Investigation into biological and biomimetic transmembrane systems



Zonghua Bo

Merton College

University of Oxford

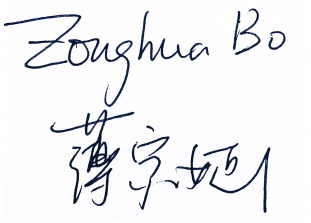
A thesis submitted for the degree of

Doctor of Philosophy

Sepetember 2023

Declaration

I, Zonghua Bo, declare that the Thesis I am submitting is entirely my own work except where indicated in the text, caption, footnote, or bibliography. This Thesis has not been submitted in whole or in part for any other academic degree or professional qualification.



September 8, 2023

Date

Signature

Acknowledgments

I express my deepest gratitude to Professor Fernanda Duarte for her exceptional supervision and support during my DPhil studies. Her guidance and expertise have been invaluable in shaping my research journey, and her commitment to academic excellence has inspired me.

I would like to express my heartfelt gratitude to my collaborators: Zhong Hui Lim, Professor Yujia Qing, and Professor Hagan Bayley on my nanopore research, and Dr. Aidan Kerckhoffs and Professor Matthew Langton on my anion transporter projects. Their active involvement and invaluable support have proven indispensable. I am truly grateful for the privilege of collaborating with them.

I want to express my sincere gratitude to the members of the Duarte group for their invaluable support. When I originally joined the group in 2014, Matina, Alistair, and Tom Young welcomed me and helped me get started. In particular, I am grateful to Tomasz and Matina for their assistance with my MD and ML work, respectively. I would also like to thank the members of the MD and ML subgroups for their help with troubleshooting and interesting discussions. I thoroughly enjoyed my time in the Duarte group and am honoured to have known such wonderful people as Hanwen, Henry, Tomasz, Shaoqi, Hafiz, Bernie, Chloe, Matina, and Aleksy, to name a few.

I also extend my gratitude to my dear friends, Yujia, Daisy, Hanwen, Yige, Scarlet, Hannah, Zheying, Chuancy, Kenny, Guanglong, Wula, Chunsen, Zhen Shao, Shuxiang,

Jim, Zhong, Hannah Hu, Eileen and Xiyan. Their constant support and encouragement have been invaluable throughout my journey. In particular, I am profoundly grateful to Yujia for being a pillar of strength during the COVID-19 pandemic. Her unwavering friendship and kindness played an indispensable role in helping me navigate the turbulent waters of depression and family difficulties.

I am grateful to my parents for their support and love throughout my academic endeavours. Their encouragement, guidance, and sacrifices enabled me to pursue this degree. I am eternally grateful for their faith in my skills.

Finally, to my partner Jon. His love, understanding, and belief in my abilities have been a constant source of motivation. Jon's patience, understanding, and willingness to help have made a significant impact on my success.

Abstract

Membranes are essential components of living organisms, which serve as effective barriers that separate distinct chemical environments on either side of the membrane. Chemists have designed biological and synthetic systems to functionalise membrane-embedded systems for a variety of applications such as sensing, sequencing, reaction mechanistic studies, and therapeutics. The continuous interest in functionalising membranes, combined with incomplete understanding of the underlying factors determining their mechanisms inspired the investigations undertaken in this work. This Thesis employs both experimental and computational methods to explore two distinct applications for sequencing and therapeutics, respectively.

(1) Engineered biological nanopores have found great success in DNA sequencing. The Bayley group previously reported a molecular hopper, which makes sub-nanometer steps by thiol-disulfide interchange along a track with cysteine footholds within a protein nanopore. In Chapter 2, the hopping rate was optimized with a view towards rapid enzymeless biopolymer characterization during translocation within nanopores. I first used a nanopore approach to systematically profile the reactivity of individual cysteine footholds along an engineered protein track at the single-molecule level. Using this approach, I calculated the pK_a of cysteine thiols and the pH-independent rate constants for the reaction between thiolates and a disulfide molecule. This reactivity profile guided site-specific mutagenesis. Together with the optimization of experimental

conditions, the overall stepping rate of a DNA cargo along a five-cysteine track was accelerated. This work extends the practical application of this enzymeless system as a sequencing method for biopolymers beyond DNA.

(2) Synthetic anion transporters have attracted significant attention as promising therapeutics for ion channel diseases. In Chapter 3, I use computational modelling to investigate the chloride binding and transmembrane transport mechanisms of *E*-/*Z*-switchable synthetic transporters. Using a model system, I developed a workflow to construct full energy profiles for the transmembrane transport process. These results revealed the importance of pre-organization of the *Z*-isomer and the balance between the energy barrier of transport and the solubility of the transporter. Additionally, in Chapter 4, I present a predictive machine-learning (ML) approach for estimating the chloride transport activity of a variety of synthetic chloride transporters. The ML models, employing both classification and regression frameworks, exhibited remarkable performance across a diverse range of systems. Moreover, they offered insights crucial for future design efforts, *e.g.*, identifying key structural features and experimental conditions that influence the observed transport activity.

Overall, this work bridges biological and molecular design, computational modelling and data-driven approaches to advance the development of two applications to functionalise membranes for sequencing and therapeutics. It provides interpretable molecular models as well as structure-activity relationships that will aid hypothesis generation and contribute to synthetic advances in both fields.

Publications

This Thesis is based on the following publication:

1. **Z. Bo**, Z. H. Lim, F. Duarte, H. Bayley, Y. Qing, Mobile molecules: reactivity profiling guides faster stepping on a cysteine track, *Angew. Chem. Int. Ed.*, **2023**, 62, e202300890.

2. A. Kerckhoffs, **Z. Bo**, S. Penty, F. Duarte, M. Langton, Red-shifted tetra-ortho-halo-azobenzenes for photo-regulated transmembrane anion transport, *Org. Biomol. Chem.*, **2021**, 19, 9058.

The following manuscripts were in preparation at the time of completion of this Thesis:

1. **Z. Bo**, A. Kerckhoffs, M. Langton, F. Duarte, Elucidating the transport activity of synthetic anion transporters through computational modelling.

2. **Z. Bo**, M. Langton, F. Duarte, Predicting chloride transport activity of synthetic anion transporters with machine learning.

3. S. Zavitsanou, **Z. Bo**, E. Casali, M. Langton, F. Duarte, *Pythia*: explainable machine learning in chemistry for non-experts.

Other publications:

1. **Z. Bo**, M. Yang, T. Xu, B. Xu, D. Wang, H. Zheng, Uni-GBSA: an open-source

and web-based automatic workflow to perform MM/GB(PB)SA calculations for virtual screening, *Briefings in Bioinformatics.*, **2023**, 24, 4, bbad218.

2. Y. Yu, C. Cai, J. Wang, **Z. Bo**, Z. Zhu, H. Zheng, Uni-Dock: GPU-accelerated docking enables ultra-large virtual screening, *J. Chem. Theory Comput.*, **2023**, 19, 11, 3336-3345.

3. Y. Guo, Y. Zhou, Q. Bai, **Z. Bo**, K. Song, J. Chang, Y. Zhang, M. Yang, Y. Deng, D. Wang, A Benchmark for Accurate GPCR Ligand Binding Affinity Prediction with Free Energy Perturbation, *ChemRxiv*, **2023**.

List of abbreviations

αHL α-Hemolysin from *Staphylococcus aureus*

AMPSO N-(1,1-Dimethyl-2-hydroxyethyl)-3-amino-2-hydroxypropanesulfonic acid

AUC Area under the ROC curve

As(III) Organoarsenic(III) molecular walker

BisBzImPy 1,3-Bis(benzimidazol-2-yl)pyrimidine

CaCC Calcium-activated chloride channel

CFTR Cystic fibrosis transmembrane conductance regulator

CIC Chloride ion channel

CV Cross-validation

Cys-113 Cysteine residue at position 113 on a single β strand in αHL

Cys-115 Cysteine residue at position 115 on a single β strand in αHL

Cys-117 Cysteine residue at position 117 on a single β strand in αHL

Cys-119 Cysteine residue at position 119 on a single β strand in αHL

Cys-121 Cysteine residue at position 121 on a single β strand in αHL

Cys-123 Cysteine residue at position 123 on a single β strand in αHL

D8 Octa-aspartate

DFT Density functional theory

DMSO Dimethylsulfoxide

DNA Deoxyribonucleic acid

DPPC 1,2-Dipalmitoyl-*sn*-glycero-3-phosphocholine

DPhPC 1,2-Diphytanoyl-*sn*-glycero-3-phosphocholine

DT Decision tree

DTNB 5,5'-Dithio-bis-(2-nitrobenzoic acid)

DTT DL-dithiothreitol

(E111S-M113C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a serine mutation at position 111 and a cysteine mutation at position 113

(E111Q-M113C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a glutamine mutation at position 111 and a cysteine mutation at position 113

(E111Q-5C)$_1$WT$_6$ Five-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a glutamine mutation at position 111 and five cysteine mutations at positions 113, 115, 117, 119 and 121

EC$_{50}$ The concentration of transporter required to obtain 50% chloride efflux

EDTA Ethylenediaminetetraacetic acid

ESP Electrostatic potential

ET Extra trees

FN False negative

FP False positive

(G119C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a cysteine mutation at position 119

G-mean Geometric mean

GP Gaussian process

HB Hydrogen bond

HEPBS N-(2-Hydroxyethyl)piperazine-N'-(4-butanesulfonic acid)

HPTS 8-Hydroxypyrene-1,3,6-trisulfonic acid

ISE Ion selective electrode

IVTT in vitro transcription and translation

kNN k-Nearest neighbours

LASSO Least absolute shrinkage and selection operator

LR Logistic regression

(M113C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a cysteine mutation at position 113

MACCS Molecular access system

MAE Mean absolute error

MBSA 3-morpholinopropane-1-sulfonic acid

MEET 2-(2-Methoxyethoxy)ethanethiol

MCC Matthew's correlation coefficient

MD Molecular dynamics

ML Machine learning

MM Molecular mechanics

MSE Mean squared error

(N121C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a cysteine mutation at position 121

(N123C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a cysteine mutation at position 123

NPT Isothermal-isobaric ensemble

NVT Canonical ensemble

OLS Ordinary least squares

OPSSG Orthopyridyl glutathionyl disulfide

PCA Principle component analysis

PME Particle mesh ewald

PMF Potential of mean force

POPC 1-Palmitoyl-2-oleoyl-glycero-3-phosphocholine

QSAR Quantitative structure-activity relationship

R^2 R-squared

RF Random forest

RMSE Root mean squared error

RNA Ribonucleic acid

ROC Receiver operating characteristic

rRBCm Rabbit red blood cell membranes

SHAP Shapley additive explanations

SMOTE Synthetic minority over-sampling technique

SPAA 4-Sulphophenylarsonous acid

ssDNA Single-stranded DNA

SVM Support vector machines

(T115C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a cysteine mutation at position 115

(T117C)$_1$WT$_6$ Single-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with a cysteine mutation at position 117

TN True negative

TP True positive

t-SNE T-distributed stochastic neighbour embedding

UMAP Uniform manifold approximation and projection

US Umbrella sampling

WHAM Weighted histogram analysis method

WT Wild-type

(WT-5C)$_1$WT$_6$ Five-cysteine nanoreactor containing six wild-type subunits and one mutant subunit with five cysteine mutations at positions 113, 115, 117, 119 and 121

Contents

1 Introduction **1**

2 Reactivity profiling guides faster movement on a cysteine track **8**

2.1 Abstract 8

2.2 Introduction 9

2.2.1 α-Hemolysin nanopore 9

2.2.2 History of nanopore DNA sequencing 10

2.2.3 Covalent chemistry in αHL nanopores 13

2.2.4 Chemical stepping along a protein track 14

2.3 Theory 17

2.3.1 αHL nanoreactor 18

2.3.2 Single-channel recording 18

2.3.3 Time resolution and detection limit 20

2.3.4 Single-molecule kinetic analysis 21

2.4 Materials and methods 21

2.4.1 General 21

2.4.2 *In vitro* transcription and translation and preparation of αHL mutants 21

2.4.3 Single-channel electrical recordings 23

2.4.4 Molecular dynamics simulations 24

2.5 Results and discussion 26
2.5.1 Current blockades and spatial blockades of αHL-adducts . . . 28
2.5.2 Single-molecule reactivity profiling of individual cysteine residues 33
2.5.3 Enhanced cysteine reactivity by mutating a neighbouring residue 41
2.5.4 Accelerated chemical stepping along engineered tracks 42
2.6 Conclusion 46

