# Development of Novel Chloride-Selective Membrane Carriers

Martin Ožbej Flerin Maver

Wadham College


University of Oxford

A thesis presented for the degree of

*Doctor of Philosophy*

Michaelmas 2025

Friday 10th October, 2025

# Acknowledgments

"It takes a village to raise a child," goes the old saying, and while the process of completing a DPhil can feel insulating at times, the saying certainly feels apt here too. For without all of those named here, this piece of work would not have come to be what it is.

Firstly, I would like to extend my biggest gratitude to my supervisors, Profs. Matthew Langton and Fernanda Duarte. From the moment they gave me the opportunity to work on this project, they have given me their unwavering support. The amount that I learned from them is immeasurable, and I will be forever thankful. To Matt for his ability to steer any meeting that began with a desperate lack of ideas into a fruitful discussion that left me feeling optimistic. To Fernanda for pushing me to always get to the bottom of things, without leaving any stone unturned and allowing my curiosity to flourish. To both of them for allowing me to find my passion in chemistry - I arrived into the DPhil as a synthetic chemist and am leaving full to the brim of computational ideas.

I would also like to thank everyone who has made my moments in Oxford some of the best of my life. Aleksy, who took me on a tour on the first day of our SBM course, and has been my closest companion through this journey since - both in the PTCL and outside it. Without him, the more challenging times would have been unbearable and the nicer times not nearly as fun. I would like to thank the St. John's contingent for making me an honorary part of their college over these four years. To Bruno, I owe eternal gratitude for being there as a partner in crime in everything, from chemistry, to music and wine, and for introducing me to so many wonderful people in Oxford - Eddie, Gemma, Joe, and others. Our card-playing nights with Ana V. reminded me of the exhilaration of debating with passion, no

matter the subject matter, and made me feel closer to home. To May, who was the second friend I made in Oxford at the safety induction, where we immediately recognised we had a common passion for electronic music, I thank for being my confidant, for all of the walks and for all of the laughs. All of the evenings spent with Joe, in which he became one of my closest friends, and where he allowed me to learn from him about everything - chemistry, coolness, and culture and most importantly how to be a good person - will stay with me forever. I am ever so grateful to him for them, as well as for introducing me to the wonderful Liam, Aisling, Emma and Lewis. While no longer being strictly St. John's during our DPhils, Ana Š. must be named with this contingent, and I must thank her for being such a caring and supportive friend, from the tarok table to the paths of uni parks.

I must also thank everyone in the Duarte group and the PTCL, who made the beautiful mid-century palace a wonderful place to work. From Tom and Alistair, who shared their brilliant brains with me when I first joined and lit a spark for computational chemistry in me, to Matthew and Tristan, who became my friends during that first rotation and made it so enjoyable. Endless thanks go to Tomasz, the ceaselessly calm and friendly mentor, who taught me the wizardry of molecular dynamics; to Veronika, the treasure chest of computational knowledge, who did the same for machine learning potentials; and to Ally, who helped me follow in her wake, modelling ion transport. Wojtek, for being an encyclopedia of everything ML and biological, as well as being a kind and supportive friend; Sara, my dear komšija, for showing me how to present technical knowledge in an approachable and engaging way, as well as for being such a fountain of laughs; James, for always sharing his impeccable knowledge of theory and his friendly and hilariously ironic demeanour alike. Shoubhik, Bastian and Hanwen, for all being on hand to answer any of my (silly) theoretical questions, and Joe H., for being my buddy in both music and nonsensical jokes.

To my friends from the CRL trenches - I thank you likewise for being there in all of the failed synthetic attempts and vesicle assays that may or may not have stretched into the wee hours. Aidan, with whom we would commandeer the aux cable, for his impeccable taste, and for all the support he gave while also being the best synthetic chemist I have ever worked with. Kelly, for being my fumehood and rotovap neighboor and never getting sick of me. To Xiangyu, for all the tears we didn't shed listening to sad indie music, and contemplating life. To Charlie, for being such a ball of energy and helping me keep going through her kindness and enthusiasm. To Elin, for letting me speak to her in my horrible German from time to time, and helping me out whenever I messed something up. To Gurshinder, Chiara and Amy, for being so kind and accommodating whenever I showed up to the lab unannounced in my latter computational years.

I would also like to thank all of my friends from back home. Those who continued a parallel scientific academic path and have always shared in shouldering all of its struggles with me - thank you Luka, Klemen, Liza and Jure. To Mitja, for staying my best buddy in my years abroad, being the first to visit wherever I went. All of the chemists, physicists and engineers from Imperial. Dario, for always being there for me, coming to pick me up in my first year when times were hard, for keeping my heart in London, for all the nights dancing, and for helping me be a better person. To Gunin and Sach, for putting up with me for the four years at Imperial and forever more, and being there for me in everything. To Pushpin, for getting onboard and putting up with me in the post-Imperial years. To Genna and Leon, the physics geniuses who showed me how much more there is beyond science. And finally to Ina, for being there for me through the whole four years of the DPhil and never letting me give up.

The biggest thanks, as for everything in my life, goes to my family. First and foremost to my mother, the single kindest and smartest person in the world, who

taught me how to dream and how to do. From the first time she sat down with me to balance a chemical reaction ($6H_2O + 6CO_2 \rightarrow C_6H_{12}O_6 + 6O_2$), she helped me recognise I had a knack for chemistry. Most importantly, she, my grandmother, and my grandfather all always taught me to nurture knowledge and kindness within me. I would like to thank all of them for always being on the other end of the line when I missed them, and when I needed a chat. And to Mártin, Jure, Nina, Miha, Vesna, Janez, Jošt and Jerca for being there to welcome me back every time I came home. To my nono, the scientist of the family, and to omica. To my father and his partner Karin for their equally unwavering support and my wonderful siblings for teaching me over and over how curiously wonderful the world is; Leonard for showing me what drive means, Miriam for showing me how to be truly kind, and Julia for being the most naturally friendly, good-natured and funny person I know.

To all of you, and anyone else I have forgotten, I will forever be thankful to you for going through life with me. This thesis would not be possible without you.

# Declaration

I, Martin Ožbej Flerin Maver, declare that the Thesis I am submitting is entirely my own work except where otherwise indicated. This Thesis has not been submitted in whole or in part for any other academic degree or professional qualification.

10/10/2025

*Signature* *Date*

# Software availability

Additions made to the *mlp-train* software package that are discussed in this thesis can be found on the public GitHub repository of the Duarte group:

https://github.com/duartegroup/mlp-train/

# Abstract

Synthetic ion transporters attempt to achieve a monumental task - carry out the role of massive, kDa- to MDa-sized protein transporters. They hold promise as both chemical probes and potential therapeutics for diseases linked to malfunctioning biological ion transport, but face several stumbling blocks in their translation to the clinic. One of these is the cytotoxicity arising from unselective ion transport. The aim of this thesis is to address this issue - by developing novel selective transporters, as well as the methods which can be used to study them, and predict novel classes. In Chapters 2 and 3, it is demonstrated that highly active and selective anionophores can be accessed by combining halogen bonding anion recognition with macrocyclic anion encapsulation. In chapter 2, their synthesis is demonstrated and transport experiments in large unilamellar vesicles (LUVs) performed, revealing record selectivity. The mechanism underpinning selectivity is then dissected to its mechanistic underpinnings in Chapter 3 through Density Functional Theory (DFT) calculations and molecular dynamics (MD) simulations at the membrane interface, demonstrating exactly how the cyclic structure imposes an energetic preference for chloride binding over hydroxide, as well as a greater desolvation of hydroxide, which further disfavours its transport. In Chapter 4, this methodology is extended to the study of stimuli-responsive transporters. The geometric and energetic components of transporter activation are decomposed, highlighting the crucial balance of explicit solvent and quality of energetic treatment. The gaps in the respective MD and DFT treatments which inhibit its applications in predictive studies are addressed by developing novel machine learning interatomic potential (MLIP) simulation methods in Chapter 5. To do this, new components are added to the *mlp-train* software package, which enable completely novel sampling strategies. These are evaluated on custom datasets to benchmark sampling quality and

efficiency. This allows for binding energy predictions to be produced using umbrella sampling simulations performed using a machine-learned potential trained on only 600 training set geometries. We anticipate that the combination of results and methods presented in this work has the potential accelerate the transition toward the use of artificial chloride transport in biology across the entire pipeline - from modelling, to synthesis and assay.

# Publications

This Thesis is based upon the following publications:

1. Chloride Selective, Non-protonophoric Ion Transport with Macrocyclic Halogen Bonding Anionophores. **M. Flerin**, F. Duarte and M. J. Langton. **Chem. Eur. J.**, 2025, DOI: https://doi.org/10.1002/chem.202502033

2. Responsive Anionophores with AND Logic Multi-Stimuli Activation. M. Ahmad, T. G. Johnson, **M. Flerin**, F. Duarte and M. J. Langton, **Angew. Chem. Int. Ed.**, 2024, DOI: https://doi.org/10.1002/anie.202403314

3. Stimuli-responsive anion transport utilizing the dynamic hydrogen bonding interactions of hydrazone-based anionophores. M. Ahmad, **M. Flerin**, H. M. Tay, F. Duarte and M. J. Langton **Nanoscale**, 2024, DOI: https://doi.org/10.1039/d4nr03220a

Contributing-author articles published or submitted for publication during the course of this degree that do not form part of this Thesis are as follows:

1. Synthesis, biological activity, and molecular dynamics simulations of LNA-neutral linkages for enhanced splice-switching antisense oligonucleotides. A. Kennett, L. Lie, **M. Flerin**, B. Z. Kurt, Y. R. Baker, A. H. El-Sagheer, F. Duarte and T. Brown. **Angew. Chem. Int. Ed.**, 2025, DOI: https://doi.org/10.1002/anie.202511386

2. Computational studies into switchable synthetic anion transporters. Z. Bo, **M. Flerin**, A. Kerckhoffs, M. J. Langton, and F. Duarte. *Manuscript in preparation.*

3. Machine learning potentials for condensed-phase reactivity. **M. Flerin**, J. Holownia, V. Juraskova, B. B. Skjelstad, H. Zhang, F. Duarte. *Manuscript in preparation.*

# Table of Contents

# List of Abbreviations

| | |
|---|---|
| Å | Ångstrom |
| AL | Active learning |
| ASL | Airway surface liquid |
| C | Celcius |
| cat. | Catalyst |
| CuAAC | Copper(I)-catalyzed azide-alkyne cycloaddition |
| COSX | Chain-of-spheres exchange |
| CPCM | Conductor-like Polarizable Continuum Model |
| DCM | Dichloromethane |
| DFT | Density functional theory |
| DMAP | N,N-dimethylaminopyridine |
| DMF | *N*,*N*-Dimethylformamide |
| DMSO | Dimethylsulfoxide |
| DPPC | 1,2-dipalmitoyl-sn-glycero-3-phosphocholine |
| E | Energy |
| $EC_{50}$ | Half maximal effective concentration |
| ECP | Effective core potential |
| EDTA | Ethylenediaminetetraacetic acid |
| eV | Electron volt |
| EDG | Electron donating group |
| EP | Extra particle (charge) |
| EtOAc | Ethyl acetate |
| ESP | Electrostatic potential |
| EWG | Electron withdrawing group |
| FCCP | Carbonyl cyanide-*p*-trifluoromethoxyphenylhydrazone |
| g | Gram |

GGA ..... Generalised gradient approximation

H ..... Enthalpy

*h* ..... Planck's constant

h ..... Hours

Ha ..... Hartree

HEPES ..... *N*-(2-hydroxyethyl)piperazine-*N*'-(2-ethanesulfonic acid)

HF ..... Hartree Fock

HOMO ..... Highest occupied molecular orbital

HPTS ..... 8-hydroxy-1,3,6-pyrenetrisulfonate

*I* ..... Intensity

*k* ..... Rate constant

*K* ..... Equilibrium constant

K ..... Kelvin

kcal $mol^{-1}$ ..... Kilocalories per mole

KOH ..... Potassium hydroxide

LED ..... Light emitting diode

LSDA ..... Local spin density approximation

LUMO ..... Lowest unoccupied molecular orbital

LUVs ..... Large unilamellar vesicles

M ..... mol $dm^{-3}$

ma ..... Minimally augmented

MAE ..... Mean absolute error

MD ..... Molecular dynamics

Me ..... Methyl

MeCN ..... Acetonitrile

MeOH ..... Methanol

MM ..... Molecular mechanics

mol ..... Mole

MP2 ..... Second-order Møller-Plesset perturbation theory

MUE ..... Mean unsigned error

NaCl ..... Sodium chloride

NBO ..... Natural bond orbital

NCI ..... Non-covalent interaction

NMR ..... Nuclear magnetic resonance

$\nu$ ..... Frequency

PES ..... Potential energy surface

Ph ..... Phenyl

PMF ..... Potential of mean force

POPC ..... 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine

Pr ..... Propyl

RESP ..... Restrained electrostatic potential

$q$ ..... Charge

QM ..... Quantum mechanics

R ..... Alkyl

$R$ ..... Gas constant

RI ..... Resolution of the identity

RMSE ..... Root-mean-squared error

RMSD ..... Root-mean-squared deviation

RRHO ..... Rigid rotor-harmonic oscillator

rt ..... Room temperature

S ..... Selectivity factor

SCF ..... Self consistent field

SMD ..... Solvent Model based on Density

T ..... Temperature

$T$ ..... Kinetic energy

TBA salt ..... Tetrabutylammonium salt

TBTA ..... Tris((1-benzyl-4-triazolyl)methyl)amine

THF ..... Tetrahydrofuran

TMT ..... Tetramethylthiourea

TS ..... Transition state

$V$ ..... Potential energy

$V_S$ ..... Electrostatic surface potential

XB ..... Halogen bonding

$E_{\mathrm{xc}}$ ..... Exchange-correlation functional

ZPE ..... Zero-point energy

# 1 |

# Introduction

## 1.1 Ion Transport in Lipid Membrane Bilayers

### 1.1.1 Role of Lipid Membranes

Since the discovery of cells as the smallest units of life by Robert Hook and the formulation of the cell theory a few centuries later, the study of cellular processes has become ever more important at the intersection of chemistry and biology.[1, 2] And it is the very boundary of the cell and its organelles, the lipid membrane, which continues to be one of the most fascinating facets of cellular life. Membranes form the fundamental structural backbone cells - acting as the barrier between extracellular spaces and the biochemical machinery within, and connect to the cytoskeleton of the cell to provide the very scaffolding on which cellular movement is organised. They also allow for sub-compartmentalisation within the cell, which enriches the multitudes of reactions at play by allowing for precise control of the organelle environment. This critical innovation of life has at its core the phospholipid bilayer, a structure held together by intermolecular forces, with hydrophobic forces between the lipid tails chief amongst them, while electrostatic repulsions, attractions, and hydrogen bonding of the head groups with water molecules additionally shape membrane curvature and shape (Figure 1.1a shows various attractive and repulsive interactions at play). Most importantly, the hydrophobic nature of the tails presents a barrier for both large and polar molecules, allowing for a careful system of checks and balances of what is and is not allowed to

pass through the membrane, while non-polar molecules and gases can freely diffuse (Figure 1.1b).[3] This creates a need for complex mechanisms of transport across the membrane and allows for establishing concentration gradients of molecules and ions, which subvert entropy and help power cellular processes. Yet when these processes fail, they cause debilitating ailments, such as cystic fibrosis, which the pharmaceutical industry has been fighting to address for decades.

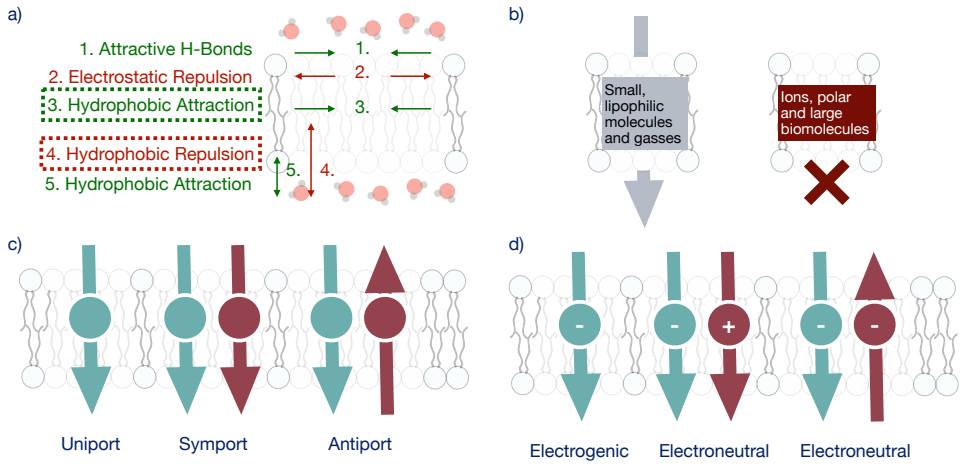



**Figure 1.1: Membrane interactions, permeability, and transport modes**. (a) Intermolecular forces at the lipid–water interface shape solute partitioning and the bilayer barrier: (1) hydrogen-bond attractions, (2) electrostatic repulsion between charged head groups, (3/4) hydrophobic attraction between tails and repulsion between tails and water favoring nonpolar species within the acyl core, which together represent the strongest forces in the membrane assembly, and (5) hydrophilic attraction of head-groups for water. (b) Consequent permeability selectivity: small, lipophilic molecules and gases diffuse, whereas ions, polar solutes, and macromolecules require protein pathways. (c) Protein-mediated transport pathways: *uniport* (single-species), *symport* (co-transport), and *antiport* (exchange). (d) Electrogenicity of transport cycles: *electrogenic* (net charge transfer) versus *electroneutral* (no net charge). Schematic summarizes concepts used throughout the chapter to contextualise chloride transport mechanisms.

### 1.1.2 Chloride Gradients and Transport in Nature

The gradients that can be found across membranes consist of both chemical components (concentration differences) and electrical components (charge differences) that together determine the thermodynamically preferred direction of molecular

movement.[3] This movement can most broadly be classified into passive and active transport. Active transport uses energy to move chemical species *against* a concentration gradient, while passive transport results in chemical species moving *with* the concentration gradient.[4] This can be simple diffusion, for hydrophobic molecules which can pass through the membrane, or facilitated diffusion, through pores and channels for bulky or polar molecules that diffuse through the membrane too slowly. Furthemore, the transport can also be classified into different categories (Figure 1.1c), based on whether it is a single species that is being transported across the membrane (uniport), whether two species are transported together in the same direction (symport), or whether two species are being exchanged across the membrane in different directions (antiport).[3] When the species being transported are charged (Figure 1.1d), the various combinations of these different directional movements can result in transport that keeps the charge balance across the membrane constant (electroneutral) or it can result in charge build-up due to more charged species moving in one direction than the other (electrogenic).

Chloride is a particularly ubiquitous and versatile participant in membrane transport processes. For example, chloride has a crucial role in epithelia which are thin tissues covering organs both on the inside (e.g. in the digestive tract) and outside (e.g. on the skin) of the body. Here, its transmembrane flux contributes to salt and water movement and to cell volume regulation. This often takes place via cation-chloride co-transporters that use the $Na^+$ and $K^+$ gradients established by the $Na^+/K^+$-ATPase to drive $Cl^-$ transport.[5, 6] In red blood cells, electroneutral $Cl^-/HCO_3^-$ exchange underlies the chloride shift - a crucial process in keeping the blood high in oxygen and low in $CO_2$. This process allows for efficient $CO_2$ transport between peripheral tissues and the lungs by coupling ion exchange to carbonic anhydrase chemistry.[7] In endosomes and lysosomes, vacuolar $H^+$-ATPases pump protons to acidify the lumen, and the charge imbalance is dissipated by a

concomitant chloride transport pathway, allowing the acidification to proceed.[8] In neuronal tissues, the chloride gradient is a key determinant of synaptic signaling, with important effects on the $GABA_A$ receptors.[6] Thus, beyond its classical role as a counterion in the background of cellular processes, the $Cl^-$ anion, through its various gradients and transport processes, has incredibly varied and important effects on cell physiology.

Based on the combination of whether the transport is active or passive, and whether the transport occurs through uniport, symport or antiport, chloride transporters can be categorised into several different types:

- **Passive chloride channels** (uniport). These protein pores allow downhill $Cl^-$ flow dictated by the electrochemical gradient, without direct coupling to another substrate, making the process overall uniport. This class includes classical ligand- or voltage-gated anion channels and the ABC protein CFTR, which, despite being gated by ATP binding and subsequent hydrolysis, functions merely as an anion channel rather than a pump.[5, 8]

- **Electroneutral anion exchange** (antiport). Carriers that alternately transport $Cl^-$ for a counter-base with a 1:1 stoichiometry, which results in charge-balanced transport across the membrane, maintaining neutrality. The red-blood-cell $Cl^-/HCO_3{}^-$ exchanger AE1 (anion exchanger 1), also known as Band 3 Protein or by the name of the gene that encodes it - SLC4A1, exemplifies this chemistry.[7]

- **Cation–chloride co-transport (CCC)** (symport and mainly electroneutral). Fixed-stoichiometry carriers of the SLC12 family (e.g., NKCCs, NCC, KCCs) that couple $Cl^-$ movement to $Na^+$ or $K^+$, and are tasked with both removing $Cl^-$ from the cell as well as loading it into the cell. These symporters harness cation gradients maintained by the $Na^+/K^+$-ATPase, and therefore

carry out what is known as secondary active transport. They redistribute chloride without direct ATP hydrolysis by the transporter itself, but the gradients they rely on to do so utilise ATP.[5, 6] These transporters have varied roles across the physiology.

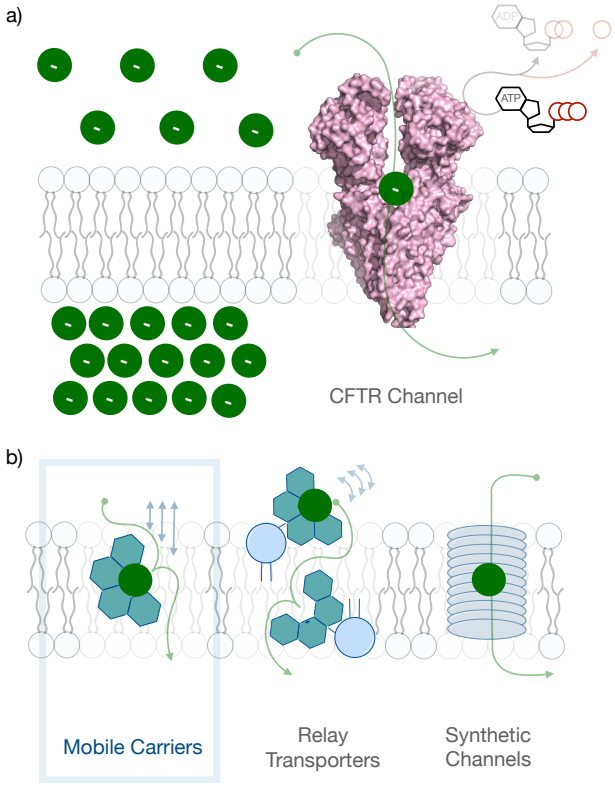



**Figure 1.2: CFTR and synthetic chloride transporters**. (a) ATP-gated CFTR anion channel in a lipid bilayer: ATP binding and hydrolysis at the nucleotide-binding domains regulate opening of a conductive pore that permits downhill $Cl^-$ flux (green). (b) Representative small-molecule strategies for transmembrane chloride transport: *mobile carriers* that shuttle anions across the hydrophobic core; *relay transporters* that pass the anion between interfacial binding sites on opposite leaflets; and *synthetic channels* that assemble a continuous pore for ion conduction.

- **CLC family** (channels and electrogenic $Cl^-/H^+$ antiporters) Proteins within the CLC fold are evolutionarily incredibly old, and function through two different mechanistic pathways. The channel type CLCs form anion pores with multi-ion occupancy and classical passive conduction along the con-

centration gradient. Transporter type CLCs, on the other hand, are $Cl^-/H^+$ antiporters that exchange two $Cl^-$ for one $H^+$ per cycle (2:1 stoichiometry), thus changing the memebrane potential difference ($\Delta\psi$) through electrogenic transport. As each turnover imports two $Cl^-$ and exports one $H^+$, it moves a net negative charge into the lumen. This provides an overall counterion movement that prevents $\Delta\psi$ build-up, and permits continued V-ATPase–driven acidification, albeit at the cost of exporting one proton per cycle.[8]

### 1.1.3 Chloride Transport in Disease

The disturbance of chloride flux across epithelia often results in disruption to normal functioning of organs with mucosal membranes, because $Cl^-$ movement sets fluid volume, pH, and mucus movement.[9] This affects surfaces such as the airways, pancreas, and intestine. One of the most important and epidemiologically widespread examples of such conditions is cystic fibrosis (CF), in which mutations in the apical (i.e. on the inner lumen surface of the epithelium) anion channel CFTR reduce $Cl^-$ secretion. This leads to dehydrated, acidic airway surface liquid (ASL) in the lungs, which causes impaired mucus clearing and chronic infection. It also leads to pancreatic insufficiency, and intestinal disease (Figure 1.2a).[10] CFTR mutations which lead to disease affect several different facets of the protein's life cycle, from defective production, to trafficking, gating, conductance and overall reduced abundance. Due to this varied nature of the causes of the disease, there are also several different modulator therapies: potentiators (e.g., ivacaftor) improve gating of the channel, whereas correctors (e.g., tezacaftor, elexacaftor) enhance folding and processing.[11–13] For the most common F508del mutation, a triple therapy of drugs developed by Vertex Therapeutics and which recently became available in the NHS (elexacaftor/tezacaftor/ivacaftor) has been shown to produce

large, rapid gains in lung function and quality of life in patients.

## 1.2 Synthetic Ion Transporters

The field of synthetic ion transport has evolved to encompass diverse mechanistic approaches that we broadly classify into three main categories in this section: **mobile carriers**, **synthetic channels** and **relay transporters** (Figure 1.2c).[4, 14–16] Each category offers distinct advantages and challenges in achieving efficacious transport of chloride across lipid bilayer membranes, which we will discuss in more detail in what follows.

### 1.2.1 Examples of Different Transporter Types

#### Mobile Carriers

Mobile carriers are perhaps the most extensively studied type of synthetic chloride transporters. They operate through a fundamentally different mechanism to structurally complex and highly ordered protein channels. They have thus far only been developed to work as passive carriers, despite somewhat provocative recent attempts at engineering light-driven "active" transport. Here, transport against a gradient active transport was achieved across liquid solvent membranes at a macroscopic scale.[15, 17, 18]. These small molecular ionophores are typically not large enough to span the entire bilayer membrane, meaning that they must associate reversibly with chloride ions to form lipophilic carrier-ion complexes which then diffuse across the hydrophobic membrane core (Figure 1.3a). The transport mechanism involves several discrete steps: anion binding at the membrane-water interface, formation of a lipophilic complex and its translocation across the membrane interior, and ion release at the opposite interface (which we discuss in more mechanistic and computational detail in Section 1.3).[19] While mobile carriers

are often highly selective for their target ions, transport rates are inherently limited by the rate of transmembrane diffusion.

Many highly successful families of mobile $Cl^-$ carriers have been developed to date, usually built around a combination of one of a handful of anion binding domains and scaffolds which together fulfil the key parameters for ion transport (Figure 1.3a and Section 1.2.5). Besides the main scaffold, which provides the overall shape and geometry of the transporter molecule, and to which the anion-binding domains are usually attached, mobile carriers also often feature appended groups which may alter the electronics of binding or the lipophilicity and size of the carrier.

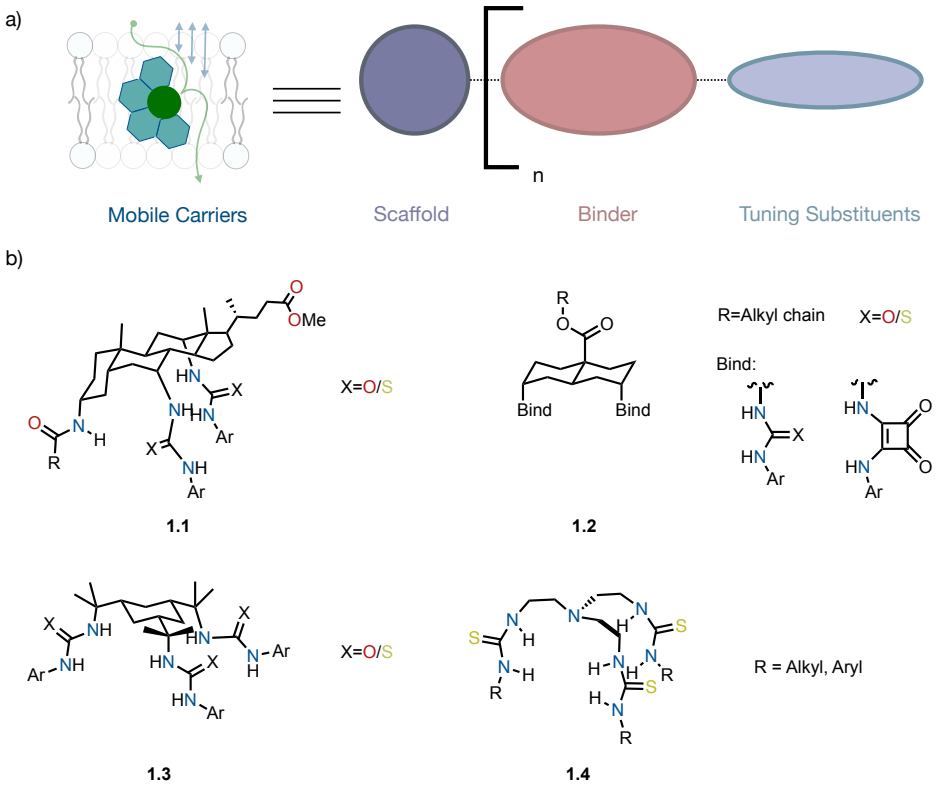



**Figure 1.3: Representative mobile chloride carriers and design logic.** (a) Schematic cartoon design of a mobile anion carrier featuring three modular elements: a hydrophobic scaffold that ensures bilayer solubility and gives the requisite geometry, a convergent H-bond binder (e.g. urea/thiourea/squaramide) that engages $Cl^-$, and tuning substituents (alkyl chains, electron-withdrawing groups) that adjust lipophilicity, acidity, and transport kinetics. (b) Examples chosen here: **1.1** a steroid-derived "cholapod" bearing multiple (thio)urea donors; **1.2** *trans*-decalin platforms presenting diaxial/multidentate (thio)ureas or squaramides; **1.3** 1,3,5-substituted cyclohexane transporters bearing (thio)ureas and **1.4** a tripodal tren tris(thio)urea.

The steroid-containing cholapod family developed by Davis and colleagues exemplifies one of the earliest cases of highly effective mobile carriers (Figure 1.3b, **1.1**). Based on a scaffold offering inherent lipohilicity, but enclosing a binding site of highly acidic hydrogen bonds, stemming from urea, thioruea or squaramide moieties, they offer excellent anion binding properties.[15, 20] These bile acid-derived molecules transport chloride via an exchange mechanism, with transport rates directly correlating with anion binding affinity across four orders of mag-

nitude (with binding constants or $K$ values ranging from $10^7$ to $10^{11}$ $M^{-1}$). [16, 21] Remarkably, unlike many carrier systems that suffer from transport inhibition at high binding affinities, cholapods maintained efficient transport even with extremely tight chloride binding.[15] Several variants of the family were developed, with a variety of substituents on the H-bond donor nitrogens offering tunable properties.

A synthetically simpler scaffold which emerged afterward was the *trans*-decalin family, which presented a more compact platform that still retained the preorganisation and rigidity of the cholapods (Figure 1.3b, **1.2**).[22, 23] They contain two axial binding groups, which can be interchanged for ureas or squaramides. The first diaxial diureidodecalins showed very high activity using $Cl^-/NO_3{}^-$ exchange assays in large unilamellar vesicles (LUVs). High degrees of tunability could be achieved using the electronics of the aryl substituents and alkyl tail length. Replacement of the urea binding groups for thiourea groups resulted in even better transport properties and bindng affinities.[23] Crucially for potential biological applications, cell-based assays confirmed strong, persistent transport with low cytotoxicity for a diureidodecalin in epithelial cell assays.[24]

Developments in chloride carriers showed that even more minimalist scaffolds can yield excellent performance, notably the cyclohexane-based anionophores from the Davis group (Figure 1.3b, **1.3**).[25] By mounting three thiourea donors around a cyclohexane core, the group produced record $Cl^-/NO_3{}^-$ antiport activities despite reduced preorganisation and lower binding affinities compared to earlier steroid and decalin systems. This indicated that a phenomenon they termed "controlled flexibility" can accelerate carrier turnover, and that the transport process and its overall rate depended on a delicate balance of the different transport steps.

Taking scaffold substitution back a notch further, both 1,2-disubstituted cyclo-

hexane and even 1,2-disubstituted benzene transporters have been shown to be effective, with both urea- and squaramide-based binding groups.[26, 27] Interestingly, they displayed a high level of toxicity in cells unlike the aforementioned decalin-based transporters. The difficulty of predicting toxicity is thus shown to remain a great challenge in developing medicinally relevant transporters.

The final example highlighted is perhaps the simplest of all discussed in this section — the tren (tris(2-aminoethyl)amine) tripodal family, in which three urea or thiourea H-bond donors are appended to a small amine core (Figure 1.3b, **1.4**). These receptors are typically made in a single step from tris(2-aminoethyl)amine and (iso)thiocyanates but despite their simplicity display very high chloride transport rates at very low loadings. Activity trends have been shown to correlate with lipophilicity (and fluorination) more than binding constants, with peak performance around clogP $\approx$ 8.[28, 29]

**Synthetic Chloride Channels**

Synthetic chloride channels and pores operate by a mechanism in which they create an opening in the membrane, with a lumen which is hydrophilic. This allows them are to transport ions at higher rates than mobile carriers, approaching those achieved by protein channels found in living cells, as evidenced by single molecule conductance measurements.[16, 30–36]. Recent reports of synthetic channels report a single molecule conductance of 20 pS, while the single molecule conductance of the CFTR channel has been measured to be around 6–10 pS.[37, 38] One issue of channels is that it is harder to achieve anion vs. cation specificity compared to mobile carriers.

Most often, channels form from stacks of monomeric, macroyclic molecules, which already contain a lumen in their monomeric form (e.g. **1.5** in Figure 1.5). For example D,L-$\alpha$- (and related $\beta/\gamma$) peptides such as **1.5** are a good example of a pore with

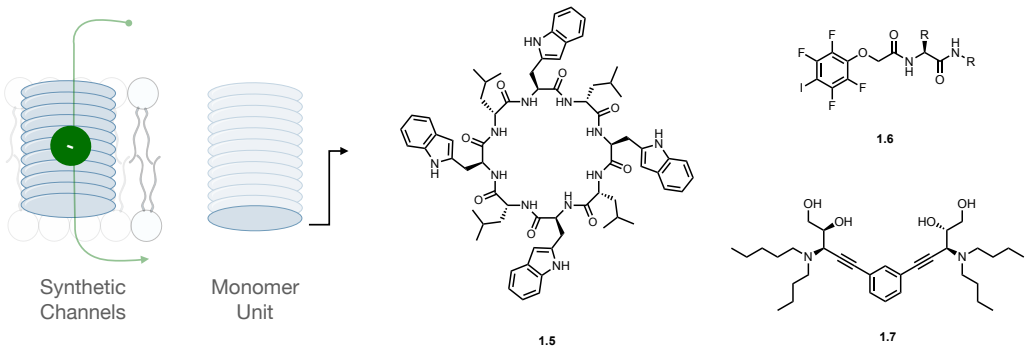



**Figure 1.4: Representative synthetic chloride channels.** Left: schematic of a self-assembled, water-filled pathway spanning across a lipid bilayer, showing how the monomers generate a continuous pore. Middle and right, clockwise: **1.5**, an alternating D/L cyclic peptide that is able to stack into nanotubes which are able to facilitate transport across membranes[39]; **1.6**, an iodinated perfluoroisophthalamide (halogen-bond donor) that assembles into chloride-selective channels [40]; and **1.7**, an amphiphilic phenylene-based monomer bearing tertiary amines and phenolic OH groups that self-assembles into membrane-spanning channels[41].

a lumen size controllable by the make-up of the peptide.[39, 42] These structures stack into nanotubes with a hydrophilic interior, and their ring size controls pore diameter. Higher selectivity in channels can be achieved using specific interactions with the chloride ion, such as the halogen-bond–driven chloride channels reported by Sharma *et al.* [40] Small 5-iodo-isophthalamide halogen-bonding motifs self-assemble into chloride-selective column structures in bilayers (**1.6**, Figure 1.5). Barrel-rosette ion channels are a variant of self-assembled membrane-spanning channels, which offer more sophisticated mechanisms of achieving selectivity.[41] They are constructed from rigid molecular building blocks such as **1.7** featuring chiral vicinal diols tethered to 1,3-diethynylbenzene cores, which are able to interact intermolecularly and form a supramolecularly assembled, petal-like structure (hence rosette). These are then able to stack on top of each other, forming a continuous lumen across the membrane interior for ions to pass through (hence the barrel). The transport activity can be precisely tuned by controlling the lipophilicity of the constituent monomers, with the most active molecules demonstrating

selective chloride transport via antiport mechanisms.[41]

### Relay Transport Mechanisms

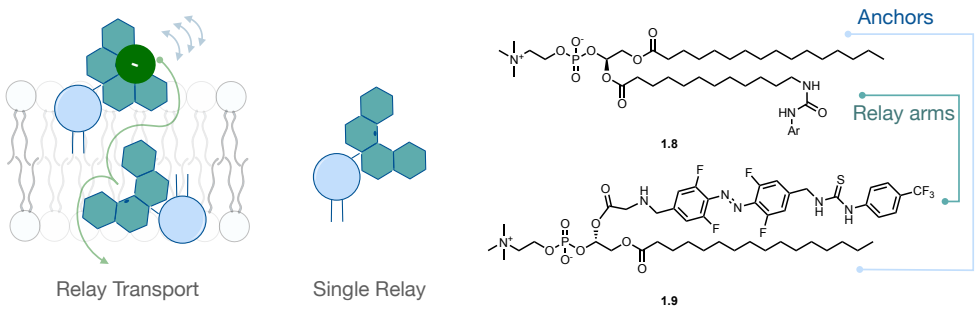



**Figure 1.5: Relay transport concept and examples.** *Left:* Cartoon of a relay mechanism in which two monolayer-anchored carriers pass an ion from one membrane leaflet to the next without moving across the entire membrane core with the ion by themselves or forming a continuous pore. *Centre and right:* A cartoon representation of a single relay, featuring a phospholipid anchor and a relay "arm" for shuttling ions across the membrane. Representative lipid-anchored relays include **(1.8)**, the groundbreaking relay transporter reported by Mcnally *et al.* [43] and **(1.9)**, the photoswitchable relay transporter reported by Johnson *et al.* [44]

Relay transport is an innovative approach where membrane-anchored binding groups (relay transporters) transfer an anion along the membrane by passing it between individual transporter molecules, rather than the groups diffusing across the entire membrane themselves. It has been shown that relays must be positioned on opposite sides of the membrane, thus proving that two relay molecules operate by passing ions.[43–45] The pioneering relay transporter design of McNally *et al.* consists of phosphatidylcholine derivatives with urea groups appended to the sn-2 acyl chain terminus.[43] These amphiphilic molecules insert into the membrane bilayer with their phospholipid head groups anchored at the membrane interfaces and their urea functionalities able to move between the membrane core and lipid-water interface. Recent advances from the Langton group have introduced halogen bonding as an alternative to hydrogen bonding for relay transport mechanisms.[44, 45] These systems exhibit chloride selectivity arising from lower kinetic barriers for chloride exchange between membrane-anchored receptors compared to hydroxide

transport.

### 1.2.2 Medicinal Applications

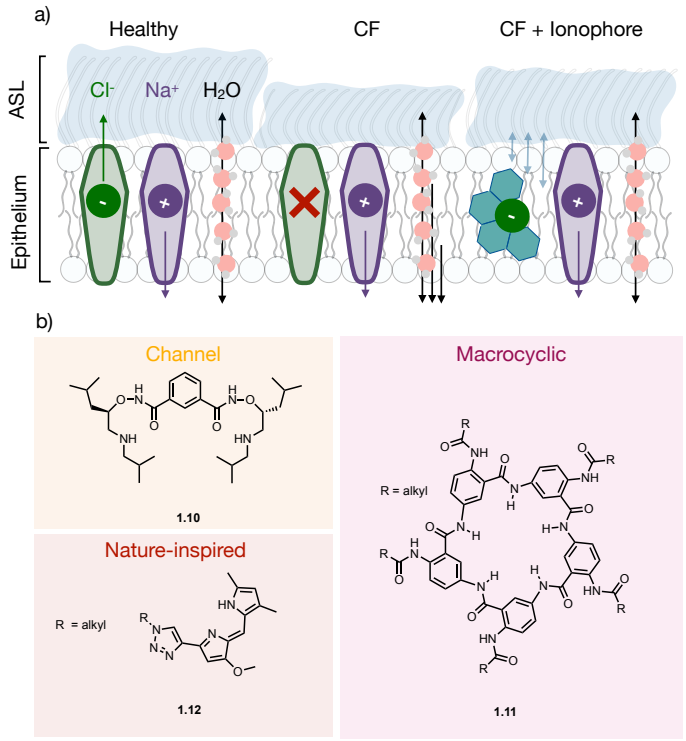



**Figure 1.6: Potential Applications of Ionophores as Cystic Fibrosis Treatments (Bypass Therapy)** a) In healthy airway epithelia, apical chloride secretion by the CFTR protein and regulated sodium absorption maintain a hydrated airway surface liquid (ASL). In cystic fibrosis (CF) patients, the loss of CFTR-mediated $Cl^-$ secretion and heightened $Na^+$ absorptions dehydrate the ASL, causing it to reduce in thickness and increase in viscosity, slowing its clearance by cilia. Introducing a membrane transporter therapeutic provides an alternative pathway for anion movement down the concentration gradient, drawing water osmotically and partially restoring ASL height. b) Representative examples of $Cl^-$ transport-restoring transporters: an isophthalamide $\alpha$-aminoxy amino acid membrane-forming structure **1.10**; a macrocyclic carrier featuring an aromatic pentaamide core in **1.11**; $R$ = alkyl); and a nature-inspired carrier, with a prodigiosin-like scaffold **1.12**.

The therapeutic application of synthetic chloride transporters has emerged as a promising area in treating diseases associated with ion transport, such as the aforementioned cystic fibrosis, as well as for cancer therapy and antimicrobial

treatments.[4, 14] In cystic fibrosis, the transport of $Cl^-$ ions across the apical membrane of the epithelium also causes dysregulation of the associated $Na^+$ transport, causing a reduction of the airway surface liquid (ASL) in the lungs and the clearance of pathogens by the cilia (Figure 1.6a). This causes mucus build-up and repeated infection. Several key studies have shown the efficacy in restoring transport in CF epithelial cell lines.[36, 46, 47] If successful in the clinic, these transporters would offer a mutation-independent therapeutic approach by providing alternative pathways for chloride transport, potentially compensating for defective CFTR function.

Studies in cellular systems have already demonstrated the feasibility of this approach, using both synthetic chloride channels and mobile carriers. Shen *et al.* showed that a synthetic chloride channel **1.10** could restore chloride conductance in human CF epithelial cells, with transport efficiency approaching that of natural CFTR at $\mu$M concentrations (Figure 1.6b).[36] Besides high activity, the channel demonstrated selective chloride permeability, which was key in restoring the transepithelial chloride transport in a cellular model.

Macrocyclic transporters **1.11** developed by Gong's group also represented a significant advancement in the field of synthetic transporters and their CF therapeutic development. These aromatic pentaamide macrocycles arrange in five-pointed star configurations that create selective chloride transport pathways across cell membranes.[46] When tested in CF lung cell models, they increased airway surface liquid thickness by over 50% compared to untreated controls (7.8 ± 1.7 $\mu$m vs. 5.1 ± 1.4 $\mu$m), indicating restoration of normal mucus hydration, which is one of the key pathological symptoms of cystic fibrosis in the lungs. The compounds also demonstrated transport efficiencies comparable to CFTR while maintaining excellent biocompatibility.

Finally, in studies of prodigiosin-inspired anionophore family **1.12**, it was shown that these transporters have particular promise for CF applications due to their optimal activity under pH conditions characteristic of CF airway epithelia.[47] These synthetic transporters, based on triazole-modified prodigiosin structures, exhibited very high transport activities while also exhibiting reduced cytotoxicity compared to the natural product. Importantly, they maintained activity under acidic pH, which is a key requirement for their application in the affected lung tissues. The approach of using $Cl^-$-transporting molecules could potentially benefit all CF patients regardless of their specific genetic defect, addressing a major limitation of existing therapies.

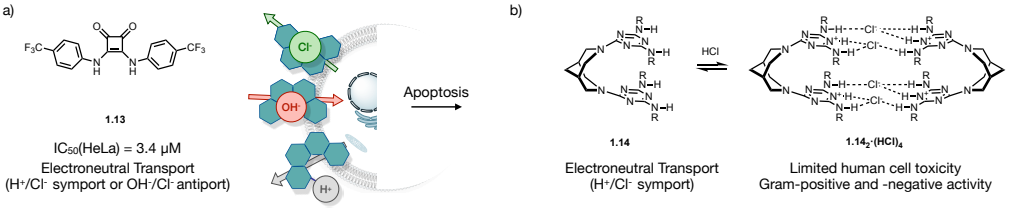



**Figure 1.7: Anticancer vs antimicrobial anionophores.** a) A bis(trifluoromethyl)–substituted squaramide anionophore (1.13) carrying out pH-dissipating, electroneutral $Cl^-$ transport, perturbing chloride homeostasis and triggering apoptosis (with an $EC_{50}$ in HeLa cells 3.4 $\mu$M). b) A macrocyclic mobile $Cl^-$ carrier (1.14) that co-transports HCl to give $(1.14)_2{\cdot}(HCl)_4$ and functions as an electroneutral $H^+/Cl^-$ carrier with broad Gram-positive as well as Gram-negative antibacterial activity and low mammalian cell toxicity.[48]

In addition to CF applications, some synthetic anion transporters have demonstrated significant potential as anticancer agents by drastically affecting cellular osmosis, both cellular and organelle pH gradients, and thus overall cell homeostasis. A bis(trifluoromethyl) squaramide **1.13** reported by Busschaert *et al.* acts as an electroneutral $H^+/Cl^-$ co-transporter (or $OH^-/Cl^-$ antiporter, *vide infra* for discussion on protonophoric selectivity), rapidly perturbing intracellular $Cl^-$ concentrations and dissipating pH gradients of both the cytosol and the lysosome, which triggers cellular death at low-$\mu$M levels.[49] Broad-spectrum

antibacterial activity was demonstrated against multidrug-resistant Gram-positive bacteria by a bispidine-based carrier **1.14** which also functions as both a $Cl^-$ carrier and protonophore ($H^+$ transporter) in a co-transporting mechanism. Ko *et al*. showed broad activity against multidrug-resistant Gram-positive pathogens including MRSA and VRSA. Key findings included a remarkable potency against the bacteria, with minimal inhibitory concentration values ranging from 0.5-2 $\mu$g/mL against resistant pathogens while maintaining excellent selectivity ($EC_{50}$ = 810 $\mu$g/mL against mammalian cells), which is key for therapeutic applications. The compound exhibited rapid kinetics of bacterial clearance, decimating bacterial cells by over 5 orders of magnitude within 12 hours.[48]

### 1.2.3 Chloride Selectivity

Any therapeutic replacement for defective CFTR channels must deliver chloride ions across epithelial membranes without disrupting the intracellular pH homeostasis. In cell and tissue models, many small-molecule anionophores move $Cl^-$ efficiently but also carry $H^+$ (via reversible deprotonation) and/or $OH^-$ (via hydrogen-bonded complexes), which causes disruption to pH gradients and drives cytotoxicity. Selective nonprotonophoric $Cl^-$ transport, where $Cl^-$ is orders of magnitude faster than any $H^+/OH^-$ conduction is therefore one of the key requirements for CF applications. Textbook systems that show such behaviour in vesicles and cells now exist, and their design principles are becoming clear.[50–55] We introduce these principles alongside the assays used to determine selectivity in the section below.

Gale and co-workers gave the first rigorous definition and read-out for $Cl^- > H^+/OH^-$ selectivity using a modified HPTS base-pulse assay[50]. Large unilamellar vesicles are loaded with a buffered sodium chloride solution, the external pH is raised with base, and the intravesicular pH is monitored by HPTS (8-hydroxypyrene-

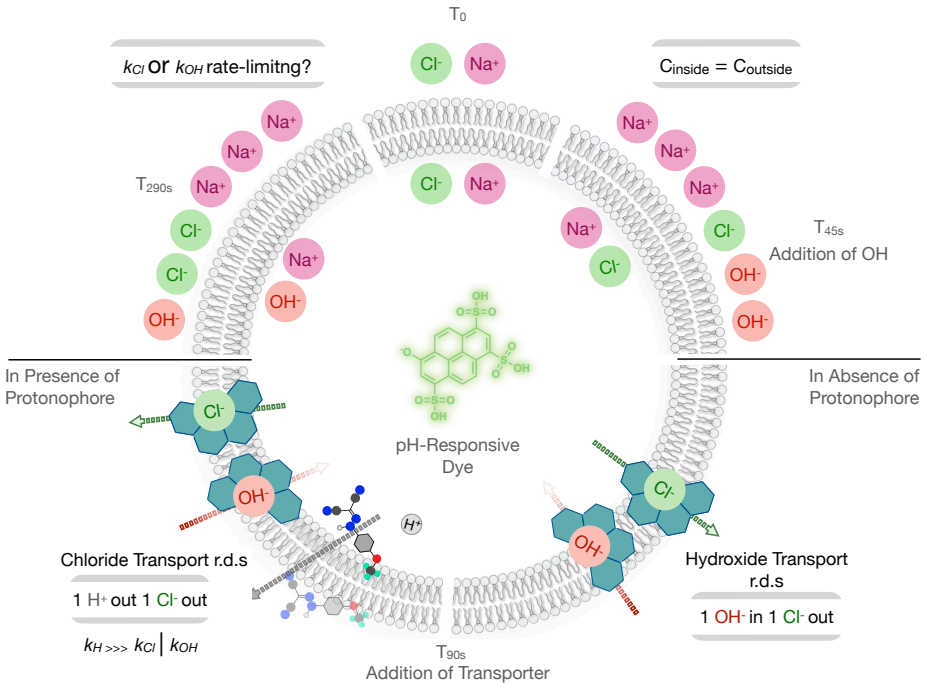



**Figure 1.8: Assay for chloride selectivity**[50, 52] Large unilamellar vesicles first experience an external base pulse. After the addition of transporter, two different regimes can operate, based on whether a protonophore (proton carrier) is present or not. In the absence of a proton carrier, the observed pH change is limited by ($H^+/OH^-$), movement; in the presence of a protonophore such as FCCP (Carbonyl cyanide-p-trifluoromethoxyphenylhydrazone, which provides a fast $H^+$ pathway), the pH response reports the intrinsic $Cl^-$ step. It should be noted that only the antiport mechanism of $Cl^-/OH^-$ transport is shown here, but the pH dissipation step and therefore the selectivity factor determined by these assays could equally be a result of $Cl^-/H^+$ symport.

1,3,6-trisulfonate). The base pulse, added after 45 seconds, creates a pH gradient, which drives $OH^-$ transport into the vesicle, with concomitant $Cl^-$ transport out of the vesicle in the case of antiport, or transport of both $Cl^-$ and $H^+$ out of the vesicle. Repeating the experiment with a proton transporter, also called a protonophore, such as the carrier FCCP or the channel gramicidin D, offers a way to distinguish mechanisms of transport and rate-limiting steps. If pH dissipation becomes markedly faster on adding a protonophore, then $H^+/OH^-$ movement was the slow step and the anionophore exhibits $Cl^->H^+/OH^-$ selectivity, i.e. $Cl^-$ transport is intrinsically faster, but needs a proton pathway to maintain electroneutrality. If

the protonophore has little effect, $Cl^-$ movement is rate-limiting or the system is intrinsically protonophoric (able to carry either $OH^-$ or $H^+$ ions to dissipate pH gradients).[50] A schematic of the logic and read-outs is shown in Figure 1.8, while a more detailed explanation of the steps of the assay and the concomitant changes in the normalised fluorescence plots is given in Figure 2.9.

Through the various systems that have been found to be selective through this validation assay, a few design rules for increasing $Cl^-$ selectivity have been elucidated.

1. **Reduce acidity of binding sites.** Acidic N–H donors (e.g. thioureas) can be deprotonated, resulting in protonophoric transport, and they may also bind and transport $OH^-$ strongly; lowering acidity and avoiding ionizable N–H groups diminishes protonophoric pathways. [50]

2. **Exploit $\sigma$-hole interactions (halogen/chalcogen bonding).** Neutral halogen- or chalcogen-bond donors (e.g., iodotriazoles, telluromethyltriazoles) couple high lipophilicity with directional, Lewis-acidic contacts, enabling intrinsically higher $Cl^-$>$OH^-$ selectivity without N–H acidity. [51, 52]

3. **Increase encapsulation / preorganisation.** Higher encapsulation correlates with improved $Cl^-$>$H^+$/$OH^-$ selectivity. [50, 54]

### 1.2.4 Halogen Bonding

A halogen bond (XB) is the net attractive interaction between an electrophilic region on a covalently bound halogen atom and a nucleophile.[56, 57] This arises from a region of positive electrostatic potential (ESP) known as $\sigma$-hole, located on the extension of the R–X bond and a nucleophile, such as a lone pair or an anion. Its strength and directionality arise from an anisotropic electrostatic potential reinforced by polarization, dispersion, as well as charge-transfer component.[56,

58, 59] XB potency increases down the halogen group (where strength follows: I > Br > Cl) as shown by the increasingly positive ESP in Figure 1.9a. Additionally when the R group in R-X is an EWG, the $\sigma$-hole is also deepened. Perfluoroaryl or perfluoroalkyl iodides and iodotriazoles have been widely used. Critically for anionophores, XBs are strongest in media of low dielectric constant and exhibit near-linear geometry.[57]

Both features are well matched to lipid bilayers, where halogen bonding has been increasingly used in transporters.[52, 53, 55]. The first examples date back to 2012, where Matile and co-workers were the first to demonstrate halogen-bond-mediated transmembrane anion transport in lipid bilayers, combining it with DFT studies on the transport-active species and its association with ions (Figure 1.9b and c).[60–62]

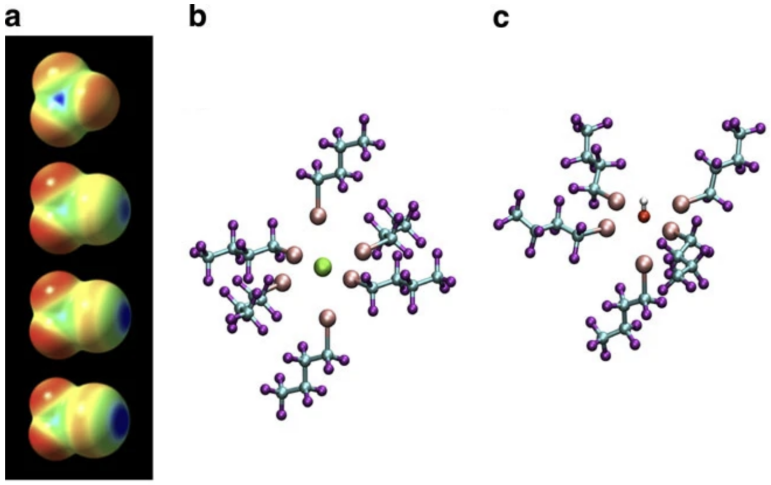



**Figure 1.9: First example of halogen bonding membrane transport developed by the Matile group.** a) Electrostatic potential surfaces of $CF_4$, $CF_3Cl$, $CF_3Br$, $CF_3I$ - blue is positive, red is negative electrostatic potential (ESP). b) and c) show 6 and 5 perfluoro-1-iodobutane molecules coordinate to $Cl^-$ and $OH^-$, respectively. Reproduced with journal permissions.[60]

### 1.2.5 Design Rules for Activity

Besides $Cl^-$ selectivity, transporters must also deliver high levels of activity in transport assays to be able to transition into cellular study and minimise other sources of toxicity arising from non-specific interactions with biomolecules. The discussion below focuses on the design parameters that have been found to determine activity.

**Chloride binding affinity ($K_a$).** Perhaps the most obvious design parameter is the ability of the transporter to bind the ion it needs to transport (Figure 1.10, row 1). But here, a picture of a delicate balance, or so-called "Goldilocks" effect emerges. It has been shown that transport turnover requires both efficient interfacial capture of $Cl^-$ from water and good release kinetics on the opposite side of the membrane. Across some families, activity rises with $K_a$ up to this "Goldilocks" window beyond which the complex binding becomes too persistent and limits the rate. For example, overly strong, highly preorganized tris-squaramides were found to trap $Cl^-$ and transport poorly once the binding affinities got too high.[22, 23, 29] On the other hand, this release-limiting behaviour did not seem to be an issue in steroidal cholapods, which showed a positive correlation between stronger (thio)urea-based halogen bonding groups, higher $Cl^-$ affinities, and faster $Cl^-/NO_3^-$ exchange over several orders of magnitude.[15, 16, 20, 43] Complex dynamics clearly also govern ion release, which may have a scaffold-dependent specific mechanism, and require further study.

**Size and shape**. The geometry that donors adopt around the anion governs both extraction and how easily the carrier crosses the hydrophobic core (Figure 1.10, row 2). Rigid scaffolds that present convergent binding groups and undergo only low conformational penalties upon binding — e.g., diaxial di(thio)ureidodecalins — were estimated to deliver record single-molecule transport rates of around

$\sim 8.5 \times 10^2$ $Cl^-/s^{-1}$. This indicates that they were able to offer both efficient capture of the anion while also retaining diffusional mobility through the bilayer.[22, 23] Increased encapsulation (within macrocycles such as cholaphanes) can further stabilise the bound state, assist with interfacial desolvation, and improve performance when release remains fast, but excessive bulk can slow lateral diffusion and offset any activity thus gained.[15]

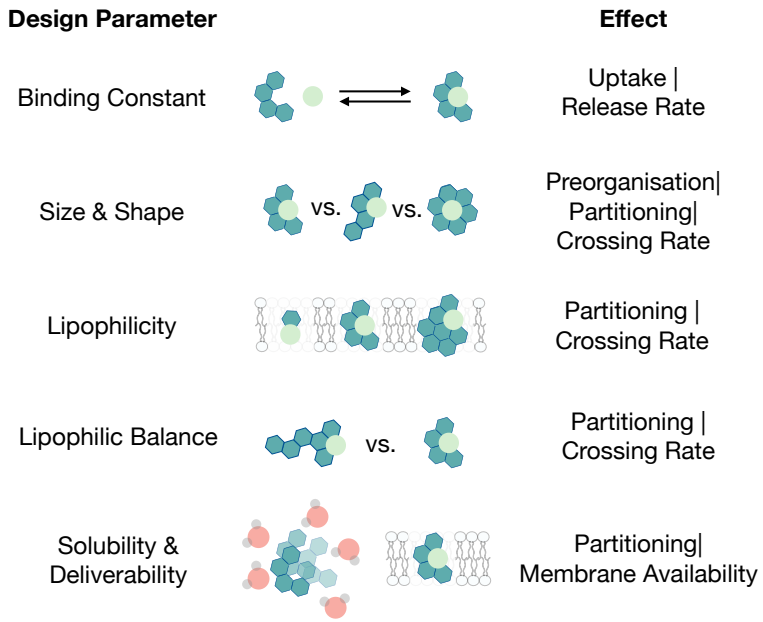



**Figure 1.10: Design parameters controlling activity of mobile chloride carriers.** Schematic summary mapping structure to mechanism. (1) Binding constant ($K_a$) tunes the extraction–release trade-off (with some families showing clear signs of "Goldilocks" behaviours - too weak: poor uptake; too strong: slow release). (2) Size & shape govern binding group convergence and preorganisation versus bilayer mobility. This affects both membrane partitioning and transmembrane crossing. (3) Lipophilicity sets bilayer uptake and residence position, as well as the rate of membrane diffusion. (4) Lipophilic balance - meaning the placement of hydrophobic moieties around the binding site — is able to bias interfacial residence positions and promote (or hinder) core crossing at fixed overall LogP. (5) Solubility & deliverability determine how much transporter reaches the membrane under external dosing, independent of intrinsic activity.

**Lipophilicity (logP).** A carrier must partition into, and remain within, the membrane (Figure 1.10, row 3). Within the tren- tris(thio)urea series, increasing lipophilicity through increased substituent fluorination boosts activity until an op-

timum around clog$P \approx$ 8, beyond which mobility and deliverability suffer.[29] This modification also modulates donor strength, as electron-withdrawing groups increase acidity and binding, and thus feeds back into the $K_a$ window above, which makes the effects difficult to disentangle. Different transporter families also appear to have different optimal lipophilicities, with other works reporting optimal lipophilicity windows closer to LogP $\sim$ 5.[63]

**Lipophilic balance**. With overall LogP held constant, activity can vary strongly depending on the distribution of hydrophobic groups appended to the transporter (Figure 1.10, row 4). Designs that surround a central binding site with balanced hydrophobic domains disfavor residence at the polar interface, and promote passage across the core, giving markedly higher turnover than analogues where lipophilicity is concentrated at one end.[64] This principle can also be thought of in tandem with preorganisation. By designing a compact, convergent binding site alongside a series of balanced hydrophobic groups, the transporter will typically outperform a diffuse site with a single large tail (refer to Figure 1.3a for a schematic of mobile carrier structure).

**Solubility and deliverability**. High intrinsic activity in vesicle transport assays where the transporter has been loaded into the lipid prior to vesicle formation can be negated in cells if the compound fails to transfer from the dosing medium into membranes ("deliverability", as showin Figure 1.10, row 5). o-Phenylene bisurea studies quantified this by comparing rates for external addition against pre-incorporated transport rates, revealing that the most lipophilic analogues had the poorest deliverability despite excellent intrinsic transport.[65] Due to this, external dosing has been the more popular method for transporters with potential medicinal applications. In order to improve deliverability, classic solubilising strategies, such as adding polar handles and attempts at preventing intermolecular association (e.g. by $\pi$-stacking) are possible options. More sophisticated options also include

proanionophore strategies (e.g., glutathione-triggered masks or peptide tags) that enhance aqueous handling and release the active transporter *in situ*.[66, 67]

## 1.3 Stages of Chloride Transport

The transport of anions by mobile ionophores is usually divided into distinct stages: anion capture, membrane crossing, anion release, and membrane recrossing by the ionophore (the three steps usually modelled are shown in Figure 1.11). The rate-determining steps have been difficult to elucidate experimentally, due to the interconnectedness of transporter parameters (*vide supra*) and the inability to directly measure individual transport steps in an experimental setting.[4, 14, 68] Computationally, the process has been modelled in its individual steps, but the difficulty of studying all steps of the transport process using a single unified computational method makes definitive conclusions based on theoretical studies alone challenging. Below, we discuss the most important computational studies into $Cl^-$ transporters, divided into the individual transport steps and the computational methods used to study them.

### 1.3.1 Ion Binding

The first step of the transport process is the binding of the ion at the transporter's binding site, which has been shown to occur at the membrane bilayer's boundary between the tail and the headgroup regions, where unbound mobile carriers generally reside at equilibrium.[71–73] This step has usually been modelled using density functional theory methods (DFT, for a theoretical introduction see Theoretical Appendix Section C.1), which allow for good estimates of binding energy trends between different molecules. DFT is also very commonly used to ascertain the binding modes of ions, which is especially important for large and flexible

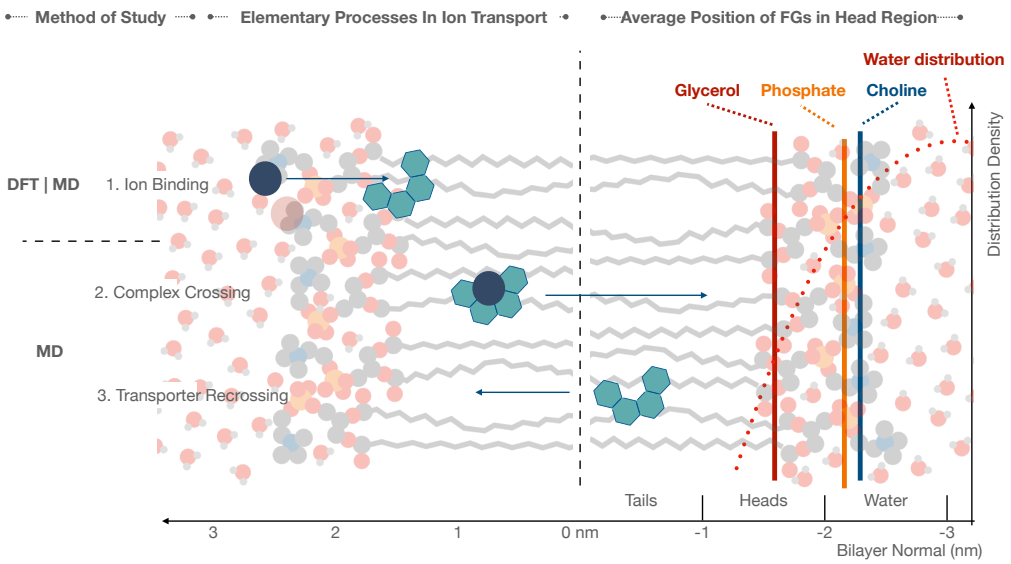



**Figure 1.11: Schematic of stepwise transport and a POPC bilayer's composition**. The ionophore's stepwise transport of an ion across the bilayer is mapped to the methods that are generally used to study the respective steps, and detail of the membrane components is shown. Left: The transport cycle is split up into: (i) interfacial capture of $Cl^-$ by the carrier, and is most commonly studied using DFT methods; (ii) [transporter·ion]$^-$ complex translocation across the membrane's hydrophobic core is usually studied using classical molecular dynamics (MD) in model membranes; and likewise (iii) the recrossing of the of the free carrier to reset the transport cycle is usually modelled using classical dynamics. Right: The bilayer normal distances ($z$) show the various the mean positions of choline, phosphate, and glycerol head groups as well as the interfacial water-density maximum.[69, 70]

transporters, and can be used to cross-reference these bound minima against crystal structures. Other indicators of binding strength, such as the maximum electrostatic potential found at the transporter's binding site atoms (e.g. acidic protons or $\sigma$-hole iodines), or energy decomposition analyses can also be used to quantify the energetic contributions to binding and guide further transporter development. For example, Bickerton *et al.* showed that the increasing transport activity in their triazole-based transporters correlated well with the increase in electrostatic potential at the binding site (Figure 1.12a), while Spooner *et al.* found similar

trends. [71, 74]

But while density functional theory is good at distinguishing the binding energy trends between very similar compounds, e.g. with different substitution patterns of electron withdrawing- and donating- groups, there are many potential problems with trying to use it to rationalise transport trends between highly disparate transporter types. When two transporters from the same family sit in the membrane in a very similar pose, this means that all conformational effects and membrane interactions are in effect held constant. Thus, their electronics alone, and thus gas-phase or implicit solvation DFT binding energies are a sufficient surrogate for binding trends. However, more significant differences in the transporter may result in its equilibrium membrane pose changing significantly, which would mean that other steric and electronic factors may influence the ability of the ion to enter the binding site. The transporter may have a different tilt, distance from the membrane core, or may associate differently with the charged headgroups if the geometry, lipophilicity, size or shape of the transporter change significantly.

This brings us to an additional issue with DFT methods for studying binding, which is the lack of explicit representation of the specific membrane environment in the DFT binding energy cycle. Anionophores feature highly Lewis acidic binding sites in order to associate with anions, and so it is highly likely that many interactions with Lewis basic and negatively charged species in the headgroup region would need to be displaced by the anion upon binding. This multitude of different interactions cannot be easily captured by explicitly including parts of these residues, and implicit solvation models can at best mimic the general dielectric constant of the area of the membrane where binding occurs. In fact, even the comparatively homogenous solvent environment is already too challenging for quantitive binding energy prediction using DFT (see section 1.5.1), which means that interpretation of DFT results for transport property analysis must be carefully

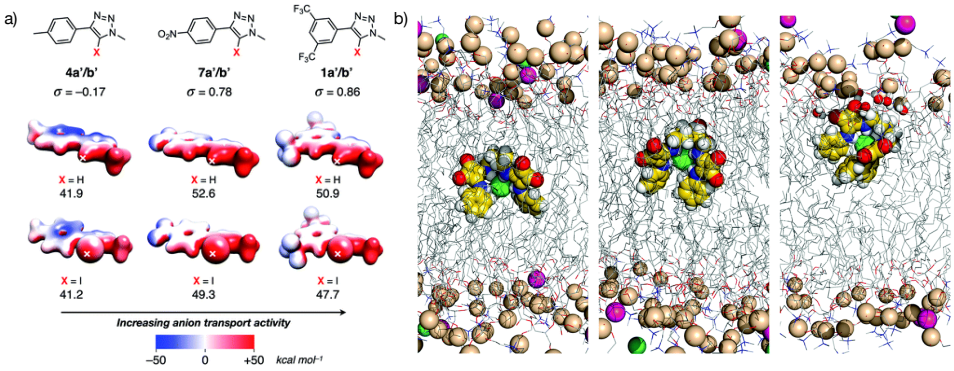



**Figure 1.12: Literature examples of how DFT and MD differ in the steps of the transport process that they are used to study.** a) DFT is usually used to study binding strength of different transporters. Here, Bickerton *et al.* showed that the electrostatic potential - a property not amenable to study by classical dynamics - correlates with transport in a closely related family of transporters. b) Molecular dynamics study of tren-based transporters by Spooner et al, showing how the translocation of the [transporter·Cl]$^-$ complex can be studied using molecular dynamics. Both parts of the scheme adapted from literature with permissions.[52, 71]

framed.

Using a membrane-mimetic implicit solvent was the approach taken by Bickerton et al, in their study of the selectivity of the bis-triazolo transporters **1.15**-**1.17** (Figure 1.13). Having found that selectivity increased greatly upon substitution from the **1.15** proto- to the **1.16** iodo- and finally **1.17** telluro- variant (see section 1.2.3), a thermodynamic cycle was constructed for the binding process to probe whether there was an inherent binding preference originating from the binding interactions of the chloride with the different Lewis acids. The thermodynamic cycle consisted of the two ions of interest, $Cl^-$ and $OH^-$, moving from the implicitly solvated aqueous phase to a chloroform phase, which mimics the intramembrane region where ion binding with the transporter occurs. The next step was binding of each ion to the transporter in $CHCl_3$, and the comparison of the associated binding energies. Though the authors noted that implicit solvation models often struggle with reproducing solvation energies for ions, the trends of the respective

ions' binding energies showed some interesting trends. While the $Cl^-$ binding energy did not change significantly from the less selective **1.16** (S = 5.3) to the more selective **1.17** (S = 67), the hydroxide binding energy became less negative in the more selective telluromethyl variant. This indicated that preferences of transporters for different ions can correlate well to experimental observables when modelled using DFT binding studies.

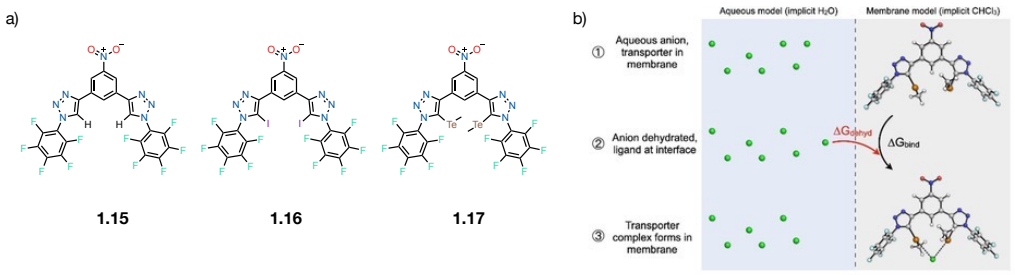



**Figure 1.13: Selectivity modelling by Bickerton et al**[74] a) Three $Cl^-$ transporters with different levels of $Cl^-$>$OH^-$/$H^+$ selectivity. b) Thermodynamic cycle showing the ion binding process using implicit solvation.[74] Part b) reproduced with permissions.

An alternative is to carry out binding studies using molecular dynamics simulations, which can be carried out in model systems with explicit consideration of thousands of atoms, for observation of the behaviour of entire sections of the membrane bilayer (e.g. Figure 1.12b). This considers of all relevant transporter-membrane interactions, including the interactions between polar sections of the transporter with the headgroups and the non-polar sections with the tails of the membrane, which will influence the distance and angle of the transporter in the membrane.

### 1.3.2 Membrane Translocation Steps

Due to the ability to capture the various transporter-membrane interactions, molecular dynamics studies in model membranes have been the method of choice for

the study of transporter crossing and recrossing steps (Figure 1.11 middle and bottom). In these simulations, POPC bilayer sections spanning a few nm of membrane (usually equalling a few hundred lipid molecules), as well as the neighbouring water slabs, are usually simulated on the order of several tens or hundreds of nanoseconds.[71–73] Lipid membranes composed purely of POPC are usually good models for simulations which aim to reproduce conditions of vesicle assays that are used to test ionophores. More complex membrane compositions, including DOPC/POPC mixtures, as well as membranes containing sphingomyelins and cholesterol molecules can be used for certain studies where permeation, packing or membrane diffusion need to be modulated in a biologically inspired manner.[75, 76]

The ionophore translocation problem is in essence a study of membrane permeability, which has precedent in literature, e.g. for the study of the permeability of drug-like molecules across biological membranes.[2, 77–79] Molecules can be compared using the the translocation free-energy profiles or potentials of mean force (PMFs) across the membrane bilayer normal. These can be obtained by performing umbrella sampling simulations along the bilayer normal axis, or alternatively through metadynamics simulations along the same coordinate. For small-molecule diffusion, very long microsecond-scale simulations have also been used to observe crossing events directly without applying biases to enhance the sampling. Unbiased dynamics are also very often to represent factors which are known to influence the transport process without directly quantifying the energetics of any particular step. For example, by finding the equilibrium positions and tilts of different anionophores in their unbound states, trends of transport rates were rationalised in the study we discuss next.

In 2016, Edwards *et al.* used a family of decalin-based bis-urea transporters with variable *n*-alkyl "tails", in order to systematically probe how increasing the

size of lipophilic groups affects transporter behaviour in the membrane.[73] They conducted a total of 200 ns unbiased POPC simulations, as well as a number of pulling simulations across the membrane. In the unbiased simulations, they extracted the equilibrium depth and tilt of the free carriers and monitored headgroup contacts and spontaneous "flip" attempts across the midplane.[73] They observed a bell-shaped dependence of experimental $Cl^-$ transport rates on tail length, which predicted an optimal alkyl substituent length of 6 carbon atoms. This was related to two opposing mechanistic tendencies using MD simulations. With increasing tail length, the transporters' binding groups moved further away from the headgroup region due to stronger anchoring in the lipophilic tail region of the membrane. This meant that anion binding competed less intensely with negatively charged headgroup residues for access to the binding site, improving transport rates. On the other hand, the barrier to rotation of the transporter in the membrane appeared to increase with increasing tail length, during the membrane translocation step, becoming a hindrance at large alkyl groups. Together, this meant that an intermediate chain length was the optimum for transport activity.

Marques *et al*. studied squaramide carriers using umbrella sampling simulations to obtain free energy profiles for the translocation of both unbound and $Cl^-$-bound transporters along the bilayer normal.[72] The resulting PMFs showed a substantially higher barrier for complex cross-over than for carrier recrossing (Figure 1.14, middle row), identifying the core translocation of the complex as the rate-determining step and therefore kinetic bottleneck of the transport process rather than the return of the empty carrier through the POPC membrane. Translocation being the r.d.s. explains why squaramides with very strong interfacial binding may nevertheless be poor transporters. They also showed that the barrier of crossing is much lower for the transporter-associated $Cl^-$ ion than the naked anion by itself, which was the first literature example of a theoretical study quantifying the

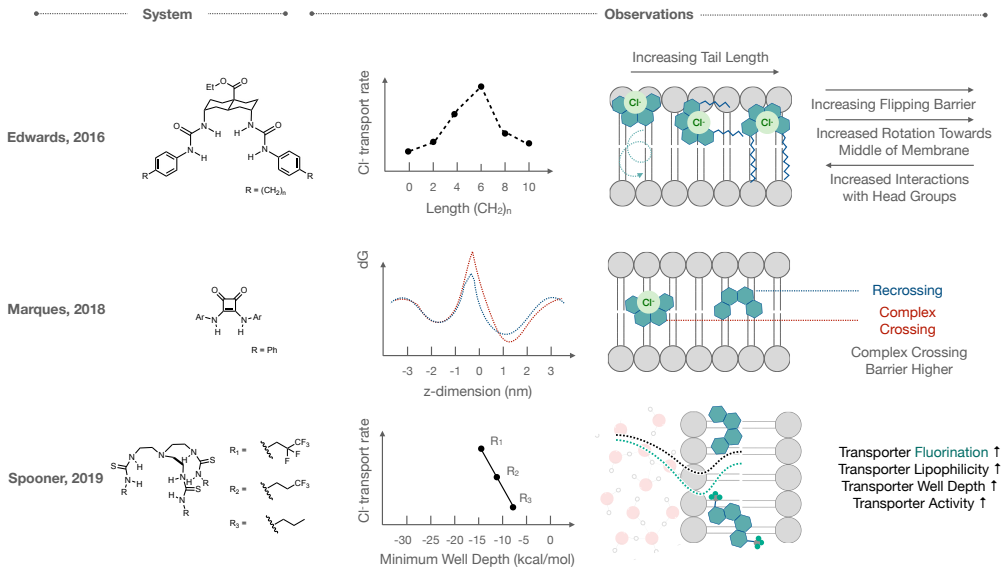



**Figure 1.14: Key molecular dynamics studies of membrane transporters, displaying schematics of the systems under investigation, the observable of interest (experimental or computational), and the observations resulting from the studies.** Top: A study by Edwards *et al.* on decalin-based ureas, which used unbiased molecular dynamics to relate alkyl tail length to $Cl^-$ transport rate through molecular dynamics observables.[73] Middle: Marques *et al.* study on the barriers of squaramide transporter membrane crossing using umbrella sampling. [72] Bottom: Spooner *et al.* study on the effect of fluorination on the transport rate, looking at the energy minima of umbrella sampling energy profiles. [71]

lowering of the membrane-crossing barrier for an ion by a transporter. Within their transporter family, they found that when the interfacial energy minimum well of the [transporter·Cl]$^-$ complex is very deep, then the barrier for reaching the midplane of the membrane rises, lowering the rate of transport. This study also highlighted many other observables which can be used to rationalise trends of transport. For example, the number of relevant H-bonds between the transporter and the headgroup phosphate and choline moieties, and its H-bonds with water. These interactions show a delicate balance, and the findings additionally indicated that a significant number of water molecules is transported across the membrane core alongside the anion.

