,…),
$\beta_0$ – reference level, $\beta_i i = 1, \ldots, m$ – regression coefficient, $x$ – input

**Equation 27. Logistic regression.**

$$y = C + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

$C$ – intercept, constant, $\beta$ – coefficient, $\varepsilon$ – noise

**Equation 28. Multiple linear regression with "n" predictors.**

### III.2.5. Neural networks

Depending on the model, neural network-based algorithms can learn from labelled or unlabelled data, as well. In other words, both supervised and unsupervised learning is possible using neural networks (NNs). NNs are inspired by the neurons of the brain and their connections. Neurons are long, asymmetric cells. They have many dendrites to collect the input information, but only one axon to forward the answer (Fig. 24a). The schematic neuron in neural networks also has many inputs but only one output (Fig. 24b). The inputs ($x_1, \ldots, x_n$) have a weight ($w_1, \ldots, w_n$) that indicates the strength and direction between two neurons. The output ($y$) is the weighted sum of the inputs passing through a nonlinear differentiable function (*e.g.*, sigmoid), called activation function (Eq. 29) [6], [16], [17].

$$y = f\left(\sum_{i=1}^{n} w_i x_i + bias\right)$$

**Equation 29. The output of a neural network.**

As for the neurons in the brain, a signal is given above a threshold. Neurons are connected; thus, the output of one neuron is the input of another one. So that, they build layers. Neurons in the brain are highly connected. Traditionally, NNs have fully connected layers – input, hidden, and output layers (Fig. 24c). In other words, each neuron is connected to every neuron in the previous layer. This can result in thousands of parameters to set; thus, the neural networks of the current techniques are often not fully connected [6], [16], [17].

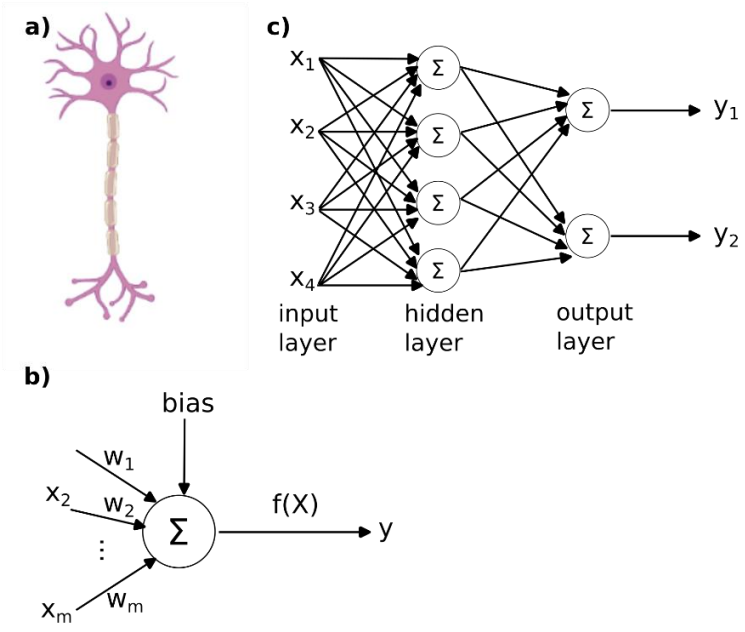



**Figure 24. Basics of neural networks.**

**a)** A neuron, **b)** a schematic neuron, and **c)** a neural network called multilayer perceptron.

### III.2.5.1. Multilayer perceptrons

Multilayer perceptrons (MLPs) are fully connected neural networks with at least one hidden layer (Fig. 24 c). In other words, there is at least one layer in the neural network between the input and the output layers. Information from the inputs is forwarded (propagated) through the hidden layers until the output layer is reached. An error is estimated by applying a so-called loss (cost) function using the calculated output and the expected output. This results in a real number that represents the error associated with the output. This error is back propagated by applying the chain rule[8] during the derivation of the activation function, and the initial weights are recalculated in each iteration of the algorithm. Thus, MLP belongs to the supervised methods.

The construction of an MLP makes possible to separate non-linearly separable outputs. Most of the biological features are non-linearly separable, as well. A typical example of a non-linearly separable problem is the XOR logistic function where the output is not separable by one linear line (Fig. 25) [3], [19].

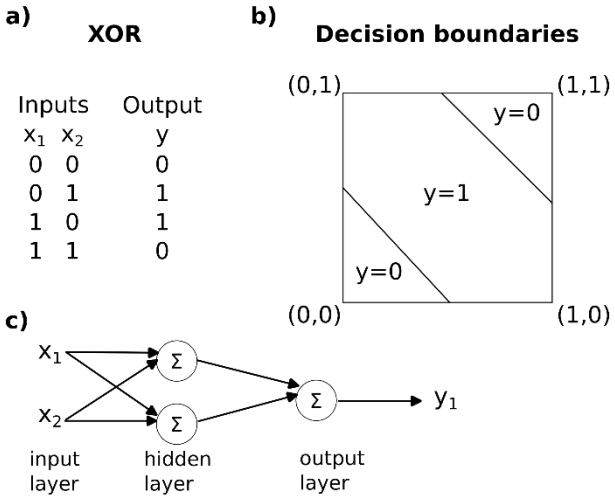



**Figure 25. Solving the XOR problem by MLP.**
**a)** Truth table, **b)** decision boundaries, and **c)** multilayer perceptron for XOR function.

### III.2.5.2. Autoencoders

Autoencoders (AEs) are unsupervised neural network-based ML models. They map the input to a lower dimensional data representation. AEs were developed as the non-linear extension of the linear PCA. An AE is built by two neural networks: an encoder and a decoder (Fig. 26). The encoder maps the input data ($x$) to a latent encoding space that has a smaller dimension than the input (called undercomplete autoencoder). In some cases, *e.g.*, for biomolecular simulation trajectories, the size of the input should be reduced before passing to the encoder. The decoder creates a reconstruction of the input from this latent space. This is the output ($x'$) of an AE. A specific loss function gives a penalty if the output differs from the input. This function is called reconstruction loss and it is often the mean squared error loss function[9]. During the training, a function, that can reconstruct the input data after a dimensionality

[8] Chain rule: $(fog)'(x) = f'(g(x))g'(x)$ [18]
[9] $Loss = \|x - x'\|^2$

reduction, is learnt. Thus, after the learning, the resulting code (***h***) can be used *e.g.*, for dimensionality reduction as PCA [3], [19].

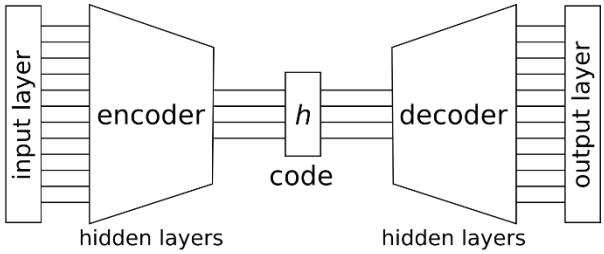



**Figure 26. A schematic autoencoder.**

### III.3. ML in theoretical chemistry

ML-based techniques become essential in computational chemistry. They have numerous advances that were out of reach because of the computational complexity. As mentioned in the previous Chapter, ML is more and more present in FF definition, as well. In this section, the focus is on those techniques that can be used to solve the CV definition problem.

#### III.3.1. Techniques to solve the CV definition problem

Many ML-based algorithms/methods were proposed to help to solve the definition of collective variables [3], [13], [20]–[22]. Most of them are based on one of the previously described algorithms (section III.2). In the next sections, two approaches are presented. As for all ML-based approaches, a relatively big amount of input data is required for the learning procedure of ML-based CV definition algorithms, as well. This means that an *a priori* guess or unbiased simulations in the states of the biological process is necessary. The idea is often to train the model on many possible driving forces to decipher the most important ones. This is dimensionality reduction of features that can be used for both analysis and CV definition. It is still not totally obvious which method suits best for which kind of system, but Fleetwood *et al.* benchmarked many algorithms that are good to demystify complex simulations. Moreover, a checklist for interpreting molecular simulations with ML is proposed (Fig. 27) [3], [8].

##### III.3.1.1. Demystifying

Demystifying, published by Fleetwood and co-workers in 2020, is a branch of advanced dimensionality reduction (analysis) techniques that can be used for CV definition, as well. From the many tested algorithms (principal component analysis – PCA, random forests – RFs, autoencoders – AEs, restricted Boltzmann machines – RBMs, and multilayer perceptrons – MLPs), only the supervised algorithms work well for more complicated inputs (*e.g.*, Cartesian coordinates) than optimal inverse interatomic distances. This statement is based on the results of a toy model and three biological processes on which the evaluation of the tested algorithms was performed. The evaluation was carried out using a metric based on the mean squared error between the estimated and the true feature importance profiles. The false positive values produced by a given method were also measured. All methods were tested using both Cartesian coordinates and internal coordinates, and on the full and a reduced feature set, as well. In general, the choice of the input features has a larger effect on the performance of the tested algorithms than the choice of the parameters of a given algorithm [3].

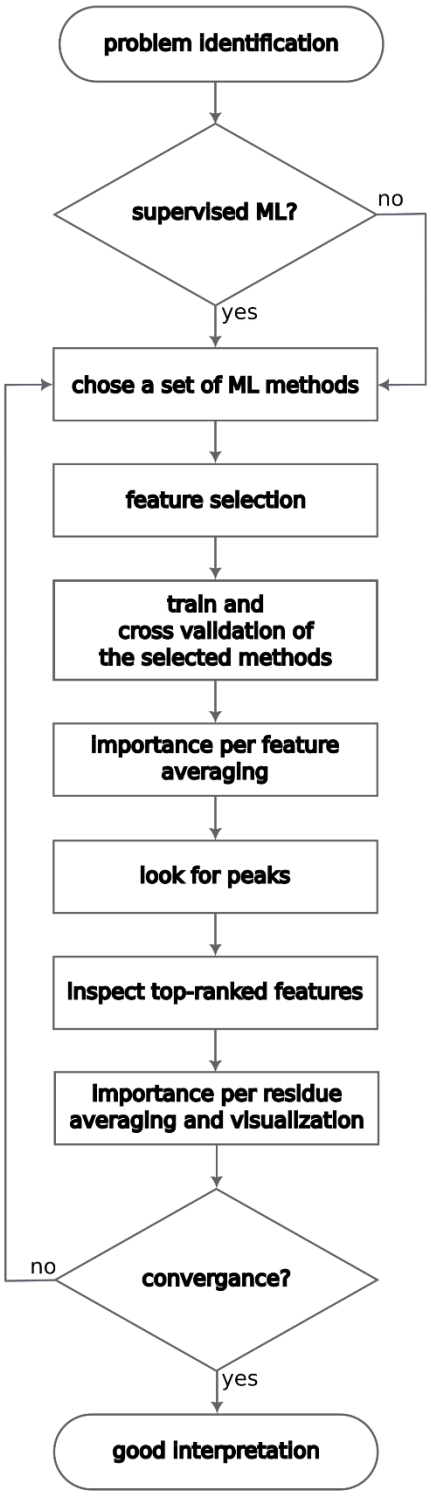



**Figure 27. Flowchart of the choice of ML technique for interpreting molecular simulations.**
[Based on Box 1 of [3].]

### III.3.1.2. Automated design of CVs

Sultan *et al.* proposed an ML-based automated design to solve the problem of selecting CVs for enhanced sampling simulations. This method suits well for those kinds of systems where the start and end states of the biological system are known. Therefore, the first step is to run short (ns up to μs long) unbiased simulations in both states. On the resulted trajectories, features (*e.g.*, dihedrals, distances) are calculated and used to train a classifier to distinguish between the two states. The classifier can be *e.g.*, a support vector machine (SVM), where the signed closest distance to its hyperplane is a possible CV. The differentiable probability output of logistic regression is also a good CV candidate. Finally, Sultan *et al.* propose to use the function learned by a neural network-based method as a CV [13]. These algorithms are promising, but *a priori* knowledge about the system is necessary for the right feature selection.

## III.4. Advanced analysis techniques

Even though not all analysis techniques use an ML approach, unsupervised ML techniques (section III.2.3) are important to analyse molecular dynamics simulations. Many state-of-the-art analysis techniques are typical cases of unsupervised algorithms as their goal is to find interesting information from a huge amount of data, from the features calculated on the trajectory. So, the output is unknown [8].

### III.4.1. Inflection core state clustering

Inflection core state (InfleCS) is a free energy clustering method along some pre-defined reaction coordinates (CVs). The goal is to find the metastable states in free energy minima based on the data, as *a priori* knowledge about the system is often not available. Clustering based on geometric criteria cannot find the boundaries well. Density-based clustering depends too much on the basis function. Gaussian mixture models (GMMs) do not always find the correct number of clusters and core boundaries. InfleCS provides metastable core states based on the density second-order derivative values from a GMM density estimation (Fig. 28). Each point on this free energy curve is either transition or metastable state which is based on the value of the second derivative. GMM is the sum of Gaussians with amplitude ($\boldsymbol{a}$), mean ($\boldsymbol{\mu}$) – centre – and covariance ($\boldsymbol{\Sigma}$) – width – values (Eq. 30). Gaussian parameters are optimized iteratively using expectation-maximization that is a maximum likelihood estimation suited for incomplete dataset. To find the final model, multiple GMMs are fit and evaluated in a range of the number of Gaussian components. The one with the smallest Bayesian information criterion is selected as the final model. InfleCS does not use a lot of assumptions about cluster shapes or dataset structure which is a big advantage [23], [24].

$$\boldsymbol{\rho}_{\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{N_{basis}}} \boldsymbol{a_i}\boldsymbol{\mathcal{N}}(\boldsymbol{x}|\boldsymbol{\mu_i},\boldsymbol{\Sigma_i})$$

$$a = (a_i)_{i=1}^{N_{basis}},\ \mu = (\mu_i)_{i=1}^{N_{basis}}$$

$$\Sigma = (\Sigma_i)_{i=1}^{N_{basis}},\ \mathcal{N}(x|\mu_i,\Sigma_i) = \frac{1}{\sqrt{(2\pi)^{N_{dims}}|\Sigma_i|}} e^{\hat{f}_i}$$

$$\hat{f}_i = -\frac{(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)}{2}$$

**Equation 30. Gaussian mixture model [23].**

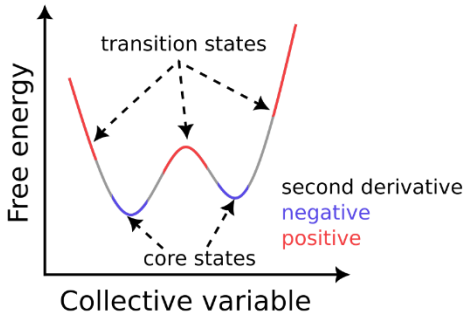



**Figure 28. 1D example of InfleCS clustering.**
[Based on Figure 1. of [23].]

### III.4.2. Allosteric pathway analysis (Allopath)

Membrane proteins are modulated by many stimuli coming from the surrounding environment that often has an impact on distinct sites from the functional site. This is called allosteric modulation. Moreover, lipids and small molecule-binding (together called cofactors) might contribute to this allosteric communication. Allostery is an important factor in membrane protein functions. Residue interaction networks can be built for isolated proteins. However, it does not take into account the dynamics of proteins and the constant movement of the surrounding cofactors. Cofactors should be translated into a static network (nodes and edges) by maintaining the network while perpetually reassigning the cofactors to the nodes in each frame of the MD trajectory. The dynamics of the protein is included in a contact map which is the average of continuous contact maps over frames. A mutual information matrix ($M$) describes the correlation of node movement around an equilibrium position (Eq. 31). Gaussian mixture model is used to get the over frames mutual information [25], [26].

$$\boldsymbol{M_{ij} = H_i + H_j - H_{ij}}$$
$i, j$ – nodes, $H_i, H_j, H_{ij}$ – entropies

**Equation 31. Mutual information calculation.**

A network (adjacency matrix – $A$) is built by the elementwise product of the contact map ($C$) and the mutual information matrix ($M$)[10]. Current/information flow betweenness and closeness are used to find the most important residues in the allosteric communication between source and sink nodes. These measures come from electrical networks. The information flow betweenness measures how much a node is involved in the connection of two other nodes (source and sink) in the allosteric path. The information flow closeness shows the distance of a node to other nodes (source and sink). To calculate these values for an average of paths, an inverse reduced Laplacian is used. The inverse reduced Laplacian is a matrix that has zeros in the rows and columns for the sink nodes in the inverse of a Laplacian (diagonal matrix minus adjacency matrix) not including the rows and columns of the sink nodes[11] [25], [26].

[10] Adjacency matrix: $A_{ij} = C_{ij}M_{ij}$

[11] Inverse reduced Laplacian: $\tilde{L}^{-1} = ((D - A) - sink\_nodes)^{-1} + zero\_for\_sink\_nodes$

## References


[1] *BioExcel Webinar #39: Accelerating sampling in GROMACS with the AWH method*, (2019). Accessed: Sep. 07, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=SN-Ompgp-mE

[2] *BioExcel Webinar #54: Applying the Accelerated Weight Histogram method to alchemical transformations*, (2021). Accessed: Sep. 07, 2022. [Online Video]. Available: https://www.youtube.com/watch?v=E5nGLcbyqTQ

[3] O. Fleetwood, M. A. Kasimova, A. M. Westerlund, and L. Delemotte, 'Molecular Insights from Conformational Ensembles via Machine Learning', *Biophys J*, vol. 118, no. 3, pp. 765–780, Feb. 2020, doi: 10.1016/j.bpj.2019.12.016.

[4] E. Rich and K. Knight, *Artificial Intelligence*, 2nd ed. McGraw-Hill.

[5] O. T. Unke *et al.*, 'Machine Learning Force Fields', *Chem Rev*, vol. 121, no. 16, pp. 10142–10186, Aug. 2021, doi: 10.1021/acs.chemrev.0c01111.

[6] A. Mechelli and S. Vieira, *Machine Learning: Methods and Applications to Brain Disorders*, First edition. Academic Press, 2019.

[7] N. Artrith *et al.*, 'Best practices in machine learning for chemistry', *Nat Chem*, vol. 13, no. 6, pp. 505–508, Jun. 2021, doi: 10.1038/s41557-021-00716-z.

[8] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, 'Unsupervised Learning Methods for Molecular Simulation Data', *Chem Rev*, vol. 121, no. 16, pp. 9722–9758, Aug. 2021, doi: 10.1021/acs.chemrev.0c01195.

[9] A. Deshpande and M. Kumar, *Artificial Intelligence for Big Data: Complete guide to automating Big Data solutions using Artificial Intelligence techniques*. Birmingham Mumbai: Packt Publishing, 2018.

[10] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, 'Machine learning for molecular and materials science', *Nature*, vol. 559, no. 7715, pp. 547–555, Jul. 2018, doi: 10.1038/s41586-018-0337-2.

[11] M. Grootendorst, 'Validating your Machine Learning Model', *Medium*, May 15, 2020. https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7 (accessed Sep. 06, 2022).

[12] A. Rodriguez and A. Laio, 'Clustering by fast search and find of density peaks', *Science*, vol. 344, no. 6191, pp. 1492–1496, Jun. 2014, doi: 10.1126/science.1242072.

[13] M. M. Sultan and V. S. Pande, 'Automated design of collective variables using supervised machine learning', *J Chem Phys*, vol. 149, no. 9, p. 094106, Sep. 2018, doi: 10.1063/1.5029972.

[14] S. Sperandei, 'Understanding logistic regression analysis', *Biochem Med*, pp. 12–18, 2014, doi: 10.11613/BM.2014.003.

[15] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, 'An Introduction to Logistic Regression Analysis and Reporting', *J Educ Res*, vol. 96, no. 1, pp. 3–14, Sep. 2002, doi: 10.1080/00220670209598786.

[16] Fonyó A. and Ligeti E., *Az orvosi élettan tankönyve*, Fourth edition. Budapest: Medicina Könyvkiadó Zrt., 2008.

[17] U. Alon, *An Introduction to Systems Biology*, 2nd edition. Boca Raton, Fla: Chapman and Hall/CRC, 2019.

[18] Z. Vágó, 'Matematikai analízis I.' Pázmány Egyetem eKiadó, 2010.

[19] S. S. Haykin, *Neural networks and learning machines*, 3rd ed. New York: Prentice Hall, 2009.

[20] F. Noé, G. De Fabritiis, and C. Clementi, 'Machine learning for protein folding and dynamics', *Curr Opin Struct Biol*, vol. 60, pp. 77–84, Feb. 2020, doi: 10.1016/j.sbi.2019.12.005.

[21] A. Mardt, L. Pasquali, H. Wu, and F. Noé, 'VAMPnets for deep learning of molecular kinetics', *Nat Commun*, vol. 9, no. 1, p. 5, Dec. 2018, doi: 10.1038/s41467-017-02388-1.

[22] D. Mitrovic, S. E. McComas, C. Alleva, M. Bonaccorsi, D. Drew, and L. Delemotte, 'Reconstructing the transport cycle in the sugar porter superfamily using coevolution-powered machine learning'. bioRxiv, p. 2022.09.24.509294, Sep. 26, 2022. doi: 10.1101/2022.09.24.509294.

[23] A. M. Westerlund and L. Delemotte, 'InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures', *J Chem Theory Comput*, vol. 15, no. 12, pp. 6752–6759, Dec. 2019, doi: 10.1021/acs.jctc.9b00454.

[24] J. Brownlee, 'A Gentle Introduction to Expectation-Maximization (EM Algorithm)', *Machine Learning Mastery*, Oct. 31, 2019. https://machinelearningmastery.com/expectation-maximization-em-algorithm/ (accessed Sep. 07, 2022).

[25] A. M. Westerlund, O. Fleetwood, S. Pérez-Conesa, and L. Delemotte, 'Network analysis reveals how lipids and other cofactors influence membrane protein allostery', *J Chem Phys*, vol. 153, no. 14, p. 141103, Oct. 2020, doi: 10.1063/5.0020974.

[26] P. W. Kang *et al.*, 'Calmodulin acts as a state-dependent switch to control a cardiac potassium channel opening', *Sci Adv*, vol. 6, no. 50, p. eabd6798, Dec. 2020, doi: 10.1126/sciadv.abd6798.

## Chapter IV. ABCC family: MRP1 in the focus

### IV.1. On the interplay between lipids and asymmetric dynamics of an NBS degenerate ABC transporter

Ágota Tóth, Angelika Janaszkiewicz, Veronica Crespi and Florent Di Meo*

Inserm UMR 1248 Pharmacology & Transplantation, Univ. Limoges, 2 rue du Prof. Descottes, 87000 F-Limoges, France

**Contact information**
Dr. Florent Di Meo
Inswem U1248 Pharmacology & Transplantation, Univ. Limoges,
2 rue du Prof. Descottes,
87000 F-Limoges, France
florent.di-meo@inserm.fr
Tel: +33(0)5 19 56 42 76

**Submitted, bioRχiv, DOI: 10.1101/2022.05.16.492073v1**

This study is focusing on ABCC1/MRP1. We investigated the ABC conformational space such as where our simulations take place compared to several resolved ABC structures. We have also studied the dynamics of MRP1 protein in different membranes. We have seen the importance of the degenerate NBS in MRP1.

To note, this work is under major revision and is expected to be resubmitted soon.

**Abstract**

Multidrug resistance-associated proteins are ABC C-family exporters. They are crucial in pharmacology as they transport various substrates across membranes. However, the role of the degenerate nucleotide-binding site (NBS) remains unclear likewise the interplay with the surrounding lipid environment. Here, we propose a dynamic and structural overview of MRP1 from *ca.* 110 μs molecular dynamics simulations. ATP binding to NBS1 is likely maintained along several transport cycles. Asymmetric NBD behaviour is ensured by lower signal transduction from NBD1 to the rest of the protein owing to the absence of ball-and-socket conformation between NBD1 and coupling helices. Even though surrounding lipids play an active role in the allosteric communication between the substrate-binding pocket and NBDs, our results suggest that lipid composition has a limited impact, mostly by affecting transport kinetics. We believe that our work can be extended to other degenerate NBS ABC proteins and provide hints for deciphering mechanistic differences among ABC transporters.

#### IV.1.1. Introduction

ATP-binding cassette (ABC) transporters belong to one of the largest trans-kingdom protein superfamilies. The structural resolution of several ABC transporters has led to different conformations (namely, inward-facing – IF, outward-facing – OF**,** and occluded), which illustrate the alternating access as the most likely model to rationalize substrate translocation along the transport cycle[1–3]. ABC transporter structures are made of at least two transmembrane domains (TMDs) consisting of six transmembrane helices (TMHs). TMDs are

bound to two nucleotide-binding domains (NBDs) which are evolutionarily conserved over species. ABC transport cycle requires the binding of two ATP molecules and the energy released from the hydrolysis of at least one of them[3–7]. ATP molecules bind at the interface of the NBD dimer which adopts a non-covalent pseudo-symmetric head-to-tail arrangement; enabling the formation of two nucleotide-binding sites (NBSs). Both NBSs are formed with the conserved Walker A- and B-motifs, the A-, Q- and H-loops of one NBD and the ABC signature sequence and the X-loop of the other NBD[4,5,7,8].

Except for a few members (CFRT, ABCA4, ABCD4, SUR1/2)[2,3,9,10], eukaryotic ABC transporters are exporters, *i.e.*, they extrude substrates to the extracellular compartment. Eukaryotic ABC transporters used to be classified into type I (ABCB, ABCC, ABCD) and type II (ABCA, ABCG) families. Recently, the structural and functional diversities of ABC transporters have led to a new folding-based classification[2,3] in which the previous type I and type II exporters adopt the type IV and V folding, respectively.

Multidrug resistance-associated proteins (MRPs) are NBS degenerate ABC transporters[3–5,7,8]. In the non-canonical NBS1, the Walker-B catalytic glutamate, the A-loop tyrosine, and the first glycine residue of the ABC-signature motif are mutated into aspartate, tryptophan, and valine residues, respectively. These mutations were associated with significantly lower ATPase activity and higher ATP-binding affinity for the degenerate NBS1[4,7]. These observations have recently led to the development of a new asymmetric model for NBD function which may affect the dynamics and function of the whole transporter[3,7]. The function and kinetics of bovine ABCC1/MRP1 (*b*MRP1) were extensively and thoroughly investigated by combining structural information from cryo-EM experiments and single-molecule Förster Resonance Energy Transfer (smFRET)[5]. Despite the robust insights provided by the resolution of *b*MRP1 structure, unexplained differences between ABCB and ABCC exporters were observed, likely due to the non-native detergent-based environment used for *b*MRP1 experiments[7,11].

ABC transporters play a crucial role in pharmacology by transporting a tremendous variety of substrates, including xenobiotics and endogenous compounds, across cell membranes. For instance, their pharmacological role has been stressed by the International Consortium Transporter (ITC) which draws out a list of transporters of “emerging clinical importance” for which interactions with new xenobiotics have to be investigated in drug development[12]. Over the past decades, ABCC transporters, in which MRPs are included, have gained a growing interest owing to their role in pharmacology including patient inter-individual responses to treatments. For instance, investigations on ABCC2/MRP2 and ABCC4/MRP4 have been recommended by the ITC to retrospectively provide a mechanistic explanation of clinical observations regarding drug dispositions[13]. Given the role of MRPs in local pharmacokinetics and pharmacodynamics relationships (PK/PD) and therefore in local drug bioavailability[14], there is still a need to decipher *in situ* MRP transport cycle to provide a comprehensive overview of xenobiotic membrane crossing events. This is particularly relevant for MRPs located in proximal tubular kidney cells and liver hepatocytes since kidneys and liver are involved in the elimination of most worldwide used xenobiotics[15].

Unfortunately, there is no experimentally resolved structure for human MRPs yet. An MD-refined protein threading MRP4 structure[16] has been proposed for which the computational resolution precludes functional investigations or thorough structural dynamics. However, several conformations of *b*MRP1 transporter have been resolved by cryogenic electron microscopy (cryo-EM)[4,5,8]. Given the high sequence similarity to the human ortholog *h*MRP1 (91%) as well as within other human MRPs (*ca.* 40-50%), the use of *b*MRP1 structures appears

relevant for investigating MRP transporter dynamics and functions[3]. MRP1 exporter adopts the type IV folding, however, it has an extra N-terminal TMD made of five TMHs. The so-called $TMD_0$ was shown not to play a role either in ABC ATPase activity or substrate transport[3,4,8]. Therefore, a $TMD_0$-less MRP1 model can be used as a prototype for ABCC exporters even for those which do not possess this domain (*e.g.*, MRP4). $TMD_0$ is connected to conventional ABC TMDs by a linker ($L_0$) which was shown to be mandatory for both trafficking and function[8,17]. The $L_0$ sequence is conserved in all members of the ABCC subfamily even in absence of $TMD_0$[8]. It is important to note that since the present model does not include $TMD_0$, the standard TMH labelling for ABC type IV will be used in the present manuscript, *i.e.*, TMH1 to TMH12.

The present work aims to map the ABC conformational space[1,18] of MRP1 considering different bound states (namely, apo, ATP and/or leukotriene C4 – LTX – bound states[8] for IF conformation and ATP bound state[4] for OF conformation) highlighting the importance of asymmetry in ABC domains. Furthermore, given the importance of surrounding lipids in the ABCC transport cycle[11,19], the interplay between the lipid bilayer membrane and protein dynamics was also investigated. This was achieved by using different computational symmetric membrane models made of (i) pure POPC (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine), (ii) pure POPE (1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine), (iii) POPC:POPE (3:1), (iv) POPC:Chol (3:1) and (v) POPC:POPE:Chol (2:1:1); the last being the closest to mimic *in situ* MRP1 dynamics. All-atom unbiased microsecond-scaled molecular dynamics (MD) simulations were conducted to address the objectives.

### IV.1.2. Results

#### Projection of MRP1 onto the ABC conformational space

To examine the conformational space sampled during the simulations in POPC:POPE:Chol (2:1:1) accounting for bound states of *b*MRP1, different structural descriptors[1,11,18] were considered. Namely, intracellular (IC) and extracellular (EC) angles were monitored for TMDs while NBD distance and NBD rocking-twist angle were used for NBDs (Fig. 29 and Supplementary Fig. 1-6). These structural descriptors were also measured on a large data set of experimentally resolved ABC proteins including *b*MRP1 cryo-EM structures[4,5,8] in order to reconstruct the ABC conformational space[1,18] (Fig. 29d and Supplementary Table 1). The known ABC conformations were efficiently featured, *i.e.*, IF open and occluded, OF as well as the recently resolved asymmetric unlock-returned (UR) turnover[1] conformations. *b*MRP1 MD simulations revealed the spontaneous closing of the intracellular cavity for all IF systems regardless of the bound state and the membrane composition (Fig. 29a-b, Supplementary Fig. 1-6 and Supplementary Tables 2-5). Average NBD distances and IC angles were significantly smaller than those calculated for *b*MRP1 cryo-EM structures. For instance, IF conformations converged toward similar IC angle and NBD distance values ranging from 26.0 ± 0.5 to 35.4 ± 1.8 and from 30.3 ± 1.9 to 55.0 ± 2.8 Å, respectively (Fig. 29b, Supplementary Fig. 1, 3-4, and Supplementary Tables 3-4). Interestingly, the spontaneous dimerization of NBDs was also observed in absence of ATP molecules. The difference between simulations and cryo-EM structure might be explained by the use of non-native detergent in experiments[7,11].