3 Elucidating the transport activity of synthetic anion transporters through computational modelling 49

3.1 Abstract 49
3.2 Introduction 50
3.2.1 Synthetic anion transporters 50
3.2.2 Computational modelling of synthetic anion transporters . . . 53
3.3 Theory 55
3.3.1 Force fields 56
3.3.2 Nonstandard parameters 59
3.3.3 MD simulation 59
3.3.4 Isobaric-isothermal ensemble 61
3.3.5 Umbrella sampling 63
3.4 Materials and methods 64
3.4.1 Transporter parametrization 64
3.4.2 MD simulation protocol 64
3.4.3 Free energy calculations of chloride transport across a POPC membrane 65
3.5 Results and discussion 66
3.5.1 Chloride binding studies 66
3.5.2 *E*-isomer dimerization 70
3.5.3 Passive diffusion in a POPC membrane 71

3.5.4 Chloride transport across a POPC membrane 73
3.6 Conclusion 78

4 Predicting chloride transport activity of synthetic anion transporters with machine learning 79

4.1 Abstract 79
4.2 Introduction 80
4.3 Theory 84
4.3.1 Data collection and preprocessing 84
4.3.2 Feature representation 85
4.3.3 Model training 89
4.3.4 Supervised learning algorithms 92
4.3.5 Model evaluation 100
4.4 Materials and method 104
4.5 Dataset 111
4.6 Feature engineering 113
4.7 Model training and evaluation 114
4.7.1 Classification models 115
4.7.2 Regression models 130
4.8 Conclusion 138

5 Conclusions 139

5.1 Thesis summary 139
5.2 Perspectives and future directions 142
5.2.1 Towards an ideal track for biopolymer translocation 142
5.2.2 Computer-aided design of synthetic anion transporters 144

Bibliography 148

Appendices **183**

A Supplementary information for Chapter 3 **184**

A.1 MD simulations of 1^Z in DMSO 185

A.2 MD simulations of 12^Z and 13^Z in DMSO 188

A.3 MD simulations of 1^E in DMSO 191

A.4 Free energy calculations of chloride transport across a POPC membrane 194

B Supplementary information for Chapter 4 **197**

B.1 Dataset analysis 198

B.1.1 Train-test split 203

B.2 Classification models 206

B.2.1 Addressing class imbalance with class weights 206

B.2.2 Model performance on Morgan fingerprints 211

B.2.3 Model performance on Mordred descriptors 213

B.3 Regression models 231

B.3.1 Permutation test on the test set 231

B.3.2 Model performance on fingerprints 233

B.3.3 Model training and evaluation on Mordred descriptors 235

C Molecular dynamics theory **246**

C.1 Molecular interactions 246

C.1.1 Bonded interactions 246

C.1.2 Non-bonded interactions 248

C.2 Molecular dynamics algorithm 251

C.2.1 Equations of motion for molecular systems 251

C.2.2 Integration of equations of motion 252

C.3 Timestep and constrains 253

C.4 Periodic boundary condition 254

C.5 Ensembles 254
C.6 Thermostats 255
C.7 Barostats 257

Appendix bibliography **258**

Chapter 1

Introduction

Membranes are essential components of living organisms, which serve as effective barriers that separate distinct chemical environments on either side of the membrane.[1] The exchange of ions across membranes is vital in numerous biological processes, including the regulation of cellular pH, the maintenance of cell volume and osmotic balance, and cellular signalling. Phospholipid membranes allow the easy diffusion of small neutral molecules like water and carbon dioxide, while charged solutes such as ions require the assistance of membrane transport proteins.

There are two primary classes of membrane transport proteins: channels and carriers.[2] Channel proteins are typically embedded within the membrane, while carriers can either be associated with the membrane or exist independently. Carriers responsible for moving cargo across the membrane must possess sufficient lipophilicity to enable their diffusion through the hydrophobic region of the membrane. In contrast, channel proteins generally span the entire width of the membrane, creating a hydrophilic pathway that facilitates ion movement. A further classification can be made based on the mechanism of action by which ion transport occurs in these proteins. This classification considers the stoichiometry and direction of transport, leading to three

categories: uniport, symport, and antiport transporters (Figure 1.1A). "Uniport" refers to the transport of a single type of ion across a membrane. "Symport" involves the transport of two types of ions in the same direction across the membrane, whereas "antiport" entails the transport of two types of ions in opposite directions across the membrane. Additionally, if transmembrane ion transport results in a net change in electrical charge across the membrane, it is described as "electrogenic"; otherwise, it is referred to as "electroneutral" (Figure 1.1B). When solute movement relies solely on diffusion, the process is termed "passive transport", while processes requiring additional energy input are termed "active transport".

In biology, in addition to membrane transport proteins, ion transport across membranes can also be achieved by ionophores, which are small hydrophobic molecules capable of dissolving in membranes. Ionophores, like membrane transport proteins, can be classified into mobile carriers and membrane-spanning channels. Mobile carriers facilitate ion movement by shuttling them through the membrane, while membrane-spanning channels establish a hydrophilic pathway for ions to traverse the membrane (Figure 1.2). Both function by shielding the charge of the transported ion, allowing it to move across the lipid bilayer driven by their electrochemical gradients.

Chemists have designed biological and synthetic systems to functionalise membranes for a variety of applications. Membrane channel proteins, combined with single-channel electrophysiology, which was originally developed for real-time monitoring of channel gating dynamics, enabled various applications including single-molecule sensing,[3–5] sequencing,[6,7] and reaction mechanistic and kinetics studies.[8–14] Bacterial toxin proteins were common membrane proteins of choice for these applications due to their stable open channels producing low-noise ionic currents when subjected to an applied electric potential.[15–17] In particular, α-hemolysin (αHL), the bacterial toxin secreted by *Staphylococcus aureus*, has played a critical role in the development

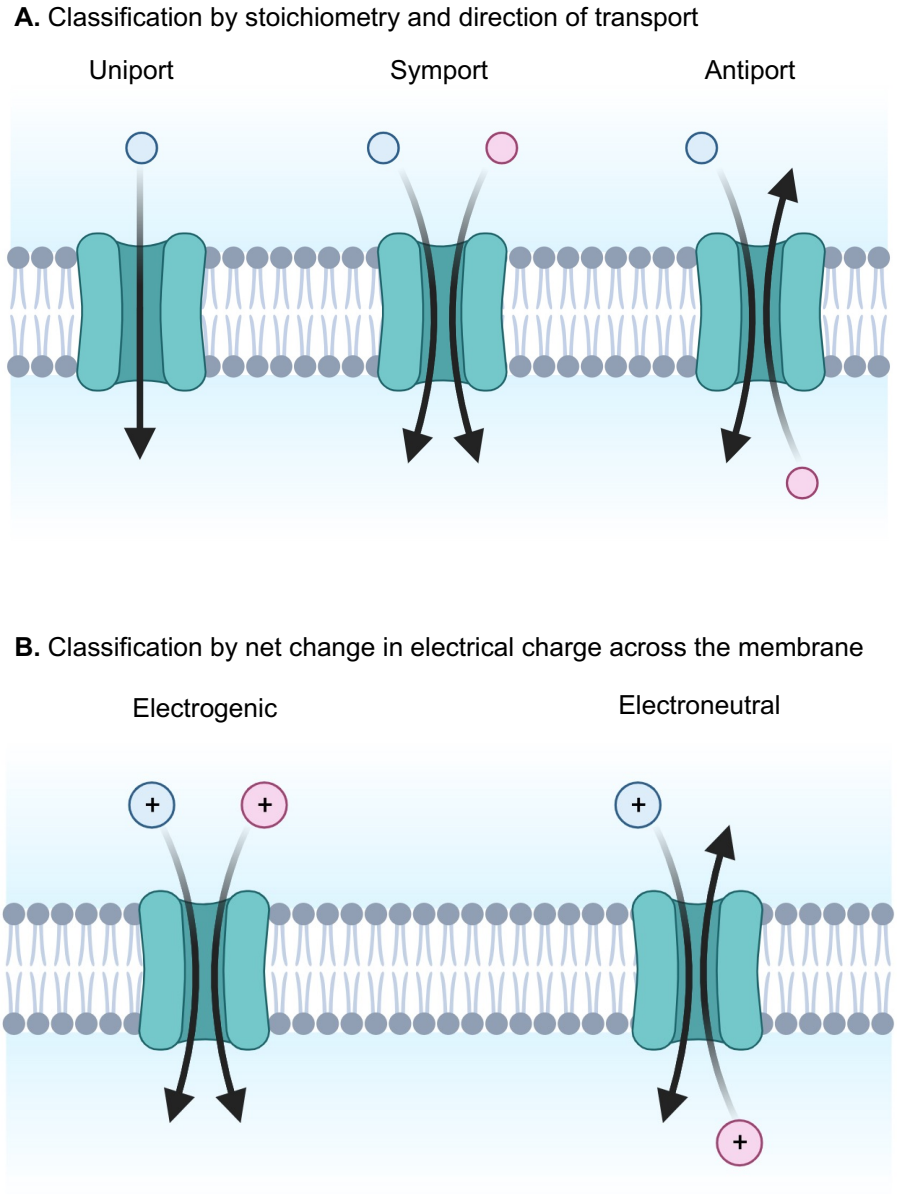


Figure 1.1: Types of transmembrane ion transport, (A) classified based on the transport stoichiometry and the direction and (B) classified based on net change in electrical charge across the membrane.

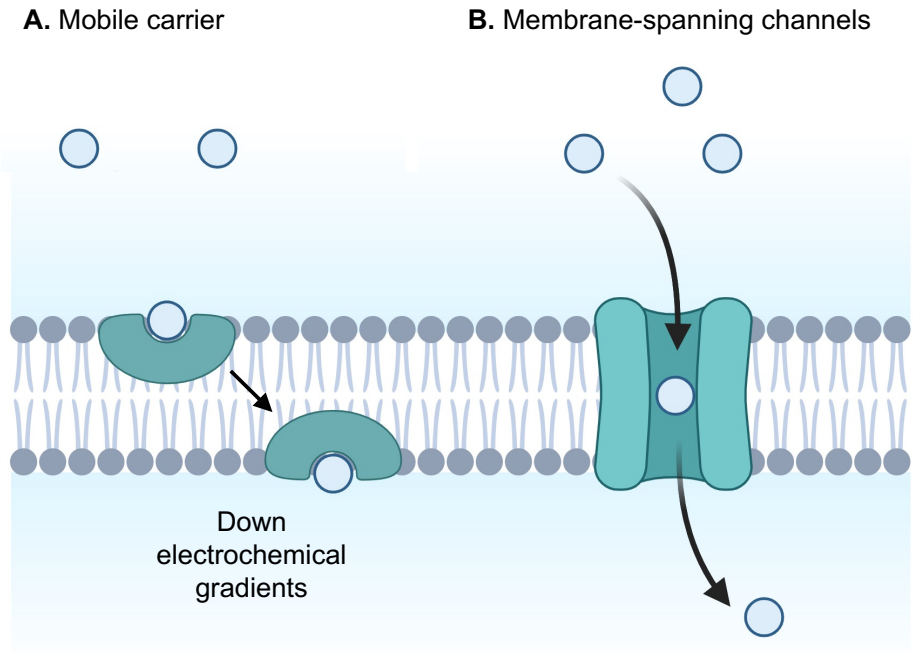


Figure 1.2: Two types of ionophores. (A) Mobile carriers bind to a particular ion, protecting its charge from the environment around it and thus facilitating its passage into the hydrophobic interior of the lipid membrane. (B) Membrane-spanning channels act by introducing a hydrophilic pore into the membrane, allowing the ion to pass and avoiding contact with the hydrophobic interior of the membrane.