The finding that the translocation step is rate-determining was followed up with a systematic study of how fluorination of tren-based thiourea carrier affects the energetics of this step. Spooner *et al.* found that increasing aryl fluorination with fluorine and trifluoromethyl groups raised lipophilicity and deepened the interfacial minimum well of the [transporter·Cl]$^-$ complex PMF profile relative to its energy in bulk water. Alongside this deepening of the minimum well, a lowering of the barrier for translocation across the membrane core was also observed, which correlated with higher $Cl^-$ transport rates across the series (Figure 1.14, bottom row). This suggested that in this transporter family's chemical space, a deeper interfacial well improved transport, but the study's results, when coupled with previous work by Edwards *et al.* and Marques *et al.* seems to suggests that excessive well depth could invert this trend by trapping the complex at the interface. Interestingly, they also examined the maximum electrostatic surface potential $V_{S,max}$ of the various transporters, and found that there was little to no correlation between the potential values and the binding constants for the chloride anion. They concluded that there are other very important factors besides Lewis acidity, such as specific interactions with solvents and conformational effects which have an important effect on binding constants, which cannot be explained by $V_{S,max}$ calculations and require dynamic treatment.

## 1.4 Stimuli-Responsive Transporters

Within the field of stimuli-responsive systems, anionophores have emerged as a specialised and highly promising class of therapeutics.[80–82] The movement of ions across biological membranes in nature is typically governed by membrane-bound proteins, which have the innate ability to function within the maze of biological signalling and respond to the needs of the cell through the ubiquitous

chemical messages that are dispatched within it.[1] With the ability to mimic this natural responsive functionality with synthetic carriers, one could theoretically address a diseased area of the body with precise control.

Within therapeutic applications, stimuli-responsive ionophores also directly address the issue of traditional carriers lacking specificity, which can lead to toxicity.[4] Precise activation of transport thus becomes particularly powerful in situations where disease states have a unique chemical signature, such as the low pH and altered redox conditions found in tumors.[83, 84] Strategic design thus allows for the targeted disruption of cellular processes critical for a disease, by making the transporter activate only in diseased tissues. This transforms them into precise, bio-mimetic therapeutic strategies.

Beyond their direct therapeutic use, stimuli-responsive systems are highly valuable as tools for fundamental research. They allow for the spatio-temporal control of ion transport and signal transduction, which can be used to study ion transport phenomena, engineer functional membranes, create complex signalling networks within artificial cells, or open up new avenues for controlled in-cell synthesis[81, 85–88].

The most common types of stimuli used to gate the activity of ion carriers are light, pH changes and redox manipulation, while others such as temperature and membrane potential have also been employed.[44, 80, 81, 89–95] Activation is generally grouped into reversible and irreversible mechanisms - with the former usually utilising molecular switches, whereas the latter employs caging groups, which are removed from the transporting molecule. Reversibility is perhaps most easily achieved using light, where photoswitches have successfully been developed in order to switch on and off the transport of ions in the Langton group.[87, 89, 96, 97] The use of light to develop therapeutic molecules falls into the broader area of

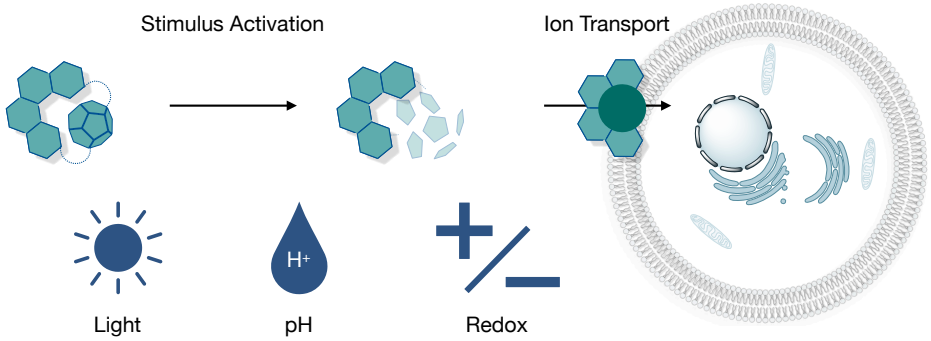



**Figure 1.15: Irreversible activation of an ionophore**. By using an external stimulus, be it light, pH changes or redox chemistry, a transporter can be activated. Upon the activation of the transporter, here shown as the removal of a blocking and solubilising group to the binding site, the active species can now partition into the membrane and is thus capable of carrying out transport (e.g. in a cell). The symbols for light, pH and redox activation used here are also used in later schemes.

photopharmacology, and is particularly attractive because low-energy near-IR light can penetrate deep into the body and is bio-orthogonal to cellular processes, which results in minimised risk from using it as an activating stimulus for transport.[98] Kerckhoffs *et al.* reported near-IR light-activatable azoarene photoswitch **1.18**, where highly enriched photostationary states of the *Z*-isomer transporter resulted in large increases in chloride transport activity relative to the *E*-isomer, with up to 10-fold increases in the $EC_{50}$ value.[89] This was based on the differences in proximities of the two squaramide anion-binding domains in the *E*- and *Z*-isomer. In the *Z*-isomer, both squaramides could bind to a single chloride ion simultaneously, increasing the binding affinity, while in the *E*-isomer, their greater intramolecular distance made this impossible, resulting in reduced affinity and transport activity.

Busschaert *et al.* reported transporter **1.19** in 2014, which utilised the high acidity of amide nitrogens on thiosquaramide groups in order to reversibly control $Cl^-$ transport (Figure 1.16b).[99] At physiological pH, the squaramide became depro-

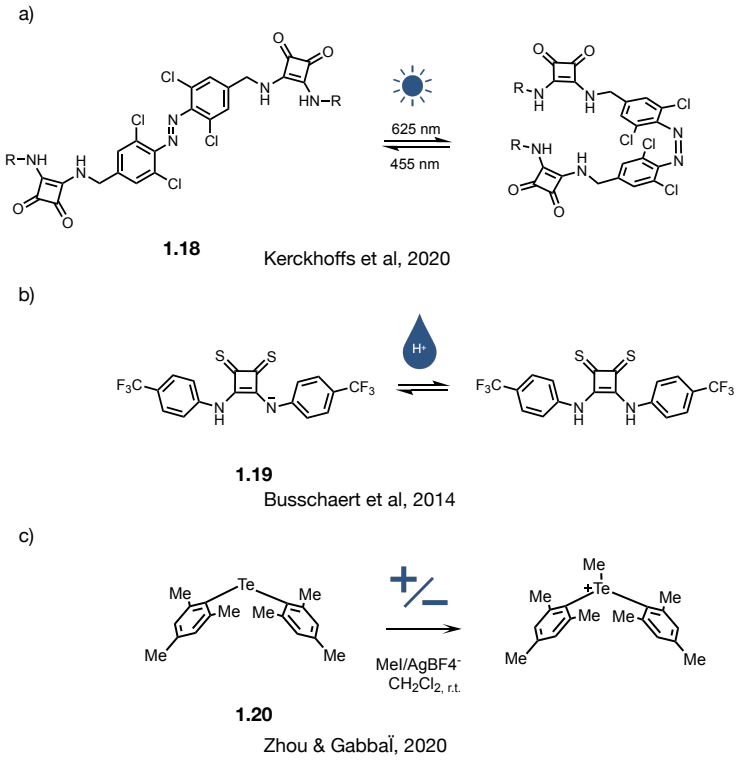



**Figure 1.16: Examples of stimuli-responsive chloride transporters**. a) Work employing bi-directional photowsitching in azobenzene **1.18** to switch transport on and off reversibly.[89] b) Reversible deprotonation of thiosquaramide **1.19**, allowing for reversible control.[99] c) Redox-activated, irreversible transport activation in **1.20**.[93]

tonated, which led to an inability to bind chloride, whereas in acidic environments, the ability to bind and transport the chloride was regained with full reversibility. An example of irreversible redox switching comes from Zhou and Gabbaï, who reported tellurium-based transporter **1.20**, which could be irreversibly methylated in a process promoted by silver salt formation from methyl iodide. This gave a positively charged tellurium salt, whose sigma hole was had a highly positive electrostatic potential, and allowed for chloride binding and transport (Figure 1.16c).[93]

Besides having good transport properties, stimuli-responsive transporters must also fulfil several design criteria related to their activation, no matter the activation

mode.[80] Foremost, they must demonstrate biocompatibility, ensuring minimal toxicity and immunogenicity for potential biological applications (of all of the inactive, active, and waste protective group species), and sufficient chemical stability under physiological conditions. The kinetics of activation are also critical, as the response must be rapid enough to be therapeutically effective but not so fast that it becomes uncontrollable and essentially negates the protection, thereby allowing for precise spatio-temporal control. This is particularly important for targeted drug delivery or localised tissue engineering.

A high degree of completeness of activation is necessary to achieve a significant transport flux when triggered, which is reflected in a high "OFF-to-ON" ratio to prevent unintended ion leakage or cargo release in the absence of the stimulus. Additionally, the transporter must exhibit high selectivity for its intended target ion to avoid disrupting the delicate homeostatic balance of other ions.[4] Finally, for advanced applications, the incorporation of multi-stimuli responsiveness can provide an orthogonal layer of control, requiring the presence of two or more independent triggers for activation, thus enhancing the system's specificity and safety.

Systems that mimic other biological processes, such as intercellular signalling, have also been developed. The system reported by Gartland *et al.* emulates signaling by utilizing a two-part process (Figure 1.17). It functions by setting up a signalling network between vesicles, where a signal can be transmitted from a small population of "sender" vesicles to a larger population of "receiver" vesicles. In the synthetic system, the caged zinc ionophore is first pre-incorporated within the membrane of the sender vesicles. The light activation signal then triggers the cleavage of the photo-cage, releasing the active ionophore species. The released ionophore then acts as a travelling chemical messenger, diffusing through the aqueous medium to the receiver vesicles. Upon reaching the receiver vesicles,

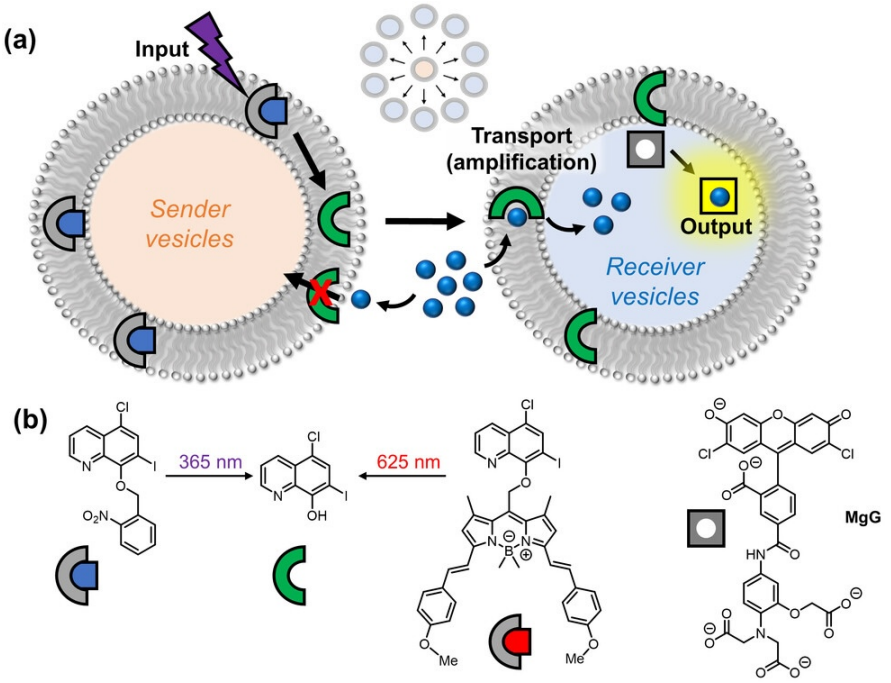



**Figure 1.17:** a)A light-sensitive caged cationophore embedded in a small population of sender vesicles. Upon photo-irradiation, the cationophore is released and diffuses into the surrounding solution. The liberated transporter then travels to a larger population of receiver vesicles, where it initiates transmembrane cation transport. This catalytic process leads to signal amplification, which is detected by a fluorescent output. b) Shows the molecular structures of the zinc transporter, the two photo-caged cationophores, and the zinc ion sensor MgG. Reproduced with journal permissions.[100]

the ionophore embeds itself in their membranes and catalytically facilitates the transmembrane transport of zinc ions. The system's design ensures that ion transport is reported only in the receiver vesicles and not in the senders. Since each ionophore molecule can catalyze the transport of many ions, the initial signal from a small number of sender vesicles is greatly amplified in the larger population of receivers. This catalytic process results in a significant signal amplification of up to 30,000-fold relative to the sender vesicles.

### 1.4.1 Computational Studies of Stimuli-Responsive Transporters

Both DFT and MD have been used to study the mechanisms of stimuli-responsive transporter activation and action. Very often, the active states of photoswitchable systems are transient, and therefore their geometries cannot be captured using analytical techniques. The use of DFT can therefore be very helpful in modelling their geometries and guiding further development to improve the activation and deactivation properties. Perhaps most ubiquitously, such studies of activation have been performed for photo-switchable transporters, which we will briefly discuss.

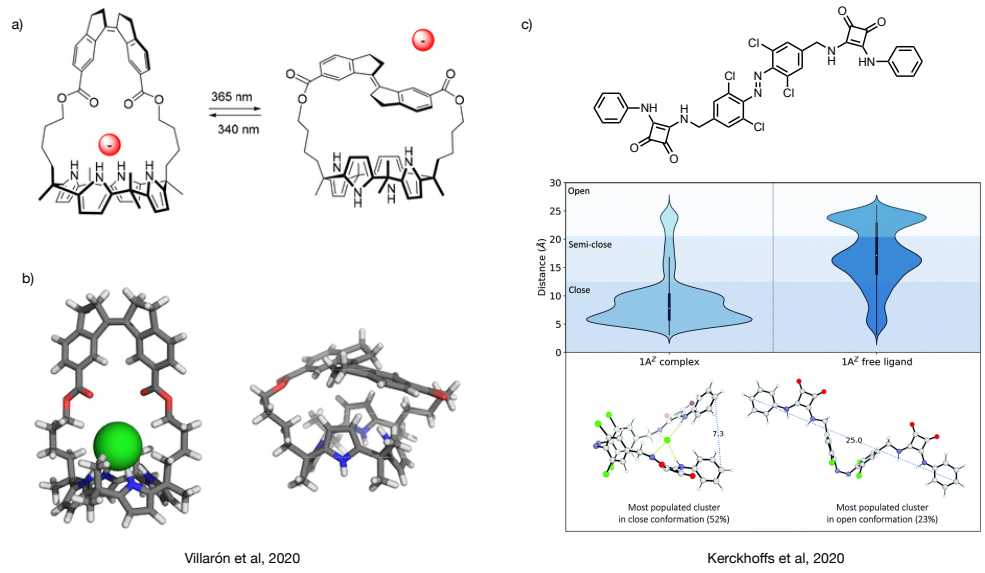



**Figure 1.18: Examples of stimuli-responsive transporter computational study**. a) Light-promoted switching between the *Z*- and *E*- stilbene calix[4]pyrrole-based photo-switches developed by Villarón *et al*. b) The corresponding 3D geometries, from DFT optimistions (left) and X-ray crystallography (right) of the same two isomers. Adapted with permission from the journal[101] c) The distribution of the two anion binding domains of the azobenzene-based chloride transporter developed by Kerckhoffs *et al*. from MD simulations in the *Z*- and *E*- isomers. Adapted with permission from the journal.[89]

In a study on a strapped calix[4]pyrrole-based transporter, Villarón *et al*. used DFT to explain how the isomerisation of a stilbene moiety in the strap surrounding the binding cavity was able to promote the binding of a chloride ion.[101] They

were able to determine that the *Z*- isomer had an 8000-fold higher binding constant than the *E*- by NMR titration experiments, which the DFT optimisations were able to explain with a cone-shaped geometry that was only accessible following light-promoted isomerisation from the *E*- to the *Z*- isomer. Indeed, in the *E*- isomer, this cone-shaped geometry was found to be energetically out of reach, and the actual ground-state conformation in this isomer lead to the exclusion of the ion.

In the context of stimuli-responsive anionophores, MD is a vital tool for understanding how the now-active molecule interacts with and moves in the membrane. These simulations can investigate atomic-scale details of the differences between the active and inactive transporter's interactions with the anion, water, and the membrane lipids in a dynamic fashion.[89] For example, Kerckhoffs and colleagues demonstrated that the *Z*- and *E*- isomers of their azobenzene transporter **1.18** had very different distributions of the end-to-end distances between the two anion-binding groups both in solvent and in the membrane (Figure 1.18b). They were also able to show that the *E*- isomer was more prone to dimerisation in the membrane core, which served as a possible explanation for its greater reduction in binding activity relative to the *Z*- than would have been expected from the binding constants alone.

## 1.5 Importance of Binding Prediction In Solution

So far, we have discussed supramolecular assemblies only in the context of membrane transport. IUPAC defines supramolecular assemblies as complexes formed of two or more molecules, which are held together by non-covalent interactions, which can include electrostatic forces, dispersive attractions and solvent-displacement effects.[102] In the transport process, the important assemblies are thus: the lipid bilayer membrane, transporter-ion complexes, stacking and

hydrogen-bonding assemblies forming pores and channels, etc. Their variety alone makes computational modelling challenging. Indeed, as mentioned in section 1.3.1, the membrane environment makes modelling the energetics of processes within it to a quantitative level almost impossible. Here, we introduce a prediction problem in slightly less convoluted systems - host-guest complexes in solution. These have great value in both method development and applications alike.

Host-guest complexes are defined by a host molecule, which contains a binding cavity or site, in which another, usually smaller, molecule (the guest) can bind.[103] Supramolecular hosts can have a variety of different applications, which require different levels of affinity and specificity for their guests. Besides the membrane transport[4] discussed at length in this thesis, anion-binding supramolecular hosts may be used to sense ions in aqueous media at very low concentrations[104, 105], which can be used to measure pollution. They can also be used to capture pollutants irreversibly and remediate the environment[106], can form part of responsive materials[107], act as catalysts in reactions with anionic substrates and products[108], and have even been in development as drug delivery vehicles[109]. For all of these applications, the affinity of the host for the guest is one of the most important properties, and its prediction is one of the grand challenges of computational supramolecular chemistry.

The discovery of novel supramolecular hosts remains a formidable experimental challenge. Synthetic methods and binding characterisation are time-consuming, laborious and material-intensive. This is especially the case with large supramolecular compounds, which may suffer from poor solubility, low synthetic yields and long synthetic pathways, which all limit easy access. Binding of large, multivalent anions, which have very high solvation energies and pre-defined geometries especially require intelligent design. By accurately predicting the binding constants of novel host molecules, a lot of trial-and-error experimentation could be

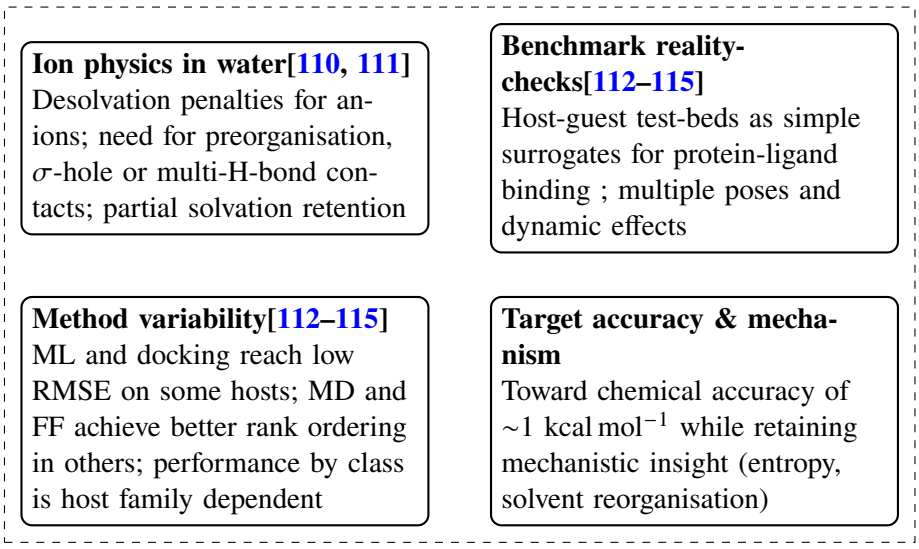



**Figure 1.19: Overview of the main motivations and challenges in predictive host-guest binding.** In clockwise order from top left: the intrinsic difficulty of binding anions in water, the role of benchmark host-guest systems as surrogates for larger biomolecular complexes, the variability in performance across computational approaches (ML, docking, MD, force fields), and the goal of achieving chemical accuracy (∼1 kcal mol$^{-1}$) while retaining mechanistic insight into entropic and solvent-reorganisation contributions.

spared, leading to great increases in the efficiency of the discovery process for novel supramolecular hosts. Additionally, supramolecular molecules with relatively pre-defined binding sites are a great surrogate for the study of larger, more conformationally complicated systems, such as proteins. This means that developing better methodologies for the study of supramolecular binding events in solution can have direct effects on medicinal chemistry.

Predicting binding remains a stringent test of modelling because it couples subtle non-covalent forces with solvent reorganisation, ion effects, and long-range electrostatics. In order to achieve this goal, shared between supramolecular and medicinal chemistry, the SAMPL (statistical assessment of the modelling of proteins and ligands) challenges, funded by the NIH, have been taken through to their 9th iteration by the Mobley group in UC Irvine.[112–115] One of the aims of the challenges is to improve the methods by which the free energies of binding

($\Delta G$) can be predicted for the binding of drug-like compounds with supramolecular hosts. Other aspects include the prediction of physicochemical properties of interest - p$K$a, log$D$, permeability - which relate to the interactions of molecules with solvents and membranes, as well as their ionisability. In order to predict all of these properties, accurate modelling of supramolecular interactions is key, including the correct treatement of charged species and proper consideration of solvation that are key to binding.

The most recent SAMPL challenge reported a machine-learning approach based on molecular descriptors to have the highest accuracy, followed by docking approaches and molecular dynamics.[115] Interestingly, there is a noticeable lack of ab-initio methods used in these systems due to their large size, and the few examples of semi-empirical and ab-initio QM-based calculations present in SAMPL challenges 6 to 8, were reported to be outperformed by docking and machine learning approaches. While each of these methodologies has its drawbacks, the main drawback of docking, and especially big data machine learning approaches, is that they do not give much insight into the mechanisms of binding. Mechanistic understanding is key for elucidating entropic effects, solvent reorganisation and other factors which may aid in the future improvement of host-guest binding. The ideal methodology would thus be able to reproduce binding energies close to the so-called chemical accuracy limit of 1 kcal·mol$^{-1}$, as well as shedding light on the mechanism of binding - combining accurate dynamics and ab-initio accuracy.

### 1.5.1 Existing Methods for Modelling of Halogen Bonding

Halogen bonding is a particularly complex non-covalent interaction to model computationally. Compared to hydrogen bonding, it exhibits an even higher degree of directionality, while also containing significant dispersive and charge-transfer contributions.[57, 58] Combined experimental and computational studies have further

demonstrated that the strength and geometry of halogen bonds depend sensitively on the surrounding medium, with solvent effects deviating from simple electrostatic trends established for hydrogen bonds.[116] Accurate treatment of solvation is therefore a key requirement for the reliable modelling of halogen bonds in solution (Figure 1.20a). Importantly, halogen bonding association is often treated differently in electronic-structure approaches and in molecular mechanics simulations: DFT studies can capture the anisotropic electrostatics and charge-transfer components, whereas classical MD force fields typically approximate halogen bonding via carefully-tuned explicit "extra point" charge models.[117, 118] These methods, including their respective strengths and weaknesses in studying halogen bonding to ions in solution, will be briefly introduced below.

In DFT studies, an implicit solvation model may be used to account for solvent effects. It has been shown that changing the DFT functional used for a chemical process like anion binding can lead to significant variations in the calculated solvation contributions.[119] This is because both the functional and the solvation model introduce their own approximations, and the combination of these choices can dramatically impact the description of charge, electronic structure, and solute-solvent interactions (Figure 1.20a). Different functionals have varying strengths and weaknesses in describing non-covalent interactions, such as hydrogen bonding or halogen bonding and dispersion forces, which are crucial for host-guest binding.[56, 120] Additionally, the functional's ability to accurately describe the electronic charge distribution of the anion and the host influences the strength of the electrostatic interactions with the implicit solvent continuum. Different implicit models (such as COSMO, CPCM, and SMD) have distinct ways of defining the solute-solvent interface, calculating the electrostatic and non-polar contributions to the solvation energy, and handling charge density.[121] DFT studies of the binding of anionic guest molecules to supramolecular hosts are not particularly

ubiquitous, which speaks volumes about the magnitude of the challenge. In a study by Houk, featuring hydrogen-bonding host molecules, it was found that the mean absolute deviation (MAD) of the binding energy with respect to the experimental value was 2.5 kcal·mol$^{-1}$.[122] They found that this error was similar using both CPCM or SMD solvation models.

In molecular dynamics simulations, the much lower computational cost of energy and force evaluations allows for the inclusion of a much larger number of molecules, which unlocks the ability to simulate systems in their completely solvated state. Besides the improvement in the description of solvent, there is also a much better treatment of entropy, as the system can explore multiple conformational states over time, and the use of long simulation times should yield a good approximation of the entire ensemble, which allows for much more accurate determination of free energies using appropriate methods, such as umbrella sampling or well-tempered metadynamics. Blind challenge benchmarks seem to agree that MD is a better predictive tool in this case. In SAMPL5, an attach–pull–release (APR) physical-pathway protocol achieved good agreement for CB7/$\beta$CD host–guests, recovering absolute free energies and enthalpies across 12 systems.[123] SAMPL8/9 show that method performance is host-family dependent. Alchemical MD excelled for some sets, while ML or docking won out in others, underscoring that sampling, force-field, and protonation-state choices still govern outcomes in dynamics simulations.[114, 115]

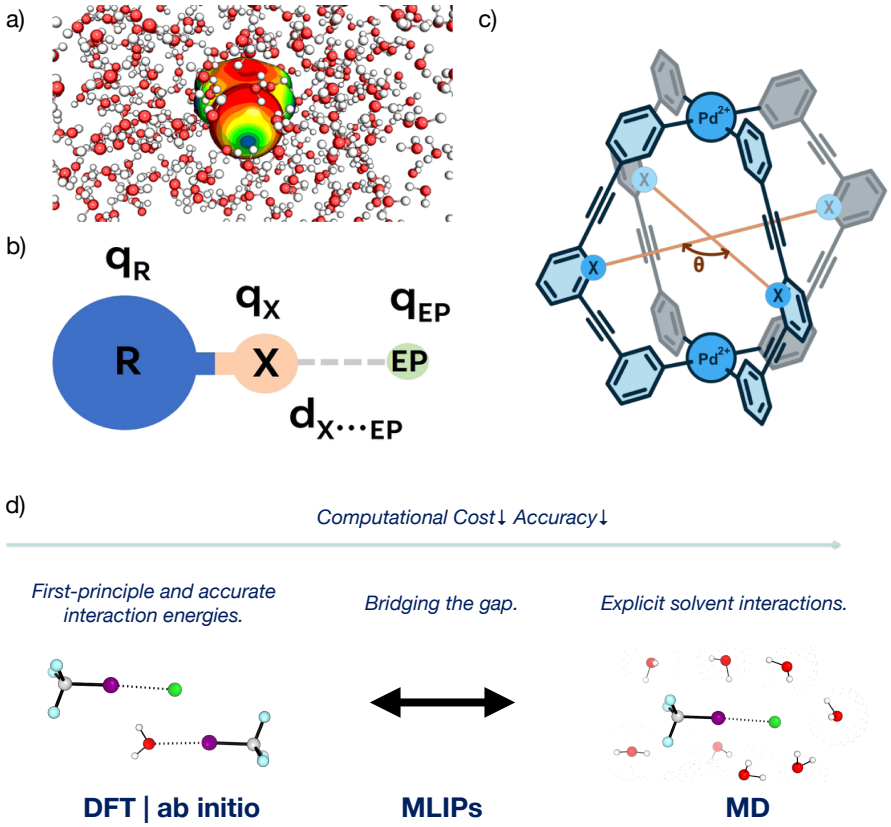



**Figure 1.20:** a) Scheme showing the electrostatic surface potential of an iodobenzene molecule, highlighting the importance of rigorous treatment of charge anisotropy to capture solvent effects.[124] b) Schematic of extra point charges introduced for MD simulations, which allow for charge anisotropy.[124] c) Recent application of machine-learning interatomic potentials (MLIPs) for supramolecular applications, studying palladium cages for drug encapsulation and delivery.[125] d) Schematic depiction of how MLPs are able to bridge the accuracy of ab-initio calculations with a cost that is much closer to classical dynamics simulations. a) and b) reproduced with permission from [124], c) reproduced with permission from [125]

On the other hand, classical force fields struggle in their out-of-the-box parameters with halogen bonds, because point-charge models cannot reproduce the anisotropic electrostatics of the $\sigma$-hole.[117, 118, 124] Standard RESP fitting often assigns halogen atoms either an overall negative charge (eliminating any possibility of halogen bonding) or only a weakly positive charge, underestimating both strength and directionality. To address this, many force fields now introduce off-center "extra particle" (EP) sites: massless charges bonded to the halogen along the C–X

axis to mimic the $\sigma$-hole (Figure 1.20b).

### 1.5.2 MLIPs

Machine learning interatomic potentials arose out of a growing need to simulate ever larger systems with increasing accuracy. They were developed to bridge the gap between the computationally demanding but high-accuracy DFT methods and the much faster but less accurate classical force fields (Figure 1.20d). By learning the DFT potential energy surface using a machine learning algorithm, it can be reproduced, in theory, to a limit of DFT accuracy at a fraction of the cost. This is because the mathematical form of the machine learning algorithm is usually much simpler than the self-consistent field equations that would need to be solved at every step of a DFT-level *ab initio* dynamics simulation. The use of MLIPs thus unlocks previously unavailable time- and length-scales for simulations of very high accuracy.

For instance, MLIPs are faster than DFT calculations on the same system, with evaluation time speed-ups of 3-5 orders of magnitude.[126]. On the other hand, these MLIPs are still much slower than classical force fields, with evaluation time differences usually around 3-4 orders of magnitude, depending on system size as well as MLIP and force-field choice (See Figure C.7 for comparisons of MLIP and classical MD evaluation steps). This is due to the fact that classical force fields have a very simple functional form and highly optimised algorithms in software packages which have been actively optimised over a period of decades. MLIPs, on the other hand, consist of more complicated mathematical functions, be it neural networks, linear regression, or kernel regression methods, which take a longer time to evaluate, and have been in development for a much shorter time. They also construct much more complicated descriptors of the atomic environment, which add to the computational cost (for in-depth discussions of MLIPs, see

the Theory section in the Theory Appendix Section C.3). It was observed that systems with around 2000 atoms are able to run in excess of 1 µs per day using an AMBER force field, while they struggle to get to even a single ns using the MACE GNN architecture MLIP, even with the most computationally affordable model hyperparameters. Nevertheless, the ability of MLIPs to model bond-forming and -breaking phenomena means that they have become increasingly prominent.

While their most successful early applications were in periodic, solid-state systems, which were geared towards materials chemistry applications, MLIPs have recently begun to expand beyond this focus.[127] They are now being used to study the properties of complex systems like organic crystals for drug discovery[128], to model the behaviour of organic molecules in solution[129, 130], to design new materials for applications such as organic redox flow batteries[131], as well as to develop new tools for TS finding [132] and modelling accurate bond dissociation profiles[133]. This expansion has been driven by the development of new models and training strategies that can better handle the diverse and often complex intermolecular interactions present in organic systems. Applications to supramolecular systems have only very recently begun to be seen, with work by Stebani *et al.* on the study of supramolecular cages (Figure 1.20c).[125]

#### 1.5.2.1 Sampling approaches

In the language of machine learning, the training process of MLIPs is called supervised learning - which means that the target values which the model needs to reproduce are known. These target values are the energies and forces (which are, in other words, the derivatives of the energies). The goal of the machine learning is to thus learn a function $f$, which maps the coordinates of the system $R$ to an energy $E$ with some function parameters $\alpha$, which are optimised during the

training process:

$$E(R) = f(R; \alpha)$$

In order for the machine learning algorithm to learn the ideal parameters for reproducing the energy surface $E$, the training data needs to cover the phase space of the simulation with sufficient density. Ideally, the data points are:

- **Diverse**: The training data must contain a wide enough variety of atomic configurations and chemical environments to ensure that the model can generalise to unseen structures that it will encounter during the simulation. This is particularly important for structures that undergo significant structural changes.
- **Realistic**: Training points should not stray too far from what the system will experience during a simulation, while still including both high- and low-energy structures.
- **Accurate**: The energies and force labels of the data must be of high quality. If the DFT reference method is inaccurate with respect to higher level reference methods or experiment, this represents an irreducible error, which the model cannot systematically overcome by reproducing the reference method itself more closely.

In early materials applications, the common way to gather training data for Machine learning interatomic potentials was to create a static training dataset. The datasets were often constructed iteratively by taking crystal structures of materials in different representative phases, performing high-temperature *ab initio* dynamics on them to gather more training points, and then using the combined training points to train an MLIP, with which more training data points could be generated.[134]

This procedure was later elaborated into algorithms that performed automated

loops of the same procedure, of which a schematic depiction can be seen in Figure 1.21. If not many configurations of the system exist, then one may take a minimum energy structure and perturb it, by randomly displacing some of the atoms. If previous active learning loops of similar systems have already been carried out, then one can leverage datasets found in those as a starting point.[129, 130, 135–137] Once the model has been trained on this data set, a dynamics simulation using the trained model is conducted. This simulation is performed until some selection criteria are fulfilled. This is where the efficiency of active learning algorithms lies. This so-called 'oracle', which selects the data that is best suited to improve the training set ensures that only the most relevant data points are added. By selecting the oracle carefully the criteria for a healthy and balance data set that were outlined above can be satisfied. This may be by selecting outliers in terms of energy - for example if a high energy configuration is encountered for the first time, the selector may get triggered, and this configuration added to the training set. Because this requires every configuration that is evaluated (usually a set fraction of configurations encountered in the dynamics trajectory) to be subjected to expensive DFT reference method calculations, other criteria have been developed. For example, Zhang *et al*. used geometry-based criteria to determine outliers (Figure 1.21, IV), which take only a fraction of the time to evaluate geometries, and show much better data efficiency.[130] Finally, if no new configurations are selected to be added to the data set within a certain time-frame, and the simulation remains stable throughout, then the MLIP is considered to be stable, and can be used for running production simulations.

#### 1.5.2.2 MACE and Foundation Models

MACE is a MLIP architecture based on graph neural networks (GNNs), which gives it several benefits in representing chemical systems.[138] GNNs are well

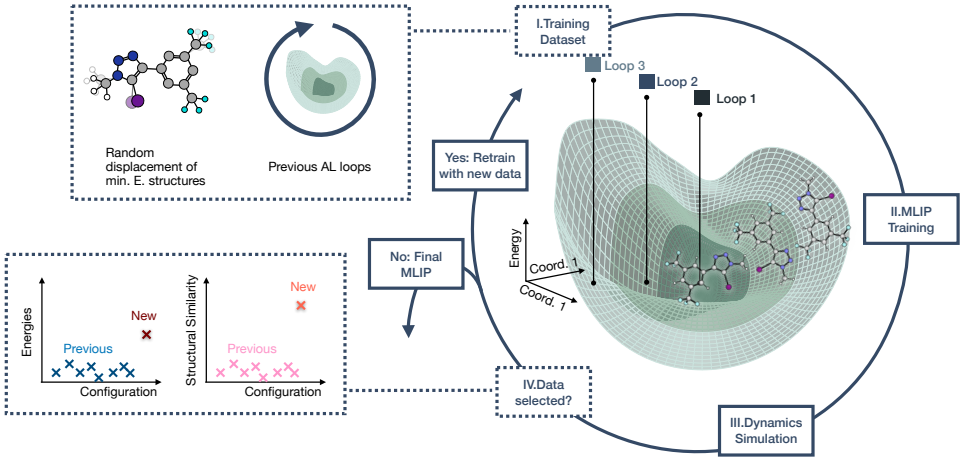



**Figure 1.21:** Showing the steps of the active learning process. I: The active learning loop begins with a data set, which can contain a small amount of randomly displaced structures, or a curated dataset already obtained through active learning. II. The MLIP is trained on this dataset. III. Dynamics simulations are run until some criteria are met by which a new geometry from the dynamics trajectory should be added to the training dataset. IV. In the case of mlp-train these criteria are either energy- or structure-based. If no new structures are added, the last iteration of the MLIP training has produced the final potential, otherwise new data is added to the training set and the loop starts again.

suited for predicting molecular properties because molecules are very efficiently represented as graphs - with the atoms acting as nodes, and the bonds between them acting as edges. GNNs operate through passing information between graph neighbours through their edges, and the relevant properties of atoms necessary for simulating them (energies and forces) directly depend on their neighbours, which makes these architectures well-suited for learning potential energy surfaces. MACE, building on the equivariant descriptors in the polynomial regression-based atomic cluster expansion (ACE) architecture, has an incredibly flexible form, where the descriptor embeddings are continuously optimised during training.

MACE has exhibited exceptional accuracy on a wide range of systems, from materials to small organic compounds, and even peptides.[139] Its equivariance means that it offers robust predictions for diverse chemical environments. It also

scales reasonably well in computational cost with the size of the system. Its great success has also been accompanied by the development of foundation models based on the architecture. These are models trained on massive and diverse databases of molecular structures. For example, the MACE MP0 model, which was trained on the entirety of the materials project database, and MACE-OFF, which was trained on the SPICE dataset containing some 130,000 compounds.[139, 140] The aim of these foundation models is to serve as a surrogate for a high-level ab-initio molecular dynamics (AIMD) simulation in the initial sampling phase. This allows a user to quickly simulate the system with a high degree of geometric accuracy, in order to gather configurations which can then be evaluated using a DFT method. Furthermore, the foundation models can be fine-tuned, which means that training starts from the model weights of the pre-trained foundation model (in so-called transfer learning), which results in much quicker training on the desired dataset, and has been shown to lead to much more stable ML models.[141, 142] For organic systems, the MACE-OFF models allow the user to simulate any organic system containing the elements H, C, N, O, P, S, F, Cl, Br, I, which greatly accelerates the development of initial, stable models for a system. These developments have quickly enabled the simulation of increasingly complicated organic molecular systems.

## 1.6 Outlook

In this chapter, the field of ion transport was introduced alongside the development of synthetic ion transporters and their application to biology. Ion transporters, and chloride transporters more specifically, represent a crucial area of disease physiology. Many synthetic transporters have been developed, and design criteria for their successful applications have been carefully elucidated in literature. Among these,

selectivity for chloride ion transport over hydroxide has been deemed especially important for biological applications. While extensive computational simulations have been carried out to justify activity trends, quantitative prediction of ion binding strengths for transporters remains a challenge. Furthermore, little work has been carried out on the modelling of selectivity. An investigation into the origins of the selectivity-defining principle of encapsulation appeared warranted, and led to the joint experimental-computational approach described in the following two chapters, followed by work on quantitative modelling of chloride binding in solution.

# 2 |

# A Novel Class of Chloride-selective Transporters

The design, synthesis, and assay characterisation of a wholly novel class of macrocyclic, chloride-selective membrane transporters are discussed in this chapter. By tracing the origins of selectivity to examples found in the literature (as discussed in Section 1.2.3), further proof of these foundational principles is offered, and thus computational study is enabled in Chapter 3. The synthesis of the transporters is convergent and modular, which should easily allow for further modifications of the electronic and physicochemical properties of the transporters in future (e.g. for property tuning - Section 1.10). These transporters are shown to be among the most selective developed to date. Their ability to be dosed externally makes them ideal candidates for biological study. This work was published in *Chemistry: A European Journal* in 2025.[54]

## 2.1 Transporter Design

A constant challenge in the application of anionophores as precise target therapeutics in channelopathy treatment has been their selectivity. For most H-bonding carriers, it is a fact that binding $Cl^-$ also enables unwanted $H^+$ or $OH^-$ transport, which destroys pH gradients key to the functioning of several organelles - including the lysosome and mitochondria - and provokes cytotoxicity. Mechanistic work from Gale and co-workers established a key principle for achieving selectivity already outlined in Section 1.2.3. By increasing the degree of anion encapsulation, higher $Cl^- > H^+/OH^-$ selectivity can be achieved. Secondly, increasing the

acidity of N–H donors correlates with more proton/hydroxide transport.[50] These relationships were extracted using HPTS-based assays configured to decouple any electrogenic $Cl^-$ transport from pH-dissipating processes, which helped to identify anionophore equivalents of valinomycin (an incredibly selective cationophore) which are able to move $Cl^-$ without carrying $H^+/OH^-$.[50] In subsequent developments, it was found that replacing hydrogen bonding moieties with halogen or chalcogen bonding groups also greatly increased selectivity. The discussion now turns to key literature examples that significantly informed the design strategy for the novel transporters explored in this chapter.

### 2.1.1 Selectivity-Enhancing Strategies

In two key publications from the Langton group, it was shown that replacing acidic N–H donors with halogen- or chalcogen-bond donors results in large increases in selectivity (Figure 2.1a).[52, 94] Not only does such an approach remove a potential handle for reversible protonation and thus proton shuttling, it also leverages the more directional, lipophilic $\sigma$-hole interactions for anion binding, which have also been shown computationally to reduce the propensity to transport $OH^-$.[55] More recently, Valkenier's iodotriazole-decorated calix[6]arene delivered fast $Cl^-$ transport with negligible $H^+/OH^-$ leak in HPTS assays, validating the non-protonophoric potential of halogen-bond donors.[53]

The Davis group developed the archetypal example of encapsulating transporters, progressing from cholapods (acyclic) to cholaphanes (macrocyclic), where closing the ring encloses the binding site (Figure 2.1b, left). In their initial works, they only demonstrated that closure increased transport rates at low loadings, which is consistent with the interpretation that a shielded and preorganized cavity improves the membrane positioning of the transporter, likely expediting membrane translocation and reducing unwanted participation of interfacial water. The latter

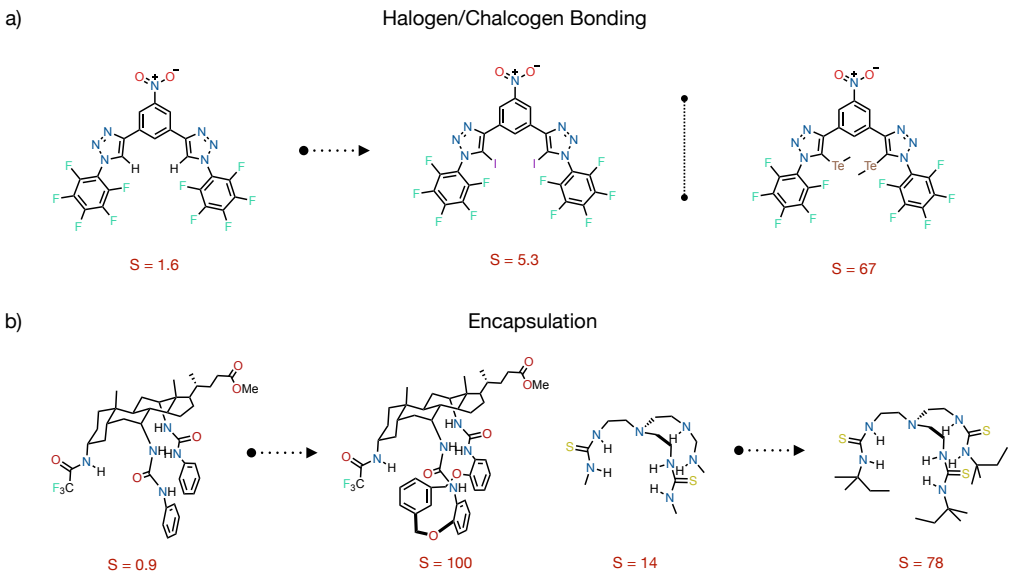



**Figure 2.1: Literature examples of design approaches to $Cl^-$ > $H^+/OH^-$ selectivity.** (a) Replacing acidic N–H donors with $\sigma$-hole donors increases directional Lewis acidity and decreases proton and hydroxide transport, giving higher selectivity values (From left to right: $S$ = 1.6 → 5.3 → 67). (b) By enclosing the anion-binding site (acyclic → macrocyclic), the transporter enhances desolvation of the anion and may shield the binding site from interfacial water. This converts non-selective systems ($S \approx 0.9$) into highly $Cl^-$-selective carriers ($S \approx 100$). Equally, increasing the steric bulk around the tren-based thiourea transporters increases the selectivity ($S \approx 14$ to $S \approx 78$) A larger $S$ indicates the multiple factor of non-protonophoric $Cl^-$ transport enhancement relative to $H^+/OH^-$.

is likely also a prerequisite for suppressing $OH^-$ transport, given the $OH^-$ ion's high hydrophilicity.[43, 143] While these studies chiefly focused on improving performance rather than $Cl^-/OH^-$ selectivity, they illustrated the principle that binding-site enclosure is beneficial for both efficacy and selectivity trends. It was only demonstrated in later work by the Gale group that these encapsulated variants also offered very large gains in selectivity.

Lastly, tren-based tris-(thio)ureas (Figure 2.1b, right) bring three flexible hydrogen bonding donor arms around a cavity that can partially enclose an ion. Structure–activity studies showed that activity scales primarily with lipophilicity rather than raw binding constants, while further selectivity assays crucially also showed

that greater encapsulation through bulkier terminal alkyl groups favours $Cl^-$ over $H^+/OH^-$ transport.[50, 144] All of these results point to encapsulation as another path to $Cl^- > H^+/OH^-$ discrimination, but the exact mechanisms through which this occurs remain under-explored.[50]

Curiously, the combination of encapsulation and halogen bonding had not been explored in literature, which begged the question of whether these two strategies can yield additive effects.

### 2.1.2 Balancing Activity–determining Parameters

The starting point for the transporter design was compound **2.1** reported by Bickerton *et al.*, which displayed excellent activity for chloride transport in assays with the protonophore gramicidin ($EC_{50} \approx 9$ nM), but only showed a moderate slowdown in assays where a protonophore was absent ($EC_{50} \approx 48$ nM).[55] The selectivity factor of 5.3 was far from the high selectivities observed by some of the most $Cl^-$-selective transporters.[50] While the increase in selectivity ($S$) of the novel macrocyclic transporter was desired, the retention of maximum activity was also sought. Since biological applications are the long-term goal of such $Cl^-$ ionophores, one of the objectivas was also to maintain the excellent deliverability properties of **2.1** in the new transporter family (Figure 2.2a). The parent compound could be delivered to vesicles *in situ* without pre-incorporation into the lipid. This indicates much better solubility and bodes well for cellular and in vivo applications. Here, it is discussed how the interplay of the various design parameters was juggled in order to come up with design variants, which were ultimately synthesised.

Halogen bonding (XB) at the iodine is strengthened by electron-withdrawing substituents on the iodotriazole moiety (Figure 2.2b). Electron withdrawal causes the

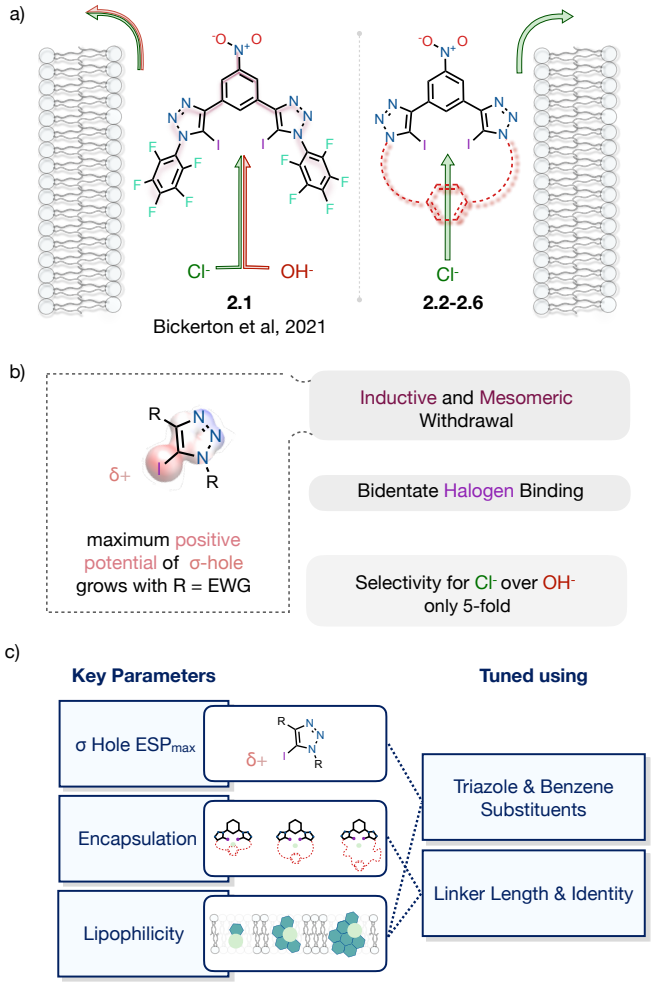



**Figure 2.2: Parameter optimisation in the macrocyclic halogen-bonding transporter family.** a) Parent compound and design starting point **2.1**, as well as the design evolutions **2.2–2.6**, highlighting the goal of achieving selective transport. b) Key features of the starting point molecule **2.1**. Electron-withdrawing aryl substituents make the $\sigma$-hole at iodine even more positive and strengthen bidentate halogen bonding, while the starting macrocycle **2.1** achieves only modest $Cl^-$ over $OH^-$ selectivity. c) Key parameters for the novel family, how they are tuned, and how they interact: $ESP_{max}$ of the $\sigma$-hole, encapsulation, and lipophilicity.

positive electrostatic potential ($\sigma$-hole) to increase 180° away from the C–I bond, improving the binding of negatively charged species.[58, 59, 145] Replacing the pentafluorophenyl substituents used in the acyclic parent compound **2.1** with less electron-withdrawing substituents necessary for tight binding site encapsulation was expected to reduce the $ESP_{max}$ at iodine and lower the intrinsic $Cl^-$ affinity and transport rate somewhat.[58] Encouragingly, encapsulation also seems to increase the activity in some cases, as demonstrated by the transformation of the trifluoroacetyl-substituted cholaphads to cholaphanes.[50]

A flexible linker was desired that could wrap around the binding site and shield the anion. A similar linker type to that employed by the Davis group in the construction of their cholaphane carriers was used, with a phenyl group offering a bulky shield to the binding site flanked by two alkyl arms (Figure 2.2a).[143] Designs where the phenyl group was replaced with a terephthalic acid derivative were also explored. These designs would allowd for modifications of the length of the linker to optimise parameters (Figure 2.2c). In attempting to make up for the expected loss in binding affinity due to $\sigma$-hole weakening, designs employing a fluorinated $\pi$-acidic group beneath the linker were also explored, so that a bound $Cl^-$ could engage in an additional anion–$\pi$ contact with the linker portion of the molecule as well. If successful, this would add a secondary stabilising interaction while preserving the anion size-dependent binding pathway in a compact binding site.[50, 108, 146]

**Table 2.1:** Consensus cLogP values (SwissADME) for macrocycles **2.1** - **2.6**.[147]

| Compound | **2.1** | **2.2** | **2.3** | **2.4** | **2.5** | **2.6** |
|---|---|---|---|---|---|---|
| Consensus cLogP | 7.1 | 4.8 | 5.4 | 6.0 | 6.6 | 6.6 |

Fluorination was expected to aid transport performance beyond binding. In tren-based transporters, increased fluorination enhances lipophilicity and was found to raise activity in vesicles and cells by improving bilayer crossing energetics.[29,

71] The beneficial effects of transporter fluorination have also been postulated to be a result of the unique solubility properties of perfluorinated groups, which reverse the regular polarity distribution of all-proton aliphatic and aromatic groups. For molecules **2.2.-2.6**, which were successfully synthesised and tested, the SwissADME consensus clogP values for **2.1**–**2.6** (Table 2.1) span 4.8–7.1.[147] Of these, compounds **2.3**–**2.6** (5.4–6.6) fall in the window which is generally accepted to be the optimum lipophilicity for membrane transport, whereas **2.1** (7.08) is nearer the upper end, where deliverability can become limiting.[147] Nevertheless **2.1**'s published activities were excellent, which suggested the novel macrocyclic family was likely in the appropriate range of lipophilicity.[29, 63, 64]

Finally, the *para*-$NO_2$ group was replaced with an ester to test three linked variables within the same topology. A slightly weaker $\sigma$-hole at iodine due to reduced mesomeric withdrawal would allow the probing of the trade-offs between XB strength and improvement in lipophilic balance around the binding cavity and potentially improved deliverability from aqueous media. Esters also offer a handle for later proanionophore strategies, which would allow for further development of the scaffold from the benzene moiety, as was the case in the photo-regulated anchored transporter reported previously by the Langton group.[74]

## 2.2 Synthesis of Macrocyclic Ionophores

Synthetic attempts towards macrocyclic variants were first initiated with aliphatic linkers, after which the synthesis of other linker variants was undertaken. The general disconnection strategies and successful attempts at the former will be discussed first, followed by a discussion of some of the unsuccessful attempts toward the latter.

### 2.2.1 Retro- and Forward Synthesis of Transporters with Aliphatic Linkers

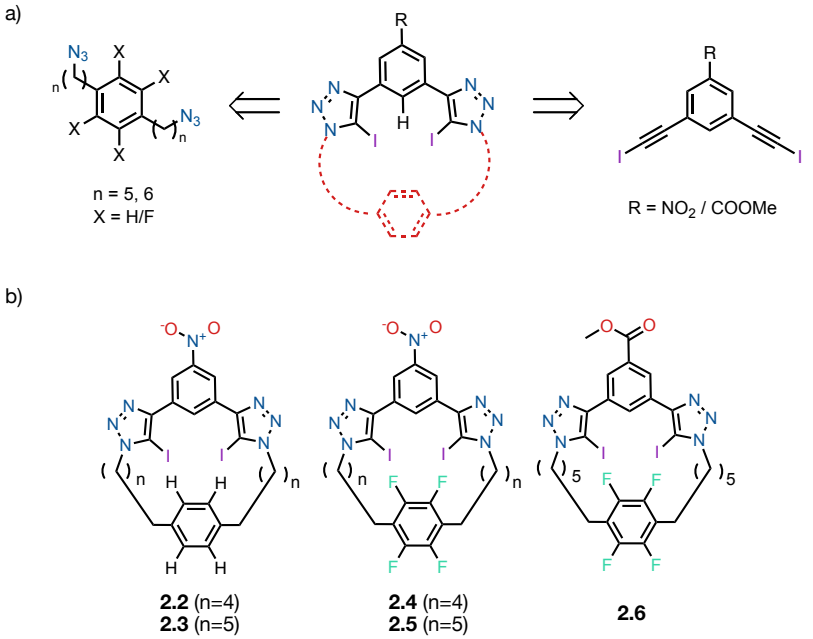
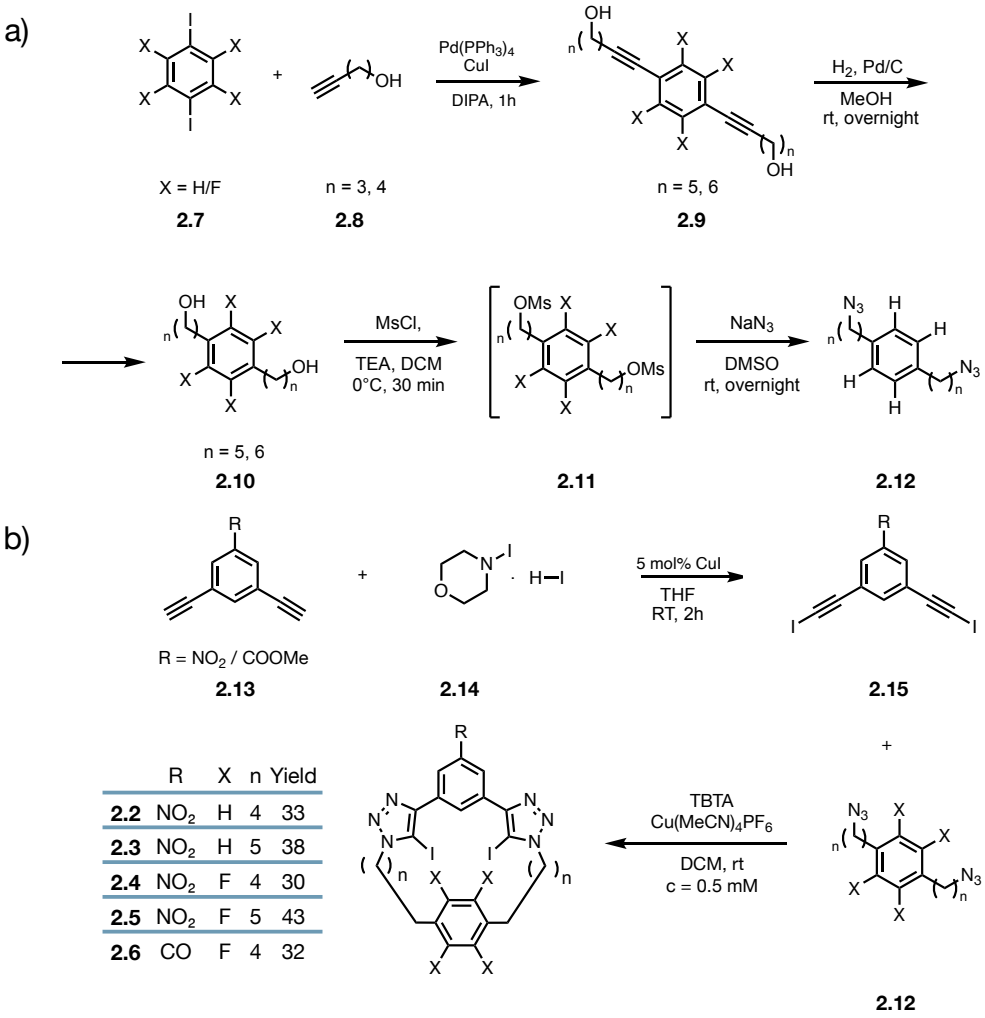



**Figure 2.3:** a) Disconnection strategy towards the family of macrocycles with aliphatic linkers **2.2**-**2.6** to give symmetric bis-azide and bis-iodotriazole precursors.b) The succesfully synthesised target structures of transporters **2.2**-**2.6**.

Macrocycles **2.2**–**2.6** were targeted by a convergent plan in which each ring is assembled from a symmetrical bis-azide and a symmetrical bis-iodoalkyne, closed in a double copper(I)-catalyzed azide-alkyne cycloaddition (CuAAC) to install two 5-iodo-1,2,3-triazoles per macrocycle (Figure 2.3). The disconnection yields two highly symmetrical and synthetically accessible precursors. The first is a *para*-disubstituted arene bearing –$(CH_2)_n$–$N_3$ termini and the second a *meta*-diiodoethynyl arene, with the two envisioned as coming together in a single step, with one intermolecular click reaction followed by an intramolecular macrocyclisation. This approach maximises convergence, allows variation of linker length ($n$ = 4, 5) and aryl electronics (X = H/F, R = $NO_2$/COOMe), and delivers both

encapsulation and $\sigma$-hole tuning in the final ring.




| | R | X | n | Yield |
|---|---|---|---|---|
| **2.2** | $NO_2$ | H | 4 | 33 |
| **2.3** | $NO_2$ | H | 5 | 38 |
| **2.4** | $NO_2$ | F | 4 | 30 |
| **2.5** | $NO_2$ | F | 5 | 43 |
| **2.6** | CO | F | 4 | 32 |

**Figure 2.4: Synthesis of key intermediates towards macrocyclic transporters 2.2-2.6.** a) Synthesis of the bis-azide precursors. b) Synthesis of the bis-iodotriazole precursors and yields for all of the final macrocyclisation steps.

A Sonogashira coupling of *para*-diiodoarenes **2.7** with the terminal alkyne alcohol **2.8** ($Pd(PPh_3)_4$, CuI, DIPA) provided diynes **2.9**.[148] Catalytic hydrogenation ($H_2$, Pd/C) reduced the triple bonds to saturated $–(CH_2)_n–$ linkers, affording diols **2.10**. Mesylation (MsCl, $NEt_3$) gave dimesylates **2.11**, which underwent substitution with $NaN_3$ (DMSO, rt) to furnish the bis-azides **2.12** in good overall efficiency

(Figure 2.4a). Reaction sequences, conditions and stoichiometries are summarised in Figure 2.4a (full synthetic details can be found in Appendix Chapter A).

The complementary iodoalkynes were obtained from the *meta*-diethynyl arenes **2.13** via copper-catalysed hydroiodination using the morpholine–HI complex **2.14** to give **2.15**.[149] In the optimised conditions, the CuAAC macrocyclisation of bis-azides **2.12** with bis-iodoalkynes **2.15** was then performed at high dilution (c = 0.5 mM) in DCM at room temperature with tris[(1-benzyl-1*H*-1,2,3-triazol-4-yl)methyl]amine (TBTA) used as an additive to enhance the reaction rate, furnishing the monomeric [1+1] macrocycles **2.2**–**2.6**. This was achieved in fairly low yields following painstaking purification, which included regular flash column chromatography followed by reverse-phase chromatography and finally recrystallisation to separate the [1+1] macrocycles from the side product (Figure 2.4b). The optimisation of this step is discussed next.

### 2.2.2 Competing [1+1] and [2+2] Macrocyclisations

The convergent disconnection (2.3) and forward route (2.4) proved reliable for accessing the aliphatic macrocycles in yields that allowed for their testing in vesicle assays. However, in the terminal CuAAC ring-closing step, competition between intramolecular [1+1] closure and bimolecular [2+2] cyclodimerisation was consistently observed. This was problematic both due to the low yields and highly difficult purification of the desired [1+1] macrocycle. Such competition is well precedented for macrocyclisations in general, as there is generally both an enthalpic and entropic cost to pre-arranging the molecule in a reactive conformation with smaller macrocycles.[150]

To explore this effect in the macrocyclic system, the crude ratios of the [1+1] monomer were compared to [2+2] dimer ratios for two different linker lengths

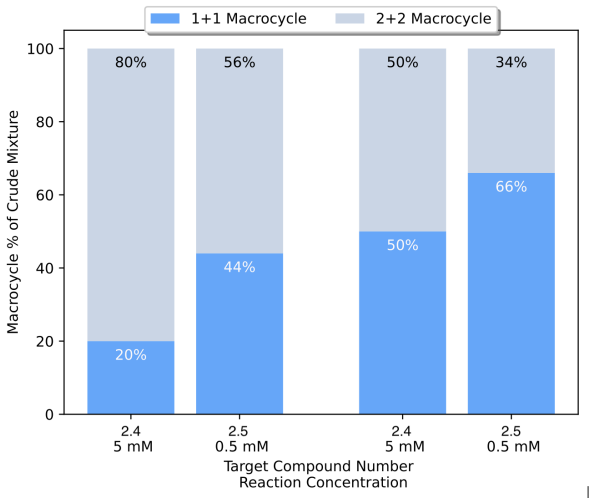



**Figure 2.5: Competition between [1+1] and [2+2] closure.** Crude product distributions for targets **2.4** (n = 4) and **2.5** (n = 5) at two precursor concentrations as determined by NMR spectroscopy of the filtered reaction mixture. Dilution from 5 mM to 0.5 mM favours intramolecular [1+1] macrocyclisation, with a stronger effect for the larger ring (**2.5**).

(targets **2.4** and **2.5**, n = 4 and n = 5). As shown in Figure 2.5, diluting the reaction from 5 mM to 0.5 mM results in a marked shift of the product ratio in favour of the [1+1] product. For the smaller macrocycle **2.4**, the monomer fraction increases from ~20% to ~44%, approaching parity with the undesired dimer, and making the mixture of products much easier to separate in practice. For the larger **2.5**, the monomer pleasingly becomes dominant (with from ~34% to ~66%). These observations align with a strain-controlled kinetic barrier that penalises short linkers in the transition state for [1+1] closure. The longer linker likely allows **2.5** to fold its two reactive ends together in the half-cyclised intermediate, and therefore cyclises more readily intramolecularly. This indication of ring strain presence in the smaller macrocycles indicated that any exploration of even shorter linkers would not be fruitful. As shown in literature, flow CuAAC macrocyclisations could thus be a particularly effective way to increase the intramolecular component without resorting to impractically large solvent volumes.[150] These reactions can be carried out at much higher temperatures, which help overcome the intramolecular

barrier, and move the product ratio to the [1+1] product at higher concentrations.

### 2.2.3 Synthetic Attempts Towards Terephthalic Acid Derivatives

Motivated by the design logic in Sections 1.2.5 and 1.2.3 an exploration of different variations of macrocycle linkers followed. In particular, a parallel family variant was desired whose length could also be easily modified but would offer different encapsulation and solubility properties. By adding electron-withdrawing groups next to the perfluorophenyl blocking group in the linker, there would also be a higher likelihood of achieving a third anion-$\pi$ interaction. It was decided that a tetrafluoroterephthalic acid would offer a good platform for these reasons (Figure 2.7). The two carbonyl substituents on the perfluorinated ring would increase electron withdrawal, thus making the aromatic ring even more electron poor and better suited for interacting with the chloride anion. The angle of the ring to the linker would also be different than in the all-aliphatic linker, offering a different encapsulation geometry. Finally, two different cyclisation strategies could be employed - both the original copper-mediated click macrocyclisation and an esterification or a condensation ring-forming pathway.

Thus, two convergent routes were pursued in parallel, an amide pathway, in which a bis-bromoethyl amide of a perfluoro-terephthalate was converted to the corresponding bis-azide and coupled to a bis-iodoalkyne in a double CuAAC macrocyclisation, and an ester pathway, in which terephthaloyl chloride derivatives were used to form diesters with the corresponding bis-alcohol partners prior to macrocyclisation, or alternatively click macrocyclisation was attempted as in the amide pathway. In both cases, no isolable [1+1] macrocycle was obtained despite extensive screening of conditions (Figures 2.6 and 2.7). Below, the observations and

the most plausible factors responsible for failure are summarised.

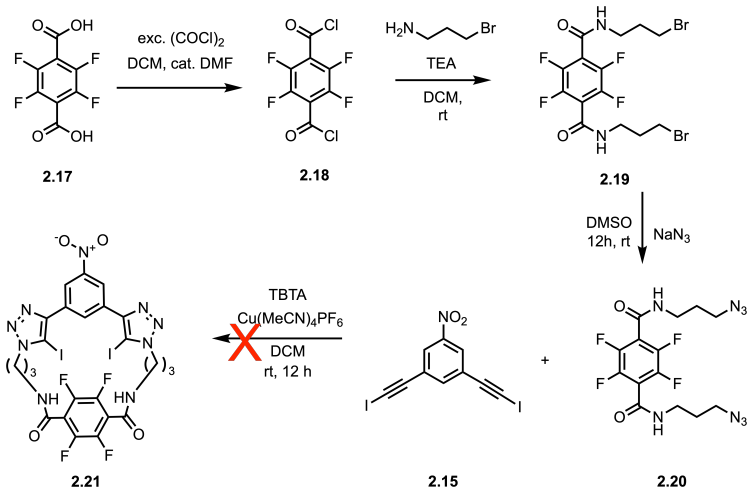
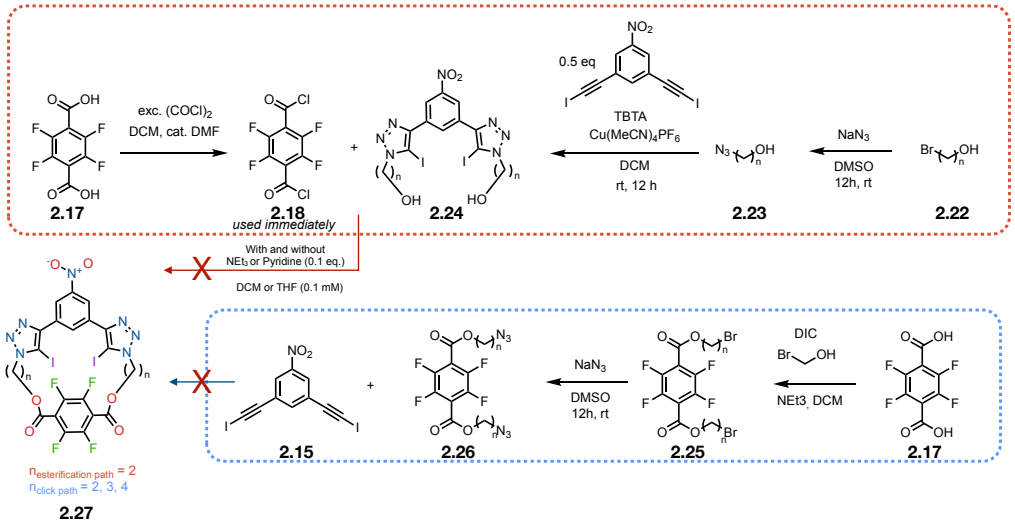



**Figure 2.6:** Attempts at synthesis of terephthalic amide derivatives that did not lead to successful coupling.

Activation of the diacid **2.17** with oxalyl chloride provided the terephthaloyl dichloride **2.18**, which coupled succesfully with $H_2N(CH_2)_nBr$ to give bis-bromoethyl amides **2.19**. Nucleophilic substitution with $NaN_3$ in DMSO furnished the bis-azide partner in good conversion **2.20**. However, double CuAAC between this bis-azide and the bis-iodoalkyne **2.15** partner failed to yield the targeted [1+1] macrocycle, and the reactions were found to return starting material or ill-defined oligomers. This is attributed partially to poor preorganisation due to the substitution of the phenyl moiety for the terephthalic amide derivative, which enjoys conjugation over a much greater area and is thus relatively more rigid. Thus the planar terephthalic unit holds the azide vectors too far apart for productive intramolecular closure at any practical concentrations. Alternatively, it could be possible that the filament-like linear intermediates, which have the ability to hydrogen bond as well as stack, may form non-productive intermolecular assemblies which further hinder macrocyclisation.




| Solvent | T (°C) | Reaction Time | TBTA | $Cu(MeCN)_4PF_6$ | Yield |
|---|---|---|---|---|---|
| DCM | 25 (then 40) | 12h (then 3h) | 0.1 eq. TBTA | 0.1 eq. $Cu(MeCN)_4PF_6$ | trace |
| DCM | 40 | 4 days | 0.1 eq. TBTA | 0.1 eq. $Cu(MeCN)_4PF_6$ | none |
| DCM | 25 - 40 | 7 days | total 0.5 eq. TBTA | total 0.5 eq. $Cu(MeCN)_4PF_6$ | trace |
| THF | 70 | 12h | 0.1 eq. TBTA | 0.1 eq. $Cu(MeCN)_4PF_6$ | none |
| EtOH | 70-120 (uW) | 10 - 120 min | 0.1 eq. TBTA | 0.1 eq. $Cu(MeCN)_4PF_6$ | none |
| Toluene | 110-130 (uW) | 10 - 30 min | none | none | none |
| Chloroform | 65 | 4 days | 0.1 eq. TBTA | 0.1 eq. $Cu(MeCN)_4PF_6$ | none |

**Figure 2.7: Synthesis of terephthalic ester derivatives that did not lead to successful coupling.** Top: acyl-chloride esterification pathway between the bis-iodoalkyne fragment and the bis-alcohol/bis-azide partners. Bottom: double CuAAC macrocyclisation pathway and associated conditions explored in the macrocyclisation step.

The synthesis of the terephthalate diesters (Figure 2.7), was attempted next via an analogous azidoalkyl and iodoalkynyl fragment route (blue dotted panel), as well as via a double esterification pathway (red dotted panel). For the latter, standard acyl chloride couplings were examined in DCM and DMF with $Et_3N$ or pyiridine and none furnished the desired diesters in isolable yield. The reaction was also hindered by poor solubility of the alcohol partners in the reaction solvent. In the double click cyclisation approach (Figure 2.7, blue panel), the double CuAAC macrocyclisation was again pursued by screening various combinations of

copper sources and additives, including $Cu(MeCN)_4PF_6$/TBTA, CuI/DIPEA, and $CuSO_4$/ascorbate systems, various solvent mixtures (DCM, THF, DMF, tBuOH / $H_2O$), high dilution (0.1–0.5 mM), and slow addition protocols. Despite all of the attempted conditions, only trace linear "half-clicked" products or intractable material were observed, consistent with a likely unfavourable pre-organisation of the compound, as was the case with the amide pathway.

Relative to the successful aliphatic series, the terephthalic linkers impose additional constraints on the approach geometry of the reacting ends and increase polarity. Both effects act to counter the macrocyclisation. The conformational constraints amplify the strain penalty of intramolecular closure, while the polarity mismatching of the substrates and the reaction solvents likely promotes aggregation and catalyst deactivation. These outcomes align with the general guidance for CuAAC macrocyclisations that favour preorganised, flexible tethers, high dilution, and minimal competitive ligation of Cu(I) — criteria met by the aliphatic family but not by the terephthalic variants. It was therefore decided not to follow any further variants of the terephthalic family.

## 2.3 Transporter Assays

### 2.3.1 $Cl^-$ Binding Titrations

Chloride anion binding affinities of compounds **2.2**-**2.6** were determined by $^1H$ NMR binding titrations with tetrabutylammonium (TBA) chloride in $d_6$-acetone containing 2.5% $D_2O$. This is a competitive solvent mixture utilised to represent anion binding to an anionophore at the hydrated lipid-aqueous interface. It was performed by monitoring the binding-induced changes in chemical shifts of protons in the benzene ring between the two triazoles (Figure 2.8a). The data in each case

could be fitted to a 1:1 stoichiometric binding isotherm using Bindfit.[151] By comparing the strength of the binding in this solvent mixture, the design changes of the macrocyclic transporters could be evaluated. They also serve as a basis for interpreting any changes in transport activities, and whether they arise due to binding differences or other transport parameters.

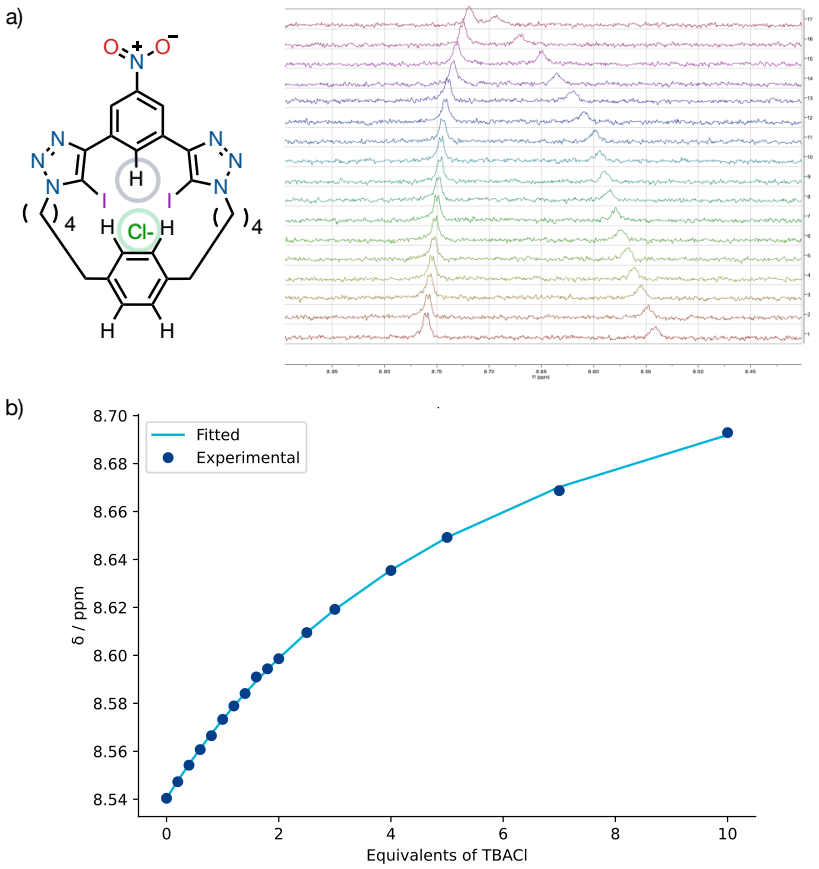



**Figure 2.8: 1:1 transporter-chloride association constants.** Binding constants were determined by titrating aliquots of $Cl^-$ as the TBA salt in 2.5% $D_2O$ (v/v) in acetone-$d_6$ and fitting with Bindfit software. a) Shows the diagnostic proton used to track $Cl^-$ binding in NMR titrations, as well as the binding titration experiment spectra for **2.2**. b) Corresponding $Cl^-$ binding isotherm. Experimental data shown by filled dark blue dots, while the fitted 1:1 binding isotherm is shown by a solid line.