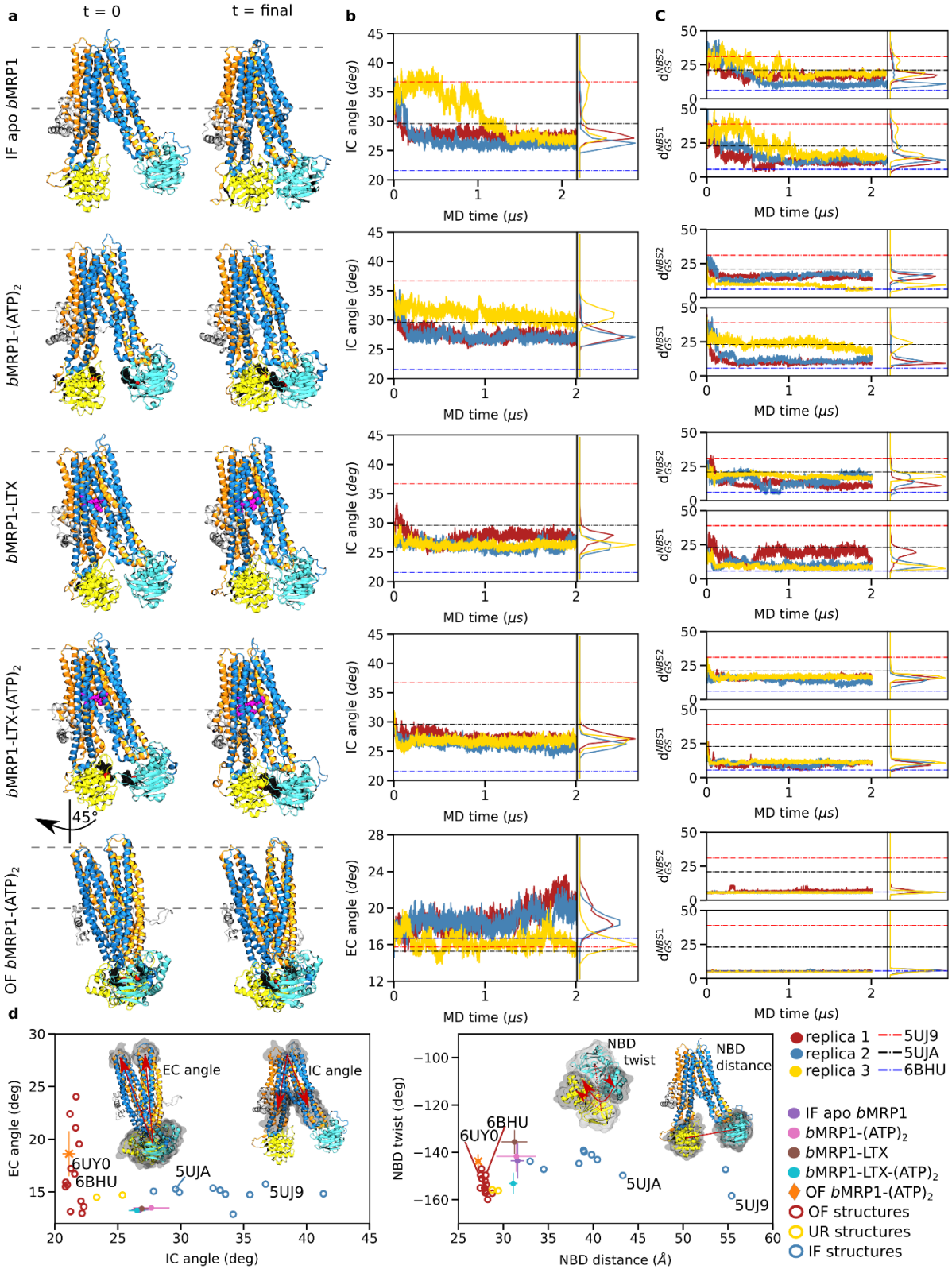



**Figure 29. Overview of *b*MRP1 milestone structures along the transport cycle in POPC:POPE:Chol (2:1:1).**

*Inward-facing conformations:* apo-state (IF apo *b*MRP1), substrate-bound (*b*MRP1-LTX), ATP-bound (*b*MRP1-$(ATP)_2$), ATP/substrate-bound (*b*MRP1-LTX-$(ATP)_2$); *Outward-facing conformation:* ATP-bound (OF *b*MRP1-$(ATP)_2$). **a)** Representative snapshots of the different *b*MRP1 systems investigated here at the beginning (left) and the end of MD simulations. **b)** Time-evolution of the IC and EC angles respectively for IF and OF structures. IC and EC angles were calculated according to the proposed ABC structural parameters defined in Methods

section[1,11,18]. **c)** Time-evolution of key NBS distances defined by inter-NBD distances between Walker A Glycine and ABC signature serine residues[8]. **d)** Projection of *b*MRP1 structural parameters, each replica was averaged over the last 800 ns, onto the ABC conformational space obtained from multiple resolved ABC structures. PDB IDs of the resolved *b*MRP1 cryo-EM structures are explicitly mentioned. The first and second TMDs are respectively depicted in orange and blue, and NBD1 and NBD2 are respectively coloured yellow and cyan.

Indeed, resolved cryo-EM structures of *b*MRP1 exhibited larger IC angles and NBD distances than other ABC structures resolved in native environments (Fig. 29d). On the extracellular side of the lipid bilayer membrane, MD simulations of OF *b*MRP1-$(ATP)_2$ displayed minor openings of the EC gate as compared to *b*MRP1 OF cryo-EM structures, suggesting the existence of a slightly more open state. However, calculated EC angles remained small (lower than 20° in POPC:POPE:Chol (2:1:1), see Fig. 29a and Supplementary Fig. 1-2 and Supplementary Table 2), which precludes the substrate re-entry[1].

Even though trajectories performed with the pre-translocation state (*i.e.*, *b*MRP1-LTX-$(ATP)_2$) tend to populate the conformational subspace of OF conformation (Fig. 29d), substrate translocation was not observed. NBD twist values calculated for *b*MRP1-LTX-$(ATP)_2$ state are similar to OF ABC conformations. However, NBD distances remain larger than for resolved substrate-free OF ABC structures and OF *b*MRP1-$(ATP)_2$ simulations. Distances between C$\alpha$ atoms of the ABC signature motif serine and a Walker A glycine were monitored (*i.e.*, Gly681-Ser1430 and Ser769-Gly1329, respectively denoted as $d_{GS}^{NBS1}$ and $d_{GS}^{NBS2}$, Fig. 29c, Supplementary Fig. 7-8). Structural differences between the pre- and the post-translocation states (*i.e.*, *b*MRP1-LTX-$(ATP)_2$ and OF *b*MRP1-$(ATP)_2$, respectively) suggest that a conformational transition of NBD dimer is required prior to the substrate translocation event. Such transition from the so-called “non-competent” to “competent” NBD dimer conformations is likely to trigger TMD conformational transitions suggesting that it might be the limiting step for IF-to-OF transition.

### Asymmetric dynamics of bMRP1 and modulation of its conformational landscape

Overall flexibilities were assessed by calculating root-mean-square fluctuations (RMSF, Supplementary Fig. 9). Per-residue RMSF confirms the asymmetric behaviour of NBDs[7], NBD2 being more flexible than NBD1. For each system, backbone-based principal component analyses (PCA) were conducted. Only the three first largest principal components were considered, revealing 85.6 % to 95.9 % of the overall structural variabilities depending on the system (Supplementary Fig. 10). For all IF simulations, the three first largest variabilities were systematically assigned to asymmetric NBD motions for which NBD2 contributed the most, from 27.0 to 59.6% of the motion (Supplementary Fig. 11). The first principal components were mostly assigned to NBD twist and rocking motions (Supplementary Movies 1-5). This is in agreement with the experimentally suggested higher flexibility of NBDs from smFRET experiments along the kinetic cycle of MRP1[5]. Regarding OF simulations, the two first principal components were associated with concerted NBD twist motion and opening of the extracellular side mediated mostly by TMH4, TMH5, TMH7 and TMH8. These TMHs have been suggested to behave as a single bundle in ABCB1/P-gp[18]. Furthermore, to a lesser extent than for IF conformations, NBD2 remained more involved in this shared motion than NBD1.

The asymmetric behaviour was also pictured by the supramolecular arrangement between NBD1 and NBD2 for which two main subpopulations were observed for apo, *b*MRP1-$(ATP)_2$ and *b*MRP1-LTX states (Fig. 30a). Interestingly, interactions within NBS1 across NBDs were maintained along our MD simulations, even in absence of ATP molecules. In contrast, NBS2

is more flexible leading to either open or closed NBD dimer arrangements. Surprisingly, both arrangements were also observed to a lesser extent in *b*MRP1-$(ATP)_2$ while ATP molecules were expected to maintain interactions at the NBD dimer interface. However, in presence of both substrate and ATP molecules, only the closed population was observed picturing the information transduction from the TMD substrate-binding pocket to NBDs in order to likely decrease the energy barrier for IF-to-OF transition in type IV folding ABC transporters[20,21].

In order to explain the asymmetric behaviour of NBDs, particular attention was paid to coupling helices (CH) which ensure the signal transduction from TMDs to NBDs[20,21]. Natively, ABC transporters exhibit four coupling helices linking intracellular domains of TMH2/TMH3, TMH4/TMH5, TMH8/TMH9, and TMH10/TMH11. The so-called $CH_{2-3}$ and $CH_{10-11}$ are in contact with NBD1 while $CH_{4-5}$ and $CH_{8-9}$ interact with NBD2 (Fig. 30a). Contacts between CHs and NBDs are not modified upon binding of ATP molecules or substrate (Supplementary Fig. 12). Interestingly, for a given NBD, one can consider a so-called “weak CH” (namely, $CH_{2-3}$ and $CH_{8-9}$ for NBD1 and NBD2, respectively) for which only a few contacts were observed. On the other hand, the so-called “strong CH” (namely, $CH_{10-11}$ and $CH_{4-5}$ for NBD1 and NBD2, respectively) exhibited more contact with the NBD and for which the “ball-and-socket” arrangement is significantly more pronounced than for the “weak CH”. Interestingly, in mammalian MRP1, the NBD1 sequence exhibits a 13-aminoacid deletion which is associated with fewer contacts and significantly weaker interactions between $CH_{2-3}$ and $CH_{10-11}$ as compared to $CH_{4-5}$ and $CH_{8-9}$ for NBD2 (Supplementary Fig. 13). Furthermore, the few contacts observed between $CH_{2-3}$ and $CH_{10-11}$ are slightly decreased in presence of ATP molecules while the contact pattern for NBD2, *i.e.*, $CH_{4-5}$ and $CH_{8-9}$, is well conserved regardless of the presence of ATP and/or leukotriene C4. This leads to the disruption of the so-called “ball-and-socket” arrangements of $CH_{2-3}$ and $CH_{10-11}$ in NBD1 which is expected (i) to be responsible for the aforementioned larger flexibility of NBS1 region and (ii) to preclude the information transduction between TMDs and NBD1[8].

IF conformations exhibited weaker NBD dimer interactions than OF conformation, explaining the overall larger flexibility. Likewise, MD simulations revealed slightly higher flexibility of IF apo state as compared to ATP- and/or LTX-bound systems. This is in agreement with previous observations suggesting that interactions between TMHs or between NBDs are modulated by the presence of substrate and/or ATP molecules[4,7,8,22]. Taking advantage of our extensive unbiased MD simulations, explored conformational subspaces were featured in terms of free energy using the InfleCS clustering methods[23] (Fig. 30b). Given the aforementioned non-competent NBD dimer conformations, the focus was paid to NBD structural parameters, namely NBD twist and NBD distance. Larger variabilities for IF conformations were observed leading to multiple plausible minima for which interconversion is possible but slow. The expected to be competent NBD twist *versus* NBD distance subspace tends to be populated in presence of ATP molecule and/or substrate. Such finding highlights the central role of ATP- and substrate-binding events in the ABCC1 transport cycle.

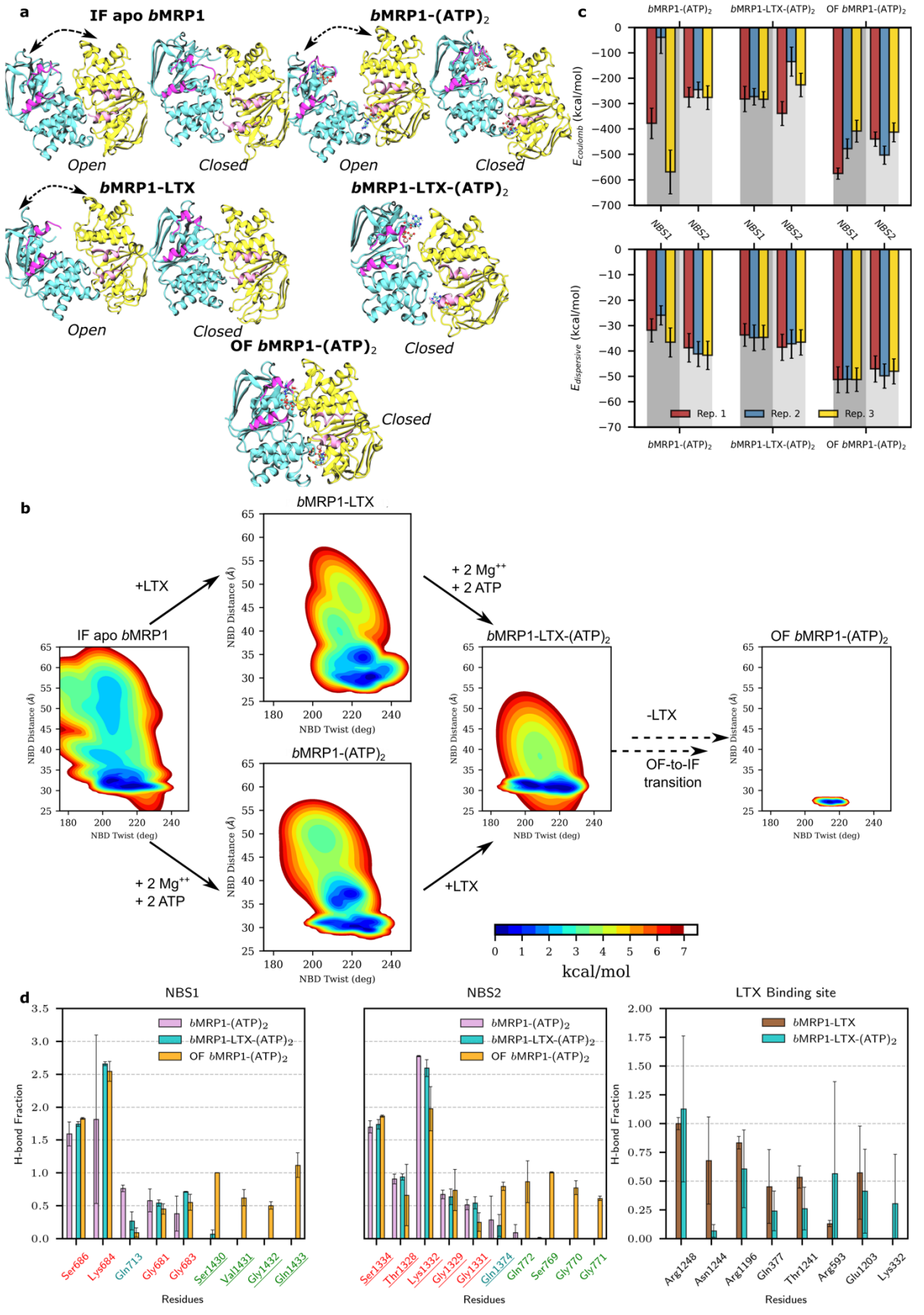



**Figure 30. Asymmetric structural dynamics of *b*MRP1 systems in POPC:POPE:Chol (2:1:1) and their conformational landscape.**

**a)** Representative snapshots picturing the open and closed conformations of NBD dimers observed during MD simulations. IF apo *b*MRP1, *b*MRP1-$(ATP)_2$ and *b*MRP1-LTX revealed two main subpopulations for which black arrows highlight the motion; while the pre- and post-translocation conformations (namely, *b*MRP1-LTX-$(ATP)_2$ and OF *b*MRP1-$(ATP)_2$) exhibited only the closed NBD dimer conformation. NBD1, NBD2 and coupling helices are respectively depicted in cyan, yellow and pink. **b)** System-dependent local conformational landscapes obtained from the GMM-based approach developed in the InfleCS method[23,25] highlighting the influence of nucleotides and the substrate on *b*MRP1 structural dynamics. **c)** Averaged Coulomb (up) and van der Waals (down) potentials calculated between nucleotides and NBS1 and NBS2, separately. **d)** Calculated H-bond networks between ATP and NBS1 (left) or NBS2 (centre) and between LTX and substrate-binding pocket residues (right).

To better decipher the NBD dimer dynamics, structural networks[24] were obtained by considering both contact maps (Supplementary Fig. 14) and dynamic cross-correlation matrices (Supplementary Fig. 15). Each NBD can be globally split into two subdomains regardless of the conformation or the bound states driven by ATP-binding (Supplementary Fig. 16). Indeed, it is worth mentioning that similar so-called communities were relatively well-conserved over the replicas and the systems investigated (Supplementary Table 6). The first community is defined by the Walker A and B as well as the A- and H-loops. On the other hand, the second community includes Q-and X-loops and the ABC signature sequence. Residues involved in these communities are highly correlated suggesting that local residue displacement upon ATP-binding will propagate. The two communities may be considered almost independent thus the binding of one ATP is not expected to strongly impact the motions of the second community. Meanwhile, given the asymmetric NBD dimer arrangement, ATP and magnesium are expected to link communities across NBDs. Interestingly, dynamic correlations in IF apo simulations exhibited larger variabilities since communities could be split into subcommunities (Supplementary Table 7), so that ATP-binding is expected to strengthen the structural cooperation within and between NBDs.

### Conformation-dependent binding modes of ATP in degenerate and canonical NBSs

Particular attention was paid to the binding modes of ATPs by assessing van der Waals and Coulomb potentials as well as H-bond networks in the NBSs (Fig. 30c-d). It is important to note that such analyses should not be considered quantitatively owing to the compensation of errors between the two potentials, especially at short distances[26]. However, they can be used to provide qualitative hints in order to compare ATP-binding driving forces. Interestingly, both Coulomb and dispersive interactions exhibited less attractive energies for all IF conformations than for OF. Given that NBDs exhibit lower flexibility in OF conformation, the subsequent tighter NBD dimer interactions can be rationalized by the proper local arrangement of ATP molecules in NBSs. For instance, calculated lower $\pi$-stacking distances between ATP molecules and NBS conserved motifs were systematically larger in IF than in OF conformations leading to lower interaction energies between ATP molecules and MRP1 residues (Fig. 30c and Supplementary Table 8). Surprisingly, in IF conformations, NBS1 tends to exhibit lower dispersive interactions between ATP molecule and Trp653 A-loop residue as compared to Tyr1301 in NBS2 A-loop, while the spatial aromatic surface of tryptophane is larger than for tyrosine. On the other hand, in the OF conformation, dispersive contributions tend to be slightly larger for NBS1 than for NBS2. H-bond networks between ATP molecules and NBSs (Fig. 30d) suggest a similar network between NBSs for IF conformations. The H-bond network is well conserved over IF and OF conformations regarding interactions with Walker A, however, ATP-bound IF systems do not exhibit the expected H-bond network with the signature motif as observed in OF simulations. Interestingly, MD simulations suggest an asymmetric behaviour

between NBS regarding H-bonds with Q-loop glutamine residues, namely Gln713 and Gln1374, respectively for NBS1 and NBS2. Calculated distances suggested that ATP $\gamma$-phosphate binding to NBS1 Q-loop was weaker than for NBS2 in the OF conformation. The opposite behaviour was observed in MD simulations for IF conformations. Therefore, present simulations underline that proper ATP-binding modes in both NBSs are key in triggering conformational changes required for substrate translocation.

**Towards deciphering the allosteric modulation between substrate- and nucleotide-binding sites**

Particular attention was paid to the substrate-binding pocket and compared to the cryo-EM structure of *b*MRP1 bound to leukotriene C4[8]. In agreement with experiments, leukotriene C4 binding mode took place in the two so-called P- and H-pockets ("P" and "H" respectively standing for polar and hydrophobic). Given the amphiphilic feature of leukotriene C4, Coulomb and H-bond networks have been shown to be central for substrate-binding[8]. MD simulations highlighted the same key residues which were experimentally observed (Fig. 30d). For instance, strong salt-bridges were observed between arginine residues (Arg1248, Arg1196 and Arg593, Fig. 30d and Fig. 31a) maintaining at least two out of the three leukotriene carboxylate groups in the P-pockets. Interestingly, variabilities in terms of H-bond fractions or interaction energies (Fig. 30d and Supplementary Fig. 17) suggest a dynamic binding mode in agreement with its expected breaking along the IF-to-OF large-scale conformational changes.

Even though differences in terms of interaction energies and H-bond networks between *b*MRP1-LTX-$(ATP)_2$ and *b*MRP1-LTX or *b*MRP1-$(ATP)_2$ remained low, they suggest a distant effect between NBSs and the substrate-binding pocket. Allosteric effect was assessed between the substrate-binding pocket and NBS1 and NBS2, independently (Fig. 31b). This was achieved by considering key residues for each binding site (Supplementary Table 9) for allosteric pathway network analyses[27,28]. Efficiencies (Fig. 31c) were calculated in presence or absence of substrate and ATP molecules. The impact of POPC:POPE:Chol (2:1:1) lipid bilayer was also considered. Natively, substrate-binding pocket and NBSs are allosterically connected through the protein as shown by calculated efficiencies without including nucleotides or substrates. As expected, the presence of substrate and/or ATP molecules substantially increased the allosteric communication from substrate-binding pocket to both NBSs. Interestingly, in spite of the aforementioned NBD asymmetric dynamics, present calculations did not exhibit a significant difference in terms of efficiencies between NBSs. Betweennesses were calculated to picture the residue and domain contributions to the allosteric pathway (Fig. 31d and Supplementary Fig. 18-20). Particular attention was paid to ATP-bound systems, *i.e.*, IF *b*MRP1-$(ATP)_2$, and *b*MRP1-LTX-$(ATP)_2$ as well as OF *b*MRP1-$(ATP)_2$ (Fig. 32d), other systems are reported in Supplementary Fig. 21. The main residues involved in the binding pocket-NBS allosteric pathways are mostly located in the intracellular part of TMHs. Interestingly, an asymmetric behaviour, for which TMH4, TMH5, TMH7 and TMH8 are significantly more involved than TMH1, TMH2, TMH10 and TMH11, was again observed. *b*MRP1 exhibits an asymmetric feature regarding the so-called "ball-and-socket" arrangements which was shown to be responsible for the structural cooperation between TMHs and NBDs through coupling helices[8]. The thirteen amino acid deletion in NBD1 leads to lower cooperation between TMH1, TMH2, TMH10 and TMH11 with NBD1, which in turn decreases the allosteric communication between substrate-binding pocket and NBS1. Furthermore, it is worth mentioning that our calculations suggest that information mostly goes through NBD2 since the direct communication with NBD1 is significantly weakened by the absence of "ball-and-socket"

conformation for $CH_{10\text{-}11}$[8]. Interestingly, POPC:POPE:Chol lipid bilayer was also shown to play a key role in the allosteric communication between substrate-binding pocket and NBSs (Fig. 31c). However, even though its impact is significant, lipid bilayer contributions appeared milder than in *e.g.*, Major Facilitator Superfamily membrane transporters[29].

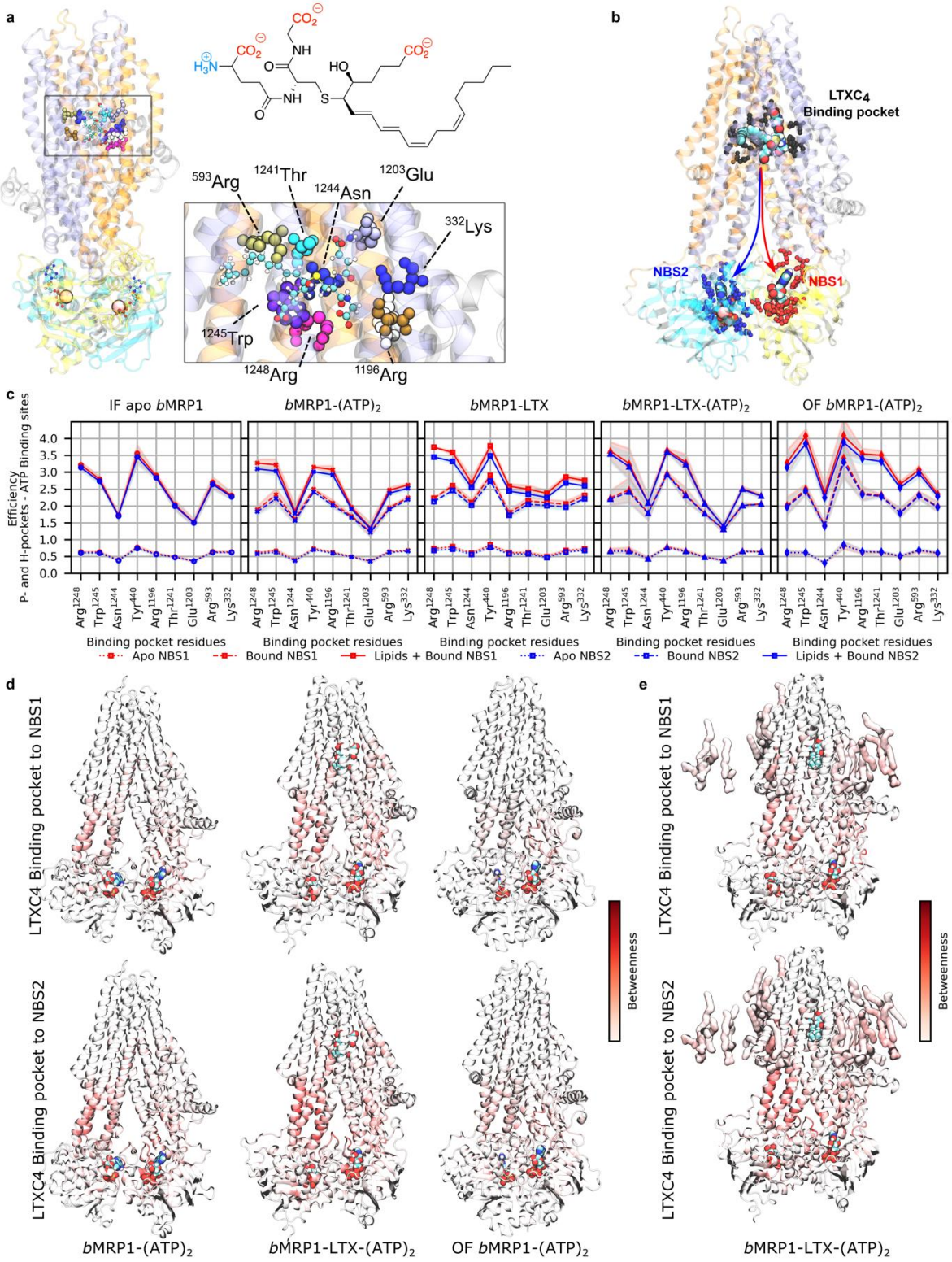



**Figure 31. Substrate-*b*MRP1 interactions and subsequent allosteric communications to nucleotide-binding sites.**

**a)** Substrate-binding pocket highlights important residues to leukotriene C4 binding. The structure of leukotriene is also shown for which amphiphilic features are stressed out. **b)** Definition of the allosteric pathway investigated in the present study for which NBS1 and NBS2 were treated separately. **c)** Calculated allosteric efficiencies of the information flow between substrate-binding pocket and NBS1 (red) or NBS2 (blue) for the different systems embedded in POPC:POPE:Chol (2:1:1). Solid, dashed and dotted lines respectively depict efficiencies considering: Protein + lipids + nucleotides/substrate, Protein + nucleotides/substrate and standalone Protein. **d)** Protein and **e)** lipid contributions to the information flow for allosteric communication from substrate-binding pocket to NBS1 (top) and NBS2 (bottom) show that NBD2 and its coupling helices ($CH_{4\text{-}5}$ and $CH_{8\text{--}9}$) are systematically involved regardless of the sink region.

### On the interplay between the lipid bilayer and bMRP1 structures and dynamics

As nowadays more attention has been paid to the interplay between the surrounding lipid environment[11,19,22,27,29] and membrane proteins, lipid-dependent protein dynamics and lipid-protein interactions were investigated. This was achieved by carrying out MD simulations in different lipid bilayers, namely POPC, POPC:Chol (3:1), POPC:POPE:Chol (2:1:1). IF apo *b*MRP1 and OF *b*MRP1-$(ATP)_2$ systems were also considered in unrealistic POPE and POPC:POPE (3:1) lipid bilayers.

Projection of lipid-dependent structural parameters onto the ABC conformational space (Supplementary Fig. 1 and 22-23) revealed that most of the present simulations tend to explore similar ABC subspaces regardless of the lipid bilayer membrane composition. Figure 4a compares structural parameter averages according to lipid bilayer compositions for all MD simulations performed in the present study. For IF conformations, only intracellular structural parameters (*i.e.*, IC angle and NBD distance) exhibited slight deviations according to the lipid composition. Systems performed in pure POPC lipid bilayer exhibited slightly more open conformations. On the other hand, our calculations suggest that only EC angle is affected by lipid bilayer composition but only for OF conformations. Likewise, it is important to note that calculated cavity radii (Supplementary Fig. 24) exhibited tiny differences while comparing lipid bilayer compositions. Even if these calculations underline a relatively limited overall impact of membrane composition on the *b*MRP1 structure, they did not sufficiently picture the dynamic variability over MD simulations and replicas (Supplementary Fig. 1-6 and 24).

Therefore, lipid-dependent conformational landscapes were calculated (Supplementary Fig. 25-26). For a given conformation and bound state, MD simulations preferentially populated similar regions regardless of lipid composition. However, structural variability was shown to be significantly affected by lipid composition. For instance, in pure POPC lipid bilayer, more open IF conformation subspaces, ranging from 30 to 55 Å NBD distance, were sampled. This effect was however reduced in presence of substrate which is expected to tend to maintain more contacts between TMHs and thus reduce the intracellular opening. MD simulations performed on POPC:POPE:Chol (2:1:1) exhibited a significantly smaller sampled region suggesting that the presence of PE lipids tends to close the intracellular gate of *b*MRP1 IF conformations, regardless of the bound state. On the other hand, regarding OF conformations and extracellular opening, global minima were interestingly observed in the same subspace of the conformational space as MD simulations performed in pure POPC and POPC:Chol (3:1). It suggests a limited impact of cholesterol on the opening of *b*MRP1 EC gate. However, the presence of PE lipids slightly shifted the calculated minima toward more opened OF structures, from 13.9 to 18.5°. Tilt angles between TMHs and lipid bilayer normal were measured in order to unravel slight differences between the distinct lipid bilayer compositions (Supplementary Fig. 27). Even though no clear conclusions can be drawn from these analyses, orientations of TMH3, TMH6 and TMH9 appeared more sensitive in absence of cholesterol. While considering

that TMHs act as bundles as shown for ABCB1/P-gp[18], the impact of lipid bilayer membrane was more pronounced on the tilt orientations of given smaller bundles (Supplementary Fig. 28), namely Bundle C & D respectively consisting in TMH3/TMH6 and TMH9/TMH12. Interestingly, these bundles are expected to undergo larger conformational changes along the transport cycle[18]. This suggests that even though lipid composition seems to have a rather limited impact on the overall structure of *b*MRP1, it may play a role in the kinetics of substrate transport by *b*MRP1.