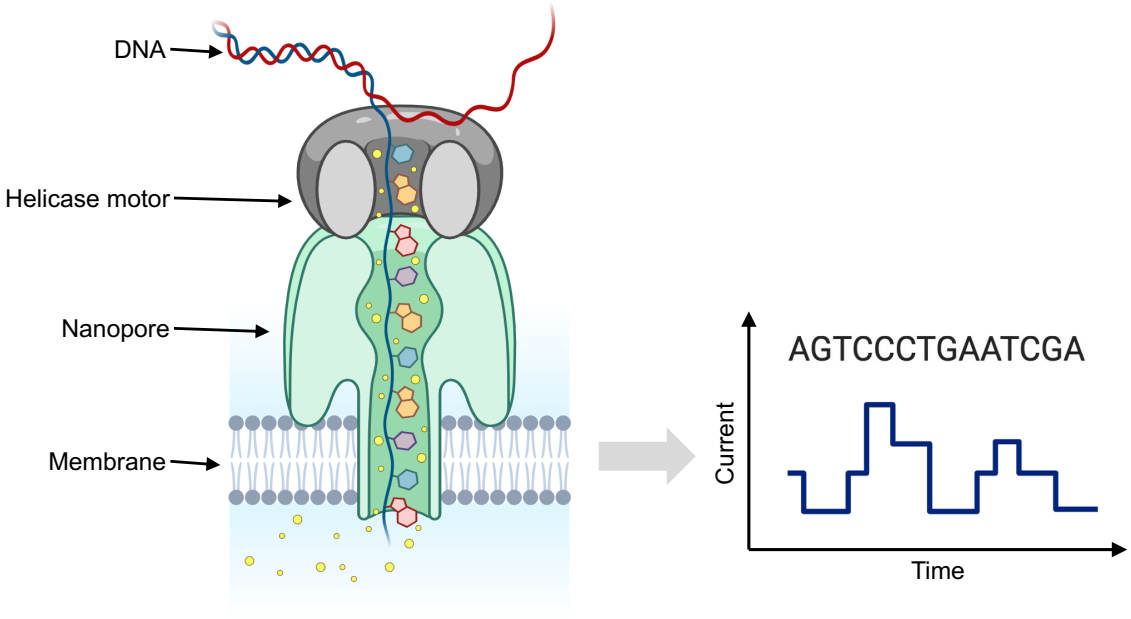


Figure 1.3: Schematic representation of a nanopore DNA sequencer in action. A DNA strand is threaded through a nanopore with a helicase motor. The resulting ionic current is measured and used to decode the DNA sequence.

and commercialization of low-cost, parallel, long-read sequencing technology that revolutionised DNA sequencing (Figure 1.3).[6,7]

Ionophores have found applications as research tools for studying membrane transport. For example, Valinomycin, which is a cyclic depsipeptide that is selective for potassium ion transport, is commonly used as control for studying cation transport.[18,19] Similarly, gramicidin A is employed as a model system in biophysical research to investigate cation channels and their interaction with membrane lipids.[19,20] Furthermore, anions play a pivotal role in numerous cellular processes. They contribute to the regulation of cellular pH, the maintenance of cell volume and osmotic balance, and serve as cellular signalling molecules.[21] Disruptions to transmembrane anion transport due to various factors, such as toxicity, autoimmune responses, or genetic mutations, can result in a range of ion channel diseases including cystic fibrosis,[22,23] Bartter's syndrome,[24,25] and Dent's disease.[26] Given the crucial role of transmembrane anion transport in cellular function, synthetic anion transporters have also attracted great interest as

therapeutics for ion channel diseases in addition to serving as research tools.[27–30]

The continuous interest in designing engineered biological systems and synthetic systems to functionalise membrane-embedded systems and our incomplete understanding of the underlying factors determining the kinetics and mechanisms of these systems have inspired the investigations undertaken here. This Thesis uses both experimental and computational methods to explore two such applications: (1) engineered biological nanopores for sequencing and as nanoreactors for single-molecule chemistry; and (2) synthetic anion transporters as promising therapeutics for ion channel diseases.

Chapter 2 presents a workflow to systematically profile the reactivity of individual cysteine footholds along an engineered protein track using a single-molecule approach. The pK_a values of cysteine thiols and the pH-independent rate constants of the thiolates with a small-molecule disulfide varied are calculated. Guided by the reactivity profile, site-specific mutagenesis is performed, and together with optimization of experimental conditions, the overall stepping rate of a DNA cargo along a five-cysteine track is accelerated. This work advances the practical application of an enzymeless system for a generalised sequencing method for biopolymers other than DNA.

Chapter 3 presents a computational modelling study of E-/Z-switchable anion transporters using a model system. Key factors that affect anion binding and transmembrane transports are revealed. A computational workflow to construct full energy profiles for the transmembrane anion transport process is also presented. This work advances the computational study and future design of synthetic anion transporters.

Chapter 4 presents the use of machine learning (ML) to predict the transmembrane transport activity of synthetic chloride transporters. Our ML models, employing both classification and regression frameworks, exhibit remarkable performance across a diverse range of systems. Moreover, they offer insights relevant to future design efforts. For example, our ML models identify key structural features and experimental

conditions that influence the observed transport activity. This work bridges data-driven approaches and molecular design, providing interpretable structure-activity relationships that will aid hypothesis generation and contribute to synthetic advances in this field.

Chapter 2

Reactivity profiling guides faster movement on a cysteine track

2.1 Abstract

A molecular hopper, which makes sub-nanometer steps by thiol-disulfide interchange along a track with cysteine footholds within a protein nanopore, was previously reported by the Bayley group. Here, the hopping rate ($\sim$0.1 s^{-1} in the previous work) was optimized with a view towards rapid enzymeless biopolymer characterization during translocation within nanopores. I first took a single-molecule approach to obtain the reactivity profiles of individual footholds. The pK_a values of cysteine thiols within a pore ranged from 9.17 to 9.85, and the pH-independent rate constants of the thiolates with a small-molecule disulfide varied by up to 20-fold. Through site-specific mutagenesis and a pH increase from 8.5 to 9.5, the overall hopping rate of a DNA cargo along a five-cysteine track was accelerated 4-fold and the rate-limiting step 21-fold. I also employed computational modelling to reveal a strong correlation between the relative spatial blockade and the relative current blockade of covalently

linked reactants in a nanoreactor.

This project was in collaboration with Zhong Hui Lim, Prof. Yujia Qing, and Prof. Hagan Bayley. I performed the experiments on reactivity profiling, including single-cysteine nanoreactor preparation, single-channel recordings on the three-step thiol-disulfide interchange reaction, linear regression analysis and calculations of the pK_a values, and the pH-independent rate constants. I also performed all the computational modelling work. Zhong Hui Lim performed the experiments concerning the five-cystine track, including track engineering via site-directed mutagenesis and single-channel recordings and kinetics studies of the chemical stepping of a DNA cargo on the engineered tracks. This work was published in *Angew. Chem. Int. Ed.* in 2023.[14]

2.2 Introduction

2.2.1 α-Hemolysin nanopore

α-Hemolysin (αHL) is a bacterial toxin secreted by *Staphylococcus aureus*.[31] It consists of a 293-residue monomeric polypeptide that assembles into a pore containing seven identical subunits.[32] This pore spans approximately 10 nm. The entrance to the pore is located on the *cis* side of the bilayer, extending approximately 6.5 nm above the membrane surface and having a diameter of approximately 3 nm (Figure 2.1). Moving from the *cis* to the *trans* side, the pore narrows to a diameter of 2.4 nm before opening into a roughly spherical vestibule measuring around 4.2 nm in diameter. Approximately 1.5 nm above the membrane surface, the vestibule narrows further and becomes a 14-stranded β-barrel structure that spans 5.2 nm in length and has an average diameter of approximately 2 nm. The *trans* entrance to the pore is defined by the turns in the transmembrane hairpins.

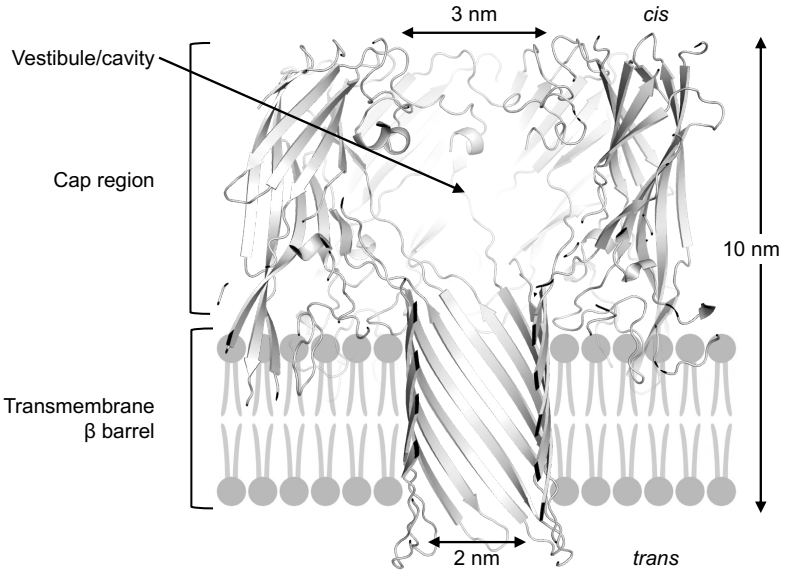


Figure 2.1: An αHL nanopore.

2.2.2 History of nanopore DNA sequencing

Nanopore sensing and sequencing rely on the single-channel electrophysiology technique,[3,6] which measures ion flow through the nanopore. The single-channel electrophysiology technique was first invented in 1976 by Neher and Sakmann for real-time monitoring of the opening and closing of single ion channels (Figure 2.2).[33] This technique significantly advanced the studies of ion channel functions. For this achievement, Neher and Sakmann were awarded the Nobel Prize in Physiology or Medicine in 1991. When a stable open channel, like the one in αHL, is used, analytes introduced into the nanopore under the influence of an applied electric potential can induce picoampere changes in the ionic current. These changes, including blockade amplitude, duration, and frequency of the ionic current, can be utilised for the stochastic sensing and characterization of various analytes at the single-molecule level, including metal ions,[34,35] small organic molecules,[36,37] and biomolecules such as DNA,[38–41] peptides[42,43] and proteins.[44,45] The most notable application is nanopore DNA sequencing.[6,7]

The first concept of nanopore sequencing was sketched by Deamer in a notebook in 1989 (Figure 2.2).[7] In 1993, Bezrukov and Kasianowicz determined the conditions necessary

to avoid spontaneous gating of αHL.[46] This was a crucial step towards nanopore sequencing as it allowed unhindered observations of DNA translocation through the nanopore. In 1996, the first observation of single-stranded DNA (ssDNA) translocation through an αHL was observed.[47] Later that year, the first crystal structure of αHL was resolved using X-ray crystallography by Gouaux *et al.* (Figure 2.2).[32] In 2005, a single adenine nucleotide was identified at a specific location on a ssDNA using αHL by the Bayley and Ghadiri groups (Figure 2.2).[48] In the same year, Oxford Nanopore Technologies (ONT) was founded by Hagan Bayley with Gordon Sanghera, Spike Willcocks, and David Norwood.