Complete results of the titrations for the entire transporter family is presented in Table 2.2. Transporters **2.2** and **2.4**, which both feature the shorter pentyl spacers,

exhibited $Cl^-$ binding constants of 153 ± 6 and 130±25 $M^{-1}$, respectively. The larger analogues **2.3** and **2.5** with hexyl linkers, on the other hand, bound more strongly with binding constants of 897±78 and 720±34 $M^{-1}$, respectively. These results indicate that shortening the aliphatic linkers most likely causes the ring to adopt a worse binder-anion geometry and increases steric strain in the bound state, possibly even directly clashing with the ion. In contrast, the longer hexyl chains may allow the donors to relax toward a less clashing geometry and thus better accommodate the anion. Finally, replacing the *para*-$NO_2$ group in **2.4** by a methyl ester in **2.6** sharply reduced affinity to 36±4 $M^{-1}$, which is consistent with the expected smaller $\sigma$-hole at iodine and diminished halogen-bond donation in this less electro-withdrawing variant. Together, the series shows how both macrocycle size and substituent electronics come together to determine the chloride affinity in this transporter family.

**Table 2.2:** Chloride binding constants ($K_a$) from $^1$H NMR titrations with aliquots of TBACl in acetone-$d_6$ with 2.5% (v/v) $D_2O$ at 298K, fitted to a 1:1 isotherm using Bindfit.[151] Titration data can be found in Appendix Section A.1.3.2

| Compound | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|
| $K_a$ ($M^{-1}$) | 153 ± 6 | 897 ± 78 | 130 ± 25 | 720 ± 34 | 36 ± 4 |

### 2.3.2 Activity and $Cl^-$ Selectivity Assays

With transporters **2.2**-**2.6** in hand, the next step was to characterise their $Cl^-$ transport activities and selectivities in model vesicle systems.[44, 50, 52, 55] Large unilamellar vesicles (LUVs) were prepared using standard procedures, where a thin film of lipid was dissolved in a small round bottom flask, followed by hydration using the NaCl and HEPES buffer carrying the pH-responsive dye HPTS. Five freeze-thaw cycles of the resulting solution ensure that the vesicles that form are unilamellar, and extrusion through polycarbonate membrane with a pore size of 200 nm ensures they are of the appropriate size. Finally, size exclusion chromatography

separates the vesicles from the external HPTS dye-containing buffer, ensuring that the pH-dependent response will be from any $OH^-$ ions entering the vesicles, and not from the pH changes on the outside of the vesicles.

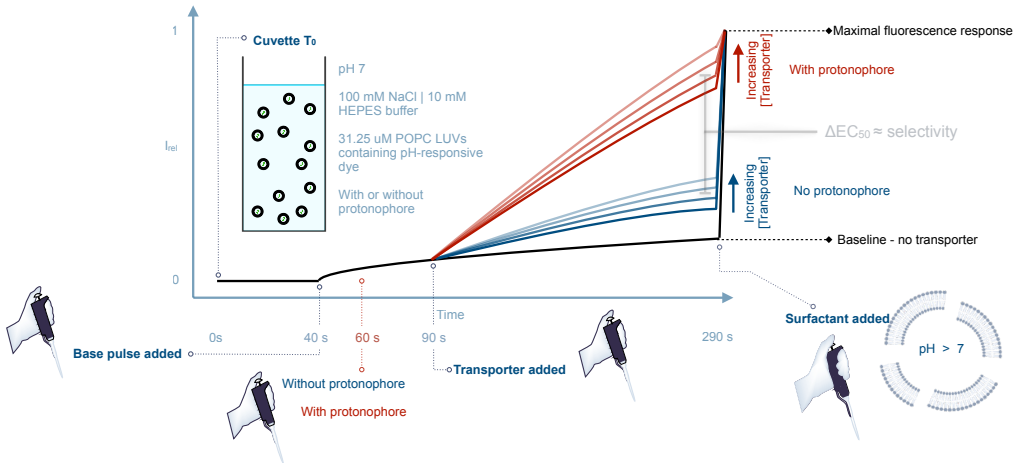



**Figure 2.9: HPTS base–pulse assay and concentration–response schematic explanation of the sequence of events.** At 0 s the measurement was begun using 3 mL of the prepared vesicle solution in a cuvette at 25 C. At 40 s, a pulse of aqueous NaOH (0.5M 30 $\mu$L) was added to initiate the experiment. In experiments utilising a protonophore, a pulse of FCCP in DMSO (100 $\mu$M, 7.5 $\mu$L) was added at 60 s . At 90 s, a pulse of the transporter in DMSO was added (varying concentrations, 7.5 $\mu$L). Finally, Triton X-100 in 7:1 (v/v) $H_20$:DMSO (11% (w%), 37.5 $\mu$L) was added to calibrate the assay. The data of of each concentration of assay was the average of at least two repeats.

Figure 2.9 shows the sequence of events unfolding during the $Cl^-$ transport assays. These assays allow us to determine the transport activity in the regimes where $Cl^-$ transport is rate limiting, as well as the regime where $OH^-/H^+$ transport is rate limiting, and determine the selectivity from the difference in activities (for more detailed explanations on the mechanistic background of the selectivity assay please refer to Section 1.2.3). Firstly, a base pulse is added to initiate a pH gradient at 40 s, followed by the optional addition of FCCP at 60 s. The pH gradient causes a small initial jump in fluorescence due to residual HPTS in the external buffer solution from any vesicles lysed prior to the start of the assays. This is controlled for in

the final concentration-response plots by using the same vesicle stock across the entire set of experiments for a single compound. Transport does not begin until the transporters are dosed to the vesicle solution at 90 s, whereupon a concentration-dependend response in the fluorescence increase is expected until vesicle lysis is performed at 290 s using the surfactant triton-X to calibrate the assay to 100% transport.

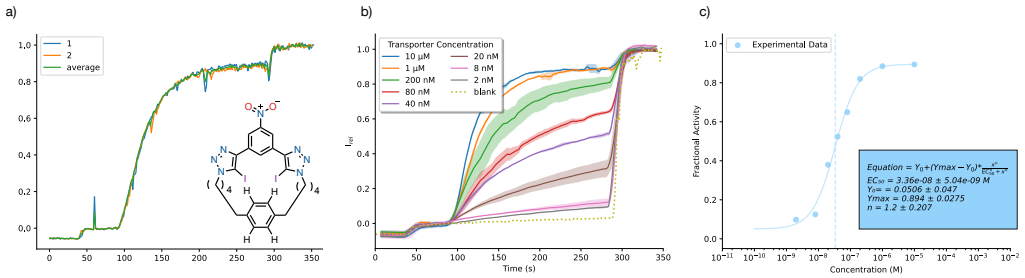



**Figure 2.10: Example anion transport data for 2.2** a) Two $Cl^-$ transport experiments with 1 $\mu$M of transporter **2.2** in the presence of protonophore FCCP are averaged. b) The entire range of concentrations is then obtained, from the minimal (2 nM) to the maximal (10 $\mu$M) fluorescence response. Shaded areas indicate standard deviations from the repeat runs. c) A logarithmic plot reveals the sigmoidal shape of the dose-response curve of the maximal absorbance prior to lysis at 290 s (scatter dots), which can be fitted to the Hill equation (full line).

For each transport activity experiment, the amount of deprotonated HPTS is monitored using the emission ratio at the emission wavelength $\lambda_{em}$ = 510 nm and the two excitation wavelengths $\lambda_{exc}$ = 460/405 nm, to give the emission ratio $R$. This is first normalised to give the relative fluorescence intensity $I_{rel}$ using the equation:

$$I_{rel} = \frac{R_t - R_0}{R_d - R_0}$$

where $R_t$ is the fluorescence ratio at time t following the start of the assay, $R_0$ is fluorescence at time 0, and $R_d$ is the fluorescence at the end of the assay, following equilibration following the addition of the detergent. For each concentration, two repeats are carried out and the relative fluorescence intensities averaged (Figure

2.11a). The entire set of concentrations is then repeated in the same manner, giving the concentration-dependent set of curves in Figure 2.11b. This can then be fitted to the Hill equation:

$$y = y_0 + (y_{max} - y_0)\frac{x^n}{EC_{50}^n - x^n}$$

where $y_0$ is the baseline fractional activity in the absence of transporter, $y_{max}$ is the fractional activity in with excess transporter and x is the transporter concentration in the cuvette. At least 7 concentrations were spanned in each case to ensure a reasonable fitting quality. From the fitting, the $EC_{50}$ value can be extracted, indicating the concentration of transporter required to give 50% transport activity under the assay conditions, meaning that a lower $EC_{50}$ value indicates a more active transporter. The value of $n$ can also be extracted, which is the Hill coefficient. This gives an indication of the stoichiometry of the transport process in case of endergonic assemblies, with a value of >1 indicating that several transporters may be acting together to bring the ions across the membrane.[152]

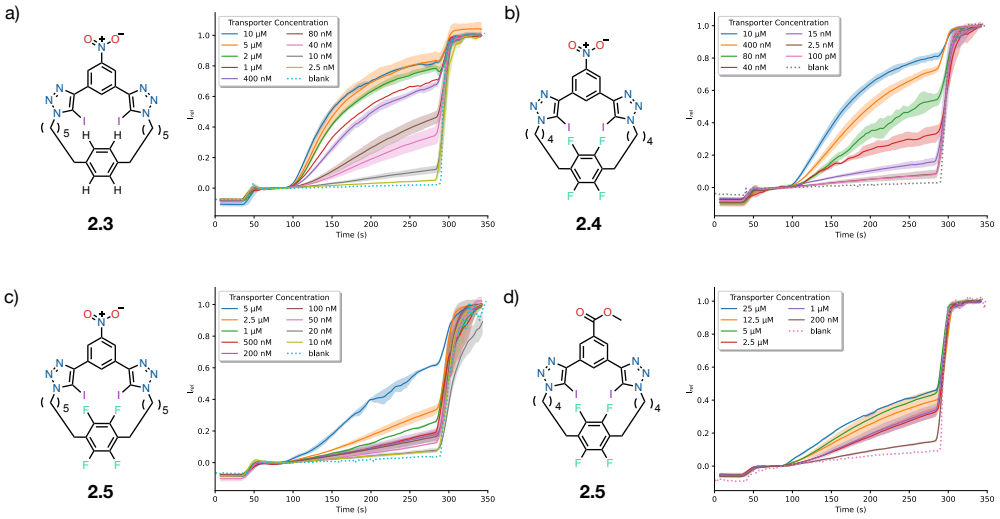



**Figure 2.11: Transport assay data for transporters 2.3-2.6 in the absence of FCCP.** The concentration dependent transport activity of the mobile carriers was measured in 200 nm POPC LUVs (31.25 $\mu$M).a) Transporter **2.3**. b) Transporter **2.4**. c) Transporter **2.5**. d) Transporter **2.6**.

The best-performing transporter **2.2** achieved an $EC_{50}$ of 34 ± 5 nM in the assays with the FCCP protonophore, where $Cl^-$ transport is the rate-limiting process, with a Hill coefficient of approx. 1.3, confirming that it likely operates in a 1:1 transport stoichiometry with the $Cl^-$ anion. Transporters **2.3** and **2.4** showed very similar activities to **2.2**, with $EC_{50}$ values of 68 ± 11 and 50 ± 7 nM, respectively, and also exhibited Hill constants of ~1, suggesting 1:1 uniport mechanisms. The larger of the two perfluorophenyl-containing macrocycles, **2.5** displayed disappointing transport properties, with fractional activities not reaching a plateau before the solubility limit, which was also the case with the bispentyl-perfluorophenyl-linked **2.6**, which differed from **2.4** only by the substitution of the nitro group in the latter with a methyl ester.

*Effective concentration to reach 50% of maximal activity in the HPTS assay, determined from Hill analysis of the relative fluorescence intensity at t = 288 s, immediately prior to vesicle lysis. Experiments conducted in LUVs of POPC (mean diameter 200 nm) loaded with HPTS (1 mM), with NaCl (100 mM) solution buffered to pH 7.0 with 10 mM HEPES in the intravesicle and extravesicle environment. b) Hill coefficient. $EC_{50}$ errors at the 95% confidence limit.

**Table 2.3: Transport metrics from HPTS assays without FCCP.***

| | **2.2** | **2.3** | **2.4** | **2.5** | **2.6** |
|---|---|---|---|---|---|
| $EC_{50}{}^{a}$ ( + FCCP) [nM] | 34 ± 5 | 68 ± 11 | 50 ± 7 | n.d. | n.d. |
| Hill $n$ $^{b}$ ( + FCCP) | ≈ 1.3 | ≈ 1.1 | ≈ 1.2 | n.d. | n.d. |

These trends add important context to the $Cl^-$ binding results discussed above. Clearly, while stronger binding can assist interfacial extraction, it is not sufficient on its own to guarantee transport activities. For example, despite relatively large $K_a$ values, **2.5** and **2.6** proved inactive in HPTS assays up to the 10 $\mu$M solubility limit (experienced by all transporters), whereas the modest binders **2.2** and **2.4** delivered nanomolar $EC_{50}$ values in the regime with the protonophore FCCP. This difference reaffirms the idea that transport is a multi-parameter optimisation task, as has been established in Section 1.2.5. Mobile carriers depend on a balance of binding strength with lipophilicity, lipophilic balance and deliverability. This suggests that that the highly lipophilic **2.5** (Table 2.1) likely violated the lipophilicity criteria.

A comparison between **2.2** and **2.3** shows that increasing binding affinity beyond a certain $K_a$ value no longer increases transport activity for this family of transporters. This suggests that uptake and release steps are likely no longer the the limiting factor at high affinities, and the membrane translocation step becomes limiting. Especially for compounds **2.5** and **2.6**, dosing at higher concentrations was limited by their solubility in the DMSO stock solution, which became turbid. This shows that membrane dynamics and solubility are too complex to be predicted by a simple line in the sand using predicted cLogP values, which were deemed to still be within the acceptable window.

The selectivities of the transporter were next investigated via transport assays conducted without FCCP. It is contended that a proton transport mechanism is highly improbable for these halogen bonding transporters, given their absence of acidic protons or basic nitrogen atoms (as the protonated triazole exhibits a p$K_a$ <1),

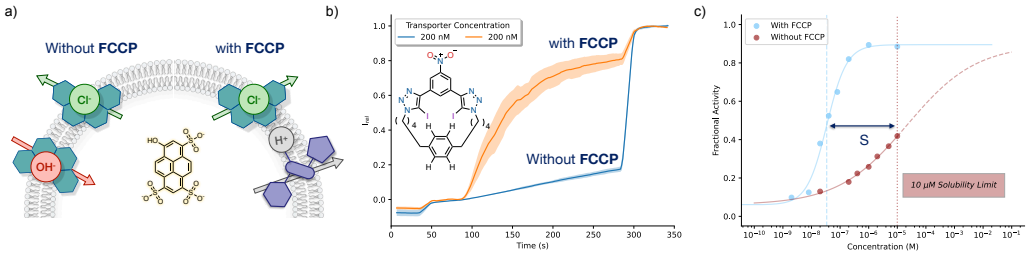



**Figure 2.12: Quantification of $Cl^-$>$H^+$/$OH^-$ selectivity for macrocycle 2.** a) Assay logic: without FCCP the pH response is limited by slow $OH^-$ movement; with FCCP a fast proton pathway uncovers the limiting rate of the intrinsic $Cl^-$ step. (b) Representative traces for fixed doses illustrate faster kinetics in the regime with FCCP. c) Dose–response at 290 s and definition of the selectivity factor $S$: $S = EC_{50}(\text{no } FCCP)/EC_{50}(\text{+FCCP})$. When 50% activity is not reached by the 10 $\mu$M solubility ceiling in the absence of FCCP, $EC_{50}(\text{no } FCCP)$ is estimated at the 10 $\mu$M solubility limit, yielding a lower bound for $S$. For **2**, $EC_{50}(\text{+FCCP}) = 34 \pm 5$ nM and $EC_{50}(\text{no } FCCP) > 10\,\mu$M, hence $S > 300$.

which prevents the reversible protonation states that facilitate such transport.[52] This means that this selectivity is a direct comparison of $Cl^-$ and $OH^-$ transport rates. The selectivity factor $S$ is obtained from the same Hill fits used for activity (Figure 2.11c). It is defined as the ratio of $EC_{50}$ values measured without and with FCCP:

$$S = \frac{EC_{50}(\text{+FCCP})}{EC_{50}(\text{no } FCCP)}.$$

For macrocycles **2.2**–**2.4**, where full activities could be obtained for experiments with the protonophore FCCP, the dose–responses curves without FCCP did not attain 50% fractional activity before the 10 $\mu$M solubility limit, which leads to an estimate of $EC_{50}(\text{no } FCCP) > 10{,}000$ nM. Using the fitted $EC_{50}(\text{+FCCP})$ values gives lower-bound selectivities of $S > 300$ for **2** (34±5 nM), $S > 100$ for **3** (68±11 nM), and $S > 150$ for **4** (50±7 nM). These results indicate that the halogen-bonding macrocycles strongly favour $Cl^-$ transport over the $OH^-$ pathways under POPC LUV conditions, while retaining high intrinsic $Cl^-$ activity in the +FCCP regime (nanomolar $EC_{50}$). Transporters **2.5** and **2.6** also displayed decreased transport rates without FCCP, suggesting that they retain some selectivity, although

this could not be effectively estimated, due to Hill plots not being obtainable in either the regime with FCCP or the regime without the protonophore.

**Table 2.4:** Operational $Cl^-$ > $H^+/OH^-$ selectivity from HPTS base–pulse assays. $S$ = $EC_{50}$(no FCCP)/$EC_{50}$(+FCCP). As 50% activity was not reached by the 10 $\mu$M solubility ceiling in the absence of FCCP, $EC_{50}$(no FCCP) is estimated at 10 $\mu$M and $S$ is reported as a lower bound.

| Compound | $EC_{50}$ (no FCCP) / nM | Selectivity $S$ |
|---|---|---|
| **2.2** | > 10,000 | > 300 |
| **2.3** | > 10,000 | > 100 |
| **2.4** | > 10,000 | > 150 |
| **2.5** | n.d. | Inactive |
| **2.6** | n.d. | Inactive |

### 2.3.3 Control assays: Fatty-acid and Cation Transport, and Membrane Fluidity Studies

Gale and co-workers have reported that anionophores appear less $Cl^-$-selective in commercial lipids because trace fatty acids catalyse $H^+$ transport via a flip-flop mechanism.[144] In short, this process consists of free fatty acids in the membrane losing the carboxylic proton, before flipping the carboxylate to the other membrane leaflet and picking up another proton, therefore in effect carrying out electrogenic transport of protons overall. By scavenging those acids with BSA, the true value of $Cl^-$>$H^+/OH^-$ selectivity can be achieved.[50, 144] Applying this control to the best-performing macrocycle **2.2**, it was found that the response without FCCP (where the proton transport step is rate-limiting) decreased by a factor of ~4 upon BSA treatment, whereas the activity with FCCP was essentially unchanged (Figure 2.13a, b). This means it can cautiously be said that, due to the four-fold drop in the no-FCCP signal after BSA addition, a selectivity $S > 1000$ can be estimated in fatty-acid-free POPC. This behaviour stands in contrast to earlier hydrogen-bonding tripods where BSA treatment was required to reveal high selectivity

(from 78 to 690).[144]

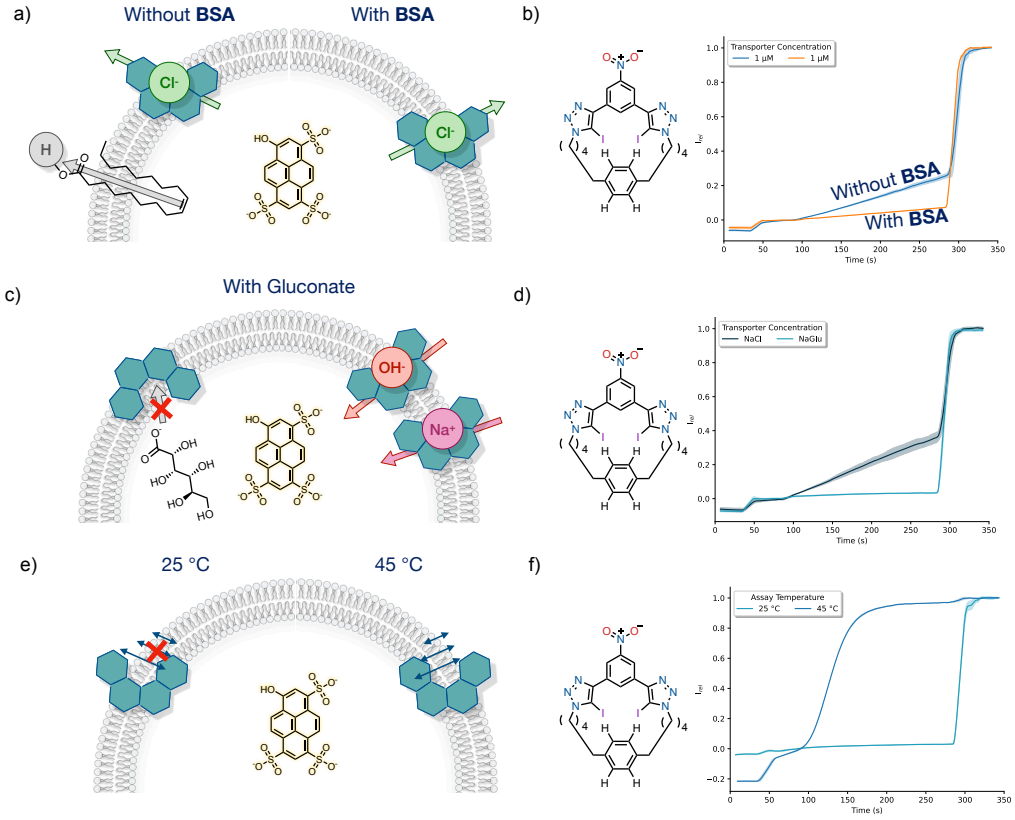



**Figure 2.13: Control experiments that validate mechanism or further quantify selectivity.** a) and b) **Bovine serum albumin (BSA) sequestration of free fatty acids.** POPC LUVs made from commercial lipid (containing traces of palmitate/oleate) were compared before and after treatment with BSA. Removing fatty acids suppresses carrier-assisted $H^+$ leak catalysed by fatty-acid flip-flop and therefore reduces the "no-FCCP" signal while leaving the +FCCP traces essentially unchanged. c) and d) **Anion substitution.** Replacing external $Cl^-$ by gluconate ($Glu^-$) kills the transport response, excluding $H^+/Na^+$ antiport and confirming that pH dissipation arises from $Cl^-$ transport. e) and f) **Phase sensitivity.** In DPPC vesicles transport is off below the gel–fluid transition (25 °C < $T_m$) and restored above it (45 °C > $T_m$), behaviour diagnostic of a mobile carrier rather than a membrane-spanning channel.

To confirm that the observed pH dissipation assay indeed reports $Cl^-$ ion antiport, the external $Cl^-$ was replaced by bulky, hydrophilic gluconate ions. Inactivity in sodium gluconate (NaGlu) (Figure 2.13c,d) rules out cation-dependent $H^+/Na^+$ antiport and confirms membrane integrity (no HPTS or non-specific proton leakage), which aligns with reported control experiments on known mobile carriers.[50]

Finally, experiments in dipalmitoylphosphatidylcholine (DPPC) vesicles below and above the gel–fluid transition (25 and 45 °C, $T_m \sim 41$ °C) provide a control for the existence of a mobile carrier mechanism. The absence of activity below $T_m$ and its restoration above $T_m$ (Figure 2.13e,f) indicate that the active transporter species performs transport by diffusing across the membrane, and is impeded by the membrane's gel phase. Synthetic channels, for example, usually show phase-independent behaviour.[50] Together, the three controls of BSA sequestration, anion substitution, and membrane phase-sensitivity demonstrate that macrocycle **2** is a highly $Cl^-$-selective, non-protonophoric carrier that remains highly selective even in complex, impurity-containing membrane mimics, an essential criterion for translation to cellular systems.

## 2.4 Conclusions and Future Work

In this chapter, the design, synthesis and characterisation of a novel family of macrocyclic anionophores are reported, which combine partial encapsulation with halogen bonding to deliver highly active non-protonophoric $Cl^-$ transport. A convergent double-CuAAC route succesfuly led to aliphatic-linker macrocycles **2.2**–**2.6**, while attempts to synthesise terephthalic variants did not cyclise productively, underscoring the importance of linker flexibility, preorganisation ability and catalyst compatibility for macrocyclisations. NMR binding studies in acetone-$d_6$/$D_2O$ showed that hexyl linkers (**2.3**, **2.5**) bind $Cl^-$ more strongly than pentyl analogues (**2.2**, **2.4**), while replacing *para*-$NO_2$ by an ester (**2.6**) weakens halogen bonding markedly. In POPC LUVs, **2.2**–**2.4** displayed nanomolar $EC_{50}$ values with FCCP, and very low activity without FCCP up to the 10 $\mu$M solubility limit, giving selectivity factor estimates of $S > 300$ (**2.2**), $> 100$ (**2.3**) and $> 150$ (**2.4**). BSA-treated vesicles further suppressed the no-FCCP response while leaving the

+FCCP response essentially unchanged, implying $S > 10^3$ for **2.2**. Gluconate substitution and DPPC phase-sensitivity controls confirmed an anion-dependent mobile-carrier mechanism. By contrast, **2.5** and **2.6** were solubility-limited and inactive under external dosing, consistent with an unpromising balance of halogen-bond strength, lipophilicity, and deliverability.

The high $Cl^-$>$H^+$/$OH^-$ selectivity arises from the cooperative features of neutral and directional $\sigma$-hole donors and a cavity that enforces partial desolvation of $Cl^-$ and penalises $OH^-$. The particular mechanistic origins of this effect are investigated in the next chapter (Chapter 3). Further gains in $S$ may be possible by modestly increasing enclosure while preserving the good binding properties, for example by shortening and rigidifying only portions of the linker to move toward greater encapsulation without reaching the strain threshold observed for very short tethers. Electronics can be refined by replacing *para*-$NO_2$ with strong inductive groups (e.g., $CF_3$, CN, $SF_5$) or by returning to more $\pi$-acidic aryls on the other triazole's nitrogen substituent while maintaining global clogP in the 5–7 range and maintaining encapsulation geometries. Enabling an anion–$\pi$ contact with a perfluorinated aryl may still be possible by careful linker construction coupled with conformational modelling to select suitable candidates.

The macrocyclic topology may also allow making the transporters stimuli-responsive by restricting the macrocyclic conformation in an unproductive state or appending a blocking group. Replacing a portion of the linker or the phenyl blocking group with photoswitches (for example by using a *cis-trans* azobenzene) would modulate encapsulation distance and $\sigma$-hole convergence. This would enable light-controlled on-off transport, or perhaps merely the switching of selectivity. Alternatively, pH-sensitive tethers could reversibly alter cavity shape, providing spatiotemporal control and act as safety checks against unintended protonophoric pathways.

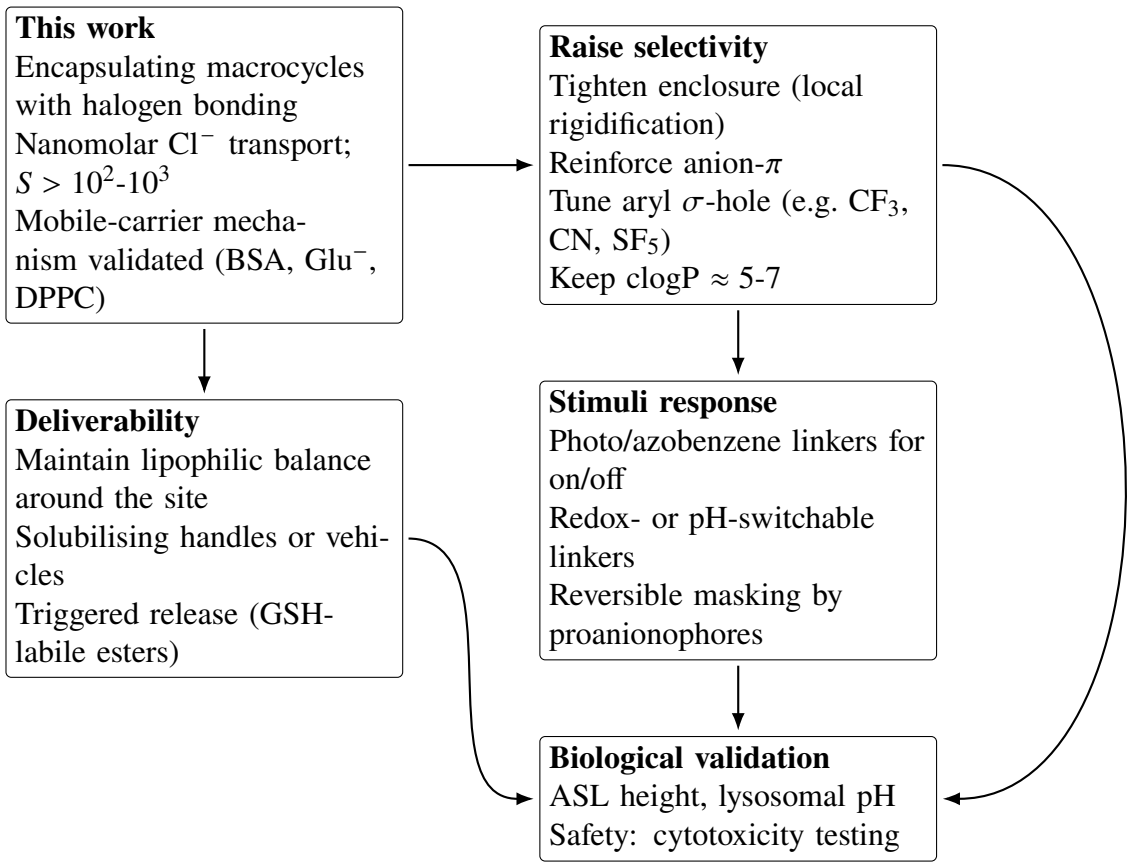



**Figure 2.14:** Text schematic of key conclusions and future directions.

In short, future designs should be directed towards varying linker length and rigidity to tune encapsulation and transport rates, as well as adjusting electronics around iodine to set $ESP_{max}$ while holding overall lipophilicity constant. Deliverability could additionally be improved in a biocompatible way by using proanionophore masks (e.g. GSH-labile esters) or mild polarity handles. Vehicles such as cyclodextrins, carriers containing polyethyleneglycol (PEG) linkers or others could also be promising.

The next stage of progressing the transporters towards the clinic would be to validate non-protonophoric $Cl^-$ transport in cellular systems that mirror CF pathophysiology. Air–liquid interface cultures of CF epithelial cells should then quantify

the amount of airway surface liquid (ASL) height recovery, as a key marker of CF treatment, as well as pH maintenance. The toxicity towards other cellular processes should also be monitored, including lysosomal pH, mitochondrial potential, membrane integrity, and inflammatory markers to exclude off-target protonophoric effects. Protein binding, mucus penetration, and stability in complex media would also need to be profiled early for development towards the clinic.

# 3 |

# Modelling Selective Chloride Transporters

In this chapter the transporters characterised in the previous chapter are modelled computationally. This is done in order to shed light on the $Cl^-$-selectivity observed. DFT modelling is used to model the inherent preference of the transporters for the two ions, finding a consistent energetic preference of the macrocycles for binding the $Cl^-$ over $OH^-$ arising from both enthalpic and entropic components. This finding links conformational preorganisation of the transporters to the ion binding selectivity. Additionally, molecular dynamics show that desolvation penalties likely also play a key role, which is a factor across all stages of the transport process, and confirms long-held beliefs about the encapsulation mechanism which had not been mechanistically investigated. Both of these findings may be used in the future to guide further optimisation of future selective transporters, be it through high-throughput screening or development of particularly promising scaffolds. This work was published in *Chemistry: A European Journal* in 2025.[54]

## 3.1 Preamble

Ion selectivity of synthetic carriers in passive membrane tranport is fundamentally a free–energy difference $\Delta\Delta G$ between the barriers of the membrane-crossing processes of different ions. In the modelling of carriers, it has often been assumed that one of the steps outlined in Section 1.3 is rate-determining, and that modelling this step is enough to compare the transporters. In the binding step, evaluation of the coordination of the anion to the transporter's binding site, just below the

lipid-water interface, is carried out. It is usually evaluated using DFT methods due to the high accuracy of intermolecular interaction energies between the transporter and different ions that the method offers. Nevertheless, this approach foregoes the assessment of conformational effects, explicit solvent stabilisations, and lipid interactions, which may all add crucial energetic contributions. The translocation step through the membrane core may also be energetically different for different ions, and can be studied most accurately using molecular dynamics, which capture interactions with the environment. This is usually done by restraining the anions to their bound pose in the transporter using a covalent model, before performing pulling across the membrane core. Both methods are discussed across this chapter.

## 3.2 Probing Inherent Ion Binding Preference

Size complementarity between a transporter's binding-preorganised pose and the ion it is transporting has previously been shown to be a driver of selectivity. In valinomycin, one of nature's great examples of selective transport, the discrimination between the larger $K^+$ and smaller $Na^+$ ions has been shown by quantum chemical calculations to be a product of the size-matching between the cavity of the transporter and the ions.[153, 154] Valinomycin's cavity has been shown to contain 6 of the oligodepsipeptide's carbonyls in an inward-facing configuration. This pose is very well adapted for binding the $K^+$ but not the $Na^+$ (Figure 3.1). Free energy perturbation studies utilising classical dynamics in $CH_2Cl_2$ have shown a preference of over 6 orders of magnitude for the larger ion. Analogously, the size difference between chloride and hydroxide ions (ionic radii of 1.8 and 1.1 Å, respectively), could, in principle, be exploited to enhance the selective binding of one over the other, if a macrocycle's minimum energy pose displays an arrangement more suitable for one of the ions.

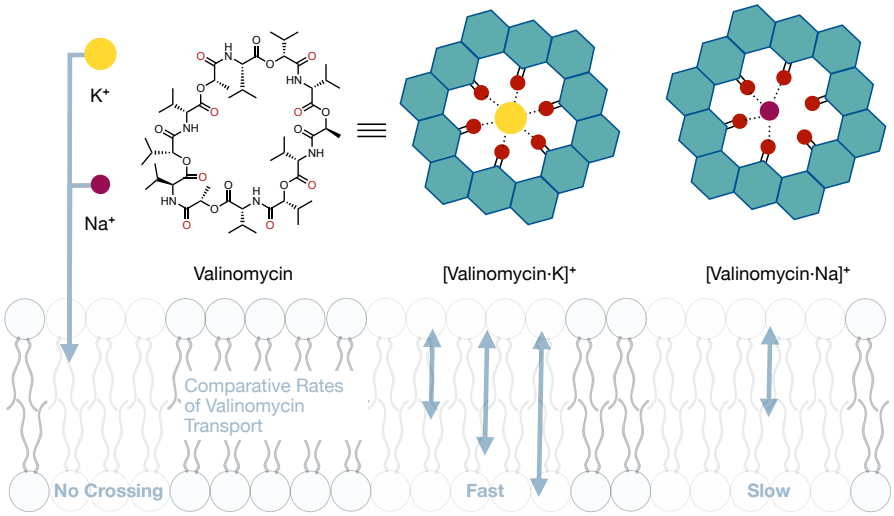



**Figure 3.1: Schematic of Valinomycin-assisted transport of $K^+$ and $Na^+$ ions.** Left: $K^+$ nor $Na^+$ is able to cross the membrane by itself. Valinomycin associates with both, but transports $K^+$ at rates much faster than $Na^+$. The size-mismatch between valinomycin and the respective ions is shown.

### 3.2.1 Binding Model

The study of $Cl^-$>$OH^-$ selectivity of the transporters described in the previous chapter (Figure 3.2a) began with modelling the binding step of the transport process. The thermodynamic model was based on that employed by the Duarte group in the study by Bickerton *et al.* and introduced in section (1.3.1). Namely, the energy differences are considered for binding of each of the $Cl^-$ and $OH^-$ to the transporter in a membrane-mimetic chloroform medium. The continuum solvation in $CHCl_3$ was originally chosen as simile of an average membrane environment, as it has a dielectric constant intermediate to the high-dielectric headgroup region and the low-dielectric core of the lipid tail region.[52, 155] The total binding energy is described by:

$$\Delta G_{binding} = \Delta G_{desolvation} + \Delta G_{association}$$

where $\Delta G_{desolvation}$ represents the movement of the ion from the aqueous to the membrane ($CHCl_3$) phase, while $\Delta G_{association}$ represents the movement of the ion from the membrane into the binding site of the transporter. The $\Delta G_{desolvation}$ component is the same for each ion in the energy cycle of all transporters, as it merely represents the transporter-independent movement of the ion from the aqueous phase to the chloroform phase. Thus, when comparing the binding energies of the transporters, only the $\Delta G_{association}$ energies of systems already solvated in chloroform were considered for each ion-transporter combination, as illustrated in Figure 3.2b. These were then compared as $\Delta G_{association}$ values, which showed the changes in ion binding strength when moving from acyclic and non-selective parent compound **3.1** to macrocycles **3.2**-**3.6**.

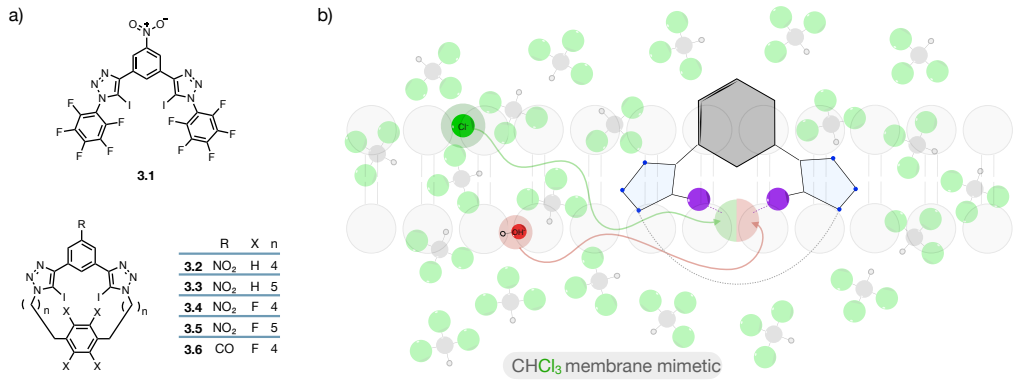



**Figure 3.2: Compounds under study and binding model.** a) Parent iodotriazole compound **3.1** and the novel iodotriazole macrocycles **3.2–3.6** differing in linker length ($n = 4, 5$), *para*- substituent on the central ring ($R = NO_2$ or COOMe), and aryl substituents ($X =$ H or F). b) Schematic of the interfacial binding model in a chloroform ($CHCl_3$) membrane-mimetic environment used to estimate relative binding strength of $Cl^-$ (green) vs. $OH^-$ (red) to the respective transporter carrier site. The drawn path indicates anion approach in an already desolvated environment, which can be used to compare different selectivities of non-macrocyclic and macrocyclic transporters.

### 3.2.2 Conformational Search Workflows Using XTB Methods

Macrocycles are well known for their vast conformational landscapes, which often have a very large number of geometries that are incredibly close in energy.[156–158] These poses may differ only by subtle non-covalent contacts, such as hydrogen bonds between distant moieties, or small steric clashes. Since the conformation of the transporter in the unbound state is very important as an energy benchmark for binding, and the [transporter·ion]$^-$ complex's pose is important both for maximising favourable interactions and minimising repulsions, it was imperative that the conformer searching workflow approaches the true minimum energy structures as closely as possible. In order to do this, a benchmark of conformer searching methods was employed in a two-tiered approach, where exhaustive, low–cost exploration with CREST was first employed to build an ensemble of low-lying conformers.[159] This was then followed by three different methods of energy refinement in order to rank the conformers as accurately as possible and find the lowest energy structure. The workflow is shown in Figure 3.3 and mirrors current best practice for medium–large organic molecules.[159–161]

CREST is based on GFN2-xTB, and uses metadynamics to explore conformational space in a fast and efficient way, giving a broad range of conformers.[159] Although it is broadly reliable for producing DFT-quality geometries and intramolecular contacts, it struggles with reproducing the energies of subtle and highly directional interactions, such as the sigma holes on the iodines of the transporters under study.[162] In contrast, modern dispersion-corrected hybrid DFT functionals capture these effects a lot better, which is why the reordering of CREST geometries was carried out using DFT calculations. Each of the reordering protocols tested had a different level of complexity and computational cost, as outlined below.

Method I utilises CENSO (Commandline ENergetic SOrting of conformer en-

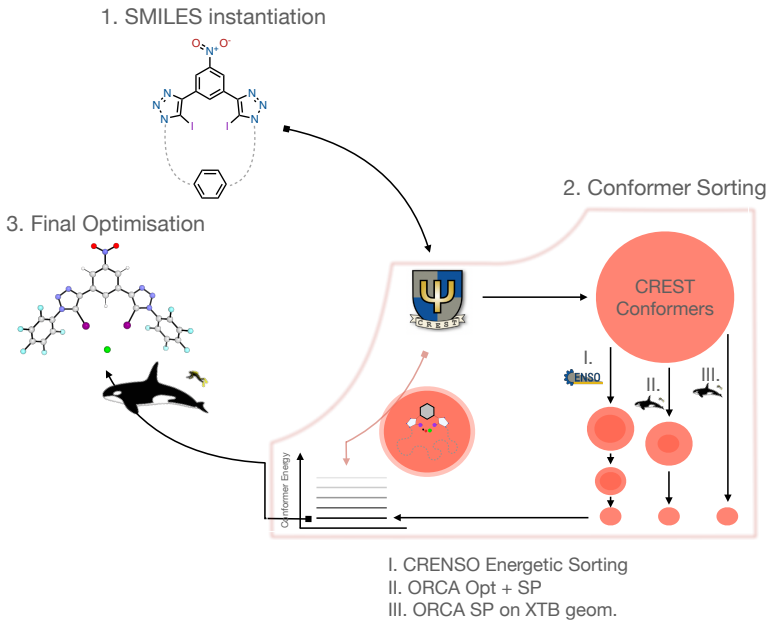



**Figure 3.3: Conformer generation and energetic sorting.** CREST, based on xTB generates a broad ensemble of conformers, from which the lowest-lying conformers are fed through to three refinement routes to be benchmarked. I a multi–tier CENSO energetic sorting, II re-optimisation at a low-level (LL) DFT followed by high-level (HL) single points, and III direct HL single points on xTB geometries. The best route III was selected for production runs; see text for details.

sembles), a method which performs an automated tiered refinement of energies of conformers from a CREST calculation.[163] First, a cheap DFT prescreening, followed by a higher-level sorting with SMD($CHCl_3$) solvation and optional re-optimisations. The conformers are pruned at each level of energetic optimisation and sorting, ending up at with the minimum energy conformer at the desired level of theory. The other two methods intended to offer a more cost-effective pathway, and relied on DFT calculations in ORCA through custom workflows.[164, 165] In method II, at most 40 conformers lying within the 6 kcal $mol^{-1}$ window of the minimum from the CREST output were reoptimised at a lower level (LL) of theory (SMD($CHCl_3$-$\omega$B97X-D3/def2-SVP), followed by single points at the higher level (HL) of theory (SMD($CHCl_3$-$\omega$B97X-D3/def2-QZVP). Finally, in method III, A

direct HL single point with the same SMD($CHCl_3$)-$\omega$B97X-D3/def2-TZVP setup was run on each of the 40 lowest-lying CREST geometries without intermediate DFT reoptimisation.

On representative molecule **3.2** routes I, II and III were found to identify the same lowest-energy conformer within $< 0.2$ kcal mol$^{-1}$. Because route III was at least 5× cheaper than both routes I and II, route III was selected for production conformer sorting.

### 3.2.3 Basis–set Benchmarking

In keeping with earlier study on **3.1** and with benchmarks that identify long range-corrected hybrids as reliable for halogen bonding, $\omega$B97X–D3 was adopted as the working functional.[56, 166] Because the systems under study feature highly polarised anions and heavy halogens which are sensitive to basis set choice, benchmarking was carried out on the basis set combination for geometry optimisations and single–point calculations, in order to assess binding energy convergence. All basis sets are from the Karlsruhe def2 family (with relativistic effective core potentials - ECPs - on I),[167]. Minimally augmented variants were considered throughout on the iodine and heavy atoms of the anions ($Cl^-$ and O of $OH^-$) to ensure that any required diffuse character necessary was present for the $\sigma$–hole interactions. SMD($CHCl_3$) implicit solvation was used in all electronic–structure steps.[168] The bar chart in Figure 3.4 reports $\Delta\Delta E_{\text{bind}} = \Delta E_{\text{bind}}(X//Y) - \Delta E^{\text{ref}}_{\text{bind}}$, where the $E_{\text{bind}}(X//Y)$ is the binding energy for a single point carried out with basis set $X$ on a molecule optimised with basis set $Y$ and $\Delta E^{\text{ref}}_{\text{bind}}$ is the binding energy of the final chosen basis set combination.

Two clear trends were observed in the benchmarking (Figure 3.4). Small or unaugmented triple–$\zeta$ bases underestimate $OH^-$ binding by ~2–3.5 kcal mol$^{-1}$

relative to the quadruple-$\zeta$ reference, while differences for $Cl^-$ are minor (0.1–0.5 kcal mol$^{-1}$). For example, TZVP//TZVP gives a 2.2 kcal mol$^{-1}$ binding underestimation for $OH^-$ but only 0.16 for $Cl^-$. A QZVP single point optimised at SVP level is energetically equivalent to higher level optimisation (QZVP//TZVP and QZVP//QZVP are within $\leq$ 0.3 kcal mol$^{-1}$ of the reference), meaning that convergence of the geometry optimisation can be achieved more cost-effectively, and optimising at the SVP level is sufficient. Extra polarisation alone (TZVPP) helps but does not make up for the additional functions available in the QZVP basis set for for $OH^-$, indicating that basis set size is the main limiting factor.

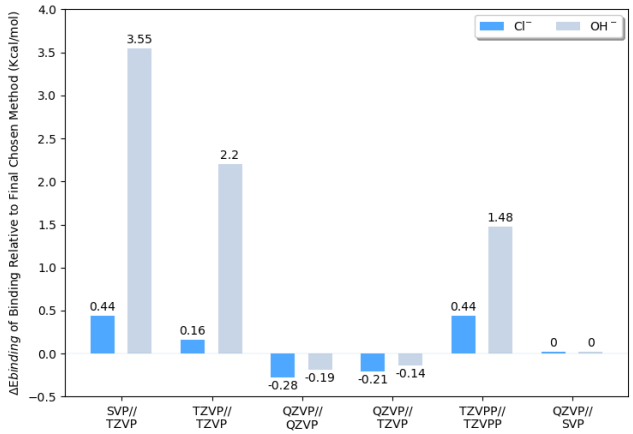



**Figure 3.4: Basis–set benchmarking for anion binding at $\omega$B97X–D with SMD($CHCl_3$).** Bars show $\Delta\Delta E_{bind}$ (kcal mol$^{-1}$) for different basis set combinations (single-point//optimisation), relative to the final chosen method (QZVP//SVP). Differences in binding energy are shown as blue for $Cl^-$ and grey fo $OH^-$.

### 3.2.4 Binding Energies

Figure 3.5 compares the optimised carrier $\cdot$ $Cl^-$ and carrier $\cdot$ $OH^-$ geometries for the acyclic parent **3.1** and macrocycles **3.2–3.4**. In the case of all transporters, the anions bind between the two iodines, with very highly conserved I$\cdots$ion distances and I-ion-I angles across all of the transporters, suggesting that ion binding ener-

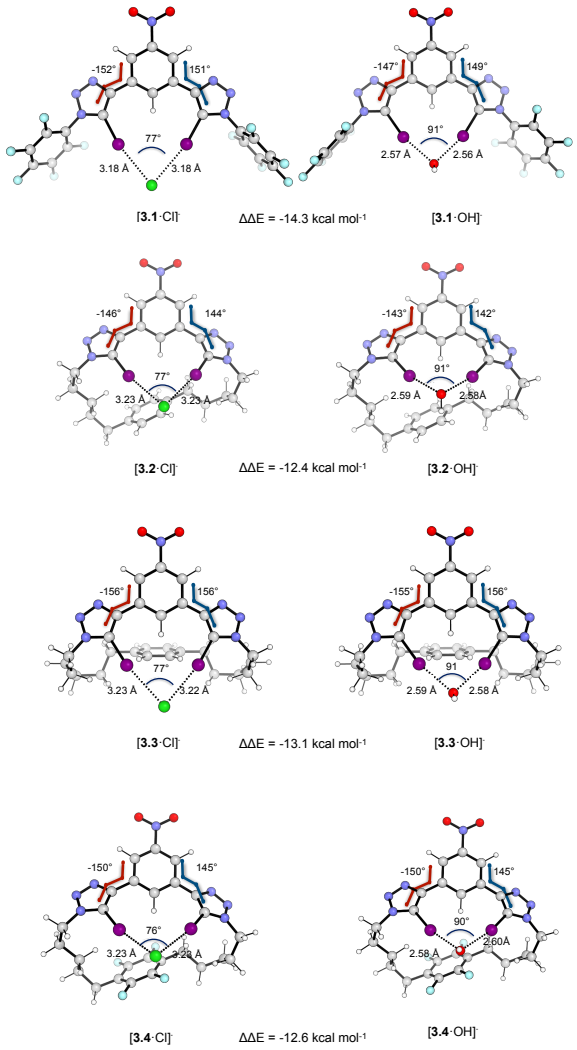



**Figure 3.5: Geometry and energetics of $Cl^-$ vs. $OH^-$ binding.** Optimised complexes for the acyclic parent **3.1** and macrocycles **3.2**, **3.3** and **3.4** illustrating I$\cdots$ion distances, I–ion–I bite angles, and iodotriazole-nitrobenzene dihedral angles for each transporter's complexes with both $OH^-$ (right of each pair) relative to $Cl^-$ (left). The difference in the singe point energies of binding are also shown. Here, $\Delta\Delta E_{\text{association}} = \Delta E_{\text{association},OH^-} - \Delta E_{\text{association},Cl^-}$, and $\Delta E_{\text{association},ion} = E_{\text{complex}(CHCl_3)} - (E_{\text{ion}(CHCl_3)} + E_{\text{transporter}(CHCl_3)})$. All calculations carried out at SMD($CHCl_3$)-$\omega$B97X–D/QZVP//$\omega$B97X–D/SVP.

getics dominates the overall geometry in the minimum energy conformation rather than transporter conformational preferences. Additionally, the conformation for the two transporters with pentyl group linker spacings (**3.2** and **3.4**) both adopt the same overall macrocyclic conformation. This behaviour, where the plane of the phenyl ring is in-line with the anion rather than perpendicular to it, demonstrates that an anion-$\pi$ interaction is unlikely to be operating. If such an interaction were occuring, it would likely result in the perfluorinated aryl ring adopting a different pose to the unfluorinated phenyl ring. This is supported by the binding constants from experiment, which showed no increase in binding affinity upon fluorination. The consistence of experimental results and DFT geometries when it comes to **3.2** and **3.4** are also a good litmus test for the quality of the conformer searching workflow. Since the sorting converged to a near-identical low–energy macrocycle pose for both transporters, which differ only in fluorination, it indicates robustness in its ability for finding the lowest energy conformer.

The larger macrocycle **3.3** with longer linkers also appears to fold back its linkers more strongly than **3.2/3.4**, which is indicated by its greater nitrobenzene-iodotriazole torsions. In the smaller rings, the linker resides closer to the binding site, which may contribute to the lower observed binding values for the smaller macrocycles relative to the larger variants as observed in Section 2.3. As expected, chloride complexes show longer $I\cdots Cl^-$ contacts across the whole series (3.18–3.23 Å) and tighter bite angles (76–77°), whereas complexes with the smaller hydroxide ion contract the $I\cdots OH^-$ distances to ~ 2.56–2.60 Å with an opened bite angle of 90–91°.

Considering the nature of the binding equation, where the ions are already solvated in the non-polar $CHCl_3$ solvent at the outset, and desolvation penalties are not considered, it is not surprising that $OH^-$ binding is strongly preferred in each transporter (-12.4 to -14.3 kcal $mol^{-1}$). But when comparing the magnitude of this

preference, it was found that it diminishes upon cyclisation and is smallest for the smaller rings **3.2/3.4**. The structural picture suggests that the macrocycles enforce near–identical donor geometry for both anions. It appears that the less flexible conformational landscape in the rings thus reduces the interaction advantage that $OH^-$ enjoys in the acyclic molecule **3.1**. In order to investigate whether the chloride's relative advantage over the hydroxide in the macrocycles stems from more highly energetic interactions with the ion or a smaller distortion of the bound conformation, Houk's distortion-interaction analysis was performed.[169] It should be noted that we group the macrocycles by ring size in subsequent energy analyses, with the 14 carbon linker **3.2** and **3.4** grouped together, followed by the 16 carbon linker **3.3**.

In the distortion–interaction analysis the energy difference between two states is decomposed (here, the unbound and the bound state, Figure 3.6a) into the stabilising (i.e. interaction) and the destabilising energetic components (i.e. distortion). In the case of binding of the ions to the transporters under study, the entirety of the distortion component results from the rearrangement of the transporter into the bound pose, as the ions cannot really undergo any conformational changes. The interaction component of the binding energy then comes from the attraction between this pre-arranged transporter and the ion.[169] Together, the interaction and distortion therefore sum to give the binding energy. In the analysis, differences are calculated in each component between the two ions, for each transporter, i.e.:

$$\Delta\Delta X = \Delta X_{\text{OH-binding}} - \Delta X_{\text{Cl-binding}}$$

where $X \in \{E_{\text{binding}}, E_{\text{interaction}}, E_{\text{distortion}}\}$. Across the parent molecule and the macrocyclic family of transporters, the distortion penalty for accommodating the tighter I-$OH^-$ contacts is very consistent: 4.7 (**3.1**), 4.0 (**3.2**), 4.5 (**3.4**), 4.9 kcal

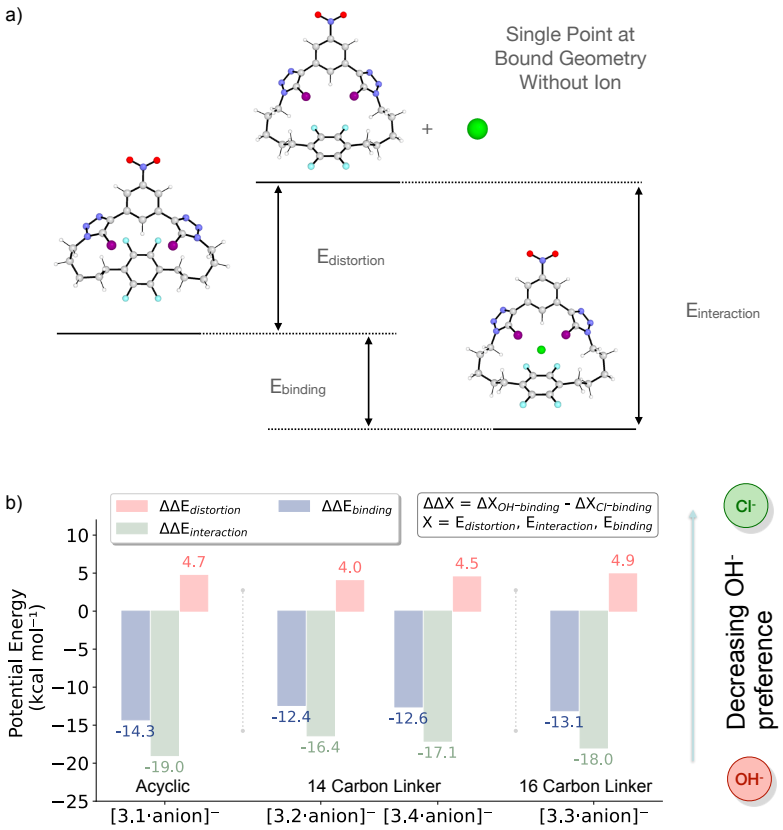



**Figure 3.6: Distortion–interaction decomposition for $OH^-$ vs. $Cl^-$ binding.** a) Schematic definition of terms, illustrated with the distortion-interaction binding cycle on molecule **3.4**. b) $\Delta\Delta E$ contributions for **3.1**, **3.2**, **3.4**, **3.3**; labels above and below bars give $\Delta\Delta E$ values. More negative values favour $OH^-$.

$mol^{-1}$ (**3.3**). In other words, the preorganisation step consistently deforms the donor geometry by a similar amount when switching from $Cl^-$ to $OH^-$, costing an additional ~4–5 kcal $mol^{-1}$. What really differentiates the receptors is the interaction term. The acyclic **3.1** enjoys the largest $OH^-$ advantage ($\Delta\Delta E_{\text{interaction}}$ of -19.0 kcal $mol^{-1}$) yielding an overall $\Delta\Delta E_{\text{binding}}$ of -14.3 kcal $mol^{-1}$. Upon cyclising, this gain is reduced by 2–3 kcal $mol^{-1}$. It lies at -16.4 and -17.1 kcal $mol^{-1}$ for the pentyl-spaced **3.2** and **3.4**, and at -18.0 kcal $mol^{-1}$ for the longer **3.3**. Consequently, the net binding preferences correlate with $\Delta\Delta E_{\text{interaction}}$: −12.4 (**3.2**), −12.6 (**3.4**), and −13.1 kcal $mol^{-1}$ (**3.3**). The most likely explanation is

that the smaller rings can only adopt a less flexible geometry where the $\sigma$-bond donors cannot tighten as effectively around $OH^-$, while the the longer **3.3** is able to fold onto the smaller anion more efficiently and partially restores some of the interaction advantage seen in **3.1**, which aligns with the analysis of the poses (*vide supra*).

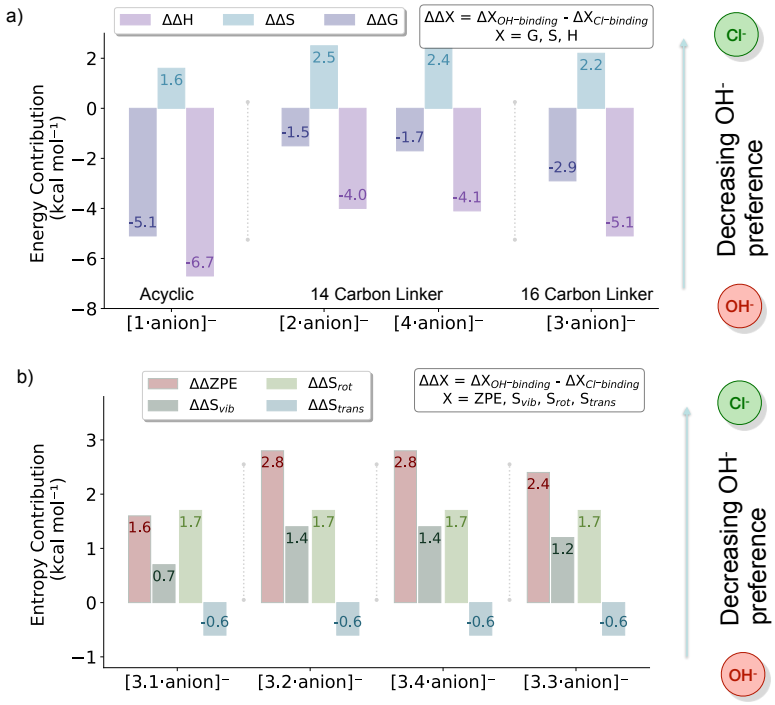



**Figure 3.7: Free energy and entropy term analysis for $OH^-$ vs. $Cl^-$ binding.** *(a)* $\Delta\Delta H$, $T\Delta\Delta S$, and $\Delta\Delta G$ for **3.1**, **3.2**, **3.4**, and **3.3**. More negative values favour $OH^-$. *(b)* Breakdown of thermal terms (zero-point, vibrational, rotational, translational).

Next, the binding energies were quantified with free energy contributions, as these additionally account for the curvature of the potential energy surface and thus the vibrations that are thermally accessible to the different transporters (Figure 3.7a). The relative difference of binding free energies between the two ions is quantified as $\Delta\Delta G = \Delta G_{\text{OH-binding}} - \Delta G_{\text{Cl-binding}} = \Delta\Delta H - T\Delta\Delta S$. As was the case with electronic energies, cyclisation decreases the intrinsic $OH^-$ preference observed

for the acyclic parent molecule. $\Delta\Delta G$ changes from $-6.7$ kcal mol$^{-1}$ for **3.1** to $-4.0$ and $-4.1$ for the smaller macrocycles **3.2** and **3.4**, and reduces less significantly to $-5.1$ for the longer **3.3**. By examining the contributions to the free energy, it can be seen that the origin is two-fold. First, the enthalpic advantage of $OH^-$ weakens on cyclisation and $\Delta\Delta H$ becomes less negative, going from $-5.1$ for **3.1** to $-1.5$, $-1.7$, and $-2.9$ kcal mol$^{-1}$ **3.2**, **3.4** and **3.3**. This mirrors the energy differences from electronic energies.

Another relative penalty for hydroxide binding arises from the entropy, which shows increases from $T\Delta\Delta S = 1.6$ (**3.1**) to 2.5 and 2.4 for **3.2** and **3.4**. The longer macrocycle shows a slightly smaller penalty, of 2.2 (**3.3**). Once again, this points to a picture of binding site preorganisation being more favourable for binding of the larger $Cl^-$, and the pentyl-linked, less flexible macrocycles encountering a higher entropic penalty when trying to accommodate binding of the smaller $OH^-$. This is confirmed by the decomposition of the entropy (Figure 3.7b), which clarifies which terms shift upon cyclisation. The translational and rotational components are essentially constant for all systems, with $\Delta\Delta S_{\text{trans}}$ standing at $\sim -0.6$ kcal mol$^{-1}$ and $\Delta\Delta S_{\text{trans}}$ at $\sim 1.7$ kcal mol$^{-1}$, respectively. The true difference comes from the zero-point term and the vibrational entropies, with a higher penalty for the smaller rings, suggesting a stiffening of the vibrational modes. If the macrocycle's low-lying breathing modes rise in energy upon replacement of the $Cl^-$ with the $OH^-$, this decreases the number of thermally-available modes, resulting in a penalty in the ZPE and vibrational entropies, which indicates another mechanism that operates to select for $Cl^-$ binding, related to the preorganisation and rigidity of the macrocyclic rings.

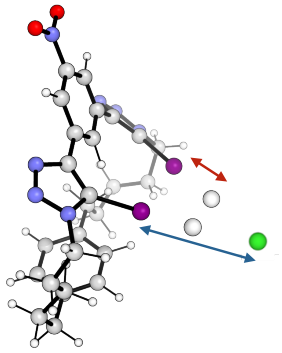
## 3.3 Membrane and Explicit Solvent Behaviour

To explore the thermodynamics of the transport process with consideration of explicit solvent and membrane environments the representative, most selective transporter **3.2** was modelled with molecular dynamics in both environments.

### 3.3.1 Halogen Bonding Parameterisation

Although the MD described in this section work does not attempt to compute binding constants, a realistic description of the anisotropic electrostatics was required at the halogen-bonding site in order to ensure realistic membrane poses, interactions with the bulk phases and to allow for close contacts between the iodine atoms and the respective ions in the restrained binding conformations. This is achieved in classical force fields by the use of a so-called extra point model (described in Section 1.5.1). These massless particles, which carry a charge and are constrained to a fixed position from the center of one of the atoms, allow for the redistribution of the charge in anisotropic atoms. For example, the partial positive character of the $\sigma$-hole 180° away from the iodine's bond to the electron-withdrawing triazole moiety can be added as a small positive charge. While the iodine would normally be only weakly positive, or even negative in some cases, the addition of the EP allows for its Lewis acidic interactions with water, lipid headgroups and the ions of interest.

Binding was calibrated in molecular mechanics optimisations without solvent, against the binding energies and distances in DFT calculations (obtained at $\omega$B97X-D3/def2-QZVP//$\omega$B97X-D3/def2-SVP). These DFT-optimised structures were also used as initial structures for the molecular mechanics optimisations, which followed a standard gradient descent algorithm in GROMACS for 10,000 steps. Different



| Fitting Parameter | | | Observable | |
|---|---|---|---|---|
| EP - I Distance (Å) | Lennard Jones (Å) | RESP Fitting Method | I-Cl Binding Distance (Å) | Binding Energy (kcal mol$^{-1}$) |
| 2.15 | 2.15 | free fit | 3.88 | -22.9 |
| 2.00 | 2.15 | free fit | 3.57 | -29.0 |
| 1.80 | 2.15 | free fit | 3.61 | -28.0 |
| 1.60 | 2.15 | free fit | 3.62 | -28.0 |
| 1.60 | 2.15 | 0.25 | 3.66 | -30.8 |
| 1.80 | 2.15 | 0.2 | 3.62 | -32.0 |
| 1.80 | 2.15 | 0.3 | 3.52 | -39.6 |
| 1.60 | 2.00 | free fit | 3.59 | -29.3 |
| 1.60 | 1.80 | free fit | 3.36 | -34.0 |
| DFT Benchmark | | | 3.07 | -38.7 |

**Figure 3.8: EP parameter optimisation for iodine.** Left: definition of the extra point (EP) on the C–I axis at distance $r_{EP}$; the iodine radius $R_{min}/2$ was optionally reduced. Right: Table showing the effect of the different EP fitting options of the iodine atom (EP distance from iodine centers, Lennard-Jones parameter, and whether RESP fitting was free or constrained to a set Q charge value) compared to the DFT reference distances and energies.

literature fitting methods were combined (Figure 3.8), testing for stability in simulations after each fitting, to obtain the closest binding energies and distances to the DFT reference while maintaining stability in molecular dynamics simulations (Figure 3.8).[118, 170, 171] The iodine atom's EP position was scanned from a distance of 2.15 to 1.60 Å, finding that distances below this lower bound produced excessively large extra point charges, which resulted in instability, but likewise finding that the I-Cl binding distance was still much too long in comparison to the DFT reference. Next, it was attempted to manually increase the EP charge by fixing it in the fitting at the distance of 1.80 Å, but this could not improve the binding distance error beyond 0.45 Å. The Lennard-Jones parameter of the iodine atom was then systematically lowered while keeping the position of the EP fixed, to allow for a closer approach of the $Cl^-$ ion, which gave the best balance of binding energies and geometries. Reductions of the Lennard-Jones parameter beyond 1.80 Å also resulted in instabilities in the simulation, likely due to overly close contacts and thus integrator instabilities. This suggested that an EP distance of 1.60 Å combined with a 1.80 Å Lennard-Jones parameter was the stability boundary.

### 3.3.2 Metadynamics Conformational Studies of a Macrocyclic Ionophore In Solvent

The high dilution conditions that were required to facilitate the cyclisation of the smaller macrocycles indicated the presence of a high level of strain. NMR studies on triazole-containing macrocycles have shown that strain can increase preorganisation, slow interconversion between conformers and thus limit the conformational flexibility of these rings. This opened up the question of whether the macrocycles studied may be getting trapped in unproductive conformations.[150] Namely, upon macrocyclisation some of the macrocycle may become trapped in a configuration where the two iodines are not converged in space (i.e. a *trans*- arrangement with respect to the iodines). In a highly strained macrocycle, interconversion between such a *trans*- arrangement to the *cis*- arrangement may not be thermally achievable. In this case, there could be a fraction of the macrocycle that is not able to transport the chloride ions effectively.

This led to an investigation of the dynamics of macrocycle **3.2**'s interconversion between the *cis*- and *trans*- conformers, with respect to iodotriazole orientation. The conformational landscape was mapped by performing two-dimensional well-tempered metadynamics. This method adds small biases in the form of Gaussian hills along the collective variable (i.e. a chosen coordinate), slowly biasing the system away from the minima. In well-tempered metadynamics (WTMetaD) the hill sizes get smaller as the overall bias accumulates over time, leading the dynamics to become tempered, meaning that they are free to explore the entirety of the collective variable space (for an example of how the free energy surface is "filled" with bias deposition over time, see Figure 3.10c). The sum of the biases can then be used to reconstruct the shape of the free energy surface. The collective variables were chosen to be the dihedral angles of the two iodotriazole rings to the

nitrobenzene moiety.

Molecule **3.2** and the DMSO solvent were modelled using GAFF parameters. **3.2** was inserted into a DMSO box and minimised using a gradient descent algorithm, and a short NPT equilibration (298 K, 1 bar, 2 ns) was conducted. NPT simulations were then conducted in GROMACS using Plumed as the driver for the 2D-metadynamics for a total of 1200 ns, until convergence was reached.[172, 173] Gaussians were deposited every 500 steps with a height of 0.3 kJ $mol^{-1}$ and a width of 0.3 radians. The biasfactor was set to 6.

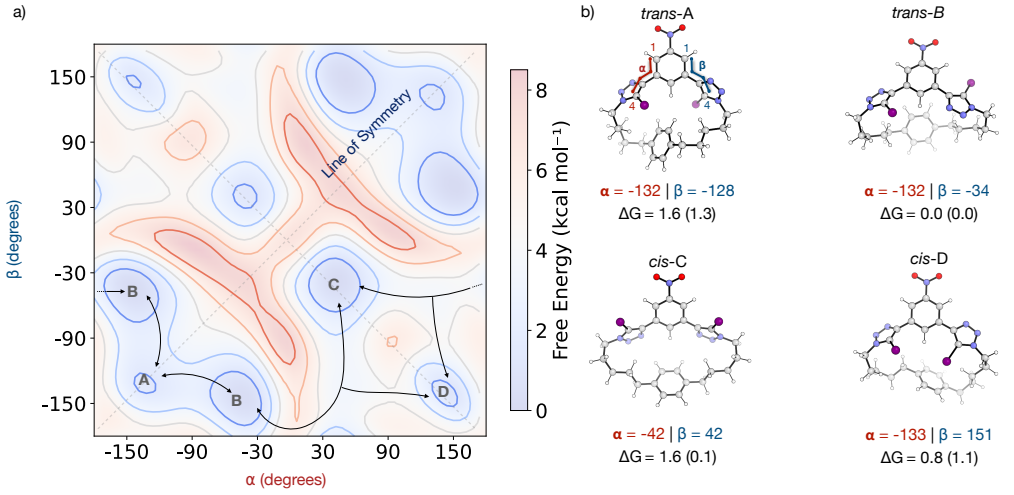



**Figure 3.9: Results of 2D well-tempered metadynamics** a) 2D energy profile as a function of torsion angles $\alpha$ and $\beta$ for 2, generated using 2D well-tempered metadynamics MD (WT-MetaD) simulations in explicit DMSO solvent at 298 K, using GAFF parameters for DMSO and the transporter, with iodine modeled using anisotropic charge following literature protocols. b) Representative structures for each of the minimum identified in the 2D energy profile computed independently using CREST followed by DFT optimisation at the SMD(DMSO)-$\omega$B97X-D3/def2-QZVP// SMD(DMSO)-$\omega$B97X-D3/def2-SVP level of theory (highlighting angles $\alpha$ (red) and $\beta$ (blue)). Energies are reported in kcal $mol^{-1}$ relative to conformer B, with energies from WTMetaD presented in parenthesis.

Four energy minima (A-D) were identified in the 2D free energy landscape, corresponding to *trans*- (A,B) or *cis*- (C,D) arrangements of the two iodotriazole moieties relative to the plane of the ring (Figure 3.9). Only the *cis*- arrangements

allow bidentate $Cl^-$ binding, with *cis*-D being the most pre-organized for binding due to its convergent halogen bonding groups. Due to the macrocycle's C2v symmetry, the energy profile also contains two lines of symmetry. Interestingly, the metadynamics predicts an ability of the iodotriazole rings to rotate around both directions of rotation - with the iodine pointing out of the ring (through the 0° angles) as well as by threading the iodines through the middle of the ring (through the 180° angles). Both pathways are thermally accessible for interconversion between the conformers, with barriers between 3 and 6 kcal $mol^{-1}$. This suggests that thermal trapping of conformers does not occur, and the molecule is able to prearrange for binding.

In order to probe the validity of the free energy surface obtained via metadynamics further, representative conformers were sought of each minimum via a dihedral angle-constrained conformational search using the conformer-rotamer ensemble sampling tool (CREST). For the lowest-energy conformers (sorted using the same workflow as in Section 3.2.2) DFT optimisation and frequency calculations were performed. Next, the energies of the conformers found using DFT were compared against the minima found by metadynamics. Both MD simulations and DFT calculations indicate that all conformers are close in energy and within 1.6 kcal $mol^{-1}$ of the *trans*-B global minimum. While both methods find B to be the minimum, they change the order of the next two lowest-energy minima, the *cis*- conformers C and D. Despite this, the close match between the MD and DFT relative energies, and the low barriers to interconversion found by the metadynamics simulations suggest that thermal interconversion is able to proceed in the DMSO solvent used to dose the transporter into the assays.

In order to ensure that the metadynamics was performed for a sufficiently long time and the analysis above is valid, the convergence of the metadynamics simulations was tracked by way of monitoring the 1D projection of the $\alpha$ angle over the course

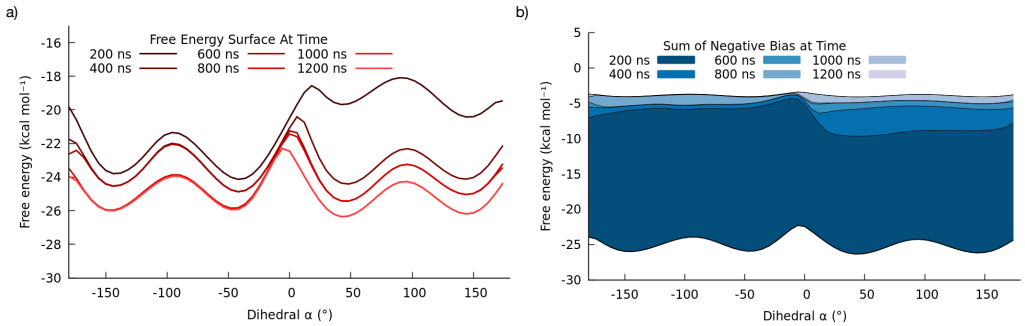



**Figure 3.10: Tracking convergence of well-tempered metadynamics**. a): A plot of the convergence of the 2D free energy surface in a single dihedral projection (dihedral angle $\alpha$) through the course of the 1.2 µs simulation. b) Plot of the cumulative negative bias deposition over time, along dihedral angle $\alpha$ during the same simulation. Metadynamics parameters: hill deposition pace of 500 steps, hill size of 0.3 kJ mol$^{-1}$, sigma of 0.3 radians, bias factor 6.

of the simulation time, and by tracking the cumulative bias deposited along the coordinate (Figure 3.10). It can be seen that the profile stabilises completely after 1000 ns of simulation time, whereupon the new biases deposited become very flat. This indicates that the free energy surface is converged at this point, allowing for estimating the relative energies of basins as well as barriers.

### 3.3.3 Umbrella Sampling Simulations in the Membrane

To investigate the energetics of the translocation step, the precedent set by Felix and co-workers in their work on $Cl^-$ transporters with the Gale group was followed.[71, 72] Due to the high computational cost of umbrella sampling, only the most selective of the studied transporters, **3.2**, was chosen, as well as the non-selective parent compound **3.1**, to serve as the points of comparison. Umbrella sampling allows the estimation of free energy profiles along a chosen coordinate by restraining the positions of molecules inside so-called umbrellas (Figure 3.11b), which are harmonic restraints that constrain the sampling to a particular region of the coordinate space.[174, 175] By performing a weighted histogram analysis (so-called WHAM), it is possible to infer the free energy from the positions of

the molecule within the umbrella-constrained space. To construct the free energy profile across several umbrellas, the umbrellas must overlap sufficiently. For convergence to be achieved, the sampling must be run for long enough for all of the degrees of freedom orthogonal to the restraining coordinate to be sufficiently explored. The restraining force of each umbrella, the umbrella spacing, and the length of time for which each trajectory is run are therefore key parameters.

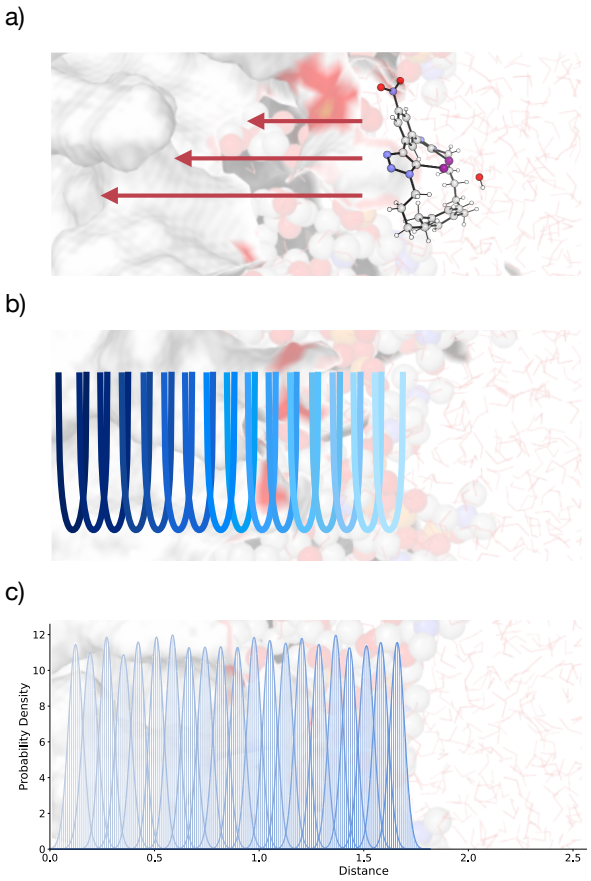



**Figure 3.11: Umbrella sampling coordinate details.** a) Pulling and umbrella sampling coordinate of [**3.2**·OH]$^-$, from the membrane-water interface at ~1.7 nm, to the membrane core at 0 nm. b) Schematic showing the steepness and the distribution of the harmonic restraints placed on the position of the complex at equal distances across the coordinate. c) Histograms of the [**3.2**·OH]$^-$ complex' positions during each of the 21 umbrella sampling runs across the membrane.