To better understand the interplay between the lipid bilayer and *b*MRP1, particular attention was paid to the lipid bilayer membrane structure. The larger structural variability in pure POPC membrane was explained by a significantly more fluid lipid bilayer structure pictured by lower order parameters ($S_{CD}$) for palmitate and oleate tails (Fig. 32b). In line with the biophysics of pure lipid bilayers, the presence of cholesterol modulates the fluidity of POPC by increasing the lipid order which in turn led to lower flexibility of the lipid bilayer membrane[30]. To a lesser extent, the presence of PE lipid potentiated the structural effect of cholesterol, which is in agreement with the slightly lower structural variability of *b*MRP1 in POPC:POPE:Chol (2:1:1) as compared to other lipid bilayer membranes. Lipid order calculations also suggest a weak impact of protein dynamics on lipid bilayer structures. Indeed, IF apo and OF *b*MRP1-$(ATP)_2$ MD simulations exhibited slightly less ordered lipid tail profiles than IF *b*MRP1-LTX, *b*MRP1-$(ATP)_2$ and *b*MRP1-LTX-$(ATP)_2$. This may be explained by larger variability of intracellular and extracellular openings for IF and OF conformation, respectively, which in turn is likely to lead to more pronounced displacements of surrounding lipids.

The distribution of surrounding lipids revealed important cholesterol and PE lipid hotspots. For instance, PE lipids were shown to preferentially bind to pre-TMH7 elbow helix as well as close to the $L_0$ domain, for more than 50% of the simulation (Fig. 32c). It is worth mentioning that electron density maps revealed three cholesterol molecules bound to the resolved OF *b*MRP1 structure. Interestingly, one by the pre-TMH7 was maintained near to its initial position with a probability higher than 80% (Fig. 32c) for which cholesterol is oriented in line with pre-TMH7 elbow helix, *i.e.*, parallel to the lipid bilayer. Pre-TMH7 cholesterol hot spots were also observed for example in IF-apo POPC:POPE:Chol (2:1:1) simulations (Fig. 32c). Moreover, pseudo-symmetrically, a cholesterol hotspot was also observed near to pre-TMH1 elbow helix in simulations carried out with IF *b*MRP1 conformations (Fig. 32c). Interestingly, allosteric pathway analyses underlined the key role of cholesterol molecules close to pre-TMH1 and -TMH7 elbow helices in the information transduction from the substrate-binding pocket to NBSs as shown in Fig. 32d.

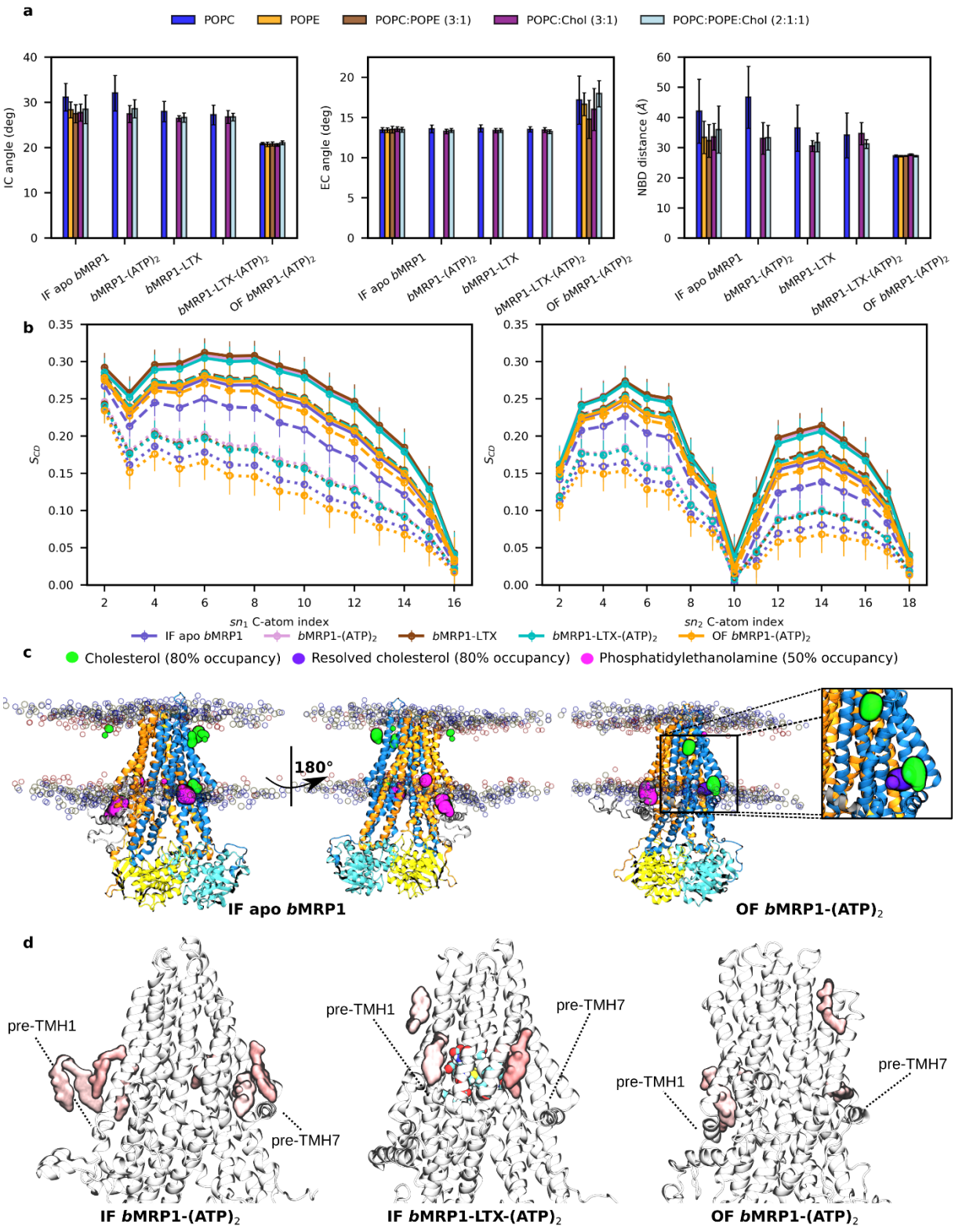



**Figure 32. Interplay between *b*MRP1 structural dynamics and lipid bilayer according to its composition.**

**a)** Average IC angle, EC angle and NBD distance for all *b*MRP1 systems embedded in different lipid bilayers. **b)** Calculated C-atom lipid tail order parameters ($S_{CD}$) for palmitic ($sn_1$, left) and oleic tails ($sn_2$, right) for all systems embedded in POPC-based models. Solid, dashed and dotted lines respectively depict lipid order parameters obtained in POPC:POPE:Chol (2:1:1), POPC:Chol (3:1) and POPC lipid bilayer models. **c)** Calculated binding hotspots obtained from cholesterol and PE lipids defined by presence likelihood higher than 80% and 50% respectively for cholesterol and PE lipids. Zoom on Cryo-EM resolved cholesterol (purple) is shown to highlight the

specific parallel orientation with respect to the lipid bilayer. **d)** Important lipid areas based on allosteric pathway analysis. The location of key cholesterol molecules involved in the allosteric communication between substrate-binding pocket and NBSs is shown. Both NBSs are considered, highlighting the expected central roles of $L_0$, pre-TMH1 and pre-TMH7 regions in specific lipid-protein interactions which might favour *b*MRP1 function.

### IV.1.3. Discussion

*b*MRP1 is the only member of the ABC drug exporter C-family which has been resolved by cryo-EM so far, given that ABCC7/CFTR is a chloride channel[31] and sulfonylurea receptors (ABCC8 and ABCC9) are involved in the regulation of potassium channels[32]. Over the past decade, particular attention has been paid to MRP transporters including MRP1, MRP2 and MRP4 given their clinically and pharmacologically relevant roles in drug disposition as pointed out by the ITC[12,13]. In contrast to its cousin ABCB1/P-gp, knowledge about MRP1 dynamics and functions still remains fragmented although it has been resolved in multiple states, such as IF apo and substrate-bound states[8] and two OF states under pre-[4] and post-hydrolysis[5] conformations, respectively bound to either two ATP molecules or to ADP/ATP pair. In the present work, an extensive set of all-atom MD simulations were performed in order to capture conformational dynamics of *b*MRP1 in different states considering different mixtures of lipid bilayer models including PC and PE lipids as well as cholesterol. We propose an MD-based computational approach in order to (i) complete the experimental observations made in detergents and (ii) highlight structural patterns which might be extended to, at least, other ABC C-family transporters.

Global conformational dynamics in POPC:POPE:Chol (2:1:1) shows significant variations of IF structures with respect to cryo-EM structures. In IF states, spontaneous closing of NBDs was systematically observed regardless of the presence of ATP molecules as pictured by IC angle and NBD distance values which all converged toward the same subspace. Regarding ABC conformational space, the presence of either ATP molecules in both NBSs or substrate in the TMD binding pocket mostly shifts the dynamics of NBD dimerization by modulating the NBD twist value (Fig. 30b and Supplementary Fig. 5). Therefore, our MD simulations are in perfect agreement with recent observations[6,7] that wide-open IF structures are expected to be unlikely in membrane environments. Wide-open structures observed in cryo-EM experiments are thus believed to be due to artifacts owing to the use of non-physiological environments for structure resolution[7], in agreement with structural differences observed *e.g.*, for P-gp reconstituted either in detergents or in nanodiscs[11]. MRP1 and likely other MRPs are expected to adopt the same "ATP-switch" model feature as suggested for TM287/288[6]. ATP-bound to NBS1 is expected to stay over multiple transport cycles sterically acting as a pivot point in absence of substrate. This is *e.g.*, pictured by the existence of two subpopulations for IF apo, ATP- and LTX-bound states regarding NBD dimer (Fig. 30a), namely open and close for which NBS1 is systematically formed. The NBD1 13 amino-acid deletion is thus expected to play a central role in this pivot model. Indeed, the absence of the "ball-and-socket" structure for NBD1 might weaken the coupling between NBD1 and TMH10/TMH11. In absence of substrate, this is associated with more flexibility regarding NBD1 leading to the opening of NBD dimer. However, in presence of ATP molecules and substrate, only close NBD dimer conformations were observed (Fig. 30a) highlighting the allosteric communication between TMDs and both NBSs. Owing to the lower coupling between TMH10/TMH11 and NBD1, allostery between TMDs and NBDs is expected to be mostly mediated through NBD2. Interestingly, NBS1 was also formed even in absence of ATP molecules (Fig. 30a). For instance, this might also explain why IF-to-OF transitions were experimentally observed even in absence of ATP molecules[5]. Our

results highlight the importance of NBD and NBS asymmetry as an evolution of ABC transporters which may result from energy saving while keeping transport function by requiring a single ATP hydrolysis.

MRP1 is expected to carry mostly anionic amphiphilic substrates contrary to P-gp which mostly transports hydrophobic substrates[2]. In agreement with structural observations, substrate access directly from either high- or low-density lipid tail regions of the membrane is very unlikely owing to the absence of an access channel in the lipid bilayer contrary to P-gp[8]. However, amphiphilic substrates might partition in the high-density polar head region. Therefore, MRP1 substrate access is expected to occur either directly from the cytoplasm or from the high-density polar head region, for which access might be possible via TMH4 and TMH6 (Supplementary Fig. 29). On the other side of *b*MRP1, our MD simulations suggest that substrate access between TMH10 and TMH12 may be less likely for bulky substrates. However, such assumptions require further biochemical and structural investigations.

In the present work, we also investigated the interplay between lipid bilayer membranes and protein dynamics. Our results first underline the particular importance of the use of a reliable membrane environment to provide robust insights into transporter dynamics and functions. By playing with different lipid bilayer models such as cholesterol-free and PE-free membranes, *b*MRP1 dynamics is modulated. In line with previous biophysical studies, the presence of cholesterol in PC-based membrane increased membrane stiffness which in turn reduced conformational dynamics of MRP1[30]. The impact of PE lipids is expected to be rather limited for IF states. However, interestingly, EC opening under OF conformation appeared more favourable in presence of PE (Fig. 32a). Globally, our results also suggest that lipid bilayer composition is very unlikely to preclude overall transporter functions, but it is expected to mostly affect kinetics. This hypothesis should be considered carefully for other MRP transporters. Indeed, in contrast to *e.g.*, MRP2 and MRP4, MRP1 is a ubiquitous exporter which was observed in different cell types for which lipid bilayer composition might differ. We can thus infer that other MRP members might become more sensitive to lipid bilayer membrane compositions.

Out of the physical role of lipid bilayer on protein dynamics, our results support that lipid components also play an active role in transporter function. In line with computational observations made on other membrane proteins[11,27,33,34], lipid components are involved in the allostery from TMDs to NBDs. More importantly, cholesterol bind to pre-TMH1 and pre-TMH7 elbow helices are strongly involved in these allosteric pathways. Such findings pave the way regarding the role of lasso pre-TMD motif ($L_0$) which was shown to be required for the MRP1 transport function[8,17]. Even if the role of lipids might sound limited on MRP1 as compared to other membrane receptors and channels, more attention should be paid to lipid-protein interactions, not only regarding the biophysical impact on the protein dynamics but also as transport modulator as recently proposed for P-gp[11].

We hope that the present work provides new insights into the function and the lipid-protein interplay of NBS degenerate ABC transporters for further investigations such as the role of MRPs in local pharmacokinetics, including *e.g.*, the impact of rare mutations/polymorphism as well as disease-based membrane lipid imbalance toward personalized medicine.

### IV.1.4. Methods

#### IV.1.4.1. Construction of *b*MRP1 models embedded in lipid bilayer membranes

In order to have an overview of milestone structures along the transport cycle of *b*MRP1, different conformations and bound states were considered in the present study: IF apo *b*MRP1, IF *b*MRP1-$(ATP)_2$, IF *b*MRP1-LTX, IF *b*MRP1-LTX-$(ATP)_2$ and OF *b*MRP1-$(ATP)_2$. The cryo-EM structures were used as starting structures for IF (PDB ID: 5UJ9[8] and 5UJA[8]) and OF-conformations (PDB ID: 6BHU[4]). It is important to note that OF conformation was resolved using E1454Q mutant which was shown to lower the rate of ATP hydrolysis thus promoting OF structure determination[4]. This mutation was reverted manually for the present study. The so-called "$TMD_0$" was not included in the present models as it has already been shown not to affect the substrate transport[3,4,8]. However, it was shown that the so-called pre-TMH1 lasso domain ($L_0$) is mandatory for MRP1 function while it was not resolved in any cryo-EM MRP1 structures[8,17]. Missing parts of $L_0$ domain was modelled using either I-Tasser (Iterative Threading ASSEmbly Refinement) server[35] or modeller v9.23[36] for IF and OF conformations, respectively. Likewise, the missing loop between TMH6 and NBD1 was also added in the present models.

IF conformations were built in different bound states, namely apo, $ATP_2$-, LTX- and LTX-$(ATP)_2$-bound states while OF conformation was solely constructed in the $ATP_2$-bound state. IF *b*MRP1-$(ATP)_2$ and IF *b*MRP1-LTX-$(ATP)_2$ were constructed by superimposing separately NBD of 5UJ9 and 5UJA, respectively onto the NBDs of OF conformation in which both ATP molecules and $Mg^{2+}$ are bound to NBSs. All final models were shortly minimized in vacuum using the Amber18 package[37,38] to avoid unphysical steric clashes.

CHARMM-GUI input generator[39,40] was used to embed the different *b*MRP1 models into different lipid bilayers, namely pure POPC, POPC:Chol (3:1) and POPC:POPE:Chol (2:1:1) taking advantage of *b*MRP1 coordinates obtained from the OPM (Orientations of Proteins in Membranes) database[41]. IF apo *b*MRP1 and OF *b*MRP1-$(ATP)_2$ structures were also embedded into pure POPE and POPC:POPE (3:1) lipid bilayers in order to specifically investigate lipid-protein interactions with PE lipids. It is worth mentioning that the resolved cryo-EM structure of OF *b*MRP1-$(ATP)_2$ also includes three resolved cholesterol molecules which were kept during all simulations in order to investigate their importance. From these different lipid bilayer compositions, POPC:POPE:Chol (2:1:1) mixture appeared the most relevant to model cell membranes. The other types of membranes were considered to help understand the role of each lipid in protein dynamics. The original total size of every system was *ca.* 120 x 120 x 180$Å^3$ (see Supplementary Tables 10-11 for system descriptions). To mimic physiological conditions, 0.15 M NaCl was used, and the systems were solvated using TIP3P explicit water molecules[42–44]. The final systems are made of *ca.* 245 000 atoms (see details in Supplementary Table 12).

#### IV.1.4.2. Molecular dynamics simulations

CHARMM-GUI[39,40] outputs were converted to Amber format using AmberTools scripts[37,38] (namely, charmmlipid2amber.py and pdb4amber**)**. Regarding $ATP_2$- and LTX-bound systems, substrate, nucleotides and $Mg^{2+}$ ions were added after building protein-lipid systems; therefore, neutrality was ensured by randomly removing the corresponding number of counterions. Amber FF14SB[45], Lipid17[46] and the modified DNA.OL15[47,48] force fields were used to respectively model protein residues, lipids and ATP molecules. Water molecules, $Mg^{2+}$ ions

and counterions were modelled using the TIP3P water model[42–44] as well as the corresponding monovalent and divalent ion parameters from Joung and Cheatham[49,50]. LTX (Leukotriene C4) substrate parameters were derived from the Generalized Amber Force Field version 2 (GAFF2)[51] using the Antechamber software[52]. LTX partial atomic charges were derived from quantum mechanical based calculations at the HF/6-31G* level of theory, using the R.E.D. server[53]. Each system was simulated with periodic boundary conditions. The cutoff for non-bonded interactions was 12 Å for both Coulomb and van der Waals potentials. Long-range electrostatic interactions were computed using the particle mesh Ewald method[54].

Minimization and thermalization of the systems and MD simulations were carried out with Amber18 and Amber20 packages[37,38] using CPU and GPU PMEMD versions. Minimization was carried out in four steps by sequentially minimizing: (i) water O-atoms (20000 steps); (ii) all bonds involving H-atoms (20000 steps); (iii) water molecules and counterions (50000 steps) and (iv) the whole system (50000 steps). Each system was then thermalized in two steps: (i) water molecules were thermalized to 100 K during 50 ps under (N,V,T) ensemble conditions using a 0.5 fs time integration; (ii) the whole system was then thermalized from 100 K to 310 K during 500 ps under (N,P,T) ensemble conditions with 2 fs timestep in semi-isotropic conditions. Then, each system was equilibrated during 5 ns under (N,P,T) ensemble conditions with 2 fs timestep in semi-isotropic conditions, using Berendsen barostat. Production runs were then carried out at the microsecond scale with 2 fs integration timestep under (N,P,T) ensemble conditions with semi-isotropic scaling. Temperature was maintained using the Langevin dynamics thermostat[55] with 1.0 $ps^{-1}$ collision frequency. Constant pressure set at 1 bar was maintained with semi-isotropic pressure scaling using either Berendsen barostat[56] for IF apo *b*MRP1 and OF *b*MRP1-$(ATP)_2$ or Monte Carlo barostat for IF *b*MRP1-$(ATP)_2$, IF *b*MRP1-LTX and IF *b*MRP1-LTX-$(ATP)_2$. The latter was used to speed up computational time.

In order to ensure the ATP docking into NBS, it is worth mentioning that restraint-MD simulations were carried out using a similar approach as proposed by Wen *et al.*[58]. Shortly, a set of distance-based restraints were applied between the A-loop tryptophan/tyrosine residues (Trp653 and Tyr1301, respectively for NBD1 and NBD2) and corresponding ATP purine moiety. $Mg^{2+}$-ATP-NBD arrangement was maintained by applying restraints between $Mg^{2+}$ ions and ATP phosphate groups as well as between $Mg^{2+}$ ions and Walker A Serine and Q-loop glutamine residue (namely, Ser685 and Gln713 for NBD1 and Ser1333, Gln1374 for NBD2). Moreover, ATP phosphate moieties were also restrained with surrounding Walker A residues. All distances were restrained using harmonic potentials for which minimal distances and force constants are reported in Supplementary Tables 13-14. Distance-based restraints were applied for thermalization and box equilibration steps. They were then smoothly removed along the first 10 ns of production runs. Restraints for $Mg^{2+}$ ions were kept during the whole simulation.

Snapshots were saved every 100 ps. For each system, three replicas were performed to better sample the local conformational space. Each production run was carried out for 2.0-2.5 μs and 1.5-2.0 μs, respectively for IF and OF models (Supplementary Table 15). Indeed, simulations performed using OF model reached the equilibrium faster than IF conformations (see time-dependent RMSDs in Supplementary Fig. 30). In the present study, the total MD time is 112.4 μs.

### IV.1.4.3. Analysis and visualization

Simulations were analysed using the CPPTRAJ[59] package, and in-house Python scripts taking advantage of MDAnalysis module[60,61]. Plots were obtained using the matplotlib v3.3.1 Python package[62]. Structure visualization and rendering were prepared using VMD software[63] (v1.9.3 and the alpha-v1.9.4). The so-called ABC structural parameters (*i.e.*, IC angle, EC angle, NDB distance, NBD twist and EC distance) were calculated using the same definition as proposed by Hofmann *et al*.[1]. Shortly, IC angle describes the IC opening of substrate entry and is defined by the angle between two vectors; both starting from the centre-of-mass of the whole extracellular region and directed toward either the IC region of TMH1, TMH2, TMH3, TMH6, TMH10 and TMH11 or the IC region of TMH4, TMH5, TMH7, TMH8, TMH9 and TMH12. Likewise, EC angle describes the EC opening for substrate release and is defined by the angle between two vectors starting from the centre-of-mass of both NBDs and directed toward either the EC region of TMH1, TMH2, TMH9, TMH10, TMH11, and TMH12 or the EC region of TMH3, TMH4, TMH5, TMH6, TMH7, and TMH8. EC distance was defined as the distance between the EC regions of TMH1, TMH2, TMH9, TMH10, TMH11, and TMH12 and the EC region of TMH3, TMH4, TMH5, TMH6, TMH7, and TMH8. NBD distance was defined as the distance between the two NBD centres-of-mass. Since these definitions were applied to the present MD models of *b*MRP1 but also to available resolved structures of other ABC transporters, intracellular and extracellular regions were defined based on the membrane thickness proposed in the OPM database[41]. Residue selections for each system and each structure parameter are reported in Supplementary Table 1. For each system and each lipid bilayer, local free energy landscape was calculated using the InfleCS approach which takes advantage of Gaussian Mixture Models (GMM)[23,25]. Structural parameters (*i.e.*, IC and EC angles, NBD distance and twist) were taken to monitor the free energy landscape using a grid size set at 80, from 2 to 12 gaussian components for each GMM obtained by a maximum of 20 iterations.

H-bond analyses were performed using CPPTRAJ[59] in which distance and angle cutoffs were set at 3.5 Å and 120°, respectively. Substrate cavities were calculated using the Hole2.2 software[64] on a trajectory made of frames selected every 10 ns of the equilibrated part of each replica separately. Profiles were then averaged and PC P-atom z-density profiles were used to define the centre of the lipid bilayer membrane (z = 0). Regarding tilt angles, all systems were aligned to the OF *b*MRP-$(ATP)_2$ model embedded in POPC:POPE:Chol (2:1:1). Lipid distributions were obtained using the CPPTRAJ[59] package while membrane thicknesses were obtained using the MEMBPLUGIN for VMD[63,65].

Principal component analyses were also performed using the CPPTRAJ[59] package, focusing on the ABC core, defined by backbone atoms of TMH1 to TMH12, NBD1 and NBD2. System variabilities were investigated by carrying out independent PCA for each system for which each replica was aligned to an average structure of the system. Network analyses were performed using the VMD Network Analysis plugin[63,66]. Dynamic cross-correlation matrices were calculated separately for each replica on which C$\alpha$-atoms were selected as nodes and all the default restrictions (notSameResidue, notNeighboringCAlpha, notNeighboringPhosphate, notNeighboringResidue) were applied. Communities were then calculated using gncommunities[66]. Allostery network pathways were determined using the recent Allopath approach developed by Westerlund *et al*.[27,28]. Shortly, the distant “communication efficiency” between two domains was obtained from the contact map and the mutual information matrix of protein residues as well as non-protein interactors such as surrounding lipid and bound molecules (*e.g.*, nucleotides and substrate). Such an approach also provides a betweenness

profile which pictures the involvement of each component (*i.e.*, residue, lipid, substrate and nucleotide). For each lipid molecule, three nodes were defined corresponding to the polar head group and the two lipid tails. ATP was also split into three nodes: purine and ribose moiety as well as triphosphate tail. LTX substrate was divided into three nodes: glutathione moiety, poly-unsaturated tail and hydroxypentanoic acid. $Mg^{2+}$ ions and cholesterol molecules were considered as one node each. Atom selections per node are reported in Supplemental Fig. 31. Allosteric pathways were calculated from substrate-binding pocket to each NBS, separately, as defined in Supplemental Table 9.

### IV.1.5. References


1. Hofmann, S. *et al.* Conformation space of a heterodimeric ABC exporter under turnover conditions. *Nature* **571**, 580–583 (2019).
2. Thomas, C. & Tampé, R. Structural and Mechanistic Principles of ABC Transporters. *Annu Rev Biochem.* **89**, 605–636 (2020).
3. Kroll, T., Prescher, M., Smits, S. H. J. & Schmitt, L. Structure and Function of Hepatobiliary ATP Binding Cassette Transporters. *Chem Rev* **121**, 5240–5288 (2021).
4. Johnson, Z. L. & Chen, J. ATP Binding Enables Substrate Release from Multidrug Resistance Protein 1. *Cell* **172**, 81-89.e10 (2018).
5. Wang, L. *et al.* Characterization of the kinetic cycle of an ABC transporter by single-molecule and cryo-EM analyses. *eLife* **9**, e56451 (2020).
6. Hutter, C. A. J. *et al.* The extracellular gate shapes the energy profile of an ABC exporter. *Nat Commun* **10**, 2260 (2019).
7. Stockner, T., Gradisch, R. & Schmitt, L. The role of the degenerate nucleotide binding site in type I ABC exporters. *FEBS Lett* **594**, 3815–3838 (2020).
8. Johnson, Z. L. & Chen, J. Structural Basis of Substrate Recognition by the Multidrug Resistance Protein MRP1. *Cell* **168**, 1075-1085.e9 (2017).
9. Scortecci, J. F. *et al.* Cryo-EM structures of the ABCA4 importer reveal mechanisms underlying substrate binding and Stargardt disease. *Nat Commun* **12**, 5902 (2021).
10. Xu, D. *et al.* Cryo-EM structure of human lysosomal cobalamin exporter ABCD4. *Cell Res* **29**, 1039–1041 (2019).
11. Kapoor, K., Pant, S. & Tajkhorshid, E. Active participation of membrane lipids in inhibition of multidrug transporter P-glycoprotein. *Chem Sci* **12**, 6293–6306 (2021).
12. Giacomini, K. M. *et al.* Membrane transporters in drug development. *Nat Rev Drug Discov* **9**, 215–236 (2010).
13. Zamek-Gliszczynski, M. J. *et al.* Transporters in Drug Development: 2018 ITC Recommendations for Transporters of Emerging Clinical Importance. *Clin Pharmacol Ther* **104**, 890–899 (2018).
14. Mateus, A. *et al.* Prediction of intracellular exposure bridges the gap between target- and cell-based drug discovery. *Proc Natl Acad Sci* **114**, (2017).
15. Nigam, S. K. What do drug transporters really do? *Nat Rev Drug Discov* **14**, 29–44 (2015).
16. Chantemargue, B. *et al.* Structural patterns of the human ABCC4/MRP4 exporter in lipid bilayers rationalize clinically observed polymorphisms. *Pharmacol Res* **133**, 318–327 (2018).

17. Bickers, S. C., Benlekbir, S., Rubinstein, J. L. & Kanelis, V. Structure of Ycf1p reveals the transmembrane domain TMD0 and the regulatory region of ABCC transporters. *Proc Natl Acad Sci* **118**, e2025853118 (2021).

18. Moradi, M. & Tajkhorshid, E. Mechanistic picture for conformational transition of a membrane transporter at atomic resolution. *Proc Natl Acad Sci* **110**, 18916–18921 (2013).

19. Immadisetty, K., Hettige, J. & Moradi, M. Lipid-Dependent Alternating Access Mechanism of a Bacterial Multidrug ABC Exporter. *ACS Cent Sci* **5**, 43–56 (2019).

20. Srikant, S. & Gaudet, R. Mechanics and pharmacology of substrate selection and transport by eukaryotic ABC exporters. *Nat Struct Mol Biol* **26**, 792–801 (2019).

21. Verhalen, B. *et al.* Energy transduction and alternating access of the mammalian ABC transporter P-glycoprotein. *Nature* **543**, 738–741 (2017).

22. Olsen, J. A., Alam, A., Kowal, J., Stieger, B. & Locher, K. P. Structure of the human lipid exporter ABCB4 in a lipid environment. *Nat Struct Mol Biol* **27**, 62–70 (2020).

23. Westerlund, A. M. & Delemotte, L. InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *J Chem Theory Comput* **15**, 6752–6759 (2019).

24. Sethi, A., Eargle, J., Black, A. A. & Luthey-Schulten, Z. Dynamical networks in tRNA:protein complexes. *Proc Natl Acad Sci* **106**, 6620–6625 (2009).

25. Westerlund, A. M., Harpole, T. J., Blau, C. & Delemotte, L. Inference of Calmodulin's Ca2+-Dependent Free Energy Landscapes via Gaussian Mixture Model Validation. *J Chem Theory Comput* **14**, 63–71 (2018).

26. Zgarbová, M., Otyepka, M., Šponer, J., Hobza, P. & Jurečka, P. Large-scale compensation of errors in pairwise-additive empirical force fields: comparison of AMBER intermolecular terms with rigorous DFT-SAPT calculations. *Phys Chem Chem Phys* **12**, 10476–10493 (2010).

27. Westerlund, A. M., Fleetwood, O., Pérez-Conesa, S. & Delemotte, L. Network analysis reveals how lipids and other cofactors influence membrane protein allostery. *J Chem Phys* **153**, 141103 (2020).

28. Kang, P. W. *et al.* Calmodulin acts as a state-dependent switch to control a cardiac potassium channel opening. *Sci Adv* **6**, eabd6798 (2020).

29. Janaszkiewicz, A. *et al.* Insights into the structure and function of the human organic anion transporter 1 in lipid bilayer membranes. *Sci Rep* **12**, 7057 (2022).

30. de Meyer, F. & Smit, B. Effect of cholesterol on the structure of a phospholipid bilayer. *Proc Natl Acad Sci* **106**, 3654–3658 (2009).

31. Liu, F., Zhang, Z., Csanády, L., Gadsby, D. C. & Chen, J. Molecular Structure of the Human CFTR Ion Channel. *Cell* **169**, 85-95.e8 (2017).

32. Lee, K. P. K., Chen, J. & MacKinnon, R. Molecular structure of human KATP in complex with ATP and ADP. *eLife* **6**, e32481 (2017).

33. Martens, C. *et al.* Lipids modulate the conformational dynamics of a secondary multidrug transporter. *Nat Struct Mol Biol* **23**, 744–751 (2016).

34. Martens, C. *et al.* Direct protein-lipid interactions shape the conformational landscape of secondary transporters. *Nat Commun* **9**, 4151 (2018).

35. Yang, J. *et al.* The I-TASSER Suite: protein structure and function prediction. *Nat Methods* **12**, 7–8 (2015).

36. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinforma* **54**, (2016).

37. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the Amber biomolecular simulation package: Amber biomolecular simulation package. *Wiley Interdiscip Rev Comput Mol Sci* **3**, 198–210 (2013).