At the time, researchers found that each nucleobase in a ssDNA strand was driven through the nanopore sensing region in <10 microseconds even with the smallest applied electric potential.[7] To enable individual nucleobase identification, each nucleobase needed to stay in the nanopore sensing region for > 100 microseconds. Thus, establishing a slow ratcheting mechanism of ssDNA through the nanopore was critical in the development of the nanopore DNA sequencing technique. This slow ratcheting was finally realised in 2012 using the bacteriophage phi29 DNA polymerase (Figure 2.2), which enabled processive nucleobase-by-nucleobase resolution of ssDNA.[49] Since then, with various optimizations including the choice and engineering of nanopores, motor proteins, membranes, conditions, and decoding algorithms, the commercialization of low-cost, parallel, long-read sequencing technology with nanopores has been hugely successful and revolutionised DNA sequencing (Figure 2.2).[6,7]

The success of nanopore DNA sequencing sparked the exploration of other biopolymers using the nanopore method. Currently, in nanopore DNA sequencing, processive nucleobase-by-nucleobase translocation of DNA strands is facilitated by enzymes like helicase motor proteins.[6] However, such enzymes do not exist for other biopolymers, limiting the generalizability of the nanopore sequencing techniques on other biopoly-

mers. Stepwise translocation by chemical ratcheting poses a promising solution.[50]

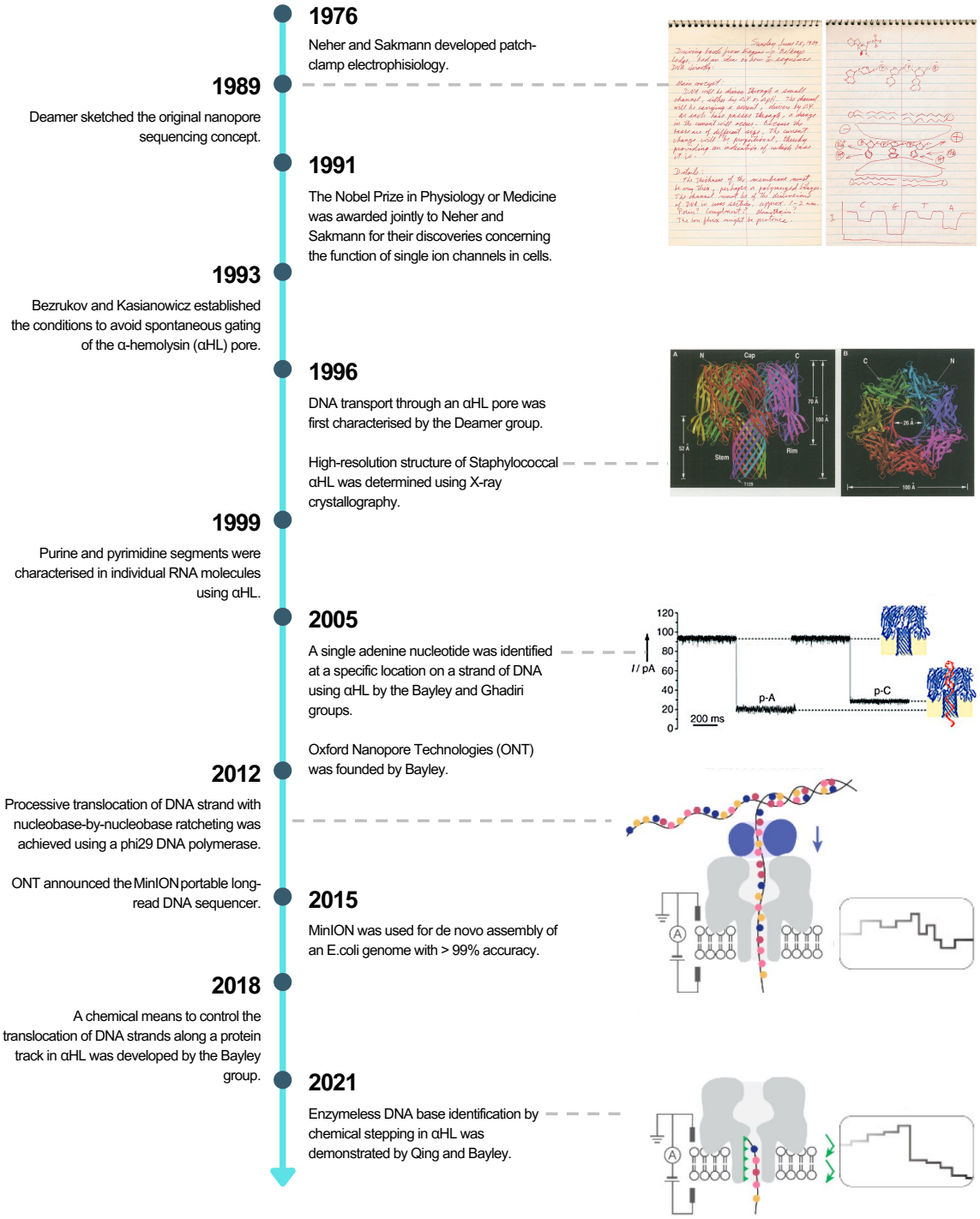


Figure 2.2: Milestones of nanopore DNA sequencing and a timeline towards a generalised sequencing technique for biopolymers. Figures are reproduced from References[7, 32, 50, 51] with permissions.

2.2.3 Covalent chemistry in αHL nanopores

To discuss stepwise translocation by chemical ratcheting, I first need to introduce how covalent chemistry is monitored in αHL nanopores. Similar to how analytes can induce picoampere changes in the ionic current going through the nanopore, individual bond-making and bond-breaking events within the nanoreactors can also be monitored by employing similar techniques.[52] The Bayley group has developed a means to monitor chemical reactions in engineered αHL nanopores known as "nanoreactors".[8,53]

αHL pores have several desirable properties that make them suitable as nanoreactors: (1) In the RCSB PBD database, there are currently 14 crystal structures available with resolutions ranging from 1.89 Å to 3.40 Å. The high-resolution X-ray structure (PDB ID 7AHL, resolution = 1.89 Å) was reported in 1997.[32] This availability of crystal structures facilitates rational engineering.[16] (2) αHL is stable over a wide range of pH values[54] and temperatures.[55] (3) The occurrences of spontaneous gating (pore opening and closing) events, seen in many other related β-barrel proteins, are rare in αHL.[52] In αHL, the pore remains stably open at transmembrane potentials below 100 mV in common buffer solutions. (4) αHL has a pore diameter of 2 nm, which is suitable for detecting chemical reactions by disrupting ionic currents.[32] This size ensures that molecules occupy a significant portion of the pore's diameter, which creates a large enough current change for effective monitoring, while not creating steric hindrance.[52]

An αHL nanoreactor typically consists of six wild-type subunits and one mutant subunit containing one or more reactive site(s) (Figure 2.3). It has been widely employed to examine various aqueous chemical reactions within the pore. Many of these reactions involve thiol chemistry with cysteine side chains.[8,10,13,14,56,57] This is due to the easy installation of this residue through mutagenesis as well as the reactivity enabled by this residue, which can act as a potent nucleophile when deprotonated and

engage in rapid and reversible chemical reactions, including thiol-disulfide interchange and arsenic(III)-thiol chemistry, facilitating stepwise molecular motion.[58,59] In 2002, reversible covalent chemistry in an αHL nanoreactor using a substitution reactions of arsenic(III) compounds with cysteine thiol was first observed by Bayley *et al.*[59] Beyond thiols, other reactive sites, such as tetrazines, can also be incorporated into the nanoreactors by site-directed chemical modification[11,12,60,61] or native chemical ligation.[62,63]

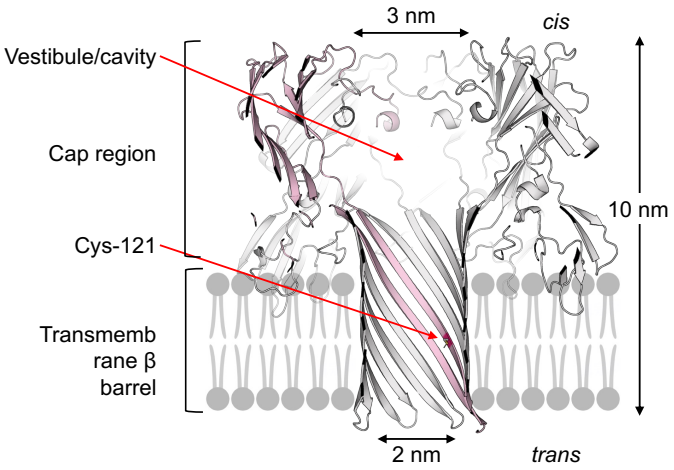


Figure 2.3: An αHL nanoreactor with a reactive cysteine site at position 121. The nanoreactor consists of six wild-type subunits (grey) and one mutant subunit (pink) containing a single reactive cysteine site at position 121.

2.2.4 Chemical stepping along a protein track

In cells, motor proteins, such as kinesins, dyneins, and myosins, move along microtubule and actin tracks.[1] In contrast to the prevalence of protein-based tracks in nature, synthetic mobile molecules tread mostly on DNA[64–66] or small-molecule tracks,[67–69] which are often designed for ease of synthesis. Recently, the ability to chemically traverse multiple reactive sites on an engineered protein track has also been realised. An engineered cysteine track, built along a single β strand within an αHL nanoreactor has been reported by Bayley *et al.*[10,70] This protein track consists of cysteine footholds strategically positioned ~6.9 Å apart along a β strand at positions 113, 115, 117, 119,

and 121 within an αHL nanopore (Figure 2.4A).[10,70] Cysteine is a desirable building block for a protein track because the side-chain thiol, when deprotonated, is a powerful nucleophile. Thiolates can participate in various rapid and often reversible chemistries (*e.g.*, thiol-disulfide interchange and arsenic(III)-thiol chemistry) that can be used to produce stepwise molecular motion.[58,59]

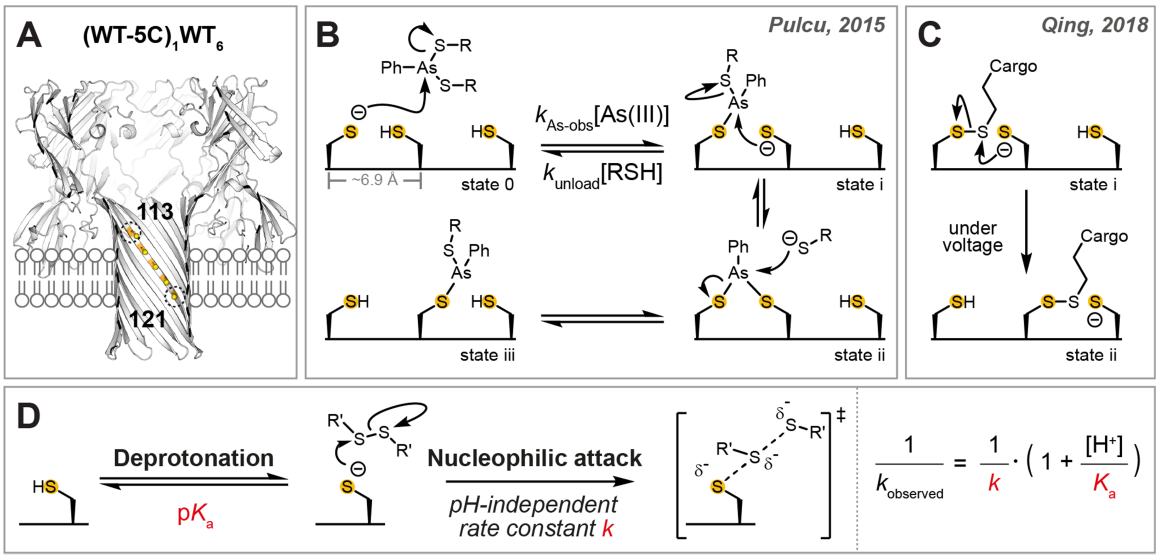


Figure 2.4: Cargo translocation along a protein track. (A) An αHL nanoreactor containing a five-cysteine track bounded by Cys-113 and Cys-121. (B) The molecular walker, formed in situ by the reaction of 4-sulphophenylarsonous acid with excess 2-(2 methoxyethoxy)ethanethiol, is referred to as As(III). The walker was first loaded onto the protein track (state 0 to i); translocation of the walker (state i to iii) was mediated by two consecutive nucleophilic substitutions. (C) Chemical stepping of the molecular hopper along the protein track proceeded through thiol-disulfide interchange (state i to ii). The directionality of translocation was controlled by an externally applied electric potential. (D) Reactivity in a thiol-disulfide interchange reaction can be separated into two components: thiol deprotonation and subsequent nucleophilic attack. The reaction parameters of individual steps were obtained by using the equation shown. Adapted from Reference[14] with permission.