Following RESP fitting (*vide supra*), and GAFF parameterisation, the [transporter·ion]$^-$

complexes were restrained at their bound geometries using additional harmonic bonding terms between the iodines and the anions. This covalent model assumes that the bound ion's position does not change drastically during the transport process through the middle of the membrane. The complexes were then inserted into the core of a pre-equilibrated POPC membrane modelled using the Stockholm lipid parameters, with surrounding water slabs modelled by TIP3P-fb water molecules. Following energy minimisation and a short NPT equilibration (298 K, 1 bar, semi-isotropic coupling) at the membrane center, the molecules were then subjected to constant force pulling using a position restraint of 2000 kJ $mol^{-1}$ $nm^{-2}$ along the membrane normal (3.11a). To obtain the starting structures for the umbrella sampling trajectories, frames spaced 0.08 nm were taken from this pulling trajectory, from the membrane core (0 nm) to the water boundary (1.68 nm). Each of the 21 umbrellas (Figure 3.11b) was first equilibrated for 5 ns (same conditions as initial equilibration) and was then subjected to a production simulation for a total of 100 ns until convergence was achieved, with three repeats carried out for each of the transporter and ion combinations (Figure 3.12, top four pannels).

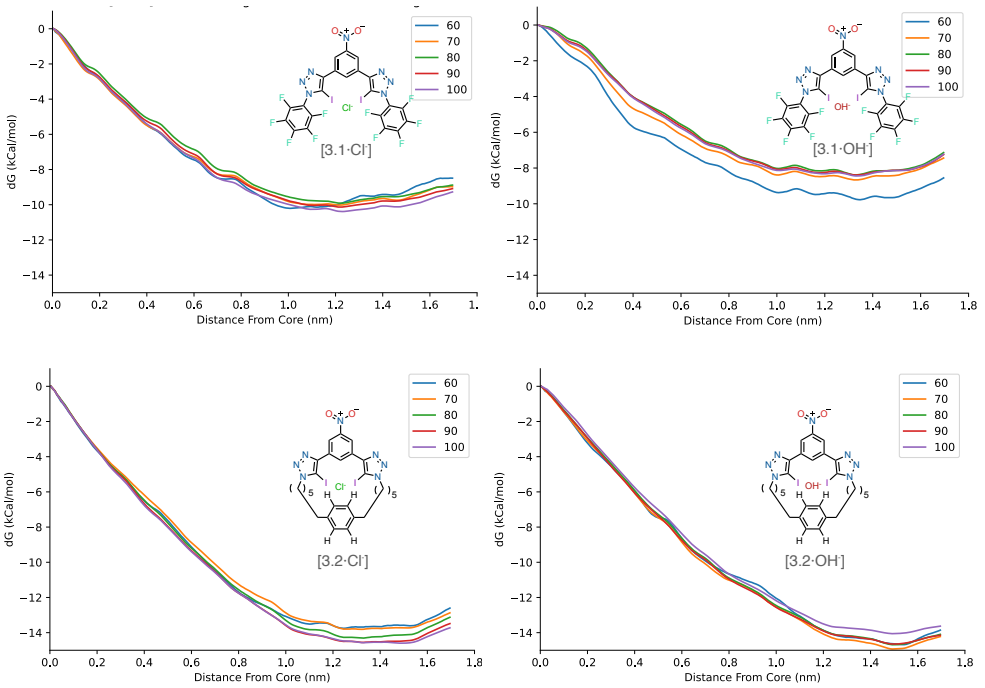
The barriers for translocation to the membrane core were calculated from the umbrella sampling runs and compared to estimate the selectivity of the translocation step, which has previously also been shown to be the rate-limiting step in unpublished work from the Duarte group. In the case of the non-selective **3.1**, it was found that the translocation of the $Cl^-$ complex had a higher barrier than that of the $OH^-$ complex, with energies of 9.6 ± 0.4 and 7.6 ± 0.4, respectively. In the case of the selective macrocycle **3.2**, the barrier ordering was found to be reversed, with the $Cl^-$ complex having a lower barrier of 13.7 ± 0.5 than the $OH^-$ complex, which had a barrier of 14.3 ± 0.6 kcal $mol^{-1}$. This correlates well with the experimental results - the less selective transporter is able to transport $OH^-$ much better relative to the $Cl^-$ than the selective transporter - but only in a qualitative

sense. If the selectivities were determined solely by this membrane translocation step, then neither **3.1** nor **3.2** would be very selective at all, given the barrier for $OH^-$ is almost level with that of the $Cl^-$ in the case of **3.2**, and even lower in the case of **3.1**. It would therefore be expected that no selectivity is observed for either transporter experimentally.

Additionally, the errors associated with the sampling mean that the difference between the two $\Delta\Delta G$ values is not likely to be statistically significant. Another piece of evidence that suggests that not everything is determined by the barrier of translocation in this case are the differences in barrier heights for $Cl^-$ transport between the smaller acyclic **3.1** and the larger macrocyclic **3.2**. Experimentally, the differences in their $EC_{50}$ values in the $Cl^-$ transport rate-limiting regime (with FCCP added), were found to be only an order of magnitude apart, but the barrier heights here indicate a difference of 4.1 ± 0.9 kcal $mol^{-1}$. This would lead to a much more drastic difference in the transport activities by simple Boltzmann-weighted extrapolation.

### 3.3.4 Unbiased Molecular Dynamics Simulations

The two ions, $Cl^-$ and $OH^-$ differ in many important aspects, but one key way in which they differ is their size and charge density. The much harder nature of the $OH^-$, stemming from its smaller radius and larger charge density, gives it very different hydration properties to the larger and relatively softer $Cl^-$. Their hydrogen bonding abilities also differ - while the $Cl^-$ can only act as a hydrogen bond acceptor, the $OH^-$ can act both as a donor and an acceptor, which also contributes to its much higher hydration energy. The free energies of hydration have been determined to be -105.0 kcal mol$^-$1 for the hydroxide and -74.6 kcal $mol^{-1}$ for the chloride.[176, 177] This means that a much greater energetic penalty is paid for desolvating the $OH^-$, and that enforcing desolvation could be a key




| Compound | $\Delta G_{translocation}$ ($Cl^-$) | $\Delta G_{translocation}$ ($OH^-$) | $\Delta\Delta G$ ($OH^-$ vs $Cl^-$) |
|---|---|---|---|
| 3.1 | 9.6 ± 0.4 | 7.6 ± 0.4 | -2.0 ± 0.8 |
| 3.2 | 13.7 ± 0.5 | 14.3 ± 0.6 | 0.6 ± 1.1 |

**Figure 3.12: Results of the umbrella sampling simulations.** The four panels show the WHAM free energy surfaces for time intervals from 0 to 60, 70, 80, 90 and 100 ns, respectively, for each of the transporter and ion complexes across three repeats. Uncertainties were estimated using WHAM analysis of the combined uncorrelated repeats. Summaries of the barriers for translocation to the membrane core are shown in the table below. $\Delta\Delta G$ is calculated as $\Delta G_{translocation(OH^-)}$ - $\Delta G_{translocation(Cl^-)}$, meaning that a lower (or more negative) number indicates more favourable $OH^-$ transport.

pathway for favouring $Cl^-$ transport, which is considered by neither the DFT binding model with implicit solvation nor the membrane translocation modelling of the [transporter·ion]$^-$ complexes. While literature reports have speculated that the differing hardness of the ions is one of the mechanistic origins behind the selectivity gains afforded by encapsulation, no modelling work has thus far been undertaken to demonstrate this. [50]

To compare differences between the carriers affect the hydration of the anions at the interface, equilibrium MD was performed for the complexes $[\mathbf{3.1} \cdot X^-]^-$ and $[\mathbf{3.2} \cdot X]^-$ with $X^- = Cl^-, OH^-$. Each system was prepared by placing the DFT-optimised structures in the core of the membrane and the ion positions were restrained to their DFT-optimised poses. The systems were equlibrated, and the complexes were then pulled towards the membrane-water interface, analogously to the umbrella sampling preparation (*vide supra*). Following 5 ns of equilibration, three independent production trajectories of 100 ns were produced for each system. This protocol allows an estimation of the hydration in the bound pose without directly interrogating the binding energetics.

Radial distribution functions $g(r)$ between water oxygens and the anion heavy atom were computed from each replica and averaged, to give the plots in Figure 3.13a. The first hydration number $n_1$ was obtained by integrating $g(r)$ up to the first minimum $r_{min}$. The functions show that the maximum of this first hydration shell is much tighter for the hydroxide, which is to be expected given its smaller size and stronger interactions with water. It is also much higher than the maximum of the $Cl^-$ ion, which is also to be expected due to the same reasons. Upon cyclisation, both of the maxima experience a decrease, leading to a change in the number of waters in the first hydration shell (Figure 3.13b). For $Cl^-$, $n_1$ decreases from 2.1 in **3.1** to 1.7 in **3.2**, representing a 19% decrease, while in the case of the $OH^-$ the decrease in $n_1$ is from 2.6 to 1.9, representing a 27% reduction. The

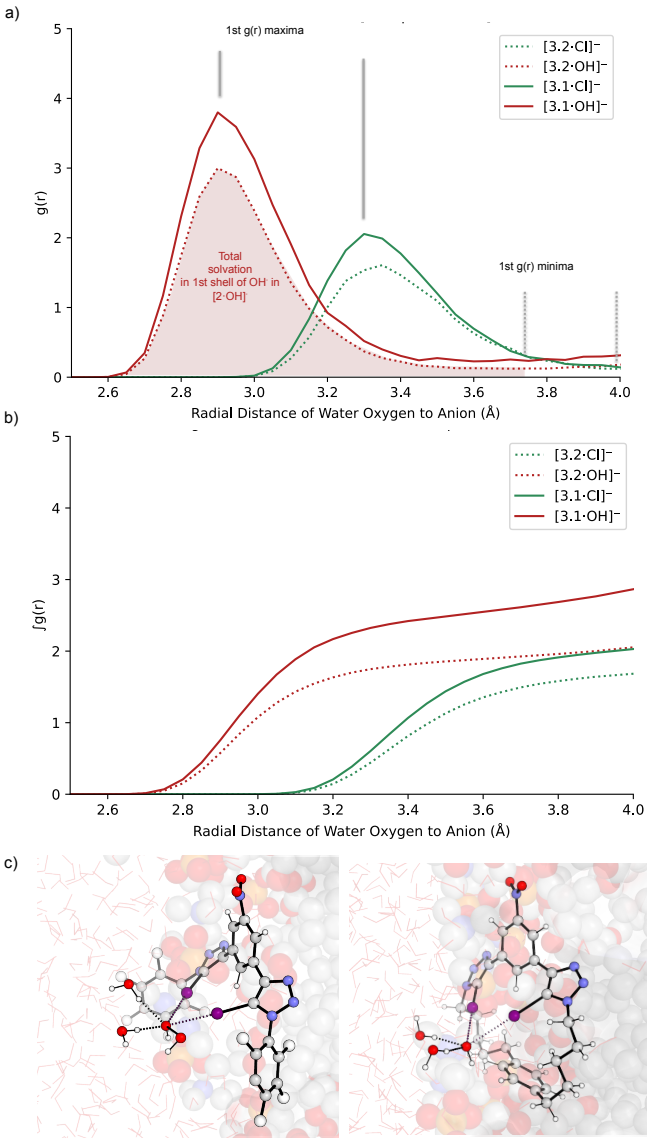



**Figure 3.13: Hydration at the membrane interface from equilibrium MD.** a) Water–anion radial distribution functions $g(r)$ for the bound complexes of the acyclic parent **3.1** (solid) and the macrocycle **3.2** (dotted) with $Cl^-$ (green) and $OH^-$ (red). First maxima and integration limits to the first minimum are indicated; the shaded area illustrates the first–shell coordination for $[\mathbf{3.2}\cdot OH]^-$. b) Integrals of the radial distribution function up to the first minimum (first solvation shell) $\int g(r)$ c) Representative snapshots at the POPC interface showing reduced first–shell waters in $[\mathbf{3.2}\cdot OH]^-$ relative to $[\mathbf{3.1}\cdot OH]^-$.

increased desolvation (Figure 3.13), will likely discourage $OH^-$ binding especially, leading to increased $Cl^-$ preference in binding. As literature studies, as well the observations from US simulations have shown, a number of water molecules are carried alongside the ion across the membrane during the translocation step, meaning that the desolvation observed here could have far reaching consequences across the entire transport process.[71, 72]

## 3.4 Conclusions and Future Work

In this chapter the origins of selectivity were probed in a coherent, multi-method and -scale approach in order to assemble a picture of all possible steps in the transport process that could be imposing energetic discrimination between the two ions of interest ($Cl^-$ and $OH^-$). This was first done by considering the inherent transporter-anion interactions without explicit interactions with other species in DFT calculations using membrane-mimetic implicit solvation. These studies showed that cyclisation decreases the intrinsic $OH^-$ preference of the acyclic parent. Distortion-interaction analysis showed that this happens by way of decreasing the interaction strength with the hydroxide relative to the chloride upon cyclisation, while the distortion remains constant. Additionally, increased ordering of the vibrational normal modes upon $OH^-$ binding affords this ion further energetic penalties, favouring the $Cl^-$ further. This all suggests that the macrocycles are uniquely preorganised for $Cl^-$ over $OH^-$ bonding, leveraging the differing sizes of the two anions.

Well–tempered metadynamics in explicit solvent allowed a mapping of the conformational landscape of the most selective macrocycle and demonstrated facile access to convergent *-cis* states via low dihedral rotation barriers, suggesting a dynamically preorganised binding site. Umbrella sampling of membrane translo-

cation were undertaken, which also exhibited the correct energy ordering with respect to the experimentally observed trends, but with differences found to be within the error margins of the method. Classical MD at the POPC interface, utilising an explicit $\sigma$–hole model for iodine allowed a closer look at the behaviour of the macrocycles in the experimentally relevant environment - both qualitatively and quantitatively. Here, it was seen that the macrocycle reduces first–shell hydration of both anions, with a larger desolvation penalty for $OH^-$, rationalising the observed $Cl^- > OH^-$ selectivity.

In order to probe whether different selectivity mechanisms may operate within different classes of compounds, the same workflow could be applied to other literature examples or variants of the scaffolds studied here. For example, larger and more encapsulating variants of the present macrocycles could be tested, as well as hydrogen-bonding analogues analogues to disentangle preorganisation from bonding type. The recently published calixarenes from the Valkenier group may also be an interesting computational case study for direct energetic prediction, given their higher levels of rigidity would likely simplify a lot of the aspects of the modelling.[53] By way of studying numerous compounds through the approaches described in this chapter: systematic exploration of $\Delta\Delta G$ comparisons at the binding site, interfacial hydration comparisons, as well as further attempts at performing pulling simulations across the membrane, it may be possible to further improve the design rules for $Cl^-$-over-$OH^-$ selectivity. This could ultimately help with transitioning these compounds towards even more biological study.

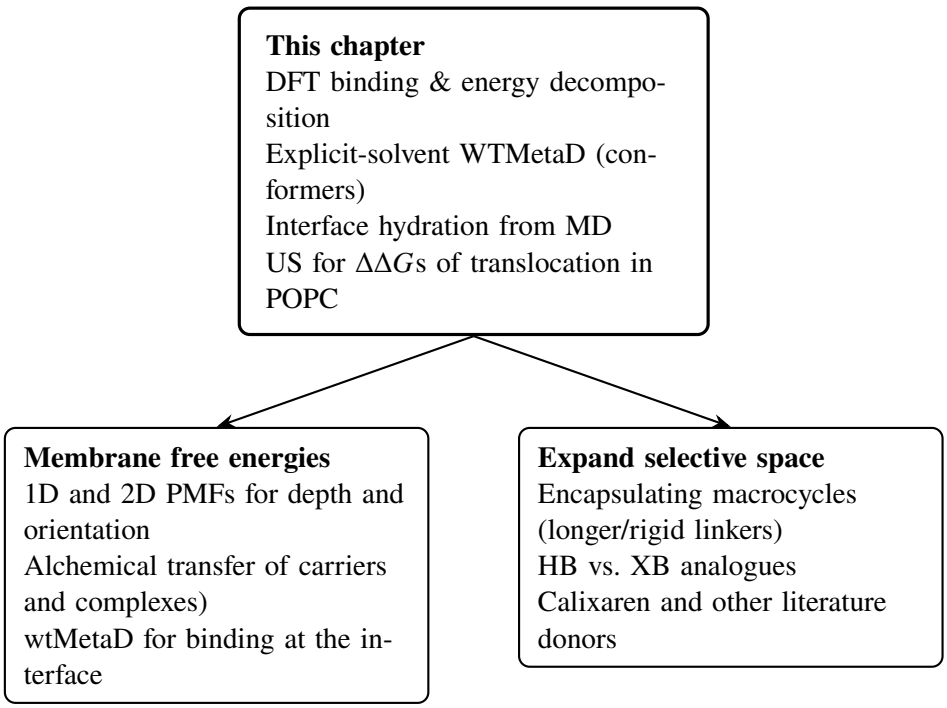



**Figure 3.14:** Text schematic of key conclusions and possible directions.

The membrane translocation and binding steps could both be quantified with rigorous pathway and alchemical free energy calculations, provided that halogen bonding parameterisation can be carried out in such a way as to reliable reproduce both $Cl^-$ and $OH^-$ binding energies and distances in the gas phase. Performing 2D metadynamics of membrane insertion distances and angles to the membrane normal for the complexes would additionally add more detail about the possible pathways that different transporters can take.

Focusing first on only the transporter-membrane interaction, and carrying out alchemical transfer free energy calculations for the free carrier's insertion from the water to the bilayer followed by that of the bound complex may help deconvolute the delivery steps from the binding. WTMetaD for ion association at the interface would additionally be incredibly helpful in characterising binding, provided the parameterisation of both the transporters and ions can be rigorously benchmarked.

For all future work, attempts should be made to frame the investigations in terms of assayable observables: beyond the transport and binding experiments undertaken, logP measurements would be an additional useful benchmark for parameters.

# 4 |

# Stimuli-Responsive Transporter Modelling

In this chapter, two families of stimuli-responsive caged transporters synthesised and characterised by Dr. Manzoor Ahmad in the Langton group are studied computationally. The isophthalamide transporter class is studied first, rationalising the stoichiometries of binding and transport observed experimentally using DFT modelling. Attention is then moved to the hydrazone-based transporter class, where the analysis is expanded to decompose the energetics, model association in explicit solvent and provide comparisons of energy binding estimates across several computational methods. Both DFT and MD provide good descriptions of the hydrazone-based class, but the strengths of the explicit solvent dynamics are found to give it an advantage over DFT methods. The work in this chapter was published in *Angewandte Chemie* and *Nanoscale* in 2024.[178, 179]

## 4.1 Isophthalamide-based Transporters

The isophthalamide motif was first developed as a chloride transporter by the Quesada group in 2007 (Figure 4.1a).[180] They found that varying the substitution patterns on the *ortho-* position to the aromatic amide substituent led to a difference in the hydrogen bonding pattern, and thus the conformation of the amide groups. Solid state structures showed that introducing a hydroxyl next to the amide of compound **4.1**, to give transporter **4.2**, changed the *syn-anti* conformation to a *syn-syn* conformation. This was confirmed to hold true in solution, as evidenced by $^{1}$H NMR signals of the hydroxyls being shifted far downfield. Upon capping of

the phenolic position with methyl groups in compound **4.3**, the hydrogen bonding pattern inverts, with the oxygen changing from being a hydrogen bond donor to being a hydrogen bond acceptor, which results in an *anti-anti* conformation. This locks both of the acidic amide protons into an intramolecular hydrogen bond, as was supported by the solid state structures as well as the characteristic NMR downfield shifts, and thus precludes their ability to bind an ion.

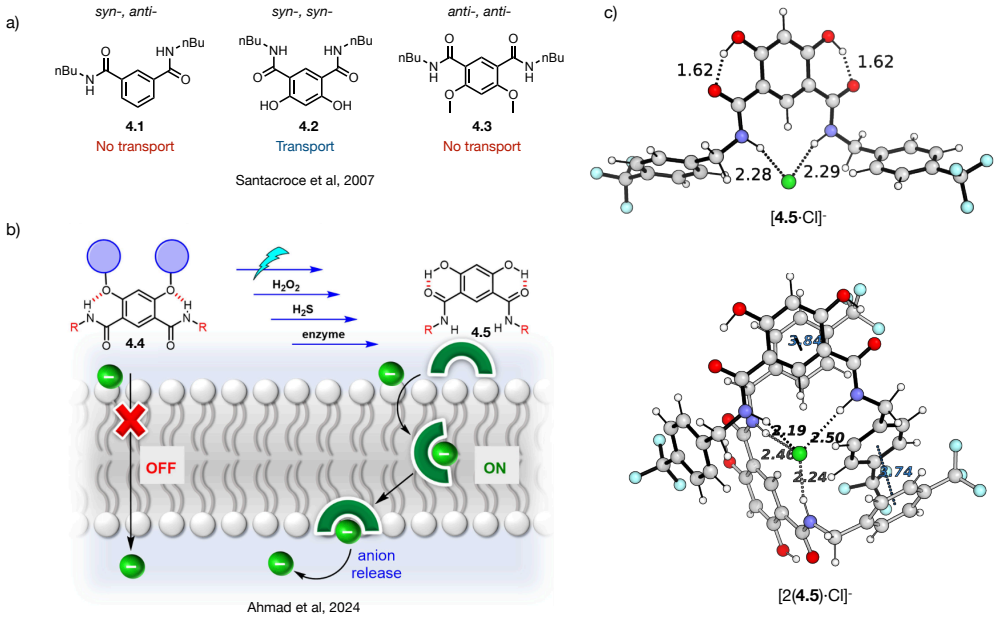



**Figure 4.1: Literature precedent of isophthalamides as transporters and the present system.** a) The three different isosphthalamide derivatives developed by Santacroce *et al.*, with the *syn-anti*, *syn-syn*, and *anti-anti* conformations held by the two amide groups, depending on the substitution of the aromatic ring, leading to differing abilities to transport chloride ions.[180] b) The design of the different isophthalamide derivatives discussed in this chapter, synthesised and tested by Dr. Ahmad, and the various stimuli that could be used to deprotect them, including: light, oxidation, reduction, enzymatic degradation. Without stimuli, the transporters were inactive, whereas following their deprotection, transport was observed. Adapted with journal permissions.[178] c) Structures of deprotected isophthalamide transporter **4.5**'s 2:1 and 1:1 complexes with the $Cl^-$ ion optimised at the $\omega$B97X-D3/def2-SVP level of theory.

By recognising that these phenolic positions had relatively low p*K*a values of around 8, Santacroce *et al.* saw an opportunity in being able to use varying protonation states to control transport.[180] The *syn-anti* conformation of **4.1**

and the *anti-anti* conformation of **4.3** were both found to result in ineffective transporters, due to the conformations not being preorganised for binding. At a pH of 6.4, **4.2** was found to display good transport properties, whereas increasing the pH gradually to 9.1 led to almost complete diminishing of transport in a pH-dependent manner. This showed that the dihydroxyisophthalamide motif could be manipulated *in situ* to switch membrane transport on and off.

### 4.1.1 Design, Synthesis and Transport*

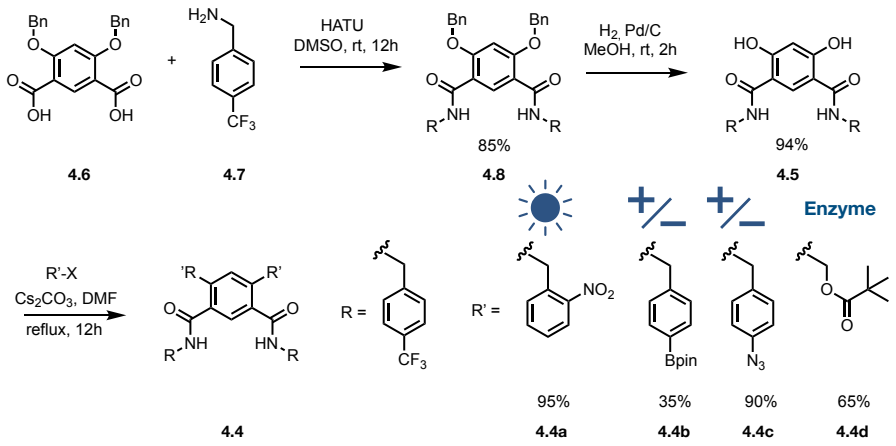



**Figure 4.2:** The synthesis of the trifluororomethylbenzyl substituted isophthalamide ionophores, and their protection with four different protecting groups, each offering a different avenue for deprotection: light, oxidation, reduction, and enzyme hydsolysis. Adapted with journal permissions.[178] Synthetic work was carried out by Dr. Manzoor Ahmad.

The scaffold designed originally by Santacroce *et al.* served as a starting point for our work on developing the isophthalic acid derivatives discussed in this chapter. The first family of these that is discussed contained the same isophthalamide motif as the ion-binding core.[178] Initial work focused on testing a variety of substituents on the amide nitrogen, in order to optimise for both lipophilicity and anion-binding ability. It was found that a *p*-trifluoromethylbenzyl substituent

*Experimental work carried out by Dr. Manzoor Ahmad.

performed best in both anion binding and chloride transport efficacy, and **4.4** was thus used as the active species. It was synthesised by coupling the benzyil-protected dihydroxyisophthalic acid **4.6** with trifluoromethylbenzylamine **4.7**, to give the bis-amide **4.8**. This was then deprotected in near-quantitative yields to give **4.5** which was then protected with various cleavable functional groups to give **4.4a**-**4.4d**. A light-cleavable *ortho*-nitrobenzyl group on **4.4a** was explored for potential photopharmacological applications. The other three protecting groups all addressed activation by species which are found to be overexpressed in cancer cells, with $H_2O_2$ used to cleave the Bpin ester **4.4b**, the phenyl azide **4.4c** cleaved by $H_2S$ reduction and the *t*-Bu ester in **4.4d** cleaved by an esterase enzyme.

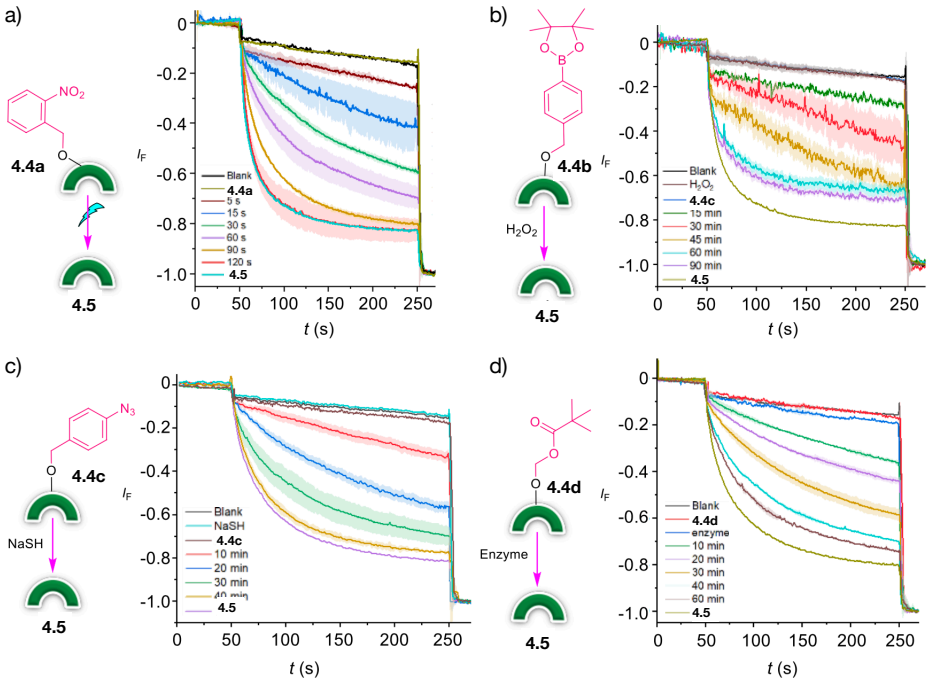



**Figure 4.3: Ion transport activities following specific activation conditions**. Activities were measured across POPC LUVs loaded with lucigenin. a) **4.4a** was activated by photo-irradiation at 405 nm using a 1 W LED. b) **4.4b** was treated with $H_2O_2$ (50 eq.). c) **4.4c** was reacted with NaSH (10 eq.). d) **4.4d** was treated with porcine liver esterase enzyme. Assay work was conducted by Dr. Manzoor Ahmad.

Figure 4.3 shows how various activation times with the respective stimuli affected

the $Cl^-$ transport activity in POPC large unilamellar vesicles (LUVs). Initially all protecting groups effectively prevented transport. The photo-deprotection of **4.4a** proceeded after mere seconds, reaching the full activity of the deprotected **4.5** transporter after a total of 120s. Meanwhile the boronic ester **4.4b** took the longest time to be decaged, with the activity of the deprotected **4.5** not fully achieved even after 90 minutes of deprotection with $H_2O_2$. Overall, these results demonstrated that a variety of different biocompatible groups can be attached to the phenolic oxygens to impede transport, and that they can effectively be removed in-situ with their respective stimuli in order to switch on the transport.

### 4.1.2 Computational Study

Interestingly, the Hill plot analysis of the $Cl^-$ transport properties of **4.5** revealed a Hill coefficient of 2, while the binding isotherm itself was best described by a simple 1:1 binding model. At first glance, this appears contradictory. It suggests that in solution, the receptor binds chloride as a monomer, yet during transport through the membrane the process seems to require the cooperative action of two receptor molecules. A possible explanation for this is that although the monomer is the dominant species in bulk solution, once embedded in the membrane the receptor can assemble into dimers or higher-order aggregates that work together to carry chloride across. To explore this possibility, Density Functional Theory (DFT) calculations were performed to model both the 1:1 [**4.5**·Cl]$^-$ complex and the 2:1 [2(**4.5**)·Cl]$^-$ complex in the gas phase and in solution (final optimised structures in 4.1c).

Following a conformational search carried out using CREST[159], and a refinement of their single point energies using XTB, the lowest energy conformers were optimised at the $\omega$B97X-D3/def2-SVP and final single points carried out at the $\omega$B97X-D3/def2-QZVP level of theory (the workflow followed is the same as de-

scribed in Section 3.2.2). The DFT optimisations were then carried out both with implicit solvation and without. The results of these calculations are summarized in Table 4.1. In DMSO, the 2:1 complex was predicted to be significantly more stable than the 1:1 complex, with a free energy change of –14.9 kcal mol$^{-1}$ compared to –3.4 kcal mol$^{-1}$. In purely energetic terms, this suggests that the dimeric complex is the dominant species in solution. This contradicts the experimentally found binding isotherm. The most likely reason for the discrepancy between the computational and experimental results lies in the way solvation was modelled. As discussed in Section 1.5.1, a persistent weakness of continuum solvation methods is that the approach treats the solvent only as a bulk dielectric medium. In this case, this means it cannot explicitly capture the strong interactions of DMSO with both chloride and the polar groups of the receptor. In reality, DMSO molecules are highly competitive hydrogen-bond acceptors and can effectively solvate the receptor's hydroxyl and amide groups, as well as chloride itself. Moreover, the polar aromatic rings of **4.5** are also well solvated by DMSO, which would destabilize $\pi$–$\pi$ stacking between receptor molecules (highlighted in Figure 4.1c). Due to the missing explicit solvent interactions, the calculations may over-stabilize the 2:1 complex by underestimating solvent competition.

**Table 4.1:** Calculated binding energies ($\Delta E$) and free energies ($\Delta G$).

| **Reaction** | **Phase** | $\Delta E$ (kcal/mol) | $\Delta G$ (kcal/mol) |
|---|---|---|---|
| Transporter + $Cl^-$ | Gas phase | -51.5 | -47.7 |
| $\rightarrow$ [Transporter·Cl]$^-$ | DMSO | -9.1 | -3.4 |
| 2Transporter + $Cl^-$ | Gas phase | -96.9 | -75.1 |
| $\rightarrow$ [2Transporter·Cl]$^-$ | DMSO | -39.5 | -14.9 |

The energetics in the gas phase were also examined in order to shed further light on the energetics. Both the 1:1 and 2:1 complexes were calculated to be highly favourable without any solvent consideration, as expected in the absence of ion desolvation penalties, with free energies of –47.7 and –75.1 kcal mol$^{-1}$, respec-

tively (Table 4.1). However, when the binding is considered stepwise, the first receptor binds chloride much more strongly than the second. The second binding event remains exergonic (–27.4 kcal $mol^{-1}$) but is noticeably weaker, reflecting the elongation of hydrogen bonds that occurs when two receptors compete to bind the same anion (Figure 4.1c). In the 1:1 complex, the hydrogen bonds are 2.28 Å and 2.29 Å long, respectively. In the 2:1 complex, on the other hand, one hydrogen bond on each transporter stays short (2.19 and 2.24 Å), while the other elongates (2.46 and 2.50 Å). This picture is consistent with the $Cl^-$-binding NMR titrations, which show clean 1:1 binding in solution, but it also illustrates how a 2:1 capsule can form under favourable conditions, which is what is suggested by the Hill constant found in transport assays.

Although the quantitative agreement with experiment is imperfect, the insights from the DFT modelling are very useful. As illustrated in Figure 4.1c, two molecules of **4.5** can come together around chloride, forming a capsule-like complex stabilized by hydrogen bonds and $\pi$–$\pi$ interactions. In this arrangement, the chloride is more deeply encapsulated and shielded from its environment than in the 1:1 complex. This encapsulation is likely to be critical in the context of transport, because it lowers the energetic penalty for moving the anion through the hydrophobic membrane interior to enhance activity, as has been demonstrated in literature studies.[4, 143] In this way, the computational results can be reconciled with experimental findings. Despite the monomer dominating chloride binding in bulk solution, the cooperative formation of a 2:1 transport-active complex in the membrane provides the mechanism for the observed Hill coefficient of 2.

## 4.2 Hydrazide-based Transporters

Initial work with amide-based transporters utilised intramolecular phenol–carbonyl hydrogen bonds to preorganize the binding site and demomstrated that it could be locked or unlocked through caging strategies to gate transport. Building on this, attention was next turned to hydrazone-based analogues of the best-performing trifluoromethylbenzyl substituted amide carrier (Figure 4.4a). Replacing the amide linkage with an acyl hydrazone introduces additional N–H and C–H hydrogen bond donors, which enhances chloride binding affinity and significantly improves transport activity. Indeed, the hydrazone derivatives consistently outperformed their amide counterparts, with the most active transporter showing almost a three-fold increase in activity compared to the amide system, a difference attributed to the stronger and more convergent hydrogen bonding interactions of the hydrazone motif.[179] Hill analysis of transport dose-response curves in LUV fluorescence assays of the most active hydrazone carriers again revealed coefficients close to 2, suggesting that two transporter molecules cooperate in the transport process, consistent with a 2:1 transporter-to-anion stoichiometry within the membrane.

Moreover, the hydrazone systems were also successfully integrated with light- and $H_2S$-responsive cages, providing robust OFF–ON control with minimal background activity in the OFF state (Figure 4.4b). The subtle differences between the the two transporter families, combined with the limitations of the modelling experienced when looking at the amide family, made the hydrazones an ideal case study for further computational analysis. This was a motivation to explore the thermodynamics of chloride binding in greater detail.

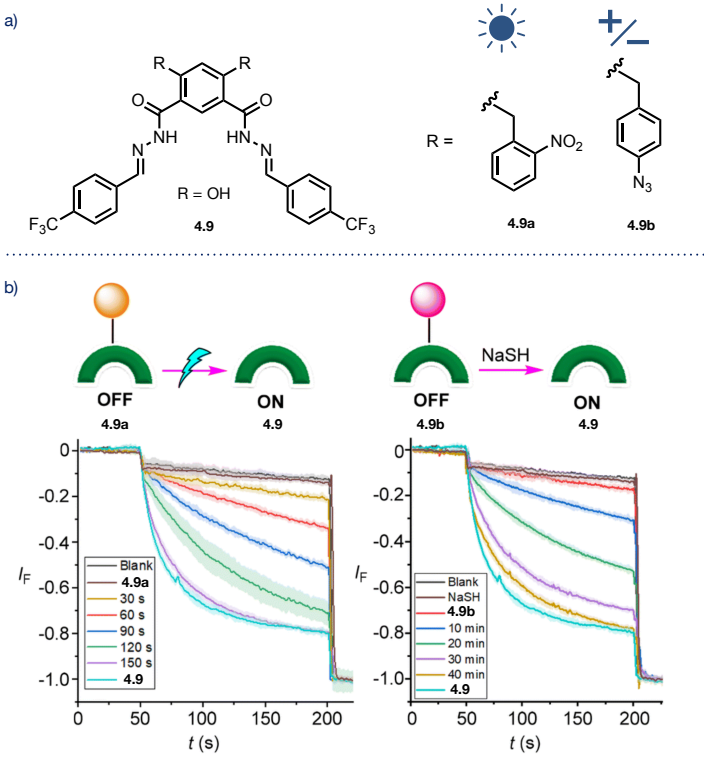



**Figure 4.4: Introduction of the hydrazone-based transporter class.**a) Structures of the hydrazone transporter **4.9** and its protected analogues **4.9a** and **4.9b**, with representative light- and $H_2S$-responsive caging groups. b) Activities after deprotecting the caged transporters were measured across POPC LUVs loaded with lucigenin. Left: **4.9a** was activated by photo-irradiation at 405 nm using a 1 W LED. Right: **4.9b** was treated with NaSH (10 eq.). Reproduced with journal permissions.[179] Synthetic and assay work was conducted by Dr. Manzoor Ahmad.

### 4.2.1 DFT Analysis

To better understand the difference between the active hydrazone transporter and its protected analogues, both **4.9** and a truncated version of the protected molecules were modelled. In the truncated protected molecule, methyl groups were used to cap the phenolic oxygens (**4.9m**). DFT calculations were performed to examine their binding interactions with chloride, and distortion-interaction analysis used to examine the energetic origins of the diminished binding and transport in the protected species. As shown in Figure 4.5a, the unprotected hydrazone **4.9**

features a strong intramolecular O–H···O hydrogen bond in its minimum energy conformation. This interaction has a calculated stabilisation energy of 80 kcal $mol^{-1}$ according to second order perturbation analysis carried out with the NBO package.[181] This preorganisation is retained upon chloride binding, where its energetic contribution remains similar at 78 kcal $mol^{-1}$. In this bound pose the hydrazone N-H, imine C–H and the aromatic C-H protons all converge on the anion to deliver a an interaction energy of 52 kcal $mol^{-1}$. Together, the binding event in **4.9** corresponds to a free energy of –2.5 kcal $mol^{-1}$.

By contrast, in the methylated analogue **4.9m**, the intramolecular hydrogen bond is substantially weaker, with a value of 20 kcal $mol^{-1}$ for N–H···O. This reflects relative hydrogen bond donor and acceptor abilities of the N and O atoms. While N is a much better H-bond donor, O is the much better acceptor. Due to the phenolic oxygen protection, the only hydrogen bonding available in **4.9m** occupies the N-H protons in an intramolecular interaction which must be broken for the $Cl^-$ ion to bind. This, along with a slightly reduced sum of all of the proton-$Cl^-$ interactions leads to the binding of chloride being endergonic (2.1 kcal $mol^{-1}$). In other words, while chloride can still interact with hydrazone N-H and the C–H donors in **4.9m**, the energetic penalty associated with reorganizing its geometry into a conformation available for binding outweighs the competing stabilisation provided by the hydrogen bonds.

Figure 4.5b shows the distortion-interaction analysis of the binding free energies (first introduced in section 3.2.4, which further helps to illustrate the sum of energetic contributions. For **4.9**, the distortion energy required to reorganize the host into the bound conformation is small, totalling only 1.6 kcal $mol^{-1}$, while the interaction energy between the preorganized host and chloride is strongly favorable –4.1 kcal $mol^{-1}$. In contrast, for **4.9m** the distortion cost is much larger at 5.7 kcal $mol^{-1}$. Although the interaction energy is only slightly weaker than in the

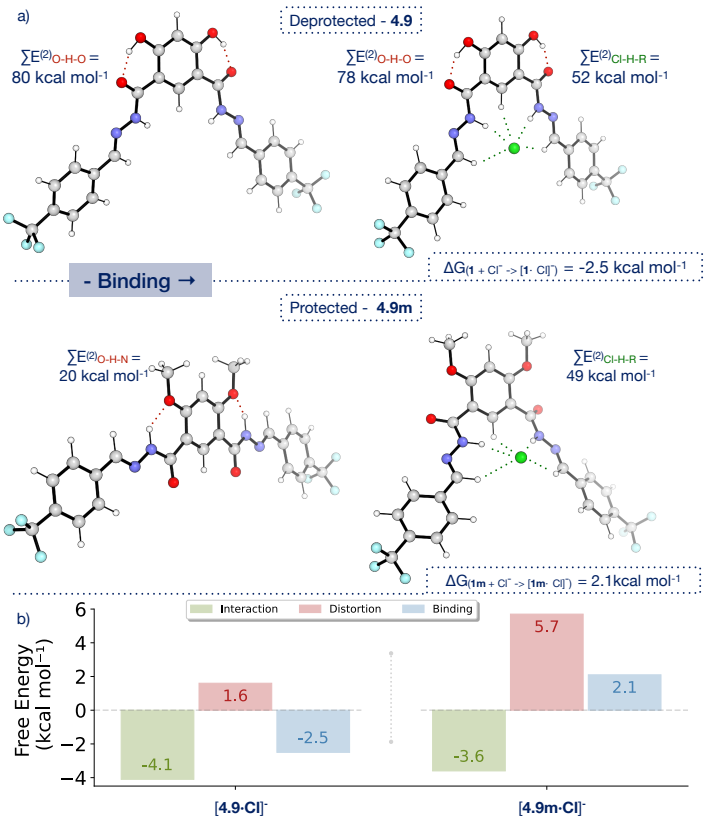



**Figure 4.5: Density Functional Theory modelling results for the hydrazone transporters.** a) DFT-optimised geometries of the deprotected **4.9** and analogue with a truncated methyl protecting group **4.9m**, both free and bound to chloride. Key intramolecular hydrogen-bonding interactions are indicated and the total second order perturbation analysis energy contributions of these interactions are displayed as$\Sigma E^{(2)}$). The calculated binding free energies ($\Delta G$) are also shown. c) Distortion-interaction analysis of chloride binding in **4.9** and **4.9m**.

deprotected variant at –3.6 kcal mol$^{-1}$, the net result is a positive binding free energy. This reinforces the experimental results, shedding light on the energetic contributions which make intramolecular hydrogen-bonding network necessary for preorganisation.

To further dissect the origins of this difference in binding behaviour, second-order perturbation analysis of natural bond orbitals (NBO) was carried out and is dissected in more detail here. This analysis was conducted on both the free receptors, as well as their chloride and DMSO complexes (Figure 4.6). This approach quan-

tifies donor–acceptor interactions as second order perturbation energies $E^{(2)}$. This energy reflect the stabilisation arising from interactions between filled and empty orbitals, and provides a useful way to compare intra- and intermolecular hydrogen bonding contributions across the series.

As outlined earlier, the free receptor **4.9**, the dominant stabilising contribution comes from the intramolecular O–H···O interaction, with a $\Sigma E^{(2)}$ of 80 kcal mol$^{-1}$. Crucially, this interaction remains essentially intact upon chloride binding ($\Sigma E^{(2)} \approx 78$ kcal mol$^{-1}$), which shows that the preorganised hydrogen-bonding scaffold survives complexation. In the chloride adduct [**4.9**·Cl]$^-$, additional strong intermolecular contributions are observed, particularly from the hydrazone N–H···Cl interaction ($\sim$ 42 kcal mol$^{-1}$), supplemented by weaker $C_{Ar}$–H···Cl and $C_{Ben}$–H···Cl contacts (a few kcal mol$^{-1}$ each). The involvement of the benzylic imine protons ($C_{Ben}$–H) is postulated to be one of the reasons for the hydrazone displaying higher binding and transport activities than its amide predecessor.

By contrast, in the competing [**4.9**·DMSO] complex the donor–acceptor interactions of these protons are are negligible ($\sim$ 1–2 kcal mol$^{-1}$ in total), while the hydrogen bonding contribution of the most highly energetic N–H···X interaction diminishes less dramatically upon DMSO binding (from 42 to 34 kcal mol$^{-1}$). This suggests that the smaller oxygen atom in the DMSO molecule is unable to sustain all of the hydrogen-bonding contacts to the hydrazide binder as well as the larger and more diffuse chloride ion, and sacrifices the less energetic C-H contacts to optimise the N-H interactions. The intramolecular hydrogen bonding O-H···O contact also slightly diminishes in the DMSO complex relative to the Cl$^-$ complex (down to 67 from 78 kcal mol$^{-1}$). This suggests that the hydrogen bonding in the minimum energy state preorganises the molecule much better for binding the larger Cl$^-$ anion than the smaller oxygen atom of the DMSO.

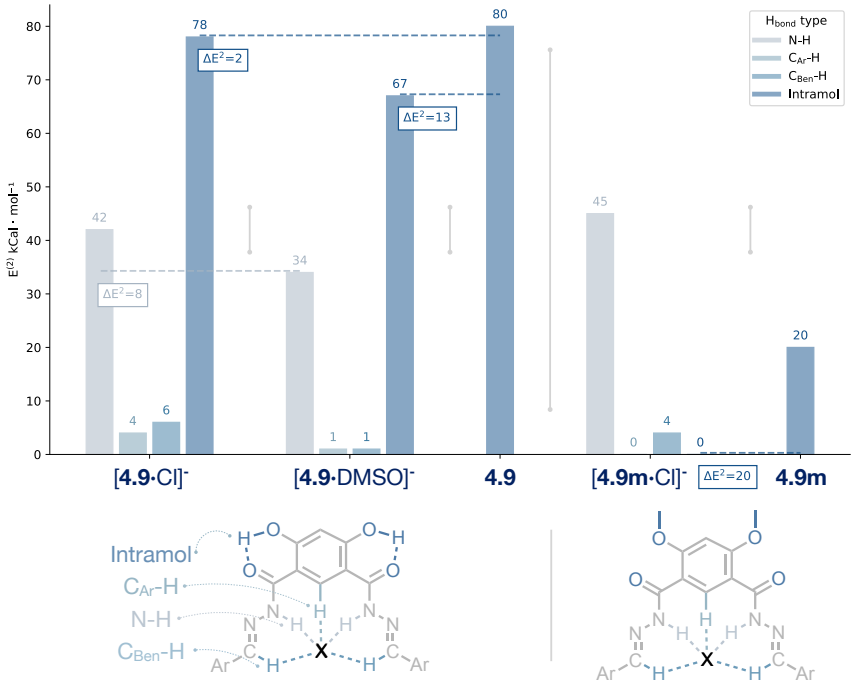



**Figure 4.6:** Second-order perturbation (NBO) analysis of key hydrogen-bonding interactions for the free receptors **4.9** and **4.9m**, and their complexes with chloride or DMSO. Bar plots show the summed $E^{(2)}$ contributions (kcal mol$^{-1}$) from intramolecular O-H···O or O-H···N interactions, as well as intermolecular N-H···Cl, $C_{Ar}$-H···Cl, and $C_{Ben}$-H···Cl contacts. Schematics below illustrate the hydrogen-bonding motifs identified in **4.9** (left) and **4.9m** (right).

In the methylated analogue **4.9m**, the picture is markedly different. The key O–H···N intramolecular interaction is much weaker ($\Sigma E^{(2)} \approx 20$ kcal mol$^{-1}$), and the transporter is preorganised in the wrong conformation, as the N-H protons are unavailable for binding. On chloride binding, a strong N–H···Cl interaction is still formed ($\sim$ 45 kcal mol$^{-1}$), but the overall stabilisation is offset by the loss of the intramolecular hydrogen bonding interactions and the significant geometric reorganisation required to accommodate the guest. The NBO results thus give important context to the distortion-interaction analysis. In **4.9**, the strong O–H···O framework allows the receptor to retain its internal stabilisation while gaining H-bonding interactions with the $Cl^-$ ion, whereas in **4.9m** the binding event

effectively trades one set of interactions for another, at high energetic cost.

### 4.2.2 Molecular Dynamics Simulations

The role of explicit solvent was further interrogated through molecular dynamics simulations, which both validate the DFT findings and expand them to the many-body environment of bulk DMSO. Whereas DFT predicted that binding of a single DMSO molecule to **4.9** was only slightly endergonic ($\Delta G$ = +0.8 kcal mol$^{-1}$), MD explicitly includes thousands of solvent molecules and thus captures the true competition between chloride and solvent. **4.9**, **4.9m** and the DMSO solvent were modelled using GAFF parameters. The transporters were inserted into a DMSO box and minimised using a gradient descent algorithm, and a short NPT equilibration (298 K, 1 bar, 2 ns) was conducted. NPT simulations were then conducted in GROMACS[173] for a total of 100 ns for each transporter.

Figure 4.7a shows violin plots of the intramolecular O–H···O hydrogen-bond distances analysed in GROMACS.[173] In **4.9** and the protected analogue **4.9m**. The distributions of the O-H···O bond in **4.9** indicate that the complex stays intramolecularly hydrogen bonded for the vast majority of the simulation time. This suggests that the preorganisation is indeed persistent even in bulk solvent. A similar observation can be made for the N-H···O hydrogen bond in **4.9m**. Here, the distribution indicates that the intramolecular hydrogen bond which blocks binding of the $Cl^-$ ion by occupying the acidic N-H protons is equally persistent in explicit DMSO. These hydrogen bond distribution support the energetic trends found by the DFT calculations and help explain the divergent binding thermodynamics. The preorganised cavity of **4.9** pays little distortion penalty, whereas **4.9m** must reorganise significantly to engage chloride.

The umbrella sampling free-energy profiles (Figure 4.7b) provide a more detailed

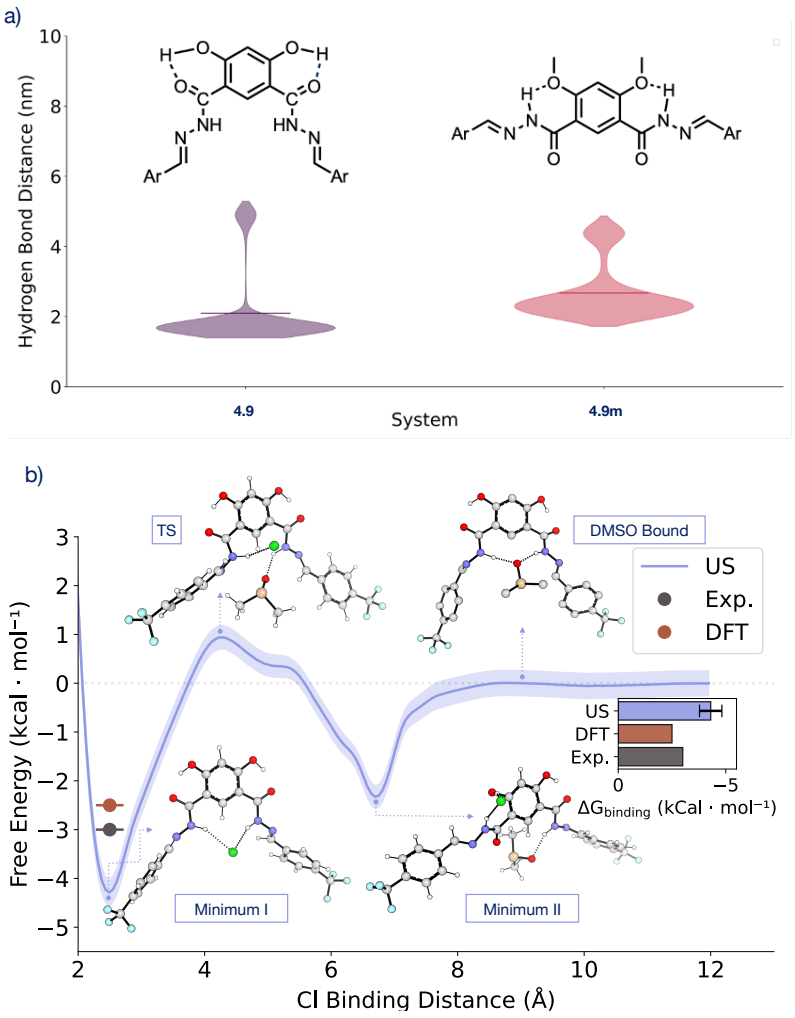



**Figure 4.7:** (a) Violin plots of intramolecular hydrogen bond distances from MD simulations of **4.9** and protected analogue **4.9m** in explicit DMSO, showing stable preorganisation in **4.9** and weaker, more labile interactions in **4.9m**. (b) Umbrella sampling free-energy profile for chloride binding to **4.9** in explicit DMSO, revealing two minima: Minimum I ($Cl^-$ bound to both hydrazides) and Minimum II ($Cl^-$ bound to one hydrazide and phenol, with DMSO insertion). The solvent-bound receptor state and transition region are also indicated. The inset compares binding free energies from umbrella sampling (US), DFT, and experiment.

mechanistic view of chloride association with the unprotected, activated transporter **4.9** in explicit DMSO. The distance between the $Cl^-$ anion and one of the two N-H binding groups was chosen as the pulling coordinate in the sampling. Starting frames for the umbrella sampling were obtained from constant force pulling simulations using a position restraint of 2000 kJ $mol^{-1}$ $nm^{-2}$ along the N-H-$Cl^-$ binding coordinate. The umbrella windows were spaced 1 Å apart, across distances spanning 2 to 12 Å, and each umbrella sampling window was run for a total of 100 ns until energy convergence was observed.

Two free-energy minima can be seen from the binding trajectory. Minimum I has a value of $\Delta G = -4.3 \pm 0.5$ kcal $mol^{-1}$ and can be found at a distance of ~2.4 Å to the acidic N-H. This minimum corresponds to chloride bound to both hydrazide donors in a geometry highly similar to the DFT-optimised structure. Minimum II, with a $\Delta G = -2.3 \pm 0.5$ kcal $mol^{-1}$, is found at the greater distance of ~2.4 Å. This minimum represents a solvent-inserted half-bound state for the anion. Here, the chloride is bound to one hydrazide and the neighbouring phenolic O-H (which is an interaction that was not observed in the DFT calculations), while a DMSO molecule inserts at the second hydrazide. The barrier between these minima corresponds to a transition region where DMSO begins to displace chloride. On full dissociation, the chloride ion is released and the intramolecular O–H···O hydrogen bond reforms, following which both hydrazides are solvated by DMSO molecules. This DMSO-bound receptor is roughly 4.3 kcal $mol^{-1}$ higher in energy than the chloride-bound state and persists at chloride distances greater than 9 Å from the binding site, consistent with the weak but appreciable solvent occupancy predicted by DFT and observed in unbiased MD trajectories.

The comparison of binding free energies by umbrella sampling (US/MD), DFT, and experiment (bar plot inset in Figure 4.7b) demonstrates excellent overall agreement between all three of these methods. The umbrella sampling tends to slightly

over-stabilise chloride binding, possibly due to mild hysteresis in the PMF reconstruction. In order to further validate the behaviour of the system, another binding estimation method compatible with MD was employed. This method was the free energy perturbation (FEP), where a thermodynamic cycle is employed estimate the energies of three species in solution using the BAR method - the bound complex, the unbound transporter **4.9** and the $Cl^-$ ion.[182] From these, the binding energy can be constructed. FEP calculations (Table 4.2) gave $\Delta G = -3.1 \pm 1.1$ kcal $mol^{-1}$, in line with the experimental binding free energy derived from titration. The agreement of the two dynamics-based binding prediction methods validates both the force-field description and the mechanistic interpretation of solvent competition provided by the simulations.

**Table 4.2:** Free energy perturbation (FEP) results for chloride binding in explicit DMSO, obtained from Bennett Acceptance Ratio (BAR) analysis. Values are given as free energy changes ($\Delta G$) in kcal $mol^{-1}$ with associated uncertainties. The binding free energy is calculated as the difference between chloride appearance in bulk DMSO and chloride appearance in the presence of **4.9**.

| **Process** | $\Delta G$ (kcal $mol^{-1}$) | Uncertainty (kcal $mol^{-1}$) |
|---|---|---|
| $Cl^-$ appearance in DMSO | –70.3 | ± 0.2 |
| $Cl^-$ appearance with **4.9** in DMSO | –74.6 | ± 0.9 |
| **Binding free energy** | –3.3 | ± 1.1 |

In summary, the MD data capture effects absent from the DFT treatment of a single DMSO adduct. In the explicit-solvent environment, many DMSO molecules continuously exchange at the binding site, leading to transient solvent-assisted chloride dissociation events. MD is also able to capture multiple solvent and transporter complex configurations that contribute to the free energy. The binding events involve significant entropic components — the cost of displacing or reorganising bulk solvent, the multiplicity of solvent arrangements, and the dynamic exchange of hydrogen-bond partners — which are inherently accounted for in MD but only approximated in continuum-solvent DFT. By discovering the solvent-assisted Min-

imum II and the transition state where a DMSO-$Cl^-$ clash dictates ion dissociation kinetics, the importance of considering explicit solvent is highlighted, beyond mere bulk entropic effects.

## 4.3 Conclusions and Future Work

In this chapter, two closely related families of stimuli-responsive chloride transporters were studied. The amide-based systems reported in our earlier work and the newer hydrazone-based scaffolds. Both employ intramolecular hydrogen bonding as a driver of preorganisation, and allow for gating this preorganisation using chemical caging. However, the hydrazone derivatives clearly demonstrate enhanced chloride binding and transport activity compared to their amide analogues.

Computational analysis provided complementary mechanistic analysis of the interactions at play in the two transporter families. In the amide family, DFT calculations were used to compare the potential 1:1 and 2:1 complexes with the chloride ion in both the gas phase and implicit solvent. Recognising the need for enhanced modelling of the energetics, the DFT calculations were enhanced with NBO and distortion–interaction analyses for the hydrazone family. It was highlighted how preorganisation lowers the distortion penalty and stabilises chloride binding in the active hydrazone, while its protected analogue is destabilised by weaker intramolecular hydrogen bonds and a much higher cost of reorganisation. Explicit-solvent molecular dynamics simulations then revealed the limitations of continuum solvation and lack of consideration of multiple binders. Free energy profiles obtained from umbrella sampling and FEP in DMSO showed how chloride binding is dynamically modulated by solvent competition. This included the discovery of solvent-assisted intermediates in the dissociation pathway, which was not visible in the static DFT description.

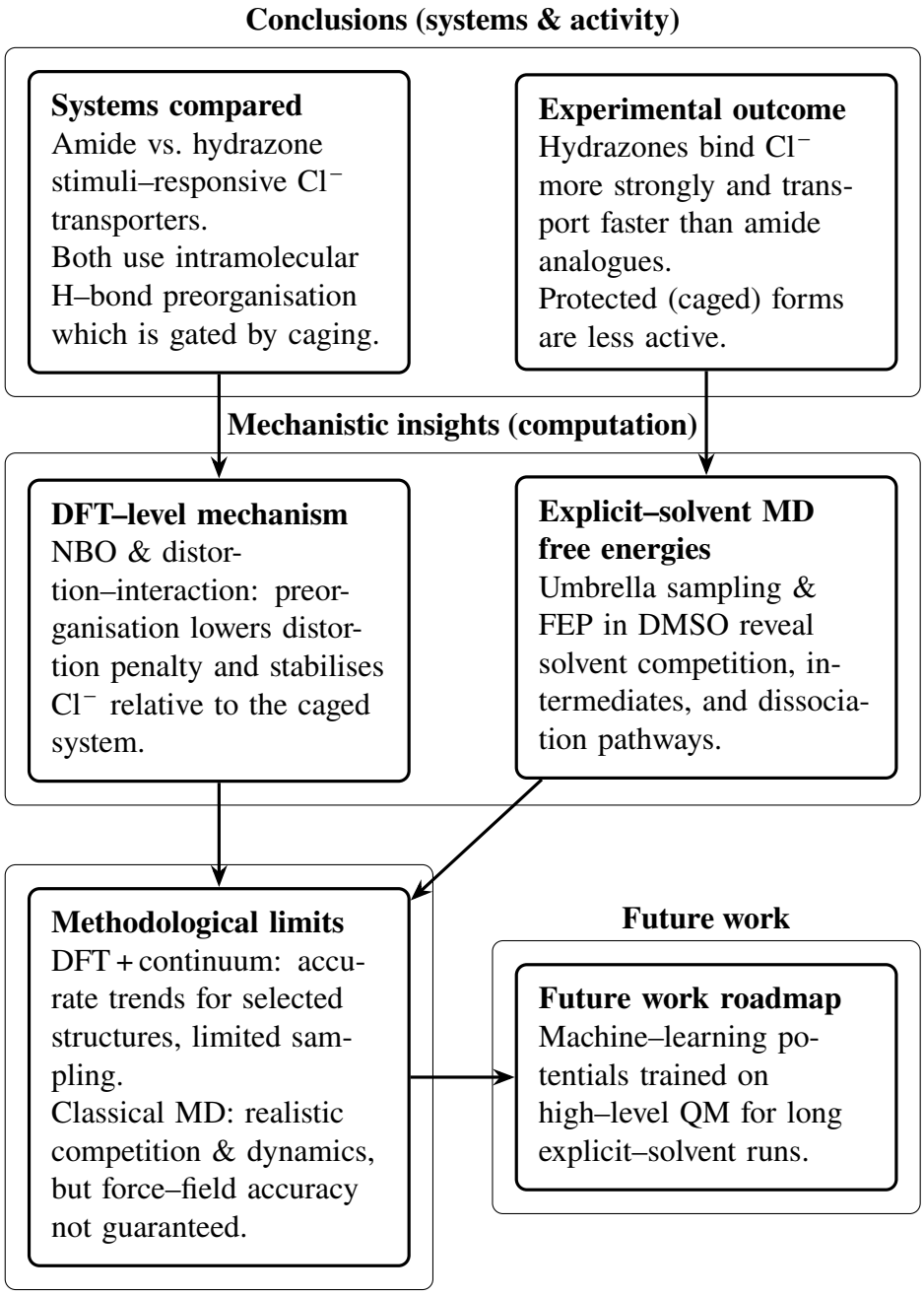



**Figure 4.8: Conclusions and future directions for modelling of stimuli–responsive chloride transporters.** Hydrazone scaffolds outperform amides; explicit–solvent calculations offer significant improvement on DFT; future avenues should focus on tight cooperation between experiment and computation, as well as the use of machine learned interatomic potentials

Together, these studies underline both the power and the limitations of current mod-

elling approaches. While DFT combined with continuum solvation can rationalise relative binding preferences and the role of intramolecular hydrogen bonds, it is inherently restricted to only be able to model a handful of complexed species at a time. Explicit-solvent MD captures the competitive environment more realistically, but relies on classical force fields whose accuracy for delicate hydrogen-bonding and solvation effects is not guaranteed. To go beyond these compromises, there is a clear need for computational approaches that can retain quantum-level accuracy while sampling the rich conformational and solvation landscapes that govern binding in solution. Machine-learning interatomic potentials (MLIPs), trained on high-level DFT data but applicable to long explicit-solvent simulations, offer an exciting route to bridge this gap. In the following chapter, attention is therefore turned to the development and application of machine-learning models for the binding of chloride anions, with the aim of capturing both the energetic subtleties of intermolecular bonding and the many-body effects of bulk solvent relevant to experiment.

# 5 |

# MLIPs For Halogen Bonding In Solution

In this chapter, a new method of modelling and predicting supramolecular binding is developed, using the state-of-the-art MACE machine learning interatomic potential (MLIP) architecture.[138] Several novel additions to the *mlp-train* package, which allow the automated training of MLIPs for chemical systems are described.[183] These aid the simulation and training of molecules in fully solvated conditions. Several methods of data acquisition are then benchmarked against test sets, before finally demonstrating stable umbrella sampling dynamics using the final MLIP. This is shown to perform better than classical dynamics in preliminary umbrella sampling simulations, reaching <1 kcal $mol^{-1}$ errors relative to the experimental reference.[52]

## 5.1 DFT and Implicit Solvation

Literature gas phase benchmarking studies of halogen bonding showed the SCS-MP2 wavefunction method to be the closest in accuracy to the gold standard of CCSD(T) with a complete basis set approximation, when it came to the geometries and energies of halogen bonded complexes.[56, 185] This same work by Kozuch and Martin showed that the GGA M06L functional[186], the M06-2X hybrid[187], and the $\omega$B97X range-separated hybrid functionals were three of the best-performing DFT functionals when it came to matching the gold standard[166]. It was decided to include both the regular and the SCS- version of MP2, the mentioned DFT functionals, as well as some others into our own benchmarking set,

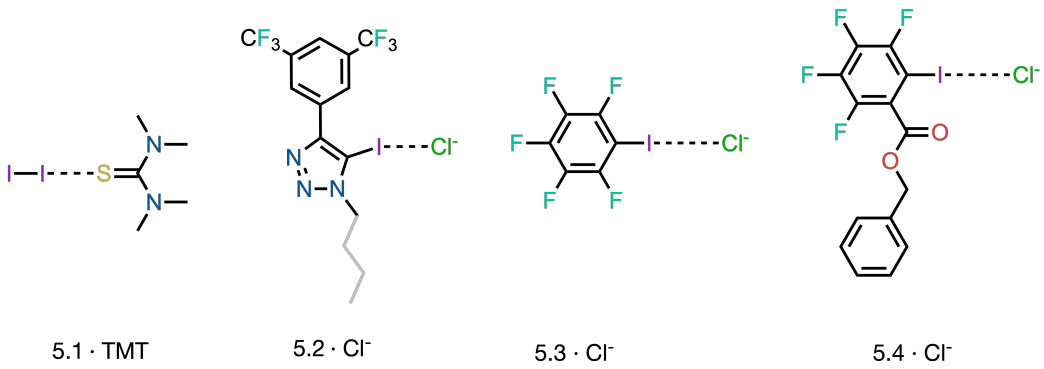



**Figure 5.1:** Four systems under study in this chapter, named from here on by the host molecule. $I_2$ **5.1** binding with tetramethylthiourea (TMT) in methanol[184], the iodotriazole **5.2** binding with $Cl^-$ in acetone[52], with gray alkyl tail indicating truncation with respect to the experimental system, perfluorobenzene **5.3** binding with $Cl^-$ in acetone[116], and the perfluoroiodobenzoate ester **5.4** binding $Cl^-$ in acetone.[116]

which was to be carried out on literature chloride transporters **5.1**-**5.4** (Figure 5.1).

The additional functionals to be included were the hybrid PBE0[188], the double hybrid B2PLYP[189], and the range-separated CAM-B3LYP[190], as well as the newer version of the Head-Gordon range-separated family, the $\omega$B97M[191]. As in the study by Kozuch and Martin, the effects of dispersion corrections on some of the functionals were also probed, for which the computationally affordable PBE0 was chosen on the one end, probed with both the D3BJ and the D4 corrections[192–194], as well as the two more expensive range-separated functionals in $\omega$B97X and $\omega$B97M[191, 195].

Across our set, the mean absolute errors (MAE) of the electronic energies of chloride binding to **5.1**-**5.4** were found to be between 1.3–5.3 kcal mol$^{-1}$, while free energy MAEs spanned a range of 1.4–5.7 kcal mol$^{-1}$ (Table 5.1). Among DFT functionals, the range-separated hybrid meta-GGA $\omega$B97M-D3BJ gives the lowest MAEs for both electronic energies (1.3 kcal mol$^{-1}$) and free energies (1.4 kcal mol$^{-1}$), followed by M06-2X (1.7 and 2.2 kcal mol$^{-1}$) and $\omega$B97X (1.7 and 2.0 kcal mol$^{-1}$). In contrast, the PBE0 functional, especially when including the D3

**Table 5.1:** Mean absolute errors (MAE) of energies of binding across four test molecules **5.1**-**5.4**, for each tested DFT functional against the RI–SCS–MP2 method, which was found to perform closest to the CCSD(T) gold standard. The errors are presented for electronic binding energies and free energies (kcal/mol), alongside the relative single-point electronic energy evaluation time (s). Rows are grouped by Jacob's ladder rungs, with wavefunction references shown first for context.

| Functional | MAE (Electronic) | MAE (Free Energy) | Eval. Time (s) |
|---|---|---|---|
| *Wavefunction references* | | | |
| SCS-MP2 | 0.00 | 0.00 | 247 |
| MP2 | 1.98 | 2.20 | 239 |
| *Rung 3: meta-GGA* | | | |
| M06-L | 2.95 | 3.70 | 59 |
| *Rung 4: hybrid GGA (global / range-separated, with/without D3/D4)* | | | |
| CAM-B3LYP | 1.96 | 2.33 | 221 |
| PBE0 | 3.35 | 3.80 | 154 |
| PBE0-D3BJ | 4.90 | 5.28 | 154 |
| PBE0-D4 | 5.30 | 5.68 | 154 |
| **$\omega$B97X** | 1.67 | 1.98 | 254 |
| $\omega$B97X-D3BJ | 2.33 | 2.72 | 241 |
| *Rung 4: hybrid meta-GGA (incl. VV10 / D3BJ)* | | | |
| **M06-2X** | 1.67 | 2.15 | 165 |
| $\omega$B97M-V | 2.15 | 2.30 | 256 |
| **$\omega$B97M-D3BJ** | 1.33 | 1.42 | 309 |
| *Rung 5: double hybrid* | | | |
| B2PLYP | 2.08 | 2.42 | 184 |

and D4 dispersion corrections, show the largest errors on this halogen-bonded set, with errors between 3.4 and 5.7 kcal mol$^{-1}$ for both electronic and free energies. The double hybrid B2PLYP is intermediate (2.1 and 2.4 kcal mol$^{-1}$), while the meta-GGA M06-L is fast but relatively inaccurate (3.0 and 3.7 kcal mol$^{-1}$). There is a significant difference between the wavefunction method MP2 and its SCS-variant (2.0 and 2.2 kcal mol$^{-1}$).

Interestingly adding D3BJ correction to the hybrid GGA $\omega$B97X functional worsens the quality of the result, while its addition to the hybrid meta-GGA $\omega$B97M

functional improves it. In the case of the latter, the D3BJ correction also outperforms the VV10 dispersion correction. This goes to underscore what was already found by Kozuch *et al.* in their benchmarking study - that dispersion corrections may not always perform well with halogen bonds, and that their performance with different functionals may differ greatly.[56]

Overall, the ordering of the free energies is similar to that of the electronic energies, but is around 0.1–0.5 kcal $mol^{-1}$ higher in value. This suggests that the PES curvature is captured similarly well by the different functionals as the overall difference between the stationary points. This is important for sampling away from minima during MLIP training and binding simulations. The ranking of the three best functionals (bold font in Table 5.1) was thus used, combined with their evaluation times, to decide on our functional of choice for performing our MLIP simulations. While $\omega$B97M-D3BJ achieves the best accuracy, it is among the slowest (309 s). Both M06-2X and $\omega$B97X offer good accuracy at lower cost (165 and 254 s) with similar energetic accuracies. M06-L, which was found to perform well by Kozuch *et al.*, is fastest (59 s) but is also found by us to be the least reliable on this chemistry.[56] It was decided that the best trade-off between accuracy and cost for training MLIPs is offered by M06-2X, which is the functional that was used for our further studies.

But before moving onto the training of our machine learned interatomic potentials, the extent to which uncertainty from implicit solvation adds to the error already present in functional choice was also quantitatively probed. As shown in Figure 5.2, small differences in solvation energies for the bound and unbound species can quickly accumulate. If these solvation differences vary across different functionals, this can lead to functional-dependent errors in solvation contribution ($\Delta G_{\text{solv}}^{\text{bind}}$). This can due to a variety of reasons. The unbound host exists partially pre-solvated in its cavity by Lewis-basic solvent, and continuum methods may not

properly represent these explicit interactions. Equally, the free anion carries a structured primary and secondary solvation shell, which is disrupted by binding, and continuum solvation only crudely addresses this. Literature studies on anion solvation energies have found implicit solvation models to have errors on the order of 3-5 kcal·mol$^{-1}$.[196, 197] The studies mention that the models also fail to account for the specific interactions that a solvent may have with the directionality of hydrogen bonds, which does not bode well for the highly anisotropic and even more directional character of the halogen bond. Due to the solvation energy's direct dependence on the charges of the solute, the functional choice may also have a high impact on the final solvation energy of the process.

The overall degree of solvation disruption caused by binding depends on binding site geometry, contact distances, and the directional character of the interaction, which all change when the geometries and charge distributions alter upon optimisation with different functionals. Because continuum models (here CPCM) couple directly to the electronic charge distribution, the functional choice thus leaves an energetic imprint on the computed solvation response. And while absolute solvation values are not available, the binding in implicit solvation can be evaluated through our calculations for each functional, along with the impact of the uncertainty of the implicit solvation model on this value.

To quantify this, each of the wavefunction and DFT methods was measured against the experimental binding values using the respective system's solvent in the CPCM solvation method (Table 5.2). The obtained values were compared directly against experimental binding constants to assess the predictive power of the different methods for host–guest binding, as well as to quantify how much the solvation contribution diverged among the different functionals. For **5.1** in MeOH, many methods significantly overbind relative to experiment (e.g., M06-L -7.0 and PBE0 -6.9 kcal mol$^{-1}$), whereas for **5.2** in acetone most methods underbind by 0.3 to

**Table 5.2:** Errors in solvated binding free energies (kcal mol$^{-1}$) for all DFT functionals, as well as for the wavefunction methods, relative to experiment, ordered by Jacob's ladder (Appendix Section C.1 for discussion). Error is defined as $\Delta G^{\text{bind}}_{\text{calc}} - \Delta G^{\text{bind}}_{\text{exp}}$ (negative values indicate more favorable binding than experiment). CPCM solvation was used (MeOH for **5.1**; acetone for **5.2**, **5.3**, **5.4**). The footer reports the per-molecule mean absolute error (MAE) across all functionals, and the standard deviation of the solvation contribution $\Delta G^{\text{bind}}_{\text{solv}}$ for each molecule across functionals, indicating the variability of solvation contributions.

| **Functional** | **5.1** | **5.2** | **5.3** | **5.4** |
|---|---|---|---|---|
| *Experimental Reference* | | | | |
| Binding Energy | -4.7 | -3.1 | -2.5 | -4.3 |
| *Wavefunction references* | | | | |
| SCS-MP2 | −3.5 | +3.2 | +2.7 | +5.4 |
| MP2 | +0.6 | +3.9 | +3.6 | +5.0 |
| *Rung 3: meta-GGA* | | | | |
| M06-L | −7.0 | +1.8 | +1.7 | +3.0 |
| *Rung 4: hybrid GGA (global / range-separated, with/without D3/D4)* | | | | |
| CAM-B3LYP | −2.5 | +3.8 | +3.0 | +4.2 |
| PBE0 | −6.9 | +2.0 | +1.8 | +2.3 |
| PBE0-D3BJ | −5.1 | +1.0 | +1.0 | +5.6 |
| PBE0-D4 | −1.9 | +0.3 | +0.6 | +5.7 |
| $\omega$B97X | +1.2 | +2.1 | +1.7 | +3.4 |
| $\omega$B97X-D3BJ | −1.9 | +1.4 | +1.6 | +3.6 |
| *Rung 4: hybrid meta-GGA* | | | | |
| M06-2X | −3.8 | +3.1 | +2.0 | +3.9 |
| $\omega$B97M-V | −0.7 | +2.0 | +1.7 | +3.6 |
| $\omega$B97M-D3BJ | −0.1 | +3.1 | +2.6 | +4.6 |
| *Rung 5: double hybrid* | | | | |
| B2PLYP | −1.6 | +2.9 | +2.8 | +4.5 |
| *MAE across functionals (\|error\|)* | **2.83** | **2.35** | **2.06** | **4.22** |
| $\sigma\left[\Delta G^{\text{bind}}_{\text{solv}}\right]$ *across functionals* | 0.21 | 1.07 | 1.21 | 1.57 |

3.9 kcal mol$^{-1}$. **5.3** shows a similar positive bias with values ranging from 1.0 to 3.6, and **5.4** exhibits the largest positive deviations 2.3 to 5.7. Across the test set, even high-level wavefunction references do not uniformly match experiment once solvent is involved, with mean absolute errors (MAEs) ranging from 2.06 to 4.22 kcal mol$^{-1}$. Furthemore, the standard deviation of $\Delta G^{\text{bind}}_{\text{solv}}$ is significant (albeit

size-dependent) for all molecules. In **5.1** the standard deviation ($\sigma$) is 0.21 kcal mol$^{-1}$ but it grows for **5.2** to 1.07 kcal mol$^{-1}$, **5.3** (1.21), and **5.4** (1.57), indicating that larger and more polar hosts amplify solvation-model dependence.

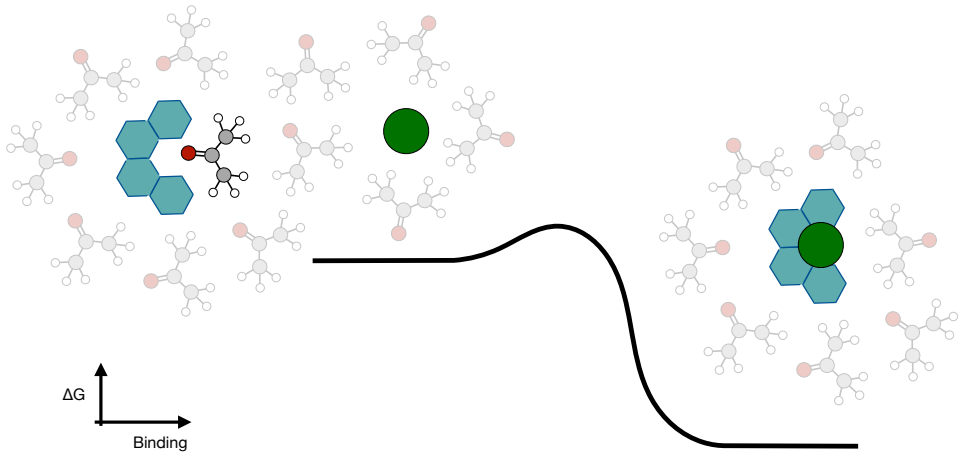



**Figure 5.2:** Schematic depiction of a supramolecular anion binder binding a chloride anion in acetone: on the left, the separated host and guest molecules interact with the solvent separately, whereas on the right they interact with the solvent as a complex.