38. Case, D. A. *et al. AMBER 2018*. (niversity of California, San Francisco).

39. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J Comput Chem* **29**, 1859–1865 (2008).

40. Wu, E. L. *et al.* CHARMM-GUI *Membrane Builder* toward realistic biological membrane simulations. *J Comput Chem* **35**, 1997–2004 (2014).

41. Lomize, M. A., Pogozheva, I. D., Joo, H., Mosberg, H. I. & Lomize, A. L. OPM database and PPM web server: resources for positioning of proteins in membranes. *Nucleic Acids Res* **40**, D370–D376 (2012).

42. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).

43. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**, 3586–3616 (1998).

44. Price, D. J. & Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. *J Chem Phys* **121**, 10096–10103 (2004).

45. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).

46. Dickson, C. J. *et al.* Lipid14: The Amber Lipid Force Field. *J Chem Theory Comput* **10**, 865–879 (2014).

47. Meagher, K. L., Redman, L. T. & Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem* **24**, 1016–1025 (2003).

48. Galindo-Murillo, R. *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* **12**, 4114–4127 (2016).

49. Joung, I. S. & Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J Phys Chem B* **112**, 9020–9041 (2008).

50. Joung, I. S. & Cheatham, T. E. Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters. *J Phys Chem B* **113**, 13279–13290 (2009).

51. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J Comput Chem* **25**, 1157–1174 (2004).

52. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J Mol Graph Model* **25**, 247–260 (2006).

53. Vanquelef, E. *et al.* R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res* **39**, W511-517 (2011).

54. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J Chem Phys* **98**, 10089–10092 (1993).

55. Davidchack, R. L., Handel, R. & Tretyakov, M. V. Langevin thermostat for rigid body dynamics. *J Chem Phys* **130**, 234101 (2009).

56. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J Chem Phys* **81**, 3684–3690 (1984).

57. Åqvist, J., Wennerström, P., Nervall, M., Bjelic, S. & Brandsdal, B. O. Molecular dynamics simulations of water and biomolecules with a Monte Carlo constant pressure algorithm. *Chem Phys Lett* **384**, 288–294 (2004).

58. Wen, P.-C., Verhalen, B., Wilkens, S., Mchaourab, H. S. & Tajkhorshid, E. On the Origin of Large Flexibility of P-glycoprotein in the Inward-facing State. *J Biol Chem* **288**, 19211–19220 (2013).

59. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

60. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

61. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proc. 15th Python Sci. Conf.* 98–105 (2016) doi:10.25080/Majora-629e541a-00e.

62. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng.* **9**, 90–95 (2007).

63. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33–38, 27–28 (1996).

64. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J Mol Graph* **14**, 354–360, 376 (1996).

65. Guixà-González, R. *et al.* MEMBPLUGIN: studying membrane complexity in VMD. *Bioinforma Oxf Engl* **30**, 1478–1480 (2014).

66. Eargle, J. & Luthey-Schulten, Z. NetworkView: 3D display and analysis of protein·RNA interaction networks. *Bioinformatics* **28**, 3000–3001 (2012).

### IV.1.6. Acknowledgements

We acknowledge Dr. Benjamin Chantemargue and Mehdi Benmameri for stimulating and fruitful scientific discussions. We thank Xavier Montagutelli from the IT department of Limoges University for technically supporting supercomputer facilities. Calculations were performed using CALI ("CAlcul en LImousin") and “Baba Yaga” supercomputers hosting in Limoges University, as well as on the “Jean-Zay” national supercomputer from IDRIS HPC resources under the allocations 2020-A0080711487 and 2021-A0100711487 made by GENCI. This work was supported by grants from the “Agence Nationale de la Recherche” (ANR-19-CE17-0020-01 IMOTEP), Région Nouvelle Aquitaine and “Institut National de la Santé et de la Recherche Médicale” (INSERM, AAP-NA-2019-VICTOR).

### IV.1.7. Author Contributions

Á.T and F.D.M conceived the study. Á.T and F.D.M. conducted all the simulations. Á.T, A.J., V.C and F.D.M analysed the simulations. Á.T. and F.D.M. interpretated the results and discussed them together with A.J. and V.C. Á.T. and F.D.M wrote the manuscript and it was edited, reviewed and approved by all authors.

## IV.2. Structural and dynamical differences shown by the pre- and post-hydrolysis states of ABCC1 (MRP1)

Ágota Tóth, Angelika Janaszkiewicz, Veronica Crespi and Florent Di Meo*

Inserm UMR 1248 Pharmacology & Transplantation, Univ. Limoges, 2 rue du Prof. Descottes, 87000 F-Limoges, France

**Contact information**
Dr. Florent Di Meo
Inswem U1248 Pharmacology & Transplantation, Univ. Limoges,
2 rue du Prof. Descottes,
87000 F-Limoges, France
florent.di-meo@inserm.fr
Tel: +33(0)5 19 56 42 76

### In preparation

An OF ATP-ADP-bound *b*MRP1 structure was published recently[1]; thus, simulations were run similarly as for the previously available *b*MRP1 structures (section IV.1.4.). During the analysis, the focus was on the differences shown by the pre- ($ATP_2$-bound) and post-hydrolysis (ATP-ADP-bound) states of OF *b*MRP1. This was done from different point of views. First the mapped conformational space was investigated, following by

Until the results will be published, supplementary information is available upon request.

**Keywords.** ABCC1/MRP1, molecular dynamics, dynamic L0, post-hydrolysis state, allosteric communication

### IV.2.1. Materials and Methods

Simulations were prepared similarly as in our previous study (section IV.1.4.)[2]. Shortly, OF *b*MRP1 ATP-ADP-bound structure was constructed using the available cryo-EM structure (PDB ID: 6UY0). As the so-called "$TMD_0$" do not affect the substrate transport[3–5], it was not included in the present model. The pre-TMH1 lasso domain ($L_0$), which is mandatory for MRP1 function[5,6], was not totally resolved in the cryo-EM structure. Therefore, the missing part of $L_0$ as well as the missing loop between TMH6 and NBD1 were modelled using modeller v9.23[7]. $L_0$ was modelled based on the $L_0$ structure from the IF structure as well as on sequence for that part which was clashing with TMHs. The final model was shortly minimized in vacuum using the Amber18 and Amber20 package[8,9] to avoid unphysical steric clashes.

OF *b*MRP1 ATP-ADP-bound structure was embedded into three different lipid bilayers, such as pure POPC, POPC:Chol (3:1), and POPC:POPE:Chol (2:1:1), using the CHARMM-GUI input generator[10,11]. The three co-crystalized cholesterol molecules were kept during the simulations. The original total size of the system was *ca.*120 x 120 x 180$Å^3$ (Supplementary Table 1). To mimic physiological conditions, 0.15 M NaCl salt concentration was used, and the system was solvated using TIP3P explicit water molecules[12]. The final systems are made of *ca.* 245 000 atoms (Supplementary Table 2).

Amber software was used to run simulations, as it was done previously[2] using the parameters in table 11. Monte Carlo barostat was used for the production run to speed up computational time. Production run was carried out for 1.5 μs.

**Table 11. Parameters used during the simulations**

| Parameter | Modelled molecules |
|---|---|
| FF14SB[13] | protein residues |
| Lipid17 | lipids |
| modified DNA.OL15[14,15] | ATP molecules |
| TIP3P water model[12,16,17],<br>monovalent and divalent ion parameters from Joung and Cheatham[18,19] | water molecules, $Mg^{2+}$ ions and counterions |

The same analysis techniques were used to monitor the simulations as previously, see details in section IV.1.4.[2]. Shortly, analyses were performed using the CPPTRAJ[20] package and in-house Python scripts taking advantage of the MDAnalysis module[21,22] and the matplotlib v3.3.1 package[23]. Structure visualization and rendering were prepared using the VMD software (v1.9.3 and the alpha-v1.9.4)[24]. So-called ABC structural parameters[25,26] were calculated and used to map the ABC conformational space. The GMM-based InfleCS clustering method was applied to decipher the free energy landscape using the ABC structural parameters as collective variables for the energy sampling.

### IV.2.2. Results

#### IV.2.2.1. MD simulations reveals slight but relevant structural differences between the pre- and post-hydrolysis states

In the present section, particular attention is paid to MD simulations performed in POPC:POPE:Chol (2:1:1). The influence of the membrane is discussed in section IV.2.2.3. Conformational spaces sampled for pre- and post-hydrolysis states (namely OF *b*MRP1-$(ATP)_2$ and OF *b*MRP1-ATP-ADP) were examined focusing on key ABC structural parameters (Supplementary Fig. 1-2,4-5). NBD distances and intracellular (IC) angles were used to monitor the intracellular opening, while the extracellular angle (EC) was used to assess the opening towards the external cell compartments. NBD rocking-twist as originally defined by Moradi *et al.*[25] was also monitored. These parameters were projected onto the ABC conformational space defined by measuring ABC structural parameters of a large data set of experimentally resolved ABC transporters (Fig. 33a). The so-called inward-facing (IF), outward-facing (OF) and asymmetric unlock-returned (UR) turnover[26] states are explicitly separated using these structural parameters (Fig. 33a). For both pre- and post-catalytic states, IC angle shows similar trends picturing small IC opening, values ranging from 19.4 to 22.6°. Furthermore, they did not show large variability suggesting stable conformations (Fig. 33a, Supplementary Fig. 1). On the other side of the membrane, post-hydrolysis state exhibits slightly smaller EC angle values than pre-catalytic states, ranging from 13.8 to 23.6 and from 12.6° to 19.3°, respectively for pre- and post-catalytic states. Interestingly, variability is larger than for IC angle depicting larger EC opening fluctuation (Fig. 33a, Supplementary Fig. 1-2).

Pre- and post-catalytic states mostly differ in NBD dimer arrangement. The NBD rocking-twist angle of the post-hydrolysis state does not deviate from the originally resolved structure while

the NBD distance value is slightly larger (Fig. 33a, Supplementary Fig. 3-4). The opposite trend was observed in pre-catalytic state for which NBDs get closer to each other in MD simulations with respect to cryo-EM resolved structure. This might be explained by stronger NBD-NBD non-covalent interactions in biomimetic membranes compared to detergent, as suggested in our previous work[2]. Furthermore, the E1454Q mutation used for cryo-EM resolution may also alter ATP binding in nucleotide binding site (NBS) 2. Interestingly, the post-catalytic OF *b*MRP1-ATP-ADP state exhibit lower NBD twist than the pre-catalytic state by *ca.* 5°. Noteworthy, the post-hydrolysis state which is expected to be before the OF-to-IF transition is getting closer to the conformational space of the UR structures which were resolved in native nanodisk environment. This is in agreement with the recent structural finding suggesting that ABC transporter might adopt UR state prior to get back to IF conformational states, resetting the transport cycle[26]. It is important to note that the use of NBD distances should be carefully considered, especially for NBD degenerate ABC transporters such as MRP1 given the expected pivotal role of ATP-bound NBS1 in maintaining NBD dimerization along transport cycle[27], in agreement with our previous findings[2].

Assuming that our MD simulations were long enough to independently sample the local conformational space of pre- and post-catalytic states, local free energy surfaces were calculated according to ABC structural parameters using the InfleCS framework[28,29] (Fig. 33b). Particular attention was paid to NBD structural parameters. Interestingly, two main populations were captured from MD simulations performed on the post-catalytic OF *b*MRP1-ATP-ADP state regarding to NBD distance, showing a slight flexibility in term of NBD opening while it is unlikely to happen in pre-catalytic state. Likewise, small difference for NBD rocking-twist parameter between pre- and post-hydrolysis states is also captured from the free energy point of view, confirming that ATP hydrolysis shift the free energy surface toward more open and less twisted NBD dimer conformation.

On the other side of the membrane, pre-hydrolysis state simulations show more flexible EC opening compared to the post-catalytic state. MD simulations performed on OF *b*MRP1-ATP-ADP state surprisingly suggest two well-separated populations. Interestingly, transition from one to the other requires to cross a relatively high energy barrier (*ca.* 3 kcal.mol$^{-1}$). Our simulations suggest that such transition may be favoured by slightly closing IC gate. However, this should be carefully considered and confirmed by further investigations.

Particular attention was paid to the lasso domain ($L_0$) which connect the $TMD_0$ and the functional core of MRP1. It is worth remembering that $TMD_0$ was not included in the MRP1 model since it was shown not to be involved in the transport function. However, experiments stressed out that $L_0$-less MRP1 is not functional. MD simulations reveals large flexibility of the $L_0$ domain as pictures by monitoring root-mean-square deviation (RMSD) for the whole protein along the simulations (Supplementary Fig. 6). This was confirmed by assessing RMSD only on the ABC core (*i.e.*, extruding $L_0$) showing much smaller deviations along the simulations of pre- and post-catalytic states (Supplementary Fig. 7). Domain flexibilities were also assessed by measuring the per-residue root-mean-square fluctuations (RMSF) over the last 400 ns (see Fig 33c). Interestingly, RMSF analyses suggests higher flexibility for the pre-hydrolysis state in membranes containing cholesterol (Fig 33c). Interestingly, in both pre- and post-catalytic states, $L_0$ flexibility suggest two distinct behaviours (Supplementary Fig. 8). On one hand, Asn203-Lys267 does not exhibit large flexibility, remaining in close contact to pre-TMH1, TMH1, TMH2, TMH10 and TMH11. On the other hand, Ser268-Lys307 display large variability. This is in agreement with resolved structure since the latter has not been totally resolved by

cryo-EM techniques. MD simulations also suggest that the flexible $L_0$-subdomain can interact with NBD1. Even though $L_0$ roughly maps the same space in the pre- and post-hydrolysis states, dynamics shows some differences.

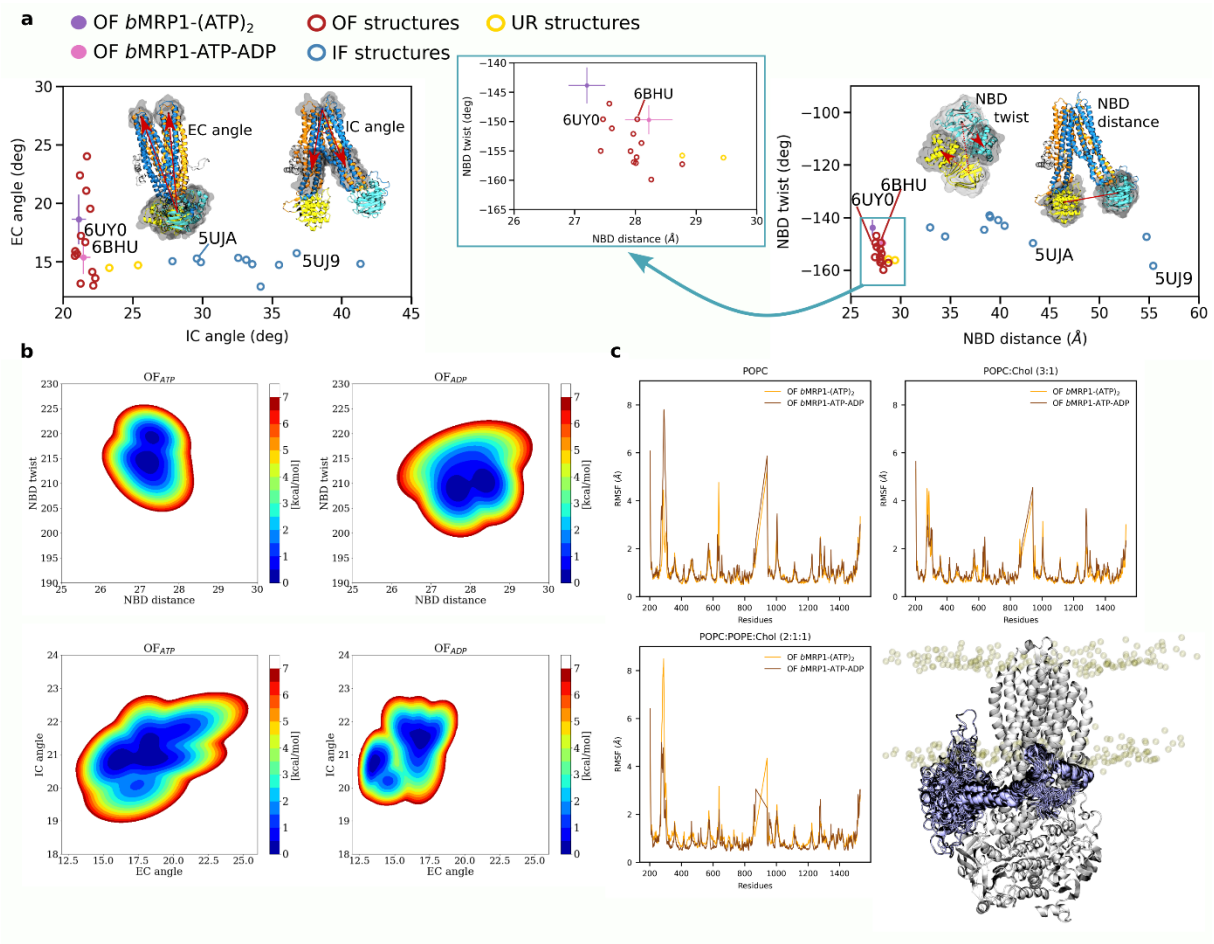



**Figure 33. Structural dynamics of OF bMRP1-ATP-ADP.**

**a)** Projection of *b*MRP1 structural parameters of the pre- and post-hydrolysis states, respectively OF-$(ATP)_2$ and OF-ATP-ADP simulations, onto the ABC conformational space obtained from multiple resolved ABC structures. Each replica was averaged over the last 800 ns. PDB IDs of the resolved *b*MRP1 cryo-EM structures are explicitly mentioned. The zoom shows the OF area of the NBD twist – NBD distance plot. The first and second TMDs are respectively depicted in orange and blue, and NBD1 and NBD2 are respectively coloured yellow and cyan. **b)** System-dependent local conformational landscapes obtained from the GMM-based approach developed in the InfleCS method[28,29] highlighting the influence of nucleotides on *b*MRP1 structural dynamics. **c)** RMSF calculated on the last 400 ns and structure representation of OF *b*MRP1-ATP-ADP POPC:POPE:Chol (2:1:1) show the $L_0$-dynamics.

### IV.2.2.2. Differences in the nucleotide-binding of the pre- and post-hydrolysis states

As NBD distance and NBD twist values showed differences in the pre- and post-hydrolysis states, special attention was paid to the NBSs. MRP1 belongs to NBD degenerate ABC transporter of which NBS1 is expected to exhibit (i) increased important ATP-binding but (ii) significantly lower ATP hydrolysis rate. Local key distances proposed to maintain the nucleotide in the binding site[2,30] were monitored between either ATP or ADP molecules and well-known conserved motifs (Fig. 34a). Distances between ATP molecules and NBS1 binding motifs are globally conserved when comparing pre- and post-hydrolysis states. Only the distance between A-loop Trp653 and adenosyl moiety is slightly larger in the post-hydrolysis

state by *ca.* 1.0 Å. This is in agreement with former observations made on NBD degenerate ABC transporters suggesting that ATP remains constantly bound in NBS1 along several transport cycles. In contrast, all distances between the nucleotide and NBS2 motifs differ. The most important deviation is observed with the Q-loop motif (respectively 6.2 ± 0.1 and 8.9 ± 0.8 Å in pre- and post-hydrolysis states).H-loop and Walker B motifs also show significant structural differences. This might be easily explained by the absence of γ-phosphate which is supposed to bind Q-loop motif. However, in presence of ADP in NBS2, motifs supposed to bind γ-phosphate tend anyway to locally reshape by getting closer to β-phosphate moiety (*e.g.*, ABC signature, H-loop and Walker B), as shown by distances measured with terminal β-phosphate O-atom ($d^*_{post}$ in Fig. 34a). This suggests a local rearrangement of NBDs after ATP hydrolysis.

More specifically, binding modes of the nucleotides were also investigated by assessing atomic contacts, H-bond networks as well as van der Waals and Coulomb potentials between ATP or ADP molecules and NBS motifs (Fig. 34b, Supplementary Fig. 9). It must be stressed that van der Waals and Coulomb potentials can only provide qualitative hints to compare the binding sites due to a compensation of errors between the two potentials especially at short distances[31]. H-bond fractions (Fig. 34b) highlights the significant difference between pre- and post-hydrolysis in NBS2 H-bond network, especially regarding the ABC signature motif and Q-loop. This is confirmed by interatomic contact analyses since there are significantly less contacts in the post-hydrolysis state in NBS2. Walker A motif in NBS2 has slightly looser connections with the ADP than ATP which might be involved in the subsequent nucleotide release. NBS-nucleotide atomic contact fraction maps confirm weaker interactions between ADP and NBS2 ABC signature Q-loop motifs, while contact remains similar with other motifs (Supplementary Fig. 9). Interestingly, ABC signature motif has similar H-bond fraction in NBS1 in pre- and post-hydrolysis states.NBS1 Walker A motif, that binds α- and β-phosphates, exhibits slightly stronger connections in the post-hydrolysis state in agreement with the expected maintain of ATP molecule along likely multiple transport cycles[27].

NBS structural differences between pre- and post-hydrolysis states is expected to modulate the NBD dimer supramolecular arrangement. In this context, NBD2 arrangement with respect to NBD1 were assessed by calculating RMSDs focusing on NBDs either separately or using NBD1 position as reference over the trajectories. Interestingly, NBD1 secondary structure is conserved during the hydrolysis (ΔRMSD being 0.86 ± 0.13 Å, see Fig. 34c). NBD2 secondary structure slightly deviates more while comparing pre- and post-hydrolysis (ΔRMSD being 1.36 ± 0.12 Å). However, NBD1-aligned structure clearly highlights the structural deviation of post-catalytic NBD dimer as compared to pre-catalytic state. The pivotal role of NBS1 is confirmed as well as the asymmetric NBD motion along the transport cycle[27]. For instance, NBD2 and NBS2 exhibit larger RMSD values, respectively 3.33 ± 0.45 Å and 2.55 ± 0.32 Å. Moreover, visual inspection clearly underlines the opening of NBD dimer on the side of NBS2, as previously suggested[2].

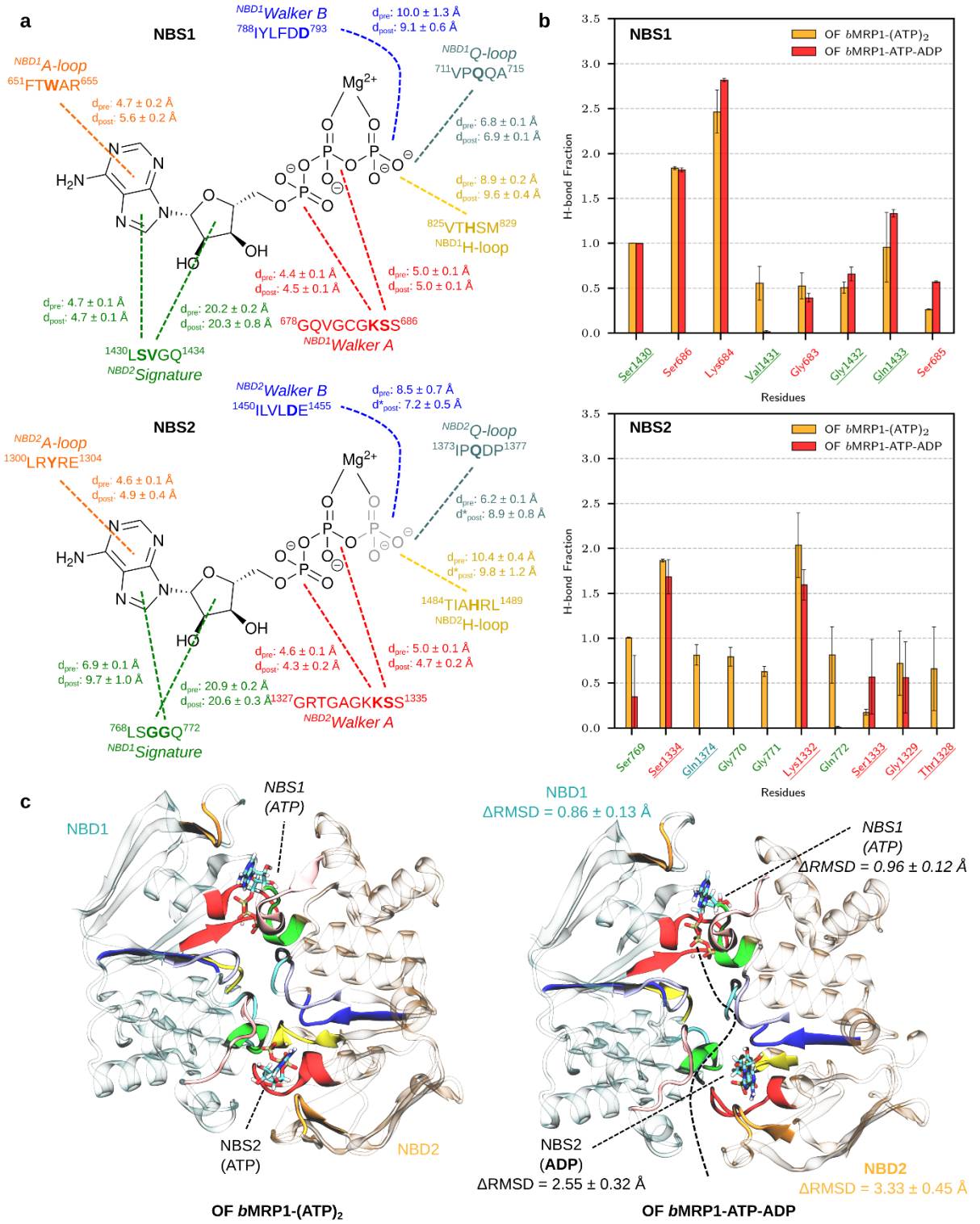



**Figure 34. Nucleotide-binding sites in the pre- and post-hydrolysis states of OF *b*MRP1.**

**a)** Important distances between the nucleotides and key residues in the nucleotide-binding sites. **b)** Calculated H-bond networks between nucleotides and NBS1 (top) or NBS2 (down). **c)** RMSD of NBD1, NBD2 and NBS1, NBS2 of the post-hydrolysis state compared to the pre-hydrolysis state. The post-hydrolysis state is more open as it is indicated by the dashed line. Conserved motives are coloured as follows: Walker A red, Walker B blue, signature motif green, A-loop orange, Q-loop teal, X-loop pink, D-loop cyan, and H-loop yellow.

### IV.2.2.3. The role of the lipids

As observed for IF conformations[2], lipid composition does not significantly affect the local structure of the present states. For instance, regardless of membrane components, local free energy minima obtained from the InfleCS[28,29] approach are located in similar conformational space. Only, slight differences were observed. For example, even though EC opening remains smaller for post-catalytic state than for pre-catalytic state, protein dynamics tends to be slightly larger for PE-free lipid bilayers than for POPC:POPE:Chol (2:1:1) lipid bilayer (Supplementary Fig. 12). IC opening was not observed in MD simulations in none of the simulations independently on the lipid bilayer (Fig. 33a, see Supplementary Fig. 10-12). Interestingly, H-bond analyses showed that both NBS nucleotide binding modes of post-hydrolysis state remains more similar than pre-hydrolysis state in PE-free lipid bilayers (Fig. 34b, Supplementary Fig. 13). Overall dynamic cross-correlation between residues show less correlation in the post-hydrolysis state in POPC:POPE:Chol (2:1:1) membrane; however, this difference is not observable in pure POPC and POPC:Chol (3:1) lipid composition (Supplementary Fig. 14). Cryo-EM OF structures were resolved including three cholesterol molecules that were suggest to actively participate to the allosteric communication between substrate binding pocket and NBD. As observed for pre-hydrolysis state, resolved cholesterol remains in close contact to pre-TMH7 elbow helix all along the simulation (Fig. 35).

Interestingly, MD simulations reveals plausible PE binding spots. It shed light on key PE-protein H-bond interactions and thus hotspots which are observed in the lower leaflet. This is in agreement with the known *in vivo* asymmetry for cell membrane composition, since PE lipids are mostly located in the lower leaflet than the upper leaflet. However, PE-binding seems to be more favourable prior to ATP hydrolysis, even though such assumption should be considered carefully given that the timescale is likely not sufficient to provide an accurate sampling of surrounding lipid motions.

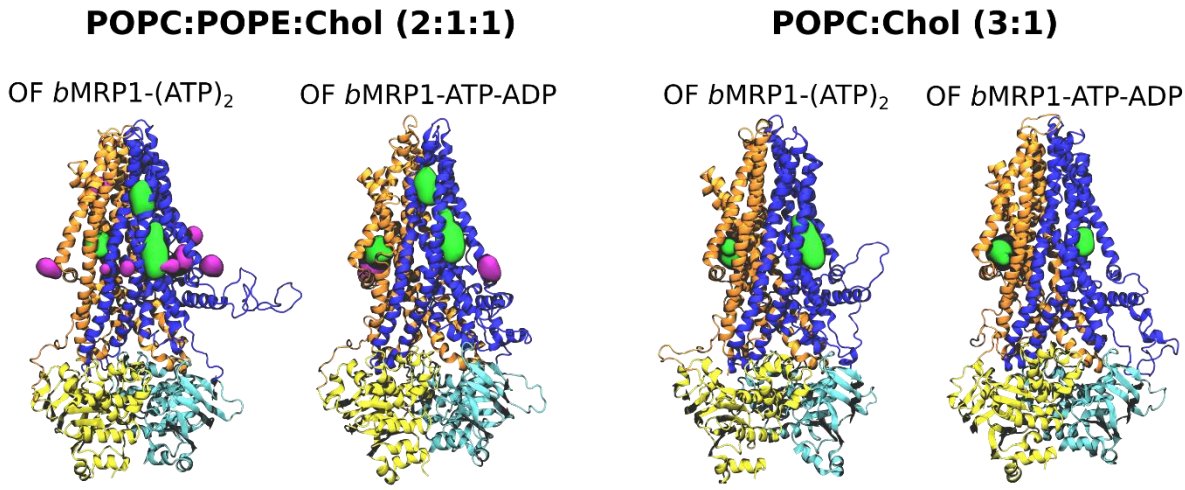



**Figure 35. Key lipids in the pre- and post-hydrolysis states.**

Cholesterol (green) and PE lipid (magenta) hotspots defined by presence likelihood higher than 50%.

### IV.2.3. Discussion

*b*MRP1 is the only drug exporter in ABCC family for which there are resolved cryo-EM structures available, so far. The International Transporter Consortium[32,33] has stressed out the clinical and pharmacological role of transporters of this family. Knowledge about MRP1 dynamics and functions is still fragmented, in contrast to its cousin ABCB1/P-gp. In the present work, all-atom MD simulations were performed in the pre- and post-hydrolysis states, respectively OF *b*MRP1-$(ATP)_2$ and OF *b*MRP1-ATP-ADP, in different bilayers including PC and PE lipids as well as cholesterol, to compare their dynamics and structures.

Global conformational dynamics in POPC:POPE:Chol (2:1:1) shows unfavoured dynamic EC opening, smaller NBD twist, and higher NBD distance indicating that the post-hydrolysis state adopt a close-conformation rather than outward-open conformation. EC closing and IC opening are likely necessary for the OF-to-IF conformational changes. Present results as well as our previous studies[2] clearly indicate that outward-facing open conformation is a transient state which may exist only for the substrate release. We here therefore suggest that the use of "OF" conformation for such states should perhaps be replaced by the so-called "close-conformation" (cc) as stated for ABCB4[34]. It is worth mentioning, that only a few V-fold ABC transporters were resolved with a clear open OF conformation (*e.g.*, the historical Sav1866[35]) as shown in Supplementary Fig. 15.