In 2015, the Bayley group reported a two-legged molecular walker that moves along this track (Figure 2.4B).[10] The walker corresponded to an organoarsenic(III) molecule, As(III), formed *in situ* through the reaction between 4-sulphophenylarsonous acid (SPAA) and an excess of 2-(2-methoxyethoxy)ethanethiol (MEET). The walking mechanism involves the displacement of thiol ligands by free thiols of the cysteine footholds, allowing the walker to take steps along the track. The As(III) walker exhibited stochastic motion with a slight bias towards the end of the track at position

121, which was likely driven by a thermodynamic sink resulting from the distinct environments of the five cysteine residues within the β-barrel. The speed of the As(III) walker could be adjusted by using different thiol ligands, as well as by changes in pH and temperature. In 2018, a single-legged molecular hopper carrying a DNA cargo was developed, utilising thiol-disulfide interchange to move along the cysteine track (Figure 2.4C).[70] The precise regioselectivity of each stepping reaction ensured that the hopper remained bound to the track at all times, which enabled highly processive and directional movement. At pH 8.5, the molecular hopper exhibited an average stepping rate of 0.13 s^{-1}, with variations in rate depending on the specific position along the track.

This molecular hopper served as a prototype for directional, processive, stepwise ssDNA translocation without the need for the helicase motor,[70] which held promise for a generalised nanopore biopolymer sequencing technique. This was finally achieved in 2021 when the molecular hopper system was developed into an enzymeless system that enables directional stepwise translocation of biopolymers for sequence analysis.[50] However, the stepping rates were much slower than those achieved with the helicase motors, which make hundreds of steps per second.[6] For the enzymeless system to find practical application in a generalised sequencing method for biopolymers beyond DNA, it will be necessary to extend the length of the track and increase the rate of stepping. The former may require exploring alternative pores beyond αHL, while the latter could be achieved through track optimization, such as enhancing the reactivity of the cysteine footholds and adjusting experimental conditions.

In this Chapter, I address the latter. I first took a nanoreactor approach to probe the reactivity of individual cysteine footholds at the single-molecule level. Reactivity in the context of the thiol-disulfide interchange that underlies chemical hopping was determined from the pK_a values of the cysteine thiols and the pH-independent rate

constants for nucleophilic attack by the corresponding thiolates on a set electrophile (Figure 2.4D). Single-cysteine mutations were made at positions 113, 115, 117, 119, 121 or 123. Six nanoreactors, each containing one mutant subunit, were prepared. Each nanoreactor was inserted into a planar lipid bilayer for electrical recording. Small-molecule disulfide and thiol reagents were spatially separated by the lipid bilayer. Thiol-disulfide interchange reactions occurred at the cysteine site in a three-step reaction cycle. Changes in the ionic current passing through the αHL nanoreactor under an applied transmembrane electric potential were utilized to monitor the reaction cycle. I calculated the pK_a values of the cysteine thiols and the pH-independent rate constants of the thiolates for attack on the disulfide using the observed rate constants, which were obtained at different pH values from the mean lifetimes of each current state. The local protein environment around Cys-113 was modified on the protein track by mutations and identified a favourable mutation that increased the rate. This mutation was later incorporated into protein tracks. Finally, the chemical stepping of a DNA molecule on a modified cysteine track was tested at two pH values; faster chemical stepping was observed at a higher pH. This work advances the practical application of the enzymeless system for sequencing biopolymers in general.

2.3 Theory

In this Thesis, I study the reactivity of individual cysteine footholds within an engineered protein track at the single-molecule level using engineered protein pores known as "nanoreactors". These nanoreactors enable us to monitor individual bond-making and bond-breaking events within the pore using single-channel recording. Here, I introduce the theory behind αHL nanoreactors and the single-channel recording technique.

2.3.1 αHL nanoreactor

An αHL nanoreactor typically consists of six wild-type subunits and one mutant subunit containing one or more reactive site(s). To achieve this, an octa-aspartate (D8) tail is introduced in the mutant polypeptide, whereas the wild-type polypeptide lacks this. After protein expression and oligomerization, homoheptamers and heteroheptamers with different numbers of mutant subunits are separated by gel-electrophoresis and located by autoradiography. Specifically, αHL heptamers composed of different numbers of mutant subunits carry different numbers of D8 tails. As a result, different heptamers have different electrophoretic mobilities and can be separated in the gel. After autoradiography, all the bands are visible. The top band corresponds to homoheptamers WT_7, the second band corresponds to $(mutant)_1WT_6$ and so on, with consecutive bands having a WT subunit replaced with a mutant subunit.

The relevant band is then excised and extracted from the gel to yield a heteroheptamer with a single mutation. To ensure only one subunit is mutated, the absence of multiple mutant subunits within a single pore can be confirmed by observing no more than one reaction within a pore.

2.3.2 Single-channel recording

To monitor single-molecule chemistry in nanopores, a single αHL nanoreactor is inserted into a synthetic lipid bilayer (Figure 2.5A). An externally applied electric potential induces ionic current flows through the pore. Chemistry occurring within the nanopore can change the mass, shape, or polarity of the attached reactant and perturb the ion flow through the pore, resulting in changes in the recorded current output (Figure 2.5B). Compared to traditional ensemble measurements, this single-molecule approach provides real-time observation of reaction intermediates in a sequential manner and without the need for any tags or labels.[53] Fast steps following slower

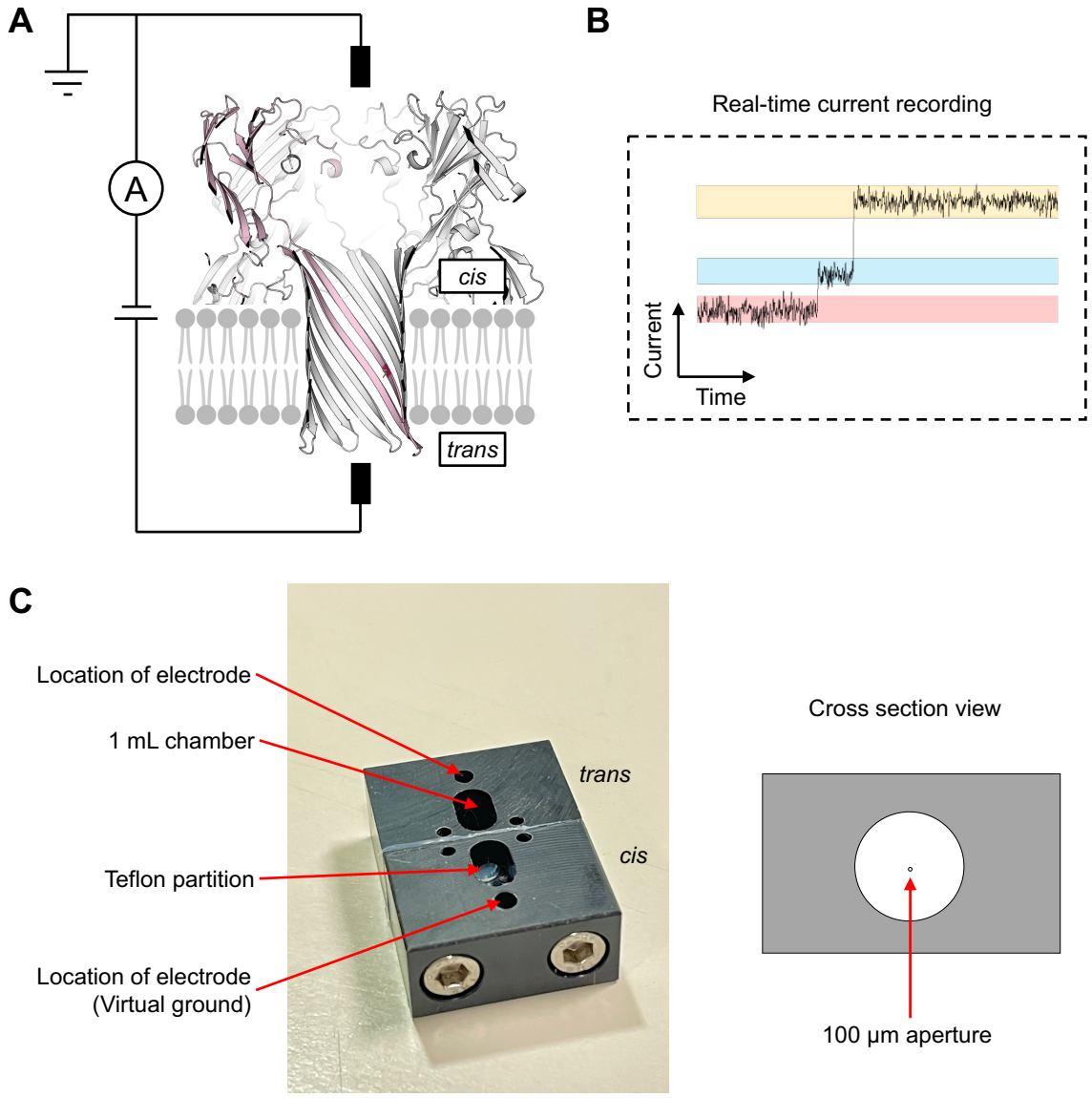


Figure 2.5: Nanopore approach to single-molecule chemistry. (A) An αHL nanoreactor, consisting of six wild-type subunits (grey) and one mutant subunit (pink) containing a single reactive cysteine site at position 121, is inserted into a synthetic lipid bilayer. An external electric potential is applied across the bilayer to generate ion flow through the pore. (B) Chemical reactions at this site disrupt the ion flow through the pore. This is reflected in real-time current recording as changes in current levels. (C) Setup for single-channel recording.

ones can be easily discerned, and complex branched pathways can be dissected at the single-molecule level. This analysis results in a comprehensive understanding of the reaction pathway.[52]

A typical single-channel recording setup consists of two chambers, referred to as *cis*

and *trans*, separated by a thin teflon barrier, which is approximately 25 μm thick (Figure 2.5C).[53] A small aperture with a diameter of around 100 μm is formed in the teflon. To form a lipid bilayer across the aperture, the lipids are first prepared by dissolving appropriate lipids in a suitable solvent, spreading the solution over a hole in a plastic partition, separating two chambers, and allowing the solvent to evaporate to form the lipid bilayer. Electrolyte solutions, containing suitable buffers to maintain pH and ionic strength, are added to both chambers. Electrodes in both chambers establish an electric potential across the bilayer, which is connected to a data acquisition system for current measurement. The *cis* chamber is grounded, while the applied potential is determined on the *trans* side. A positive current indicates the movement of cations from the *trans* to the *cis* chamber or anions from the *cis* to the *trans* chamber through the pore. Finally, a small amount of protein pores are added to the *cis* chambers and current flow measurements are recorded when a single pore spontaneously inserts itself into the bilayer.

2.3.3 Time resolution and detection limit

In single-channel current recording, the instrument noise limits the time resolution to the microsecond domain.[52] A small molecule freely diffusing through the pore cannot be detected. This is because the diffusion coefficient (D) for a small molecule like glucose and lactose in water is $\sim 0.5 \times 10^{-5}$ cm^2 s^{-1}.[71] For these molecules, the mean residence time halfway in the β-barrel ($x = 2.5$ nm) can be calculated as $\frac{x^2}{2D} = 6.25$ ns, which is too quick to be detected using single-channel recording. As a result, a reactant is typically attached to the lumen of the pore by a covalent bond to ensure the molecule resides in the pore for long enough to be detected.

2.3.4 Single-molecule kinetic analysis

At the single-molecule level, the observed rate constants (k_{obs}) can be calculated from the mean lifetimes ($< \tau >$) of states.[53] For a bimolecular reaction between the pore and a molecule, A, such as step 1 and step 2, the concentration of A ($[A]$) remains relatively unchanged in the single-molecule experiment. k_{obs} is pseudo-first order dependent on $[A]$.

$$< \tau >= \frac{1}{k_{obs}} = \frac{1}{k_A[A]} \tag{2.1}$$

For a simple unimolecular reaction, such as step 3, k_{obs} is the reciprocal of $< \tau >$.

$$< \tau >= \frac{1}{k_{obs}} \tag{2.2}$$

2.4 Materials and methods

2.4.1 General

All reagents were purchased from Sigma-Aldrich unless stated otherwise. 1,2-Diphytanoyl-*sn*-glycerol-3-phosphocholine (DPhPC) was purchased from Avanti Polar Lipids. Orthopyridyl glutathionyl disulfide (OPSSG) was synthesized.[72]

2.4.2 *In vitro* transcription and translation and preparation of αHL mutants

αHL monomers were prepared by the *E. coli in vitro* transcription and translation (IVTT) system (*E. coli* T7 S30 Extract System for Circular DNA, Cat #L1130, Promega). Before use, the T7 S30 extract provided in the kit was treated with 1 μL rifampicin (1 μg mL^{-1}, final concentration) to suppress transcription by *E. coli*

RNA polymerase. A standard reaction comprised: DNA plasmid mixture (< 4 μg, DNA plasmids in a 7:1 ratio of background: mutant), amino acid mixture without methionine (5 μL, as supplied in the kit), S30 premix without amino acids (20 μL, as supplied in the kit), [35S]methionine (2 μL) and T7 S30 extract, circular (15 μL, as supplied in the kit). The reaction mixture was supplemented with nuclease-free water to bring the final volume to 50 μL. The mixture was incubated at 37°C for 1 h.