Many studies show that for the solvation energies of species with strong solvent interactions, the introduction of at least a small amount of solvent molecules is necessary for achieving a good match with experiment. For example, some models, such as hybrid or discrete-continuum models, where the first solvation shell is treated explicitly, whereas the bulk solvent is still treated implicitly, have been shown to outperform implicit solvation models in such systems.[198] The explicit consideration of a larger number of solvents using DFT is normally outside of the domain of available computational resources for most research groups, but the combination of exploring solvated conformations using semi-empirical methods (e.g. using the XTB-based CREST) or using molecular dynamics, before optimising a microsolvated configuration with only a few solvents, can be a popular approach. In order to explore to what extent explicit consideration of solvent

molecules improves the treatment of anion binding (as previously observed in Chapter 4), a molecular dynamics approach was next explored.

## 5.2 MD of Halogen Bonding Molecules in Explicit Solvent

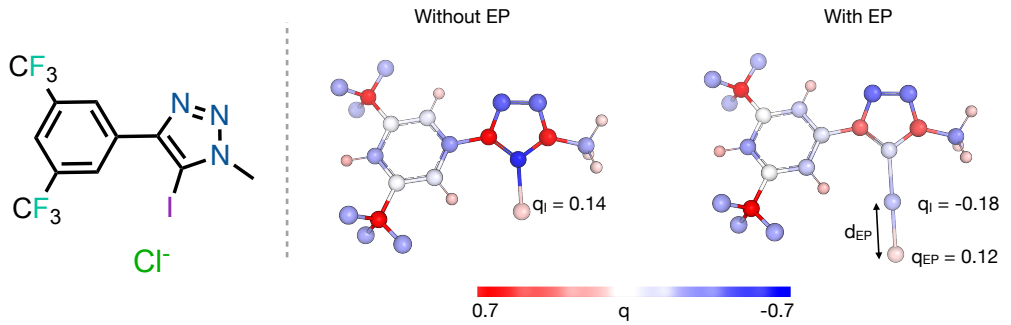



**Figure 5.3:** Example of the fitting of an EP charge to the halogen bonding iodine of the **5.2** system (left). The middle and right depictions show the RESP fitted charges with and without the EP, showing the overall charge distribution.

In this work, a single EP charge was used to model the sigma hole of the halogen bond. The iodine atom was assigned a positive charge in the absence of the extra particle ($q_I$=0.14), which changed to a negative charge ($q_I$=-0.18) after the addition of the EP. The final charge distribution of the molecule can be seen to only change in the immediate vicinity of the iodine atom, with the additional positive charge ($q_{EP}$=0.12) of the extra particle being countered by modified charges of a range of its neighbours.

The weakness of this approach is that it is still inherently parametric (the parameters chosen were the same parameters as those optimised in Section 3.3.1). One must choose the distance of this EP charge from the middle of the iodine atom. On top of the distance, two different approaches exist depending on whether the EP charge is allowed to vary with the rest of the particles in the RESP fitting, or whether

it should be held constant. All of these parameters mean that the final binding energy will be highly dependent on their specific choice.

The binding profile of the $Cl^-$ anion (Figure 5.4) features a minimum of -1.9 kcal·mol$^{-1}$ at 3.8 Å, with a strong repulsive barrier at shorter distances and a dissociation curve at larger distances, which is followed by another shallow minimum of -0.7 kcal·mol$^{-1}$ at a distance of 6.2 Å. This was identified as an electrostatic interaction with the positively charged protons of the N-Me group on the triazole ring. The binding process is almost barrierless, with the movement of the chloride being bound at the methyl group and it binding at the iodine being obstructed by only a 0.4 kcal·mol$^{-1}$ of an activation energy. This corresponds to the energy needed to displace the acetone bound to the iodine, allowing the binding of the $Cl^-$ to the $\sigma$-hole.

The umbrella sampling binding trajectory in Figure 5.4b can be broadly split up into three different regions - the bound region, the transition state region and the unbound region (regions I, II and III in Figure 5.4a, which will be discussed once more when MLIP umbrella sampling simulations are performed). The binding energy prediction of -1.9 kcal·mol$^{-1}$ using this method, is reasonably close to the -3.2 kcal·mol$^{-1}$ obtained using chloride binding titrations experimentally. Its error of 1.3 kcal·mol$^{-1}$ also outperforms all but one of the functionals trialed in the DFT binding prediction of molecule **3.2**. This shows the value of classical dynamics in binding prediction already demonstrated in the SAMPL challenges.[112–115] Unfortunately, the approach has a few inherent limitations which are difficult to overcome, especially when dealing with halogen bonding systems. Namely, the parametrisation of the additional charge particle remains an expert task, and is inherently a parametric method, meaning that the optimal approach of fitting the EP can differ from system to system. Ion force-field parameters are also usually tuned to reproduce the solvation properties of the ion in a particular solvent, and

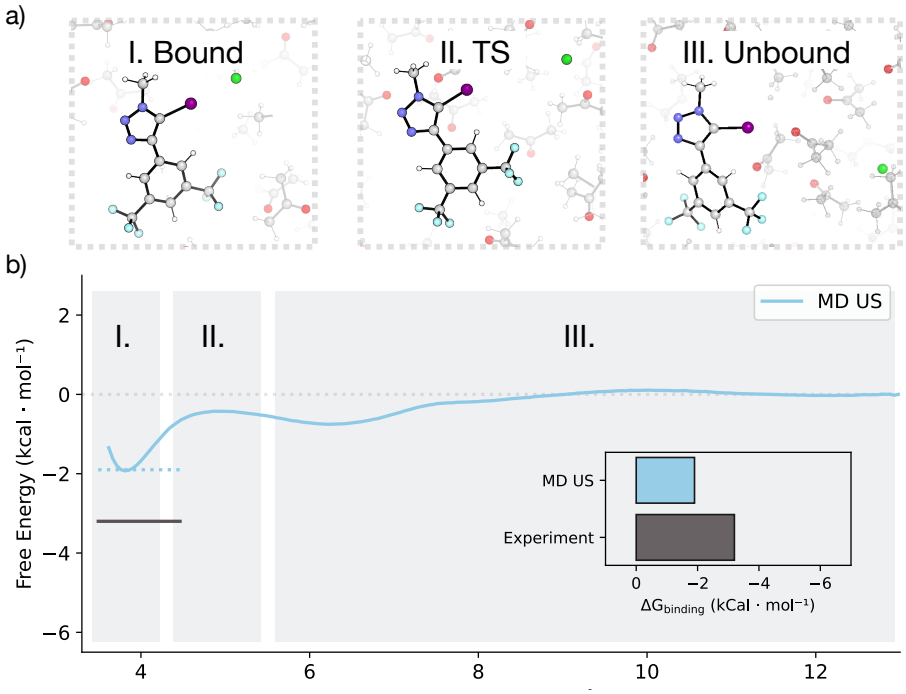



**Figure 5.4:** a) The three regions of $Cl^-$ pulling during US simulations including: I. the bound $Cl^-$ complex, II. the transition region of the $Cl^-$ pulling away, and III. the unbound $Cl^-$ and host. Representative snapshots taken from simulations. b)Umbrella sampling energy profile of **5.2**·$Cl^-$ binding in acetone, performed in GROMACS. Windows were spaced 1 Å apart, with a 2000 kJ $mol^{-1}$ restraint placed on the I-Cl distance. Each window was run for 100 ns until convergence.

their transferability may be limited.

Additionally, when it comes to making the transition beyond mere non-covalent interactions, the limitation of classical force fields being unable to model bond-breaking and -making rears its head once more. This makes them incapable of modelling reactions with supramolecular halogen bonding catalysts, for example. The use of MLIPs to model ion binding would represent a stride towards a method where the binding energy result does not depend on the choice of a small number of input parameters. With them, intermolecular interactions could be modelled at the high level of DFT methods, while the computational efficiency of classical dynamics is retained, allowing the consideration of a large number of solvent molecules and permitting covalent bond formation.

## 5.3 Machine Learning Interatomic Potentials

### 5.3.1 Algorithmic Additions to mlp-train

Three key algorithmic additions were made to *mlp-train* in order to facilitate some of the new sampling approaches employed in this work: periodic boundary reconstruction, solvation, and microsolvation. In this section, key aspects of the new functionalities are introduced, which are necessary for understanding the novel sampling approaches employed that follow.

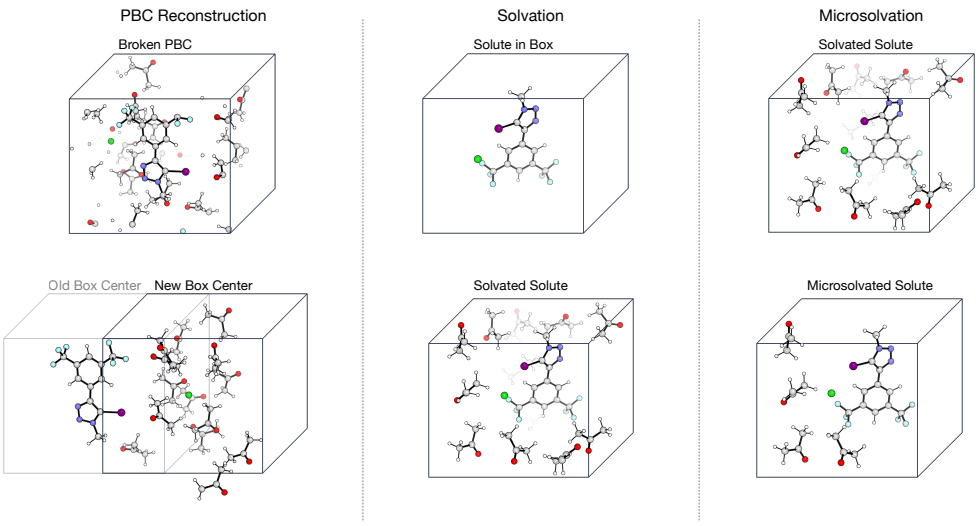



**Figure 5.5:** Depictions of the three novel algorithms, enabling new functionalities within mlp-train, to allow for different sampling strategies, as well as new simulation conditions which were not available in the package before. From left to right: periodic boundary reconstruction, with the option of centering atoms; solvation function; microsolvation function.

One of the key challenges in analyzing molecular dynamics trajectories with periodic boundary conditions (PBC) is the artificial fragmentation of molecules across simulation box boundaries. During MD simulations, individual atoms can cross periodic boundaries independently, resulting in molecules that appear broken when

visualized or analyzed. This issue particularly affects post-processing tasks such as trajectory visualisation, carrying out non-periodic DFT calculations and analysing interactions across PBCs.

Unlike classical MD, MLIP simulations do not require bonds to be defined prior to beginning simulation. This means that they also do not have lists of bonds, which are used to reconstruct broken molecules in classical MD codes. In order to enable the use of the *mlp-train* package beyond simple spherical clusters of molecules, it was necessary to implement a function which would be able to take a list of molecules and reconstruct those that had been broken across PBCs.

**Table 5.3:** Summary of the timing of the periodic boundary reconstruction function depending on the box size, including the number of tests per box size, the average time taken to fill the box, the average number of molecules per test and the average time needed per molecule. The tests were carried out on 12 Xeon Gold 6330 CPU cores.

| Box Size | Tests | Avg Time (s) | Avg Mol/Test | Avg Time/Mol (s) |
|---|---|---|---|---|
| 15.0 | 2 | 0.000072 | 55.5 | 0.000001 |
| 20.0 | 2 | 0.000167 | 151.0 | 0.000001 |
| 25.0 | 2 | 0.000321 | 310.0 | 0.000001 |
| 30.0 | 2 | 0.000513 | 545.0 | 0.000001 |
| 35.0 | 2 | 0.000730 | 874.0 | 0.000001 |
| 40.0 | 2 | 0.001045 | 1309.0 | 0.000001 |

Simply taking the wrapping function of the ASE molecular dynamics engine in use within mlp-train, the atoms are wrapped back into the box without consideration of which atoms belong to which molecule, and should not be broken across box boundaries (5.5, top left).[183, 199] This is solved in our algorithm by storing the integer indexes of the starting atoms of each molecule. This allows the periodic boundary reconstruction algorithm to determine whether a group of atoms belonging to the same molecule has crossed any of the box boundaries in "unwrapped" structures. In an unwrapped simulation box, atoms are not forced back into the central cell when they cross periodic boundaries; instead, their positions continue smoothly into the neighbouring periodic images. This ensures that molecules

which would otherwise appear broken across a boundary remain continuous, allowing their geometric centres (COMs) $\mathbf{R}_m^{\text{unwrap}}$ to be computed without artificial discontinuities. By finding these geometric centres, molecules can then simply be translated by however many integer multiples of the box dimensions they have moved away from the original box, resulting in a reconstructed box:

$$\mathbf{R}_m^{\text{recon}} = \mathbf{R}_m^{\text{unwrap}} - \mathbf{N}_m \cdot \mathbf{L}$$

where:

- $\mathbf{R}_m^{\text{unwrap}} = \frac{1}{N_m} \sum_{i \in m} \mathbf{r}_i^{\text{unwrap}}$ is the geometric centre (centroid) of molecule $m$ in the unwrapped structure,
- $\mathbf{N}_m \in \mathbb{Z}^3$ is the integer vector recording how many box lengths the molecule's centre has moved,
- $\mathbf{L} = (L_x, L_y, L_z)$ is the box dimension vector,
- $\mathbf{R}_m^{\text{recon}}$ is the reconstructed centre of the molecule.

Then each atom is shifted consistently with that:

$$\mathbf{r}_i^{\text{recon}} = \mathbf{r}_i^{\text{unwrap}} - \mathbf{N}_m \cdot \mathbf{L}, \quad i \in m$$

That way, all atoms in the molecule are translated by the same integer multiple of the box vectors to rebuild the periodic image. Additionally, one has the option of centering any of the atoms in the system (Figure 5.5, bottom left). For example, if one simulates a small box of a solute and only a small number of solvents, then by centering the atoms which crucially require solvation in the following DFT calculations, this preferential solvation can be achieved. In benchmarks of boxes

of **5.2** solvated in various box sizes with either acetone and water, it was confirmed that the algorithm scales roughly in an $O(N)$ fashion with the number of atoms 5.3, and that the evaluation times per atom are roughly on the order of a millisecond on 12 CPU cores.

In order to facilitate the construction of solvated boxes (Figure 5.5, centre), a solvation algorithm was implemented next, based on the random insertion algorithm suggested by Triviño [200]:

- For the given box size, the number of solvent molecules required is calculated based on the density of the solvent. Several different modes of specifying the solvent and density exist, from completely automatic for a large database of supported solvents, as well as completely manual specification of custom solvents.
- The volume of the solute is then removed, assuming that it has the same density as the solvent.
- Solvent molecules are added until the number is achieved or the function time-out is reached, while checking for contacts between incoming atoms and existing atoms in the box using a k-d-trees algorithm, using either a rigid contact threshold or one based on the sum of van der Waals radii of proximate atoms.

**Table 5.4:** Summary of solvation function performance by solvent, including the number of tests per solvent, the average time it takes to fill the box, the average number of molecules and atoms added for the given solvent, and the average success rate, defined as the fraction of molecules added to the solvent box divided by the desired number of molecules to be added to achieve the desired density.

| Solvent | No. Tests | Avg. Time (s) | Avg. Mol. | Avg. Atoms | Avg. Succ. % |
|---|---|---|---|---|---|
| acetonitrile | 12 | 0.6 | 129.8 | 779.0 | 100.0% |
| water | 12 | 0.6 | 381.9 | 1145.8 | 100.0% |
| toluene | 12 | 7.6 | 54.5 | 817.5 | 89.6% |

We benchmarked the algorithm solvating each of **5.1**, **5.2**, **5.3** and **5.4** with three different solvents (acetonitrile, toluene and water) in boxes with lengths of 10, 20 and 30 Å. The full results of all 36 tests can be found in B.14. A comparison of the three different solvents shows that water and acetonitrile boxes are solvated much more quickly than the toluene, despite having a smaller size and thus a larger number of molecules that need to be added (5.4). This shows that the bottleneck is the random insertion of molecules, which makes larger and less globular solvents more difficult. This is also reflected in the success rate of solvation - i.e. how many of the desired number of solvents were actually added by the algorithm. Due to the time-out of the solvent addition function, if solvents cannot be packed into the box after 5000 attempts, the box is simply returned with as many solvents as the algorithm was able to insert. The lower success rate for toluene again indicates that packing larger solvents using such an approach is more difficult.

**Table 5.5:** Summary of solvation function performance by box size, including the number of tests per solvent, the average time it takes to fill the box, the average number of molecules and atoms added for the given solvent, and the average success rate, defined as the fraction of molecules added to the solvent box divided by the desired number of molecules to be added to achieve the desired density.

| Box Size | Tests | Avg. Time (s) | Avg. Mol. | Avg. Atoms | Avg. Succ. % |
|---|---|---|---|---|---|
| 10.0 | 12 | 0.100 | 6.3 | 29.0 | 100.0% |
| 20.0 | 12 | 1.800 | 122.3 | 590.8 | 94.6% |
| 30.0 | 12 | 6.900 | 437.6 | 2122.5 | 95.0% |

The final algorithmic addition to the *mlp-train* package was a microsolvation functionality. Using a pre-defined list of solvent molecules produced by the solvation function or manually specified by the user, this function has the ability to remove all but the most key solvents around the solue in the following manner:

- The number of solvent molecules $N_{microsolvation}$ to be kept is specified, and their geometric centres and distances from the solute's centre are calculated.
- Alternatively, atoms of interest, which are the most important to consider

in interactions with the solvent, can be specified as the geometric centre of interest in the solute for the distance list.

- All but the nearest $N_{microsolvation}$ solvent molecules specified are removed from the simulation box, leaving a microsolvated configuration.

This function can be used to produce entirely microsolvated trajectories, which can be used for labelling, plugged into active learning (AL; *vide infra*) or used for the development of further functionalities within.

### 5.3.2 Comparison of Sampling Approaches

We decided to use the **5.2** system as an initial case study for developing an MLIP for chloride binding in solution, which would aim to overcome the limitations posed by DFT and molecular dynamics approaches. The system that would need to be simulated would be composed of three distinct species – the supramolecular host **5.2**, the chloride anion, and the acetone solvent. All three components have complex intermolecular interactions with each other, namely strong repulsion between the oxygens of the acetone and the chloride anion and strong attractions between the sigma hole of the transporter and both the chloride anion and the acetone's oxygens. With such complex systems, attempts to train the entire system at once are not usually fruitful, which means that they are trained slowly in a component-wise manner, which can lead to a long acquisition time of the training data. Thus, two strategies were compared in detail in order to determine the most efficient way to train the final MLIP.

In the first strategy, the subcomponents of the systems were trained using active learning loops, according to the strategy developed in our group by Zhang *et al.*, with one small modification to their procedure[130]. Their training was carried out using gas phase clusters, which need to be restrained using a spherical potential

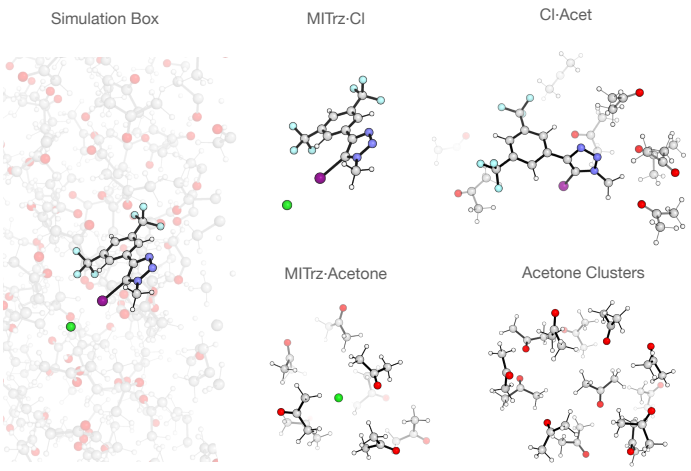



**Figure 5.6:** Breaking down the system to be simulated with the final MLIP (left) into smaller components (right), which can be run through an active learning loop in order to generate useful new configurations.

in order to prevent solvent molecules from diffusing away from the solute over time. This can lead to several unphysical effects and other issues. Firstly, a radius of the restraining potential needs to be carefully defined, which somewhat accurately captures the density of the solvated system. This is difficult, because the solvent cannot diffuse from one side of the cluster to the other, creating a boundary region, where there must be additional space in order not to compress the system unphysically, and lead to deformation of bonds. This is a trial-and-error procedure, which can take a significant amount of time. It also does not allow for the reorganisation of solvent to the same extent as true periodic boundary conditions, where solvents can diffuse across all of the box sides. The strategy was therefore changed to proceed through active learning simulations in small solvent boxes, featuring a minimal number of solvent molecules, using our solvation and periodic boundary reconstruction functionalities (*vide supra*, Figure 5.5, left and middle).

In the second strategy, random sampling would be employed using foundation mod-

els at 500 K. Due to the out-of-the-box stability of foundation models, they can be utilised at high temperatures to generate a broader range of system configurations, with the higher kinetic energies leading to exploration of closer intermolecular contacts and out-of-equilibrium molecular geometries. This may lead to better stability of the final MLIP, due to having explored a larger proportion of phase space and therefore having a smaller chance of encountering a configuration in a simulation which it has trouble interpolating the energies and forces for. It can also lead to perils - for example, if the system has functionalities that were under-represented in the training set of the foundation model, it may generate unreasonable configurations, which violates the criteria outlined in 1.5.2.1.

The same number of solvent molecules and the same box size would be used for each sub-component sampling in both approaches. In order to compare the two approaches fairly, the same number of training configurations would also be generated for each sub-component of the **5.2** system before being analysed. Their individual merit would be measured by the extent of their phase space exploration and ability to reproduce energies and forces of an independent test set.

We used the GNN-based MACE machine learning architecture and the Duarte group's *mlp-train* package in order to carry out both sets of sampling. For each strategy, the system was broken down into smaller components in the training: the [**5.2**·Cl]$^-$ complex, the Cl$^-$ anion surrounded by 10 acetone molecules, the **5.2** transporter surrounded by 15 acetone molecules, and a cluster of 10 acetone molecules (Figure 5.6). In the case of the active learning strategy, the process was started with 12 randomly perturbed structures of the initial cluster, while the random sampling was conducted using the MACE-OFF foundation model (with the S parameters, see SI) from a box with of exactly the same atoms, obtained using the solvation function in the cases where acetone was present. It should be noted that for the [**5.2**·Cl]$^-$ complex sampling with MACE-OFF, a moving restraint was

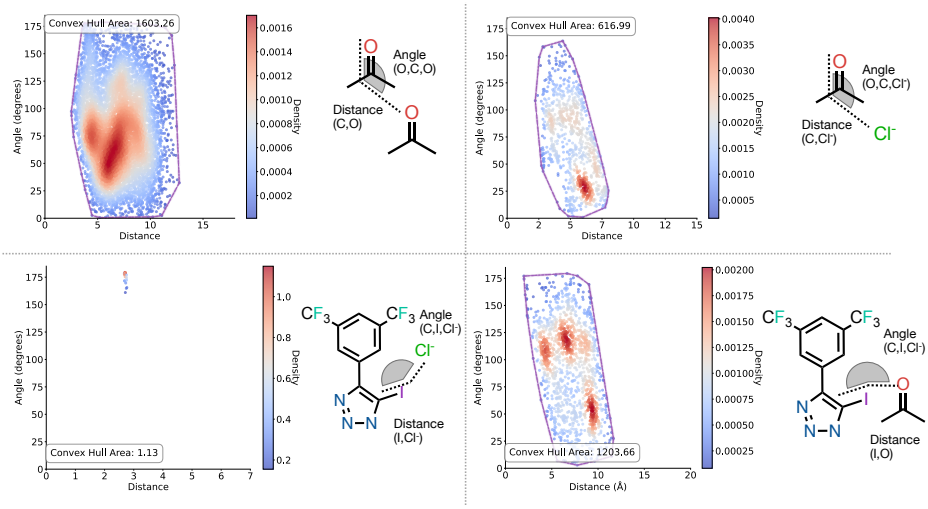



**Figure 5.7:** 2D density plots of a characteristic angle and distance are shown for each subsystem — [**5.2**·Cl]$^-$, acetone-Cl$^-$, **5.2**-acetone, and acetone-acetone in the active learning strategy. Convex hull outlines indicate the total coverage of the sampled points. 100 configurations each were analysed per subsystem.

employed in order to explore the relevant angles and distances, yet the MACE-OFF sampling group will still be referred to as "random" sampling In each case, the sampling was run until 100 structures were obtained.

In order to examine how effective each method was in providing configurations that most efficiently trained an MLIP of the system, the breadth and uniformity of the phase space coverage was examined first. For each of the four sets of data, a characteristic angle and distance were chosen, which are important for capturing the interactions at play and the training data was plotted using using a density plot (Figures 5.7 and 5.8). In order to facilitate comparison between the two different methods, a convex hull was also plotted around the boundary points of the sampled data, to give a total sampling area.

Across the four subsystems, the MACE-OFF sampling produces a much broader exploration of the chosen angle–distance spaces than the active learning strategy.

This is immediately clear from the convex hull areas annotated on the plots. For the acetone cluster, acetone-$Cl^-$ contact, I-$Cl^-$ contact, and I-O contact panels the AL hull areas are 1603, 617,1, and 1204 (in units of °·Å), respectively, whereas MACE-OFF expands these to 2435, 1962, 422, and 2551. This represents a roughly 1.5–3 times larger coverage in three of the four cases and a particularly large increase in sampling space for the I-$Cl^-$ motif (AL = 1 vs MACE-OFF = 422), where the regular AL strategy in particular was seen to fail. In this last case, the AL strategy could not be carried out beyond a total of 24 structures, as the MLIP learned that the $Cl^-$ anion binds tightly to the iodine, and was not able not explore any geometries beyond the bound conformation. When metadynamics were attempted to move the chloride ion away from the iodine, severe distortions of the C-I bond were observed. This issue was not present in the random sampling strategy utilising MACE-OFF, where a bias ensuring regular sampling across the required coordinate space could be applied without causing severe distortions of the structures in the simulations due to MACE-OFF's higher out-of-the-box stability.

The AL distributions also appear denser and more localized than the MACE-OFF sampling structures, with high-probability "islands" forming around the areas where the starting structures in the active learning loops were located. This is consistent with the AL selection strategy, which iteratively refines around initial structures, using uncertainty as a selection criterion. By contrast, MACE-OFF is more uniform and boundary-reaching, allowing it to explore more extreme angles, as well as both long and short contacts that enlarge the areas of the convex hulls. With increased AL sampling times, broader phase space may be accessed, which would decrease the risks of coverage holes, while in some cases (such as the I-$Cl^-$ contact), it may also need to be combined with enhanced sampling methods. MACE-OFF provides broader phase-space support, capturing rare geometries

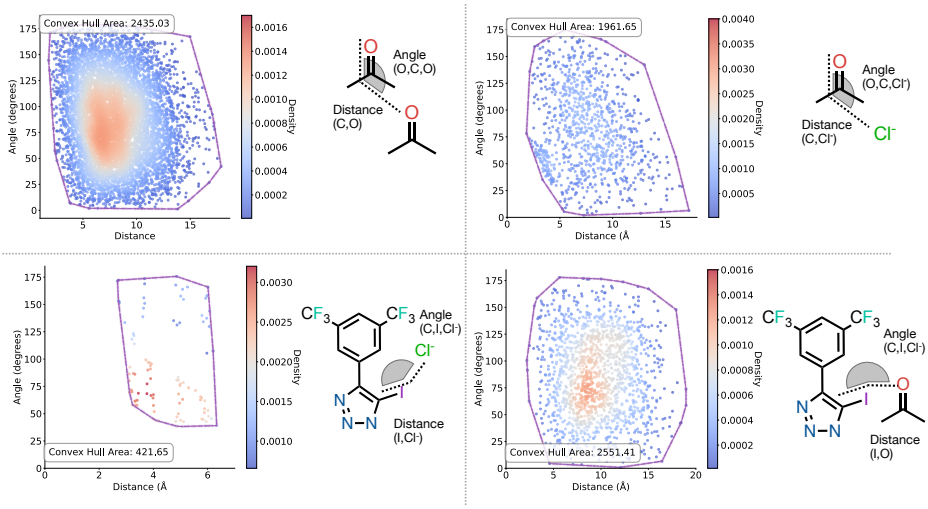



**Figure 5.8:** 2D ensity plots of a characteristic angle and distance are shown for each subsystem[**5.2**·Cl]$^-$, acetone-Cl$^-$, **5.2**-acetone, and acetone-acetone in the random sampling strategy. Convex hull outlines indicate the total coverage of the sampled points. 100 configurations each were analysed per subsystem.*

that may improve robustness and extrapolation, but may also inherently bias the system with any errors already present from its large foundational training set. For example, while the SPICE dataset on which MACE-OFF is trained is large, featuring well over 100,000 configurations, not many of these are likely to feature high-quality descriptions of the directionality-dependence of halogen bonding.

To assess the utility of the two sampling strategies more directly, model error landscapes were compared against DFT single point reference maps for the four representative interactions(Figure 5.9). For acetone–acetone contact (distance C–O, angle O–C–O), both sampling strategies perform very well. The chosen DFT reference PES features a minimum with subtle dipole interactions between the protons of one acetone's $\alpha$ carbons (4.5 Å, 145 °), and the oxygen of another acetone, and a maximum where the oxygen approaches the central carbon too

*Except for the I-Cl$^-$ contact in the AL strategy, where biased sampling was not possible due to instability and unbiased active learning trapped the Cl$^-$ in the bound minimum energy well.

closely (3.8 Å, 140 °). Mean unsigned errors for both sampling strategies lie at around 0.2 (AL) to 0.3 (Random) kcal $mol^{-1}$, and both methods slightly underestimate the energy at close contact regions. While AL performs slightly better here, it appears that both strategies capture the energetics of this interaction very well.

The acetone-chloride contact (distance O-$Cl^-$, angle O–C-$Cl^-$), presents itself as a more challenging target. At high angles and short distances, there is again a subtle attraction between the $\alpha$ protons and the $Cl^-$ ion, while at low angles the repulsion between the oxygen and the $Cl^-$ begins to dominate. In this strongly directional and steeply varying landscape, AL is clearly more effective within the chemically relevant region, with the lower unsigned error of 3.7 kcal $mol^{-1}$. In particular, the AL sampling strategy performs much better at the border between the repulsive and attractive regions, where the Random Sampling strategy highly overestimates the energy. On the other hand, the Random sampling strategy perfoms much better in the repulsive regime at intermediate and long distances, indicating that the high-temperature sampling performed may be able to force exploration of some of these regions better than active learning. The differences between the two are consistent with AL adding data in high-curvature zones identified by outlier identification much bette. The random sampling on the other hand, which uses a pre-trained model that already has a tendency to explore low-energy configurations, will be prone to avoiding these regions.

One issue which is apparent in the I-$Cl^-$ contact is the limited range of vision of the MACE potential, which is 4.5 Å without any message-passing neighbours between the I and $Cl^-$. Hence, the energy curvature in the reference PES levels off beyond this distance. This is also exemplified by the large errors at the boundary of this range for both sampling strategies. Unsurprisingly, the AL approach, which could not explore outside of the bound basin (Figure 5.7), performs much worse, with an average error of 10.2 kcal $mol^{-1}$, compared to the MACE-OFF sampling's

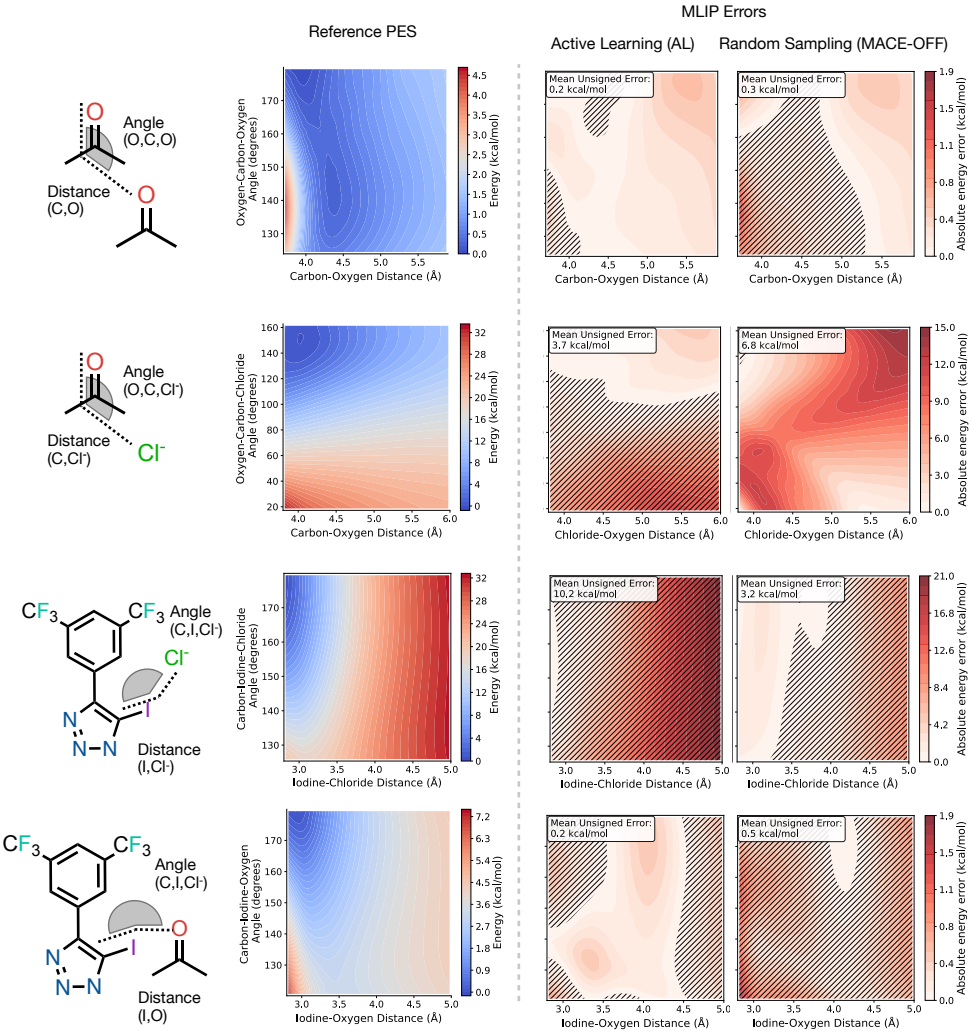



**Figure 5.9:** Error benchmarking for both AL and random sampling methods vs. 2D DFT single point scans of relevant space of four representative interactions. For each subsystem (rows), the left panel sketches the interaction and defines the reaction coordinates (distance, angle); the second panel shows the DFT single-point potential-energy surface $E_{\mathrm{DFT}}$(distance,angle). The two right panels report absolute error maps $|E_{\mathrm{ML}} - E_{\mathrm{DFT}}|$ for models trained with AL and random/MACE-OFF sampling, respectively. Insets give the mean unsigned error. Color scales are in kcal mol$^{-1}$; axes and contour levels of the error plots are matched along the rows for direct comparison. In the reference PES plots, the energy scale is from blue (energy minimum) to red (energy maximum). In the error plots, the absolute energies are plotted from white (no error) to red (maximum error), with hashes indicating areas of underestimated energies (negative error).

3.2 kcal $mol^{-1}$. In the I-O contact, which is an energetically weaker interaction, both methods perform well across the entire PES, as with the acetone–acetone contact. AL again delivers a small improvement over the Random Sampling strategy, though both schemes achieve very low energy errors of 0.2 and 0.5 kcal $mol^{-1}$, respectively.

Overall, it appears that a combination of the two sampling strategies might be most effectively employed to leverage their strengths. In the cases of solvent contacts, both strategies appear to be equally capable of exploring the PES, as is the case with the supramolecular host in solvent. Contacts of the charged ion in solvent, as well as the binding surface of the ion with the host, appear to be more challenging. For the former, the contrasting error regions may result in a more stable potential if the two datasets are used synergistically, while for the latter, the number of restrained sampling data points must increase in order to increase the coverage of the space. While there are many acetone-acetone, $Cl^-$-acetone, and I-acetone contacts in the solvated structures, the gas phase sampling of the I-$Cl^-$ contact only gets a single reference for the interaction in each training data point. A combination of active learning and periodic boundary condition simulations may also be beneficial, in order to overcome some of the issues outlined earlier with the cluster strategy.

### 5.3.3 Benchmarking Ternary Mixtures

Having benchmarked both the AL and the random sampling approaches on binary interactions that compose the experimental system, it was next important to see how these data sets could be efficiently combined to train an MLIP capable of simulating the entire ternary system of the binder and chloride in solution. The true validation task of the final MLIP is to reproduce some experimental observables, most crucially the binding energy of the chloride ion to **5.2**. Unfortunately, the large computational cost of carrying out umbrella sampling each time new data

is added to the overall training set in order to gauge convergence makes using umbrella sampling as a convergence metric completely untenable. This means that a different, independent validation set had to be constructed as a surrogate for validating the MLIP's ability to correctly simulate the entire experimental solvated system.

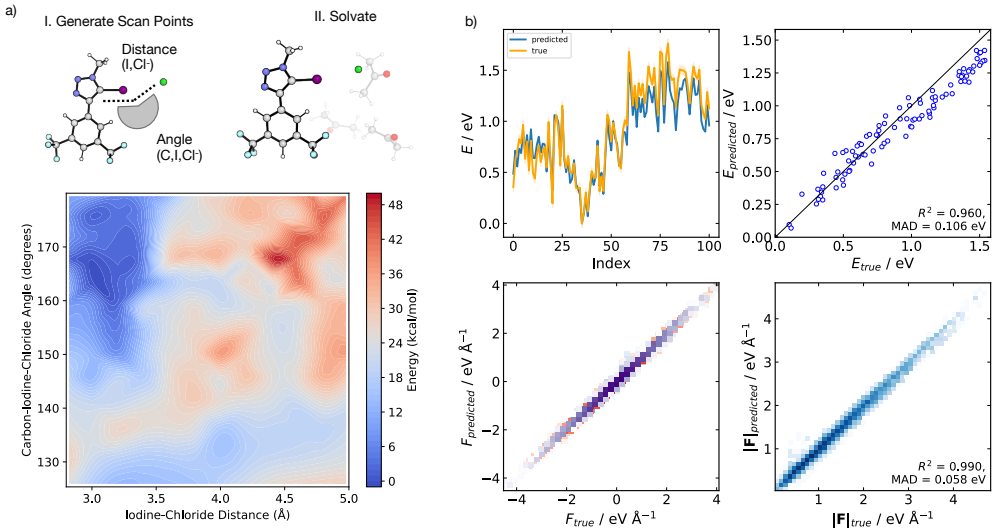



**Figure 5.10:** Ternary mixture benchmarking. a) The process of constructing the randomly microsolvated 2D scan of the $Cl^-$ at different distances and angles away from the iodine binding site, as well as the reference 2D energy of this benchmark. b) The parity plot showing a comparison of the true and predicted energies and forces on the microsolvated $Cl^-$ pulling validation set.

In the umbrella sampling, there are three key different regions, where the MLIP would have to perform well in order to capture the binding curve correctly (Figure 5.4a). Firstly, it would need to correctly predict the forces and energies acting on the bound **5.2**·$Cl^-$ complex, in order to estimate the depth of well and the force needed to pull the chloride away from this bound pose. Next, the geometries of the chloride being displaced by the acetone as it pulls away would need to be well-sampled, as inaccuracies in this region of space may have a great effect on both the enthalpic and the entropic components of the free energy, and thus the height of the barrier for the chloride's displacement. Finally the interactions of both the

chloride and the transporter with the bulk solvent would need to be represented in the training set, as this may have an impact on the unbound energy baseline, and thus impact the overall binding energy.

To asses how the different training sets described in section 5.3.2 would fare on the umbrella sampling regions, two different validation strategies were employed. The first consisted of sampling the region of space where the ion is near the binding site (Figure 5.10a), by performing a combined distance and angle scan in 2D, in the same manner as the binary component energy benchmark. These structures were then solvated up to the experimental density in a full box of solvent, whereupon all but the three solvent molecules nearest to the mid-point of the iodine and chlorine atoms were removed (using the microsolvation functionality) and single point energies calculated. Despite the random placement of solvent and lack of optimisation of these structures, the 2D energy profile in 5.10b qualitatively shows the expected appearance, with a minimum at 180deg and 2.7 Å, followed by a barrier as the chloride moves away, and another minimum at larger distances and more acute angles. This indicates that it should be a valid simulacrum of a 2D pulling trajectory of the chloride ion from **5.2**'2 binding site in acetone. The second validation set consisted of performing linear pulling simulations of the chloride away from the **5.2** host in an acetone box over 10 ps, from 2.7 to 10 Å, gathering 100 configurations. These were then microsolvated, by removing all but the nearest three acetones from each frame, and each MLIP's predictions of energies and forces on the microsolvated configurations were compared to the single point energies (Figure 5.10c).

We tested the combined four different datasets (results in Table 5.7). The first two were the combined active learning configurations (AL) and the combined random sampling configurations (Random). The combined training set of all of the active learning data, together with the **5.2**·$Cl^-$ binding distance and angle restrained scan

using MACE-OFF (Best of), was then also tested. Finally, the combination of all AL and random sampling data was tested (Total Combined).

**Table 5.6:** Energies and Forces Mean Absolute Error (MAE) for different test sets in microsolvation and MD scenarios. Lower energies indicate better performance.

| Test Set | Microsolvated Rigid Scan | | Biased MD Pulling | |
|---|---|---|---|---|
| | Energies MAE (meV) | Forces MAE (meV/Å) | Energies MAE (meV) | Forces MAE (meV/Å) |
| AL | 442 | 60 | 388 | 63 |
| Random | 271 | 58 | 106 | 58 |
| Best Of | 247 | 56 | 182 | 57 |
| Total Combined | 271 | 49 | 174 | 52 |

The ternary benchmarks confirm the complementary roles of the two sampling strategies (Table 5.7). In the microsolvated rigid scan validation set, training on Random data lowers the energy MAE substantially relative to AL (from 442 to 271 meV), which is possibly a result of the AL training set's catastrophic lack of data on the I-$Cl^-$ binding coordinate. Augmenting AL with a targeted bound–unbound scan in the "Best Of" set confirms this suspicion and yields the lowest microsolvation energy MAE (247 meV) as well as an improvement in force MAE (from 60 to 56 meV/Å). Crucially, combining all data in the "Total Combined" set preserves the low energy error of the Random and Best Of training sets (271 meV, on par with Random) while delivering the best force accuracy (49 meV/Å). Besides including ternary interactions not present in the training sets, microsolvated configurations may accentuate short-range, anisotropic interactions and local solvent structuring. It appears that in the Microsolvated Rigid Scan test set, the Random Sampling approach may yield a good enough starting point, given its significantly cheaper cost than active learning, to be used as a starting point for gathering further configurations featuring all three chemical species.

In the biased pulling MD validation, which is intended to test the model along

the full $Cl^-$ departure pathway, *as it is captured by each model*, as opposed to being a static set like the Rigid Scan — the same pattern is amplified. Training on Random configurations achieves the lowest energy MAE (error of 106 meV) and clearly outperforms AL (388 meV). This reflects the strength of Random sampling at covering the phase space encountered during the pull. The Best Of set (182 meV) sits between Random and AL, indicating that the restrained scan of I-$Cl^-$ binding added to AL in this set helps improve the model, but that the larger amount of phase space covered by the random sampling is still better for stability. Again, the Total Combined dataset offers the best forces (52 meV/Å), which is desirable balance for free-energy workflows in which stable dynamics and accurate mean forces along the reaction coordinate are pivotal. Nonetheless, these improvements likely do not justify the greater cost of using AL in such a great capacity. In summary these benchmarks argue that a hybrid training curriculum of training on two-component systems is likely the best for providing an accurate MLIP for a ternary system at this stage. Following seeding with Random Sampling to secure global coverage, exploring key areas of interest with denser AL sampling (and selective restrained scans) in the bound basin and displacement region is the ideal training regimen. For a cheaper alternative, using purely Random Sampling is likely to lead to similarly accurate results.

### 5.3.4 Umbrella Sampling

In order to carry out umbrella sampling, the binding coordinate was divided into 8 windows. The chloride binding coordinate restraints were centered on distances from 2.5 Å to 9.5 Å in 1 Å intervals, and 100 ps of sampling was performed for each window.

When attempting the umbrella sampling on the different sets tested in Table 5.7, none of the better-performing sets - Random, Best Of or Total Combined - could

run stable dynamics for all 8 windows over 100 ps. Two final strategies were therefore devised to add data of the entire combined system over the three regions of the US sampling coordinate (regions shown in Figure 5.4a).

The first method (MACE-OFF pull) was to perform a pulling trajectory using MACE-OFF over 10 ps, from 2.5 to 10 Å, and sample configurations from this trajectory, spread uniformly over the entire span of pulling. These configurations are then microsolvated, keeping seven solvent molecules, before being labelled with single point calculation energies and forces, and added to the training data. This approach is analogous to the construction of the ternary test set in the previous section (*vide supra*).

As an alternative, data were also added in a more context-dependent manner (Active Learning), using a combination of metadynamics and microsolvation. Specifically, short well-tempered metadynamics of the full ternary system were propagated using an MLIP trained using the "Random" dataset from the previous section as an initial training set. The collective variable for the metadynamics was the I-Cl distance, and each round of new trajectories in the AL sampling was begun from a bound configuration, in order to facilitate the AL exploration slowly moving through the three regions of interest - from the bound to the unbound (Figure 5.4a). Each trajectory was microsolvated, and the microsolvated frames were added to the training set based on similarity criteria. This represents (to the author's best knowledge) the first example of incorporating microsolvation into active learning.

We combined 100 or 200 configurations from the two ternary sampling strategies to the Random dataset and used the Microsolvated Rigid Scan and Biased MD Pulling test sets from Section 5.3.3 to gauge convergence. Across both validation sets the Active Learning strategy leveraging metadynamics clearly outperforms the

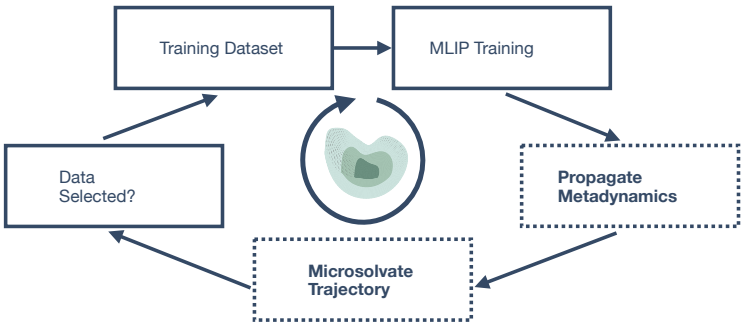



**Figure 5.11:** Metadynamics-microsolvation data augmentation loop for the ternary system. Starting from the current Training Dataset (initially the "Random" dataset, the MLIP is retrained and then used to propagate metadynamics in the full solvent. Trajectory frames are microsolvated (retain the nearest solvent molecules around the I/Cl midpoint) and screened (Data selected) using the similarity criterion. Selected configurations are labelled with single-point DFT energies and forces and merged back into the training set, closing the loop.

MACE-OFF Pulling strategy. Interestingly, adding 200 points degraded average accuracy relative to 100 for both sampling routes, going from 195 to 265 meV in the AL strategy on the microsolvated scan and 90 to 127 meV on biased MD, while the MACE-OFF Pulling deteriorates from 232 to 390 meV and 152 to 237 meV in the same test set. This pattern suggests a trade-off rather than just overfitting. The second hundred configurations likely introduce harder, boundary-region geometries that broaden space coverage and reduce the likelihood of catastrophic simulation failures errors but increase the error on the particular test sets used here.

Consistent with that interpretation, umbrella sampling (US) stability shows the opposite trend: 100-point models failed to sustain all eight windows, whereas 200-point models were able to simulate the full umbrella sampling trajectory. In other words, the extra data appear to improve the robustness of the models. By mitigating catastrophic excursions in challenging regions, they perform much better in the use case, showing a limitation of using microsolvated validation sets

**Table 5.7:** Benchmarking incremental data augmentation for the ternary MLIP. Models were retrained with either 100 or 200 additional configurations gathered by (i) MACE-OFF pulling along the binding coordinate or (ii) active learning. Performance is reported as mean absolute errors (MAE) in energies and forces on two independent validation sets: a microsolvated rigid 2D scan near the binding geometry and biased MD pulling trajectories. Lower values indicate better performance.

| Test Set | Microsolvated Rigid Scan | | Biased MD Pulling | |
|---|---|---|---|---|
| | Energies MAE (meV) | Forces MAE (meV/Å) | Energies MAE (meV) | Forces MAE (meV/Å) |
| MACE-OFF Pull (100) | 232 | 53 | 152 | 40 |
| MACE-OFF Pull (200) | 390 | 54 | 237 | 42 |
| Active Learning (100) | 195 | 45 | 90 | 34 |
| Active Learning (200) | 265 | 47 | 127 | 38 |

to extrapolate performance for performance at this stage. The final, production umbrella sampling runs were carried out with the better-performing MLIP trained on the Active Learning dataset with 200 added structures, and the 400 structures previously gathered using Random Sampling using MACE-OFF.

The umbrella sampling coordinate in Figure 5.12 indicates that this final MLIP's binding energy prediction of -4.0 kcal $mol^{-1}$ is better than the classical forcefield's prediction of -1.9 kcal $mol^{-1}$, when benchmarked against the experimental binding free energy of -3.2 kcal $mol^{-1}$. Most astonishingly, the MLIP was trained on a total of only 600 structures, which speaks volumes about the efficiency of the MACE framework. It also indicates that the data aquisition strategy that was employed in this work is an efficient way of ensuring stable and accurate dynamics, which can possibly be used predict ion binding within the limits of chemical accuracy. The current limitation is that these umbrella sampling trajectories utilising MACE-based MLIPs can only feasibly be run for 100s of picoseconds per window, if

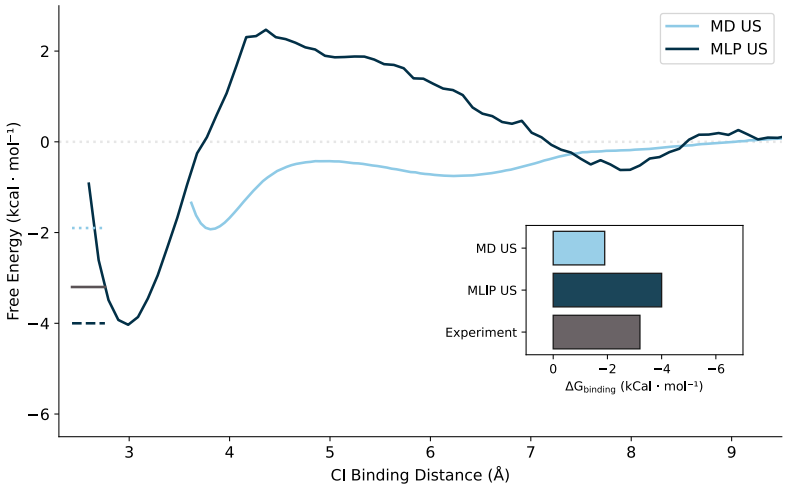



**Figure 5.12:** Plot showing the comparison of umbrella sampling free energy profiles from classical molecular dynamics (light blue) and the final MLIP (dark blue), as well as the experimental binding value (gray). The binding coordinate was divided into 8 windows, spaced 1 Å apart, from 2.5 Å to 9.5 Å. Classical MD was performed for 100 ns per window, whereas the MLIP trajectories were only performed for 100 ps.

they wish to compete with traditional FFs on cost. On the same Nvidia A40 GPU architecture, the MACE MLIP dynamics with the parameters from the present work were able to complete around 100 ps per day on a 2000 atom simulation box, while the same size of simulation box would be able to run close to a $\mu$s using the CUDA-accelerated GROMACS package.

## 5.4 Conclusions and Future Work

In this chapter, the task of predicting supramolecular anion binding in solution was tackled in a novel way, using machine learned potentials. By thoroughly benchmarking DFT methods, the appropriate level of theory was chosen across a representative set of literature halogen-binding candidates (**5.1**-**5.4**) with well-characterised binding energies.[52, 116, 184] Novel functionalities were developed

in the *mlp-train* package in order to assist in sampling relevant configurations for the MLIP training set as efficiently as possible. Acquisition strategies were benchmarked across custom test sets in order to increase this efficiency further.

The final umbrella sampling energy profile in Fig. 5.12 shows that the final MACE–MLIP produces a binding free energy closer to experiment than the classical force field. Such a level of agreement is surprising given the compact training set, since only 600 structures were required to reach stable dynamics and a realistic well depth. The result is as much a statement about the efficiency of the MACE architecture as it is about the data acquisition strategy, with which broad space coverage was demonstrated across two different sampling strategies. These strategies combined foundation model-based random sampling and microsolvation-assisted active learning. This produced a model that remains stable across the bound, displacement, and unbound regions of the coordinate. Two main limitations remain computational cost and wall time. It is also possible that instabilities caused by gaps in the data would still appear when longer simulation times are possible. In conclusion, the MLIP narrows the accuracy gap to *ab initio* methods while still trailing classical MD in throughput by orders of magnitude.

The next step is to make the approach more robust as well as generalise it. Firstly, by extending the full workflow to the remaining targets in this series (**5.1**, **5.3** and **5.4**), making sure that our conclusions do not hinge on a single binder and solvent environment. Secondly, it is crucial to quantify the uncertainty by repeating the umbrella windows with independent velocities. Thirdly, there is a need to assess convergence more robustly by simulating windows for longer and adding further configurations across both sampling streams (the random sampling of binary mixtures and active learning of ternary system configurations). This would give an indication of whether the amount of data gathered is truly the convergence limit, or if the umbrella sampling energy surface will transform with further

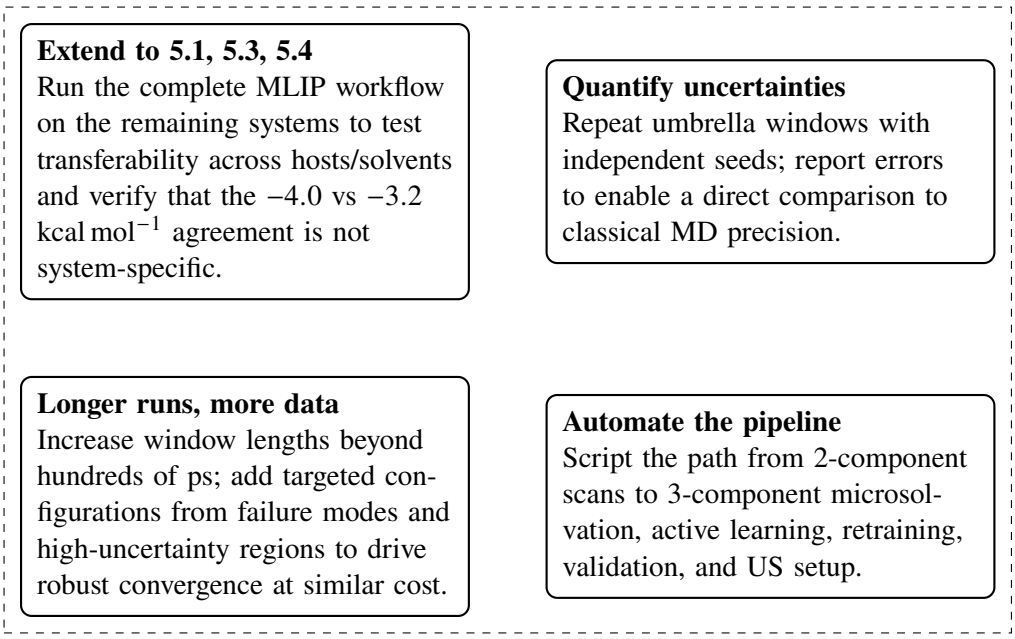



**Figure 5.13:** Planned extensions of the MLIP study. Scope, uncertainties, convergence of sampling and automation should all be advanced to ensure the practicability and accuracy of these methods.

additions to the training set. Finally, the pipeline should be changed from a semi-manual workflow to an end-to-end automated utility, which can be used easily by computational chemists. This should include preparing the two-component scans and random sampling simulations, as well as three-component, microsolvated training, model retraining, validation, and production umbrella sampling. All of these steps together will truly show whether the cost curve of MACE–MLIPs can be bent far enough to make routine umbrella sampling practical at scale.

# 6 |

# Conclusions And Outlook

## 6.1 Thesis Overview

The first aim of this thesis was to understand how small molecules can be designed to move $Cl^-$ across lipid bilayers with high activity without also facilitating $OH^-$ transport. In order to address this issue at its core, the route taken was two–pronged. The first avenue involved the iterative synthesis and testing of new transporters. Alongside this, a computational framework was built that could explain those observations mechanistically, expose their limits, and open avenues to the next set of research questions.

The macrocycles described in Chapter 2 demonstrated the power of this approach. By marrying partial encapsulation with directional $\sigma$–hole interactions, these molecules delivered nanomolar $Cl^-$ transport and successfully achieved the goal of strong preference for $Cl^-$ over $OH^-$ transport in POPC vesicles. In their synthesis, important facts were gleamed about the strain-dependent nature of macrocyclisation reactions. Flexible aliphatic linkers closed efficiently under double click cyclisations, whereas terephthalic variants did not, already hinting that cyclisation is difficult when preorganisation and strain must be balanced. Binding in mixed acetone–$D_2O$ showed improved binding in the larger rings, yet activity ultimately depended on more than affinity. Solubility, lipophilic balance, and deliverability to pre–formed membranes differentiated the viable macrocycles from those that failed to transport despite respectable $Cl^-$ binding.

The computational study assembled in Chapter 3 carefully built a holistic mechanistic picture around these findings. Extensive conformational sampling, followed by DFT optimisations with membrane–mimetic solvation first compared the intrinsic ion binding preference for the two different ions. It was shown that cyclisation decreases the preference for $OH^-$ found in an acyclic parent by decreasing its interaction advantage rather than by increasing distortion costs. Entropic analysis added important context about the way that vibrational modes in the macrocycles stiffen more for $OH^-$ than for $Cl^-$, imposing an extra penalty that the larger halide circumvents. Explicit–solvent metadynamics confirmed that the most selective macrocycle accesses convergent *cis*- poses over low barriers, consistent with a dynamically preorganised pocket rather than the transporter being forced into a single rigid conformation. Equilibrium MD at the membrane interface quantified the desolvation experienced by the ions following binding site encapsulation. Both ions lose water upon binding, but dehydration penalties dictate that $OH^-$ should suffer higher penalties. These results offer an explanation for high $Cl^-$>$OH^-$ selectivity that does not rely on any single effect but on a combination of geometric preorganisation, encapsulation and desolvation.

Chapter 4 applied the methods developed in studying the chloride-selective transporters to expand the scope to stimuli–responsive systems. Hydrazone–based transporters were found to outperform the amide analogues while preserving the same design logic. In both, intramolecular hydrogen bonding was used as a stimuli-responsive trigger for preorganisation, which turned on chloride binding and membrane transport. In this chapter, it was seen how using a single method to study binding processes can be a limiting factor in understanding the mechanism at play. DFT and distortion–interaction analysis captured how the active hydrazone requires less energy reorganise and gains more from binding, while the protected form is trapped in inactivity due to its higher reorganisational higher barriers. In

explicitly solvated classical dynamics, free–energy profiles told a more nuanced story, with solvent–assisted intermediates and dissociation pathways becoming visible. Together these results argue that predictive modelling in this field must combine both elements of DFT and MD.

The final chapter turned this idea into reality. By training machine–learning interatomic potentials on DFT data and utilising them to perform umbrella sampling in explicit solvent, the gap between *ab initio* accuracy and molecular–dynamics environment realism was bridged. To achieve this, several new functionalities were added to the Duarte group's *mlp-train* Python package, and novel sampling strategies and benchmarking methods trialled. An MLIP trained on a set of only 600 structures was able to perform umbrella sampling, with stable dynamics across the entire sampling trajectory, which suggests very high data efficiency of the methods employed. This also underlines the importance of careful data acquisition strategies. The cost of MLIP simulations, which remain orders of magnitudes slower than classical MD, remains a challenge. Yet when time comes for the simulations to be quick enough to challenge the status quo, the methods outlined here may be used to generate robust and accurate machine learned potentials with predictive power.

## 6.2 Future Directions

Looking at the combined work on chloride transporters presented here, it can be seen that the recipe for selectivity is composed of two main ingredients: how the binding site is preorganised for $Cl^-$ rather than $OH^-$ and how much desolvation is enforced on the ion at the interface. The best transporters found here do not exhibit the strongest binding but instead the right type of binding. Encapsulation helps when it enforces geometry and excludes water, but it becomes a problem when

the means of achieving it incites strained geometries or moves the molecule away from the lipophilic optimum needed for delivery. Electronics sharpen the $\sigma$–hole and can tune $ESP_{max}$, yet they must be deployed with an eye on clogP, solubility, and conformational effects in the future.

The outlook on the experimental side looks to the grand aims of the field of using transporters to address pharmacological needs. To develop even more selective transporters, the macrocyclic platform demands further increases in enclosure. Further rigidification may also be attempted to match the binding site size more closely to a specific ion, without employing overly strained geometries that result in failed cyclisations. Stimuli–responsive linkers or appendages to the macrocycles could be used to gate macrocycle conformation and binding site availability in time and space, with light, redox or pH as triggers. Translation will require work in epithelial models that report on airway surface liquid height and pH, alongside a broad safety panel that rules out protonophoric liabilities and tracks behaviour in complex media.

In aiding these developments of chloride transporters towards the clinic, the computational side must move from providing plausible explanations to giving quantitative prediction. Machine–learned potentials offer a route to bring quantum chemical accuracy and fidelity into longer and larger simulations. Their success will depend on disciplined training, uncertainty quantification, and tight coupling to experimental observables. The macrocycles developed here, the mechanistic picture assembled across theory and simulation, and the first steps toward ML–accelerated free energies together suggest that small, neutral transporters can be further engineered to move $Cl^-$ with precision in complex membranes. The next advances will come from doing the same work across a broader chemical space, with more realistic models and more direct links to assay read–outs, until modelling proposes not only why a transporter works but how to make the next

one better.

# Bibliography


(1) Cooper, G. M., *The cell: a molecular approach*, 2nd ed; ASM Press: Washington, D.C. Sunderland, Mass, 2000.

(2) Banfalvi, G., *Permeability of Biological Membranes*, 1st ed. 2016; Springer International Publishing : Imprint: Springer: Cham, 2016.

(3) Stillwell, W., *An introduction to biological membranes: composition, structure and function*, 2nd ed; Academic press, an imprint of Elsevier: London, 2016.

(4) Gale, P. A.; Davis, J. T.; Quesada, R. Anion transport and supramolecular medicinal chemistry. *Chem. Soc. Rev.* **2017**, *46*, 2497–2519.

(5) Alberts, B., *Molecular Biology of the Cell*, 6th ed.; W.W. Norton & Company: New York, 2017.

(6) Blaesse, P.; Airaksinen, M. S.; Rivera, C.; Kaila, K. Cation-chloride cotransporters and neuronal function. *Neuron* **2009**, *61*, 820–838.

(7) Alper, S. L. Molecular physiology of SLC4 anion exchangers. *Experimental Physiology* **2006**, *91*, 153–161.

(8) Jentsch, T. J.; Pusch, M. CLC Chloride Channels and Transporters: Structure, Function, Physiology, and Disease. *Physiological reviews* **2018**, *98*, 1493–1590.

(9) Planells-Cases, R.; Jentsch, T. J. Chloride channelopathies. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2009**, *1792*, 173–189.

(10) Elborn, J. S. Cystic fibrosis. *The Lancet* **2016**, *388*, 2519–2531.

(11) Ramsey, B. W.; Davies, J.; McElvaney, N. G.; Tullis, E.; Bell, S. C.; Dřevínek, P.; Griese, M.; McKone, E. F.; Wainwright, C. E.; Konstan, M. W.; Moss, R.; Ratjen, F.; Sermet-Gaudelus, I.; Rowe, S. M.; Dong, Q.; Rodriguez, S.; Yen, K.; Ordoñez, C.; Elborn, J. S. A CFTR Potentiator

in Patients with Cystic Fibrosis and the *G551D* Mutation. *New England Journal of Medicine* **2011**, *365*, 1663–1672.

(12) Keating, D.; Marigowda, G.; Burr, L.; Daines, C.; Mall, M. A.; McKone, E. F.; Ramsey, B. W.; Rowe, S. M.; Sass, L. A.; Tullis, E.; McKee, C. M.; Moskowitz, S. M.; Robertson, S.; Savage, J.; Simard, C.; Van Goor, F.; Waltz, D.; Xuan, F.; Young, T.; Taylor-Cousar, J. L. VX-445–Tezacaftor–Ivacaftor in Patients with Cystic Fibrosis and One or Two Phe508del Alleles. *New England Journal of Medicine* **2018**, *379*, 1612–1620.

(13) Middleton, P. G.; Mall, M. A.; Dřevínek, P.; Lands, L. C.; McKone, E. F.; Polineni, D.; Ramsey, B. W.; Taylor-Cousar, J. L.; Tullis, E.; Vermeulen, F.; Marigowda, G.; McKee, C. M.; Moskowitz, S. M.; Nair, N.; Savage, J.; Simard, C.; Tian, S.; Waltz, D.; Xuan, F.; Rowe, S. M.; Jain, R. Elexacaftor–Tezacaftor–Ivacaftor for Cystic Fibrosis with a Single Phe508del Allele. *New England Journal of Medicine* **2019**, *381*, 1809–1819.

(14) Davis, J. T.; Gale, P. A.; Quesada, R. Advances in anion transport and supramolecular medicinal chemistry. *Chem. Soc. Rev.* **2020**, *49*, 6056–6086.

(15) Davis, A. P.; Sheppard, D. N.; Smith, B. D. Development of synthetic membrane transporters for anions. *Chem. Soc. Rev.* **2007**, *36*, 348–357.

(16) Mcnally, B. A.; Leevy, W. M.; Smith, B. D. Recent Advances in Synthetic Membrane Transporters. *Supramolecular Chemistry* **2007**, *19*, 29–37.

(17) Shao, B.; Fu, H.; Aprahamian, I. A molecular anion pump. *Science* **2024**, *385*, 544–549.

(18) Nguyen, B.-N. T.; Grommet, A. B.; Tron, A.; Georges, M. C. A.; Nitschke, J. R. Heat Engine Drives Transport of an FeII4L4 Cage and Cargo. *Advanced Materials* **2020**, *32*, 1907241.

(19) Kaushik, V.; Yakisich, J. S.; Kumar, A.; Azad, N.; Iyer, A. K. V. Ionophores: Potential Use as Anticancer Drugs and Chemosensitizers. *Cancers* **2018**, *10*, 360.

(20) Ayling, A. J.; Pérez-Payán, M. N.; Davis, A. P. New "Cholapod" Anionophores; High-Affinity Halide Receptors Derived from Cholic Acid. *J. Am. Chem. Soc.* **2001**, *123*, 12716–12717.

(21) McNally, B. A.; Koulov, A. V.; Lambert, T. N.; Smith, B. D.; Joos, J.-B.; Sisson, A. L.; Clare, J. P.; Sgarlata, V.; Judd, L. W.; Magro, G.; Davis, A. P. Structure–Activity Relationships in Cholapod Anion Carriers: Enhanced Transmembrane Chloride Transport through Substituent Tuning. *Chem. Eur. J.* **2008**, *14*, 9599–9606.

(22) Hussain, S.; Brotherhood, P. R.; Judd, L. W.; Davis, A. P. Diaxial Diureido Decalins as Compact, Efficient, and Tunable Anion Transporters. *J. Am. Chem. Soc.* **2011**, *133*, 1614–1617.

(23) Valkenier, H.; Judd, L. W.; Li, H.; Hussain, S.; Sheppard, D. N.; Davis, A. P. Preorganized Bis-Thioureas as Powerful Anion Carriers: Chloride Transport by Single Molecules in Large Unilamellar Vesicles. *J. Am. Chem. Soc.* **2014**, *136*, 12507–12512.

(24) Li, H.; Valkenier, H.; Judd, L. W.; Brotherhood, P. R.; Hussain, S.; Cooper, J. A.; Jurček, O.; Sparkes, H. A.; Sheppard, D. N.; Davis, A. P. Efficient, non-toxic anion transport by synthetic carriers in cells and epithelia. *Nat. Chem.* **2016**, *8*, 24–32.

(25) Cooper, J. A.; Street, S. T. G.; Davis, A. P. A Flexible Solution to Anion Transport: Powerful Anionophores Based on a Cyclohexane Scaffold. *Angew. Chem. Int. Ed.* **2014**, *53*, 5609–5613.

(26) Karagiannidis, L. E.; Hiscock, J. R.; Gale, P. A. The influence of stereochemistry on anion binding and transport. *Supramolecular Chemistry* **2013**, *25*, 626–630.

(27) J. Moore, S.; E. Haynes, C. J.; González, J.; L. Sutton, J.; J. Brooks, S.; E. Light, M.; Herniman, J.; John Langley, G.; Soto-Cerrato, V.; Pérez-Tomás, R.; Marques, I.; J. Costa, P.; Félix, V.; A. Gale, P. Chloride, carboxylate and carbonate transport by ortho -phenylenediamine-based bisureas. *Chem. Sci.* **2013**, *4*, 103–117.

(28) Olivari, M.; Montis, R.; N. Berry, S.; E. Karagiannidis, L.; J. Coles, S.; N. Horton, P.; K. Mapp, L.; A. Gale, P.; Caltagirone, C. Tris-ureas as transmembrane anion transporters. *Dalton Transactions* **2016**, *45*, 11892–11897.

(29) Busschaert, N.; Wenzel, M.; Light, M. E.; Iglesias-Hernández, P.; Pérez-Tomás, R.; Gale, P. A. Structure–Activity Relationships in Tripodal Transmembrane Anion Transporters: The Effect of Fluorination. *J. Am. Chem. Soc.* **2011**, *133*, 14136–14148.

(30) Picci, G.; Marchesan, S.; Caltagirone, C. Ion Channels and Transporters as Therapeutic Agents: From Biomolecules to Supramolecular Medicinal Chemistry. *Biomedicines* **2022**, *10*, 885.

(31) Ren, C.; Ding, X.; Roy, A.; Shen, J.; Zhou, S.; Chen, F.; Li, S. F. Y.; Ren, H.; Yang, Y. Y.; Zeng, H. A halogen bond-mediated highly active artificial chloride channel with high anticancer activity. *Chem. Sci.* **2018**, *9*, 4044–4051.

(32) Li, X.; Shen, B.; Yao, X.-Q.; Yang, D. Synthetic Chloride Channel Regulates Cell Membrane Potentials and Voltage-Gated Calcium Channels. *J. Am. Chem. Soc.* **2009**, *131*, 13676–13680.

(33) Mondal, A.; Barik, G. K.; Sarkar, S.; Mondal, D.; Ahmad, M.; Vijayakanth, T.; Mondal, J.; Santra, M. K.; Talukdar, P. Nontoxic Artificial Chloride Channel Formation in Epithelial Cells by Isophthalic Acid-Based Small Molecules. *Chem. Eur. J.* **2023**, *29*, e202202887.

(34) Duran, C.; Thompson, C. H.; Xiao, Q.; Hartzell, C. Chloride Channels: Often enigmatic, rarely predictable. *Annual review of physiology* **2010**, *72*, 95–121.

(35) Li, X.; Shen, B.; Yao, X.-Q.; Yang, D. A Small Synthetic Molecule Forms Chloride Channels to Mediate Chloride Transport across Cell Membranes. *J. Am. Chem. Soc.* **2007**, *129*, 7264–7265.

(36) Shen, B.; Li, X.; Wang, F.; Yao, X.; Yang, D. A Synthetic Chloride Channel Restores Chloride Conductance in Human Cystic Fibrosis Epithelial Cells. *PLoS One* **2012**, *7*, e34694.

(37) Zeng, Z. W.; Linsdell, P.; Pomès, R. Molecular dynamics study of Cl permeation through cystic fibrosis transmembrane conductance regulator (CFTR). *Cellular and Molecular Life Sciences* **2023**, *80*, 51.

(38) Huang, W.-L.; Wang, X.-D.; Ao, Y.-F.; Wang, Q.-Q.; Wang, D.-X. Artificial Chloride-Selective Channel: Shape and Function Mimic of the ClC Channel Selective Pore. *Journal of the American Chemical Society* **2020**, *142*, 13273–13277.

(39) Ghadiri, M. R.; Granja, J. R.; Buehler, L. K. Artificial transmembrane ion channels from self-assembling peptide nanotubes. *Nature* **1994**, *369*, 301–304.

(40) Sharma, R.; Sarkar, S.; Chattopadhayay, S.; Mondal, J.; Talukdar, P. A Halogen-Bond-Driven Artificial Chloride-Selective Channel Constructed from 5-Iodoisophthalamide-based Molecules. *Angew. Chem. Int. Ed.* **2024**, *63*, e202319919.

(41) Saha, T.; Gautam, A.; Mukherjee, A.; Lahiri, M.; Talukdar, P. Chloride Transport through Supramolecular Barrel-Rosette Ion Channels: Lipophilic Control and Apoptosis-Inducing Activity. *J. Am. Chem. Soc.* **2016**, *138*, 16443–16451.

(42) Yan, T.; Liu, J. Transmembrane Ion Channels: From Natural to Artificial Systems. *Angew. Chem. Int. Ed.* **2025**, *64*, e202416200.

(43) McNally, B. A.; O'Neil, E. J.; Nguyen, A.; Smith, B. D. Membrane Transporters for Anions That Use a Relay Mechanism. *J. Am. Chem. Soc.* **2008**, *130*, 17274–17275.

(44) Johnson, T. G.; Sadeghi-Kelishadi, A.; Langton, M. J. A Photo-responsive Transmembrane Anion Transporter Relay. *J. Am. Chem. Soc.* **2022**, *144*, 10455–10461.

(45) Johnson, T. G.; Docker, A.; Sadeghi-Kelishadi, A.; Langton, M. J. Halogen bonding relay and mobile anion transporters with kinetically controlled chloride selectivity. *Chem. Sci.* **2023**, *14*, 5006–5013.

(46) Cao, R.; Rossdeutcher, R. B.; Zhong, Y.; Shen, Y.; Miller, D. P.; Sobiech, T. A.; Wu, X.; Buitrago, L. S.; Ramcharan, K.; Gutay, M. I.; Figueira, M. F.; Luthra, P.; Zurek, E.; Szyperski, T.; Button, B.; Shao, Z.; Gong, B. Aromatic pentaamide macrocycles bind anions with high affinity for transport across biomembranes. *Nat. Chem.* **2023**, *15*, 1559–1568.

(47) Hernando, E.; Capurro, V.; Cossu, C.; Fiore, M.; García-Valverde, M.; Soto-Cerrato, V.; Pérez-Tomás, R.; Moran, O.; Zegarra-Moran, O.; Quesada, R. Small molecule anionophores promote transmembrane anion permeation matching CFTR activity. *Sci. Rep.* **2018**, *8*, 2608.

(48) Mukherjee, S.; Shinde, S. V.; Talukdar, P.; Haldar, J. Unveiling the potent activity of a synthetic ion transporter against multidrug-resistant Gram-

positive bacteria and biofilms. *RSC Medicinal Chemistry* **2024**, *15*, 2127–2137.

(49) Ko, S.-K.; Kim, S. K.; Share, A.; Lynch, V. M.; Park, J.; Namkung, W.; Van Rossom, W.; Busschaert, N.; Gale, P. A.; Sessler, J. L.; Shin, I. Synthetic ion transporters can induce apoptosis by facilitating chloride anion transport into cells. *Nat. Chem.* **2014**, *6*, 885–892.

(50) Wu, X.; Judd, L. W.; Howe, E. N. W.; Withecombe, A. M.; Soto-Cerrato, V.; Li, H.; Busschaert, N.; Valkenier, H.; Pérez-Tomás, R.; Sheppard, D. N.; Jiang, Y.-B.; Davis, A. P.; Gale, P. A. Nonprotonophoric Electrogenic Cl Transport Mediated by Valinomycin-like Carriers. *Chem* **2016**, *1*, 127–146.

(51) Busschaert, N.; Park, S.-H.; Baek, K.-H.; Choi, Y. P.; Park, J.; Howe, E. N. W.; Hiscock, J. R.; Karagiannidis, L. E.; Marques, I.; Félix, V.; Namkung, W.; Sessler, J. L.; Gale, P. A.; Shin, I. A synthetic ion transporter that disrupts autophagy and induces apoptosis by perturbing cellular chloride concentrations. *Nat. Chem.* **2017**, *9*, 667–675.

(52) Bickerton, L. E.; Sterling, A. J.; Beer, P. D.; Duarte, F.; Langton, M. J. Transmembrane anion transport mediated by halogen bonding and hydrogen bonding triazole anionophores. *Chem. Sci.* **2020**, *11*, 4722–4729.