Our present work also highlighted the structural dynamics of the $L_0$ domain. MD simulations suggest the existence of a dynamic interaction with NBD1, which tends to be more pronounced in the post-hydrolysis state. Even though no clear clue was obtained to better understand its role, previous experiments have shown that the lasso motif is required for MRP1 function[5,6]. Surprisingly, it has been shown that co-expression of $L_0$ is however sufficient for substrate transport by MRP1[5]. Altogether with our present results, this may suggest that simply the $L_0$-NBD1 non-covalent interactions may play a role in triggering the OF-to-IF large scale conformational change. However, such assumption will require further joint experimental and computational investigations.

Particular attention has been paid to the NBDs. NBD rocker-switch angle slightly but significantly deviate by *ca.* 5° between pre- and post-hydrolysis state. In agreement with structural insights observed in P-gp, hydrolysis induces local conformational changes in the conserved motifs, mostly in the signature motif[36]. This is in agreement with P-gp for which two EC opening main populations were observed depending on A-loop position[37]. However, in MRP1, calculated EC opening deviation remain lower than P-gp. This might be explained by a different inherent behaviour of MRP1, which may more pronounced owing to the fact the NBD1 is degenerated. Indeed, as shown in our former study[2], NBD1 coupling helix does not adopt the "ball-and-socket" conformation, leading to lower transduction signal from NBD to TMDs. Our present calculations also confirm that ATP-binding in NBS1 acts likely as pivot since NBD1 and NBS1 are almost structurally identical regardless of the hydrolysis state. NBD2 exhibit pseudo-asymmetric structural deformation on the NBS2 side after ATP-hydrolysis. This was confirmed by the monitoring of the local distances, as well as non-covalent interactions between the nucleotides and the protein (including H-bond, van der Waals and Coulomb potentials) that suggest stronger interaction between the ATP and conserved motifs in NBS1 than in NBS2 even in the post-hydrolysis state. This strengthen the hypothesis suggesting that ATP release from NBS1 is unlikely even during several transport cycles[27]. Finally, our present MD simulations also strengthen the expected central role of PE lipids in the inner leaflet as well as cholesterol in the so-called Cholesterol recognition amino acid consensus (CRAC) or

inverted (CARC) motifs. Indeed, these motifs were also observed in other ABC transporters as well for MRP1[38]. However, their active role still requires to be elucidated.

### IV.2.4. References


1. Wang, L. *et al.* Characterization of the kinetic cycle of an ABC transporter by single-molecule and cryo-EM analyses. *eLife* **9**, e56451 (2020).
2. Tóth, Á., Janaszkiewicz, A., Crespi, V. & Meo, F. D. On the interplay between lipids and asymmetric dynamics of an NBS degenerate ABC transporter. 2022.05.16.492073 Preprint at https://doi.org/10.1101/2022.05.16.492073 (2022).
3. Kroll, T., Prescher, M., Smits, S. H. J. & Schmitt, L. Structure and Function of Hepatobiliary ATP Binding Cassette Transporters. *Chem Rev* (2020) doi:10.1021/acs.chemrev.0c00659.
4. Johnson, Z. L. & Chen, J. ATP Binding Enables Substrate Release from Multidrug Resistance Protein 1. *Cell* **172**, 81-89.e10 (2018).
5. Johnson, Z. L. & Chen, J. Structural Basis of Substrate Recognition by the Multidrug Resistance Protein MRP1. *Cell* **168**, 1075-1085.e9 (2017).
6. Bickers, S. C., Benlekbir, S., Rubinstein, J. L. & Kanelis, V. Structure of Ycf1p reveals the transmembrane domain TMD0 and the regulatory region of ABCC transporters. *Proc Natl Acad Sci USA* **118**, e2025853118 (2021).
7. Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Prot Bioinforma* **54**, (2016).
8. Case, D. A. *et al.* AMBER 2020. (2020).
9. Case, D. A. *et al.* AMBER 2018.
10. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J Comput Chem* **29**, 1859–1865 (2008).
11. Wu, E. L. *et al.* CHARMM-GUI *Membrane Builder* toward realistic biological membrane simulations. *J Comput Chem* **35**, 1997–2004 (2014).
12. Price, D. J. & Brooks, C. L. A modified TIP3P water potential for simulation with Ewald summation. *J Chem Phys* **121**, 10096–10103 (2004).
13. Maier, J. A. *et al.* ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **11**, 3696–3713 (2015).
14. Meagher, K. L., Redman, L. T. & Carlson, H. A. Development of polyphosphate parameters for use with the AMBER force field. *J Comput Chem* **24**, 1016–1025 (2003).
15. Galindo-Murillo, R. *et al.* Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* **12**, 4114–4127 (2016).
16. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J Chem Phys* **79**, 926–935 (1983).
17. MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**, 3586–3616 (1998).
18. Joung, I. S. & Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *J Phys Chem B* **112**, 9020–9041 (2008).

19. Joung, I. S. & Cheatham, T. E. Molecular Dynamics Simulations of the Dynamic and Energetic Properties of Alkali and Halide Ions Using Water-Model-Specific Ion Parameters. *. Phys Chem B* **113**, 13279–13290 (2009).

20. Roe, D. R. & Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **9**, 3084–3095 (2013).

21. Gowers, R. J. *et al.* MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. *Proceedings of the 15th Python in Science Conference* 98–105 (2016) doi:10.25080/Majora-629e541a-00e.

22. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **32**, 2319–2327 (2011).

23. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng* **9**, 90–95 (2007).

24. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33–38, 27–28 (1996).

25. Moradi, M. & Tajkhorshid, E. Mechanistic picture for conformational transition of a membrane transporter at atomic resolution. *Pros Natl Acad Sci USA* **110**, 18916–18921 (2013).

26. Hofmann, S. *et al.* Conformation space of a heterodimeric ABC exporter under turnover conditions. *Nature* **571**, 580–583 (2019).

27. Stockner, T., Gradisch, R. & Schmitt, L. The role of the degenerate nucleotide binding site in type I ABC exporters. *FEBS Lett* **594**, 3815–3838 (2020).

28. Westerlund, A. M. & Delemotte, L. InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *J Chem Theory Comput* **15**, 6752–6759 (2019).

29. Westerlund, A. M., Harpole, T. J., Blau, C. & Delemotte, L. Inference of Calmodulin's Ca2+-Dependent Free Energy Landscapes via Gaussian Mixture Model Validation. *J Chem Theory Comput* **14**, 63–71 (2018).

30. Wen, P.-C., Verhalen, B., Wilkens, S., Mchaourab, H. S. & Tajkhorshid, E. On the Origin of Large Flexibility of P-glycoprotein in the Inward-facing State. *J Biol Chem* **288**, 19211–19220 (2013).

31. Zgarbová, M., Otyepka, M., Šponer, J., Hobza, P. & Jurečka, P. Large-scale compensation of errors in pairwise-additive empirical force fields: comparison of AMBER intermolecular terms with rigorous DFT-SAPT calculations. *Phys Chem Chem Phys* **12**, 10476–10493 (2010).

32. Zamek-Gliszczynski, M. J. *et al.* Transporters in Drug Development: 2018 ITC Recommendations for Transporters of Emerging Clinical Importance. *Clin Pharmacol Ther* **104**, 890–899 (2018).

33. Giacomini, K. M. *et al.* Membrane transporters in drug development. *Nat Rev Drug Discov* **9**, 215–236 (2010).

34. Olsen, J. A., Alam, A., Kowal, J., Stieger, B. & Locher, K. P. Structure of the human lipid exporter ABCB4 in a lipid environment. *Nat Struct Mol Biol* **27**, 62–70 (2020).

35. Dawson, R. J. P. & Locher, K. P. Structure of a bacterial multidrug ABC transporter. *Nature* **443**, 180–185 (2006).

36. Dehghani-Ghahnaviyeh, S., Kapoor, K. & Tajkhorshid, E. Conformational changes in the nucleotide-binding domains of P-glycoprotein induced by ATP hydrolysis. *FEBS Lett* **595**, 735–749 (2021).

37. Verhalen, B. *et al.* Energy transduction and alternating access of the mammalian ABC transporter P-glycoprotein. *Nature* **543**, 738–741 (2017).

38. Thangapandian, S., Kapoor, K. & Tajkhorshid, E. Probing cholesterol binding and translocation in P-glycoprotein. *Biochim Biophy Acta Biomembr* **1862**, 183090 (2020).

## Chapter V. Effect of CFTR correctors on the traffic and the function of intracellularly retained ABCB4 variants

*Amel Ben Saad[1,2], Virginie Vauthier[2,3], Ágota Tóth[4], Angelika Janaszkiewicz[4], Anne-Marie Durand-Schneider[2], Alix Bruneau[2,5], Jean-Louis Delaunay[2], Martine Lapalus[1], Elodie Mareux[1], Isabelle Garcin[1], Emmanuel Gonzales[1,6], Chantal Housset[2,7], Tounsia Aït-Slimane[2], Emmanuel Jacquemin[1,6], Florent Di Meo[4], Thomas Falguières[1]*

[1] Inserm, Université Paris-Saclay, Physiopathogénèse et traitement des maladies du foie, UMR_S 1193, Orsay, France
[2] Inserm, Sorbonne Université, Centre de Recherche Saint-Antoine (CRSA), UMR_S 938, Institute of Cardiometabolism and Nutrition (ICAN), Paris, France
[3] Université de Paris, Institut Cochin, Inserm U1016, CNRS UMR 8104, Paris, France
[4] Inserm, Université de Limoges, UMR 1248 IPPRITT, Limoges, France
[5] Department of Hepatology & Gastroenterology, Charité Universitätsmedizin Berlin, Berlin, Germany
[6] Assistance Publique -Hôpitaux de Paris, CHU Bicêtre, Paediatric Hepatology & Paediatric Liver Transplant Department, Reference Center for Rare Paediatric Liver Diseases, FILFOIE, ERN Rare-Liver, Faculté de Médecine Paris-Saclay, Le Kremlin-Bicêtre, France
[7] Assistance Publique -Hôpitaux de Paris, Hôpital Saint-Antoine, Reference Center for Inflammatory Biliary Diseases and Autoimmune Hepatitis, FILFOIE, ERN Rare-Liver, Paris, France

**Correspondence**
Thomas Falguières, PhD, UMR_S 1193 Inserm/Université Paris Saclay, Rue des Adèles –Bâtiment 440 –F-91405 Orsay cedex, France.
Email: thomas.falguieres@inserm.fr

**Published, DOI: 10.1111/liv.14839**

This study investigates the role of 16 (used alone or in combination) previously reporter CFTR correctors that rescue the plasma membrane targeting of F508del-CFTR on another ATP-binding cassette transporter, namely on ABCB4 from the ABC B-family. It was a side project, strongly connected to my main project as ABCB4 is also important in pharmacology, and since it is an ABC transporter, it shows similarities to MRP1. Moreover, during this project, I had the opportunity to learn about molecular docking method using AutoDock Vina.[25] ABCB4 transport phosphatidylcholine to the bile canaliculi. Genetic variations of ABCB4 are associated to diseases from which the progressive familial intrahepatic cholestasis type 1 and 3 are the most important. The aim of this study is to find new therapeutical approaches. To achieve it the effect of known CFTR/ABCC7 correctors on ABCB4 was investigated using experimental techniques supported by theoretical simulations. As I was participating in the theoretical part, details of the experiments are not included in this thesis.

**Abbreviations**: ABC, ATP-binding cassette; ABCB4cc, ABCB4-closed conformation; ABCB4if, ABCB4-inward facing conformation; ATP, adenosine triphosphate; CFTR, cystic fibrosis transmembrane conductance regulator; CH, coupling helix; ER, endoplasmic reticulum; HEK, human embryonic kidney; NBD, nucleotide binding domain; PC, phosphatidylcholine; PFIC3, progressive familial intrahepatic cholestasis type 3; TM, transmembrane helix; UDCA, ursodeoxycholic acid; WT, wild type.

**Abstract**

**Background & aim**: ABCB4 is expressed at the canalicular membrane of hepatocytes. This ATP-binding cassette (ABC) transporter is responsible for the secretion of phosphatidylcholine into bile canaliculi. Missense genetic variations of ABCB4 are correlated with several rare cholestatic liver diseases, the most severe being progressive familial intrahepatic cholestasis type 3 (PFIC3). In a repurposing strategy to correct intracellularly retained ABCB4 variants, we tested 16 compounds previously validated as cystic fibrosis transmembrane conductance regulator (CFTR) correctors.

**Methods**: The maturation, intracellular localization and activity of intracellularly retained ABCB4 variants were analyzed in cell models after treatment with CFTR correctors. In addition, *in silico* molecular docking calculations were performed to test the potential interaction of CFTR correctors with ABCB4.

**Results**: We observed that the correctors C10, C13, and C17, as well as the combinations of C3 + C18 and C4 + C18, allowed the rescue of maturation and canalicular localization of four distinct traffic-defective ABCB4 variants. However, such treatments did not permit a rescue of the phosphatidylcholine secretion activity of these defective variants and were also inhibitory of the activity of wild type ABCB4. *In silico* molecular docking analyses suggest that these CFTR correctors might directly interact with transmembrane domains and/or ATP-binding sites of the transporter.

**Conclusion**: Our results illustrate the uncoupling between the traffic and the activity of ABCB4 because the same molecules can rescue the traffic of defective variants while they inhibit the secretion activity of the transporter. We expect that this study will help to design new pharmacological tools with potential clinical interest.

**KEYWORDS**

ABC transporters, bile secretion, cell models, cholestatic liver diseases, molecular docking, targeted pharmacotherapy

**Lay summary**

- CFTR correctors rescue the maturation and the in vitro localization of four distinct ER-retained ABCB4 variants identified in patients.
- CFTR correctors do not rescue the function of these variants even if they are relocalized at the plasma membrane.
- CFTR correctors inhibit the function of wild type ABCB4.
- *In silico* molecular docking analyses suggest direct interactions of CFTR correctors with functional domains of ABCB4.

## V.1. Introduction

Bile secretion is an essential function of the liver for lipid digestion and absorption as well as the elimination of xenobiotics and endogenous metabolites. This function mainly depends on transporters localized at the apical (or canalicular) plasma membrane of hepatocytes such as ATP-binding cassette subfamily B member 4 (ABCB4), also known as multidrug resistance protein type 3 (MDR3). ABCB4 belongs to the superfamily of ATP-binding cassette (ABC) transporters which are transmembrane proteins able to bind and hydrolyze ATP in order to fulfil their biological functions[1,2]. The expression of ABCB4 is restricted to the canalicular membrane of hepatocytes[3] and its role is to ensure secretion of phosphatidylcholine (PC) into bile[4]. With the co-secreted cholesterol and bile salts, PC forms mixed micelles in the aqueous environment of bile, thus avoiding the formation of cholesterol gallstones in the biliary tract as well as cell membrane damage by free bile acids on the canalicular membrane of hepatocytes and the apical membrane of cholangiocytes (for reviews, see[1,5]). Until now, more than 220 distinct variations of the ABCB4-encoding gene have been reported, mostly in patients with cholestatis and cholelithiasis (see https://evs.gs.washington.edu/EVS/ and http://abcmutations.hegelab.org/). These genetic variations of ABCB4 can affect the expression, the traffic, the function or the stability of the protein. Indeed, we have previously proposed a classification of these variants into five distinct classes: class I with no protein expression, class II with intracellular retention; class III with functional defects, class IV with stability impairment and class V with no apparent defect[6,7]. More details about ABCB4 function, genetic disorders and subsequent diseases can be found in the recent review by Kroll *et al.*[8].

The most severe ABCB4-related disease is progressive familial intrahepatic cholestasis type 3 (PFIC3), which is a rare autosomal recessive disease affecting homozygous or compound heterozygous patients during childhood[9]. PFIC3 appears during the first months of life and is characterized by chronic cholestasis, jaundice and pruritus[10]. The only pharmacological treatment for PFIC3 patients is the administration of ursodeoxycholic acid (UDCA), a bile acid with low hydrophobicity which renders the bile less toxic[11,12]. However, more than half of PFIC3 patients display no or little response to UDCA treatment and worsening of the disease most often requires liver transplantation[10,13]. Therefore, the therapeutic challenge is to identify new targeted pharmacotherapies as an alternative to liver transplantation for patients with severe forms of ABCB4-related diseases.

Looking for correctors of class II endoplasmic reticulum (ER)-retained ABCB4 variants, we have recently shown that structural analogues of roscovitine are able to partially rescue the traffic, localization and function of these variants[14]. In the present study, in order to pursue this repositioning strategy, we were interested in correctors previously shown to rescue the plasma membrane targeting of the F508del variant of ABCC7/CFTR (cystic fibrosis transmembrane conductance regulator), the ABC transporter mutated in patients with cystic fibrosis (see Supplementary Table S1, and references therein). While we observed that CFTR correctors are able to rescue the maturation and the localization of several ER-retained ABCB4 variants, we report that they are not able to rescue their PC secretion activity and that they also inhibit the activity of the wild type transporter. These two effects might be due to direct interactions of CFTR correctors with functional domains of ABCB4 such as transmembrane domains and/or nucleotide binding domains, as suggested by *in silico* molecular docking calculations.

## V.2. Materials and methods

*Details are reported in sections 2.1-2.4 of the article.*

*Shortly, plasmids encoding the wild type ABCB4 or an ER-retrained variant produced by site-directed mutagenesis were used in human embryonic kidney (HEK-293) cells and human hepatocellular carcinoma HepG2 cells. Cytotoxicity of the CFTR correctors was assessed. Immunoblots and indirect fluorescence analyses were performed to compare wild type and mutant ABCB4 expression. The ABCB4-mediated phosphatidylcholine secretion was measured.*

### V.2.1. *In silico* molecular docking and calculations

Two conformations of human ABCB4 (*h*ABCB4) were considered for molecular docking calculations, namely inward-facing and closed conformations ($ABCB4^{if}$ and $ABCB4^{cc}$, respectively). $ABCB4^{if}$ was built by homology modeling using the inward-facing human ABCB1 (*h*ABCB1) protein data bank structure (PDB ID: 6QEX),[20] given the high sequence identity between *h*ABCB1 and *h*ABCB4 (76.8%, see Fig. S1). $ABCB4^{cc}$ model was built using the recently resolved cryo-EM structure of $ABCB4^{cc}$ structure trapped in ATP-bound state (PDB ID: 6S7P).[21] The non-resolved extracellular loop (residues 85-104) of $ABCB4^{cc}$ was built based on the resolved hABCB1 structure. For both models, L1-linker connecting nucleotide binding domain (NBD) 1 to transmembrane helix (TM) 7, and the N- and C-terminal domains were not included in the present study owing to their absence in both resolved *h*ABCB1 and *h*ABCB4 structures (Fig. S1). Present models did not include the ATP molecules. Homology modeling was performed using the Modeller software version 9.23.[22-24] Prior to docking calculations, both models were minimized in solvated membrane lipid bilayers that were removed for docking calculations (for further details, see Appendix S1).

Molecular docking calculations were carried out using the Autodock Vina software.[25] Given the absence of a priori knowledge about the interaction mechanism of CFTR correctors, the following procedure was performed to extensively sample plausible binding sites of CFTR correctors to both $ABCB4^{if}$ and $ABCB4^{cc}$ conformers. First, so-called "blind" docking calculations were performed with all ligands in which the whole protein was considered as intentionally too large volume search space centered on the protein center of mass for both $ABCB4^{if}$ and $ABCB4^{cc}$ conformations (Table S2A). A set of 20 replicas per CFTR corrector was performed, using different random seeds, providing a total of approximately 6400 poses per ABCB4 conformation. These molecular poses exhibited clusters located at different spatial regions (see Fig. S2). Refined docking calculations were then performed in which smaller space search volumes were used (Fig. S3), which do not exceed the recommended search volume of 27 000 Å3 to ensure the reliability of docking calculation results. Again, 20 replicas of refined molecular docking calculations were performed for each CFTR corrector. For blind and refined docking calculations, exhaustiveness was set to 40 and 100, respectively, to increase the computational effort used during molecular pose search.[26] The maximum number of poses was set to 100 by replica, and the initial affinity cutoff was defined at 15 $kcal.mol^{-1}$ with respect to the top pose of a given replica. The list of flexible residues allowed for each volume search space is reported in Table S2B. Given the large number of plausible flexible residues, only sidechain $C_\beta$-$C_\gamma$ bonds were allowed to rotate as a compromise within the limit of 32 rotatable bonds allowed by Autodock Vina software.

Additional information about ABCB4 and ligand preparations as well as analyses is provided in Appendix S1.

### V.2.2. Statistics

Graphics and one-way ANOVA tests were performed using Prism version 7.00 (GraphPad Software). A P value of less than 0.05 was considered significant with *P <.05; **P <.01; ***P <.005; ns: not significant. Symbols always indicate the comparison between the control (vehicle-treated) and the other tested conditions.

## V.3. Results

*Details are reported in sections 3.1-3.3 of the article.*

*To sum up, the first step was to make sure that none of the components is cytotoxic. That was followed by the study of the maturation and the localization of the wild type and four mutants. The results show that some of the correctors are able to rescue the traffic and maturation of all the tested variants. However, CFTR correctors do not significantly restore the function, moreover they inhibit the wild type.*

### V.3.1. Molecular docking calculations reveal different plausible binding regions of CFTR correctors within ABCB4

We hypothesized that the inhibition of ABCB4 activity by CFTR correctors could be related to their direct interaction with functional domains of the transporter. To test this possibility, we performed *in silico* molecular docking simulations. Blind calculations were first performed using large box parameters (Table S2A) with inward-facing ABCB4 ($ABCB4^{if}$) and its closed conformation ($ABCB4^{cc}$) wherein no a priori plausible structural binding regions were defined (for details, see Materials and Methods). This provided 6398 and 6395 molecular poses for $ABCB4^{if}$ and $ABCB4^{cc}$, respectively. Molecular poses were then spatially clustered, revealing a single binding site for CFTR correctors within $ABCB4^{if}$ while three binding sites were obtained for $ABCB4^{cc}$ (Fig. S2). In $ABCB4^{if}$, 95% of molecular poses are located in ABCB4 protein chamber (Table S3), expected to be the canonical phospholipid binding site region.[21] Regarding $ABCB4^{cc}$, 81% of molecular poses can be initially divided into three distinct regions, namely an alternative site at the lipid–protein interface (30% of molecular poses) and the two known ATP-binding sites at both NBD interfaces (15 and 36% for ATP-binding sites 1 and 2, respectively – see Table S3). Given the low sensitivity of blind docking calculations, regions obtained from blind docking calculations on $ABCB4^{if}$ were also explored for refined docking calculations on $ABCB4^{cc}$ and vice versa (Fig. S3). It is worth mentioning that no molecular poses were observed for ATP-binding site regions on $ABCB4^{if}$, which makes sense considering the large distance between the two NBDs in this conformation. Altogether, refined molecular docking calculations on $ABCB4^{if}$ highlight the relevance of the so-called protein chamber (Fig. 36A). Regarding $ABCB4^{cc}$, molecular docking calculations suggested six possible regions for CFTR corrector binding; namely protein chamber, ATP-binding sites 1 and 2, and three alternative sites 1, 2, and 3 at the lipid–protein interface (Fig. S3 and Fig. 36A).

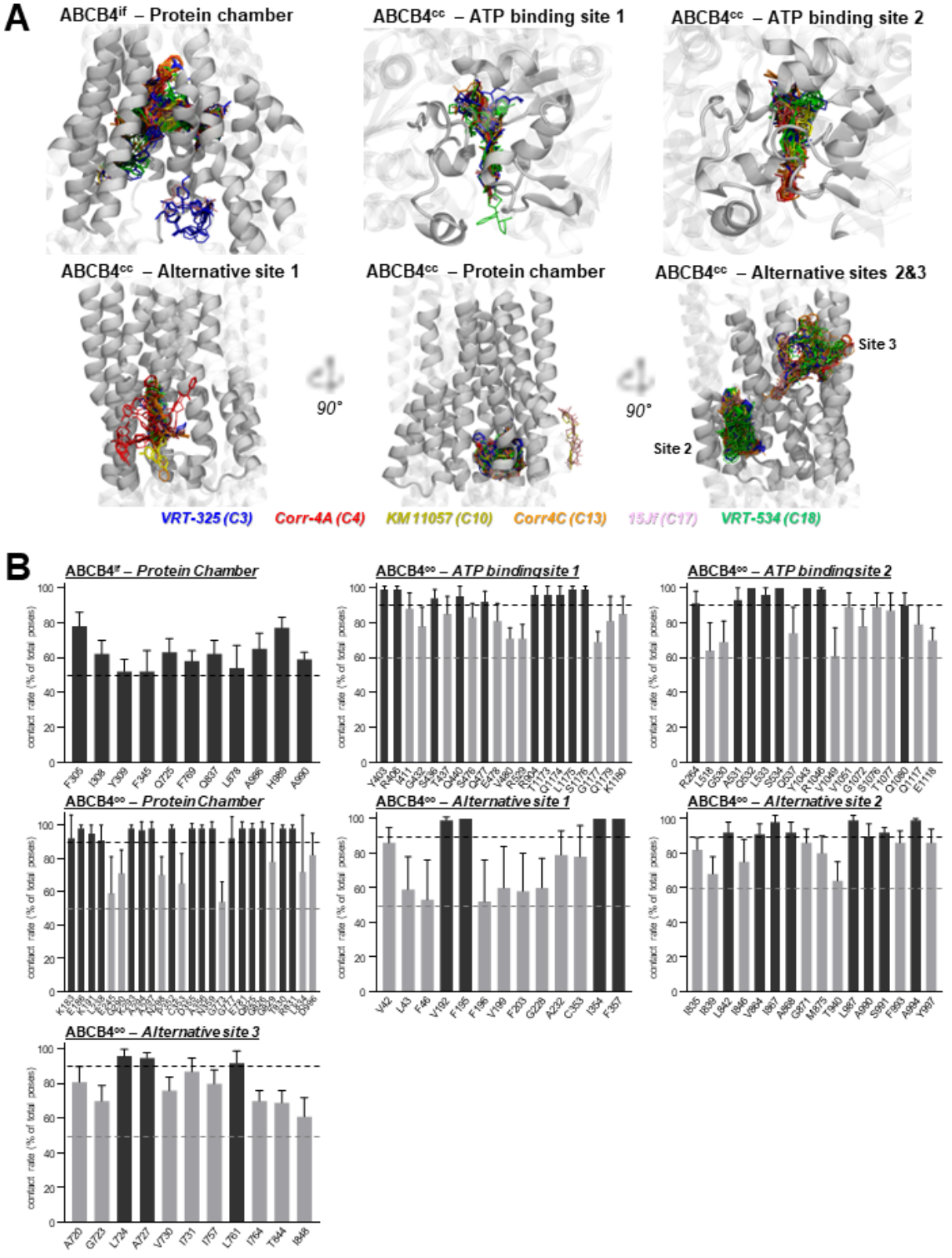



**Figure 36. In silico molecular docking of CFTR correctors suggests their direct interaction with ABCB4.**

**(A)** In silico molecular docking of the six color-coded CFTR correctors was performed with ABCB4 in inward-facing (ABCB4if) and closed conformation (ABCB4cc) models (see Materials and Methods for details). For ABCB4if, a single site of potential interaction was identified in the protein chamber, whereas three sites were identified for

ABCB4cc in an alternative site and the two ATP-binding sites. For each CFTR corrector and for each site, the 50 poses with the highest affinities from 20 independent simulations are shown. **(B)** For each of the six CFTR correctors of interest, the contact rate of ABCB4 residues was calculated for the four sites of interest shown in A (see Materials and Methods for details). For the sake of clarity, only residues for which the average contact rates (± SD) were higher than the thresholds indicated by the dashed lines are shown.

Thorough analysis over all molecular poses was then performed for each binding site to stress out key residues possibly involved in CFTR corrector binding to ABCB4. Only poses exhibiting binding affinity difference with respect to top-ranked molecular poses below 2.0 kcal.mol-1 were considered as already described.[31] In total, a set of 38 670 molecular poses over 50 482 were obtained from refined molecular docking calculations (Table S4). Accounting aforementioned in vitro results, special attention was paid to selected CFTR correctors, namely C3, C4, C10, C13, C17, and C18 (Fig. 36A). Contact analyses using a 4.5 Å cutoff were performed allowing to decipher key residues to CFTR corrector binding in the four regions of interest in ABCB4 (Fig. 36B). It is worth mentioning that contact analyses performed over the whole set of CFTR correctors exhibit similar profiles with respect to selected CFTR correctors (Fig. S4). Eleven residues are in contact with the selected CFTR correctors in at least 50% of calculated molecular poses in ABCB4$^{if}$ protein chamber (Fig. 36B). Most of these residues are either aromatic (histidine, phenylalanine, and tyrosine) or aliphatic (alanine, leucine), and presumably involved in substrate binding (Fig. 37A), as proposed recently for His989 and Ala99021 in ABCB4$^{if}$ protein chamber (underlined residues in Fig. 37A). Molecular docking calculations in ABCB4$^{cc}$ alternative site 1 clearly suggested the key role of Val192, Phe195, Ile354, and Phe357, for which contacts are higher than 90% (Fig. 36B and 37A). Likewise, binding to alternative sites 2 and 3 involved mostly contacts with aliphatic residues (Leu842, Val864, Ile867, Ala868, Leu987, Ala990, Ser991, and Ala994 for alternative site 2 and Leu724, Ala727 and Leu761 for alternative site 3). Contact analyses of ABCB4$^{cc}$ ATP-binding sites and protein chamber exhibited more residues per site owing to smaller volumes as well as more important steric hindrance. Indeed, molecular docking calculations revealed seventeen, ten, and seven residues for which contact rate is higher than 90% in ABCB4$^{cc}$ protein chamber, ATP-binding sites 1 and 2, respectively (Fig. 36B). Regarding ATP-binding sites, residue profile again included not only aromatic amino acids (*e.g.*, NBD A-loop residues Tyr403 and Tyr1043) but also polar residues such as arginine (Arg406 and Arg1046) (Fig. 37A). Residue profile for ABCB4$^{cc}$ protein chamber included mostly polar and aliphatic residues (see Fig. 36B and 37A). For all ABCB4 binding sites, thorough analyses of non-covalent interactions (H-bond and van der Waals interactions) were performed, highlighting the key role of $\pi$-stacking interactions between CFTR correctors' aromatic rings and ABCB4 aromatic residues (Table S5). This was clearly shown for ABCB4$^{cc}$ alternative site 1 and ATP-binding sites 1 and 2 for which Phe195, Phe357, Tyr403 or Tyr1043 are involved in all π-stacked conformations. To a lesser extent, H-bonding were also shown to be relevant. For instance, H-bonding with Gln725 were involved in 21% of molecular poses in ABCB4$^{if}$ protein chamber. Likewise, H-bonding between CFTR correctors and Arg904 (ABCB4cc ATP-binding site 1), Arg1046, or Glu1080 (ABCB4$^{cc}$ ATP-binding site 2) are involved in 27, 21, and 21% of molecular poses, respectively (Table S5).