After the incubation period, rabbit red blood cell membranes (rRBCm) (2 μL, $\sim$1 mg protein/mL) were added to the reaction mixture to induce heptamerization of the αHL monomers. The mixture was then incubated at 37°C for another 1 h.

After the second incubation period, MBSA buffer (1 mL; 3-morpholinopropane-1-sulfonic acid (10 mM), NaCl (150 mM), bovine serum albumin (1 mg/mL), pH 7.4) was added to the reaction mixture. The mixture was centrifuged at 12000 RCF for 10 min at 4°C. The supernatant was removed, and the pellet was solubilized at room temperature with 2 $\times$ Laemmli sample buffer (25 μL) which had been supplemented with 2-mercaptoethanol. Before loading the samples, a 5.5% SDS/PAGE gel was pre-run with 1 $\times$ Tris-Glycine SDS (TGS) running buffer containing DTT (2 mM) and sodium thioglycolate (1 mM). The resuspended pellets were then loaded onto the gel and electrophoresed at 70 V overnight (13 h).

αHL heptamers composed of different numbers of mutant subunits were separated in the gel based on their different electrophoretic mobilities which were determined by the number of octa-aspartate (D8) tails. Hence, the top band corresponded to homoheptamers WT_7, the second band corresponded to $(mutant)_1WT_6$ and so on, with consecutive bands having a WT subunit replaced with a mutant subunit. In this work, the second band from the top contained the desired protein pore.

To extract the protein pores, the gel was first dried under vacuum onto Whatman 3MM filter paper for 5 h at 50 °C. The dried gel was then exposed to photographic film

(Kodak Bio Max MR autoradiography film) for 8 h and the developed film was used to locate the target protein bands in the gel. The desired protein bands were excised and rehydrated in TE buffer (300 μL; Tris·HCl (10 mM), ethylenediaminetetraacetic acid (EDTA) (1 mM), pH 8.0) for 1 h at room temperature. The paper was then removed, and the gel was crushed with a plastic pestle. The resulting suspension was filtered through a 0.2 μm hydrophilic membrane filter (Proteus Mini Clarification Spin Column, Generon). The filtrate was stored in 10 μL aliquots at −80 °C.

2.4.3 Single-channel electrical recordings

A single αHL pore was inserted into the bilayer. DTNB or OPSSG was introduced from the *cis* compartment and DTT from the *trans* compartment. Pre-weighed samples of solid DTT and DTNB were prepared and stored at −20 °C. Before an experiment, a buffer was added to generate a 500 mM DTT stock and a 20 mM DTNB stock, which were kept on ice. The experiments were conducted at different pH values: pH 8.0, 8.4 and 8.5 (2 M KCl, 100 mM HEPBS, 20 μM EDTA), and 8.8, 9.0 and 9.5 (2 M KCl, 50 mM AMPSO, 20 μM EDTA). All reactions were carried out with $[\mathrm{DTNB}]_{cis}$ = $[\mathrm{DTT}]_{trans}$ = 50 μM, with the exception of Cys-113 where $[\mathrm{DTNB}]_{cis}$ = $[\mathrm{DTT}]_{trans}$ = 200 μM, and Cys-123 where $[\mathrm{DTNB}]_{cis}$ = 100 μM and $[\mathrm{DTT}]_{trans}$ = 50 μM.

Ionic currents were recorded by using a patch clamp amplifier (Axopatch 200B, Axon Instruments), and filtered with a low-pass Bessel filter (80 dB/decade) with a corner frequency of 10 kHz. Signals were digitized with a Digidata 1320A digitizer (Molecular Devices) at an acquisition frequency of 50 kHz. The current traces were processed with Clampfit 10.7 (Molecular Devices). Current traces were idealized by using Clampfit 10.3 (Molecular Devices). The idealized data were analyzed with QuB 2.0 software (`www.qub.buffalo.edu`).[73] Dwell time analysis and rate constant determinations were performed by using the maximum interval likelihood algorithm of QuB.[74]

2.4.4 Molecular dynamics simulations

Because the thiol-disulfide interchange reaction occurs only within the β-barrel of the αHL nanopores, a minimal β-barrel model was constructed.[75] In this model, residues Glu-111 to Asn-123, which were in or close to the β-barrel region were kept. All residues outside the β-barrel region were removed. To further reduce model complexity, all membrane-facing sidechains were removed. Six systems, αHL-TNB or αHL-DTT adducts at positions 113, 117, or 121, were prepared.

Force field parameters for the cysteine residues carrying TNB or DTT molecules were obtained from the GAFF force field.[76,77] For the modified residues, partial charges were determined with the Restrained Electrostatic Potential (RESP) method at the HF/6-31G* level using Gaussian16 and Antechamber.[76–79] An αHL-DTT adduct at position 113 in truncated pore is shown as an example in Figure 2.6.

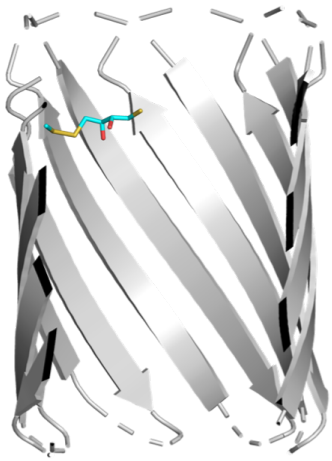


Figure 2.6: Model pore system setup. A αHL-DTT adduct (cyan) at position 113 in the truncated pore is (grey) shown as an example. Reprinted from Reference[14] with permission.

Molecular dynamics (MD) simulations were performed with GROMACS (v2019.2).[80] For each system, the nanopore was modeled with the AMBER99SB-ILDN force field[81] and inserted into a cubic box of TIP3P water molecules.[82] The system was neutralized and 0.15 M NaCl was added. The system was energy minimized using the steepest

descent algorithm until the maximum force was below 1000 kJ mol^{-1} nm^{-1}. Positional restraints of 1000 kJ mol^{-1} nm^2 were subsequently applied to the protein backbone atoms. Three independent runs were performed with random initial velocities generated according to the Maxwell–Boltzmann distribution at 298 K. The system was equilibrated under a constant volume and temperature (NVT) ensemble at 298 K (100 ps, 2 fs timestep), followed by equilibration under a constant pressure and temperature (NPT) ensemble at 1 bar and 298 K (100 ps, 2 fs timestep). Production simulations were performed using the NPT ensemble at 1 bar and 298 K (200 ns, 2 fs timestep). An electric field of 0.008 V nm^{-1} (equivalent to $\sim$50 mV across the membrane) was applied in the Z direction across the simulation box. The temperature of the system was maintained at 298 K using the V-rescale thermostat.[83] The pressure was controlled by the Parrinello-Rahman barostat at 1.0 bar, with an isothermal compressibility of 4.5 $\times$ 10^{-5} bar^{-1}.[84] All simulations were performed with three-dimensional periodic boundary conditions. Long-range electrostatics was described with the Particle Mesh Ewald (PME) algorithm.[85,86] All bond lengths involving hydrogen atoms were constrained using the LINCS algorithm.[87] Structures and input files are available on GitHub at `https://github.com/duartegroup/Current-blockade-in-nanopore`.

2.5 Results and discussion

I first investigated the reactivity of individual cysteine footholds at the single-molecule level using a nanoreactor approach. Single-cysteine mutations were made at positions 113, 115, 117, 119, 121 or 123. Nanoreactors containing one mutant subunit were prepared, and named $(M113C)_1WT_6$, $(T115C)_1WT_6$, $(T117C)_1WT_6$, $(G119C)_1WT_6$, $(N121C)_1WT_6$, and $(N123C)_1WT_6$. Each nanoreactor was inserted into a planar lipid bilayer for electrical recording.

In the presence of small-molecule disulfide and thiol reagents, spatially separated by the lipid bilayer, thiol-disulfide interchange reactions occurred at the cysteine site in a cycle of three steps (Figure 2.7A). In Step 1, a disulfide molecule (RSSR') entered from the *cis* side to react with the cysteine side chain (state 1) to generate an αHL-SR' mixed disulfide (state 2). In Step 2, the mixed disulfide was attacked by DL-dithiothreitol (DTT) introduced from the *trans* side to generate an αHL-DTT mixed disulfide (state 3), which then self-cyclized in Step 3 (Figure 2.7B). These thiol-disulfide interchange reactions were monitored by changes in the ionic current passing through the αHL nanoreactor under an applied transmembrane electric potential; the current reflected the bond-making and bond-breaking events in real-time (Figure 2.7C). Rate constants were obtained at different pH values from the mean lifetimes of each current state and used to calculate the pK_a values of the cysteine thiols and the pH-independent rate constants of the thiolates for attack on the disulfide.