(53) Singh, A.; Torres-Huerta, A.; Vanderlinden, T.; Renier, N.; Martínez-Crespo, L.; Tumanov, N.; Wouters, J.; Bartik, K.; Jabin, I.; Valkenier, H. Calix[6]arenes with halogen bond donor groups as selective and efficient anion transporters. *Chem. Commun.* **2022**, *58*, 6255–6258.

(54) Flerin, M.; Duarte, F.; Langton, M. J. Chloride Selective, Nonprotonophoric Ion Transport with Macrocyclic Halogen Bonding Anionophores. *Chem. Eur. J.* **2025**, *31*, e202502033.

(55) Bickerton, L. E.; Docker, A.; Sterling, A. J.; Kuhn, H.; Duarte, F.; Beer, P. D.; Langton, M. J. Highly Active Halogen Bonding and Chalcogen Bonding Chloride Transporters with Non-Protonophoric Activity. *Chem. Eur. J.* **2021**, *27*, 11738–11745.

(56) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *J. Chem. Theory Comput.* **2013**, *9*, 1918–1931.

(57) Cavallo, G.; Metrangolo, P.; Milani, R.; Pilati, T.; Priimagi, A.; Resnati, G.; Terraneo, G. The Halogen Bond. *Chem. Rev.* **2016**, *116*, 2478–2601.

(58) Politzer, P.; Murray, J. S.; Clark, T. Halogen bonding: an electrostatically-driven highly directional noncovalent interaction. *Phys. Chem. Chem. Phys.* **2010**, *12*, 7748–7757.

(59) Clark, T.; Hennemann, M.; Murray, J. S.; Politzer, P. Halogen bonding: the -hole. *Journal of Molecular Modeling* **2007**, *13*, 291–296.

(60) Jentzsch, A. V.; Emery, D.; Mareda, J.; Nayak, S. K.; Metrangolo, P.; Resnati, G.; Sakai, N.; Matile, S. Transmembrane anion transport mediated by halogen-bond donors. *Nat. Commun.* **2012**, *3*, 905.

(61) Jentzsch, A. V.; Matile, S. In *Halogen Bonding I: Impact on Materials Chemistry and Life Sciences*, Metrangolo, P., Resnati, G., Eds.; Springer International Publishing: Cham, 2015, pp 205–239.

(62) Benz, S.; Macchione, M.; Verolet, Q.; Mareda, J.; Sakai, N.; Matile, S. Anion Transport with Chalcogen Bonds. *J. Am. Chem. Soc.* **2016**, *138*, 9093–9096.

(63) Saggiomo, V.; Otto, S.; Marques, I.; Félix, V.; Torroba, T.; Quesada, R. The role of lipophilicity in transmembrane anion transport. *Chem. Commun.* **2012**, *48*, 5274–5276.

(64) Valkenier, H.; Haynes, C. J. E.; Herniman, J.; Gale, P. A.; Davis, A. P. Lipophilic balance – a new design principle for transmembrane anion carriers. *Chem. Sci.* **2014**, *5*, 1128–1134.

(65) M. Dias, C.; Li, H.; Valkenier, H.; E. Karagiannidis, L.; A. Gale, P.; N. Sheppard, D.; P. Davis, A. Anion transport by ortho -phenylene bis-ureas across cell and vesicle membranes. *Organic & Biomolecular Chemistry* **2018**, *16*, 1083–1087.

(66) Akhtar, N.; Pradhan, N.; Saha, A.; Kumar, V.; Biswas, O.; Dey, S.; Shah, M.; Kumar, S.; Manna, D. Tuning the solubility of ionophores: glutathione-mediated transport of chloride ions across hydrophobic membranes. *Chem. Commun.* **2019**, *55*, 8482–8485.

(67) Srimayee, S.; Badajena, S. R.; Akhtar, N.; Kar, M. K.; Dey, S.; Mohapatra, P.; Manna, D. Stimuli-responsive release of active anionophore from RGD-peptide-linked proanionophore. *Chem. Commun.* **2023**, *59*, 12759–12762.

(68) Feo, E.; Gale, P. A. Therapeutic synthetic anion transporters. *Curr. Opin. Chem. Biol.* **2024**, *83*, 102535.

(69) Kučerka, N.; Tristram-Nagle, S.; Nagle, J. F. Structure of Fully Hydrated Fluid Phase Lipid Bilayers with Monounsaturated Chains. *The Journal of Membrane Biology* **2006**, *208*, 193–202.

(70) Murzyn, K.; Róg, T.; Jezierski, G.; Takaoka, Y.; Pasenkiewicz-Gierula, M. Effects of Phospholipid Unsaturation on the Membrane/Water Interface: A Molecular Simulation Study. *Biophysical Journal* **2001**, *81*, 170–183.

(71) Spooner, M. J.; Li, H.; Marques, I.; Costa, P. M. R.; Wu, X.; Howe, E. N. W.; Busschaert, N.; Moore, S. J.; Light, M. E.; Sheppard, D. N.; Félix, V.; Gale, P. A. Fluorinated synthetic anion carriers: experimental and computational insights into transmembrane chloride transport. *Chem. Sci.* **2019**, *10*, 1976–1985.

(72) Marques, I.; Costa, P. M. R.; Q. Miranda, M.; Busschaert, N.; Howe, E. N. W.; Clarke, H. J.; Haynes, C. J. E.; Kirby, I. L.; Rodilla, A. M.; Pérez-Tomás, R.; Gale, P. A.; Félix, V. Full elucidation of the transmembrane anion transport mechanism of squaramides using *in silico* investigations. *Phys. Chem. Chem. Phys.* **2018**, *20*, 20796–20811.

(73) Edwards, S. J.; Marques, I.; Dias, C. M.; Tromans, R. A.; Lees, N. R.; Félix, V.; Valkenier, H.; Davis, A. P. Tilting and Tumbling in Transmembrane Anion Carriers: Activity Tuning through n-Alkyl Substitution. *Chem. Eur. J.* **2016**, *22*, 2004–2011.

(74) Bickerton, L. E.; Langton, M. J. Controlling transmembrane ion transport via photo-regulated carrier mobility. *Chem. Sci.* **2022**, *13*, 9531–9536.

(75) Khodadadi, E.; Khodadadi, E.; Chaturvedi, P.; Moradi, M. Comprehensive Insights into the Cholesterol-Mediated Modulation of Membrane Function Through Molecular Dynamics Simulations. *Membranes* **2025**, *15*, 173.

(76) Doole, F. T.; Kumarage, T.; Ashkar, R.; Brown, M. F. Cholesterol Stiffening of Lipid Membranes. *The Journal of Membrane Biology* **2022**, *255*, 385–405.

(77) Carpenter, T. S.; Lau, E. Y.; Kirshner, D. A.; Lightstone, F. C. Prediction of Blood-Brain Barrier Permeability from Molecular Dynamics Simulations. *Biophysical Journal* **2014**, *106*, 83a.

(78) Wang, Y.; Gallagher, E.; Jorgensen, C.; Troendle, E. P.; Hu, D.; Searson, P. C.; Ulmschneider, M. B. An experimentally validated approach to calculate the blood-brain barrier permeability of small molecules. *Sci. Rep.* **2019**, *9*, 6117.

(79) Thai, N. Q.; Theodorakis, P. E.; Li, M. S. Fast Estimation of the Blood–Brain Barrier Permeability by Pulling a Ligand through a Lipid Membrane. *J. Chem. Inf. Model.* **2020**, *60*, 3057–3067.

(80) Langton, M. J. Engineering of stimuli-responsive lipid-bilayer membranes using supramolecular systems. *Nature Reviews Chemistry* **2021**, *5*, 46–61.

(81) Ahmad, M.; Gartland, S. A.; Langton, M. J. Photo- and Redox-Regulated Transmembrane Ion Transporters. *Angew. Chem. Int. Ed.* **2023**, *62*, e202308842.

(82) Chattopadhayay, S.; Talukdar, P. Stimuli-responsive synthetic ionophores for therapeutic applications. *Curr. Opin. Chem. Biol.* **2025**, *85*, 102582.

(83) Kar, M. K.; Mahata, R.; Srimayee, S.; Haloi, N.; Kumar, R.; Lindahl, E.; Santra, M. K.; Manna, D. -Carboline-based light and pH dual stimuli-responsive ion transporters induce cancer cell death. *Chem. Commun.* **2024**, *60*, 8419–8422.

(84) Xie, F.; Wang, M.; Chen, Q.; Chi, T.; Zhu, S.; Wei, P.; Yang, Y.; Zhang, L.; Li, X.; Liao, Z. Endogenous stimuli-responsive nanoparticles for cancer therapy: From bench to bedside. *Pharmacological Research* **2022**, *186*, 106522.

(85) Chao, X.; Johnson, T. G.; Temian, M.-C.; Docker, A.; Wallabregue, A. L. D.; Scott, A.; Conway, S. J.; Langton, M. J. Coupling Photoresponsive Transmembrane Ion Transport with Transition Metal Catalysis. *J. Am. Chem. Soc.* **2024**, *146*, 4351–4356.

(86) Bai, Y.; Chen, J.; Zimmerman, S. C. Designed transition metal catalysts for intracellular organic synthesis. *Chem. Soc. Rev.* **2018**, *47*, 1811–1821.

(87) Wang, X.; Kerckhoffs, A.; Riexinger, J.; Cornall, M.; Langton, M. J.; Bayley, H.; Qing, Y. ON–OFF nanopores for optical control of transmembrane ionic communication. *Nature Nanotechnology* **2025**, *20*, 432–440.

(88) Hardy, M. D.; Yang, J.; Selimkhanov, J.; Cole, C. M.; Tsimring, L. S.; Devaraj, N. K. Self-reproducing catalyst drives repeated phospholipid synthesis and membrane growth. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 8187–8192.

(89) Kerckhoffs, A.; Bo, Z.; Penty, S. E.; Duarte, F.; Langton, M. J. Red-shifted tetra-ortho-halo-azobenzenes for photo-regulated transmembrane anion transport. *Org. Biomol. Chem.* **2021**, *19*, 9058–9067.

(90) Baliarsingh, B.; Madhavan, N. Selective and pH-Responsive Macrocyclic Anionophores. *ChemBioChem* **2025**, 202500415.

(91) Kumar, N.; Madhavan, N. Small molecule-derived pH-gated ion transporters. *Organic & Biomolecular Chemistry* **2023**, *21*, 5892–5905.

(92) Li, Z.; Yuan, L.; Chang, W.; Liu, J.; Shen, J.; Zeng, H. A controllable photoresponsive potassium transporter. *Nat. Commun.* **2025**, *16*, 6926.

(93) Zhou, B.; P. Gabbaï, F. Redox-controlled chalcogen-bonding at tellurium: impact on Lewis acidity and chloride anion transport properties. *Chem. Sci.* **2020**, *11*, 7495–7500.

(94) Docker, A.; Johnson, T. G.; Kuhn, H.; Zhang, Z.; Langton, M. J. Multistate Redox-Switchable Ion Transport Using Chalcogen-Bonding Anionophores. *J. Am. Chem. Soc.* **2023**, *145*, 2661–2668.

(95) Ahmad, M.; Muir, A.; Langton, M. J. Off–On Photo- and Redox-Triggered Anion Transport Using an Indole-Based Hydrogen Bond Switch. *ACS Omega* **2024**, *9*, 45572–45580.

(96) Kerckhoffs, A.; E. Christensen, K.; J. Langton, M. Fast relaxing red and near-IR switchable azobenzenes with chalcogen and halogen substituents: periodic trends, tuneable thermal half-lives and chalcogen bonding. *Chem. Sci.* **2022**, *13*, 11551–11559.

(97) Kerckhoffs, A.; Langton, M. J. Reversible photo-control over transmembrane anion transport using visible-light responsive supramolecular carriers. *Chem. Sci.* **2020**, *11*, 6325–6331.

(98) Kobauri, P.; Dekker, F. J.; Szymanski, W.; Feringa, B. L. Rational Design in Photopharmacology with Molecular Photoswitches. *Angew. Chem. Int. Ed.* **2023**, *62*, e202300681.

(99) Busschaert, N.; P. Elmes, R. B.; D. Czech, D.; Wu, X.; L. Kirby, I.; M. Peck, E.; D. Hendzel, K.; K. Shaw, S.; Chan, B.; D. Smith, B.; A. Jolliffe, K.; A. Gale, P. Thiosquaramides: pH switchable anion transporters. *Chem. Sci.* **2014**, *5*, 3617–3626.

(100) Gartland, S. A.; Johnson, T. G.; Walkley, E.; Langton, M. J. Inter-Vesicle Signal Transduction Using a Photo-Responsive Zinc Ionophore. *Angew. Chem. Int. Ed.* **2023**, *62*, e202309080.

(101) Villarón, D.; Siegler, M. A.; Wezenberg, S. J. A photoswitchable strapped calix[4]pyrrole receptor: highly effective chloride binding and release. *Chem. Sci.* **2021**, *12*, 3188–3193.

(102) Jones, R. G.; Ober, C. K.; Hayakawa, T.; Luscombe, C. K.; Stingelin, N. *Pure and Applied Chemistry* **2020**, *92*, 1861–1891.

(103) Beer, P. D.; Barendt, T. A.; Lim, J. Y. C., *Supramolecular chemistry: fundamentals and applications*, 2nd ed; Oxford chemistry primers; Oxford university press: Oxford, 2022.

(104) Gale, P. A.; Caltagirone, C. Anion sensing by small molecules and molecular ensembles. *Chem. Soc. Rev.* **2015**, *44*, 4212–4227.

(105) Beer, P. D.; Gale, P. A. Anion Recognition and Sensing: The State of the Art and Future Perspectives. *Angew. Chem. Int. Ed.* **2001**, *40*, 486–516.

(106) Busschaert, N.; Caltagirone, C.; Van Rossom, W.; Gale, P. A. Applications of Supramolecular Anion Recognition. *Chem. Rev.* **2015**, *115*, 8038–8155.

(107) Yan, X.; Wang, F.; Zheng, B.; Huang, F. Stimuli-responsive supramolecular polymeric materials. *Chem. Soc. Rev.* **2012**, *41*, 6042–6065.

(108) Liu, L.; Cotelle, Y.; Klehr, J.; Sakai, N.; R. Ward, T.; Matile, S. Anion–catalysis: bicyclic products with four contiguous stereogenic centers from otherwise elusive diastereospecific domino reactions on -acidic surfaces. *Chem. Sci.* **2017**, *8*, 3770–3774.

(109) Singh, P.; Mahar, R. Cyclodextrin in drug delivery: Exploring scaffolds, properties, and cutting-edge applications. *International Journal of Pharmaceutics* **2024**, *662*, 124485.

(110) Patrick, S. C.; Beer, P. D.; Davis, J. J. Solvent effects in anion recognition. *Nature Reviews Chemistry* **2024**, *8*, 256–276.

(111) Langton, M. J.; Serpell, C. J.; Beer, P. D. Anion Recognition in Water: Recent Advances from a Supramolecular and Macromolecular Perspective. *Angew. Chem. Int. Ed.* **2016**, *55*, 1974–1987.

(112) Rizzi, A.; Murkli, S.; McNeill, J. N.; Yao, W.; Sullivan, M.; Gilson, M. K.; Chiu, M. W.; Isaacs, L.; Gibb, B. C.; Mobley, D. L.; Chodera, J. D. Overview of the SAMPL6 host–guest binding affinity prediction challenge. *J. Comput. Aided Mol. Des.* **2018**, *32*, 937–963.

(113) Amezcua, M.; El Khoury, L.; Mobley, D. L. SAMPL7 Host–Guest Challenge Overview: assessing the reliability of polarizable and non-polarizable methods for binding free energy calculations. *J. Comput. Aided Mol. Des.* **2021**, *35*, 1–35.

(114) Amezcua, M.; Setiadi, J.; Ge, Y.; Mobley, D. L. An overview of the SAMPL8 host–guest binding challenge. *J. Comput. Aided Mol. Des.* **2022**, *36*, 707–734.

(115) Amezcua, M.; Setiadi, J.; L. Mobley, D. The SAMPL9 host–guest blind challenge: an overview of binding free energy predictive accuracy. *Phys. Chem. Chem. Phys.* **2024**, *26*, 9207–9225.

(116) Beale, T. M.; Chudzinski, M. G.; Sarwar, M. G.; Taylor, M. S. Halogen bonding in solution: thermodynamics and applications. *Chem. Soc. Rev.* **2013**, *42*, 1667–1680.

(117) Wolters, L. P.; Schyman, P.; Pavan, M. J.; Jorgensen, W. L.; Bickelhaupt, F. M.; Kozuch, S. The many faces of halogen bonding: a review of theoretical models and methods. *WIREs Comput. Mol. Sci.* **2014**, *4*, 523–540.

(118) Ibrahim, M. A. A. Molecular mechanical study of halogen bonding in drug discovery. *J. Comput. Chem.* **2011**, *32*, 2564–2574.

(119) E. Skyner, R.; L. McDonagh, J.; R. Groom, C.; Mourik, T. v.; O. Mitchell, J. B. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* **2015**, *17*, 6174–6191.

(120) Sure, R.; Grimme, S. Comprehensive Benchmark of Association (Free) Energies of Realistic Host–Guest Complexes. *J. Chem. Theory Comput.* **2015**, *11*, 3785–3801.

(121) 2.13. Implicit Solvation — ORCA 6.1 Manual.

(122) Young Lee, G.; Bay, K. L.; Houk, K. N. Evaluation of DFT Methods and Implicit Solvation Models for Anion-Binding Host-Guest Systems. *Helvetica Chimica Acta* **2019**, *102*, e1900032.

(123) Yin, J.; Henriksen, N. M.; Slochower, D. R.; Shirts, M. R.; Chiu, M. W.; Mobley, D. L.; Gilson, M. K. Overview of the SAMPL5 host–guest challenge: Are we doing better? *J. Comput. Aided Mol. Des.* **2017**, *31*, 1–19.

(124) Fortuna, A.; Costa, P. J. Assessment of Halogen Off-Center Point-Charge Models Using Explicit Solvent Simulations. *J. Chem. Inf. Model.* **2023**, *63*, 7464–7475.

(125) Stebani, J. A.; Iribarren Aguirre, Í.; Siddiqui, G. A.; Wragg, D.; Gagliardi, A.; Casini, A. ML-assisted Computational Workflow to the Robust Modeling of Supramolecular Metallacages in Solution. *ChemRxiv* **2025**.

(126) Mueller, T.; Hernandez, A.; Wang, C. Machine learning for interatomic potential models. *J. Chem. Phys.* **2020**, *152*, 050902.

(127) Xia, J.; Zhang, Y.; Jiang, B. The evolution of machine learning potentials for molecules, reactions and materials. *Chem. Soc. Rev.* **2025**, *54*, 4790–4821.

(128) Butler, P. W. V.; Hafizi, R.; Day, G. M. Machine-Learned Potentials by Active Learning from Organic Crystal Structure Prediction Landscapes. *J. Phys. Chem. A* **2024**, *128*, 945–957.

(129) Vitartas, V.; Zhang, H.; Juraskova, V.; Johnston-Wood, T.; Duarte, F. Active learning meets metadynamics: Automated workflow for reactive machine learning potentials. *ChemRxiv* **2025**.

(130) Zhang, H.; Juraskova, V.; Duarte, F. Modelling chemical processes in explicit solvents with machine learning potentials. *Nat. Commun.* **2024**, *15*, 6114.

(131) Wang, F.; Ma, Z.; Cheng, J. Accelerating Computation of Acidity Constants and Redox Potentials for Aqueous Organic Redox Flow Batteries by Machine Learning Potential-Based Molecular Dynamics. *J. Am. Chem. Soc.* **2024**, *146*, 14566–14575.

(132) Zhao, Q.; Han, Y.; Zhang, D.; Wang, J.; Zhong, P.; Cui, T.; Yin, B.; Cao, Y.; Jia, H.; Duan, C. Harnessing Machine Learning to Enhance Transition State Search with Interatomic Potentials and Generative Models. *Advanced Science* **2025**, e06240.

(133) F. dos Santos, L. G.; Nebgen, B. T.; Allen, A. E. A.; Hamilton, B. W.; Matin, S.; Smith, J. S.; Messerly, R. A. Improving Bond Dissociations of

Reactive Machine Learning Potentials through Physics-Constrained Data Augmentation. *J. Chem. Inf. Model.* **2025**, *65*, 1198–1210.

(134) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(135) Kang, K.; Purcell, T. A. R.; Carbogno, C.; Scheffler, M. Accelerating the training and improving the reliability of machine-learned interatomic potentials for strongly anharmonic materials through active learning. *Physical Review Materials* **2025**, *9*, 063801.

(136) Pacini, A.; Ferrario, M.; Righi, M. C. Accelerating Data Set Population for Training Machine Learning Potentials with Automated System Generation and Strategic Sampling. *J. Chem. Theory Comput.* **2025**, *21*, 7102–7110.

(137) Yang, Y.; Zhang, S.; Ranasinghe, K. D.; Isayev, O.; Roitberg, A. E. Machine Learning of Reactive Potentials. *Annual Review of Physical Chemistry* **2024**, *75*, 371–395.

(138) Batatia, I.; Kovács, D. P.; Simm, G. N. C.; Ortner, C.; Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, 2023.

(139) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Pu, Y.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF: Short-Range Transferable Machine Learning Force Fields for Organic Molecules. *J. Am. Chem. Soc.* **2025**, *147*, 17598–17611.

(140) Batatia, I.; Benner, P.; Chiang, Y.; Elena, A. M.; Kovács, D. P.; Riebesell, J.; Advincula, X. R.; Asta, M.; Avaylon, M.; Baldwin, W. J.; Berger, F.; Bernstein, N.; Bhowmik, A.; Bigi, F.; Blau, S. M.; Cărare, V.; Ceriotti, M.; Chong, S.; Darby, J. P.; De, S.; Pia, F. D.; Deringer, V. L.; Elijošius, R.; El-Machachi, Z.; Falcioni, F.; Fako, E.; Ferrari, A. C.; Gardner, J. L. A.;

Gawkowski, M. J.; Genreith-Schriever, A.; George, J.; Goodall, R. E. A.; Grandel, J.; Grey, C. P.; Grigorev, P.; Han, S.; Handley, W.; Heenen, H. H.; Hermansson, K.; Holm, C.; Ho, C. H.; Hofmann, S.; Jaafar, J.; Jakob, K. S.; Jung, H.; Kapil, V.; Kaplan, A. D.; Karimitari, N.; Kermode, J. R.; Kourtis, P.; Kroupa, N.; Kullgren, J.; Kuner, M. C.; Kuryla, D.; Liepuoniute, G.; Lin, C.; Margraf, J. T.; Magdău, I.-B.; Michaelides, A.; Moore, J. H.; Naik, A. A.; Niblett, S. P.; Norwood, S. W.; O'Neill, N.; Ortner, C.; Persson, K. A.; Reuter, K.; Rosen, A. S.; Rosset, L. A. M.; Schaaf, L. L.; Schran, C.; Shi, B. X.; Sivonxay, E.; Stenczel, T. K.; Svahn, V.; Sutton, C.; Swinburne, T. D.; Tilly, J.; Oord, C. v. d.; Vargas, S.; Varga-Umbrich, E.; Vegge, T.; Vondrák, M.; Wang, Y.; Witt, W. C.; Wolf, T.; Zills, F.; Csányi, G. A foundation model for atomistic materials chemistry. *ChemRxiv* **2025**.

(141) Röcken, S.; Zavadlav, J. Enhancing Machine Learning Potentials through Transfer Learning across Chemical Elements. *J. Chem. Inf. Model.* **2025**, *65*, 7406–7414.

(142) Zaverkin, V.; Holzmüller, D.; Bonfirraro, L.; Kästner, J. Transfer learning for chemically accurate interatomic neural network potentials. *Phys. Chem. Chem. Phys.* **2023**, *25*, 5383–5396.

(143) Judd, L. W.; Davis, A. P. From cholapod to cholaphane transmembrane anion carriers: accelerated transport through binding site enclosure. *Chem. Commun.* **2010**, *46*, 2227–2229.

(144) Jowett, L. A.; Howe, E. N. W.; Wu, X.; Busschaert, N.; Gale, P. A. New Insights into the Anion Transport Selectivity and Mechanism of Tren-based Tris-(thio)ureas. *Chem. Eur. J.* **2018**, *24*, 10475–10487.

(145) Zhou, X.; Xing, P. Beyond halogen bonding: -hole interactions. *Trends in Chemistry* **2025**, *7*, 460–473.

(146) Gamez, P. The anion– interaction: naissance and establishment of a peculiar supramolecular bond. *Inorganic Chemistry Frontiers* **2014**, *1*, 35–43.

(147) Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Rep.* **2017**, *7*, 42717.

(148) Neenan, T. X.; Whitesides, G. M. Synthesis of high carbon materials from acetylenic precursors. Preparation of aromatic monomers bearing multiple ethynyl groups. *The Journal of Organic Chemistry* **1988**, *53*, 2489–2496.

(149) Hein, J. E.; Tripp, J. C.; Krasnova, L. B.; Sharpless, K. B.; Fokin, V. V. Copper(I)-Catalyzed Cycloaddition of Organic Azides and 1-Iodoalkynes. *Angew. Chem. Int. Ed.* **2009**, *48*, 8018–8021.

(150) Bogdan, A. R.; Jerome, S. V.; Houk, K. N.; James, K. Strained Cyclophane Macrocycles: Impact of Progressive Ring Size Reduction on Synthesis and Structure. *J. Am. Chem. Soc.* **2012**, *134*, 2127–2138.

(151) BindFit v0.5 | Supramolecular.

(152) Bhosale, S.; Matile, S. A simple method to identify supramolecules in action: Hill coefficients for exergonic self-assembly. *Chirality* **2006**, *18*, 849–856.

(153) Varma, S.; Sabo, D.; Rempe, S. B. K+/Na+ Selectivity in K Channels and Valinomycin: Over-coordination *Versus* Cavity-size constraints. *J. Mol. Biol.* **2008**, *376*, 13–22.

(154) Hirata, K.; Sato, E.; Lisy, J. M.; Ishiuchi, S.-i.; Fujii, M. Ion Recognition beyond Size Matching: Cooperative Hydration Effect on the K+ Selectivity of Valinomycin over Na+ Revealed by Cryogenic Double Ion Trap Infrared Spectroscopy. *The Journal of Physical Chemistry Letters* **2023**, *14*, 5567–5572.

(155) Table of Dielectric Constants of Liquids - Alfa Chemistry.

(156) Wang, S.; Witek, J.; Landrum, G. A.; Riniker, S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. *J. Chem. Inf. Model.* **2020**, *60*, 2044–2058.

(157) Sethio, D.; Poongavanam, V.; Xiong, R.; Tyagi, M.; Duy Vo, D.; Lindh, R.; Kihlberg, J. Simulation Reveals the Chameleonic Behavior of Macrocycles. *J. Chem. Inf. Model.* **2023**, *63*, 138–146.

(158) Diaz, D. B.; Appavoo, S. D.; Bogdanchikova, A. F.; Lebedev, Y.; McTiernan, T. J.; dos Passos Gomes, G.; Yudin, A. K. Illuminating the dark conformational space of macrocycles using dominant rotors. *Nat. Chem.* **2021**, *13*, 218–225.

(159) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(160) Lin, S.; Elanany, M.; Khawaji, M. XTBDFT: Automated workflow for conformer searching of minima and transition states powered by extended tight binding and density functional theory. *SoftwareX* **2022**, *20*, 101242.

(161) Zurek, C.; Mallaev, R. A.; Paul, A. C.; van Staalduinen, N.; Pracht, P.; Ellerbrock, R.; Bannwarth, C. Tensor Train Optimization for Conformational Sampling of Organic Molecules. *J. Chem. Theory Comput.* **2025**, *21*, 1459–1475.

(162) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(163) Grimme, S.; Bohle, F.; Hansen, A.; Pracht, P.; Spicher, S.; Stahn, M. Efficient Quantum Chemical Calculation of Structure Ensembles and Free Energies for Nonrigid Molecules. *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

(164) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(165) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1606.

(166) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(167) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(168) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(169) Bickelhaupt, F. M.; Houk, K. N. Analyzing Reaction Rates with the Distortion/Interaction-Activation Strain Model. *Angew. Chem. Int. Ed.* **2017**, *56*, 10070–10086.

(170) Kolář, M.; Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 1325–1333.

(171) Wolters, L. P.; Schyman, P.; Pavan, M. J.; Jorgensen, W. L.; Bickelhaupt, F. M.; Kozuch, S. The many faces of halogen bonding: a review of theoretical models and methods. *WIREs Comput. Mol. Sci.* **2014**, *4*, 523–540.

(172) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009**, *180*, 1961–1972.

(173) Lindahl; Abraham; Hess; Spoel, v. d. GROMACS 2021.3 Source code, 2021.

(174) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187–199.

(175) Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713–3720.

(176) Marcus, Y. Ionic radii in aqueous solutions. *Chem. Rev.* **1988**, *88*, 1475–1498.

(177) Marcus, Y. Volumes of aqueous hydrogen and hydroxide ions at 0 to 200 °C. *J. Chem. Phys.* **2012**, *137*, 154501.

(178) Ahmad, M.; Johnson, T. G.; Flerin, M.; Duarte, F.; Langton, M. J. Responsive Anionophores with AND Logic Multi-Stimuli Activation. *Angew. Chem. Int. Ed.* **2024**, *63*, e202403314.

(179) Ahmad, M.; Flerin, M.; Tay, H. M.; Thompson, A. L.; Duarte, F.; Langton, M. J. Stimuli-responsive anion transport utilising caged hydrazone-based anionophores. *Nanoscale* **2024**, *16*, 21545–21553.

(180) Santacroce, P. V.; Davis, J. T.; Light, M. E.; Gale, P. A.; Iglesias-Sánchez, J. C.; Prados, P.; Quesada, R. Conformational Control of Transmembrane Cl- Transport. *J. Am. Chem. Soc.* **2007**, *129*, 1886–1887.

(181) Glendening, E. D.; Landis, C. R.; Weinhold, F. *NBO 7.0* : New vistas in localized and delocalized chemical bonding theory. *J. Comput. Chem.* **2019**, *40*, 2234–2241.

(182) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **1976**, *22*, 245–268.

(183) Young, T.; Johnston-Wood, T.; Zhang, H.; Duarte, F. Reaction dynamics of Diels-Alder reactions from machine learned potentials. *Phys. Chem. Chem. Phys.* **2022**.

(184) Robertson, C. C.; Perutz, R. N.; Brammer, L.; Hunter, C. A. A solvent-resistant halogen bond. *Chem. Sci.* **2014**, *5*, 4179–4183.

(185) Grimme, S. Improved second-order Møller-Plesset perturbation theory by separate scaling of parallel and antiparallel spin pair correlation energies. *J. Chem. Phys.* **2003**, *118*, 9095–9102.

(186) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

(187) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(188) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(189) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(190) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(191) Mardirossian, N.; Head-Gordon, M. $\omega$B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(192) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(193) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(194) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122.

(195) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272.

(196) Takano, Y.; Houk, K. N. Benchmarking the Conductor-like Polarizable Continuum Model (CPCM) for Aqueous Solvation Free Energies of Neutral and Ionic Organic Molecules. *J. Chem. Theory Comput.* **2005**, *1*, 70–77.

(197) Stahn, M.; Ehlert, S.; Grimme, S. Extended Conductor-like Polarizable Continuum Solvation Model (CPCM-X) for Semiempirical Methods. *J. Phys. Chem. A* **2023**, *127*, 7036–7043.

(198) Pliego, J. R. J. Hybrid Cluster-Continuum Method for Single-Ion Solvation Free Energy in Acetonitrile Solvent. *J. Phys. Chem. A* **2024**, *128*, 6440–6449.

(199) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson, A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(200) Triviño, J. Á. d. G. Revisiting random insertion for packing molecules in a box. *ChemRxiv* **2025**.

# A |

# Experimental and Computational Methods

## A.1 Chapter 2 Experimental Methods

### A.1.1 Materials and Methods

All reagents and solvents were purchased from commercial sources and used without further purification. Lipids were purchased from Avanti polar lipids and used without further purification. Where necessary, solvents were dried by passing through an MBraun MPSP-800 column and degassed with nitrogen. Triethylamine was distilled from and stored over potassium hydroxide. Silica gel flash column chromatography was performed either manually using Merck® silica gel 60 under a positive pressure of nitrogen or on a Buchi Pure C-815 Flash automated column chromatography system using FlashPure EcoFlex silica cartridges. Where mixtures of solvents were used, ratios are reported by volume. NMR spectra were recorded on a Bruker AVIIIHD 400 Nanobay and Bruker NEO 600 spectrometers. Chemical shifts are reported as $\delta$ values in ppm. Mass spectra were carried out on a Waters RDa bench-top TOF used with an Acquity LC system for reverse-phased chromatography. Fluorescence spectroscopic data were recorded using an Agilent Cary Eclipse fluorescence spectrophotometer, equipped with a Peltier temperature controller and stirrer. Experiments were conducted at 25 °C unless otherwise stated.

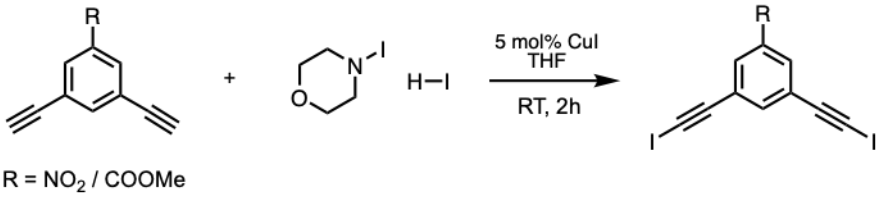
Vesicles were prepared as described below using an Avestin "LiposoFast" extruder apparatus, equipped with polycarbonate membranes with 200 nm pores. GPC

purification of vesicles was carried out using GE Healthcare PD-10 desalting columns prepacked with Sephadex G 25 medium.

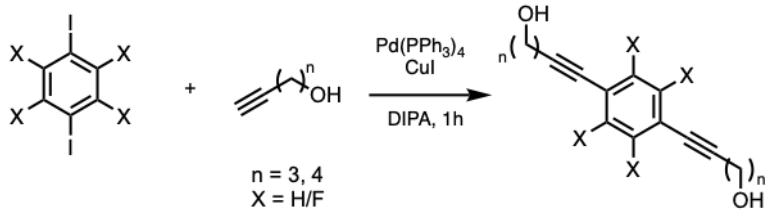
### A.1.2 Synthetic Methods

Warning! Low molecular weight organic azides used in this study are potentially explosive and should be used on a small scale. Appropriate protective measures should always be taken when handling these compounds.

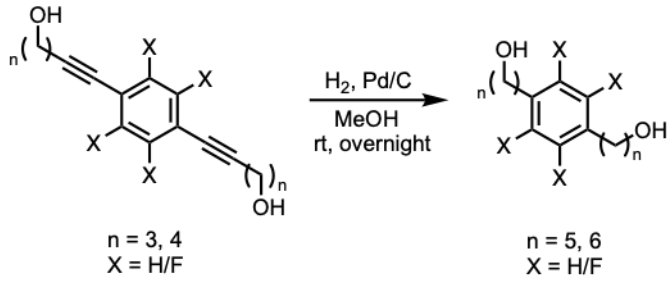
Tris[(1-benzyl-1H-1,2,3-triazol-4-yl)methyl]amine (TBTA)[1] and diethynylnitrobenzene were prepared according to literature procedures.[2]



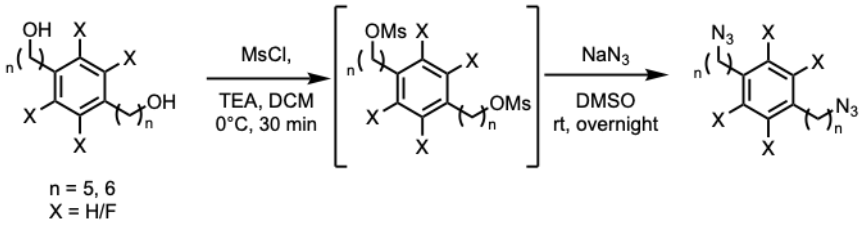

**General Procedure 1:** Di-iodoethynyl-arenes were prepared in accordance with literature precedent.[3, 4] The requisite di-ethynyl-arene (0.29 mmol, 1 eq.) was dissolved in dry THF (10 mL) at room temperature. CuI (55 mg, 0.29 mmol, 1 eq.) was added, and the mixture was left to stir for 5 minutes. Following this, *N*-iodomorpholine (325 mg, 0.957 mmol, 3.3 eq.) was added portionwise, and the reaction was left to stir at room temperature overnight. The reaction mixture was diluted with DCM (30 mL) and washed with saturated Na2S2O3 (cca. 20 mL), until the reddish colour had mostly disappeared from the organic phase. The aqueous layer was back-extracted with DCM (1 x 20 mL) and the combined organics were dried over $MgSO_4$ and concentrated to afford the respective di-iodoethynyl-arene.



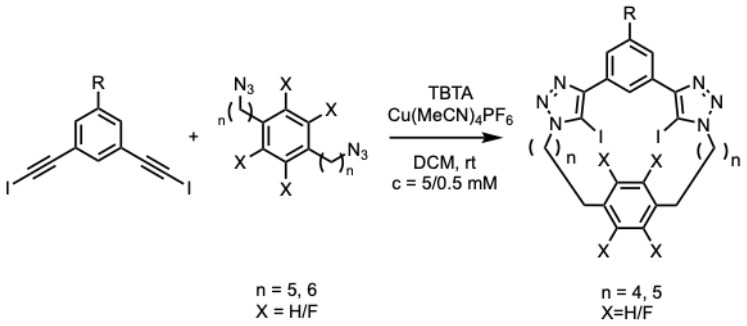

**General Procedure 2:** Bis-1,4-alkynol arenes were prepared from the corresponding p-substituted diiodoarene by a double Sonogashira reaction adapted from literature precedent.[5] The diiodoarene (1 mmol, 1 eq.) was dissolved in DIPA (10 mL). $Pd(PPh_3)_4$ (42 mg, 5 mol%) and CuI (11 mg, 5 mol%) were added and the solution was degassed. The alkynol (2.2 eq.) was then added and the reaction mixture was heated to 90 °C for approximately 45 minutes, after which significant decomposition was observed if the reaction was left at this temperature for a longer time. The reaction mixture was diluted with DCM (20 mL) and filtered through a pad of Celite, washed with more DCM (3 x 30 mL) and adsorbed onto silica. Flash column chromatography (3:2, EtOAc:Pentane) afforded the respective products.



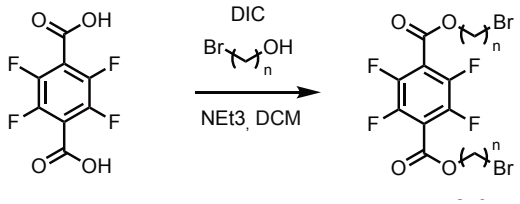

**General Procedure 3:** Bis-1,4-alkanol arenes were prepared from the corresponding p-substituted bis-1,4-alkynols prepared using General Procedure 2 by standard catalytic hydrogenation. The bis-1,4-alkynol (1 mmol, 1 eq.) was dissolved in MeOH (20 mL), and the 10% Pd/C (50 mg, 5 mol%) was added as a suspension in deionised water (0.5 mL). The vessel was filled with hydrogen with one balloon and a second balloon was used to establish a hydrogen atmosphere overnight at room temperature, following which the reaction mixture was filtered through a pad of silica, dried over MgSO4, and concentrated to afford the product.



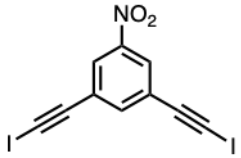

**General Procedure 4:** Bis-1,4-(azidoalkyl)arenes were prepared from the corresponding p-substituted bis- 1,4-alkanols by standard tosylation and S*N*2 displacement by an azide. The bis-1,4-alkanol (1 mmol, 1 eq.) and MsCl (0.47 mL, 2.4 mmol, 2.4 eq.) were dissolved in DCM (20 mL), at 0 °C, and $NEt_3$ (2 mL) was added dropwise. The reaction was allowed to warm to room temperature and left until monitoring by TLC indicated full completion. It was then diluted with DCM (50 mL) and washed with water (50 mL). The aqueous phase was washed with DCM (2 x 50 mL), and the combined organics were dried over $MgSO_4$ and concentrated to afford the crude intermediate, which was immediately redissolved in DMSO (5 mL), to which $NaN_3$ (155 mg, 2.2 mmol, 2.2 eq.) was added portion-wise and the mixture was left to react overnight. The reaction was quenched with brine (20 mL), and back-extracted with $Et_2O$ (5x 25 mL). The combined organic layers were once again washed with brine (100 mL), dried over $MgSO_4$ and concentrated to yield the bis-azide.



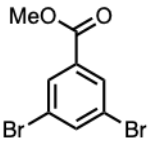

**General Procedure 5:** Macrocycles with alkyl linkers were prepared using standard iodoalkyne azide click reaction conditions at varying concentrations (5/0.5 mM).[3, 4, 6] $Cu(MeCN)_4PF_6$ (9 mg, 50 mol%) and TBTA (13 mg, 50 mol%) were dissolved in dry DCM (100 or 10 mL) and stirred for 30 minutes in order to precomplex. Di-iodoethynyl arene (0.05/0.005 mmol, 1 eq.) and bis-azide (0.05/0.005 mmol, 1 eq.) were added, and the reaction was left to stir in the dark for 48 hours or until complete conversion. The reaction mixture was washed with EDTA/$NH_4OH$ solution, dried over $MgSO_4$, adsorbed onto silica and purified with flash column chromatography in 2% MeOH in DCM to give a mixture of [1+1] and [2+2] macrocycles. This mixture was adsorbed onto silica once more and subjected to reverse-phase flash column chromatography in 60:40 to 100:0 MeCN:Water mixtures to give the desired [1+1] macrocycle as the first elution band, which was then further purified by recrystalisation from a hot 1:4 mixture of EtOAc:heptane to afford the pure [1+1] macrocycle.



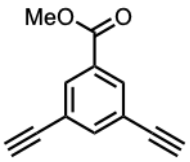

**General Procedure 6:** To dry DCM (5 mL) were added tetrafluoroterephthalic acid (238 mg, 1 mmol, 1 eq.) and the respective bromoalkanol (2 mmol, 2 eq.). Then DIC (0.332 mL) in 3 mL dry DCM and a catalytic amount of DMAP were added and the reaction was left to stir overnight. The reaction mixture was then diluted with DCM (10 mL) and filtered, washed with 5% sodium bicarbonate solution, brine, and dried over $MgSO_4$. It was then adsorbed onto silica in order to be purified by flash column chromatography using a 20% EtOAc in pentane mixture to yield the corresponding dibromide.



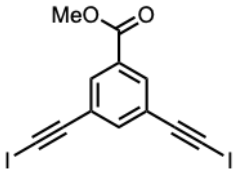

Known compound 1,3-bis(iodoethynyl)-5-nitrobenzene **7** was prepared according to General Procedure 1 and literature precedent as a pale yellow solid (108 mg, 0.26 mmol, 88%).



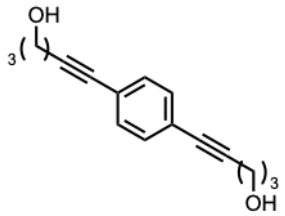

Literature compound methyl 3,5-dibromobenzoate[7] **8** was prepared from 3,5-dibromobenzoic acid (279 mg, 1 mmol, 1 eq.) suspended in dry DCM under a nitrogen atmosphere at 0 °C. A catalytic amount of DMF was added to the reaction mixture first, followed by oxalyl chloride (0.686 mL, 8 mmol, 8 eq.) being added dropwise. All volatiles were removed, and MeOH was immediately added to the dry acid chloride intermediate. After 30 minutes, the reaction mixture was concentrated to afford the yellow crystalline product (280 mg, 0.96 mmol, 96%)



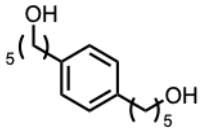

Known compound methyl 3,5-diethynylbenzoate **9** was synthesised according to an adapted literature procedure.[8] Compound **8** (0.292 g, 1 mmol, 1 eq.) was dissolved in a mixture of toluene and diisopropylamine (10 mL : 3 mL) and degassed. Pd(PPh3)Cl2 (0.35 g, 5 mol%), CuI (10 mg, 5 mol%) and TMS-acetylene (0.55 mL, 4 eq.) were added, and the reaction mixture was heated to 80 °C overnight. The crude mixture was filtered through celite, which was washed with DCM (3 x 30 mL) and the combined organic phase adsorbed onto silica. Column chromatography using 10% DCM in hexane afforded the TMS-protected product, which was immediately dissolved in methanol (5 mL). To this solution, KOH in 3 mL of a MeOH:water mixture (2:1) was added dropwise. After 30 minutes, the mixture was diluted with DCM (50 mL), which was washed with water to remove inorganic salts, dried, and concentrated to a brown solid (342 mg, 2 mmol, 70%).



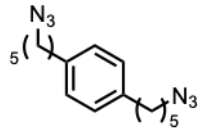

Methyl 3,5-bis(iodoethynyl)benzoate **10** was prepared according to **General Procedure 1** as an off-white solid (55%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 8.03 (d, J = 1.6 Hz, 2H), 7.64 (t, J = 1.6 Hz, $^{1}$H), 3.91 (s, 3H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 165.6, 139.8, 133.6, 130.8, 124.3, 92.3, 52.7, 9.5. $[C_{12}H_7I_2O_2]$+ calculated 436.8530 ; found 436.8525.



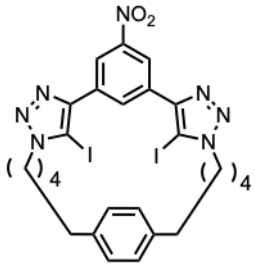

5,5'-(1,4-phenylene)bis(pent-4-yn-1-ol) **11** was prepared according to **General Procedure 2** as a white solid (80%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 7.29 (s, 4H), 3.81 (t, J = 6.1 Hz, 4H), 2.52 (t, J = 6.9 Hz, 4H), 1.89 – 1.82 (m, 4H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 131.5, 123.2, 91.1, 81.0, 61.9, 31.5, 16.2. $[C_{16}H_{19}O_2]$+ calculated 243.1380, found 243.1384



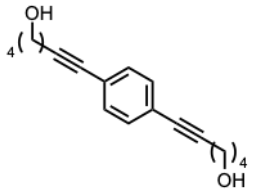

5,5'-(1,4-phenylene)bis(pentan-ol) **12** was prepared according to **General Procedure 3** as a white solid (80%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 7.08 (s, 4H), 3.63 (t, J = 6.6 Hz, 4H), 2.58 (t, J = 7.7 Hz, 4H), 1.71 – 1.54 (m, 8H), 1.43 – 1.36 (m, 4H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 140.0, 128.4, 63.1, 35.6, 32.8, 31.4, 25.5. $[C_{16}H_{27}O_2]$+ calculated 251.2006, found 251.2018



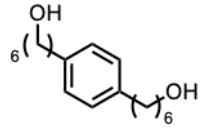
5,5'-(1,4-phenylene)bis(5-azidopentyl) **13** was prepared according to **General Procedure 4** as a thick orange oil (80%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 7.09 (s, 4H), 3.26 (t, J = 6.9 Hz, 4H), 2.59 (t, J = 7.5 Hz, 4H), 1.68 – 1.60 (m, 8H), 1.45 – 1.38 (m, 4H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 139.8, 128.5, 51.6, 35.5, 31.2, 28.9, 26.5. $[C_{16}H_{24}N_6K]$+ calculated 339.1694, found 339.1670.



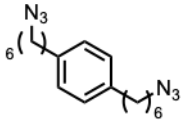
Macrocycle **2** was prepared according to **General Procedure 5** as an off-white solid (33%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 8.71 (d, J = 1.6 Hz, 2H), 8.41 (t, J = 1.6 Hz, $^{1}$H), 6.75 (s, 4H), 4.66 – 4.59 (m, 4H), 2.34 – 2.26 (m, 4H), 2.17 – 2.07 (m, 4H), 1.47 (p, J = 6.9 Hz, 4H), 1.00 (p, J = 7.9 Hz, 4H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 148.6, 139.4, 134.5, 132.8, 128.3, 122.5, 51.9, 35.1, 30.5, 28.3, 26.3. $[C_{26}H_{28}I_2N_7O_2]$+ calculated 724.0389; found 724.0388



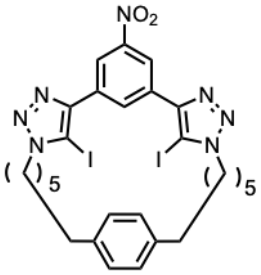
5,5'-(1,4-phenylene)bis(hex-5-yn-1-ol) **14** was prepared according to General Procedure 2 as a white solid (75%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 7.26 (s, 4H), 3.67 (q, J = 5.6 Hz, 4H), 2.42 (t, J = 6.6 Hz, 4H), 1.80 – 1.62 (m, 8H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 131.5, 123.2, 91.5, 80.9, 62.6, 32.0, 25.1, 19.4. $[C_{18}H_{23}O_2]$+ calculated 271.1693, found 271.1700



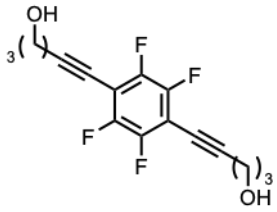
5,5'-(1,4-phenylene)bis(hexan-1-ol) **15** was prepared according to General Procedure 3 as a white solid (95%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 7.08 (s, 4H), 3.63 (t, J = 6.6 Hz, 4H), 2.61 – 2.53 (m, 4H), 1.66 – 1.51 (m, 8H), 1.41 – 1.33 (m, 8H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 139.9, 128.3, 63.0, 35.5, 32.7, 31.5, 25.6. $[C_{18}H_{31}O_2]$+ calculated 279.2319, found 279.2352



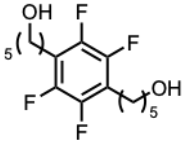
5,5'-(1,4-phenylene)bis(azidohexanyl) **16** was prepared according to General Procedure 4 as light yellow oil (65%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 7.06 (d, J = 1.8 Hz, 4H), 4.19 (td, J = 6.5, 2.5 Hz, 4H), 2.56 (td, J = 8.0, 1.9 Hz, 4H), 1.79 – 1.67 (m, 4H), 1.67 – 1.54 (m, 4H), 1.45 – 1.30 (m, 8H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 139.8, 128.3, 51.5, 35.4, 31.4, 28.8, 28.8, 26.6. $[C_{18}H_{29}N_6]$+ calculated 351.2268, found 351.2264



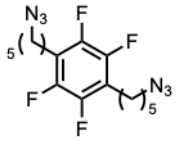
Macrocycle **3** was prepared according to **General Procedure 5** as an off-white solid (42%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 8.79 (d, J = 1.5 Hz, 2H), 8.68 (t, J = 1.5 Hz, $^{1}$H), 6.99 (s, 4H), 4.50 (t, J = 6.4 Hz, 4H), 2.51 – 2.43 (m, 4H), 1.95 (p, J = 6.9 Hz, 4H), 1.59 – 1.48 (m, 4H), 1.39 – 1.31 (m, 4H), 1.30 – 1.22 (m, 4H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 149.3, 140.2, 132.5, 131.9, 128.9, 122.5, 52.7, 50.6, 34.9, 31.3, 29.3, 27.5, 25.0. $[C_{28}H_{32}I_2N_7O_2]$+ calculated 752.0702; found 752.0699



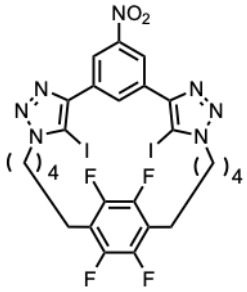
5,5'-(perfluoro-1,4-phenylene)bis(pentyn-1-ol) **17** was prepared according to **General Procedure 2** as an off- white solid (63%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 3.87 – 3.79 (m, 4H), 2.66 (t, J = 6.9 Hz, 4H), 1.90 (d, J = 12.9 Hz, 4H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 148.0 – 146.1 (m), 104.6, 104.3, 66.9 – 66.7 (m), 61.5, 31.0, 16.6. $[C_{16}H_{15}F_4O_2]$+ calculated 315.1003, found 315.1015



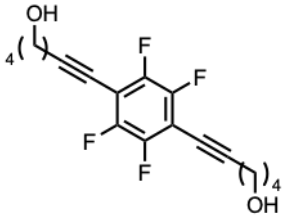
5,5'-(perfluoro-1,4-phenylene)bis(pentan-1-ol) **18** was prepared according to **General Procedure 3** as a white solid (95%). $^{1}$H NMR (600 MHz,$CDCl_3$) $\delta$ 3.64 (q, J = 6.1 Hz, 4H), 2.70 (t, J = 7.6 Hz, 4H), 1.66 – 1.55 (m, 8H), 1.41 (m, 6H, CH2 + OH). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 145.8 - 143.8 (m), 118.2 – 117.8 (m), 62.9, 32.5, 29.2, 25.5, 22.7. $[C_{16}H_{23}F_4O_2]$+ calculated 323.1629, found 323.1640



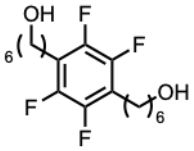
1,4-bis(5-azidopentyl)-2,3,5,6-tetrafluorobenzene **19** was prepared according to **General Procedure 4** as a lightly coloured oil (80%). $^{1}$H NMR (600 MHz,$CDCl_3$) $\delta$ 3.27 (t, J = 6.9 Hz, 4H), 2.71 (t, J = 7.6 Hz, 4H), 1.68 – 1.59 (m, 8H), 1.47 – 1.39 (m, 4H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 145.8 – 143.4 (m), 117.9 – 117.7 (m), 51.4, 28.9, 28.7, 26.4, 22.6. $[C_{16}H_{21}F_4N_6 - Na]$+ calculated 395.1578, found 395.1590.



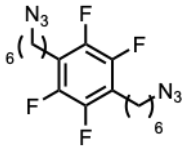
Macrocycle **4** was prepared according to **General Procedure 5** as an off-white solid (30%). $^{1}$H NMR (600 MHz,$CDCl_3$) $\delta$ 8.72 (d, J = 1.6 Hz, 2H), 8.33 (t, J = 1.6 Hz, 1H), 4.65 – 4.60 (m, 4H), 2.42 (t, J = 7.5 Hz, 4H), 2.10 (p, J = 6.6 Hz, 4H), 1.48 – 1.40 (m, 4H), 1.15 – 1.07 (m, 4H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 148.9, 148.8, 145.5 – 143.3 (m), 134.9, 132.8, 122.4, 117.5, 78.4, 51.5, 28.6, 28.4, 25.9, 22.1. $[C_{26}H_{24}F_4I_2N_7O_2]$+ calculated 796.0012, found 795.9975



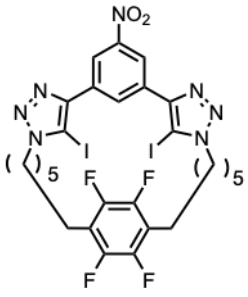
5,5'-(perfluoro-1,4-phenylene)bis(hex-5-yn-1-ol) **20** was prepared according to **General Procedure 2** as a very light pink-white solid (50%). $^{1}$H NMR (600 MHz,$CDCl_3$) $\delta$ 3.64 (q, J = 6.1 Hz, 4H), 2.70 (t, J = 7.6 Hz, 4H), 1.61 (dt, J = 13.6, 7.0 Hz, 8H), 1.48 – 1.36 (m, 8H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 147.8 – 145.9 (m), 104.6, 104.6 – 104.2 (m), 68.4 – 64.7 (m), 62.3, 31.7, 24.5, 19.7. $[C_{18}H_{19}F_4O_2]$+ calculated 343.1316, found 343.1337



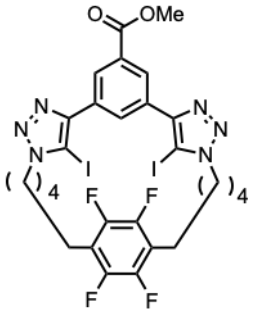
5,5'-(perfluoro-1,4-phenylene)bis(hexan-1-ol) **21** was prepared according to **General Procedure 3** as a white solid (90%). $^{1}$H NMR (600 MHz,$CDCl_3$) $\delta$ 3.64 (t, J = 6.6 Hz, 4H), 2.69 (t, J = 7.7 Hz, 4H), 1.63 – 1.54 (m, 8H), 1.45 – 1.33 (m, 8H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 145.6 – 143.9 (m), 118.3 – 117.9 (m), 63.1, 32.8, 29.4, 29.1, 25.5, 22.7. $[C_{18}H_{27}F_4O_2]$+ calculated 351.1942, found 351.1917



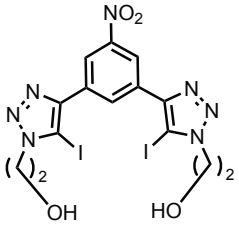
.1,4-bis(6-azido)-2,3,5,6-tetrafluorobenzene **22** was prepared according to **General Procedure 4** as light yellow oil (87%). $^{1}$H NMR (600MHz,CDCl3) $\delta$ 3.26 (t, J = 7.0 Hz, 4H), 2.70 (t, J = 7.7 Hz, 4H), 1.60 (p, J= 7.7 Hz, 8H), 1.45 – 1.33 (m, 8H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 146.0 – 144.0 (m), 118.7 – 117.2 (m), 51.5, 29.2, 28.9, 28.8, 26.5, 22.7. $[C_{18}H_{24}F_4N_6Na]$+ calculated 423.1891, found 423.1884



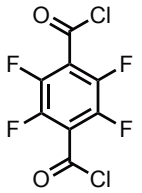
Macrocycle **5** was prepared according to **General Procedure 5** as an off-white solid (38%). $^{1}$H NMR (600 MHz,$CDCl_3$) $\delta$ 8.78 (d, J = 1.6 Hz, 2H), 8.63 (t, J = 1.6 Hz, $^{1}$H), 4.59 – 4.54 (m, 4H), 2.57 (t, J = 7.3 Hz, 4H), 1.98 – 1.91 (m, 4H), 1.54 – 1.49 (m, 3H), 1.32 – 1.27 (m, 8H). $^{13}$C NMR (151 MHz,$CDCl_3$) $\delta$ 149.2, 148.8, 145.7 – 143.7 (m), 133.3, 132.7, 122.8, 117.9, 77.6, 50.9, 29.7, 28.7, 27.5, 24.9, 21.6. $[C_{28}H_{28}F_4I_2N_7O_2]$+ calculated 824.0325, found 824.0324



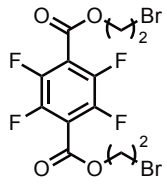
Macrocycle **6** was prepared according to **General Procedure 5** as an off-white solid (32%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 8.50 (d, J = 1.8 Hz, 2H), 8.09 (t, J = 1.7 Hz, 1H), 4.64 – 4.57 (m, 4H), 3.98 (s, 3H), 2.42 (t, J = 7.5 Hz, 4H), 2.08 (p, J = 6.6 Hz, 4H), 1.43 (p, J = 7.1 Hz, 4H), 1.15 (p, J = 7.6 Hz, 4H). $^{13}$C NMR (126 MHz,$CDCl_3$) 166.6, 133.6, 131.5, 131.1, 129.1, 78.1, 52.5, 51.4, 28.6, 28.5, 25.9, 22.1. $[C_{28}H_{27}F_4I_2N_6O_2]$+ calculated 809.0216; found 809.0220.



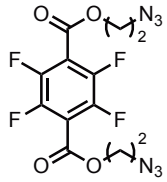
2,2'-((5-nitro-1,3-phenylene)bis(5-iodo-1H-1,2,3-triazole-4,1-diyl))bis(ethan-1-ol **23** was prepared using general procedure 5, with the exception of the excess of 3-azidopropan-1-ol (4 eq.) being used and without the requirement for reverse phase flash column chromatography. $^{1}$H NMR (400 MHz, MeOD) $\delta$ 8.99 (t, J = 1.6 Hz, 1H), 8.8 9 (d, J = 1.6 Hz, 2H), 4.67 (t, J = 5.6 Hz, 4H), 4.08 (t, J = 5.6 Hz, 4H). [M-H]+, 597.9191; found 597.9205



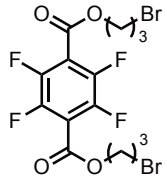
2,2'-(2,3,5,6-tetrafluoroterephthaloyl dichloride **24** was prepared from tetrafluoroterephthalic acid (238 mg, 1 mmol, 1 eq.) suspended in dry DCM under a nitrogen atmosphere at 0°C. A catalytic amount of DMF was added to the reaction mixture first, followed by $(COCl)_2$ (0.686 mL, 8 mmol, 8 eq.) being added dropwise. The compound was always used immediately without further purification. $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -136.13.



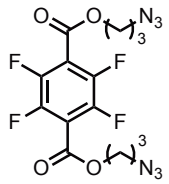
bis(2-bromoethyl) 2,3,5,6-tetrafluoroterephthalate **25** was prepared according to General Procedure 6 as a white solid (85%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 4.71 (t, J = 6.1 Hz, 1H), 3.62 (t, J = 6.1 Hz, 1H). $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -137.25. $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 158.6, 146.4 – 145.5 (m), 144.0 – 142.1 (m), 66.1, 27.5. [M-H]+, calculated 450.8798; found 450.8753



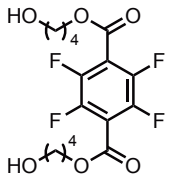
bis(2-azidoethyl) 2,3,5,6-tetrafluoroterephthalate **26** was prepared according to General Procedure 7 as a white solid (80%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 4.52 (t, J = 5.2 Hz, 4H), 3.65 (t, J = 5.2 Hz, 4H). $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -128.82. $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 162.7, 128.3, 123.5, 64.9, 49.4. [M-H]+ calculated 377.0616; found 377.0101



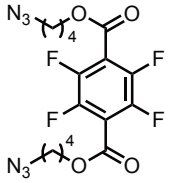
bis(3-bromopropyl) 2,3,5,6-tetrafluoroterephthalate **27** was prepared according to General Procedure 6 as an off-white solid (68%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 4.56 (t, J = 5.9 Hz, 4H), 3.52 (t, J = 6.4 Hz, 4H), 2.36 – 2.26 (m, 4H). $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -137.84. $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 158.8, 146.3 – 145.7 (m), 143.5 – 143.3 (m), 64.8, 31.4, 28.9. [M-H]+ calculated 480.9091, found 480.9027



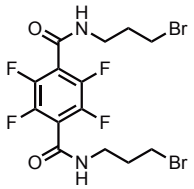
bis(3-azidopropyl) 2,3,5,6-tetrafluoroterephthalate **28** was prepared according to General Procedure 7 as a white solid (95%). $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 4.54 – 4.42 (m, 4H), 3.52 – 3.43 (m, 4H), 2.10 – 1.98 (m, 4H). $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -129.18. $^{13}$C NMR (101 MHz,$CDCl_3$) 161.1, 145.2 - 142.9 (m), 116.9, 48.3, 38.6, 28.4. $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 160.6, 154.3 - 153.0 (m), 117.7, 63.5, 47.9, 28.1. [M-H]+ calculated 405.0929, found 405.1548



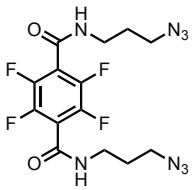
bis(4-hydroxybutyl) 2,3,5,6-tetrafluoroterephthalate **29** (274 mg, 1 mmol, 1 eq.) was dissolved in dry THF (50 mL) and to this was added a solution of butane-1,4-diol and $NEt_3$ in dry THF (25 mL) at 0°C and the mixture allowed to come to room temperature. All volatiles were evaporated and the crude mixture was adsorbed onto silica, and flash column chromatography in a 5% mixture of MeOH in DCM afforded the product as a white solid (71%, 270 mg, 0.71 mmol). $^{1}$H NMR (400 MHz, $CDCl_3$) $\delta$ 4.45 (t, J = 6.4 Hz, 4H), 3.71 (t, J = 6.3 Hz, 4H), 1.95 – 1.81 (m, 4H), 1.76 – 1.64 (m, 4H). $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -138.27. [M-H]+ calculated 383.1112, found 383.1121



bis(4-azidobutyl) 2,3,5,6-tetrafluoroterephthalate **30** was prepared according to general procedure 4 from the diol 28 to afford a yellow oil (95%) $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 4.42 (t, J = 6.3 Hz, 4H), 3.36 (t, J = 6.6 Hz, 4H), 1.94 – 1.81 (m, 4H), 1.80 – 1.69 (m, 4H). $^{19}$F NMR (377 MHz,$CDCl_3$) $\delta$ -129.27. $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 162.8, 127.9, 123.8, 65.9, 50.7, 25.5, 25.3. [M-H]+ calculated 433.1242, found 433.1860



bis(3-bromopropyl) 2,3,5,6-tetrafluoroterephthalamide **31** (548 mg, 2 mmol, 1 eq.) was dissolved in dry DCM (10 mL) under a nitrogen atmosphere and added to a suspension of 3-bromopropan-1-amine (872 mg, 4 mmol, 2 eq.) and $NEt_3$ (1 mL) in dry DCM (10 mL) at 0°C and allowed to come to room temperature. All volatiles were evaporated and the crude mixture was adsorbed onto silica, and flash column chromatography in a 2-3% mixture of MeOH in DCM afforded the product as a white solid (76%, 726 mg, 1.52 mmol). $^{1}$H NMR (400 MHz, MeOD) $\delta$ 3.63 – 3.47 (m, 8H), 2.22 – 2.11 (m, 4H). $^{19}$F NMR (377 MHz, MeOD) $\delta$ -143.55. $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 168.0, 144.0, 113.5, 55.8, 43.3, 29.7. [M-H]+ calculated 477.2147, found 476.9431



bis(3-azidopropyl) 2,3,5,6-tetrafluoroterephthalamide **32** was prepared according to general procedure 7 to afford a white solid (40%) $^{1}$H NMR (400 MHz,$CDCl_3$) $\delta$ 3.59 (q, J = 6.4 Hz, 4H), 3.50 – 3.45 (m, 4H), 1.92 (p, J = 6.5 Hz, 4H). $^{13}$C NMR (101 MHz,$CDCl_3$) $\delta$ 165.8, 148.2, 105.9, 49.1, 29.7, 28.4. [M-H]+ calculated 403.1249, found 403.1261

## A.1.3 Anion Transport Experiments

### A.1.3.1 Vesicle Preparation

A thin film of lipid (1-Palmitoyl-2-oleoyl-sn-glycero-3-phosphatidylcholine POPC, egg-yolk phosphatidylglycerol EYPG or dipalmitoyl phosphatidylcholine DPPC) was formed by evaporating a chloroform solution under reduced pressure on a rotary evaporator (40 °C) and then under high vacuum for 6 hours. The lipid film was hydrated by vortexing with the prepared buffer (100 mM NaCl, 10 mM HEPES, 1 mM 8-Hydroxypyrene-1,3,6-trisulfonic acid trisodium salt (HPTS), pH 7.0). The lipid suspension was then subjected to 5 freeze-thaw cycles using liquid nitrogen and a water bath (40°C) followed by extrusion 19 times through a polycarbonate membrane (pore size 200 nm). Extravesiclular components were removed by size exclusion chromatography on a Sephadex G-25 column with 100 mM NaCl, 10 mM HEPES, pH 7.0. Final conditions: LUVs (2.5 mM lipid); inside 100 mM NaCl, 10 mM HEPES, 1 mM HPTS, pH 7.0; outside: 100 mM NaCl, 10 mM HEPES, pH 7.0. Vesicles for the sodium gluconate assay were prepared by the same procedure, substituting NaCl for NaGluconate in the buffer solution.

### A.1.3.2 Transport Assays with HPTS

In a typical experiment, the LUVs containing HPTS (37.5 $\mu$L, final lipid concentration 31.3 $\mu$M) were added to buffer (2925 $\mu$L of 100 mM NaCl, 10 mM HEPES, pH 7.0) at 25°C under gentle stirring. A pulse of NaOH (30 $\mu$L, 0.5 M) was added at 40 secs to initiate the experiment. At 90 s the test transporter (various concentrations, in 7.5 $\mu$L DMSO) was added, followed by detergent (37.5 $\mu$L of Triton X-100 in 7:1 (v/v) H2O-DMSO) at 300 secs to calibrate the assay. The fluorescence emission was monitored at $\lambda$em = 510 nm ($\lambda$ex = 460/405 nm). The fractional fluorescence intensity (Irel) was calculated from equation:

$$I_{rel} = \frac{R_t - R_0}{R_d - R_0}$$

Rt is the fluorescence ratio at time t, R0 is the fluorescence ratio at time 0, and Rd is the fluorescence ratio after the addition of detergent.

The fractional fluorescence intensity (Irel) at 288 s just prior to lysis, defined as the fractional activity y, was plotted as a function of the ionophore concentration (M). Hill coefficients (n) and EC50 values were calculated by fitting to the Hill equation : $y = y_0 + (y_{max} - y_0)\frac{x^n}{EC_{50}{}^n - x^n}$ where $y_0$ is the baseline fractional activity in the absence of transporter, $y_{max}$ is the fractional activity in with excess transporter, x is the transporter concentration in the cuvette. Where full Hill plots were fitted, at least 7 data points spanning the required concentration range were considered, and each individual concentration was repeated at least twice and averaged. Experiments with DPPC lipids were conducted in the same way. For elevated temperature studies, the buffer was equilibrated at 45 °C (using the Peltier temperature controller) for 5 minutes prior to initiating the experiment. Experiments in the presence of protonophore trifluoromethoxy carbonylcyanide phenylhydrazone (FCCP) were carried out using the above procedure, except that a DMSO solution of FCCP (7.5 $\mu$L of 100$\mu$M solution, final concentration 0.25 $\mu$M/0.8 mol%) was added to the vesicle suspension following the addition of the NaOH pulse (at 60s).

## A.2 NMR Titration Experiments

All binding constants were measured by $^1$H NMR titrations in a Bruker AVIII 500 spectrometer at 500 MHz and 298 K. The host (anionophores **2.2, 2.3, 2.4, 2.5** and **2.6**) was dissolved in 2.5% $D_2O$-acetone-$d_6$ (v/v) mixtures, at 1 mM concentration

and a known volume (0.5 mL) added to the NMR tube. Known volumes of anion guest (added as the TBA salt, 50 mM) in 2.5% $D_2O$-acetone-$d_6$ (v/v) mixtures were added and the spectra were recorded after each addition. The chemical shift perturbations of the host spectra were monitored as a function of guest concentration. In all cases the downfield perturbation of the internal aryl signal of the bis-iodotriazole motif was monitored (Section B.1.5). The data was analysed using a global fit procedure using the Bindfit program, using non-linear least squares analysis to obtain the best fit between observed and calculated chemical shifts for the 1:1 binding stoichiometry.[9] In all experiments the association of guest and host was fast on the NMR timescale.

## A.3 Chapter 3 Computational Methods

Using the crystal structure of **3.1** as a starting point (CCDC deposition number 2075766)[4], initial structures of **3.2**-**3.6** were generated in Avogadro (version 1.20), and optimised using the built-in General Amber Force-Field, followed by optimisation using XTB (version 6.4.1).[10–12] These structures were used as the starting points for the conformer searching algorithm (Figure A.7). In the first step, a conformer search was carried out using the CREST program (version 2.11) in NCI mode using the default metadynamics parameters.[13] From the conformers generated, forty conformers of lowest energy were selected for single point energy calculations in ORCA (version 5.0.3) at the $\omega$B97X-D3/def2-TZVP level of theory to improve the reliability of their energy ranking.[14–17] The lowest energy conformers following this re-ranking were used in all subsequent DFT calculations. The coordinates of the final optimised structures can be found in the public GitHub repository https://github.com/martinzola/chloride_selective_non-protonophoric_ion_transport.

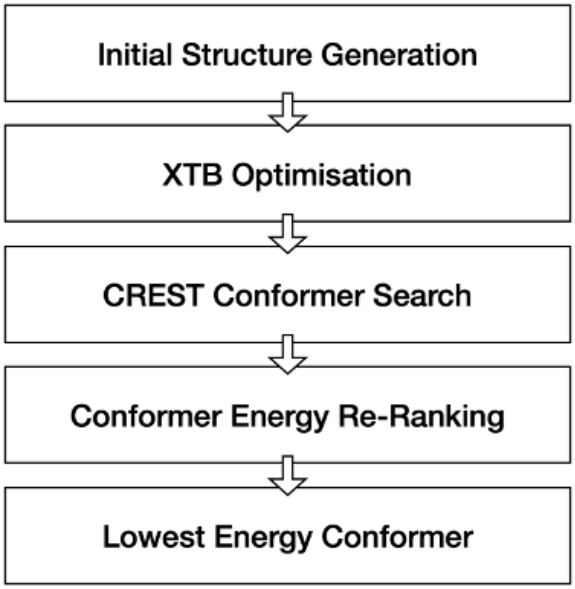



**Figure A.7:** Conformer searching and sorting workflow: Generation of transporter structure from crystal structure of **3.1** in Avogadro; Optimisation using XTB; Conformer search using metadynamics in CREST; Re-ranking of 40 lowest-energy conformers at the $\omega$B97X-D3/def2-TZVP in ORCA giving the lowest energy conformer.

Optimisations and energy calculations were carried out with the ORCA suite of programs (version 5.0.3), with the $\omega$B97X-D3 functional chosen to match that used in our previous studies on related systems.[4, 6] The choice was previously motivated by benchmarking studies showing that the $\omega$B97X family of functionals outperforms most other functionals in reproducing coupled cluster and MP2-level energies of halogen bonding interactions.[18] Lowest energy conformers of the transporters were optimised at the SMD(CHCl3)-$\omega$B97X-D3/def2-SVP level of theory (ma-def2-SVP on underlined atoms: I, $Cl^-$ and $OH^-$), referred here to as Low-Level (LL), followed by a frequency calculation at the same level.[16, 17, 19] Very tight convergence criteria were employed (2 × 10-7 Ha for the optimisation step and 10-8 Ha for the SCF energy change). The resolution of identity approximation and chain of sphere integration (RIJCOSX keyword) were used to speed up calculations, using the default auxiliary basis sets.[20] Thermochemistry was

calculated using the otherm program, using the default quasi-RRHO approximation ($\omega$0 = 45 cm-1, $\alpha$ = 4, T = 298.15 K).[21] By default, the otherm programme corrects the standard state from 1 atm to 1 M, which adds 1.89 kcal $mol^{-1}$ (equivalent to RT·ln(1 mol $dm^{-3}$/ (1/24.5 mol $dm^{-3}$)) e.g. [GHL,1M = GHL,1atm + RTln(24.5)]. Single point energies at a higher level (HL) were calculated at the CPCM(CHCl3)-$\omega$B97X-D3/def2-QZVP (ma-def2-QZVP on underlined atoms: I, $Cl^-$ and $OH^-$), level of theory. To approximate the free energies at this level, the thermodynamic contributions from the LL level were added to the single point energies at the HL level, i.e. GHL = EHL + (GLL - ELL). The binding interaction is then calculated as the energy difference of the complex [Transporter·Cl]$^-$ or [Transporter·OH]$^-$ and the energy of the separated Transporter and $Cl^-$/$OH^-$.

Additionally, in order to ascertain the balance between stabilising and destabilising interactions in all of the studied complexes, distortion interaction analysis was carried out by carrying out frequency calculations at the LL level and single point calculations at the HL level of theory on the geometries of each transporter with the host ion removed.[22] The difference between the energies of the complex in its bound pose and its minimal energy pose is the distortion energy, and quantifies how much energy must be input into the reorganisation of the molecule in order for an interaction to take place. The difference between this distortion energy and the overall binding energy is then the interaction energy, which shows how strongly the molecule is able to interact with its guest in its bound conformation. We also carried out second order perturbation analysis using NBO 7.0 in ORCA for all [transporter ·ion]$^-$ complexes at the HL level of theory to give interaction energies between the ions and the halogen bonding iodines We report the sum of all interaction energies E(2) between lone pairs on the ions and the anti-bonding orbitals of the bond between the halogen bonding iodine and the C4 carbon of the iodotriazole to which it is attached.[23]

### A.3.1 Molecular Dynamics

#### A.3.1.1 General System Preparation

To obtain the structures for fitting of the electrostatic parameters for molecular dynamics simulations, the five lowest energy conformers of transporters **3.1** and **3.2** found through the conformer searching workflow (Figure A.7) were optimised at the LL level of theory (cf. DFT Calculations). The electrostatic potential calculations were carried out in Gaussian 09 at the HF/6-31G* level of theory, using the Merz-Singh-Kollman scheme for charge calculation with 6 density points in each layer (IOp(6/33=2, 6/42=6, 6/50=1)).[24, 25] RESP fitting was then employed using the antechamber suite with equal weighting of all five conformers, alongside the assignment of GAFF atom types.[11, 26, 27] The sigma hole was parameterised as a positively charged particle with no mass, angled 180° away from the carbon-iodine bond, according to literature precedent.[28, 29] The distance of the extra particle from the iodine atom's center used in the RESP fitting and the Lennard-Jones distance parameter $\sigma$ were then altered until DFT binding distances could be approximated, without compromising the stability of simulations. Molecular mechanics interaction distances between **3.2** and $Cl^-$, obtained by the optimising 10,000 steps in the steepest descent algorithm, were compared to those in the LL level of theory for the default and optimised sigma hole and Lennard-Jones parameters (Table A.1). DMSO parameters were taken from Caleman et al, chloride parameters from Sengupta et al, hydroxide parameters from Han et al, and default AMBER parameters were used for sodium.[30–33] For POPC, the parameters used were the Stockholm lipids (Slipids-2020).[34, 35] Water was modelled using the tip3p-fb model.[36] GAFF parameters for the simulated compounds, as well as the parameter files and starting GROMACS geometries can be found in the public GitHub repository https://github.com/martinzola/chloride_selective_non-

protonophoric_ion_transport.