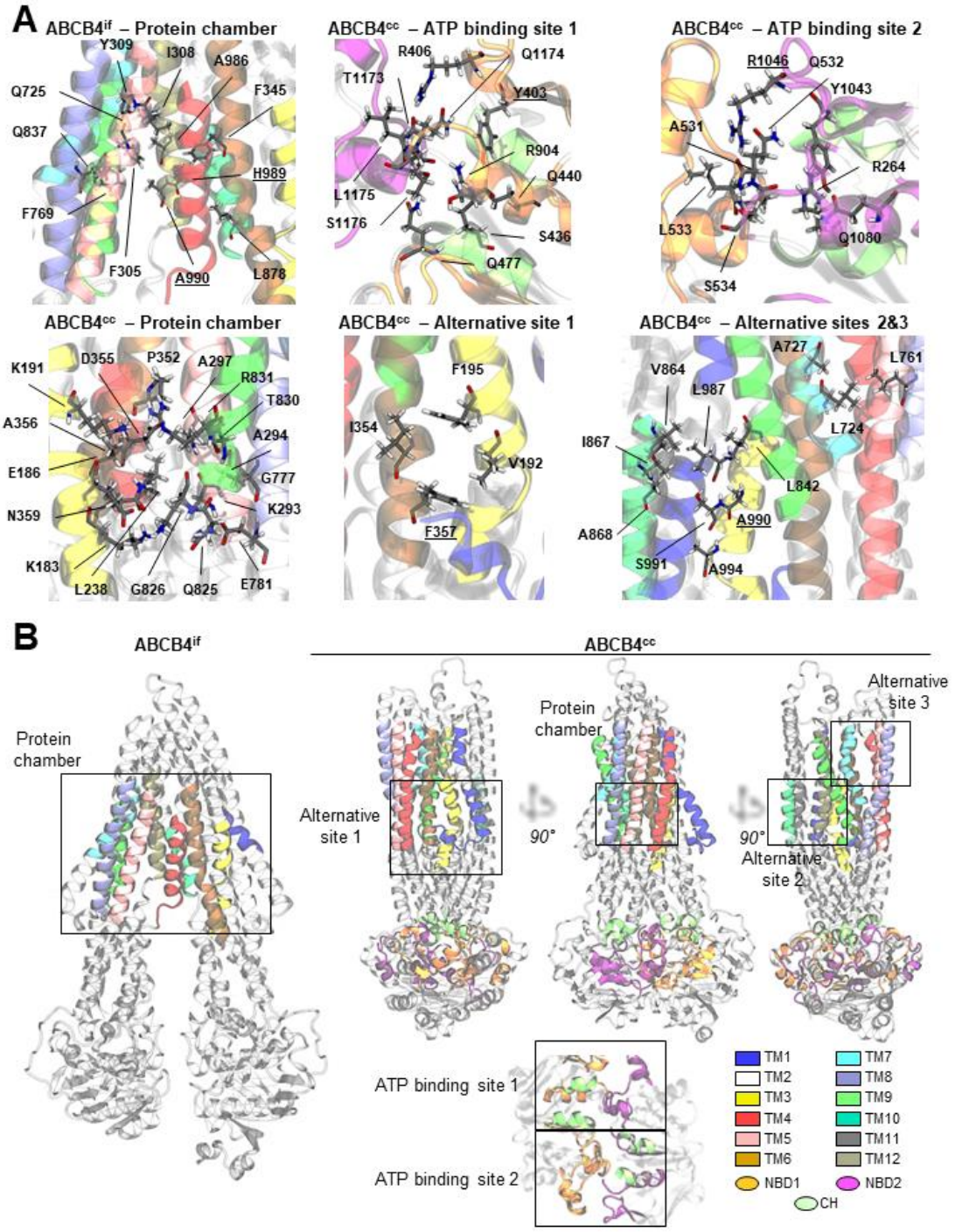



**Figure 37. In silico molecular docking of CFTR correctors reveals putative interaction residues and domains in ABCB4.**

**(A)** ABCB4 residues indicated in Fig. 36B with dark grey bars for which the contact rate is greater than 50% (ABCB4if -Protein chamber) or 90% (ABCB4cc) are represented with their lateral chains in the 3D structures of ABCB4if and ABCB4cc. Each transmembrane domain and nucleotide binding domain are represented in different colors, as well as the coupling helices (see Table S6). Underlined residues are further discussed in the Discussion section. **(B)** ABCB4 transmembrane domains (TM), nucleotide-binding domains (NBD), and coupling helices (CH), for which potential interaction with CFTR correctors is suggested in Fig. 36B and 37A, are highlighted with the indicated color code.

Finally, molecular docking calculations and aforementioned atomic-scaled analyses allowed to define putative binding sites in terms of secondary structure (Table S6). ABCB4$^{if}$ protein chamber involved the lipid-embedded regions of all TMs but TM11 (Fig. 37B). ABCB4$^{cc}$ alternative site 1 is defined by TM1, TM3, TM4, and TM6 (Fig. 37B). Alternative sites 2 and 3 are located on the other side of ABCB4$^{cc}$, involving TM1, TM4-6, TM7-9, and TM12 for site 2 and TM9, TM10-12 for site 3. Alternative sites 1 and 2 are interfacing with lower leaflet membrane while alternative site 3 is in contact with upper leaflet membrane. In addition to known residues involved in ATP binding, ABCB4$^{cc}$ ATP-binding site regions were also defined, including the coupling helices (CH), namely CHs between TM2 and 3 and TM10 and 11 for ATP-binding site 1, and CHs between TM4 and 5 and TM8 and 9 for ATP-binding site 2 (Fig. 37B). Altogether, these *in silico* molecular docking analyses suggest that CFTR correctors might directly interact with functional domains of ABCB4, which could explain their effect on ABCB4 activity.

### V.4. Discussion

Genetic variations of the phospholipid floppase ABCB4 are correlated with rare cholestatic liver diseases, the most severe form being PFIC3. Some of these genetic variations can cause ABCB4 misfolding, its retention in the ER and, as a result, the loss of its phospholipid floppase function.[6] In previous studies, we and others have demonstrated that small molecules such as cyclosporins,[6,16] 4-phenylbutyrate (4-PBA), curcumin,[32] and structural analogs of roscovitine[14] are able to partially rescue the intracellular traffic and the cell surface localization of ER-retained ABCB4 variants. However, cytotoxicity, inhibitory effect on ABCB4 function,[33] bioavailability, or the high used concentration of these molecules are major limitations for their therapeutic use. In the present study, to identify new correctors for ER-retained ABCB4 variants, we explored the potential effect of several molecules, developed as CFTR correctors, on four distinct ER-retained ABCB4 variants identified in patients (I490T, I541F, R545H, L556R mapped in Fig. S5). CFTR correctors, which were used in this study, are small molecules, mostly identified by high-throughput screening strategies to correct the traffic of F508del-CFTR, the most frequent genetic variation detected in patients with cystic fibrosis.[34] The extended use of these correctors is an attractive option for multiple rare diseases associated with protein misfolding and missorting. Interestingly, some of these molecules have been shown to successfully correct the intracellular localization and the function of other defective ABC transporters such as ABCA3[35] and ABCA4[28] and also ER-retained variants of ATP8B1[36] and α-sarcoglycan,[37] proteins without CFTR similarity. Here, we show that the correctors C10, C13, and C17 as well as the combinations of C3 + C18 and C4 + C18 can partially rescue the maturation and the plasma membrane targeting of four defective ABCB4 variants. Further investigations at the molecular level will be required to determine the capacity of these CFTR correctors to rescue a proper folding of the class II ER-retained ABCB4 variants studied here, thus allowing their plasma membrane targeting. The fact that the floppase activity of the I541F, I490T, and L556R variants of ABCB4 can be partially rescued by roscovitine analogs[14] indicates that these missense variants have intrinsic activity once rescued at the plasma membrane. However, the CFTR correctors tested here are not able to restore the function of the ER-retained ABCB4 variants and they also inhibit the floppase activity of ABCB4-WT. Although none of the compounds used in this study have been tested in vivo or in clinics, it is interesting to note that rare cases of cholestasis have been reported in patients taking sildenafil,[38,39] the parent molecule of C9 and C10, as well as cystic fibrosis patients treated with Orkambi, which contains VX-809/Lumacaftor – a derivative from C18/VX-534 (https://www.ema.europa.eu/en/documents/product-information/orkambi-epar-product-

information_en.pdf), arguing for a potential inhibitory effect of these molecules on canalicular ABC transporters. It is noteworthy to mention that this is not the first time that inhibitory effects of correctors on ABCB4 function are reported. Indeed, cyclosporin A, roscovitine and triazole compounds such as itraconazole were shown to inhibit ABCB4 floppase activity.[14,33,40] Sildenafil, the C10 parent molecule, has also been shown to inhibit the transport function of ABCB1,[41] which exhibits 86% sequence-based similarity with ABCB4.42 The PgpRules prediction tool (https://pgprules.cmdm.tw/)43 indicates that C3, C4, C10, C13, C17, and C18 might inhibit ABCB1 (data not shown), suggesting similar inhibitory mechanisms on ABCB4. It is important to note that PgpRules prediction tool is initially designed to predict ABCB1 inhibition. Given (i) the high sequence-based similarity between both transporters and (ii) the inhibitor overlaps for both transporters, one can expect that they might, at least partially, share inhibition mechanisms, despite the divergence of substrates translocated by these two transporters.[8] Furthermore, even if CFTR correctors belong to different chemical classes (quinolines, bithiazoles, quinazolines, pyrimidines),[34] they share features with pharmacophore models of ABCB1 inhibitors previously described as hydrophobic, hydrogen bond acceptor, aromatic ring centered with positive ionizable moieties.[44]

To explain how these CFTR correctors could correct missorting of ABCB4 variants and at the same time inhibit the floppase activity of the transporter, we suggest that these small molecules may directly interact with ABCB4. Such an interaction has already been reported for C18 with CFTR.45 To virtually test this hypothesis, we investigated the putative binding sites of CFTR correctors in ABCB4 by *in silico* docking simulations. Using two models of human ABCB4 in inward-facing and closed conformations ($ABCB4^{if}$ and $ABCB4^{cc}$, respectively), we identified seven distinct plausible binding regions for CFTR correctors: protein chambers in both $ABCB4^{if}$ and $ABCB4^{cc}$ models; three alternative sites at the protein–lipid interface and two ATP-binding sites in ABCB4cc. Analysis of contact residues with a 4.5 Å cutoff around the docked poses of CFTR correctors indicated that several key residues could be involved in CFTR corrector binding. In ABCB4if, we identified His989 and Ala990 as potential interactors with CFTR correctors, these residues having been recently shown to be crucial for PC binding and ABCB4 function.[21] We also detected Gln725, a residue conserved in ABCB1, and implicated in the binding of tariquidar.[46] Concerning ABCB4cc, we identified Phe357, Tyr403, and Arg1046 located in the alternative site 1, the ATP binding sites 1 and 2, respectively. These three residues were previously reported to be mutated in patients suffering from PFIC3 and low-phospholipid associated cholelithiasis syndrome,[47-49] suggesting their important role in ABCB4 expression and function. Moreover, Tyr403 and Arg1046 were also reported to be essential for ATP binding and hydrolysis.[21,50] It is thus tempting to speculate that the predicted direct interaction of CFTR correctors with key residues of ABCB4 may preclude its phospholipid floppase activity. This view is supported by the fact that sildenafil (the C10 parent molecule) and C3 impair the function of ABCB1 and CFTR, respectively.[41,51] It is important to note that present molecular docking calculations should be considered as an “idea generator” rather than a strong prediction. Thus, these aspects will require further joint experimental and theoretical confirmations, including μs-scaled molecular dynamics simulations. However, *in silico* findings might help in the design of site-directed mutagenesis variants to decipher small molecule-dependent inhibition mechanisms of ABCB4.

We propose that the trafficking rescue of ABCB4 variants mediated by CFTR correctors may be conferred through their direct interaction with critical domains of ABCB4, as suggested by our *in silico* docking analyses. We expect that our findings will help to design new non-inhibitory ABCB4 correctors that may lead to the preclinical development of pharmacological alternative

to UDCA treatment and liver transplantation. Indeed, if our hypothesis is confirmed, the next challenge will be to identify small molecules that are able to promote folding and trafficking of defective variants without compromising their activity. Thus, further investigations are eagerly needed to optimize bioavailability and affinity of these correctors, not only to promote traffic correction but also to maintain the function of the transporter.

*Acknowledgements, Conflict of interest, Autor contribution, Data availability, and ORCID are reported in the article.*

### V.5. References

[1] Falguières T, Aït-Slimane T, Housset C, Maurice M. ABCB4: insights from pathobiology into therapy. Clin Res Hepatol Gastroenterol. 2014;38(5):557-563.

[2] Srikant S, Gaudet R. Mechanics and pharmacology of substrate selection and transport by eukaryotic ABC exporters. Nat Struct Mol Biol. 2019;26(9):792-801.

[3] Smit JJ, Schinkel AH, Mol CA, et al. Tissue distribution of the human MDR3 P-glycoprotein. Lab Invest. 1994;71(5):638-649.

[4] Smit JJ, Schinkel AH, Oude Elferink RP, et al. Homozygous disruption of the murine mdr2 P-glycoprotein gene leads to a complete absence of phospholipid from bile and to liver disease. Cell. 1993;75(3):451-462.

[5] Boyer JL. Bile formation and secretion. Compr Physiol. 2013;3(3):1035-1078.

[6] Delaunay JL, Durand-Schneider AM, Dossier C, et al. A functional classification of ABCB4 variations causing progressive familial intrahepatic cholestasis type 3. Hepatology. 2016;63(5):1620-1631.

[7] Vauthier V, Housset C, Falguieres T. Targeted pharmacotherapies for defective ABC transporters. Biochem Pharmacol. 2017;136:1-11.

[8] Kroll T, Prescher M, Smits SHJ, Schmitt L. Structure and Function of Hepatobiliary ATP Binding Cassette Transporters. Chem Rev. 2020.

[9] Jacquemin E, De Vree JM, Cresteil D, et al. The wide spectrum of multidrug resistance 3 deficiency: from neonatal cholestasis to cirrhosis of adulthood. Gastroenterology. 2001;120(6):1448-1458.

[10] Jacquemin E. Progressive familial intrahepatic cholestasis. Clin Res Hepatol Gastroenterol. 2012;36(Suppl 1):S26-S35.

[11] Poupon R. Ursodeoxycholic acid and bile-acid mimetics as therapeutic agents for cholestatic liver diseases: An overview of their mechanisms of action. Clin Res Hepatol Gastroenterol. 2012;36(Suppl1):S3-S12.

[12] Achufusi TGO, Safadi AO, Mahabadi N. In: StatPearls. Treasure Island (FL), ed. Ursodeoxycholic Acid. StatPearls Publishing; 2020.

[13] Davit-Spraul A, Gonzales E, Baussan C, Jacquemin E. The spectrum of liver diseases related to ABCB4 gene mutations: pathophysiology and clinical aspects. Semin Liver Dis. 2010;30(2):134-146.

[14] Vauthier V, Ben Saad A, Elie J, et al. Structural analogues of roscovitine rescue the intracellular traffic and the function of ER-retained ABCB4 variants in cell models. Sci Rep. 2019;9(1):6653.

[15] Delaunay JL, Durand-Schneider AM, Delautier D, et al. A missense mutation in ABCB4 gene involved in progressive familial intrahepatic cholestasis type 3 leads to a folding defect that can be rescued by low temperature. Hepatology. 2009;49(4):1218-1227.

[16] Gautherot J, Durand-Schneider AM, Delautier D, et al. Effects of Cellular, chemical and pharmacological chaperones on the rescue of a trafficking-defective mutant of the ATP-binding cassette transporters ABCB1/ABCB4. J Biol Chem. 2012;287(7):5070-5078.

[17] Poupon R, Rosmorduc O, Boelle PY, et al. Genotype-phenotype relationships in the low-phospholipid associated cholelithiasis syndrome. A study of 156 consecutive patients. Hepatology. 2013;58(3):1105-1110.

[18] van Meerloo J, Kaspers GJ, Cloos J. Cell sensitivity assays: the MTT assay. Methods Mol Biol. 2011;731:237-245.

[19] Gautherot J, Delautier D, Maubert MA, et al. Phosphorylation of ABCB4 impacts its function: insights from disease-causing mutations. Hepatology. 2014;60(2):610-621.

[20] Alam A, Kowal J, Broude E, Roninson I, Locher KP. Structural insight into substrate and inhibitor discrimination by human P-glycoprotein. Science. 2019;363(6428):753-756.

[21] Olsen JA, Alam A, Kowal J, Stieger B, Locher KP. Structure of the human lipid exporter ABCB4 in a lipid environment. Nat Struct Mol Biol. 2019;27(1):62-70.

[22] Šali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234(3):779-815.

[23] Fiser A, Do RKG, Šali A. Modeling of loops in protein structures. Protein Sci. 2000;9(9):1753-1773.

[24] Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct. 2000;29(1):291-325.

[25] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455-461.

[26] Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, Olson AJ. Computational protein–ligand docking and virtual drug screening with the AutoDock suite. Nat Protoc. 2016;11(5):905-919.

[27] Lopes-Pacheco M, Boinot C, Sabirzhanova I, Morales MM, Guggino WB, Cebotaru L. Combination of correctors rescue deltaF508-CFTR by reducing Its association with Hsp40 and Hsp27. J Biol Chem. 2015;290(42):25636-25645

[28] Sabirzhanova I, Lopes-Pacheco M, Rapino D, et al. Rescuing trafficking mutants of the ATP binding cassette protein, ABCA4, with small molecule correctors as a treatment for Stargardt eye disease. J Biol Chem. 2015;290(32):19743-19755.

[29] Sormunen R, Eskelinen S, Lehto VP. Bile canaliculus formation in cultured HEPG2 cells. Lab Invest. 1993;68(6):652-662.

[30] Tougeron D, Fotsing G, Barbu V, Beauchant M. ABCB4/ MDR3 gene mutations and cholangiocarcinomas. J Hepatol. 2012;57(2):467-468.

[31] Cosconati S, Forli S, Perryman AL, Harris R, Goodsell DS, Olson AJ. Virtual Screening with AutoDock: theory and practice. Expert Opin Drug Discov. 2010;5(6):597-607.

[32] Gordo-Gilart R, Andueza S, Hierro L, Jara P, Alvarez L. Functional rescue of trafficking-impaired ABCB4 mutants by chemical chaperones. PLoS One. 2016;11(2):e0150098.

[33] Andress EJ, Nicolaou M, Romero MR, et al. Molecular mechanistic explanation for the spectrum of cholestatic disease caused by the S320F variant of ABCB4. Hepatology. 2014;59(5):1921-1931.

[34] Spanò V, Montalbano A, Carbone A, Scudieri P, Galietta LJV, Barraja P. An overview on chemical structures as ΔF508-CFTR correctors. Eur J Med Chem. 2019;180:430-448.

[35] Kinting S, Höppner S, Schindlbeck U, et al. Functional rescue of misfolding ABCA3 mutations by small molecular correctors. Hum Mol Genet. 2018;27(6):943-953.

[36] van der Woerd WL, Wichers CG, Vestergaard AL, et al. Rescue of defective ATP8B1 trafficking by CFTR correctors as a therapeutic strategy for familial intrahepatic cholestasis. J Hepatol. 2016;64(6):1339-1347.

[37] Carotti M, Marsolier J, Soardi M, et al. Repairing folding-defective α-sarcoglycan mutants by CFTR correctors, a potential therapy for limb-girdle muscular dystrophy 2D. Hum Mol Genet. 2018;27(6):969-984.

[38] Wolfhagen FH, Vermeulen HG, de Man RA, Lesterhuis W. Initially obscure hepatotoxicity attributed to sildenafil. Eur Journal Gastroenterol hepatol. 2008;20(7):710-712.

[39] Enomoto M, Sakaguchi H, Ominami M, et al. Sildenafil-induced severe cholestatic hepatotoxicity. Am J Gastroenterol. 2009;104(1):254-255.

[40] Yoshikado T, Takada T, Yamamoto T, et al. Itraconazole-induced cholestasis: involvement of the inhibition of bile canalicular phospholipid translocator MDR3/ABCB4. Mol Pharmacol. 2011;79(2):241-250.

[41] Shi Z, Tiwari AK, Shukla S, et al. Sildenafil reverses ABCB1-and ABCG2-mediated chemotherapeutic drug resistance. Cancer Res. 2011;71(8):3029-3041.

[42] Morita SY, Terada T. Molecular mechanisms for biliary phospholipid and drug efflux mediated by ABCB4 and bile salts. Biomed Res Int. 2014;2014:954781.

[43] Wang PH, Tu YS, Tseng YJ. PgpRules: a decision tree based prediction server for P-glycoprotein substrates and inhibitors. Bioinformatics. 2019;35(20):4193-4195.

[44] Ekins S. Drug transporter pharmacophores. Transporters as Drug Carriers. 2009;44:215-221.

[45] Molinski SV, Shahani VM, Subramanian AS, et al. Comprehensive mapping of cystic fibrosis mutations to CFTR protein identifies mutation clusters and molecular docking predicts corrector binding site. Proteins. 2018;86(8):833-843.

[46] Loo TW, Clarke DM. Mapping the binding site of the inhibitor Tariquidar that stabilizes the first transmembrane domain of P-glycoprotein. J Biol Chem. 2015;290(49):29389-29401.

[47] Degiorgio D, Colombo C, Seia M, et al. Molecular characterization and structural implications of 25 new ABCB4 mutations in progressive familial intrahepatic cholestasis type 3 (PFIC3). Eur J Hum Genet. 2007;15(12):1230-1238.

[48] Colombo C, Vajro P, Degiorgio D, et al. Clinical features and genotype-phenotype correlations in children with progressive familial intrahepatic cholestasis type 3 related to ABCB4 mutations. J Pediatr Gastroenterol Nutr. 2011;52(1):73-83.

[49] Nakken KE, Labori KJ, Rødningen OK, et al. ABCB4 sequence variations in young adults with cholesterol gallstone disease. Liver Int. 2009;29(5):743-747.

[50] Degiorgio D, Corsetto PA, Rizzo AM, et al. Two ABCB4 point mutations of strategic NBD-motifs do not prevent protein targeting to the plasma membrane but promote MDR3 dysfunction. Eur J Hum Genet. 2014;22(5):633-639.

[51] Kim Chiaw P, Wellhauser L, Huan LJ, Ramjeesingh M, Bear CE. A chemical corrector modifies the channel function of F508del-CFTR. Mol Pharmacol. 2010;78(3):411-418.

## Chapter VI. Phospholipid-porphyrin conjugates: deciphering the driving forces behind their supramolecular assemblies

Louis-Gabriel Bronstein[1#], Ágota Tóth[2#], Paul Cressey[1], Véronique Rosilio[1], Florent Di Meo[2*] and Ali Makky[1*]

[1] Université Paris-Saclay, CNRS, Institut Galien Paris-Saclay, 92296, Châtenay-Malabry, France
[2] INSERM U1248 Pharmacology & Transplantation, Univ. Limoges, 2 rue du Prof. Descottes, F-87025, Limoges, France

[#] Equally contributed
[*] Corresponding authors: florent.di-meo@inserm.fr, ali.makky@universite.paris-saclay.fr

**Published, DOI: 10.1039/d2nr01158a**

This study investigates different phospholipid-porphyrin conjugates and the effect *e.g.,* of the length of the linker on the self-assembly properties by means of experimental and molecular dynamics simulations. Thanks to the self-assembly capability, these conjugates show heigh interest in the usage for *e.g.*, photo-activatable drug delivery systems. Although, the subject of this study is a bit distant from my main topic but thanks to this research, I have learnt more about the lipid bilayer properties described in Section I.2.3. Similarly, as in Chapter V, the details of the experiments are not included in this thesis

### Abstract

Phospholipid–porphyrin conjugates (PL–Por) are nowadays considered as a unique class of building blocks that can self-assemble into supramolecular structures that possess multifunctional properties and enhanced optoelectronics characteristics compared to their disassembled counterparts. However, despite their versatile properties, little is known about the impact of the packing parameter of PL–Por conjugates on their assembling mechanism and their molecular organization inside these assemblies. To gain a better understanding on their assembling properties, we synthesized two new series of PL–Por conjugates with different alkyl sn2-chain lengths linked via an amide bond to either pheophorbide-a ($Ph_x$LPC) or pyropheophorbide-a ($Pyr_x$LPC). By combining a variety of experimental techniques with molecular dynamics (MD) simulations, we investigated both the assembling and optical properties of the PL–Por either self-assembled or when incorporated into lipid bilayers. We demonstrated that independently of the linker length, $Ph_x$LPC assembled into closed ovoid structures, whereas $Pyr_x$LPC formed rigid open sheets. Interestingly, $Pyr_x$LPC assemblies displayed a significant red shift and narrowing of the Q-band indicating the formation of ordered J-aggregates. The MD simulations highlighted the central role of the interaction between porphyrin cores rather than the length difference between the two phospholipid chains in controlling the structure of the lipid bilayer membranes and thus their optical properties. Indeed, while $Ph_x$LPC have the tendency to form inter-leaflet $\pi$-stacked dimers, $Pyr_x$LPC conjugates formed dimers within the same leaflet. Altogether, this work could be used as guidelines for the design of new PL–Por conjugates that self-assemble into bilayer-like supramolecular structures with tunable morphology and optical properties.

## VI.1. Introduction

Amphiphilic compounds such as surfactants and phospholipids are molecules that possess both hydrophilic and hydrophobic blocks. When dispersed in water at sufficiently high concentrations (*i.e.*, higher than their critical micelle concentration), they tend to self-assemble into a wide variety of supramolecular structures with a broad range of sizes and shapes such as micelles, vesicles, planar bilayers, nanotubes or nanofibers.[1–4] These organized assemblies offer several functional properties with respect to their disassembled counterparts In this context, amphiphilic compounds have attracted great attention for biomedical, pharmaceutical, and industrial applications. This is particularly true regarding the assembly of amphiphilic porphyrins[5] or porphyrinoids[6,7] into supramolecular structures designed for photoactivatable drug delivery devices[6–11] as well as light-harvesting nanosystems.[12,13] Amphiphilic porphyrin derivatives appear as very promising scaffolds since they can self-assemble into supramolecular structures[5] with adjustable photophysical properties and biomedical outcomes.[14] Among the amphiphilic porphyrins that have been designed so far for the development of photo-activatable drug delivery systems, the phospholipid–porphyrin (PL– Por) conjugates initially proposed by Gang Zheng's group[9,15,16] belong to the most versatile compounds for biomedical applications. These conjugates are made of pyropheophorbide-a or bacteriochlorophyll-a photosensitizers (PSs) linked to the sn2 hydroxyl group of 1-lysophosphatidylcholine (C16) via an esterification reaction. Such PL–Por conjugates self-assemble into liposome-like nanostructures named "porphysomes"[9] which exhibit a highly organized packing of the porphyrin molecules with improved optical and photophysical properties[9,14] when compared to unconjugated porphyrin derivatives. In addition, such packed porphyrin organization enables the use of these assemblies in multiple applications including photothermal therapy (PTT), photo-triggered drug release, photoacoustic imaging or fluorescence imaging[9] and photodynamic therapy (PDT) following their disassembly at the tumor site.[9,17] The shape and morphology of the self-assembled structures of amphiphiles are usually predicted using simple models of the dimensionless geometric packing parameter ($P = v/a \cdot l$) which is defined as the ratio between the volume of the hydrophobic part (v) of the amphiphiles and the optimal cross sectional surface area (a) of the hydrophilic headgroup times the length (l) of hydrophobic chains in their all trans conformation.[1,2] Amphiphiles with packing parameters lower than 1/3 or between 1/3 and 1/2 form spherical and worm-like micelles, respectively. On the other hand, while double-tailed phospholipids exhibiting packing parameters lying between 1/2 and 1 form bilayers, those with P values larger than 1 self-assemble into inverted hexagonal phases ($H_{II}$).[2,4]

Despite the versatile feature of PL–Por conjugates as building blocks for the design of multifunctional nanoplatforms,[18–21] the impact of the geometrical packing parameter on the formation and properties of their assemblies, as well as on their mixing with host phospholipids, remains unclear.[8,22–24] PL–Por building blocks initially consisting of a bulky porphyrin core directly attached to the hydroxyl group in sn2 position of 1-lysophosphatidylcholine (C16)[9] would exhibit a large area of the polar headgroup with concomitant mismatch between the length of the alkyl chain in sn1 position and the adjacent porphyrin. Thus, along the same line of thought of phospholipids assemblies, PL–Por conjugates are not expected to form bilayers, unless other driving forces or parameters are involved. Besides, the enlargement of the polar headgroup area caused by presence of the porphyrin core in its vicinity can be likely overcome by adding a linker between phosphatidylcholine headgroup and the porphyrin moiety. Indeed, changing the length of the alkyl chain in the sn2 position is expected to reduce the chain length difference between the two hydrophobic tails of the conjugate. It was expected that

chromophores grafted to a longer chain should align with the sn1 C16 alkyl chain improving the lateral packing properties. Indeed, Carter *et al.*[22] used molecular dynamics simulations to demonstrate that conjugating devinyl hexyloxyethy-pyropheophorbide (HPPH, a pyro-a derivative) which possesses a hexyl ether moiety at its extremity, enabled the formation of stable bilayers with superior packing properties when compared to the pyro-a conjugate. Such high packing properties of HPPH-conjugates were related to the presence of hexyl ether moiety which provided a better space filling between the two bilayer leaflets.[22] Beside the length difference between the two chains, changing the porphyrin structure should also play an important role in the intermolecular interaction between porphyrin cores which in turn might impact their stability and their photophysical properties.

In the present work, we describe the syntheses of two new series of PL–Por conjugates exhibiting different alkyl chain lengths in sn2 position and linked via an amide bond to either pheophorbide-a ($Ph_xLPC$) or pyropheophorbide-a ($Pyr_xLPC$) (Fig. 38). The self-assemblies of the synthesized PL–Por conjugates as well as their incorporation into liposomes were then characterized in terms of morphology, optical properties and thermodynamics using cryogenic transmission electron microscopy (Cryo-TEM), absorbance/fluorescence spectroscopy and differential scanning calorimetry (DSC). Experimental observations were supported by molecular dynamics simulations providing an atomic-scale and dynamic pictures of porphyrin conjugate self-assemblies. The present joint experimental and computational investigations aim to provide reliable insights for understanding the role of (i) the porphyrin structure, (ii) the length of the linker and (iii) the role of surrounding lipid environment on the properties of PL-Por assemblies.