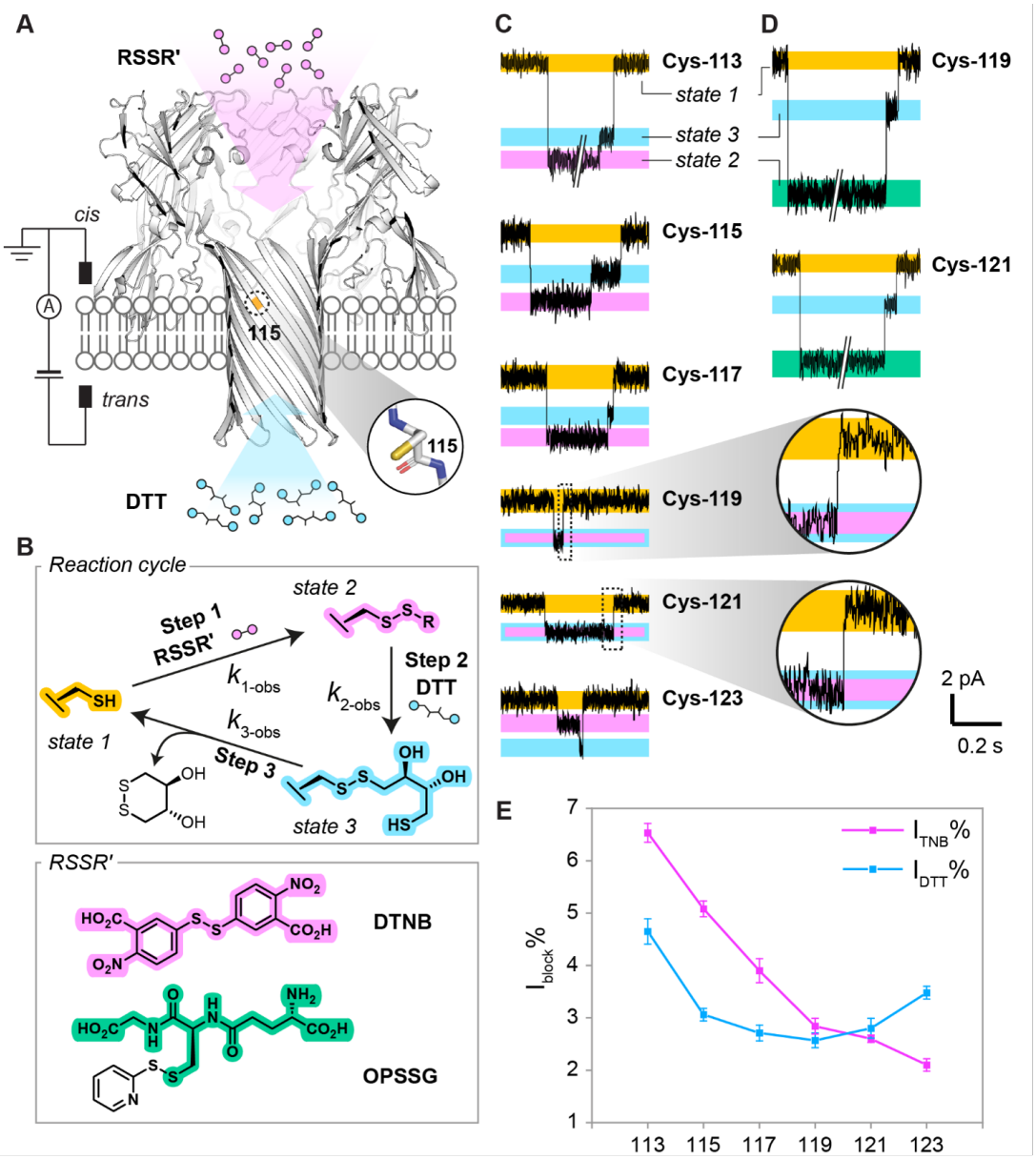


Figure 2.7: Single-molecule approach to the determination of pK_a values for cysteine residues. (A) Reactants RSSR' and DTT, spatially separated by the lipid bilayer, were introduced from the *cis* and *trans* compartments respectively. Cys-115 is shown as an example of a residue with a reactive side chain. (B) A thiol-disulfide interchange reaction cycle is used for pK_a determination. In each experiment, either DTNB or OPSSG was introduced into the cis compartment. (C) Single-channel current recordings depicting complete reaction cycles with DTNB for each single-cysteine nanoreactor. (D) Reaction cycles for OPSSG. The levels in C and D were colour-coded: free cysteine side chain (yellow),αHL-TNB (pink),αHL-SG (green) andαHL-DTT (blue). For Cys-119 and Cys-121, theαHL-DTT level could be discriminated from theαHL-SG level, but not from theαHL-TNB level (pink with blue outline). (E) Percentage current blockades (I_{block}%) forαHL-TNB andαHL-DTT for positions 113 to 123. Adapted from Reference[14] with permission.

2.5.1 Current blockades and spatial blockades of αHL-adducts

To monitor the three-step reaction cycle repeatedly at individual cysteine sites (*i.e.*, Cys-113, Cys-115, Cys-117, Cys-119, Cys-121 and Cys-123), I first treated individual cysteine nanoreactors with 5,5'-dithiobis(2-nitrobenzoic acid) (DTNB) and DTT (Figure 2.7C). Current blockades of the mixed disulfides (I_{TNB}% or I_{DTT}%) were calculated as a percentage of the open pore current (state 1). I found that the αHL-TNB and αHL-DTT adducts formed at positions 119 and 121 were indistinguishable in terms of current blockade ($I_{TNB}\% = I_{DTT}\%$). To resolve the current levels of αHL-SR' and αHL-DTT at Cys-119 or Cys-121, I treated $(G119C)_1WT_6$ or $(N121C)_1WT_6$ with orthopyridyl glutathionyl disulfide (OPSSG) and DTT. The αHL-SG adduct blocked significantly more current than the αHL-DTT adduct, allowing current level resolution (Figure 2.7D).

The percentage current blockade of αHL-TNB (I_{TNB}%) and αHL-DTT (I_{DTT}%) adducts varied at the six positions tested along the β-barrel (Figure 2.7E and Figure 2.8A). I_{TNB}% showed a gradual decrease from position 113 to 123, whereas I_{DTT}% exhibited a U-shaped behaviour with a minimum at position 119. This was accompanied by a switch in relative magnitude (*i.e.*, $I_{TNB}\% > I_{DTT}\%$ at positions 113-117; $I_{TNB}\% = I_{DTT}\%$ at positions 119 and 121; $I_{TNB}\% < I_{DTT}\%$ at position 123). This shows that current blockade by molecules within an αHL pore is not simply determined by molecular mass.

Spatial blockades of the αHL-TNB or αHL-DTT adducts at Cys-113, Cys-117, and Cys-121, defined as the end-to-end distance of the adduct divided by the radius of the unreacted pores at the cysteine position, were calculated over 600 ns of cumulative MD simulations. The spatial blockade was calculated as the end-to-end distance of an adduct divided by the radius of the unreacted pore at the cysteine positions in the unreacted pores (r_{113} = 5.8 Å, r_{117} = 5.5 Å, r_{121} = 6.4 Å) and labelled as

d_{TNB}% and d_{DTT}%. The radius of the unreacted pore at each position was calculated using the HOLE program.[88–90] A positive linear correlation was found between Δd% (d_{TNB}% – d_{DTT}%) and ΔI% (I_{TNB}% – I_{DTT}%) (R^2 = 0.98, Figure 2.8C), suggesting a strong correlation between the relative spatial blockade and the relative current blockade.

The αHL-DTT adducts have multiple conformational states, categorized by the end-to-end distances of the adducts (d_{EE}) (Figure 2.9). The conformations of the αHL-DTT adducts were predominantly stabilized by hydrogen bonding (HB) interactions between suitably oriented hydroxyl groups of DTT and neighbouring residues. For example, at position 113, an HB interaction between a hydroxyl group of DTT and Thr-145 on the adjacent strand (occupancy = 24%) resulted in an extended conformation of αHL-DTT (d_{EE} = 10 Å), whereas an HB interaction with Glu-111 on the same strand (occupancy = 10%) led to a folded conformation (d_{EE} = 4.9 Å). At positions 117 and 121, the αHL-DTT adduct adopted an extended conformation with both hydroxyl groups pointing towards the water-filled channel (d_{EE} = 7.5 Å) with HB interactions with the protein backbone on the adjacent strand: Ser-141 at position 117 (occupancy = 5%), and Gly-122 at position 121 (occupancy = 10%). In comparison, the αHL-TNB adducts had only one major state (Figure 2.10). This was attributed to the rigidity of the benzene ring, resulting in conformational flexibility mainly coming from the cysteine side chain. An HB interaction was found between the αHL-TNB adduct at position 113 and Thr-145 on the adjacent strand (occupancy = 18%), while no significant HB interactions were found for adducts at positions 117 and 121.

A

Cysteine position	Experimental (I_{block}%)		Computational (d_{EE}%)	
	I_{TNB}%	I_{DTT}%	d_{TNB}%	d_{DTT}%
Cys-113	6.53 ± 0.18	4.65 ± 0.24	15.3 ± 0.1	13.2 ± 0.7
Cys-115	5.08 ± 0.15	3.06 ± 0.12	/	/
Cys-117	3.90 ± 0.23	2.71 ± 0.15	15.3 ± 0.1	13.9 ± 0.3
Cys-119	2.84 ± 0.15	2.57 ± 0.14	/	/
Cys-121	2.60 ± 0.07	2.80 ± 0.19	12.3 ± 0.1	13.4 ± 0.2
Cys-123	2.10 ± 0.12	3.48 ± 0.12	/	/

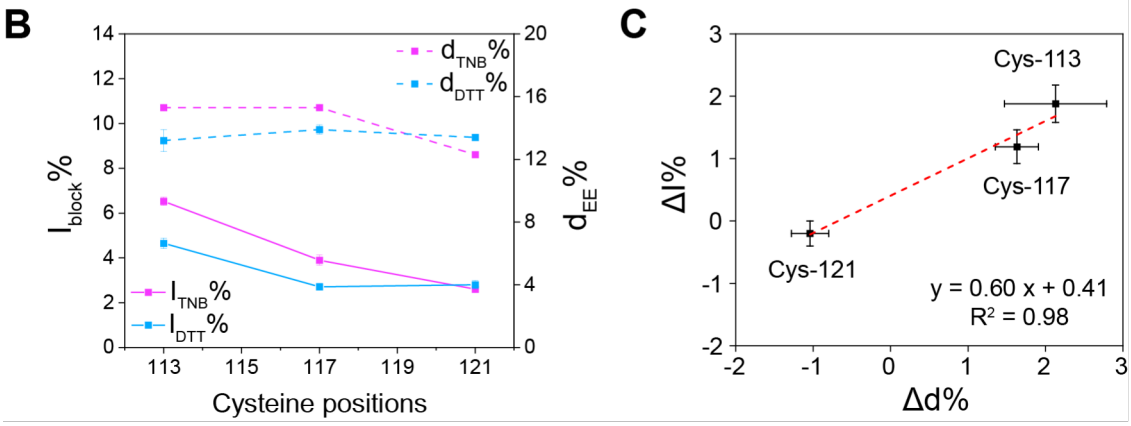


Figure 2.8: Relationship between current blockade and spatial blockade of the adducts. (A) Summary of percentage current blockades (I_{block}%): I_{TNB}% and I_{DTT}%) and end-to-end distances of the αHL-TNB and αHL-DTT adducts at positions 113, 117 and 121 (d_{EE}%: d_{TNB}% and d_{DTT}%) as a percentage of the radius at the cysteine position of the unreacted pore. For each adduct, the values were calculated over 600 ns of cumulative MD simulations. (B) Comparison between I_{block}% (solid line) and d_{EE}% (dashed line). (C) A linear correlation between Δd% (d_{TNB}% – d_{DTT}%) and ΔI (I_{TNB}% – I_{DTT}%). Reprinted from Reference[14] with permission.

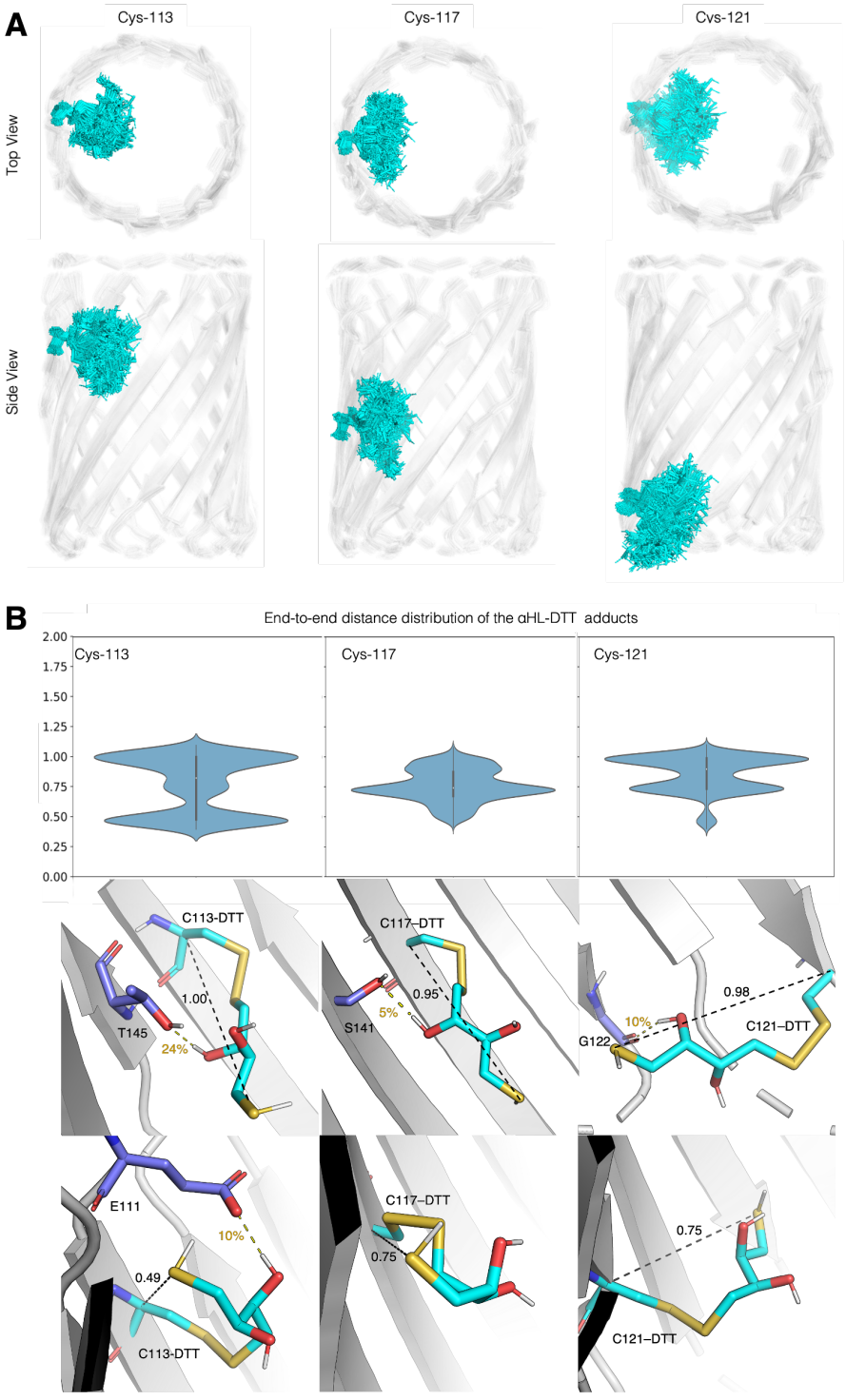


Figure 2.9: Conformations of the αHL-DTT adducts at positions 113, 117 and 121. (A) Overlap of all conformations of the αHL-DTT adducts (cyan) over 600 ns of cumulative MD simulations. (B) Top: Probability density of the end-to-end distance (d_{EE}) of an αHL-DTT adduct, calculated as the distance between the Cα atom of the cysteine residue and the terminal sulfur atom of the αHL-DTT adduct. Bottom: Two most populated conformations of the αHL-DTT adducts (cyan). d_{EE} is indicated with a black dashed line. Hydrogen bonding to a neighbouring residue (purple) and its occupancy is indicated in yellow. Reprinted from Reference[14] with permission.