**Table A.1:** Standard vs. optimised EP parameters: GROMACS iodine $\sigma$, corresponding AMBER $R_{min}$ (converted via $R_{min} = 10\,\sigma\,2^{1/6}$), EP distance from iodine, and the resulting MM I$\cdots$Cl$^-$ distance in [**2**·Cl]$^-$. The DFT reference distance is shown for comparison.

| | $\sigma_I$ (nm) | $R_{min}$ (Å) | $r_{EP}$ (Å) | MM $d_{I\cdots Cl^-}$ (Å) | DFT $d_{I\cdots Cl^-}$ (Å) |
|---|---|---|---|---|---|
| Initial parameters | 0.383 | 4.30 | 2.15 | 3.87 | 3.23 |
| Initial parameters | 0.320 | 3.60 | 1.60 | 3.36 | 3.23 |

Throughout all simulations, long-range electrostatics were described with the Particle Mesh Ewald (PME) algorithm.[37] The temperature of the system was maintained at 298 K using the V-rescale thermostat.[38] Pressure was controlled by the C-rescale barostat at 1.0 bar, with an isothermal compressibility of $4.5 \cdot 10^{-5}$ bar$^{-1}$.[39] In the case of the membrane simulations, the pressure was controlled in a semi-isotropic manner, where $P_{xy} = P_x = P_y$ is the pressure in the plane of the membrane and $P_z$ is the pressure in the direction of the membrane normal. All bond lengths involving hydrogen atoms were constrained using the LINCS algorithm.[40]

Simulations were carried out in the GROMACS simulation software package (v.2021.3 in all subsequent mentions).[41] Plumed (v.2.7.2) was used to apply biasing forces in GROMACS for metadynamics simulations.[42]

#### A.3.1.2 Metadynamics

The structure of **3.2** was inserted into a cubic box with a 1 nm buffer to the edge of the box and solvated with DMSO. The systems were minimised using a steepest descent algorithm until the maximum force in the system was below 1000 kJ mol-1. This was followed by 500 ps of equilibration in the NPT ensemble (2 fs timestep, 298 K, 1 bar) starting from random velocities sampled from the

Maxwell-Boltzmann distribution.

The equilibrated system was then simulated for 1.2 $\mu$s, until convergence was found in the free energy profile, employing the well-tempered metadynamics (WTMetaD) method using Plumed.[42] The pace of Gaussian hill was set to 500 steps and the initial height of the hills was set to 0.3 kJ mol 1. The sigma values of the Gaussian hills were set to 0.3 radians for both $\alpha$ and $\beta$ angles, respectively, and the bias factor was set to 6. To speed up the calculations, the biases were stored on a grid ranging from -$\pi$ to  for each angle, with the default grid spacing.

Rate constants were estimated using the Eyring equation, where dG was the activation barrier from metadynamics simulations and the transition coefficient was taken to be equal to 1, kb is Boltzmann's constant, T is 298 K and h is Planck's constant and R is the gas constant. The half-life t1/2 was then calculated.

$$k = \kappa \frac{k_b T}{h} e^{\frac{-dG}{RT}}$$

$$t_{1/2} = \frac{\ln(2)}{k}$$

To verify the free energy surface obtained using metadynamics simulations the relative energies of the four conformational minima were benchmarked against their DFT energies. The starting points for the DFT optimisations were the four lowest energy conformers from the crest conformational search for 2, which had the representative dihedral angles found in each minimum well (Figure 3.9). They were optimised using the LL level of theory followed by single point calculations at the HL level of theory. Relative energies of these conformers found by DFT were then compared to the relative energies of the four conformer minimum wells found by Metadynamics.

#### A.3.1.3 Umbrella Sampling

The first 10 ns of each umbrella sampling window was discarded, and convergence was monitored in increments of 10 ns from this 10 ns starting point using the GROMACS wham utility, which carries out weighted histogram analysis.[43] All of the three repeats of each window were then analysed together from 50 ns to 100 using bootstrapping analysis with 100 bootstraps to give an estimate of the uncertainty.

## A.4 Chapter 4 Computational Methods

### A.4.1 General DFT Workflow

Starting geometries for the transporters were generated and visualised in Avogadro (v.1.20) and minimised using the General Amber Force Field (GAFF).[10, 11] These were further optimised using XTB (v.6.4.1).[12] A conformer search was performed using CREST (v.2.11) in NCI mode with default parameters.[13] Forty conformers were selected for single-point energy calculations at the $\omega$B97X-D3/def2-TZVP level in ORCA (v.5.0.3), to improve energy ranking.[14–17, 19] CREST geometries were used directly, as CREST has been benchmarked for reliable DFT geometries but less so for energies.

The lowest-energy conformer was optimised at the $\omega$B97X-D3/def2-SVP level of theory (LL). Very tight convergence criteria ($2\times10^{-7}$ Ha for optimisation and $10^{-8}$ Ha for SCF) were employed, using the Defgrid3 integration grid. RIJCOSX with auxiliary basis sets was applied.[20] Thermochemistry was obtained using the **otherm** program with the quasi-RRHO approximation ($\omega_0$ = 45 cm$^{-1}$, $\alpha$ = 4, T = 298.15 K), corrected to the 1 M standard state (+1.89 kcal mol$^{-1}$).[21]

Single point energies were calculated at the $\omega$B97X-D3/def2-QZVP level of theory (HL). Free energies were obtained by adding LL thermochemical corrections to HL single points:

$$G_{\mathrm{HL}} = E_{\mathrm{HL}} + (G_{\mathrm{LL}} - E_{\mathrm{LL}})$$

Binding energies were calculated as the difference between the complex ([Transporter·guest] or [2Transporter·guest]) and its separated components .

Second-order perturbation (NBO 7.0) analyses of unbound hydrazone transporters and their complexes were carried out to quantify donor–acceptor stabilisation energies $E^{(2)}$.[23] Distortion–interaction analyses partitioned binding into distortion penalties and host–guest interaction terms.[22]

### A.4.2 Molecular Dynamics

Systems were solvated in cubic DMSO boxes (1 nm buffer), neutralised, and minimised. Equilibration (500 ps, NPT, 298 K, 1 bar) preceded 100 ns production runs. Umbrella sampling was performed on the chloride–receptor COM distance (3–13 Å, 1 Å spacing, 12 windows, 100 ns/window, triplicates). Weighted histogram analysis (WHAM) was used with bootstrapping (100 resamples) to estimate uncertainties.[43]

The first 10 ns of each umbrella sampling window was discarded, and convergence was monitored in increments of 10 ns from this 10 ns starting point using the GROMACS wham utility, which carries out weighted histogram analysis.[43] All of the three repeats of each window were then analysed together from 50 ns to 100 using bootstrapping analysis with 100 bootstraps to give an estimate of the uncertainty. Finally, the converged energy profile was corrected to account for

the entropy loss of distance restrictions, equal to 2*Kb*T*Ln(r), where Kb is the Boltzmann constant, T is equal to 298 K and r is the distance from the binding site.

Free energy perturbation (FEP) was performed by gradually switching on Van der Waals and Coulombic interactions of $Cl^-$ in bulk DMSO and in the $[4.16{\cdot}Cl]^-$ complex.[44] $\lambda$-windows (10 ns each) were analysed with the Bennett Acceptance Ratio (BAR).

## A.5 Chapter 5 Computational Methods

DFT calculations were performed using ORCA (v 5.0.3), running optimisations with the "very tight" criteria for convergence ($2 \cdot 10^{-7}$ Ha for the optimisation step and $10^{-8}$ Ha for the SCF energy change).[14] Thermochemistry was calculated using the otherm program, using the default quasi-RRHO approximation ($\omega_0$ = 45 $cm^{-1}$, $\alpha$ = 4, T = 298.15 K).[21] By default, the otherm programme corrects the standard state from 1 atm to 1 M, which adds 1.89 kcal $mol^{-1}$ (equivalent to RT·ln(1 mol $dm^{-3}$/ (1/24.5 mol $dm^{-3}$)) i.e. [$GHL_{1M}$ = $GHL_{1\ atm}$ + RTln(24.5)]). The following DFT functionals were compared against the SCS-MP2 and MP2 methods[45] in the benchmarking study: B2PLYP[46], CAM-B3LYP[47], M06-2X[48], M06-2L[49], PBE0[50] (with and without D3BJ[51, 52] and D4[53]), $\omega$B97X[16], $\omega$B97X-D3BJ[54], $\omega$B97M[55], $\omega$B97M-V[55]. The resolution-of-identity chain-of-spheres exchange (RIJCOSX) approximation was used in all calculations.[20] The def2-TZVP basis set was used in all calculations, and Coulomb integrals were approximated using the def2/J auxiliary basis set.[17] Among the tested DFT functionals, the M06-2X functional displayed the highest level of fidelity to the reference method. The CPCM solvation model was used throughout when implicit solvation was employed.[56] All calculations carried out within the active learning loops were carried out using the same methods, using ORCA

wrapped in *mlp-train* and *autodE*.[57, 58]

**Table A.2:** Hyperparameter setting for MACE potential, SOAP descriptor, and selectors.

| Type | Parameter | Description | Value |
|---|---|---|---|
| MACE | $\nu$ | Maximum body order of atomic relationships | 3 |
| | $k$ | Number of chemical feature channels within each layer | 96 |
| | $r_{\text{cut}}$ | Cutoff of atomic environment centered at each atom | 4.5 Å |
| | $L_{\text{max}}$ | Symmetry order of the features | 0 |
| SOAP descriptor | $\sigma_{\text{at}}^{\text{SOAP}}$ | Gaussian width added to the atomic density | 1.0 Å |
| | $\eta_{\text{max}}$, $l_{\text{max}}$ | Parameters for the radial and angular basis of the SOAP descriptor | 6 |
| | $r_{\text{cut}}$ | Cutoff of the atomic environment radius | 5.0 Å |
| similarity selector | $k_{\text{T}}$ | Selection threshold | 0.999 |

The MACE machine learning interatomic potential (MLIP) was used within the mlp-train Python package.[59] The hyperparameters for the MACE models, the SOAP descriptors used in the similarity selector method within the active learning, and the similarity selector threshold are presented in Table A.2.[59–62] Dynamics simulations were propagated using the atomic simulation environment (ASE) as implemented within mlp-train, with a 0.5 fs timestep.[63] All simulations employed the NVT ensemble, utilising the Langevin thermostat (with a 0.02 friction coefficient in ASE units).[38] Initial velocities were sampled from a Maxwell-Boltzmann distribution at 298 K.

The molecular mechanics benchmark umbrella sampling study was run using GROMACS (v. 2021.3), using the same overall procedure as that described in the

methods for Chapter 4.[41]

# B |

# Supplementary Results

## B.1 Chapter 2 Assay Results

### B.1.1 HPTS Assay Data

In the following figures: Left: change in relative fluorescence intensity over time in the HPTS assay (LUVs (31.25 µM lipid); inside 100 mM NaCl, 10 mM HEPES, 1 mM HPTS, pH 7.0; outside: 100 mM NaCl, 10 mM HEPES, pH 7.0). Shaded area indicates two standard deviations. Right: dependence of the fractional transport activity y in the HPTS assay on the concentration of transporter (blue circles squares), and fitted to the Hill equation (blue line).

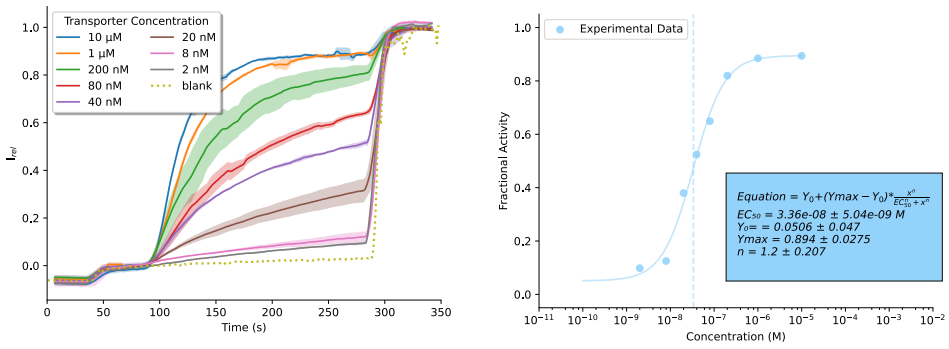



**Figure B.1:** HPTS assay data for carrier **2.2** with FCCP.

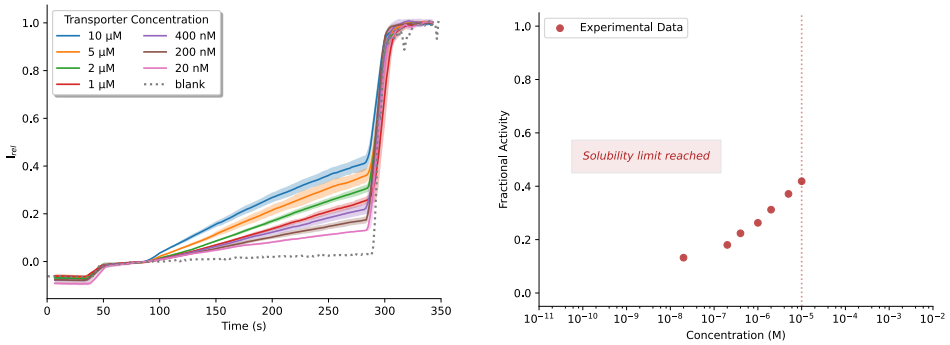



**Figure B.2:** HPTS assay data for carrier **2.2** without FCCP.

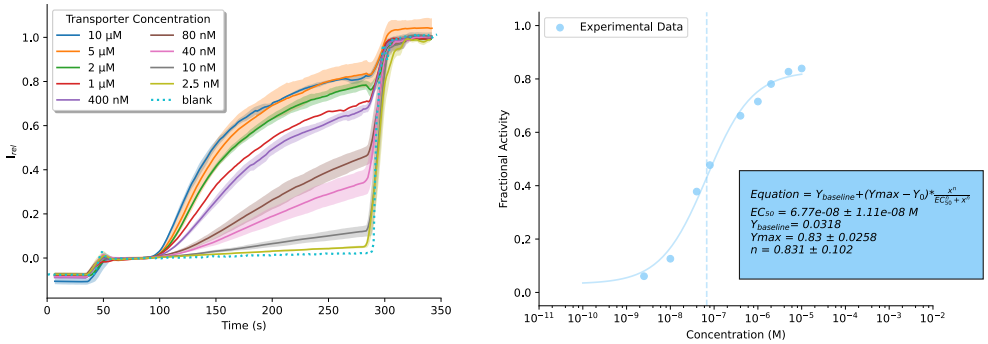



**Figure B.3:** HPTS assay data for carrier **2.3** with FCCP.

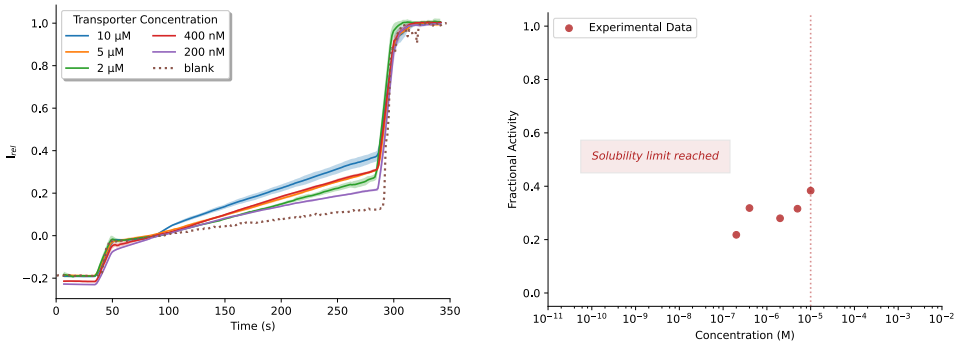



**Figure B.4:** HPTS assay data for carrier **2.3** without FCCP.

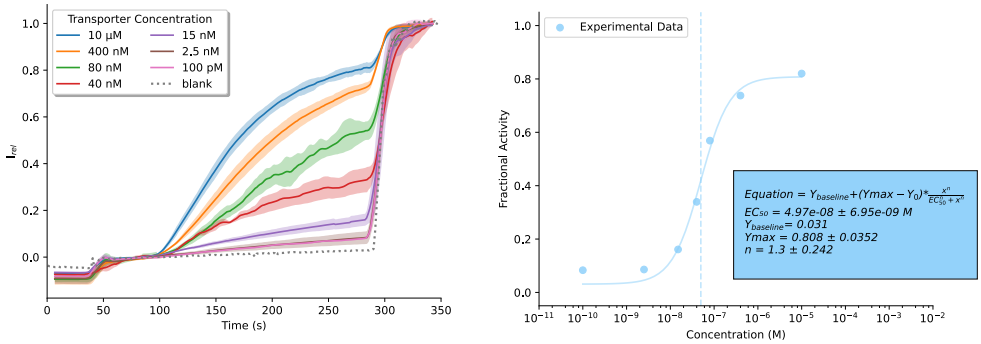



**Figure B.5:** HPTS assay data for carrier **2.4** with FCCP.

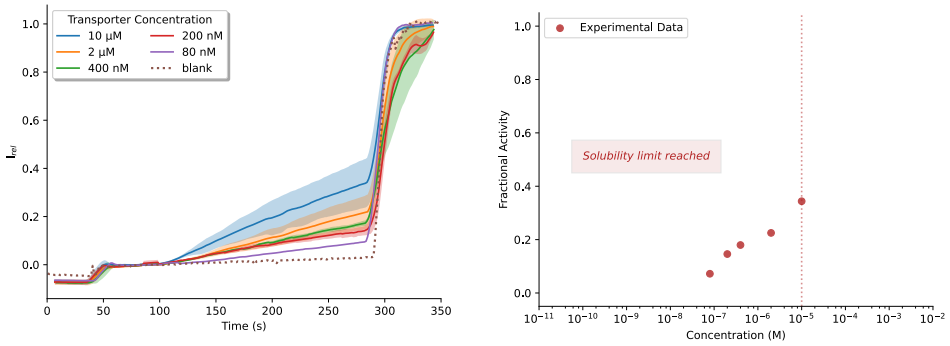



**Figure B.6:** HPTS assay data for carrier **2.4** without FCCP.

## B.1.2 Data for Transporter 2 with BSA

In experiments with BSA, fatty acid-free BSA was dissolved in the vesicle stock suspension to a final BSA concentration of 1 mol% (with respect to lipid). The BSA-containing vesicle stock suspension was stirred for 20 min before being used for membrane transport studies.

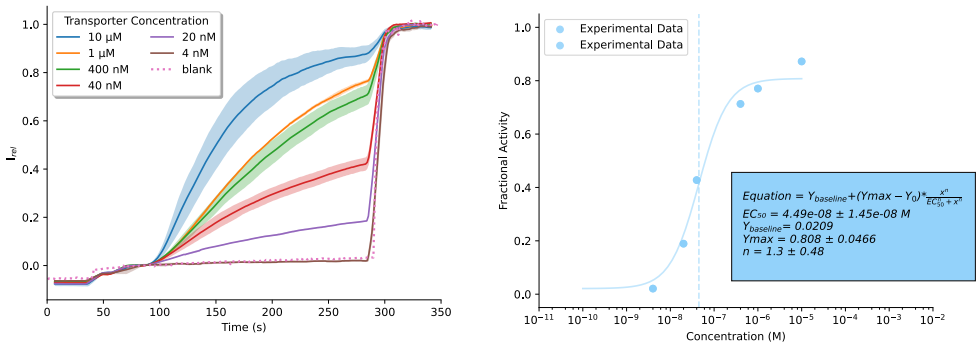



**Figure B.7:** HPTS assay data for carrier **2.2** with BSA and with FCCP.

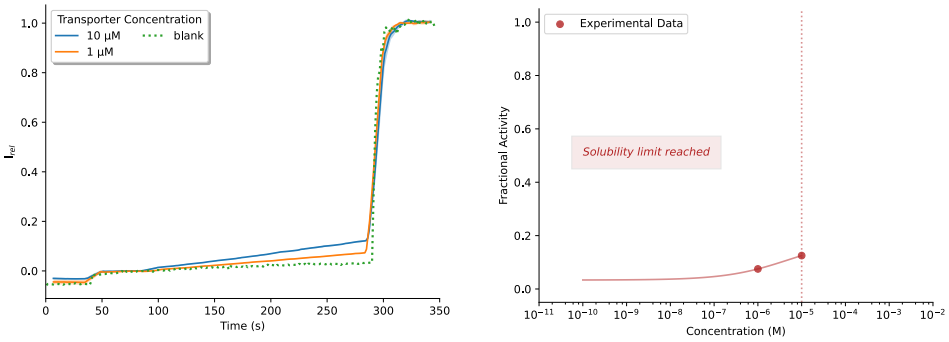



**Figure B.8:** HPTS assay data for carrier **2.2** with BSA and without FCCP.

### B.1.3 Sodium Gluconate Assay Data for all Transporters

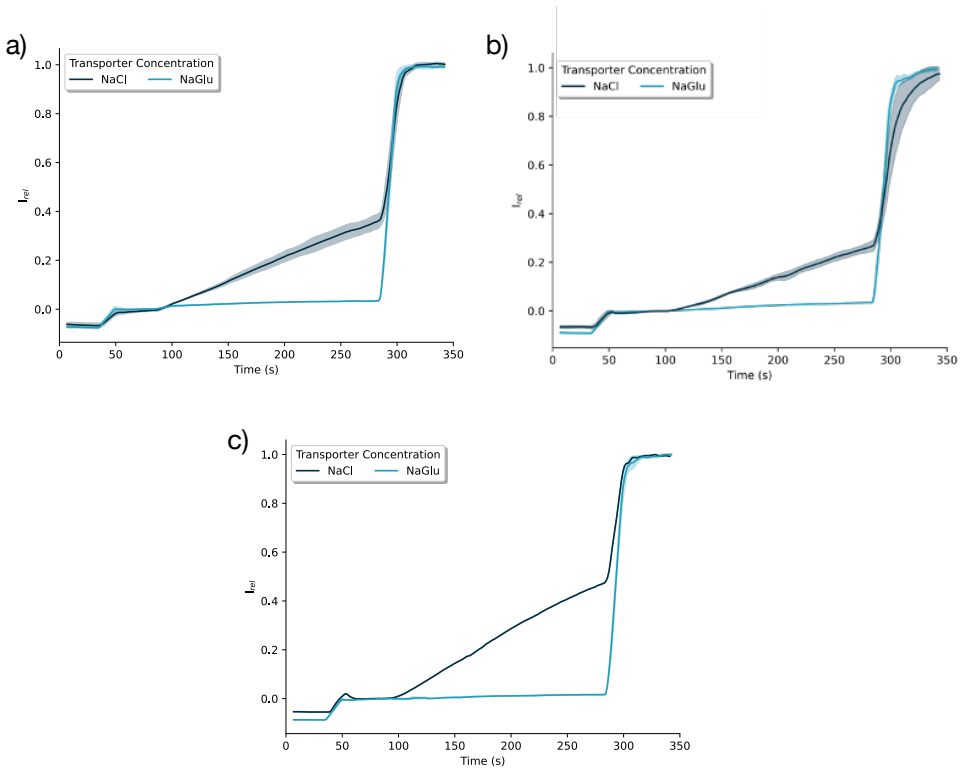



**Figure B.9:** The time-dependent change in fluorescence intensity within the gluconate variation of the HPTS assay. Carriers were administered at 5 uM. a) **2.2**, b) **2.3** c) **2.4**.

The decrease in transport with NaGlu for all carriers indicated they do not facilitate sodium cation transport (via $H^+/Na^+$ antiport) and must therefore operate via an anion transport mechanism.

### B.1.4 Membrane Fluidity Studies

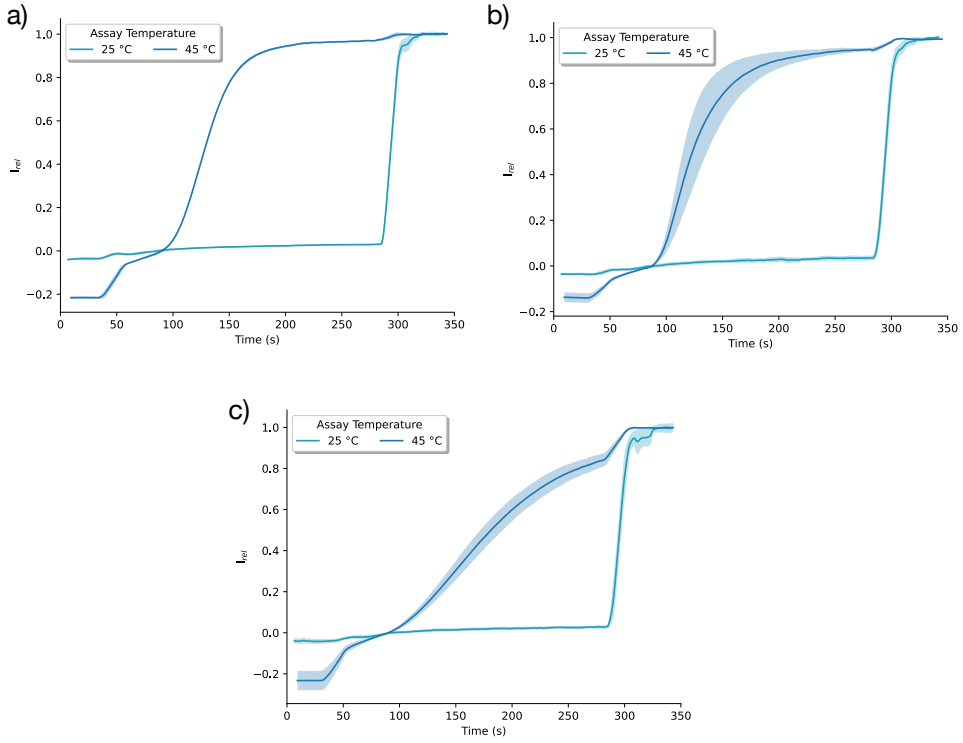



**Figure B.10:** Change in relative fluorescence over time in the HPTS assay utilizing DPPC LUVs at 25 °C and 45 °C, with the temperature controlled by a Peltier temperature controller. Carriers were administered at 1 µM at 90s following the addition of FCCP at 40s. a) **2.2**, b) **2.3** c) **2.4**.

### B.1.5 NMR Titration Experiments

$^1$H NMR titration spectra (left) and fitted 1:1 binding isotherms (right) fitted to concentration dependence of the downfield perturbation of the internal aryl signal of the bis-iodotriazole motif.

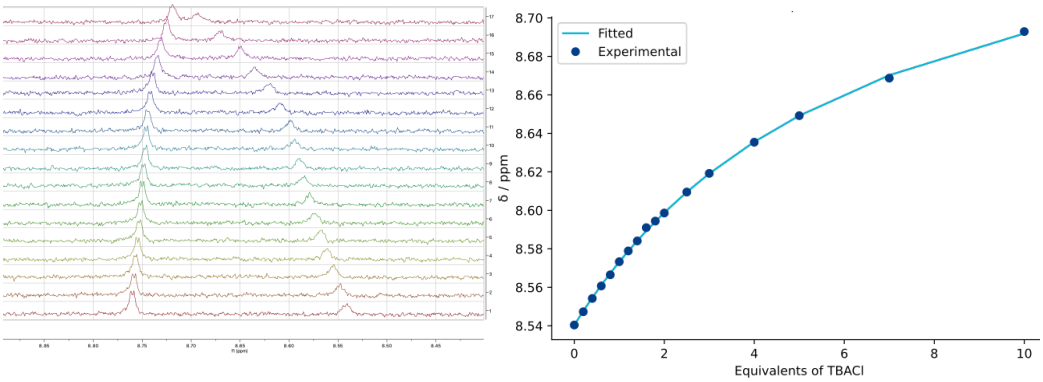



**Figure B.11:** Left: Truncated 1H NMR spectra of **2.2** upon titration with TBACl in 2.5% $D_2O$-acetone-$d_6$ (v/v) from 0 to 10 equivalents. Right: Corresponding $Cl^-$ binding isotherm. Experimental data shown by ·, fitted 1:1 binding isotherm shown by solid line.

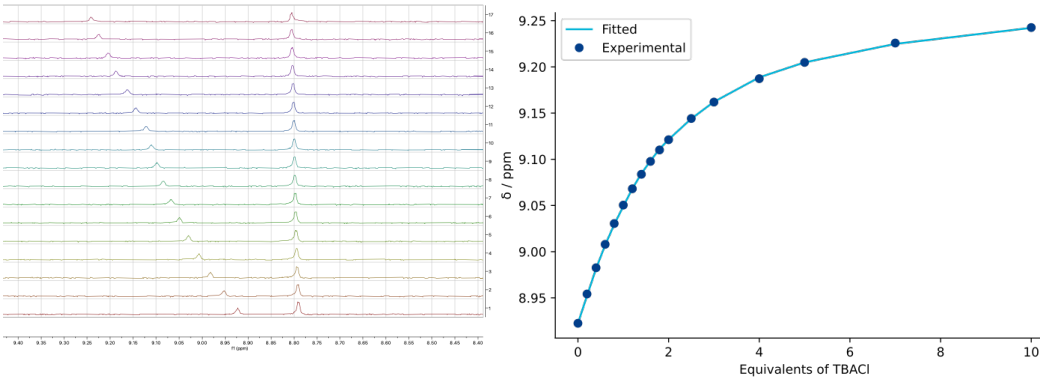



**Figure B.12:** Left: Truncated 1H NMR spectra of **2.3** upon titration with TBACl in 2.5% $D_2O$-acetone-$d_6$ (v/v) from 0 to 10 equivalents. Right: Corresponding $Cl^-$ binding isotherm. Experimental data shown by ·, fitted 1:1 binding isotherm shown by solid line.

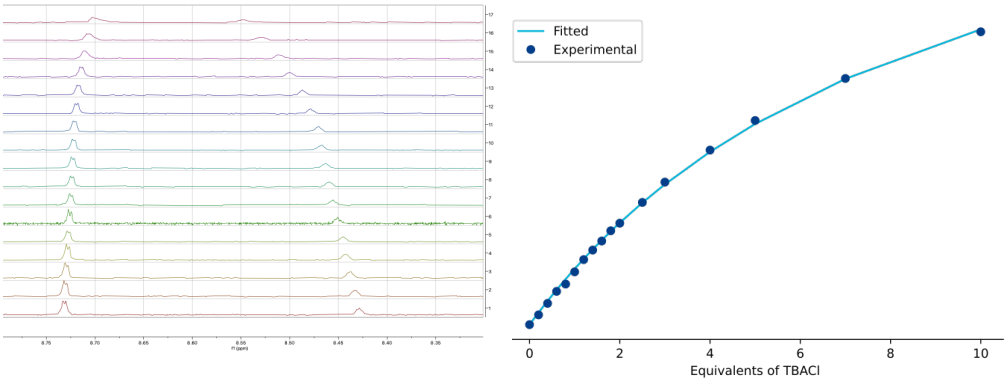



**Figure B.13:** Left: Truncated 1H NMR spectra of **2.4** upon titration with TBACl in 2.5% $D_2O$-acetone-$d_6$ (v/v) from 0 to 10 equivalents. Right: Corresponding $Cl^-$ binding isotherm. Experimental data shown by ·, fitted 1:1 binding isotherm shown by solid line.

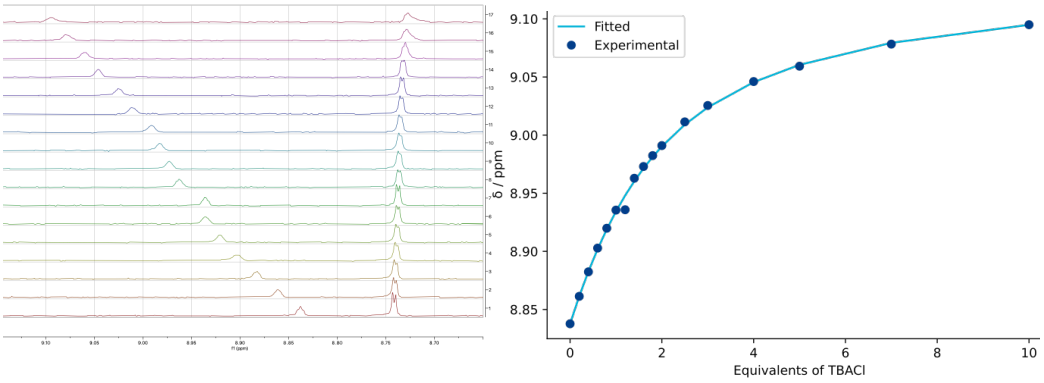



**Figure B.14:** Left: Truncated 1H NMR spectra of **2.5** upon titration with TBACl in 2.5% $D_2O$-acetone-$d_6$ (v/v) from 0 to 10 equivalents. Right: Corresponding $Cl^-$ binding isotherm. Experimental data shown by ·, fitted 1:1 binding isotherm shown by solid line.

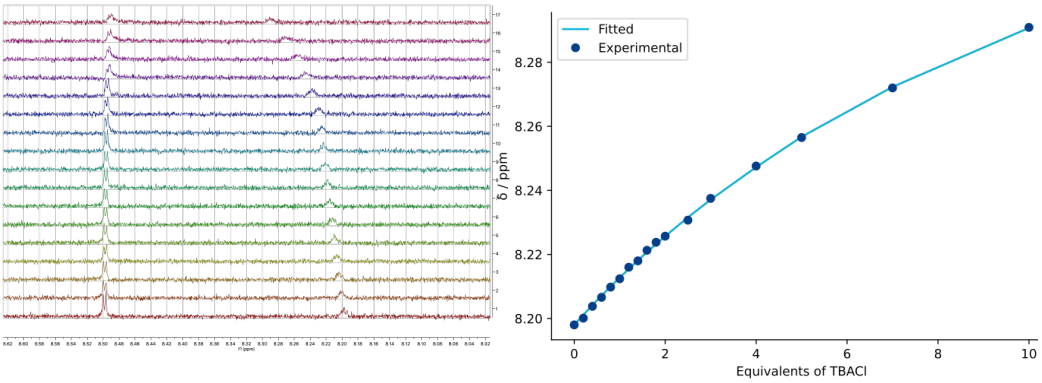



**Figure B.15:** Left: Truncated 1H NMR spectra of **2.6** upon titration with TBACl in 2.5% $D_2O$-acetone-$d_6$ (v/v) from 0 to 10 equivalents. Right: Corresponding $Cl^-$ binding isotherm. Experimental data shown by ·, fitted 1:1 binding isotherm shown by solid line.

## B.2 Chapter 3 Computational Results

### B.2.1 DFT Calculations Appendix

**Table B.1:** Complete details of computational calculations, including electronic energy ($E_{el}$), the zero point energy correction (ZPE), enthalpy (H), quasi-RRHO entropy contribution at 298.1 K (Tqh-S) and the total correction to the electronic energy giving the quasi-RRHO Gibbs free energy (qh-G) acalculated at the SMD($CHCl_3$)-$\omega$B97X-D3/def2-SVP level of theory (LL) or SMD($CHCl_3$)-$\omega$B97X-D3/def2-QZVP// $\omega$B97X-D3/def2-SVP level of theory (HL/LL). Free energies were calculated at 298.1 K and 1M. All values are in units of Eh.

| Species | $E_{el}$ | ZPE | H | Tqh-S | Total Corr | qh-G | $E_{el}$ (HL/LL) | qh-G (HL/LL) | qh-H (HL/LL) |
|---|---|---|---|---|---|---|---|---|---|
| $Cl^-$ | -460.2207 | 0.00000 | -460.2169 | 0.01438 | -0.01060 | -460.2313 | -460.3808 | -460.3914 | -460.3770 |
| $OH^-$ | -75.8195 | 0.00886 | -75.8162 | 0.01594 | -0.01263 | -75.8322 | -75.9109 | -75.9236 | -75.9076 |
| 3.1 | -2965.169 | 0.24568 | -2964.886 | 0.10176 | 0.18096 | -2964.988 | -2968.033 | -2967.852 | -2967.750 |
| $[3.1{\cdot}Cl]^-$ | -3425.413 | 0.24641 | -3425.128 | 0.10641 | 0.17903 | -3425.234 | -3428.436 | -3428.257 | -3428.151 |
| $[3.1{\cdot}OH]^-$ | -3041.036 | 0.25775 | -3040.739 | 0.10534 | 0.19161 | -3040.845 | -3043.989 | -3043.797 | -3043.692 |
| 3.2 | -2134.694 | 0.51295 | -2134.146 | 0.09390 | 0.45394 | -2134.240 | -2136.459 | -2136.005 | -2135.911 |
| $[3.2{\cdot}Cl]^-$ | -2594.931 | 0.51511 | -2594.380 | 0.09869 | 0.45211 | -2594.478 | -2596.857 | -2596.405 | -2596.307 |
| $[3.2{\cdot}OH]^-$ | -2210.551 | 0.52849 | -2209.987 | 0.09633 | 0.46751 | -2210.083 | -2212.407 | -2211.940 | -2211.843 |
| 3.3 | -2213.251 | 0.57280 | -2212.642 | 0.09819 | 0.51029 | -2212.741 | -2215.104 | -2214.593 | -2214.495 |
| $[3.3{\cdot}Cl]^-$ | -2673.491 | 0.57298 | -2672.880 | 0.10336 | 0.50767 | -2672.984 | -2675.503 | -2674.996 | -2674.892 |
| $[3.3{\cdot}OH]^-$ | -2289.112 | 0.58562 | -2288.489 | 0.10137 | 0.52196 | -2288.590 | -2291.054 | -2290.532 | -2290.431 |
| 3.4 | -2531.220 | 0.48219 | -2530.700 | 0.10082 | 0.41854 | -2530.801 | -2533.492 | -2533.073 | -2532.972 |
| $[3.4{\cdot}Cl]^-$ | -2991.460 | 0.48237 | -2990.939 | 0.10646 | 0.41539 | -2991.045 | -2993.891 | -2993.476 | -2993.370 |
| $[3.4{\cdot}OH]^-$ | -2607.081 | 0.49568 | -2606.546 | 0.10412 | 0.43079 | -2606.650 | -2609.442 | -2609.011 | -2608.907 |

**Table B.2:** Complete details of computational calculations, including electronic energy ($E_{el}$), zero point energy correction (ZPE), enthalpy (H), quasi-RRHO entropy contribution at 298.1 K (Tqh-S), and total correction to electronic energy giving the quasi-RRHO Gibbs free energy (qh-G). Calculations performed at the SMD($CHCl_3$)-$\omega$B97X-D3/def2-SVP level of theory (LL) or SMD($CHCl_3$)-$\omega$B97X-D3/def2-QZVP//$\omega$B97X-D3/def2-SVP level of theory (HL/LL) at 298.1 K and 1M. All values are in kcal $mol^{-1}$.

| **Process** | **LL** | | | | | **HL/LL** | | |
|---|---|---|---|---|---|---|---|---|
| | $E_{el}$ | **ZPE** | **H** | **Tqh-S** | **Total Corr** | $E_{el}$ | **qh-G** | **qh-H** |
| 3.1 + $Cl^-$ → [3.1·Cl]$^-$ | -14.9 | 0.5 | -15.6 | -6.1 | 5.4 | -14.1 | -8.7 | -14.8 |
| 3.1 + $OH^-$ → [3.1·OH]$^-$ | -30.1 | 2.0 | -23.2 | -7.8 | 14.6 | -28.4 | -13.8 | -21.5 |
| 3.2 + $Cl^-$ → [3.2·Cl]$^-$ | -9.8 | 1.4 | -10.3 | -6.0 | 5.5 | -10.9 | -5.4 | -11.4 |
| 3.2 + $OH^-$ → [3.2·OH]$^-$ | -23.2 | 4.2 | -15.2 | -8.5 | 16.4 | -23.4 | -6.9 | -15.4 |
| 3.3 + $Cl^-$ → [3.3·Cl]$^-$ | -12.5 | 0.1 | -13.2 | -5.8 | 5.0 | -11.7 | -6.7 | -12.5 |
| 3.3 + $OH^-$ → [3.3·OH]$^-$ | -26.3 | 2.5 | -19.0 | -8.0 | 15.3 | -24.8 | -9.6 | -17.6 |
| 3.4 + $Cl^-$ → [3.4·Cl]$^-$ | -12.6 | 0.1 | -13.4 | -5.5 | 4.7 | -11.9 | -7.2 | -12.7 |
| 3.4 + $OH^-$ → [3.4·OH]$^-$ | -26.3 | 2.9 | -18.7 | -7.9 | 15.6 | -24.4 | -8.8 | -16.7 |

**Table B.3:** Details of single point calculations of transporters distorted to their ion-binding poses with the anion removed, showing the electronic energy ($E_{el}$) for each transporter in its [transporter·ion]$^-$ binding pose. The energies were calculated at the $\omega$B97X-D3/def2-QZVP level of theory (HL). All values are in units of Eh.

| **HL Pose** | $E_{el}$ **(Eh)** |
|---|---|
| 3.1 in pose of [3.1·Cl]$^-$ | -2968.0309 |
| 3.1 in pose of [3.1·OH]$^-$ | -2968.0235 |
| 3.2 in pose of [3.2·Cl]$^-$ | -2136.4555 |
| 3.2 in pose of [3.2·OH]$^-$ | -2136.4491 |
| 3.3 in pose of [3.3·Cl]$^-$ | -2215.101 |
| 3.3 in pose of [3.3·OH]$^-$ | -2215.0931 |
| 3.4 in pose of [3.4·Cl]$^-$ | -2533.4905 |
| 3.4 in pose of [3.4·OH]$^-$ | -2533.4833 |

**Table B.4:** Details of distortion-interaction calculations of transporters distorted to their ion-binding poses without the anion. Energies calculated at the $\omega$B97X-D3/def2-QZVP level of theory (HL).

| **HL** | $\mathbf{E_{distortion}}$ **(kcal mol$^{-1}$)** | $\mathbf{E_{interaction}}$ **(kcal mol$^{-1}$)** |
|---|---|---|
| 3.1 in pose of [3.1·Cl]$^-$ | 1.3 | -15.5 |
| 3.1 in pose of [3.1·OH]$^-$ | 6.0 | -34.4 |
| 3.2 in pose of [3.2·Cl]$^-$ | 2.6 | -13.5 |
| 3.2 in pose of [3.2·OH]$^-$ | 6.5 | -29.8 |
| 3.3 in pose of [3.3·Cl]$^-$ | 1.9 | -13.7 |
| 3.3 in pose of [3.3·OH]$^-$ | 6.9 | -31.7 |
| 3.4 in pose of [3.4·Cl]$^-$ | 1.0 | -12.9 |
| 3.4 in pose of [3.4·OH]$^-$ | 5.5 | -29.9 |

**Table B.5:** Partition of vibrational energy between vibrational, rotational, and translational contributions for optimized structures at their energy minima calculated at the SMD($CHCl_3$)-$\omega$B97X-D3/def2-SVP level of theory.

| Species | $S_{\text{vibrational}}$ (Eh) | $S_{\text{rotational}}$ (Eh) | $S_{\text{translational}}$ (Eh) |
|---|---|---|---|
| $Cl^-$ | 0 | 0 | 0.01438342 |
| $OH^-$ | 0 | 0.00259858 | 0.01334352 |
| 3.1 | 0.06385005 | 0.01904634 | 0.01886851 |
| $[3.1{\cdot}Cl]^-$ | 0.0686549 | 0.01899347 | 0.01892698 |
| $[3.1{\cdot}OH]^-$ | 0.06750316 | 0.01894841 | 0.01889686 |
| 3.2 | 0.05706642 | 0.01818101 | 0.01865491 |
| $[3.2{\cdot}Cl]^-$ | 0.06166911 | 0.01829929 | 0.01872268 |
| $[3.2{\cdot}OH]^-$ | 0.05942489 | 0.01822567 | 0.01868782 |
| 3.3 | 0.06100278 | 0.01848747 | 0.0187088 |
| $[3.3{\cdot}Cl]^-$ | 0.06616269 | 0.01842617 | 0.01877409 |
| $[3.3{\cdot}OH]^-$ | 0.06427021 | 0.01836454 | 0.01874049 |
| 3.4 | 0.06344561 | 0.01859035 | 0.01878923 |
| $[3.4{\cdot}Cl]^-$ | 0.06914003 | 0.0184741 | 0.018851 |
| $[3.4{\cdot}OH]^-$ | 0.06689737 | 0.01840606 | 0.0188192 |

## B.3 Chapter 4 Computational Results

### Energy Details of Calculations - Amide Transporters

**Table B.6:** Electronic energy ($E_{el}$), zero-point correction (ZPE), enthalpy (H), quasi-RRHO entropy contribution ($T_{qh}S$), correction term, and Gibbs free energy ($G_{qh}$) at the $\omega$B97X-D3/def2-SVP level (LL) and HL/LL level. Energies in Hartrees.

| Species | $E_{el}$ | ZPE | H | $T_{qh}S$ | Corr | $G_{qh}$ | HL $E_{el}$ | HL $G_{qh}$ |
|---|---|---|---|---|---|---|---|---|
| $Cl^-$ | -460.0824 | 0.0000 | -460.0786 | 0.0144 | -0.0106 | -460.0930 | -460.2922 | -460.3028 |
| 4.12 | -1933.0507 | 0.4004 | -1932.6191 | 0.0879 | 0.3437 | -1932.7070 | -1935.3247 | -1934.9810 |
| $[4.12{\cdot}Cl]^-$ | -2393.2492 | 0.4010 | -2392.8148 | 0.0953 | 0.3391 | -2392.9101 | -2395.6990 | -2395.3599 |
| $[2(4.12)){\cdot}Cl]^-$ | -4326.3962 | 0.8036 | -4325.5278 | 0.1568 | 0.7116 | -4325.6846 | -4331.0960 | -4330.3844 |

**Table B.7:** Energy changes for binding processes. Energies in kcal $mol^{-1}$.

| Process | $\Delta E_{el}$ | $\Delta$ZPE | $\Delta$H | $\Delta T_{qh}S$ | $\Delta$Corr | $\Delta G_{qh}$ | HL $\Delta E_{el}$ | HL $\Delta G_{qh}$ |
|---|---|---|---|---|---|---|---|---|
| 4.12 + $Cl^- \rightarrow [4.12{\cdot}Cl]^-$ | -72.8 | 0.4 | -73.4 | -4.4 | 3.8 | -69.1 | -51.5 | -47.7 |
| 2(4.12) + $Cl^- \rightarrow [2(4.12){\cdot}Cl]^-$ | -133.2 | 1.7 | -132.4 | -20.9 | 21.8 | -111.5 | -96.9 | -75.1 |

### B.3.1 Final Geometries - Amide Transporters

Figures B.16–B.18 show geometries of the unbound transporter, [Transporter·Cl]$^-$, and [2Transporter·Cl]$^-$ at the $\omega$B97X-D3/def2-SVP level. Hydrogen bonds and possible $\pi$–$\pi$ interactions are indicated.

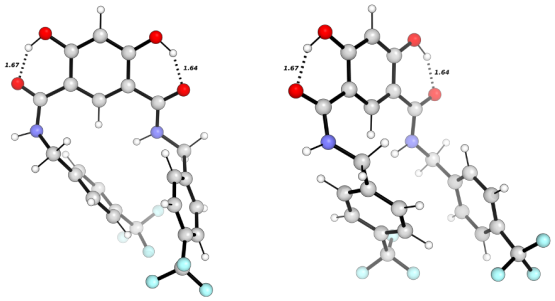


**Figure B.16:** Calculated geometry of the unbound transporter **4.12** at the $\omega$B97X-D3/def2-SVP level of theory. Hydrogen bond distances are shown as dashes.

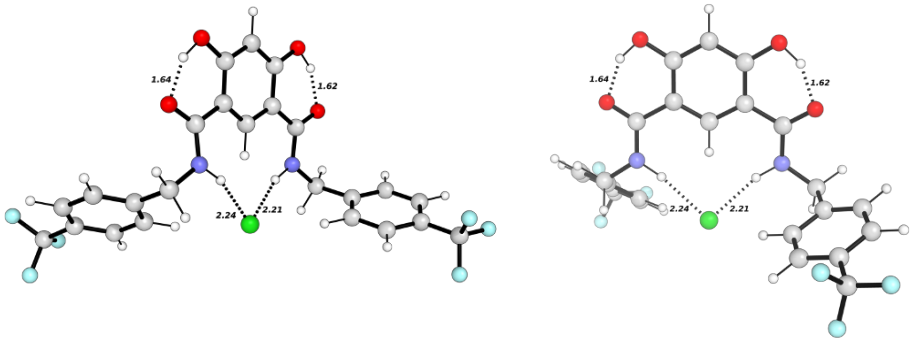



**Figure B.17:** Calculated geometry of the $Cl^-$-bound transporter **4.12** at the $\omega$B97X-D3/def2-SVP level of theory. Hydrogen bond distances are shown as dashes.

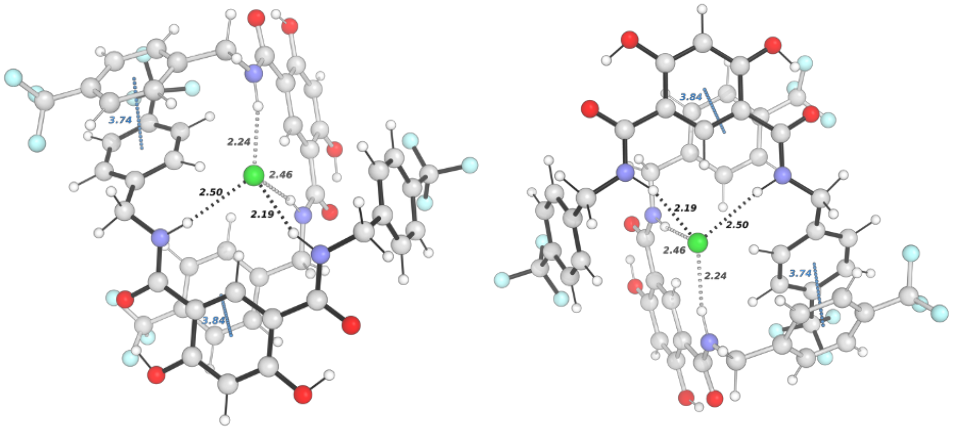



**Figure B.18:** Calculated geometry of the 2:1 complex of $Cl^-$-bound transporters **4.12** at the $\omega$B97X-D3/def2-SVP level of theory. Hydrogen bond distances are shown as dashes.

### B.3.2 Energy Details of Calculations - Hydrazone Transporters

Benchmarking with diffuse basis sets and SMD solvation confirmed minimal differences in binding energies (Figure B.19).

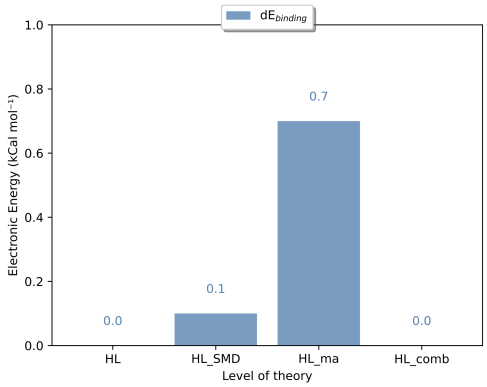



**Figure B.19:** Benchmarking of CPCM and SMD solvation models, as well as the diffuse ma-def2-QZVP basis sets, on the binding process of 4.16 + $Cl^-$ $\rightarrow$ [4.16·Cl]$^-$. The structure of unbound 4.16 optimised at the LL level of theory (the CPCM(DMSO)-$\omega$B97X-D3/def2-SVP), with energy calculations subsequently carried out at the HL (CPCM(DMSO)-$\omega$B97X-D3/def2-QZVP), HL$_{SMD}$ (SMD(DMSO)-$\omega$B97X-D3/def2-QZVP), HL$_{ma}$ (CPCM(DMSO)-$\omega$B97X-D3/ma-def2-QZVP), HL$_{comb}$ (SMD(DMSO)-$\omega$B97X-D3/ma-def2-QZVP). All energies are displayed relative to the binding energy at the HL level of theory.

**Table B.8:** Electronic energies and thermochemical corrections for hydrazone species (Hartree).

| **Species** | $E_{\mathbf{el}}$ | **ZPE** | **H** | $T_{qh}S$ | **Corr.** | $G_{qh}$ | **HL** $G_{qh}$ |
|---|---|---|---|---|---|---|---|
| $Cl^-$ | -460.20644 | 0 | -460.20066 | 0.01431870 | 0.01431700 | -460.21506 | -460.21785 |
| DMSO | -552.30947 | 0.07987531 | -552.28135 | 0.01464160 | 0.06523371 | -552.85476 | -553.21877 |
| 4.16 | -2041.16800 | 0.03380800 | -2040.75100 | 0.09338870 | 0.03273780 | -2040.94500 | -2041.10115 |
| $[4.16{\cdot}Cl]^-$ | -2801.41520 | 0.81391130 | -2800.90991 | 0.32023210 | 0.26874710 | -2804.95045 | -2805.02182 |
| $[4.16{\cdot}Cl]_d$ | -2801.41830 | 0.84579490 | -2800.90792 | 0.25525540 | 0.39253440 | -2805.31740 | -2805.61717 |
| [4.16·DMSO] | -2041.16170 | 0.04663647 | -2040.72990 | 0.09594184 | 0.03278640 | -2041.19940 | -2041.25954 |
| [4.16·DMSO] distorted | -2041.16570 | 0.81733701 | -2040.70622 | 0.13736690 | 0.03307600 | -2046.72427 | -2047.13779 |
| 4.16m | -2579.90560 | 0.04682182 | -2579.45904 | 0.10774843 | 0.03751780 | -2579.35450 | -2582.29029 |
| $[4.16m{\cdot}Cl]^-$ | -2119.26590 | 0.08946652 | -2119.19270 | 0.11104400 | 0.07490660 | -2119.59542 | -2122.14711 |
| $[1m{\cdot}Cl]_d$ distorted | -2119.66130 | 0.19490430 | -2119.18700 | 0.11408400 | 0.07455450 | -2119.28580 | -2121.78186 |

**Table B.9:** Complete details of computational calculations for hydrazone processes. LL values at CPCM(DMSO)-$\omega$B97X-D3/def2-SVP; HL/LL values from CPCM(DMSO)-$\omega$B97X-D3/def2-QZVP single points with LL thermochemistry. Energies in kcal mol$^{-1}$ (298.1 K, 1 M).

| | LL | | | | | | HL/LL | |
|---|---|---|---|---|---|---|---|---|
| **Process** | $E_{el}$ | ZPE | H | $T_{qh}S$ | Total Corr | $G_{qh}$ | $E_{el}$ | $G_{qh}$ |
| 4.16 + $Cl^-$ → [4.16·Cl]$^-$ | -26.7 | 0.9 | -27.2 | -6.9 | 6.4 | -20.3 | -8.8 | -2.5 |
| [4.16·Cl]$^-$ Binding Distortion | 0.7 | 0.4 | 0.7 | -0.8 | 0.8 | 1.5 | 0.8 | 1.6 |
| 4.16 + $DMSO^-$ → [4.16·DMSO]$^-$ | -21.7 | 1.8 | -19.9 | -11.5 | 13.3 | -8.4 | -12.5 | 0.8 |
| [4.16·DMSO] Binding Distortion | 1.6 | 0.0 | 1.5 | -0.1 | 0.0 | 1.6 | 1.4 | 1.3 |
| [4.16·DMSO] + $Cl^-$ → [4.16·DMSO] + $Cl^-$ | -5.0 | -0.9 | -7.3 | 4.6 | -7.0 | -12.0 | 3.7 | -3.3 |
| [4.16·Cl]$^-$ from [4.16·DMSO] Binding Distortion | -0.9 | 0.4 | -0.7 | -0.7 | 0.9 | 0.0 | -0.6 | 0.3 |
| 4.16m + $Cl^-$ → [4.16*m*·Cl]$^-$ | -17.1 | 0.2 | -18.0 | -6.0 | 5.1 | -12.0 | -3.0 | 2.1 |
| [4.16*m*·Cl]$^-$ Binding Distortion | 8.6 | -0.3 | 8.2 | -0.3 | -0.1 | 8.5 | 5.7 | 5.7 |

### B.3.3 Hydrazone Chloride Binding Umbrella Sampling Convergence

The first 10 ns of each umbrella sampling window was discarded, and convergence was monitored in increments of 10 ns from this 10 ns starting point using the GROMACS wham utility, which carries out weighted histogram analysis.[43] All of the three repeats of each window were then analysed together from 50 ns to 100 using bootstrapping analysis with 100 bootstraps to give an estimate of the uncertainty. Finally, the entropy uncorrected energy profile in Figure [**hydrazone_US_MD_convergence**] was corrected to account for the entropy loss of distance restrictions, equal to $2 * Kb * T * ln(r)$, where Kb is the Boltzmann constant, T is equal to 298 K and r is the distance from the binding site.

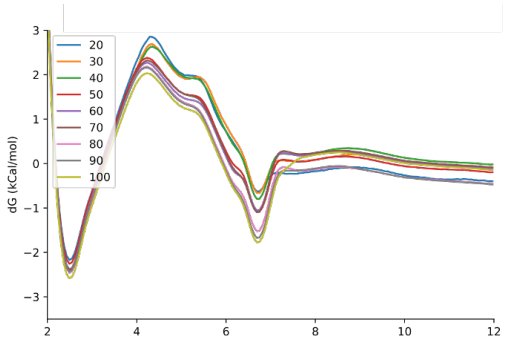



**Figure B.20:** Convergence analysis of the US simulations from 10 ns onwards, in increasing increments of 10 ns, with the line colour indicating the end-point of the wham analysis (e.g. at 100 ns this indicates the whole 10-100 ns).

## B.4 Chapter 5 Computational Results

**Table B.10:** Calculated gas phase single point binding energies (kcal/mol) for different functionals, as well as their MAE to the SCS-MP2 reference method. Binding energies are calculated as $E_{Binding} = E_{Host \cdot Guest} - (E_{Host} + E_{Guest})$. The subscript indicates the structures of either the complex or the individual species at the indicated level of theory using a def2-TZVP basis set in ORCA.

| **Functional** | **5.1** | **5.2** | **5.3** | **5.4** | **MAE**$_{vs.Ref}$ | **Eval. Time** |
|---|---|---|---|---|---|---|
| SCS-MP2 | -12.3 | -32.8 | -27.3 | -28.5 | 0.0 | 247 |
| MP2 | -15.6 | -34.5 | -28.7 | -27.0 | 1.98 | 239 |
| B2PLYP | -12.3 | -35.6 | -30.2 | -29.6 | 2.08 | 184 |
| CAM-B3LYP | -9.1 | -34.3 | -29.1 | -28.4 | 1.96 | 221 |
| M062X | -11.8 | -34.7 | -29.6 | -29.0 | 1.67 | 165 |
| M06L | -14.0 | -37.0 | -30.6 | -29.6 | 2.95 | 59 |
| PBE0 | -13.2 | -37.2 | -31.6 | -30.8 | 3.35 | 154 |
| PBE0 D3BJ | -16.5 | -38.2 | -32.4 | -31.9 | 4.90 | 154 |
| PBE0 D4 | -16.4 | -38.9 | -32.8 | -32.5 | 5.30 | 154 |
| wB97M-D3BJ | -10.5 | -33.8 | -28.5 | -28.3 | 1.33 | 309 |
| wB97M-V | -11.5 | -35.4 | -29.9 | -29.6 | 2.15 | 256 |
| wB97X | -10.1 | -34.2 | -28.8 | -28.6 | 1.67 | 254 |
| wB97X-D3BJ | -13.1 | -35.5 | -30.1 | -30.0 | 2.33 | 241 |

**Table B.11:** Calculated gas phase binding Gibbs free energies (kcal/mol) for different functionals, as well as their MAE to the SCS-MP2 reference method. Binding energies are calculated as $G_{Binding} = G_{Host \cdot Guest} - (G_{Host} + G_{Guest})$. The subscript indicates the structures of either the complex or the individual species at the indicated level of theory using a def2-TZVP basis set in ORCA.

| **Functional** | **5.1** | **5.2** | **5.3** | **5.4** | **MAE**$_{vs.Ref}$ |
|---|---|---|---|---|---|
| SCS-MP2 | -3.0 | -26.4 | -21.5 | -21.0 | 0.0 |
| MP2 | -6.1 | -28.7 | -22.9 | -23.0 | 2.20 |
| B2PLYP | -3.5 | -29.9 | -24.6 | -23.6 | 2.42 |
| CAM-B3LYP | -0.3 | -28.8 | -23.5 | -23.2 | 2.33 |
| M062X | -2.1 | -29.1 | -24.0 | -23.5 | 2.15 |
| M06L | -4.9 | -32.2 | -25.2 | -24.4 | 3.70 |
| PBE0 | -3.7 | -31.9 | -26.0 | -25.5 | 3.80 |
| PBE0 D3BJ | -6.9 | -32.7 | -26.9 | -26.5 | 5.28 |
| PBE0 D4 | -6.8 | -33.5 | -27.2 | -27.1 | 5.68 |
| wB97M-D3BJ | -1.8 | -27.9 | -22.9 | -22.6 | 1.42 |
| wB97M-V | -2.6 | -29.5 | -24.2 | -24.0 | 2.30 |
| wB97X | -0.4 | -28.3 | -23.2 | -22.7 | 1.98 |
| wB97X-D3BJ | -4.1 | -29.6 | -24.4 | -24.7 | 2.72 |

**Table B.12:** Experimental binding energies and calculated solvated binding Gibbs free energies (kcal/mol) for different functionals, SCS-MP2 reference method. Binding energies are calculated as $G_{Binding} = G_{Host \cdot Guest} - (G_{Host} + G_{Guest})$, with CPCM solvation used throughout (methanol for **5.1** and acetone for **5.2**, **5.3** and **5.4**). The subscript indicates the structures of either the complex or the individual species at the indicated level of theory using a def2-TZVP basis set in ORCA.

| **Functional** | **5.1** | **5.2** | **5.3** | **5.4** |
|---|---|---|---|---|
| Experiment | -4.7 | -3.1 | -2.5 | -4.3 |
| SCS-MP2 | -8.2 | 0.1 | 0.2 | 1.1 |
| MP2 | -4.1 | 0.8 | 1.1 | 0.7 |
| B2PLYP | -6.3 | -0.2 | 0.3 | 0.2 |
| CAM-B3LYP | -7.2 | 0.7 | 0.5 | -0.1 |
| M062X | -8.5 | 0.0 | -0.5 | -0.4 |
| M06L | -11.7 | -1.3 | -0.8 | -1.3 |
| PBE0 | -11.6 | -1.1 | -0.7 | -2.0 |
| PBE0-D3BJ | -9.8 | -2.1 | -1.5 | 1.3 |
| PBE0-D4 | -6.6 | -2.8 | -1.9 | 1.4 |
| wB97M-D3BJ | -4.8 | -0.0 | 0.1 | 0.3 |
| wB97M-V | -5.4 | -1.1 | -0.8 | -0.7 |
| wB97X | -3.5 | -1.0 | -0.8 | -0.9 |
| wB97X-D3BJ | -6.6 | -1.7 | -0.9 | -0.7 |

**Table B.13:** Solvation contributions to binding free energies (kcal mol$^{-1}$) across functionals. $\Delta G^{\text{bind}}_{\text{solv}} = G_{\text{solv}}(\text{Host} \cdot \text{Guest}) - G_{\text{solv}}(\text{Host}) - G_{\text{solv}}(\text{Guest})$

| **Functional** | **5.1** | **5.2** | **5.3** | **5.4** |
|---|---|---|---|---|
| SCS-MP2 | 5.5 | 14.8 | 15.1 | 20.4 |
| MP2 | 5.5 | 14.2 | 14.6 | 19.6 |
| B2PLYP | 5.3 | 16.0 | 16.9 | 22.3 |
| CAM-B3LYP | 5.2 | 15.4 | 16.4 | 21.7 |
| M062X | 5.2 | 14.6 | 15.4 | 20.2 |
| M06L | 4.7 | 16.1 | 16.9 | 23.2 |
| PBE0 | 5.1 | 17.2 | 18.1 | 24.0 |
| PBE0 D3BJ | 5.1 | 17.3 | 18.1 | 24.1 |
| PBE0 D4 | 5.1 | 17.0 | 18.1 | 23.9 |
| wB97M-D3BJ | 5.1 | 14.6 | 15.5 | 20.7 |
| wB97M-V | 5.1 | 15.3 | 16.0 | 21.5 |
| wB97X | 5.1 | 15.7 | 16.1 | 21.7 |
| wB97X-D3BJ | 5.2 | 14.3 | 15.0 | 20.5 |
| Std. Dev. | 0.21 | 1.07 | 1.21 | 1.57 |

**Table B.14:** Solvent box insertion statistics by structure, solvent, and box size. Tests carried out on 12 Xeon Gold 6330 CPU cores.

| Structure | Solvent | Box | Time(s) | Attempted | Added | Success% |
|---|---|---|---|---|---|---|
| **5.1** | acetonitrile | 10.0 | 0.132 | 5 | 5 | 100.0 |
| **5.1** | acetonitrile | 20.0 | 0.370 | 86 | 86 | 100.0 |
| **5.1** | acetonitrile | 30.0 | 1.263 | 305 | 305 | 100.0 |
| **5.1** | toluene | 10.0 | 0.222 | 3 | 3 | 100.0 |
| **5.1** | toluene | 20.0 | 4.692 | 43 | 36 | 83.7 |
| **5.1** | toluene | 30.0 | 17.354 | 150 | 126 | 84.0 |
| **5.1** | water | 10.0 | 0.073 | 19 | 19 | 100.0 |
| **5.1** | water | 20.0 | 0.356 | 253 | 253 | 100.0 |
| **5.1** | water | 30.0 | 1.483 | 888 | 888 | 100.0 |
| **5.2** | acetonitrile | 10.0 | 0.096 | 1 | 1 | 100.0 |
| **5.2** | acetonitrile | 20.0 | 0.387 | 82 | 82 | 100.0 |
| **5.2** | acetonitrile | 30.0 | 1.293 | 301 | 301 | 100.0 |
| **5.2** | toluene | 10.0 | 0.138 | 1 | 1 | 100.0 |
| **5.2** | toluene | 20.0 | 4.173 | 41 | 34 | 82.9 |
| **5.2** | toluene | 30.0 | 17.398 | 148 | 128 | 86.5 |
| **5.2** | water | 10.0 | 0.042 | 10 | 10 | 100.0 |
| **5.2** | water | 20.0 | 0.348 | 244 | 244 | 100.0 |
| **5.2** | water | 30.0 | 1.590 | 879 | 879 | 100.0 |
| **5.3** | acetonitrile | 10.0 | 0.089 | 4 | 4 | 100.0 |
| **5.3** | acetonitrile | 20.0 | 0.364 | 85 | 85 | 100.0 |
| **5.3** | acetonitrile | 30.0 | 1.358 | 304 | 304 | 100.0 |
| **5.3** | toluene | 10.0 | 0.219 | 2 | 2 | 100.0 |
| **5.3** | toluene | 20.0 | 5.257 | 42 | 35 | 83.3 |
| **5.3** | toluene | 30.0 | 17.649 | 150 | 128 | 85.3 |
| **5.3** | water | 10.0 | 0.046 | 17 | 17 | 100.0 |
| **5.3** | water | 20.0 | 0.353 | 251 | 251 | 100.0 |
| **5.4** | acetonitrile | 10.0 | 0.089 | 2 | 2 | 100.0 |
| **5.4** | acetonitrile | 20.0 | 0.319 | 82 | 82 | 100.0 |
| **5.4** | acetonitrile | 30.0 | 1.489 | 301 | 301 | 100.0 |
| **5.4** | toluene | 10.0 | 0.140 | 1 | 1 | 100.0 |
| **5.4** | toluene | 20.0 | 4.400 | 41 | 35 | 85.4 |
| **5.4** | toluene | 30.0 | 19.222 | 149 | 125 | 83.9 |
| **5.4** | water | 10.0 | 0.052 | 11 | 11 | 100.0 |
| **5.4** | water | 20.0 | 0.372 | 245 | 245 | 100.0 |
| **5.4** | water | 30.0 | 1.550 | 880 | 880 | 100.0 |

# C |

# Theoretical Appendix

## C.1 Electronic Structure Theory

### C.1.1 The Hartree-Fock Method and Self-Consistent Field

Before the rise of Density Functional Theory (DFT), the Hartree-Fock (HF) method was the cornerstone method for determining the properties of quantum systems in computational chemistry and physics. It provides an approximate solution to the electronic Schrödinger equation and introduces many concepts which are also helpful for understanding the procedures used in DFT computations. The formalism for this method is discussed in textbooks such as Atkins and Friedman's "Molecular Quantum Mechanics", which is the main source for this theoretical section, alongside Koch's "Chemist's Guide to Density Functional Theory".[64, 65]

#### C.1.1.1 The Hartree-Fock Ansatz and One-Electron Orbitals

The fundamental approximation in the Hartree-Fock method, or the Hartree-Fock Ansatz, is that the exact $N$-body electronic wave function of a system can be approximated by a single Slater determinant. This determinant is constructed from $N$ individual one-electron spin-orbitals, which can then be combined to varying degrees by fine-tuning their coefficients using the variational principle. Each electron occupies a spin-orbital, $\psi(x_i)$, which is a product of a spatial function,

$\phi(i)$, and a spin function, $\sigma(i)$:

$$\psi(x_i) = \phi(i)\sigma(i)$$

where $\phi(i)$ depends only on the spatial coordinates of the $i$-th electron, and $\sigma(i)$ denotes the spin (e.g., $\alpha(i)$ for spin-up and $\beta(i)$ for spin-down). These spin-orbitals are required to satisfy an orthonormality condition: $\langle\psi_i|\psi_j\rangle = S_{ij} = \delta_{ij}$, where $\delta_{ij}$ is the Kronecker delta, which is simply a function that can equal either 0, if the two input variables i and j are non-equal, or 1 if i=j. For a system with $n$ electrons, the Hartree-Fock wavefunction is expressed as an antisymmetrized product of $n$ spin-orbitals:

$$|\Psi(x_1, \ldots, x_n)\rangle \approx |\Psi_{HF}\rangle = \hat{A}(x_1, \ldots, x_n) \prod_i^n \psi_i(x_i)$$

Here, $x_i$ represents the combined spatial and spin coordinates of the $i$-th electron. The operator $\hat{A}(x_1, \ldots, x_n)$ is the antisymmetrizing operator, which ensures that the resulting many-electron wavefunction is antisymmetric with respect to the permutation of any pair of electrons. This antisymmetry is a direct mathematical consequence of the Pauli exclusion principle, which dictates that no two identical fermions (in this case electrons) can occupy the same quantum state simultaneously. The antisymmetriser can also be written out as a combination of all permutations:

$$\mathcal{A} \equiv \frac{1}{N!} \sum_{P \in S_N} (-1)^\pi \hat{P}$$

Where the permutation operator $\hat{P}$ represents the switching of electron labels in the Hartree product $\prod_i^n \psi_i(\mathbf{x}_i)$. When this permutation $P$ is combined with the sign factor $(-1)^\pi$ within the antisymmetrizing operator $\hat{A}$, it ensures the resulting Hartree-Fock wavefunction is antisymmetric with respect to the exchange of any

pair of electrons.

### C.1.1.2 Slater Determinants

For systems containing multiple electrons, the Hartree-Fock wavefunction is most elegantly represented as a Slater determinant, first introduced by John Slater as a way of maintaining antisymmetry of the wavefunction:

$$|\Psi_{HF}\rangle = \frac{1}{\sqrt{n!}} \begin{vmatrix} \psi_1(1) & \psi_1(2) & \dots & \psi_1(n) \\ \psi_2(1) & \psi_2(2) & \dots & \psi_2(n) \\ \vdots & \vdots & \ddots & \vdots \\ \psi_{n-1}(1) & \psi_{n-1}(2) & \dots & \psi_{n-1}(n) \\ \psi_n(1) & \psi_n(2) & \dots & \psi_n(n) \end{vmatrix}$$

This determinantal form inherently satisfies the Pauli exclusion principle because if any two rows or columns are identical (meaning two electrons occupy the same spin-orbital), the determinant becomes zero, making the wavefunction unacceptable. This construction also naturally accounts for the indistinguishability of electrons, as each electron is associated with every orbital within the determinant, which becomes apparent when one writes out the full product contained within the determinant. Let us illustrate this for a specific system, where only two electrons are present, each occupying a spin-orbital, $\chi_a(\mathbf{x})$ and $\chi_b(\mathbf{x})$. The general form of the Slater determinant for this system is:

$$\Psi(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{\sqrt{2!}} \det \begin{pmatrix} \chi_a(\mathbf{x}_1) & \chi_b(\mathbf{x}_1) \\ \chi_a(\mathbf{x}_2) & \chi_b(\mathbf{x}_2) \end{pmatrix}$$

where $\mathbf{x}_i$ denotes the combined spatial and spin coordinates of electron $i$. Expanding this determinant, we obtain:

$$\begin{aligned}\Psi(\mathbf{x}_1, \mathbf{x}_2) &= \frac{1}{\sqrt{2}} \left[\chi_a(\mathbf{x}_1)\chi_b(\mathbf{x}_2) - \chi_b(\mathbf{x}_1)\chi_a(\mathbf{x}_2)\right] \\ &= \frac{1}{\sqrt{2}} \left[(\text{electron 1 in } \chi_a,\ \text{electron 2 in } \chi_b) - (\text{electron 1 in } \chi_b,\ \text{electron 2 in } \chi_a)\right]\end{aligned}$$

This expanded form clearly shows a superposition of two configurations:

1. Electron 1 occupies spin-orbital $\chi_a$ and electron 2 occupies spin-orbital $\chi_b$.
2. Electron 1 occupies spin-orbital $\chi_b$ and electron 2 occupies spin-orbital $\chi_a$.

The presence of both terms ensures that it is impossible to distinguish which electron occupies which spin-orbital. Instead, the wavefunction describes a state where one electron is in $\chi_a$ and the other is in $\chi_b$, without specific assignment. Connecting this back to the fundamental requirement that the total probability of finding the electrons must be 1 (i.e., the integral of the square of the wavefunction $\int \Psi(\mathbf{x}_1, \mathbf{x}_2)^2$ over all space for all electrons must equal 1), we can see that the normalisation factor of $\frac{1}{\sqrt{2}}$ (or more generally $\frac{1}{\sqrt{N!}}$ for an $N$-electron system) ensures this.

#### C.1.1.3 The Self-Consistent Field (SCF) Method

The SCF method is an algorithmic way of improving the Hartree-Fock one-electron orbitals, and thus the overall wavefunction, in an iterative manner and in accordance with the variational principle. The reason why it needs to be done in an iterative way is that the Fock operator is defined as:

$$\hat{f}(1) = \hat{h}(1) + \sum_{j}^{n} (\hat{J}_j(1) - \hat{K}_j(1))$$

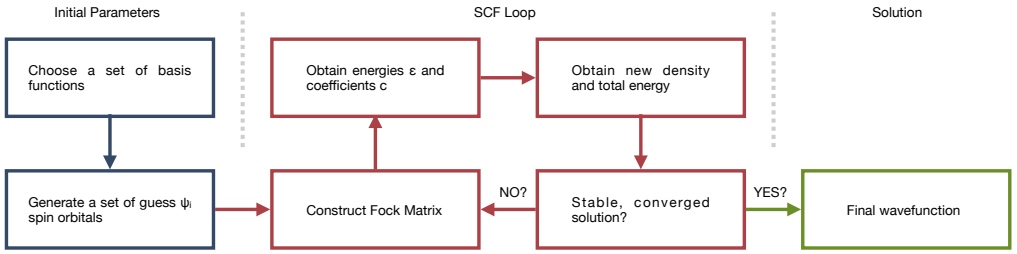



**Figure C.1:** Illustration of the SCF algorithmic procedure, where the SCF loop steps are repeated until the convergence criteria are met.

where $\hat{h}(1)$ is the one-electron Hamiltonian, $\hat{J}_j(1)$ is the Coulomb operator (representing classical electron-electron repulsion), and $\hat{K}_j(1)$ is the exchange operator (a purely quantum mechanical effect arising from the antisymmetry of the wavefunction). Since $\hat{f}$ defines the effective potential experienced by an electron, it itself then depends on the very spin-orbitals it is meant to determine. This is because the electron-electron repulsion $\hat{J}_j(1)$ is defined by the repulsion of a single electron with the effective field of all of the other electrons. In each loop of the SCF procedure, this effective field changes, as the electron distributions mapped by the wavefunctions are changed in order to minimise the energy. This creates a feedback loop. The SCF procedure (Figure C.1) is generally composed of the following algorithmic steps:

1. An initial guess for the spin-orbitals (and thus the electron density) is made.
2. Using this initial guess, the Fock matrix is constructed, which includes contributions from one-electron integrals (kinetic energy and nuclear attraction) and two-electron integrals (electron-electron repulsion, both Coulomb and exchange)
3. The Hartree-Fock equations, $\hat{f}|\psi_i\rangle = \epsilon_i|\psi_i\rangle$, are solved one by one to obtain a new set of spin-orbitals and their corresponding orbital energies
4. These newly obtained spin-orbitals are used to generate an updated electron

density and, consequently, a new Fock matrix.

5. The process is repeated (steps 2-4) until the spin-orbitals, the electron density, and the total energy converge to a stable solution, meaning the change in these values between successive iterations falls below a predefined threshold value, which is deemed acceptable.

In implementations of the SCF procedure, spin-orbitals are often expanded linear combinations of atom-centered basis functions (e.g., Slater or Gaussian-type orbitals), and the iterative procedure optimizes the coefficients of these basis functions. By incresing the size and flexibility of the basis sets used (by increasing the number of constituent functions), the calculation of the energy approaches the "Hartree-Fock limit," which represents the lowest possible energy achievable within the HF approximation for a given system. For closed-shell molecules (where all electrons are paired), the Restricted Hartree-Fock (RHF) theory can be employed to reduce computational cost by restricting $\alpha$ and $\beta$ spin-orbitals to share the same spatial orbitals.