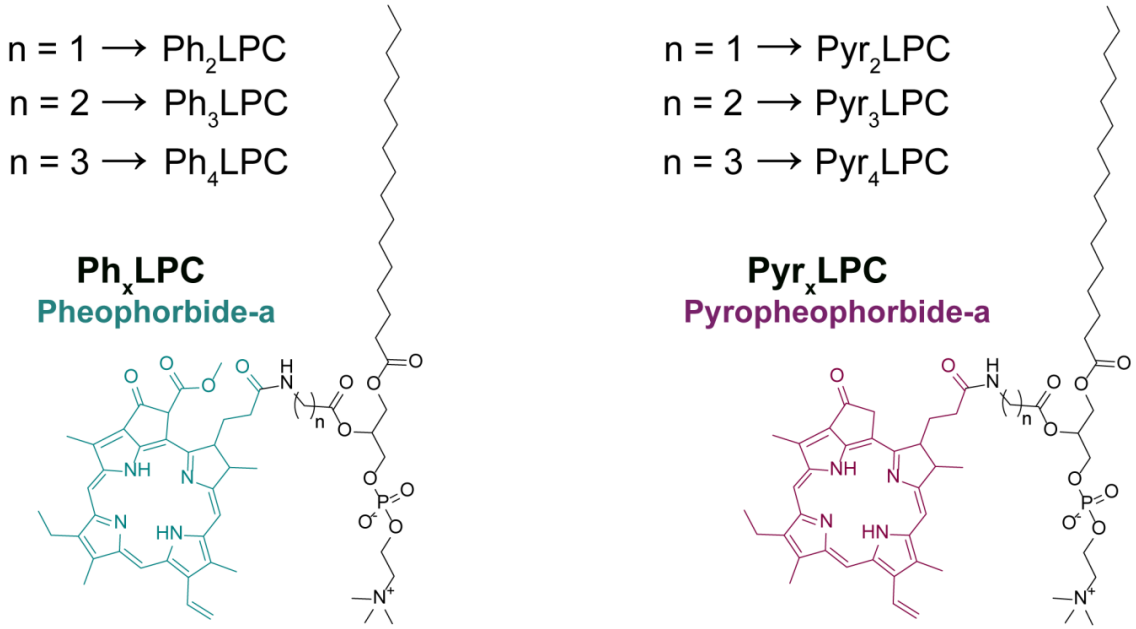



**Figure 38. Structure of phospholipid–porphyrin conjugates $Ph_xLPC$ and $Pyr_xLPC$ bearing pheophorbide-a and pyropheophorbide-a chromophore, respectively.**

## VI.2. Experimental section

*Details are reported in the sub-sections of the article.*

*To sum up, the experimental section starts with the description of the used chemicals and the synthesis of all phospholipids and the lipid–porphyrin conjugates. In the experimental part Cryogenic transmission electron microscopy (Cryo-TEM) was used. Absorption and fluorescence were measured and differential scanning calorimetry (DSC) experiments were performed.*

### Force fields and system builder

The Amber force field (FF) Lipid17 [27,28] was used to described DPPC and POPC lipid molecules. $Ph_x$LPC and $Pyr_x$LPC parameters were derived from GAFF2 [29] and Lipid17 FF and are available in ESI. † Atomic charges of porphyrin core and linker moieties were derived applying the AM1-bcc method using the antechamber package. Lipid bilayer membranes were solvated using the TIP3P water model [30] using a hydration level set up at *ca.* 40 water molecules per lipid. Given the size of porphyrin core and possible interdigitation of lipid tails in porphyrin rings, lipid bilayer membranes were all built by a homemade script (available upon request) and by taking advantage of the packmol-membrane package. Briefly, 3 + 3, 13 + 13 and 64 + 64 molecules of PL–Por respectively corresponding to 2.5 mol%, 10 mol% and pure PL–Por systems were put on two grids for upper and lower leaflets. Initial box and grid sizes were defined to prevent steric clash between porphyrins accounting DPPC and POPC area per lipids.[31] Initial area per lipid used to build pure PL–Por system was set up to 90 $Å^2$ to avoid steric clashes. Packmol was also used to (i) build DPPC or POPC lipids around $Pyr_x$LPC and $Ph_x$LPC molecules and (ii) solvate with water molecules. Na+ and Cl− ions were then randomly added in water to match NaCl = 0.154 M. It is worth mentioning that for pure PL–Por systems, both *trans* and *cis* configurations were considered regarding porphyrin A-ring.

### Molecular dynamics simulations

MD simulations were carried out using the CPU- and GPU-PMEMD (Particle Mesh Ewald Molecular Dynamics) codes available in Amber20 [32] (see https://ambermd.org/). Systems were first minimized. Then, they were thermalized in (N,V,T) ensemble up to 298 K for 250 ps steps in which lipid PC polar head groups were restrained at 2.5 kcal $mol^{-1}$ $Å^2$, using a 1 fs integration timestep. Temperature was maintained using the Langevin thermostat with a collision frequency set to 1 $ps^{-1}$. Pressure equilibration MD simulations were then performed in semi-isotropic (N, P, T) ensemble in 5 steps in which PC polar head restraints were smoothly switched off as follow: (i) 125 ps, 1 kcal $mol^{-1}$ $Å^2$, (ii) 500 ps, 0.5 kcal $mol^{-1}$ $Å^2$, (iii) 500 ps, 0.1 kcal $mol^{-1}$ $Å^2$, (iv) 500 ps, no restraints and (v) 1000 ps, no restraints, using a 2 fs integration timestep. Except for the last step, semi-isotropic Berendsen barostat was used to control pressure. Then, Montecarlo barostat was used for the last equilibration step as well as production runs of DPPC- and POPC-LPC mixtures. Pressure was maintained using Berendsen barostat regarding pure PL–Por bilayer membranes owing to the large deviation of box sizes occurring during equilibration from initially built systems. Production MD simulations were then carried at 298 K, 1 atm for 1 and 2 μs respectively for mixture and pure LPC lipid bilayers. It is worth mentioning that our present simulations are in line with the computational results obtained for similar systems accounting FF families as well as chemical differences (*e.g.,* porphyrin core structures, presence of metal, lipid composition).[22,33,34]

### Analysis

Analyses were carried out over the last 500 ns of each run, ensuring stable equilibrated sampling. Most of the analyses were performed using the cpptraj and pytraj packages35 (https://amberhub.chpc.utah.edu/) from the Amber20 suite. C-Atom lipid order parameters $S_{CD}$ were extrapolated from angles between C–H bonds and the bilayer normal.[36] Likewise, orientations of porphyrin were obtained using the angle $\theta$ between the lipid bilayer and the porphyrin core normal vectors. Order parameters were then obtained as follows:[37]

$$P_z = \frac{1}{2}\langle 3\cos^2\theta - 1\rangle$$

Given that $\theta$ correspond to porphyrin core normal, order parameter values range from 1 to -0.5 respectively indicating that porphyrin core is aligned either to membrane plane or to lipid tails.

$\pi$-Stacking events were also calculated using similar approach used for H-bond analysis in cpptraj package. Briefly, for each porphyrin pair, inter-porphyrin core distance and angle were calculated over MD simulations. It is worth mentioning that porphyrin core was hereby defined by the tetrapyrrole moiety only. Inter-porphyrin core angles were measured by calculating the angle between the vectors perpendicular to porphyrin cores. For each porphyrin pair, $\pi$-stacking event was counted as 1 if both the inter-porphyrin core distance and angle were below 8.0 Å and 10°, respectively. For each pair of porphyrins, fractions were then obtained by summing the total count of $\pi$-stacking events over the number of frames considered during the analysis. Bilayer thicknesses and thicknesses maps were obtained based on the density of PC head P-atoms. Thickness maps were calculated on MD trajectories with 2 Å resolution on the xy-plane, using VMD Membplugin tool.[38] Thicknesses were then calculated by averaging thickness map values over the three replicas. Assessment of the packing parameter requires the accurate calculation of molecular volumes for porphyrin cores and sn1 lipid tail. However, such calculations are still relatively challenging for non-globular structures. Therefore, we here propose to assess packing parameters by measuring the projected distances between the porphyrin core and palmitic acid center-of-mass onto the xy-plane. By comparing these distances with the known PC cross sectional diameter extrapolated from cross sectional area parameter, it is thus possible to roughly estimate the shape of PL–Por blocks. Molecular representations were rendered using the VMD-1.9.4-alpha package.[39]

## VI.3. Results and discussion

*Details are reported in the sub-sections of the article.*

*Shortly, six different phospholipid–porphyrin conjugates were synthesized. The differences were the alkyl chain lengths in sn2 position and the porphyrinoid derivatives. It is important to mention that pheophorbide-a has an extra methyl ester group in ortho position relative to the ketone functionality, therefor, racemic mixture was used. The supramolecular assembly was monitored through structural and optical properties.*

### Insights into the supramolecular assembly of PL–Por conjugates from MD simulations

MD simulations were shown to be relevant to investigate supramolecular properties of conventional lipid bilayers[48] as well as those made of PL–Por conjugates.[22] MD simulations were performed to get an atomic-scaled understanding about the structural influence of the different conjugates on the molecular organization, packing parameters of the conjugates, as

well as on the driving forces. Each system was made of 128 $Ph_x$LPC or $Pyr_x$LPC molecules (Fig. 40). Three replicas were run for 2 µs, each. MD simulations reveal structural differences between $Ph_x$LPC and $Pyr_x$LPC bilayers, as depicted by calculated thickness maps, averaged area per lipid (APL) and density profiles reported in Table 11 and Fig. S21–23. † For instance, the bilayers made of $Pyr_x$LPC are more ordered than $Ph_x$LPC ones, as shown by the calculated palmitate chain lipid order parameters (SCD, see Fig. 40B), regardless of the linker size. This is in line with calculated thicknesses which are slightly larger for $Pyr_x$LPC. Such result is also consistent with our recent monolayer study which showed that $Pyr_x$LPC conjugates form a film in the liquid-condensed state with subsequent appearance of well-structured domains at the air/water interface.[49]

Likewise, the areas per lipid obtained for $Ph_x$LPC systems are larger than those for the $Pyr_x$LPC bilayers, regardless of the linker length. Interestingly, MD simulations reveal that $Ph_x$LPC and $Pyr_x$LPC molecules self-assemble to maximize dispersive interactions between porphyrin cores. Therefore, $Ph_x$LPC and $Pyr_x$LPC molecules are involved in a strong network of π-stacking interactions which lead to highly ordered supramolecular assemblies in agreement with aforementioned experimental observations. We quantified π-stacking events by using geometric criteria, namely (i) the distance between porphyrin cores and (ii) the angle between the two normal vectors of tetrapyrrole cores. π-Stacking events were observed along simulations, for which the fractions of each porphyrin pair over time were calculated and summed in Fig. 40C. It is worth mentioning that π-stacking interactions are sufficiently strong to be sometimes maintained for more than 70% of simulations (see Table S1 †). Furthermore, MD simulations show two types of π-stacked dimers. On one hand, dimers are formed within the same leaflet in which dipole–dipole interactions are maximized by systematic structural shift between porphyrin cores along the stacking axis (Fig. 40D).

On the other hand, inter-leaflet π-stacked dimers were also observed for which dipole–dipole interactions were maximized by adopting an anti-parallel conformation. This provides robust hints to rationalize the overall structural differences between $Ph_x$LPC and $Pyr_x$LPC bilayers. Indeed, regardless of linker size, $Ph_x$LPC systems exhibit less π-stacking events with respect to $Pyr_x$LPC ones (see Fig. 40C). $Ph_x$LPC molecules simply differ from $Pyr_x$LPC by the presence of carboxymethyl moiety on porphyrin E-ring which is associated to a less planar porphyrin core. Furthermore, multiple stacking on both faces of porphyrin cores was not observed during simulations owing to the steric bulk arising from the axial orientation of carboxymethyl moiety in E-ring (see Fig. 40D). Altogether, $Ph_x$LPC self-assemble into less ordered lipid bilayers than $Pyr_x$LPC, as pictured by both experimental and computational results. This is in perfect agreement with the absorbance spectra where only pure $Pyr_x$LPC molecules form J-aggregates. It is worth mentioning that the present MD simulations did not show the experimentally observed mixture of highly ordered and disordered porphyrin aggregates within PL–Por bilayers. This is likely due to the limited size of modeled systems (128 lipids) and the relatively limited timescale (2 µs). This would require the modeling of much larger membranes by means of coarsegrained MD simulations.

MD simulations provided some hints about the role of linker size on the self-assembly of porphyrin conjugates. Even though such observations require further but challenging experimental validations, linker size seems to play a crucial role regarding π-stacking events. Indeed, both $Ph_3$LPC and $Pyr_3$LPC systems exhibit (i) less ordered bilayers (Fig. 40B) and (ii) less frequent π-stacking events (Fig. 40C) compared to PL–Por conjugates with either two or four carbon atoms in the linker. This highlights the balance between π-stacking driving forces

and linker-driven porphyrin flexibility. The shortest linker may lead to more ordered PL–Por bilayer owing to a lower flexibility of porphyrin core. Whereas the longest linker may be sufficiently flexible to favor proper stacking of porphyrin cores which in turn increase the structural order of the PL–Por bilayer. The intermediate linker ($Ph_3$LPC and $Pyr_3$LPC) adopts then an intermediate behavior in which ideal π-stacking is more difficult to obtain, because the linker is too short but sufficiently long to allow enough flexibility to the porphyrin.

Lateral porphyrin-palmitate distances were calculated as an approximation of packing parameters (Fig. 40E). Both porphyrin conjugates exhibit similar behavior. Taking into account these results and assuming that PC headgroup cross section diameter is 9.71 Å,50 $Pyr_2$LPC and $Ph_2$LPC should adopt slightly negative curvature, owing to the lower flexibility of the linker. Moreover, MD simulations suggest that porphyrin conjugates with the longer linker tend to adopt a planar curvature since the ratio between lateral porphyrin-palmitate distance and PC head cross section diameter are close to 1, which is particularly true for $Pyr_4$LPC and $Ph_4$LPC systems. However, the cryo-TEM results revealed that all PL–Por conjugates assemble into bilayers. This demonstrates that the calculation of the packing parameter is not sufficient to predict assembly of the PL–Por conjugates given their shapes in contrast to conventional phospholipids and detergents. Such findings highlight that the interactions between porphyrin cores plays a central role in controlling both the structure of the lipid bilayer membranes and their optical properties.

**Table 12. PL–Por bilayer thickness (in Å), averaged area per lipid (APL, in Å) and lateral porphyrin-palmitate distance (in Å).**

| | | Thickness (Å) | APL (Å$^2$) | Lateral porphyrin-palmitate distance (Å) |
|---|---|---|---|---|
| $Ph_x$LPC | x = 2 | 36.8 ± 1.8 | 92.8 ± 0.4 | 13.2 ± 3.0 |
| | x = 3 | 34.4 ± 1.9 | 99.8 ± 1.6 | 12.2 ± 3.6 |
| | x = 4 | 37.6 ± 2.0 | 94.5 ± 1.6 | 9.6 ± 2.7 |
| $Pyr_x$LPC | x = 2 | 38.8 ± 1.6 | 82.5 ± 0.7 | 12.7 ± 2.8 |
| | x = 3 | 38.0 ± 1.6 | 88.1 ± 1.4 | 11.8 ± 3.6 |
| | x = 4 | 39.9 ± 1.6 | 80.9 ± 0.4 | 8.9 ± 1.9 |

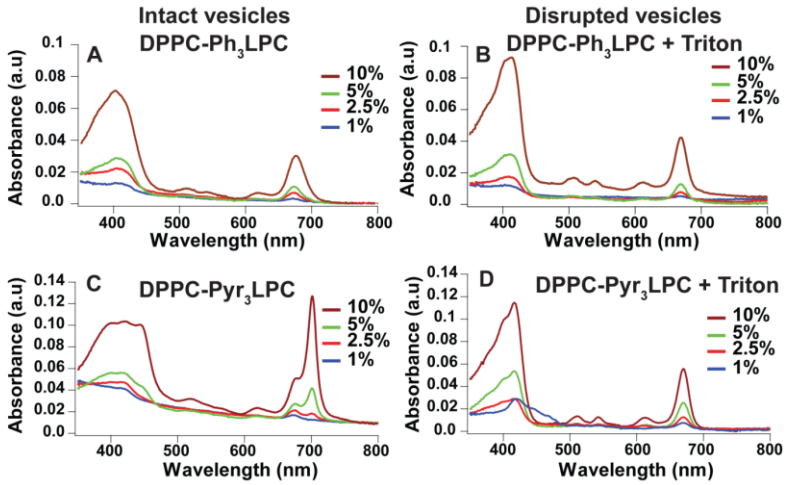



**Figure 39. Absorbance spectra of DPPC assemblies**
incorporating different molar percentage (1–10 mol%) of either Ph3LPC (A and B) or Pyr3LPC (C and D), before (A and C) and after (B and D) their disruption by adding Triton X-100 (1% v/v).

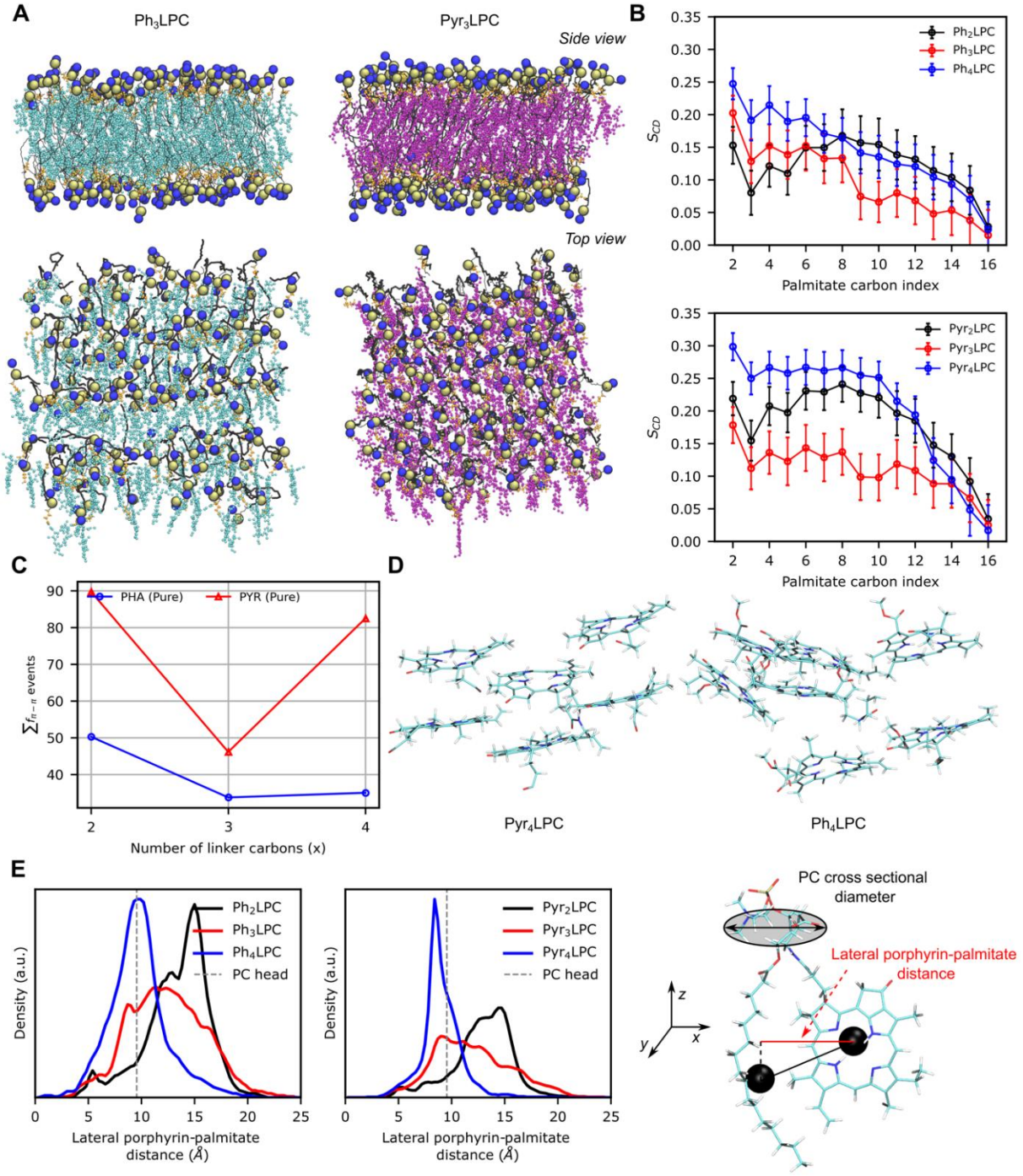



**Figure 40. Insights into the supramolecular assembly of PL–Por conjugates from MD simulations.**

**(A)** Representative snapshots of self-assembled $Ph_3LPC$ (left) and $Pyr_3LPC$ into bilayers from MD simulations. Both side (top) and top (bottom) view are displayed. PC N- and P-atoms are depicted in blue and brown spheres, respectively. Pheo-a and Pyro-a porphyrin cores as well as palmitate chains and linkers are colored in cyan, purple, black and orange respectively. **(B)** Calculated lipid order parameters (SCD) for palmitate carbons in $Ph_xLPC$ (top) and $Pyr_xLPC$ (bottom) simulations. **(C)** Overall sum of fractions of π-stacking events between porphyrin cores calculated along MD simulations and **(D)** representative snaphots (top view) of Pyr4LPC (left) and Ph4LPC (right) π-stacking organizations within the PL–Por bilayers. **(E)** Definition and distributions of lateral porphyrin-palmitate distance within PL–Por bilayers for $Ph_xLPC$ (left) and $Pyr_xLPC$ (bottom). Later, porphyrin-palmitate distance was used to approximate the shape of porphyrin conjugates and compare to PC cross sectional diameter extrapolated from cross sectional area parameter.

*Details are reported in the sub-sections of the article.*

*As only a few structures look like vesicles were found, lipid components incorporating phospholipid–porphyrin conjugates were also investigated. 2.5% and 10% of conjugates were embedded into POPC or DPPC membranes. The optical properties were examined and cryo-TEM analyses were performed.*

**Atomic-scaled structures of DPPC and POPC bilayers incorporating PL–Por conjugates**

To get a better understanding of the experimental observations, DPPC-$Ph_x$LPC, DPPC-$Pyr_x$LPC, POPC-$Ph_x$LPC and POPC-$Pyr_x$LPC systems were investigated by means of MD simulations considering either 2.5 mol% or 10 mol% of PL– Por. DPPC bilayers doped with PL–Por conjugates exhibit the typical structural behavior of DPPC gel phase given that simulations were performed at T = 25 °C (Fig. 41A). The calculated thicknesses and the APL (Table 12 and Fig. S24–S27 †) are in line with experiments and previous studies in which e.g., the APL of DPPC was shown to be *ca*. 48 $Å^2$ and the thickness ranged from 47 to 55 Å.[59,60] Lipid tail order $S_{CD}$ parameters were also computed supporting the gel and liquid phase of DPPC- and POPC-based lipid bilayer membranes (Fig. S29 and S30 †).

In all simulations, PL–Por conjugates partition in the lipid bilayer membrane, regardless of their concentration (see Table 12). This agrees with the aforementioned experiments except for the incorporation of $Ph_x$LPC compounds in DPPC lipid bilayer membranes at 2.5% mol. This is explained by the computational setup for which PL–Por conjugates were directly all incorporated in a DPPC lipid bilayer membrane. Thereby, joint computational and experimental investigations suggest that, at low concentration, $Ph_x$LPC is kinetically less likely to insert into the DPPC lipid bilayer membrane with respect to $Pyr_x$LPC, owing to the presence of polar carboxymethyl moiety. This chemical group may increase (i) the dipole moment of porphyrin core and/or (ii) the steric hindrance while penetrating in ordered lipid bilayer membrane. The latter is strongly supported by (i) POPC-experiments in which PL–Por conjugates partition in the lipid bilayer membrane regardless of the PL–Por type and (ii) the expected higher energy cost to insert in DPPC rather than in POPC, as suggested by recent computational investigations.[61,62]

Interestingly, MD simulations strongly support the importance of PL–Por structure and concentration. $Ph_x$LPC at 10 mol% concentration are more likely to modify DPPC bilayer structure than at 2.5 mol%. For instance, calculated thickness maps revealed local bilayer deformations for which thickness shrinking events are more pronounced at 10 mol% than 2.5 mol% (Fig. S24 †). In contrast, the presence of $Pyr_x$LPC molecules is expected to have a lesser effect on DPPC lipid bilayer, even locally. This might be explained by the interleaflet interactions between porphyrin cores of $Ph_x$LPC compounds. Experiments suggested that DPPC/$Pyr_x$LPC systems undergo phase separation events for concentration higher than 2.5 mol%. Unfortunately, this was not observed in MD simulations owing to the limited μs-timescale. Indeed, phase separation is driven by phenomena for which timescales range from dozens of microseconds to seconds[63] which require the use of coarse-grained MD simulations. However, present simulations provide robust hints about plausible underlying mechanisms which might ultimately lead to phase separation events.

Simulations performed with POPC bilayer show that lipid bilayer membranes are almost not impacted by the incorporation of PL–Por conjugates, regardless of the concentration (Table 12

and Fig. S26, 27 †). This supports the experimental findings where the fluidity of lipid bilayer membrane seemed to govern the dynamic behavior of the PL–Por conjugates, rather than the porphyrin core structure or the linker length. This is further confirmed by assessing the distribution function of porphyrin depths of insertion in the lipid bilayers (see Fig. S32 †).

While average distances of porphyrin core with respect to lipid bilayer membrane center-of-mass (COM) does not exhibit significant differences between the different lipid bilayer compositions, the distribution function reveals that porphyrin cores are more likely to move in fluid POPC bilayers than in DPPC-based bilayers. The more ordered the lipid bilayer membrane, the more localized the depth of penetration. Porphyrin orientations within a lipid bilayer were also investigated suggesting that the porphyrin moieties lay along the lipid tails as pictured by the calculation of order parameters $P_z$ (see Fig. S31 †).

**Table 13. Lipid bilayer thickness.**

Bilayer thickness based on distance between P-atom density peaks (in Å), averaged area per lipid (APL, in Å$^2$) and averaged distances of porphyrin core with respect to lipid bilayer membrane center of mass ($d_z$, in Å) for DPPC- and POPC-$Ph_x$LPC and -$Pyr_x$LPC systems.

| % mol | | | DPPC/PL-Por | | | POPC/PL-Por | | |
|---|---|---|---|---|---|---|---|---|
| | | | Thickness (Å) | APL (Å$^2$) | $d_z$ (Å) | Thickness (Å) | APL (Å$^2$) | $d_z$ (Å) |
| 2.5 % | $Ph_x$LPC | x = 2 | 46.7 ± 3.7 | 45.8 ± 1.2 | 10.1 ± 6.0 | 39.8 ± 2.0 | 62.2 ± 1.0 | 11.3 ± 2.7 |
| | | x = 3 | 46.7 ± 3.7 | 45.9 ± 1.6 | 7.3 ± 5.8 | 39.6 ± 1.9 | 62.4 ± 0.9 | 11.3 ± 2.8 |
| | | x = 4 | 47.3 ± 3.6 | 45.3 ± 0.7 | 10.6 ± 5.5 | 39.6 ± 1.9 | 62.7 ± 1.0 | 11.2 ± 2.7 |
| | $Pyr_x$LPC | x = 2 | 48.2 ± 3.4 | 44.3 ± 1.0 | 11.7 ± 6.0 | 39.9 ± 2.0 | 61.9 ± 1.0 | 11.0 ± 2.7 |
| | | x = 3 | 48.2 ± 3.3 | 44.3 ± 0.5 | 9.9 ± 6.0 | 39.8 ± 1.9 | 61.9 ± 1.0 | 10.6 ± 2.8 |
| | | x = 4 | 48.2 ± 3.4 | 44.1 ± 0.7 | 10.6 ± 6.1 | 39.8 ± 1.9 | 62.1 ± 0.9 | 10.8 ± 2.6 |
| 10 % | $Ph_x$LPC | x = 2 | 46.5 ± 3.8 | 49.2 ± 0.6 | 10.3 ± 6.2 | 40.0 ± 1.9 | 63.7 ± 0.9 | 10.8 ± 3.6 |
| | | x = 3 | 46.0 ± 3.9 | 49.9 ± 0.8 | 9.0 ± 5.0 | 39.8 ± 2.2 | 64.3 ± 0.9 | 9.5 ± 4.0 |
| | | x = 4 | 45.8 ± 4.1 | 50.2 ± 0.7 | 8.8 ± 4.9 | 39.7 ± 1.9 | 64.5 ± 1.0 | 10.3 ± 3.6 |
| | $Pyr_x$LPC | x = 2 | 47.2 ± 3.3 | 47.7 ± 0.6 | 10.5 ± 5.4 | 40.7 ± 1.9 | 61.9 ± 0.9 | 11.1 ± 2.8 |
| | | x = 3 | 46.6 ± 4.0 | 48.7 ± 1.1 | 9.9 ± 5.7 | 40.2 ± 2.2 | 62.6 ± 0.9 | 10.8 ± 2.9 |
| | | x = 4 | 47.0 ± 3.2 | 48.2 ± 0.6 | 10.6 ± 5.3 | 40.3 ± 1.9 | 63.1 ± 1.0 | 11.3 ± 3.2 |

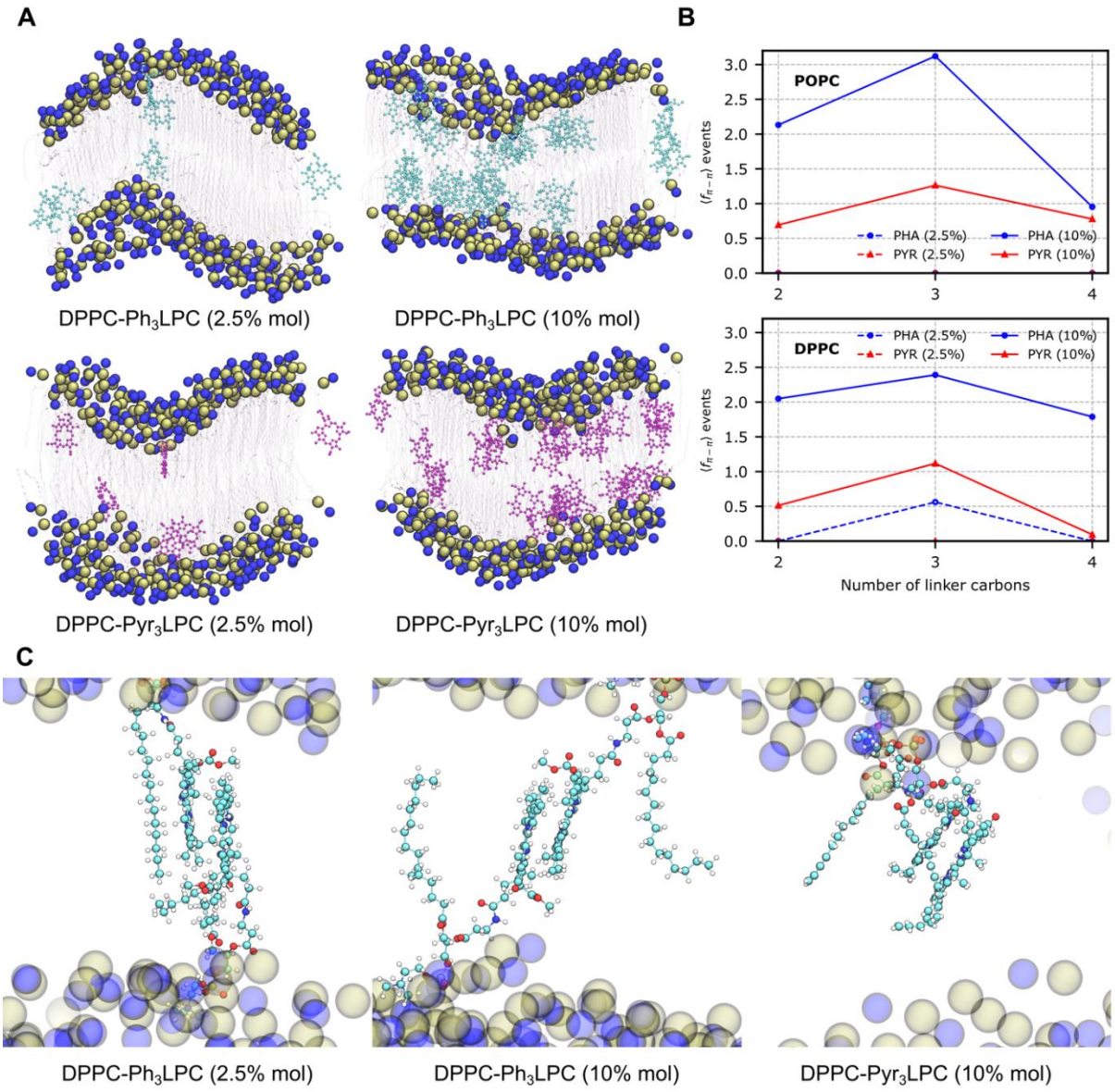



**Figure 41. Atomic-scaled structures of DPPC and POPC bilayers incorporating PL–Por conjugates.**

**(A)** Representative snapshots of DPPC-$Ph_3$LPC (top) and DPPC-$Pyr_3$LPC system from MD simulations. PC N- and P-atoms are depicted in blue and brown spheres, respectively. $Ph_3$LPC and $Pyr_3$LPC porphyrin cores as well as palmitate chains and linkers are colored in cyan, purple, orange and black. **(B)** Overall sum of π-stacking event fractions for all DPPC- and POPC-LPC systems. **(C)** Representative examples of DPPC-$Ph_3$LPC (left: 2.5 mol% and center: 10 mol%) and DPPC-$Pyr_3$LPC (right: 10 mol%) π-stacked conformations, highlighting inter-leaflet and intra-leaflet events respectively for $Ph_3$LPC and $Pyr_3$LPC.