### C.1.2 Early Density-Based Models

An alternative approach to dealing with wavefunctions is to describe the properties of atoms and molecular systems by considering the electron density directly. The earliest attempt of such an approach was the Thomas-Fermi model, proposed in 1927, and represented a significant conceptual leap, demonstrating the potential of density-based approaches as precursors to modern DFT. It offered an easier way to calculate most properties of electrons, but neglected to include the quantum effects of exchange and correlation. Following this early work, two sets of theorems were necessary to provide the theoretical foundation for DFT. In simple terms, the first Hohenberg-Kohn theorem establishes the fact that the electron density of

an electronic ground state can be itself used to define the external potential (i.e. knowing the electron density, one can know the identity and the position of the nuclei). The second theorem states that there is a functional that exists, for which the ground state density will give the minimum energy, and thus minimising the energy of the functional by varying the density will lead to better approximations of the ground state density. A deeper discussion of these theorems and their theoretical proofs can be found in Atkins and Friedman's Molecular Quantum Mechanics.[64]

### C.1.3 The Kohn-Sham Equations

The Hohenberg-Kohn theorems establish the existence of a density functional but do not provide a practical method for its calculation. The ground-breaking approach proposed by Kohn and Sham addresses this by introducing a fictitious system of non-interacting electrons that, by definition, has the same ground-state electron density as the real, interacting system. This transformation significantly simplifies the kinetic energy term, which is the most challenging component to express directly as a functional of the density for interacting systems.

The total energy functional $E[n]$ is then decomposed into several physically interpretable terms and one unknown term:

$$E[n] = T_s[n] + E_{ext}[n] + E_C[n] + E_{xc}[n]$$

where:

- $T_s[n]$ is the kinetic energy of the non-interacting system, and is determined by the standard $-\frac{\hbar^2}{2m}\nabla^2$ kinetic energy operator.
- $E_{ext}[n] = \int v_{ext}(\mathbf{r})n(\mathbf{r})d\mathbf{r}$ is the energy due to the external potential $v_{xc}(\mathbf{r})$

(representing the attraction between the electrons and nuclei).

- $E_C[n] = \frac{1}{2}\iint \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|} d\mathbf{r}d\mathbf{r}'$ is the classical electrostatic repulsion or Coulumb energy $v_c(\mathbf{r})$, representing the classical electron-electron repulsion from electrons moving in a mean field.

- $E_{xc}[n]$ is the **exchange-correlation energy**. This crucial term is defined as the difference between the true kinetic energy of the interacting system and $T_s[n]$, plus the difference between the true electron-electron interaction energy and $E_H[n]$. It represents the quantum correction to these classical terms, encapsulating all the complex many-body effects they do not account, which is composed of the electron exchange, electron correlation and self-interaction error. It is the magic component of DFT, and has been the source of intense research in the field. It is the component which makes or breaks a DFT functional.

Solving the Kohn-Sham equations follows an analogous self-consistent field procedure as in Hartree-Fock. Instead of using guess wavefunctions as the starting point, trial densities are used to construct the initial Kohn-Sham potential. For example, one could use the atomic densities for a simple system. This potential can then be used to solve the Kohn-Sham equations, varying the coefficients of the basis sets by diagonalising in order to minimise the energy and giving rise to an improved set of orbital coefficients. The process is repeated iteratively, as in the case of the Hartree-Fock procedure, until convergence criteria are satisfied.

### C.1.4 Approximations to the Exchange-Correlation Functional

Due to the unknown nature of the *"true"* exchange-correlation functional, several generations of increasingly sophisticated approximations have been developed. These were spoken about by John P. Perdew as a "Jacob's Ladder" of DFT func-

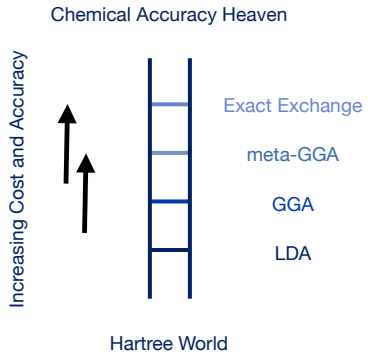



**Figure C.2:** Jacob's ladder of exchange-correlation functionals, spanning the accuracy and cost range from Hartree-Fock to so-called chemical accuracy, which would be attained by the perfect functional.

tionals, where each rung of the ladder gets increasingly more accurate, but also represents a much more expensive calculation in practice.[66] As computational resources get increasingly more powerful, the higher rungs on the ladder become increasingly accessible to the computational chemistry field for larger and more complex systems.

### C.1.4.1 Local Density Approximation (LDA)

The local density approximation or the LDA was the first approximation of the exchange-correlation functional to be developed. It approximates the XC energy density as a direct function of the electron density at the same position. This approximation is said to be equivalent to treating the local electron density as being that of a uniform electron gas of the same density.[67] The derivation of this energy is known, the method is inexpensive and can be used effectively for solids and surfaces with reasonable accuracy. Functionals belonging to this rung have been found to overbind molecules and provide equilibrium bond lengths that are much too short.[68] Most popular parameterisations of the LDA functional are known as VWN5, PZ81 and PW92.[68]

### C.1.4.2 Generalised Gradient Approximation (GGA)

The next rung on the ladder is occupied by the generalised gradient approximation, which improves upon the LDA by making the XC energy depend also on the gradient of the density at a point in space. This addition, corrects for the inhomogeneous nature of the electron cloud and allows for much more diverse chemical environments to be studied, giving more accurate bond lengths and binding energies, despite still suffering from overbinding.[69] A very popular example of a GGA functional is the PBE (or the Perdew-Burke-Ernzerhof approximation).[70]

### C.1.4.3 Meta-GGAs and Hybrid Functionals

There are two different ingredients which have been used to further improve upon the GGA functionals. The first of these again depend on higher order descriptions of either the density or the kinetic energy density, with functionals generally parameterising for the use of one or the other second-order derivative.[68] These functionals are termed meta-GGAs, and significantly improve on GGAs, but can have a tendency to underbind for some weakly-interacting systems.

Finally, one of the key issues present in all rungs discussed thus far - the self-interaction error - was addressed by introducing a portion of exact Hartree-Fock exchange to functionals, in so-called hybrid functionals. These functionals, of which PBE0 and $\omega$B97X are two examples, can also be coupled with dispersion corrections, which improve on dynamic correlation, to give the gold standard of modern functionals for use with molecular systems.

## C.2 Molecular Mechanics Theory

In molecular mechanics, the energy of a system is modelled using classical approximations. This allows for any quantum calculations of electronic structure to be bypassed, greatly speeding up the simulations. The potential energy function which models the energy of the system is known as a force field (FF), and the interactions of the system are generally broken down into bonded and non-bonded components:

$$U = U_{\text{bonded}} + U_{\text{non-bonded}}$$

The terms which make up each of these components will be discussed in the following sections. The discussion is broadly based on the textbooks Molecular Dynamics by Ruben Santamaria and Understanding Molecular Simulation by Drenkel and Smit.[71, 72]

Since interactions in classical force-fields depend on pre-defined functions, there must exist a definition for each combination of atoms which are interacting. This set of definitions is what defines the force field. This means that in order to model the diverse chemical environments within a molecule, each atom is assigned a specific atom type, which goes beyond simply its elemental identity, in order to pinpoint how it will interact with other atom types. For example, a carbon atom in an alkane will have a different atom type than a carbon in an aromatic ring or a carbonyl group. This detailed classification allows for the generation of specific parameters (e.g., equilibrium bond lengths, force constants, partial charges, Lennard-Jones parameters, which will all be discussed below) for every possible interaction, such as bond stretching between two defined atom types, angle bending involving three, or non-bonded interactions between any pair. This systematic parameterization, based on atom types, enables the calculation of the system's total potential energy.

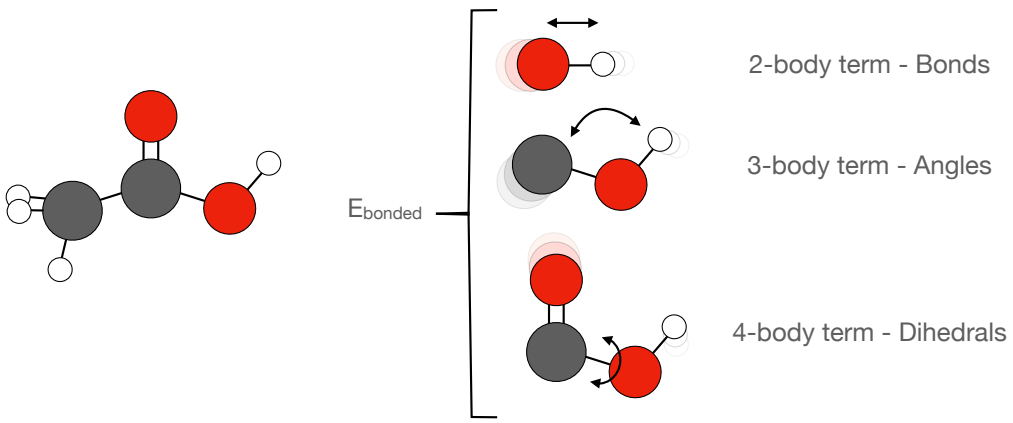



**Figure C.3:** The three different bonded terms which commonly compose the bonded component of a molecular mechanics force field.

### C.2.1 Bonded interactions

While chemical bonds are fundamentally a property of the quantum mechanical behaviour of electrons, they are nevertheless approximated classically in molecular dynamics. One concession of such a model is that chemical reactions cannot be modelled using most conventional force fields. In order to capture the different possible degrees of freedom that are characteristic of molecular motion, different bonded interactions are constructed from several different terms. These include two-body bond stretches, three-body bond bending, and four-body torsional interactions:

$$U_{\text{bonded}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{torsion}}$$

These are then summed together at each step of the simulation, alongside the non-bonded interactions.

#### C.2.1.1 Bond Stretching

The bond stretching potential is generally a harmonic potential centered about a bond's equilibrium distance $r_0$:

$$U_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} k_b (r - r_0)^2$$

with the stiffness of the bond dictated by the spring constant k. Harmonic potentials directly obey Hooke's law, which means that they are symmetrical about the equilibrium distance exactly like a classical spring. This means that they cannot model bond-breaking and -making, and that the potential is inaccurate at both short- and long-range interatomic distances. In certain forcefields, the bond potential can be improved using an additional anharmonic term, such as is the case in the MMFF94 and UFF forcefields, producing a bond stretching potential of the form:[71]

$$U_{\text{bond}} = \sum_{\text{bonds}} \frac{1}{2} k_b ((r - r_0)^2 - \alpha (r - r_0)^3)$$

with the $\alpha$ parameter dictating the strength of the anharmonicity of the bond's vibration. Another way of accounting for anharmonic effects is to use the Morse potential:

$$U_{\text{morse}} = D_e \left[1 - e^{-\alpha (r - r_0)}\right]^2$$

where the $\alpha$ parameter is related to the bond's force constant, and $D_e$ is the bond's dissociation energy. Both of these approaches increase the number of evaluations at each step of the calculation, therefore increasing the cost and potentially the timescale of simulations.

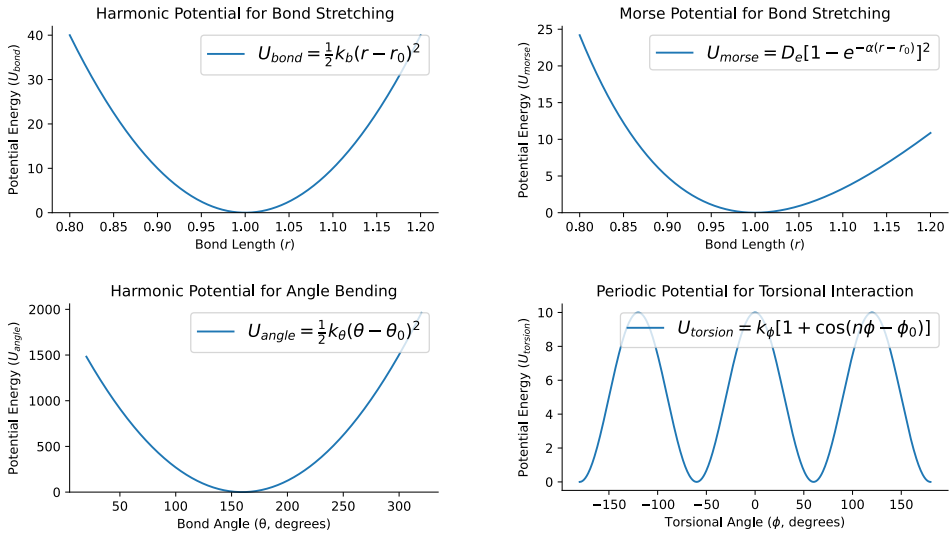



**Figure C.4:** The different mathematical forms of the 2-, 3-, and 4-body terms that make up the bonded portion of a force field, which are discussed in this section.

### C.2.1.2 Bond Bending

The scissor-like motion of three-atom groups along the angle between them is likewise generally modelled as a harmonic spring, with the only difference being that the angle takes the place of the distance found in the stretching harmonic potential:

$$U_{\text{angle}} = \sum_{\text{angles}} \frac{1}{2}k_\theta(\theta - \theta_0)^2$$

More complicated forms of bond-angle potentials may also include cubic and quartic correction terms, or they may be combined with both of the bond stretching terms to produce a product depending on all three values, giving rise to a mixed potential.[71]

#### C.2.1.3 Bond Torsions

Torsional terms describe the rotation of four atoms about their most central of the three bonds. This can also be imagined as plane bending of the two planes described by the two sets of three atoms that are bonded together. The behaviour of dihedral torsions can be described by several different types of potentials, but is very commonly modelled as a cosine function:

$$U_{\text{torsion}} = \sum_{\text{torsions}} k_\phi [1 + \cos(n\phi - \phi_0)]$$

where the $k_\phi$ value once again represents the constant, which will dictate the height of the barrier of rotation about the dihedral bond, n is the periodicity of the cosine function, and $\phi_0$ is the dihedral angle's equilibrium value. Alternative forms of the dihedral potential can also be quadratic functions to give a harmonic potential or a combined cosine quadratic function, such as:[71]

$$V_p = k[\cos(\theta_{ijkl}) - \cos(\theta^0_{ijkl})]^2/2$$

### C.2.2 Non-bonded Terms

Non-bonded terms are comprised of interactions between atoms which are not directly bonded to each other. The main non-bonded interactions captured by force fields are the Coulomb or electrostatic potential ($U_{\text{elec}}$) and the Van der Waals interaction, which is normally modelled by the Lennard-Jones poential ($U_{\text{LJ}}$):

$$U_{\text{non-bonded}} = U_{\text{LJ}} + U_{\text{elec}}$$

#### C.2.2.1 Electrostatic Potential

Due to the polar nature of atoms, which have a pre-defined and fixed charge in most forcefields, they experience a mutual level of attraction or repulsion at all times. These interactions might be with other atoms within the same molecule, or it could be with the atoms of neighbouring molecules, such as solvents or non-covalently interacting host/guest molecules. The well-known form of the Coulumb potential, which states that the force acting on two particles is proportional to their charges, and inversely proportional to the distance between them, is:

$$U_{\text{elec}} = \sum_{i<j} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}}$$

where $q_i$ and $q_j$ are the charges of the respective interacting atoms, $r_{ij}$ is the interatomic separation, and $epsilon_0$ is the permitivity of the medium . The calculation of electrostatic contributions between all particles in the box, as well as across the periodic boundary conditions, is a very computationally expensive process, because of the 1/r decay of electrostatic interactions. In principle, this would mean calculating the contributions not only from neighbouring boxes, but even boxes much further away in the periodic image scheme, leading to a calculation of infinite sums.[71] The Ewald summation scheme, which attempts to split the interactions into short-range and long-range parts by considering the former in real space and the latter in reciprocal Fourier space in order to speed up convergence.

#### C.2.2.2 Van der Waals (VdW) Interactions

Van der Waals forces are most important when the molecules are at an intermediate distance to each other. They act repulsively when two atoms get too close, become slightly attractive at the minimum, and then rapidly decrease to 0 at longer distances.[71] The repulsive part of the function is a result of Pauli exclusion, which

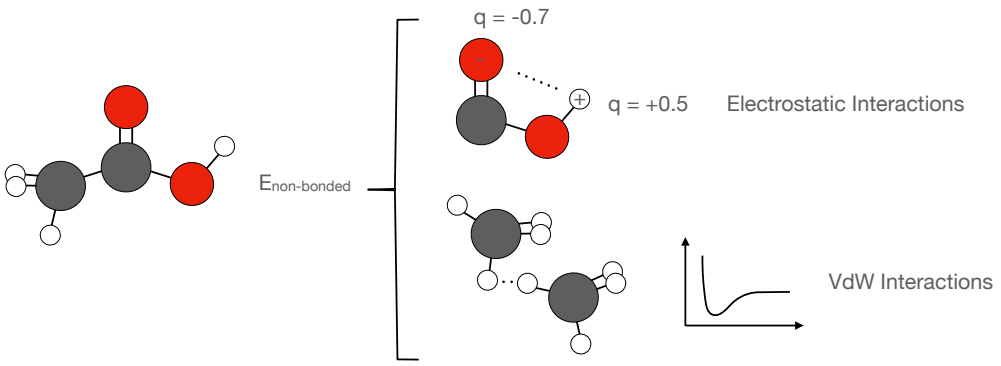



**Figure C.5:** The two main types of non-bonded interactions present in molecular mechanics force fields.

prevents the electrons of same spin occupying the same spatial regions, which happens when atoms get too close together. The longer-range contributions to the interaction are due to fluctuations in electron density which create instantaneous dipoles and lead to weak attraction. The function which is generally used to model this combined short-range repulsion and medium- and long-range attraction is the Lennard-Jones (LJ) potential, which has an minimum point at an $r_o$ internuclear distance. This is usually calculated from the atom types in the force field as the sum of the two atoms' Van der Waals radii. The LJ potential has the formula:

$$U_{\mathrm{LJ}} = \sum_{i<j} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]$$

where $\epsilon_{ij}$ dictates the strength of the attractive interaction; $\sigma_{ij}$ is the point where the potential is equal to 0; and $rij$ is the internuclear distance. This can also be reformulated in terms of the equilibrium distance, where the potential reaches its

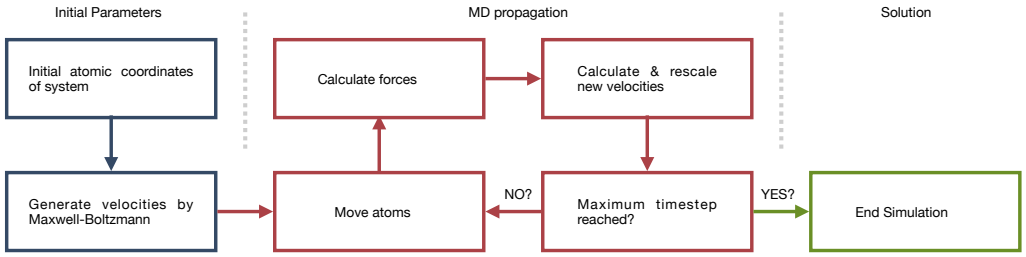



**Figure C.6:** General schematic of the steps in a molecular dynamics propagation algorithm.

minimum, $r_0$:

$$U_{\mathrm{LJ}} = \sum_{i<j} \epsilon_{ij} \left[ \left( \frac{r_0}{r_{ij}} \right)^{12} - 2 \left( \frac{r_0}{r_{ij}} \right)^{6} \right] \; ; r_0 = 2^{1/6}\sigma$$

Different molecular mechanics software may have different conventions for specifying the Van der Waals parameters, and care must be taken when converting the parameters e.g. from AMBER to Gromacs.

### C.2.3 Molecular Dynamics Propagation

Several different integrators exist for numerically solving the Newtonian equations of motion. Their task is to update the velocities and positions of atoms over discrete time-steps.[71] The balance of choosing the correct time-step is an important one. The time-step should be long enough to prevent calculating the forces too many times, while it should not be so long as to cause the bonds to vibrate to extreme positions. A bond's vibrational frequency (related to its force constant) can thus be a limiting factor when choosing the timestep for a simulation. There are many different families of integrators, but they all go through the same general cycle of updating positions and velocities, over and over again, until the desired timescale of the full simulation has been reached. The most commonly used family of integrators are the Verlet algorithms, which will be described in more detail.

#### C.2.3.1 Verlet Algorithm

The Verlet algorithm predicts the positions of atoms purely from their current and past positions, and its acceleration. The equation that is used to do this is:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + a(t)(\Delta t)^2$$

Let us explain these terms in a little more detail.

- $r(t + \Delta t)$ is the atom's position at the next time-step
- $r(t)$ is the atom's current position
- $r(t - \Delta t)$ is the atom's position one time-step before
- $a(t)$ is the atom's acceleration, which is calculated from the forces acting on the atom - i.e. by summing the forces from all of the bonded and non-bonded terms
- $\Delta t$ is the size of the time-step

This algorithm infers the position after each time-step without actually explicitly calculating or storing the velocities, which complicates the calculation of the total kinetic energy of the system, requiring subsequent algorithms like Leap-frog.[71, 72] As many thermostats and barostats act directly on velocities, this means that the basic Verlet algorithm cannot be used in combination with these. The leap-frog algorithm solves some of these challenges.

#### C.2.3.2 Leap-frog Verlet Algorithm

The leap-frog algorithm solves the issue of a lack of velocities by updating the velocities and positions in an alternating, so-called leap-frog fashion.[71] First, the velocities are updated to a half-step in time, then these new velocities are used to

propagate the positions for a full time-step, following which the velocities are then updated again to a full step. The reason for the staggered updating of positions is to ensure numerical stability, while also improving energy conservation and ensuring that the simulation remains reversible as in the case of the regular Verlet algorithm.[71]

The steps in the leap-frog algorithm have the following mathematical formulas:

- Update velocities to half-step: $v\left(t + \frac{\Delta t}{2}\right) = v(t) + a(t)\frac{\Delta t}{2}$
- Update positions to full step: $r(t + \Delta t) = r(t) + v\left(t + \frac{\Delta t}{2}\right)\Delta t$
- Calculate new accelerations $a(t + \Delta t)$: This is the **force calculation**, the most expensive step.
- Update velocities to full step: $v(t + \Delta t) = v\left(t + \frac{\Delta t}{2}\right) + a(t + \Delta t)\frac{\Delta t}{2}$

This staggered calculation uses the acceleration at the initial position to advance the velocity halfway, and then uses the force from the final position to complete the velocity update. This staggered way of updating both quantities minimizes errors in the propagation of both. The expensive calculation of the forces and with them the accelerations is only carried out once, meaning that there is no significant increase in cost compared to, say, the velocity Verlet algorithm, which carries out both updates simultaneously.[72]

### C.2.4 Ensembles

An ensemble in the context of statistical mechanics, and in the use case of molecular dynamics simulations, refers to the collection of all possible microscopic states (arrangements and energetic states of all simulated particles) that a system occupies over time.[73] The ensembles differ between each other in the macroscopic properties that are held constant, and those that are allowed to vary. Ensembles in

classical dynamics are generally closed systems, with $N$, the number of particles remaining constant, with another two of the volume $V$, temperature $T$, pressure $P$, or energy $E$ being held constant. This corresponds to different experimental setups - for example a reaction in a closed container with unvarying volume would correspond to the NVT ensemble, giving Helmholtz energies, while the NPT ensemble where the volume is allowed to vary gives Gibbs free energies. The different ensembles will be briefly discussed.

#### C.2.4.1 Microcanonical (NVE) Ensemble

In the microcanonical ensemble, the particles are simulated in a box with a fixed volume, and no energy is exchanged with the surroundings. The steps are propagated exclusively through Newton's equations of motion, and the forces acting on all particles are exclusively a result of the interactions between particles dictated by the forcefield. The microstates are described by:

$$\Omega(N, V, E) \text{ and the associated entropy; } S = k_b ln(\Omega)$$

The microcanonical ensemble simulates an isolated system, and therefore cannot reproduce realistic temperature or pressure conditions, leading to fluctuations which do not mimic any laboratory setup.

#### C.2.4.2 Canonical (NVT) Ensemble

In the canonical ensemble, the closed system of constant volume is now coupled to a heat bath in order to maintain a constant temperature. This means that the energy of the system is now allowed to fluctuate with time evolution. The way to achieve this constant temperature behaviour is by the use of a thermostat. The probability of finding a particular microstate in the canonical ensemble is dictated

by the following partition function:

$$Q_{NVT}(N,V,T) = \sum_i e^{-E_i/k_BT}$$

which is related to the Helmholtz free energy by the relation:

$$A = -k_BT \ln Q_{NVT}$$

NVT simulations are widely used to equilibrate systems before proceeding to NPT simulations, and are also often used for production simulations for systems where the volume is not expected to fluctuate significantly.

#### C.2.4.3 Isothermal-Isobaric (NPT) Ensemble

In the NPT ensemble, the system must now also be coupled to a barostat, which ensures a constant pressure by rescaling the box size during the course of the simulation. Its partition function is:

$$Q_{NPT}(N,P,T) = \sum_j \sum_i e^{-(E_i(V_j)+PV_j)/k_BT}$$

which sums over all possible microstates in all possible volumes, and is related to the Gibbs free energy $G$ via:

$$G = -k_BT \ln Q_{NPT}$$

The NPT ensemble is the standard ensemble used in production simulations of systems with fluctuating volumes, which includes most systems in solution.

### C.2.5 Thermostats

Thermostats are computational algorithms which control the kinetic energy of the system by changing the velocity of the particles at each time step, and thus control the temperature. This ensures that the system is exploring the relevant range of configurations for its thermal state.

There are different types of thermostats with different levels of complexity. Their differences lie in how they couple the motion of the particles in the simulation to the "heat bath" with which the heat is exchanged. Three common types of thermostats used in simulations, in order of their level of complexity are:

- Velocity rescaling thermostats, such as the Berendsen thermostat, which uses the equation:[74]

$$\frac{dT}{dt} = \frac{T_0 - T}{\tau}$$

  where the current temperature $T$ is relaxed to the target temperature $T_0$ in an exponential fashion according to some time parameter $\tau$. This is done by rescaling the forces acting on the atoms in the following Newton's equation of motion:

$$m_i \ddot{r}_i = -\frac{\partial U(\mathbf{R}^N)}{\partial r_i} - \frac{1}{2\tau_T}\left(\frac{T}{T(t)} - 1\right) m_i \dot{r}_i$$

  where $m_i \ddot{r}_i$ represents the thermostat-adjusted force acting on particle $i$, $-\frac{\partial U(\mathbf{R}^N)}{\partial r_i}$ is the force calculated to be acting on the particle by the force field only, and $-\frac{1}{2\tau_T}\left(\frac{T}{T(t)} - 1\right) m_i \dot{r}_i$ is the thermostat correction term. We can see that this term depends on the fraction $\frac{T}{T(t)}$. While the Berendsen thermostat has been shown to not properly reproduce the canonical (NVT) ensemble, it does couple the system to the heat bath very strongly, making it appropriate for equilibrating systems to the target temperature before initialising production simulation runs. An interesting artefact of the Berendsen thermostat is

that it can result in the "flying ice cube" effect, where energy is funneled from higher-energy vibrational degrees of motion to translational degrees of motion, resulting in a so-called flying ice cube.[75, 76]

- Stochastic thermostats act slightly differently, in that they add a random component to the velocity rescaling. Rescaling all velocities equally at each step as in the Berendsen thermostat does not actually correspond to the physical reality of temperature coupling. Kinetic energy redistribution does not simply correlate to the difference in temperature, as collisions are not perfect, and therefore the temperature coupling cannot be a simple correlation either. Examples of thermostats which add a stochastic component to account for these imperfect collisions and correctly sample the canonical ensemble include the Andersen and the V-rescale thermostats.[38, 77]

- A so-called extended system thermostat, the Noose-Hoover thermostat extends the system's dynamics to the heat bath as well.[78] The heat bath particles now have their own mass, and the equations of motion for the system now include extra degrees of freedom. While this thermostat samples the canonical ensemble well and minimises disruptive corrections to the velocity, it can also lead to temperature oscillations if improperly tuned, making it an overall more complex choice of thermostat.

### C.2.6 Barostats

If the thermostat adjusts the kinetic energy and thus the temperature of the system by rescaling velocity, then the barostat makes use of the thermodynamic equation for pressure:

$$P = -\left(\frac{\partial A}{\partial V}\right)_{T,N}$$

which tells us that the pressure will vary with volume. Barostats therefore adjust the pressure by varying the volume of the simulation box interacting with a pressure reservoir, analogously to how a heat bath is used to control the temperature. Common types of barostats include:

- The Berendsen barostat, which functions in almost exactly the same way as the Berendsen thermostat, and adjusts the pressure in accordance with the deviation of the current pressure $P_{current}$ from the target pressure $P_{target}$ at a rate proportional to the relaxation parameter $\tau_P$:

$$\frac{dP}{dt} = \frac{P_{\text{target}} - P_{\text{current}}}{\tau_P}$$

  It suffers from the same drawbacks in correctly sampling the isobaric-isothermal (NPT) ensemble as the Berendsen thermostat does in reproducing the NVT ensemble, which renders it most suitable for equilibration runs, and reaching the target pressure quickly.[74]

- The C-rescale barostat can be thought of as a stochastically-rescaling version of the Berendsen barostat, analogously to how the stachastically-rescaling thermostats (Andersen and V-rescale) improve the Berendsen thermostat. This avoids the suppresion of volume fluctuations and results in a more correct sampling of the ensemble.[39]

- The Parinello-Rahman barostat also allows the box vectors to change anisotropically, allowing for the box to change its shape.[79] The equations of motion contain an additional term when using this thermostat. The force on particle $i$ now also depends on a "friction-like" term which corrects the forces acting

on the system as it is coupling to the box with a changing shape:

$$m_i\ddot{\mathbf{r}}_i = -\frac{\partial U}{\partial \mathbf{r}_i} - W_i\dot{\mathbf{r}}_i$$

where the forces calculated from the force field terms -$\frac{\partial U}{\partial \mathbf{r}_i}$ are corrected by the correction term $W_i\dot{\mathbf{r}}_i$. This thermostat is useful when simulating anisotropic systems, for example membrane bilayers, where the forces acting in the plane of the bilayer differ from those acting perpendicularly to it differ.

## C.3 Machine Learning Potentials Background

Machine learning potentials are analytical functions of the potential energy surface. [80] This means that they map atomic positions of a molecule to the molecule's energy, and that the derivative of these positions represents the forces acting on the molecule in that position. They represent a paradigm shift, as they are computationally much more efficient than *ab-initio* dynamics, and yet with careful training of the machine learning algorithm they are able to achieve similarly high accuracies. This introduction on machine learning (interatomic) potentials, or MLIPs, will be split up into a brief history of their development across different generations, an introduction of their underlying machine learning (ML) architectures, the environment descriptors used for descriptions of the atomic surroundings, and the strategies used for training the potentials to fit energies and forces to the descriptors. But what makes MLIPs so revolutionary in computational chemistry?

The difference between MLIPs and other methods routinely used in the calculations of the energies and propagations of dynamics of molecular systems is that MLIPs themselves do not contain any physically-derived descriptions of a system's energy and forces. DFT, though a simplification of the complex Schrödinger

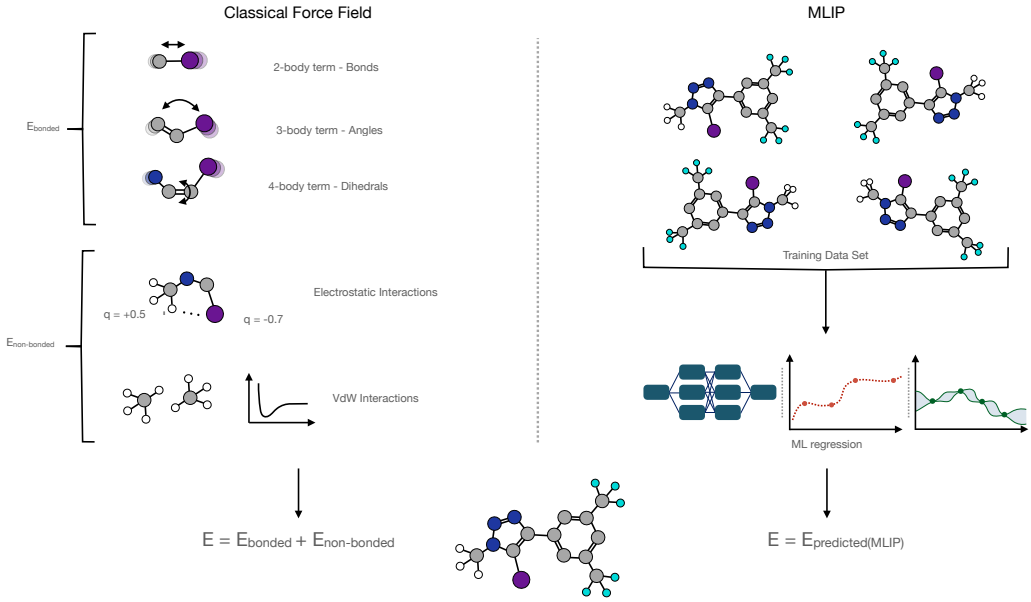



**Figure C.7:** Evaluation of an example geometry (bottom) during a dynamics simulation by a classical force field (left) and a machine learning potential (right).

equation, directly links a physically meaningful quantity—the electron density—to the molecule's energy. The iterative solution of the system's energy, while yielding highly accurate results for simple organic systems, represents a significant computational cost, and therefore precludes lengthy dynamic simulations of large, solvated systems on small computational clusters. Molecular mechanics operate on a heuristic combination of Newtonian, Coulumbic and Van der Waals terms, and therefore represent a much cheaper way of evaluating forces and energies at every step of a dynamics simulation, but they sacrifice a lot of accuracy, neglect quantum effects, and generally don't allow for bond-breaking and -making (vide supra for descriptions of both DFT and MM theory, and Figure C.7 for a comparison of MLIP and MM evaluation steps).

In contrast, MLIPs fit entirely abstract functions to reference data - usually high-level DFT single point energies and forces-to try and reproduce the underlying

function. They are entirely data-driven and free from explicit physical descriptors of the system. In their limit of accuracy, they are theoretically able to reproduce the DFT-level PES, while the evaluation cost of the potential energy at each step of the simulation is much closer to the cost of classical (MM) force fields.

### C.3.1 Development of Generations of Machine Learning Potentials

Having discussed the unique position of MLIPs amongst the methods that can be used to evaluate a molecular system's energetics, we must now look at how the methods have evolved over the period of the last 30 years. The formalism we will use is based on Behler's 2021 review "Four Generations of High-Dimensional Neural Network Potentials", which provides an excellent in-depth overview of the field.[80]

The first attempts of using neural networks (NNs) to represent potential energy surfaces, termed first-generation MLIPs, were revolutionary but often limited in scope. They relied on feed-forward neural networks to fit the energies of simple systems, with a general structure of one hidden layer, and a separate bias node acting as a computationally inexpensive surrogate for an activation function (C.8, top). The degrees of freedom being mapped to energies had to be very carefully considered because of several limiting factors in the construction of these early potentials, namely:

- A fixed number of input nodes meant that the number of coordinates on which the system's energy depended had to be kept constant, or else the NN would be unable to compute the energies.
- The coordinates had to be kept invariant to rotation and translation, as these have no physical impact on the energy of a physical system in reality. In

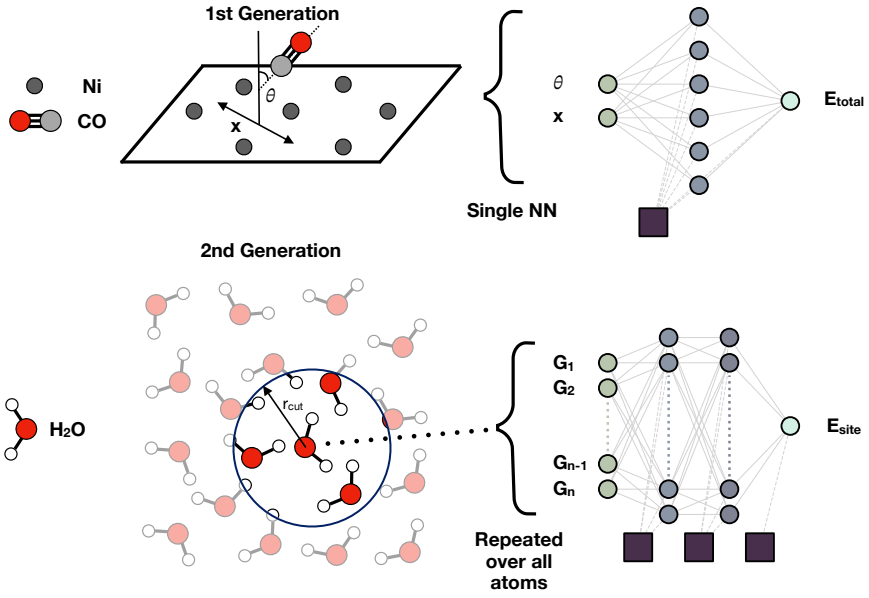



**Figure C.8:** Example first and second generation MLIPs both using a neural network to predict the energy of the system. The scheme depicts the differences in how atomic information is input into the neural network. The well-known example 1st generation potential by Blank et al [81] shows CO interacting with a Ni(111)surface with greatly simplified input coordinates. A second generation potential using the Behler-Parinello ACSFs studying water would instead input symmetry functions G for a neural network at each atomic site in order to construct the total energy $E_{total}$ from site energies $E_{site}$.

practice, feeding enough configurations of all possible rotational and translational permutations should be able to teach a flexible enough neural network the rotational and translational invariance of a whole system's energies and forces, but this is not a data-efficient way of training neural network potentials.

The use of internal coordinates was one of the ways of solving both of these problems in some early attempts NN potentials. Another was to limit the number of coordinates to a very small number of atoms involved in a reaction. Both of these were used in one of the most important early works by Blank et al. They first studied the diffusion of CO through Ni(111), where the angle and distance of the CO molecule to the Ni surface were the degrees of freedom (C.8, top). In their study of the interaction of $H_2$ with a Si surface, they instead chose to focus on the three spatial coordinates of the two hydrogen atoms and the two silicon atoms involved in the adsorption, deciding to model a very small region of reaction space. These examples make it clear that scaleable, general, and chemically diverse simulations would not be possible with such architectures.

It was not until 2007, with the publication of the Behler-Parinello seminal work and introduction of the locality approximation, that larger systems could begin to be tackled. This approximation decomposed the coordinate space of the atomic species being studied into the local environments surrounding each individual atom (C.9, Environment Descriptors) - and thus also decomposed the total energy into a sum of site energies:

$$E = \sum_i E_i(G_i)$$

where the total system energy $E$ is then obtained by summing over all of the

site energies $E_i$, which depend on the symmetry functions $G_i$, which describe the atomic environment within a cut-off radius $R_{cutoff}$. The introduction of atom-centered symmetry functions (ACSFs) meant that the system could now be decomposed into many small pieces, each of which would be evaluated using an element-specific neural network, instead of evaluating the whole system with a single network. Because each atom had its own energetic contribution, the order of the atoms in the system no longer mattered - giving permutation invariance. The clever construction of the symmetry functions from a radial and angular component also made them rotationally and translationally invariant (*vide infra* for discussions of atomic descriptors). With this breakthrough, three challenges associated with first generation potentials could thus be overcome:

- Permutational, rotational and translational invariance
- Simulating systems of variable sizes
- Avoiding the exponential scaling of model fitting with system size

In combination, this meant that these potentials could now be trained on atomic configurations of varying numbers of atoms. For example, training on small model portions of atomic interactions could result in a potential that could simulate a much larger system in the end and afforded the opportunity to develop novel training strategies (*vide infra* for Data Generation). This allowed systems of thousands of atoms to be simulated, including initial case studies of silicon in both high-temperature and high-pressure regimes by Behler and Parinello, which achieved excellent agreement with the underlying DFT and the experimental data.[82] Second generation potentials continue to thrive today, and have become increibly varied in the underlying architectures used to fit the PES, beyond just neural networks. This has allowed the study of bulk materials, surface reactions, phase transitions, condensed phases, and more recently an increasing number of organic

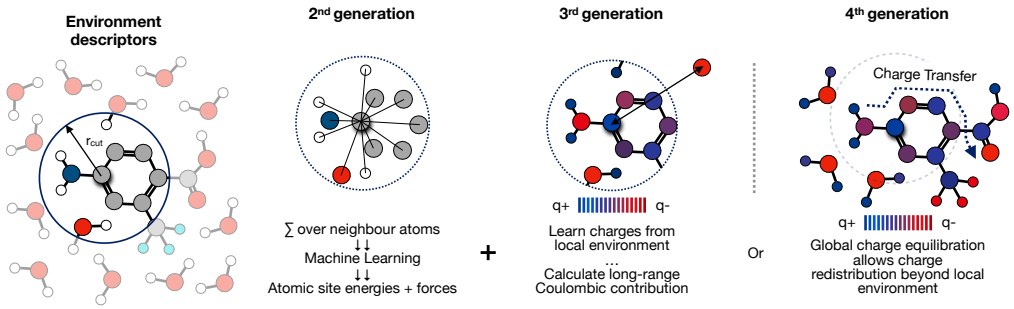



**Figure C.9:** Comparison of the second, third and fourth generations of MLIPs as defined by Behler, schematically depicted on a cartoon example system. The charges represented are fictitious but show how atoms outside the cutoff sphere and their contributions to coulubmic contributions and charge transfer processes are accounted for in the third and fourth generation potentials, respectively.

systems. Some of the most common architectures, including the graph neural network-based potentials used in this thesis, will be discussed in the following section.

An inherent problem of the locality assumption and calculating the energy only from contributions within the local cutoff sphere is that this ignores any long-range contributions. This is particularly a problem with electrostatics, and needed to be fixed for correctly describing ionic systems and polar molecules. In third-generation potentials, the charges are learned from the local environment using a separate set of neural networks (C.9, 3rd generation), and their contributions to the energies and forces of the system are then summed using standard schemes, such as the Ewald sum described in the MM theory section (*vide supra*).[80] This also allows for the study of systems with short-range charge transfer effects.

Finally, the fourth generation of MLIPs allows for the study of systems where charge transfer effects extend beyond the radial cutoff of the symmetry functions. This can happen in metalic solids, systems undergoing photochemical or redox reactions, or highly conjugated organic systems.[80] One method ensuring that charge transfer is accounted for globally across the system is to employ a global

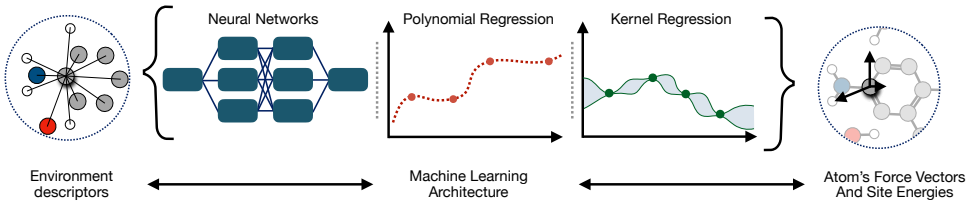



**Figure C.10:** Different machine learning architectures have been used in order to fit atomic energies to environment descriptor inputs, most commonly including Neural Networks, linear and polynomial regression, as well as kernel regression methods. In each case, the ML architecture must output the energies and forces acting on the atoms.

charge redistribution scheme, where the charges are self-consistently determined across the entire system (C.9, 4th generation).

### C.3.2 Machine Learning Architectures

Following the first implementation of atom-centered symmetry functions, the stage was set for innovations both in the space of novel atomic environment descriptors (see next section), as well as novel machine learning frameworks to use for prediction, which we discuss here. A variety of architectures could now be employed to regress the energy and forces onto the atomic environments (C.10), and several groups took on the task of developing approaches for doing so. We discuss the basic architectures of three different approaches - the original neural networks, polynomial regression, and finally Gaussian kernel regression.

Neural networks are known as universal function approximators, as it has been shown that they are theoretically able to approximate any continuous function.[83] Their task is to take an input vector, in this case an atom-centered symmetry function, and pass it through a series of neurons, made up of so-called hidden nodes, before outputting the desired quantities, here the energies and forces.

In each layer, all neurons are connected to all of the neurons in the following layer.

This connection to the neuron $z_j^{(l)}$ in layer $l$ works in the following way: the scalar of the output of each neuron $a_k$ in a layer $l-1$ is multiplied by a weight connecting the neurons j and k $w_{jk}^{(l)}$, summed together and added to a bias value $b_j^{(l)}$:[80]

$$z_j^{(l)} = \sum_k w_{jk}^{(l)} a_k^{(l-1)} + b_j^{(l)}$$

This value is then passed through a non-linear activation function $f$, which allows for modelling non-linear dependencies between different input variables:

$$a_j^{(l)} = f(z_j^{(l)})$$

The task of the training process is to then adjust the weights and biases incrementally in each pass through the data, until the predictions are satisfactory.[84] This is done by minimising a loss function $L$, which gives its dependence on the weights $\frac{\partial L}{\partial w_{jk}^{(l)}}$ and on the biases $\frac{\partial L}{\partial b_j^{(l)}}$, which then allows for adjusting $w_{jk}$ and $b_j$ to find the minimum. A stochastic gradient descent algorithm is usually employed to prioritise finding global rather than local minima. The whole process of finding the loss function via a forward pass using each iteration's current weights and biases, before performing a backward pass to find the new derivatives of the loss function, which are used to update the weights and biases of the next iteration, is called back-propagation. This process is possible because of the chain rule, where the derivative can be back-propagated from the output layer, all the way to each weight and bias in the neural network.

Whereas the ACSF descriptors which are fed into NN-based MLIPs contain vectors of the atomic environment composed of relatively simple scalars, the information fed into polynomial regressors such as ACE is much more dense.[85, 86] Here,

the atomic basis vector is constructed as a much more complex product of 2-, 3-, and 4-body terms. This means that the feature space of the descriptors themselves is much larger and more complex, and includes very detailed information about the angular and radial interdependence of the atomic environment. This leaves the learning task to a relatively more straightforward linear regression algorithm.

The construction of the atomic basis $A_{(nlm)}(r_{ij})$ is similar than with ACSFs, with radial functions and spherical harmonics projected onto a local atomic density. These are then combined into many-body terms, called the ACE basis functions $B_\nu(\mathcal{E}_i)$, which are then related to the energy of the atom by the equation:

$$E_i = \sum_\nu c_\nu B_\nu(\mathcal{E}_i)$$

The loss function can then be defined using a matrix $X$, which contains the ACE atomic descriptors in each row, and $y$ is the vector of reference energies or forces. The regression task is then to find coefficients $c$ which minimise sum of squared errors:

$$\min_{\mathbf{c}} ||\mathbf{Xc} - \mathbf{y}||^2 + \lambda||\mathbf{c}||^2$$

The Gaussian approximation potential, which was developed soon after the first Behler-Parinello neural networks, utilise Gaussian process regression, which is a Bayesian approach which also contains uncertainty estimations of its predictions.[60, 61] This framework assumes that the training data follows a multivariate Gaussian distribution, and uses a so-called "kernel trick", where a kernel function $k(x, x')$ in order to quantify the similarity between two atomic environments $x$ and $x'$. The input $x$ here is again a function describing the atomic environment, called the smooth overlap of atomic positions (SOAP) kernel. The kernel function is then

tasked with measuring how similar these two atomic environments are - the higher the similarity of the environment, the more similar the energy contribution.

### C.3.3 Local Environment Descriptors

As mentioned in the description of the evolution of the machine learning potentials, what truly enabled the development of efficient approximators of energy and forces was the development of good environment descriptors.[87] They all share the same task of transforming the atomic coordinates into translation-, rotation-, and permutation-invariant scalars, but differ somewhat in they way they achieve this, due to differences in the ML frameworks which regress target properties onto them. Here we will briefly compare the angular and radial functions which compose the ACSF, SOAP and ACE descriptors (C.11) of the NN, GAP and ACE potentials described in the previous section.

In ACSF descriptors, the angular terms have the mathematical form $2^{1-\zeta}(1 + \lambda cos\theta)^{\zeta}$, where the term $\lambda$ can be used to specify the position of the maximum of the function, and $\zeta$ is used to tune the steepness of the function.[80] By placing several functions with various $\lambda$ and $\zeta$ combinations, the different terms that sum over neighbours will have different sensitivities to the angles between atoms, and will capture angles of different sizes. The radial functions are Gaussians placed at different distances, multiplied by a cutoff function, which ensures that the contribution goes to 0 at the radial cutoff distance. For each atom, a fixed-length vector will be produced by concatenating all of the values of the angular and radial terms. Careful parameterisation of the various terms that dictate the nature of the radial and angular basis terms is necessary for them to capture the atomic environment in sufficient detail.

Unlike ACSFs, SOAP and ACE both use spherical harmonics $Y_{lm}(\theta, \phi)$ as the

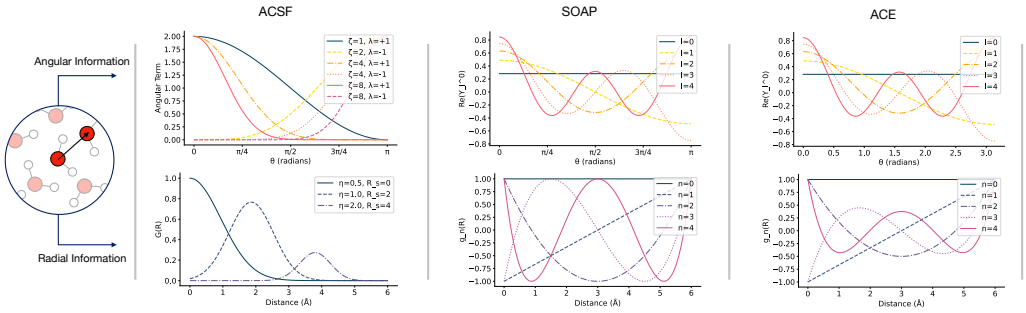



**Figure C.11:** Comparison of the mathematical functions encoding the radial and angular parts of the atomic environment descriptors in the case of the ACSF, SOAP and ACE descriptors.

angular components of the atomic environment descriptors, whereas for their radial functions they use orthogonal polynomials, which can be spherical Bessel functions or Chebyshev polynomials, combined with some cutoff functions.[60, 61, 85] The differences in their collation for feeding the descriptors into the respective ML architecture comes from how these functions are combined. In both, the functions are projected onto the density of the atoms in the neighbouring environment. In SOAP, this is the smoothed atomic density, which consists of 3D Gaussians placed at the center of atomic positions, whereas in ACE these are Dirac delta functions, which give sharp points at atomic positions. The values of these projections are then integrated to find the contribution of each radial and angular function's projection onto the atomic density. In SOAP, the vector of all of these scalars, also called the SOAP power spectrum, is then used as the input for the Gaussian kernel regression, whereas the atomic basis $A$ of ACE is combined into higher-order terms in a rotationally-invariant fashion using Clebsch-Gordan coefficients to give the many-body basis $B$, which is then used as an input for the regression.

### C.3.4 Graph Neural Networks

In this section we will discuss the structure of the graph neural network (GNN) MLIP called MACE, which is in many ways an extension of the ACE architecture adapted for use with GNNs.[59] GNNs are particularly suited for representing chemical systems, as molecules have a graph-like structure. Graphs are a way of describing objects and the connections between them - the objects being called nodes and the connections being called edges.[59] Coincidentally, nodes connected by edges are also how we think of molecules - with the atoms (nodes) connected by bonds (edges). The nodes and edges are attributed features which describe their characteristics - for example, the atoms may have a feature vector describing their atomic number, hybridisation and ionisation energy, whereas the bonds may have a feature vector denoting their hybridisation and equilibrium distance. These features are then passed along the connections in the GNN in a so-called message passing step in each layer of the network to give transformed vectors, which are then used for property prediction after the message passing layers in a final, readout layer.

In more complicated architectures, such as MACE, the descriptors become wholly flexible, and are optimised or "learned" during the training process, making the structure of the network incredibly flexible. To illustrate this, we will briefly discuss its graph neural network structure more broadly.

Regular neural network potentials will decompose a molecular system into individual atomic sites and construct fixed atomic descriptors, which are then evaluated using a NN. There is no cross-talk between atoms during the process of training the network, and all of the information about each atom's neighbours is purely captured in the construction of the descriptors. A big conceptual difference in GNN-based MLIPs is that the individual atomic environment descriptors are passed between

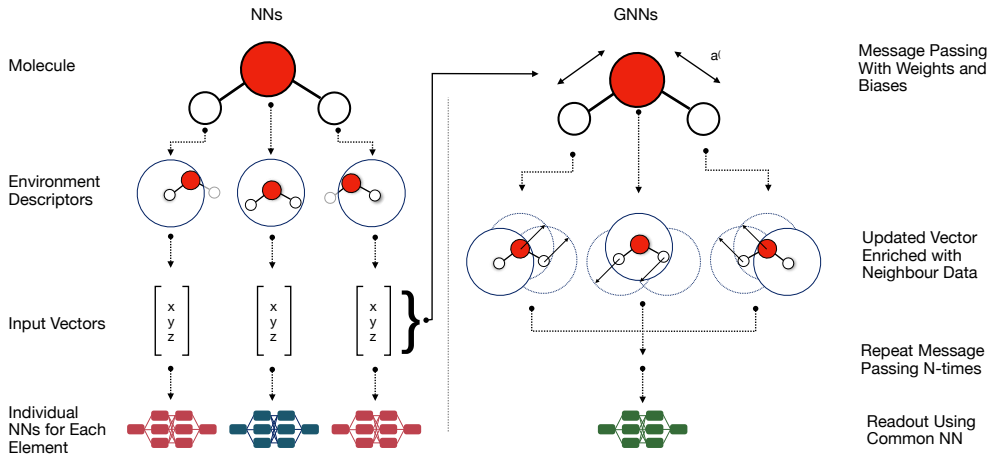



**Figure C.12:** A comparison of how classical NNs (left) and GNNs (right) handle the flow of information.

neighbours (C.12), which gives each node (atom) a much richer description of its immediate environment after each message passing layer. Let us break this process down on the example of the water molecule in C.12:

- In the initial layer, each atom possesses an initial node feature $h^{(0)}$, which is simply a vector embedding of the atomic number $z_j$ of the element. Embeddings of simple features like atomic numbers into vectors are: (1) necessary for their further manipulation by the network and (2) can be made learnable as well, meaning that the network can learn the optimal vector for each atom through the training of the network.
- At each layer $t$ of the MACE GNN, a message is constructed from the features of the nodes and the edges up to that point, to give the new feature $h^{(t+1)}$ according to the equation:

$$\mathbf{h}_{i,kLM}^{(t+1)} = \sum_{k'} W_{kL,k'}^{(t)} \mathbf{m}_{i,k'LM}^{(t)} + \sum_{k'} W_{z_i kL,k'}^{(t)} \mathbf{h}_{i,k'LM}^{(t)}$$

where $W^{(t)}_{kL,k'}$ and $W^{(t)}_{z_i kL,k'}$ are learnable weight matrices for the message and the previous node's features. The most complex step in the architecture is the construction of the message $\mathbf{m}^{(t)}_{i,k'LM}$. Here, 2-body $A^{(t)}$ features are composed by combining radial functions $R^{(t)}_{kl_1l_2l_3}(r_{ji})$ and spherical harmonics $Y^{m_1}_{L_1}$ (similar to those in ACE) with the incoming neighbouring node's features $h^{(t)}_j$. The previous layer's node features are first multiplied by learnable weight matrix, and then summed over $k$ channels. This modified feature is then multiplied by the radial and spherical harmonic terms. The whole product tensor is then multiplied by Clebsch-Gordan coefficients in order to maintain equivariance, and the $A^{(t)}$ features are then given as:

$$A^{(t)}_{i,kL_3M_3} = \sum_{l_1m_1,l_2m_2} C^{L_3m_3}_{l_1m_1,l_2m_2} \sum_{j\in\mathcal{N}(i)} R^{(t)}_{kl_1l_2l_3}(r_{ji}) Y^{m_1}_{L_1}(\hat{\mathbf{r}}_{ji}) \sum_{\tilde{k}} W^{(t)}_{k\tilde{k}l_2} h^{(t)}_{j,\tilde{k}l_2m_2}$$

These are then expanded into $B^{(t)}$ features by taking products of the $A^{(t)}$ features up to some correlation order $\nu$, before linearly combining them and passing it to the next layer.

- In the readout phase of the MACE GNN, the energy is predicted in a hierarchical readout, where the energy is a sum of contributions from each message-passing layer $t$:

$$E_i = E^{(0)}_i + E^{(1)}_i + \cdots + E^{(T)}_i$$

This can be interpreted as contributions from various levels of complexity. The zeroth layer is an element-dependent contribution, the first message passing layer adds information from the receptive fields of the immediate neighbours of the atom, and the second layer adds information from the neighbours of those neighbours.

As can be inferred from the presence of weight matrices $W$ in the message aggregation, feature update steps and readout steps, the GNN can be thought of as being made up of several smaller NNs with learnable weights and biases. The major difference is that GNNs are structure-aware and therefore better at interpreting relationships between input entities (atoms). They leverage adaptive descriptors and share information between atoms during the training process in order to best adapt themselves to the prediction task. Finally, they are able to do all of this while maintaining the ability for the size of the graph to change - as the size of the input vectors and the number of learnable weights and biases remain fixed hyperparameters.

### C.3.5 Data Generation Strategies

In order to train an MLIP to be able to simulate a system in a region of chemical space, an adequate number of training data points covering this space needs to be fed to the model during training.[88, 89] Several key properties need to be fulfilled for this process to be efficient. For example, the data needs to be diverse enough for the MLIP to be able to interpolate between it for the vast majority of the configurations that will be explored during the simulation, else the models may begin to predict erroneous forces, leading to instability. Secondly, the data should cover the PES with a sufficient sparseness to avoid feeding too many training data points to the model, thus wasting computational time in generating the data and slowing the training process down. It should also not be too sparse, as interpolation issues may then arise. The data should also contain both high- and low-energy configurations (a term used generally in the MLIP field to denote a group of atoms with co-ordinates and here incorporating both conformations and configurations in the organic chemistry sense). This ensures that barriers are predicted accurately as well as equilibrium positions. With this in mind, there are several different

approaches to generating training data, which are discussed briefly here.

- The first method of generating training data for MLIPs is to run ab initio molecular dynamics (AIMD). Here, the forces at each time step are calculated using DFT. This makes AIMD a very expensive computational method, and yet it is the method which generates the most physically realistic structures, which naturally explore the phase space which the MLIP should also be able to explore.
- Minimum energy structures of systems of interest can also be perturbed randomly to generate diverse static structures, which are then evaluated using DFT. The hope is that these perturbed structures will explore the PES around the minimum well sufficiently to propagate dynamics. This is a much more computationally efficient option than performing AIMD, but may introduce structures into training which are far from equilibrium and therefore irrelevant for the model's performance in the simulation, and which may even harm the accuracy of the MLIP. Additional selection criteria for structures' inclusion into the training set can be used to ensure only relevant data are added.
- Active learning (C.13) is one of the more popular methods. It adopts an iterative approach, where the model is first trained on a small number of configurations, following which dynamics are run to identify novel configurations which should be added to the dataset, based on some selection criteria. The model is then retrained, and the loop is repeated until no new novel configurations fulfil the selection criteria, the simulations run stably for a long time, or some accuracy metric of the MLIP on the validation dataset is fulfilled. This approach is highly efficient, as it explores phase space using the MLIP model itself, reducing the cost compared to AIMD

significantly, and only performs DFT calculations on structures which are deemed likely to improve the training dataset. The added complications are that one must carefully choose the metrics by which new structures are added and the strategy by which the phase space is sampled.

- Existing databases containing large amounts of pre-computed DFT training data can also be used to train MLIPs. This avoids the need to calculate high-level reference structures and saves a lot of computational time, but the sizes of these databases mean that the training cost is significant. In recent years, so-called foundation models, which should be able to simulate a very large breadth of different chemical elements comprising highly diverse systems at reasonable accuracy, have been developed using such databases. These can then be refined, or fine-tuned, using novel training data points, which can also be generated using the foundation model. One problem with foundation models trained on large databases is that the databases usually contain equilibrium structures, which means that the models trained on them are unlikely to be able to simulate barrier crossings, and therefore cannot predict chemical reactions accurately.

In reality, these approaches are often combined in order to generate an accurate MLIP. For example, one might start with a small amount of *ab-initio* data, which is expensive to generate, and enrich it with some randomly perturbed equilibrium structures to train an initial potential. This can then be used to start an active learning loop, which adds data in an iterative fashion and provides structures from areas of the PES not yet covered to the training set (Figure C.13).

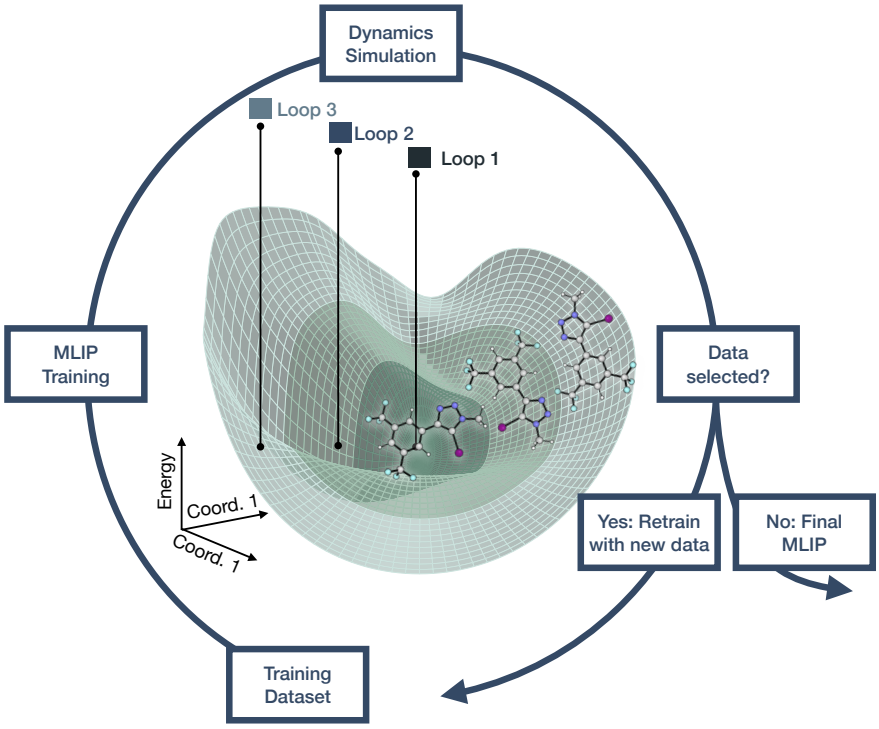



**Figure C.13:** Schematic depiction of on-the-fly active learning as implemented in mlptrain [57]. In each loop, a machine learning potential is trained on training data before running dynamics until: a selector criterion is fulfilled and new data is added, starting the new loop; or no new data are added and a final potential is output, thus exiting the active learnign loop.

# Appendix bibliography


(1) Brassard, C. J.; Zhang, X.; Brewer, C. R.; Liu, P.; Clark, R. J.; Zhu, L. Cu(II)-Catalyzed Oxidative Formation of 5,5-Bistriazoles. *The Journal of Organic Chemistry* **2016**, *81*, 12091–12105.

(2) Charoenpattarapreeda, J.; Tan, Y. S.; Iegre, J.; Walsh, S. J.; Fowler, E.; Eapen, R. S.; Wu, Y.; Sore, H. F.; Verma, C. S.; Itzhaki, L.; Spring, D. R. Targeted covalent inhibitors of MDM2 using electrophile-bearing stapled peptides. *Chem. Commun.* **2019**, *55*, 7914–7917.

(3) Hein, J. E.; Tripp, J. C.; Krasnova, L. B.; Sharpless, K. B.; Fokin, V. V. Copper(I)-Catalyzed Cycloaddition of Organic Azides and 1-Iodoalkynes. *Angew. Chem. Int. Ed.* **2009**, *48*, 8018–8021.

(4) Bickerton, L. E.; Docker, A.; Sterling, A. J.; Kuhn, H.; Duarte, F.; Beer, P. D.; Langton, M. J. Highly Active Halogen Bonding and Chalcogen Bonding Chloride Transporters with Non-Protonophoric Activity. *Chem. Eur. J.* **2021**, *27*, 11738–11745.

(5) Neenan, T. X.; Whitesides, G. M. Synthesis of high carbon materials from acetylenic precursors. Preparation of aromatic monomers bearing multiple ethynyl groups. *The Journal of Organic Chemistry* **1988**, *53*, 2489–2496.

(6) Bickerton, L. E.; Sterling, A. J.; Beer, P. D.; Duarte, F.; Langton, M. J. Transmembrane anion transport mediated by halogen bonding and hydrogen bonding triazole anionophores. *Chem. Sci.* **2020**, *11*, 4722–4729.

(7) Luca, V.; Tejada, J. J.; Vega, D.; Arrachart, G.; Rey, C. Zirconium(IV)–Benzene Phosphonate Coordination Polymers: Lanthanide and Actinide Extraction and Thermal Properties. *Inorganic Chemistry* **2016**, *55*, 7928–7943.

(8) Sun, Y.; Li, S.; Zhou, Z.; Saha, M. L.; Datta, S.; Zhang, M.; Yan, X.; Tian, D.; Wang, H.; Wang, L.; Li, X.; Liu, M.; Li, H.; Stang, P. J. Alanine-Based

Chiral Metallogels via Supramolecular Coordination Complex Platforms: Metallogelation Induced Chirality Transfer. *J. Am. Chem. Soc.* **2018**, *140*, 3257–3263.

(9) BindFit v0.5 | Supramolecular.

(10) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminform.* **2012**, *4*, 17.

(11) 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.* **2004**, *25*, 1157–1174.

(12) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(13) Grimme, S. Exploration of Chemical Compound, Conformer, and Reaction Space with Meta-Dynamics Simulations Based on Tight-Binding Quantum Chemical Calculations. *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

(14) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78.

(15) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1606.

(16) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615.

(17) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(18) Kozuch, S.; Martin, J. M. L. Halogen Bonds: Benchmarks and Theoretical Analysis. *J. Chem. Theory Comput.* **2013**, *9*, 1918–1931.

(19) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal Solvation Model Based on Solute Electron Density and on a Continuum Model of the Solvent Defined by the Bulk Dielectric Constant and Atomic Surface Tensions. *J. Phys. Chem. B* **2009**, *113*, 6378–6396.

(20) Neese, F.; Wennmohs, F.; Hansen, A.; Becker, U. Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. A 'chain-of-spheres' algorithm for the Hartree–Fock exchange. *Chem. Phys.* **2009**, *356*, 98–109.

(21) Young, T. duartegroup/otherm: Major symmetry improvements, 2020.

(22) Bickelhaupt, F. M.; Houk, K. N. Analyzing Reaction Rates with the Distortion/Interaction-Activation Strain Model. *Angew. Chem. Int. Ed.* **2017**, *56*, 10070–10086.

(23) Glendening, E. D.; Landis, C. R.; Weinhold, F. *NBO 7.0* : New vistas in localized and delocalized chemical bonding theory. *J. Comput. Chem.* **2019**, *40*, 2234–2241.

(24) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts,

R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09 Revision D.01, 2009.

(25) Singh, U. C.; Kollman, P. A. An approach to computing electrostatic charges for molecules. *J. Comput. Chem.* **1984**, *5*, 129–145.

(26) Cieplak, P.; Cornell, W. D.; Bayly, C.; Kollman, P. A. Application of the multimolecule and multiconformational RESP methodology to biopolymers: Charge derivation for DNA, RNA, and proteins. *J. Comput. Chem.* **1995**, *16*, 1357–1377.

(27) Case, D. A.; Aktulga, H. M.; Belfon, K.; Cerutti, D. S.; Cisneros, G. A.; Cruzeiro, V. W. D.; Forouzesh, N.; Giese, T. J.; Götz, A. W.; Gohlke, H.; Izadi, S.; Kasavajhala, K.; Kaymak, M. C.; King, E.; Kurtzman, T.; Lee, T.-S.; Li, P.; Liu, J.; Luchko, T.; Luo, R.; Manathunga, M.; Machado, M. R.; Nguyen, H. M.; O'Hearn, K. A.; Onufriev, A. V.; Pan, F.; Pantano, S.; Qi, R.; Rahnamoun, A.; Risheh, A.; Schott-Verdugo, S.; Shajan, A.; Swails, J.; Wang, J.; Wei, H.; Wu, X.; Wu, Y.; Zhang, S.; Zhao, S.; Zhu, Q.; Cheatham, T. E. I.; Roe, D. R.; Roitberg, A.; Simmerling, C.; York, D. M.; Nagan, M. C.; Merz, K. M. J. AmberTools. *J. Chem. Inf. Model.* **2023**, *63*, 6183–6191.

(28) Kolář, M.; Hobza, P. On Extension of the Current Biomolecular Empirical Force Field for the Description of Halogen Bonds. *J. Chem. Theory Comput.* **2012**, *8*, 1325–1333.

(29) Ibrahim, M. A. A.; Telb, E. M. Z. -Hole and Lone-Pair Hole Interactions in Chalcogen-Containing Complexes: A Comparative Study. *ACS Omega* **2020**, *5*, 21631–21640.

(30) Caleman, C.; van Maaren, P. J.; Hong, M.; Hub, J. S.; Costa, L. T.; van der Spoel, D. Force Field Benchmark of Organic Liquids: Density, Enthalpy of Vaporization, Heat Capacities, Surface Tension, Isothermal Compressibility, Volumetric Expansion Coefficient, and Dielectric Constant. *J. Chem. Theory Comput.* **2012**, *8*, 61–74.

(31) Sengupta, A.; Li, Z.; Song, L. F.; Li, P.; Merz, K. M. J. Parameterization of Monovalent Ions for the OPC3, OPC, TIP3P-FB, and TIP4P-FB Water Models. *J. Chem. Inf. Model.* **2021**, *61*, 869–880.

(32) Kim, Y. C.; Chaloux, B. L.; Rolison, D. R.; Johannes, M. D.; Sassin, M. B. Molecular dynamics study of hydroxide ion diffusion in polymer electrolytes. *Electrochemistry Communications* **2022**, *140*, 107334.

(33) qvist, J. Ion-water interaction potentials derived from free energy perturbation simulations. *The Journal of Physical Chemistry* **1990**, *94*, 8021–8024.

(34) Grote, F.; Lyubartsev, A. P. Optimization of Slipids Force Field Parameters Describing Headgroups of Phospholipids. *J. Phys. Chem. B* **2020**, *124*, 8784–8793.

(35) Jämbeck, J. P. M.; Lyubartsev, A. P. An Extension and Further Validation of an All-Atomistic Force Field for Biological Membranes. *J. Chem. Theory Comput.* **2012**, *8*, 2938–2948.

(36) Wang, L.-P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *The Journal of Physical Chemistry Letters* **2014**, *5*, 1885–1891.

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

(38) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(39) Bernetti, M.; Bussi, G. Pressure control using stochastic cell rescaling. *J. Chem. Phys.* **2020**, *153*, 114107.

(40) Hess, B.; Bekker, H.; Berendsen, H. J. C.; Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **1997**, *18*, 1463–1472.

(41) Lindahl; Abraham; Hess; Spoel, v. d. GROMACS 2021.3 Source code, 2021.

(42) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R. A.; Parrinello, M. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications* **2009**, *180*, 1961–1972.

(43) Hub, J. S.; de Groot, B. L.; van der Spoel, D. g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Theory Comput.* **2010**, *6*, 3713–3720.

(44) Bennett, C. H. Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics* **1976**, *22*, 245–268.

(45) Grimme, S. Improved second-order Møller-Plesset perturbation theory by separate scaling of parallel and antiparallel spin pair correlation energies. *J. Chem. Phys.* **2003**, *118*, 9095–9102.

(46) Grimme, S. Semiempirical hybrid density functional with perturbative second-order correlation. *J. Chem. Phys.* **2006**, *124*, 034108.

(47) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange-correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(48) Zhao, Y.; Truhlar, D. G. The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals. *Theor. Chem. Acc.* **2008**, *120*, 215–241.

(49) Zhao, Y.; Truhlar, D. G. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.* **2006**, *125*, 194101.

(50) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.

(51) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(52) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465.

(53) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122.

(54) Lin, Y.-S.; Li, G.-D.; Mao, S.-P.; Chai, J.-D. Long-Range Corrected Hybrid Density Functionals with Improved Dispersion Corrections. *J. Chem. Theory Comput.* **2013**, *9*, 263–272.

(55) Mardirossian, N.; Head-Gordon, M. $\omega$B97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **2016**, *144*, 214110.

(56) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995–2001.

(57) Young, T.; Johnston-Wood, T.; Zhang, H.; Duarte, F. Reaction dynamics of Diels-Alder reactions from machine learned potentials. *Phys. Chem. Chem. Phys.* **2022**.

(58) Young, T. A.; Silcock, J. J.; Sterling, A. J.; Duarte, F. autodE: Automated Calculation of Reaction Energy Profiles— Application to Organic and Organometallic Reactions. *Angew. Chem. Int. Ed.* **2021**, *60*, 4266–4274.

(59) Batatia, I.; Kovács, D. P.; Simm, G. N. C.; Ortner, C.; Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields, 2023.

(60) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* **2010**, *104*, 136403.

(61) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121*, 10073–10141.

(62) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Pu, Y.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF: Short-Range Transferable Machine Learning Force Fields for Organic Molecules. *J. Am. Chem. Soc.* **2025**, *147*, 17598–17611.

(63) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; Hermes, E. D.; Jennings, P. C.; Bjerre Jensen, P.; Kermode, J.; Kitchin, J. R.; Leonhard Kolsbjerg, E.; Kubal, J.; Kaasbjerg, K.; Lysgaard, S.; Bergmann Maronsson, J.; Maxson, T.; Olsen, T.; Pastewka, L.; Peterson,

A.; Rostgaard, C.; Schiøtz, J.; Schütt, O.; Strange, M.; Thygesen, K. S.; Vegge, T.; Vilhelmsen, L.; Walter, M.; Zeng, Z.; Jacobsen, K. W. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(64) Atkins, P.; Friedman, R., *Molecular quantum mechanics*, 5th; Oxford University Press: Oxford, 2011.

(65) Koch, W.; Holthausen, M. C., *A Chemist's Guide to Density Functional Theory*, 2nd; Wiley-VCH: Weinheim, 2001.

(66) Perdew, J. P.; Schmidt, K. Jacob's ladder of density functional approximations for the exchange-correlation energy. *AIP Conference Proceedings* **2001**, *577*, 1–20.

(67) Car, R. Fixing Jacob's ladder. *Nat. Chem.* **2016**, *8*, 820–821.

(68) Mardirossian, N.; Head-Gordon, M. Thirty years of density functional theory in computational chemistry: an overview and extensive assessment of 200 density functionals. *Molecular Physics* **2017**, *115*, 2315–2372.

(69) Perdew, J. P. Climbing the ladder of density functional approximations. *MRS Bulletin* **2013**, *38*, 743–750.

(70) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Physical Review Letters* **1996**, *77*, 3865–3868.

(71) Santamaria, R., *Molecular Dynamics*; Springer Nature Switzerland: Cham, 2023.

(72) Frenkel, D.; Smit, B., *Understanding molecular simulation: from algorithms to applications*, 2nd ed; Computational science series 1; Academic Press: San Diego, 2002.

(73) Atkins, P. W.; de Paula, J.; Keeler, J. J., *Atkins' physical chemistry*, Twelfth edition; Oxford University Press: Oxford, 2023.

(74) 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.* **1984**, *81*, 3684–3690.

(75) Harvey, S. C.; Tan, R. K.-Z.; Cheatham III, T. E. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J. Comput. Chem.* **1998**, *19*, 726–740.

(76) Braun, E.; Moosavi, S. M.; Smit, B. Anomalous Effects of Velocity Rescaling Algorithms: The Flying Ice Cube Effect Revisited. *J. Chem. Theory Comput.* **2018**, *14*, 5262–5272.

(77) Andersen, H. C. Molecular dynamics simulations at constant pressure and/or temperature. *J. Chem. Phys.* **1980**, *72*, 2384–2393.

(78) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* **1985**, *31*, 1695–1697.

(79) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **1981**, *52*, 7182–7190.

(80) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037–10072.

(81) Blank, T. B.; Brown, S. D.; Calhoun, A. W.; Doren, D. J. Neural network models of potential energy surfaces. *J. Chem. Phys.* **1995**, *103*, 4129–4137.

(82) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(83) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **1989**, *2*, 359–366.

(84) James, G.; Witten, D.; Hastie, T.; Tibshirani, R.; Taylor, J. E., *An introduction to statistical learning: with applications in Python*; Springer texts in statistics; Springer: Cham, 2023.

(85) Kovács, D. P.; Oord, C. v. d.; Kucera, J.; Allen, A. E. A.; Cole, D. J.; Ortner, C.; Csányi, G. Linear Atomic Cluster Expansion Force Fields for Organic Molecules: Beyond RMSE. *J. Chem. Theory Comput.* **2021**, *17*, 7696–7711.

(86) Dusson, G.; Bachmayr, M.; Csányi, G.; Drautz, R.; Etter, S.; van der Oord, C.; Ortner, C. Atomic cluster expansion: Completeness, efficiency and stability. *Journal of Computational Physics* **2022**, *454*, 110946.

(87) Yang, Y.; Zhang, S.; Ranasinghe, K. D.; Isayev, O.; Roitberg, A. E. Machine Learning of Reactive Potentials. *Annual Review of Physical Chemistry* **2024**, *75*, 371–395.

(88) P. A. Subramanyam, A.; Perez, D. Information-entropy-driven generation of material-agnostic datasets for machine-learning interatomic potentials. *npj Computational Materials* **2025**, *11*, 218.

(89) Pacini, A.; Ferrario, M.; Righi, M. C. Accelerating Data Set Population for Training Machine Learning Potentials with Automated System Generation and Strategic Sampling. *J. Chem. Theory Comput.* **2025**, *21*, 7102–7110.