## Rationalizing non-covalent interactions of PL–Por conjugates when incorporated in DPPC or POPC bilayers

To rationalize the formation of J-aggregates within DPPC lipid bilayer, inter-porphyrin π-stacking events were monitored along MD simulations of DPPC and POPC bilayers doped with PL–Por conjugates (see Fig. 41B and Table S1 †). In agreement with experimental observations made on J-aggregate formation, the higher the content of PL–Por incorporated in DPPC bilayers, the more likely the π-stacking between porphyrin cores. However, the

calculated π-stacking fractions remain low considering the number of porphyrin cores. This is consistent with the absorbance results since J-aggregate formation was shown to be only partial (Fig. 39 and 41A). However, the use of calculated π-stacking fractions alone is not sufficient to rationalize the experimentally observed difference between POPC and DPPC vesicles containing $Pyr_x$LPC at 10 mol%. Indeed, while the calculated π-stacking fractions remain in the same order of magnitude (e.g., 1.263 and 1.117 for POPC- and DPPC-$Pyr_3$LPC, respectively, see Table S1 †), optical experiments revealed the absence of J-aggregate Q-bands in the former. Interestingly, the number of porphyrin pairs involved in π-stacking events is systematically larger for POPC-based lipid bilayer membranes than for DPPC ones. In other words, a similar number of π-stacking events is observed in both POPC and DPPC bilayers incorporating 10 mol% of PL–Por but more porphyrins are involved in the former. This strongly suggests that π-stacked conformation lifetime is lower in POPC than in DPPC. J-aggregates are thus expected to be formed in POPC but for a significantly lower lifetime which may explain the absence of J-aggregate Q-band in the experiments. Both experiments and MD simulations suggest that more ordered DPPC bilayers are likely to decrease the diffusion of PL–Por with respect to POPC, which in turn increases π-stacked conformation lifetime.

Finally, particular attention was paid to the interplay between the conformation of π-stacked dimers and the structure of a porphyrin core. Interestingly, when incorporated in DPPC bilayer at 10 mol%, $Ph_x$LPC compounds favor the formation of inter-leaflet dimers (Fig. 41C), while $Pyr_x$LPC dimers are mostly observed within the same leaflet. For instance, up to 73% of observed π-stacked dimers involved both leaflet for $Ph_x$LPC (see Table S1 †). This could be related to the structural differences between the two porphyrin moieties. Indeed, the absence of bulky E-ring carboxymethyl moiety in $Pyr_x$LPC system favors the formation of π-stacked dimers, regardless of which porphyrin face interacts with another porphyrin (*i.e.*, *Si/Re*-like, see Fig. 41A). In contrast, the presence of carboxymethyl moiety in $Ph_x$LPC core (i) sterically decreases the likelihood of intra-leaflet π-stacked dimers and (ii) increases the dipole moment. Both events favor the formation of inter-leaflet antiparallel π-stacked dimers. The presence of inter-leaflet dimers leads to strong local deformation of DPPC lipid bilayer membrane in which *e.g.*, thickness is lower (as pictured in Table 12 and Fig. 41A). This is further supported by the membrane undulations and fluctuations observed in the cryo-TEM experiments. Finally, DPPC-$Pyr_x$LPC simulations can shed light on the underlying mechanism for the suggested phase separation observed in both DSC thermodynamics and cryo-TEM experiments. Indeed, MD simulations showed that $Pyr_x$LPC cores can π-stack regardless of *Si/Re* arrangement. Therefore, it is likely that more than two $Pyr_x$LPC partners can stack upon time. This event may be considered as a self-assembly nucleation event. It is worth mentioning that this event is expected to be controlled by the lateral diffusion of $Pyr_x$LPC for which the size of the linker is likely to play a role. This might then explain the difference observed between the DPPC-$Pyr_x$LPC vesicles.

### VI.4. Conclusions

PL–Por conjugates are versatile molecules that can self-assemble into supramolecular structures while exhibiting unique multifunctional properties. However, the driving forces behind their assembly remain unclear. Owing to their structural similarities with phospholipids, we expected that the shape and the morphology of the PL–Por conjugates self-assembled structures could be simply predicted from the calculation of their geometric packing

parameters. Hence, six new PL–Por conjugates possessing various geometrical packing parameters were synthesized. The new $Ph_x$LPC and $Pyr_x$LPC conjugates exhibit different linker lengths in sn2 position and bear either Pheo-a or Pyro-a at the linker extremity, respectively. It was thought that changing the linker length between the polar headgroup and the porphyrin core may modulate the chain length difference between the sn1 and sn2 chains and thus dictate the morphological structures of the assemblies. Both $Ph_x$LPC and $Pyr_x$LPC were able to self-assemble into supramolecular structures consistent with bilayers morphology and exhibiting different optical properties that were not dependent on the linker length. Indeed, while $Ph_x$LPC assembled into closed ovoid structures, $Pyr_x$LPC led to the formation of rigid open sheets. In addition, $Pyr_x$LPC assemblies displayed a significant red shift and narrowing of the Q-band which was related to the formation of ordered J-aggregates. The experimental data were strongly supported by MD simulations highlighting the central role of the interaction between porphyrin cores rather than the length difference between the two phospholipids chains in controlling the structure of the lipid bilayer membranes and thus their optical properties. Indeed, while $Ph_x$LPC have the tendency to form inter-leaflet π-stacked dimers, $Pyr_x$LPC conjugates formed dimers within the same leaflet. This explained the formation of hard open sheet observed for $Pyr_x$LPC and the undulated bilayers in the case of $Ph_x$LPC assemblies. In addition, the MD simulations shed light on the role of the linker in controlling the number of π-stacking events. For instance, the longer linker allowed the optimal π-stacking between the porphyrin moieties. Finally, both experimental and computational results demonstrated that these conjugates can be efficiently inserted in the lipid bilayer matrix with higher penetration depth for $Pyr_x$LPC compounds. Moreover, it appeared that the fluidity of the phospholipid bilayer is an important parameter to control whether J-aggregates of $Pyr_x$LPC can be formed or not. Altogether, this work could be used as guide for the design of new PL–Por conjugates that self-assemble into supramolecular structures with tunable morphology and optical properties by playing with (i) the planarity and dipole moment of porphyrin core, (ii) the linker length and (iii) the structure of host lipid bilayer membrane.

*Autor contribution, Conflict of interest, and Acknowledgements are reported in the article.*

### VI.5. Notes and references


1. J. N. Israelachvili, Intermolecular and Surface Forces, Academic Press, San Diego, 3rd edn, 2011.
2. J. N. Israelachvili, D. J. Mitchell and B. W. Ninham, Biochim. Biophys. Acta, 1977, 470, 185–201.
3. V. V. Kumar, Proc. Natl. Acad. Sci. U. S. A., 1991, 88, 444–448.
4. M. Ramanathan, L. K. Shrestha, T. Mori, Q. Ji, J. P. Hill and K. Ariga, Phys. Chem. Chem. Phys., 2013, 15, 10580–10611.
5. X. Li, S. Lee and J. Yoon, Chem. Soc. Rev., 2018, 47, 1174–1188.
6. I. Paramio, T. Torres and G. de la Torre, ChemMedChem, 2021, 16, 2441–2451.
7. J.-Q. Cai, X.-M. Liu, Z.-J. Gao, L.-L. Li and H. Wang, Mater. Today, 2021, 45, 77–92.

8. M. Shakiba, K. K. Ng, E. Huynh, H. Chan, D. M. Charron, J. Chen, N. Muhanna, F. S. Foster, B. C. Wilson and G. Zheng, Nanoscale, 2016, 8, 12618–12625.

9. J. F. Lovell, C. S. Jin, E. Huynh, H. Jin, C. Kim, J. L. Rubinstein, W. C. Chan, W. Cao, L. V. Wang and G. Zheng, Nat. Mater., 2011, 10, 324–332.

10. E. Huynh, B. Y. C. Leung, B. L. Helfield, M. Shakiba, J.-A. Gandier, C. S. Jin, E. R. Master, B. C. Wilson, D. E. Goertz and G. Zheng, Nat. Nanotechnol., 2015, 10, 325–332.

11. E. Huynh and G. Zheng, Wiley Interdiscip. Rev.: Nanomed. Nanobiotechnol., 2013, 5, 250–265.

12. S. Matsubara and H. Tamiaki, J. Photochem. Photobiol., C, 2020, 45, 100385.

13. J. Otsuki, J. Mater. Chem. A, 2018, 6, 6710–6753.

14. M. A. Rajora, J. W. H. Lou and G. Zheng, Chem. Soc. Rev., 2017, 46, 6433–6469.

15. E. Huynh and G. Zheng, Nano Today, 2014, 9, 212–222.

16. T. D. MacDonald and G. Zheng, Photonics Lasers Med., 2014, 3, 183–191.

17. C. S. Jin, J. F. Lovell, J. Chen and G. Zheng, ACS Nano, 2013, 7, 2541–2550.

18. L. Cui, Q. Lin, C. S. Jin, W. Jiang, H. Huang, L. Ding, N. Muhanna, J. C. Irish, F. Wang, J. Chen and G. Zheng, ACS Nano, 2015, 9, 4484–4495.

19. M. A. Rajora, L. Ding, M. Valic, W. Jiang, M. Overchuk, J. Chen and G. Zheng, Chem. Sci., 2017, 8, 5371–5384.

20. R. J. Paproski, A. Forbrich, E. Huynh, J. Chen, J. D. Lewis, G. Zheng and R. J. Zemp, Small, 2016, 12, 371–380.

21. E. Chang, J. Bu, L. Ding, J. W. H. Lou, M. S. Valic, M. H. Y. Cheng, V. Rosilio, J. Chen and G. Zheng, J. Nanobiotechnol., 2021, 19, 154.

22. K. A. Carter, S. Shao, M. I. Hoopes, D. Luo, B. Ahsan, V. M. Grigoryants, W. Song, H. Huang, G. Zhang, R. K. Pandey, J. Geng, B. A. Pfeifer, C. P. Scholes, J. Ortega, M. Karttunen and J. F. Lovell, Nat. Commun., 2014, 5, 3546.

23. D. M. Charron, G. Yousefalizadeh, H. H. Buzzá, M. A. Rajora, J. Chen, K. G. Stamplecoskie and G. Zheng, Langmuir, 2020, 36, 5385–5393.

24. J. Massiot, V. Rosilio, N. Ibrahim, A. Yamamoto, V. Nicolas, O. Konovalov, M. Tanaka and A. Makky, Chem. – Eur. J., 2018, 24, 19179–19194.

25. A. D. Bangham, M. M. Standish and J. C. Watkins, J. Mol. Biol., 1965, 13, 238–252.

26. M. M. da Cunha, S. Trepout, C. Messaoudi, T. D. Wu, R. Ortega, J. L. Guerquin-Kern and S. Marco, Micron, 2016, 84, 23–36.

27. B. D. Madej, I. R. Gould and R. C. Walker, J. Phys. Chem. B, 2015, 119, 12424–12435.

28. C. J. Dickson, B. D. Madej, Å. A. Skjevik, R. M. Betz, K. Teigen, I. R. Gould and R. C. Walker, J. Chem. Theory Comput., 2014, 10, 865–879.

29. J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, J. Comput. Chem., 2004, 25, 1157–1174.

30. D. J. Price and C. L. Brooks, J. Chem. Phys., 2004, 121, 10096–10103.

31. L. Martínez, R. Andrade, E. G. Birgin and J. M. Martínez, J. Comput. Chem., 2009, 30, 2157–2164.

32. R. Salomon-Ferrer, D. A. Case and R. C. Walker, Wiley Interdiscip. Rev.: Comput. Mol. Sci., 2013, 3, 198–210.

33. S. Shao, T. N. Do, A. Razi, U. Chitgupi, J. Geng, R. J. Alsop, B. G. Dzikovski, M. C. Rheinstädter, J. Ortega, M. Karttunen, J. A. Spernyak and J. F. Lovell, Small, 2017, 3, 1602505.

34. J. Federizon, C. G. T. Feugmo, W.-C. Huang, X. He, K. Miura, A. Razi, J. Ortega, M. Karttunen and J. F. Lovell, Pharmaceutics, 2021, 13, 98.

35. D. R. Roe and T. E. Cheatham, J. Chem. Theory Comput., 2013, 9, 3084–3095.

36. D. P. Tieleman, S. J. Marrink and H. J. C. Berendsen, Biochim. Biophys. Acta, Rev. Biomembr., 1997, 1331, 235–270.

37. C. J. Dickson, V. Hornak, R. A. Pearlstein and J. S. Duca, J. Am. Chem. Soc., 2017, 139, 442–452.

38. R. Guixà-González, I. Rodriguez-Espigares, J. M. Ramírez-Anguita, P. Carrió-Gaspar, H. Martinez-Seara, T. Giorgino and J. Selent, Bioinformatics, 2014, 30, 1478–1480.

39. W. Humphrey, A. Dalke and K. Schulten, J. Mol. Graphics, 1996, 14, 33–38.

40. R. Rosseto and J. Hajdu, Chem. Phys. Lipids, 2010, 163, 110–116.

41. R. Rosseto and J. Hajdu, Chem. Phys. Lipids, 2014, 183, 110–116.

42. R. Rosseto and J. Hajdu, Tetrahedron, 2014, 70, 3155–3165.

43. E. E. Jelley, Nature, 1936, 138, 1009–1010.

44. F. Würthner, T. E. Kaiser and C. R. Saha-Möller, Angew. Chem., Int. Ed., 2011, 50, 3376–3410.

45. E. E. Jelley, Nature, 1937, 139, 631–631.

46. G. Scheibe, L. Kandler and H. Ecker, Naturwissenschaften, 1937, 25, 75–75.

47. J. Massiot, V. Rosilio and A. Makky, J. Mater. Chem. B, 2019, 7, 1805–1823.

48. S. J. Marrink, V. Corradi, P. C. T. Souza, H. I. Ingólfsson, D. P. Tieleman and M. S. P. Sansom, Chem. Rev., 2019, 119, 6184–6226.

49. L.-G. Bronstein, P. Cressey, W. Abuillan, O. Konovalov, M. Jankowski, V. Rosilio and A. Makky, J. Colloid Interface Sci., 2022, 611, 441–450.

50. A. Wnętrzak, K. Łątka and P. Dynarowicz-Łątka, J. Membr. Biol., 2013, 246, 453–466.

51. J. F. Nagle, Annu. Rev. Phys. Chem., 1980, 31, 157–196.

52. D. Marsh, Chem. Phys. Lipids, 1991, 57, 109–120.

53. K. A. Riske, R. P. Barroso, C. C. Vequi-Suplicy, R. Germano, V. B. Henriques and M. T. Lamy, Biochim. Biophys. Acta, Biomembr., 2009, 1788, 954–963.

54. L. M. Ickenstein, M. C. Sandström, L. D. Mayer and K. Edwards, Biochim. Biophys. Acta, Biomembr., 2006, 1758, 171–180.

55. M. J. Bowick and R. Sknepnek, Soft Matter, 2013, 9, 8088–8095.

56. G. Vernizzi, R. Sknepnek and M. Olvera de la Cruz, Proc. Natl. Acad. Sci. U. S. A., 2011, 108, 4292–4296.

57. R. Sknepnek, G. Vernizzi and M. Olvera de la Cruz, Soft Matter, 2012, 8, 636–644.

58. C.-Y. Leung, L. C. Palmer, B. F. Qiao, S. Kewalramani, R. Sknepnek, C. J. Newcomb, M. A. Greenfield, G. Vernizzi, S. I. Stupp, M. J. Bedzyk and M. Olvera de la Cruz, ACS Nano, 2012, 6, 10901–10909.

59. J. F. Nagle and S. Tristram-Nagle, Biochim. Biophys. Acta, Rev. Biomembr., 2000, 1469, 159–195.

60. Z. V. Leonenko, E. Finot, H. Ma, T. E. S. Dahms and D. T. Cramb, Biophys. J., 2004, 86, 3783–3793.

61. C. Hoffmann, A. Centi, R. Menichetti and T. Bereau, Sci. Data, 2020, 7, 51.

62. A. Ghysels, A. Krämer, R. M. Venable, W. E. Teague, E. Lyman, K. Gawrisch and R. W. Pastor, Nat. Commun., 2019, 10, 5616.

63. E. L. Elson, E. Fried, J. E. Dolbow and G. M. Genin, Annu. Rev. Biophys., 2010, 39, 207–226.

## Conclusion

The present work aimed to provide a comprehensive structural overview of several systems of importance in the broad field of pharmacology. The present thesis has clearly demonstrated that computational methods, mostly MD simulations, are powerful tools to rationalize (i) structural dynamics of ABCC drug membrane transporters, (ii) plausible binding modes of small molecules with ABCB4 transporters and (iii) self-assembly of original phospholipid-porphyrin conjugates into lipid bilayers.

In chapter IV, structural dynamics of ABCC1/MRP1 (as an ABCC family prototype) was investigated by means of μs-long MD simulations. These investigations point out the importance of the resolution method. Our simulations consistently showed the spontaneous closing of the NBDs in case of the IF conformations. This suggests that the wide-open conformation may not exist physiologically. Likewise, simulations performed on OF structures are in agreement with previous observations suggesting that the EC open conformation is only a transition state, and not a stable state. Regarding the catalytic cycle, our results support the pivotal role of the degenerate NBS1 based on both the dynamics and measured important distances between the binding site motifs and nucleotides. ATP molecule remains stable in NBS1 even in the post-catalytic state. In contrast, significant structural deviations were observed for the canonical NBS2 in which ADP logically exhibits lower binding than ATP, respectively in post- and pre-hydrolysis state. Such deviation is mandatory for ADP release prior to ATP substitution. However, further investigations are required to understand if the ADP release may trigger the transition to IF conformation. Indeed, in a similar idea, the phosphate release has been shown to be of utmost importance for large-scale conformational transition in another ABC transporter [1]-[2]. We might hypothesize that ADP release may also play a role in the overall conformation of MRP1. It is important to note that (i) the current timescale of unbiased MD simulations is far beyond the reach of large-scale transition of ABC transporters and (ii) an ATP-ADP bound state IF narrow conformation has been recently resolved for a bacterial ABC transporter [3]. In other words, further investigations are required to decipher if OF-to-IF transitions may be shared or if ABC subfamilies exhibit significantly different transport cycle dynamics. In line with former experiments, MD simulations have also suggested the functional role of $L_0$ for which particular attention should be paid in ABCC family. It may have different roles depending on the transporter of interest, as well as the presence or absence of $TMD_0$. For instance, it has been suggested that $L_0$ possibly locates in the binding pocket in other ABC transporters (*e.g.*, human ABCB11 [4], C. elegans P-gp (PDBID: 4F4C) [5]). However, such location is very unlikely in ABCC transporter with $TMD_0$ such as MRP1 and MRP2.

Finally, the interplay between surrounding lipids and MRP1 was investigated in the present work. This is in line with the recent literature in which membrane composition and its importance has gained growing interest. Our results showed that the membrane composition seems to have lesser impact on the local conformation of MRPs, in contrast to what have been suggested in SLC transporters. We can hypothesize that it may be related to ABC location in cell membranes. Indeed, the number of human drug ABC transporters is way smaller than SLC (*ca.* 80 *vs* 400, respectively). Therefore, ABC transporter location in tissue is broader in cell. Interestingly, we can suggest that the lower dependence on membrane lipid structure may be correlated with the presence of the same ABC transporter at different poles of cell membranes, *e.g.*, MRP4 has been observed in apical and basolateral membrane of hepatocyte and kidney proximal tubular cells. Another example is the ubiquitous MRP1 cell distribution over human

tissues. However, it does not mean that the impact and role of lipids are negligible. Our MD simulations has shown that lipids play an active role in transport function, such as in the allosteric communication between the different ABC transporter domains, namely substrate-binding pocket and NBDs. This may be achieved thanks to specific cholesterol and PE-lipid binding spots, the latter being observed more specifically in the inner leaflet, in agreement with the known location of such lipids in cell membranes.

Even tough further joint experimental and computational investigations are required, the present work pave the way for better understanding on ABCC transporter function and, more importantly in the difference between ABC transporters.

Chapter V described the plausible non-covalent binding of CFTR correctors to ABCB4 transporter. Even though it is not a drug membrane transporter, molecular findings are of interest. We focused on the plausible binding modes of the so-called CFTR correctors to provide hints on the experimentally overserved rescued function of ER-retained ABCB4 variants. In spite of a careful investigation, our molecular docking approach was not accurate enough to ensure the rescue mechanism. However, we observed that such CFTR correctors may bind to ABCB4 NBS domains. Interestingly, these compounds were also shown to possibly rescue ABCC7/CFTR channel trafficking. ABCB4 and CFTR strongly differ in the ABC kingdom. However, they mostly share patterns regarding NBDs. This is why we can suggest the central role of NBD in term of binding. This may also explain why, if direct binding is confirmed, the function is challenging to rescue. Indeed, we may expect a binding to NBD, which may preclude the required ATP-binding for ABCB4 function. This pave the way for further chemico-physical investigations for which competition between ATP and correctors should be considered. Such assumptions also suggest that such therapeutic strategy may lead to only a partial recovered function.

Finally, Chapter VI reported a study which is out of the field of ABC transporters, but of interest in pharmacology. The self-assembly of lipid-porphyrin conjugates was investigated from both experimental and computational points of view. MD simulations were used in order to better understand the structure and biophysics of such original structures. The central role of non-covalent interactions, more specifically $\pi$–$\pi$ dispersive interactions, was highlighted. Interestingly, self-assembly structure seems to be correlated first by the linker size rather than favourable $\pi$-stacking arrangements. This work paves the way for further chemical optimization in which lipid-porphyrin conjugates may be used for drug delivery devices. The main advantage of porphyrin conjugates is the ability to specially release encapsulated drugs to specific tissues upon light.

As the works presented in the current thesis, as well as the studies done on SLC transporters [6], [7], in which I had also a contribution, support the relevance of computational models in the field of pharmacology. Moreover, thanks to the continuous development of computational devices and frame works, especially the growing use of machine learning, I strongly believe that computational methods will keep on complementing experimental techniques. One striking example is the release of Alpha Fold 2 [8] that makes possible to create sufficiently robust models for many proteins. Last, but not least, I believe that the present thesis provides useful hints about the dynamics of ABC proteins and the importance of lipids in biologically relevant systems.

## References


[1] H. Göddeke *et al.*, 'Atomistic Mechanism of Large-Scale Conformational Transition in a Heterodimeric ABC Exporter', *J Am Chem Soc*, vol. 140, no. 13, pp. 4543–4551, Apr. 2018, doi: 10.1021/jacs.7b12944.

[2] M. Prieß, H. Göddeke, G. Groenhof, and L. V. Schäfer, 'Molecular Mechanism of ATP Hydrolysis in an ABC Transporter', *ACS Cent Sci*, vol. 4, no. 10, pp. 1334–1343, Oct. 2018, doi: 10.1021/acscentsci.8b00369.

[3] Hofmann, S. *et al.* Conformation space of a heterodimeric ABC exporter under turnover conditions. *Nature* **571**, 580–583 (2019).

[4] L. Wang *et al.*, 'Structures of human bile acid exporter ABCB11 reveal a transport mechanism facilitated by two tandem substrate-binding pockets', *Cell Res*, vol. 32, no. 5, pp. 501–504, May 2022, doi: 10.1038/s41422-021-00611-9.

[5] M. S. Jin, M. L. Oldham, Q. Zhang, and J. Chen, 'Crystal structure of the multidrug transporter P-glycoprotein from C. elegans', *Nature*, vol. 490, 2013, doi: 10.1038/nature11448.

[6] A. Janaszkiewicz *et al.*, 'Substrate binding and lipid-mediated allostery in the human organic anion transporter 1 at the atomic-scale'. bioRxiv, p. 2022.07.14.500056, Jul. 15, 2022. doi: 10.1101/2022.07.14.500056.

[7] A. Janaszkiewicz *et al.*, 'Insights into the structure and function of the human organic anion transporter 1 in lipid bilayer membranes', *Sci Rep*, vol. 12, no. 1, p. 7057, Dec. 2022, doi: 10.1038/s41598-022-10755-2.

[8] M. Akdel *et al.*, 'A structural biology community assessment of AlphaFold 2 applications', Biophysics, preprint, Sep. 2021. doi: 10.1101/2021.09.26.461876.

## Appendices

## Appendix 1. List of publications

1) **Structural and dynamical differences shown by the pre- and post-hydrolysis states of ABCC1 (MRP1)**

   In preparation

   **Ágota Tóth**, Angelika Janaszkiewicz, Veronica Crespi and Florent Di Meo

2) **Membrane dependency on the structural dynamics of human major facilitator superfamily transporters**

   In preparation

   Angelika Janaszkiewicz, **Ágota Tóth**, Veronica Crespi and Florent Di Meo

3) **Substrate binding and lipid-mediated allostery in the human organic anion transporter 1 at the atomic-scale**

   Submitted, bioRχiv, DOI: 10.1101/2022.07.14.500056

   Angelika Janaszkiewicz, **Ágota Tóth**, Quentin Faucher, Hélène Arnion, Nicolas Védrenne, Chantal Barin-Le Guellec, Pierre Marquet, Florent Di Meo

4) **On the interplay between lipids and asymmetric dynamics of an NBS degenerate ABC transporter**

   Submitted, bioRχiv, DOI: 10.1101/2022.05.16.492073v1

   **Ágota Tóth**, Angelika Janaszkiewicz, Veronica Crespi and Florent Di Meo

5) **Phospholipid-porphyrin conjugates: deciphering the driving forces behind their supramolecular assemblies**

   May 2022, Nanoscale, DOI: 10.1039/d2nr01158a

   Louis-Gabriel Bronstein, **Ágota Tóth**, Paul Cressey, Véronique Rosilio, Florent Di Meo, Ali Makky

6) **Insights into the structure and function of the human organic anion transporter 1 in lipid bilayer membranes**

   April 2022, Scientific Reports, DOI: s41598-022-10755-2

   Angelika Janaszkiewicz, **Ágota Tóth**, Quentin Faucher, Marving Martin, Benjamin Chantemargue, Chantal Barin-Le Guellec, Pierre Marquet, Florent Di Meo

7) **Effect of CFTR correctors on the traffic and the function of intracellularly retained ABCB4 variants**

March 2021, Liver international, DOI: 10.1111/liv.14839

Amel Ben Saad, Virginie Vauthier, **Ágota Tóth**, Angelika Janaszkiewicz, Anne-Marie Durand-Schneider, Alix Bruneau, Jean-Louis Delaunay, Martine Lapalus, Elodie Mareux, Isabelle Garcin, Emmanuel Gonzales, Chantal Housset, Tounsia Aït-Slimane, Emmanuel Jacquemin, Florent Di Meo, Thomas Falguières

8) **The transport pathway in the ABCG2 protein and its regulation revealed by molecular dynamics simulations**

September 2020, Cellular and Molecular Life Sciences, DOI: 10.1007/s00018-020-03651-3

Tamás Nagy, **Ágota Tóth**, Ágnes Telbisz, Balázs Sarkadi, Hedvig Tordai, Attila Tordai and Tamás Hegedűs

9) **Analyzing Biomolecular Ensembles**

Augustus 2019, Biomolecular Simulations, DOI: 10.1007/978-1-4939-9608-7_18

Matteo Lambrughi, Matteo Tiberti, Maria Francesc Allega, Valentina Sora, Mads Nygaard, **Ágota Tóth**, Juan Salamanca Viloria, Emmanuelle Bignon, and Elena Papaleo

**Étude biophysique et structurale des transporteurs ABC par la pharmacologie computationnelle**

Le passage membrane des xénobiotiques est un événement important dans les processus pharmacologiques, en particulier dans l'absorption, la distribution, le métabolisme et l'élimination. Les transporteurs membranaires des cellules hépatiques et rénales sont des acteurs clés. Parmi elles, les exporteurs de type ABC (ATP-binding cassette) qui appartiennent à la famille ABCC, sont cruciales en pharmacologie car ils transportent divers substrats à travers les membranes. Cependant, les connaissances sur leurs dynamiques et les mécanismes de translocation sont encore limitées. Cette thèse, basée sur des simulations de dynamique moléculaire, fournit un aperçu dynamique et structural en se focalisant sur un prototype MRP1, pour laquelle des structures cryo-EM sont disponibles.
L'asymétrie des NBD de ce transporteur, mais également d'autres membres des ABCC est probablement associé à une moindre transduction du signal du NBD1 vers le reste de la protéine, en raison de l'absence de conformation en boule entre le NBD1 et les hélices de couplage. En parallèle, le rôle central de NBS1 est souligné par nos simulations, lesquelles suggèrent un rôle pivot de l'ATP. Une attention particulière est accordée au rôle de l'environnement lipidique, ainsi qu'au rôle du site dégénéré de liaison aux nucléotides, typique des MRPs. Les lipides jouent un rôle actif dans la communication allostérique entre la poche de liaison au substrat et les NBD. Cependant, si les molécules de cholestérol et les lipides PE semblent se fixer sur certains sites particuliers, l'impact de la composition de la membrane sur les minima locaux semblent être limité. Elle semble surtout affecter la cinétique de transport.
Nous pensons que notre travail peut être étendu à d'autres protéines ABC NBS dégénérées et fournir des indices pour déchiffrer les différences mécanistiques entre les transporteurs ABC.

**Mots-clés :** dynamique moléculaire, transporteurs de liaison à l'ATP, protéines associées à la multirésistance aux médicaments

**Biophysical and Structural Investigation of ABC Transporters by Computational Pharmacology**

Membrane crossing by xenobiotics is an important event in pharmacological processes, especially in Absorption, Distribution, Metabolism, and Elimination. Transporters located in liver and kidney cells are key players of metabolism and elimination. Multidrug resistance-associated proteins, belonging to ABC C-family exporters, are crucial in pharmacology as they transport various substrates across membranes. However, knowledge about their transport mechanisms is still limited. This thesis, based on molecular dynamics simulations, provides a dynamic and structural overview of MRP1, for which cryo-EM structures are available along it transport cycle.
Asymmetric NBD behaviour is ensured by lower signal transduction from NBD1 to the rest of the protein owing to the absence of ball-and-socket conformation between NBD1 and coupling helices. The pivotal role of NBS1 is confirmed by our post-catalytic simulations. Moreover, our results provide hints about the important role of $L_0$. Particular attention is paid to the role of the surrounding lipid environment, as well as the role of the degenerate nucleotide-binding site, typical for MRPs. Lipids play an active role in the allosteric communication between the substrate-binding pocket and NBDs. However, cholesterol molecules and PE-lipids highlight some hotspots, the impact of the membrane composition is limited. It seems to mostly affect transport kinetics.
We believe that our work can be extended to other degenerate NBS ABC proteins and provide hints for deciphering mechanistic differences among ABC transporters.

**Keywords:** molecular dynamics, ATP-binding cassette transporters, multidrug resistance-associated proteins