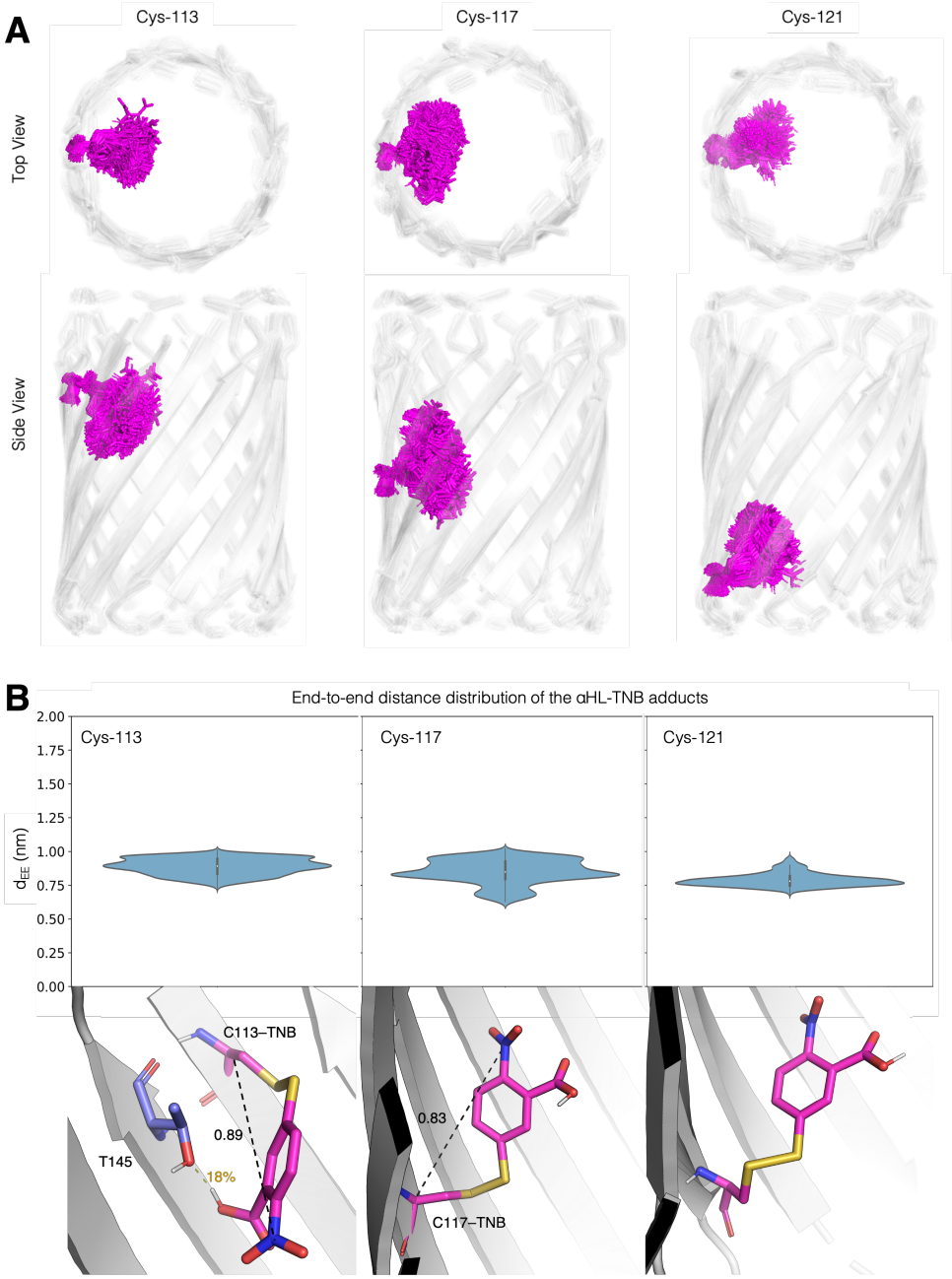


Figure 2.10: Conformations of the αHL-TNB adducts at positions 113, 117 and 121. (A) Overlap of all conformations of the αHL-TNB adducts (magenta) over 600 ns of cumulative MD simulations. (B) Top: Probability density of the end-to-end distance (d_{EE}) of an αHL-TNB adduct, calculated as the distance between the Cα atom of cysteine residue and the terminal nitrogen atom of the αHL-TNB adduct. Bottom: The most populated conformation of the αHL-TNB adducts (magenta). d_{EE} is indicated with a black dash. Hydrogen bonding to a neighbouring amino acid (purple) and its occupancy is indicated in yellow. Reprinted from Reference[14] with permission.

2.5.2 Single-molecule reactivity profiling of individual cysteine residues

Similar to the foothold thiolates involved in chemical stepping,[70] the nanoreactor cysteine thiolate acted both as the nucleophile (Step 1) and the leaving group (Step 3) during a reaction cycle. To estimate the reactivity of each cysteine thiolate, reaction cycles were recorded using DTNB and DTT at a minimum of three pH values between pH 8.0 and 9.5 using either HEPBS or AMPSO as the buffering agent. Given that the pK_a of a cysteine depends on its environment within a protein,[58] an increase in the pH during a reaction could in principle alter the thiol pK_a by (1) deprotonating neighbouring ionizable amino acid side chains (if present); (2) eliciting protein conformational changes. I expect that the pK_a values of individual cysteine thiols are minimally affected within the tested pH range (8.0 to 9.5), because (1) the inward-facing residues within the β-barrel either are non-ionizable or require a higher pH to deprotonate (*e.g.*, serine and threonine ($\mathrm{p}K_a > 13$)[91,92] or lysine ($\mathrm{p}K_a > 10.5$)[93,94]); (2) αHL nanopore is stable up to pH 11, with unfolding of the cap domain occurring at pH 12.[52]

The observed rate constants for Step 1 or Step 3 (k_{i-obs}, i = 1 or 3) were derived from the mean lifetimes of the associated chemical states ($< \tau_i >$, i = 1 or 3). For the bimolecular Step 1, $k_{1-obs} = 1/(< \tau_1 >[\mathrm{DTNB}]_{cis})$, while for the unimolecular Step 3, $k_{3-obs} = 1/< \tau_3 >$. To ensure there was no buffer catalysis, the reaction cycles in $(\mathrm{T115C})_1\mathrm{WT}_6$ were recorded at pH 8.8, buffered by both HEPBS and AMPSO. The intramolecular cyclization is an unimolecular reaction, the rate of which is less prone to experimental error than the bimolecular Step 1. A two-tailed test was performed to compare the difference in observed rate constants. The p-value was 0.64 (Figure 2.11). Hence, any observed differences between the two sets of observed rate constants obtained under different buffers are not statistically significant, *i.e.*, no buffer catalysis is observed.

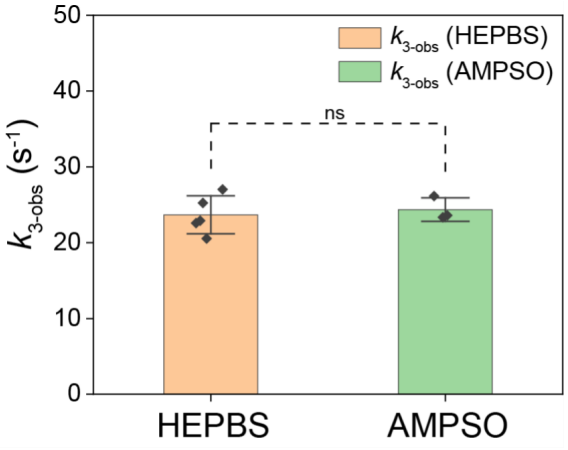


Figure 2.11: Observed rate constants of intramolecular cyclization of αHL-DTT measured with two buffering agents. Intramolecular cyclization of the αHL-DTT adduct at Cys-115 was investigated at pH 8.8 with either HEPBS or AMPSO. k_{3-obs} data for Step 3 are plotted; error bars represent the standard deviations of each set of data. Reprinted from Reference[14] with permission.

The pH-independent rate constants of the thiolate (k_i, i = 1 or 3) and the pK_a values of the respective thiols for each nanoreactor were determined by fitting $1/k_{i-obs}$ on [H^+] using linear regression. In a thiol-disulfide interchange reaction, a thiolate reacts with a disulfide by nucleophilic substitution. The Henderson-Hasselbalch equation gives the ratio of the deprotonated thiolate to the conjugate acid thiol, which is dependent on the pK_a of the thiol and the pH of the reaction medium. Therefore, a linear regression of the observed rate across different pH conditions allows the determination of the pK_a of a thiol and the pH-independent rate constants of the deprotonated thiolate k_i for thiol-disulfide interchange (Eq 2.3).

$$\frac{1}{k_{i-obs}} = \frac{1}{k_i} \cdot (1 + \frac{[H^+]}{K_a}) \tag{2.3}$$

To calculate k_i and pK_a values, I performed linear regression fitting of $1/k_{i-obs}$ on [H^+] for each thiol. The [H^+] values were calculated from the pH values of the reaction medium. The observed rates for either Step 1 or Step 3 (k_{i-obs}, i = 1 or 3) were derived from the mean lifetime of the corresponding chemical state, $< \tau_i >$ (i = 1 or 3). There were ten datasets comprising [H^+] and $1/k_{i-obs}$ values: six datasets for

the cysteine thiol of Cys-113, Cys-115. Cys-117, Cys-119, Cys-121, and Cys-123 were involved in Step 1 and four datasets for the free thiols in the αHL-DTT adducts of Cys-113, Cys-115, Cys-117, and Cys-123 involved in Step 3. For each dataset, linear regression analysis with 3-fold cross-validation was performed (Figure 2.12).

Specifically, a dataset was split into 3 groups randomly (labelled A, B and C). To fit Model 1, group C was taken as the test dataset, while the remaining two groups A and B were taken as the training dataset. Similarly, to fit Model 2, group B was taken as the test dataset, while the remaining two groups A and C were taken as the training dataset. To fit Model 3, group A was taken as the test dataset, while the remaining two groups B and C were taken as the training dataset. This means that each data point in the dataset was used in a test dataset once and used for linear regression model training twice.

To fit each model, the input values $[H^+]$, denoted as X_{train}, were fitted to the target values $1/k_{i-obs}$, denoted as y_{train}, using least square linear regression (Eq 2.4). This yielded a gradient and a y-intercept, denoted as m_j and b_j respectively. To validate the model on the test dataset, predicted target values were calculated (Eq 2.5) and then compared to the actual target values. The coefficient of determination ($R^2 > 0.7$) and root mean squared error were used as the accuracy metrics for model evaluation.

$$y_{train} = m_j \cdot X_{train} + b_j (j = 1, 2, 3) \tag{2.4}$$

$$y_{pred} = m_j \cdot X_{test} + b_j (j = 1, 2, 3) \tag{2.5}$$

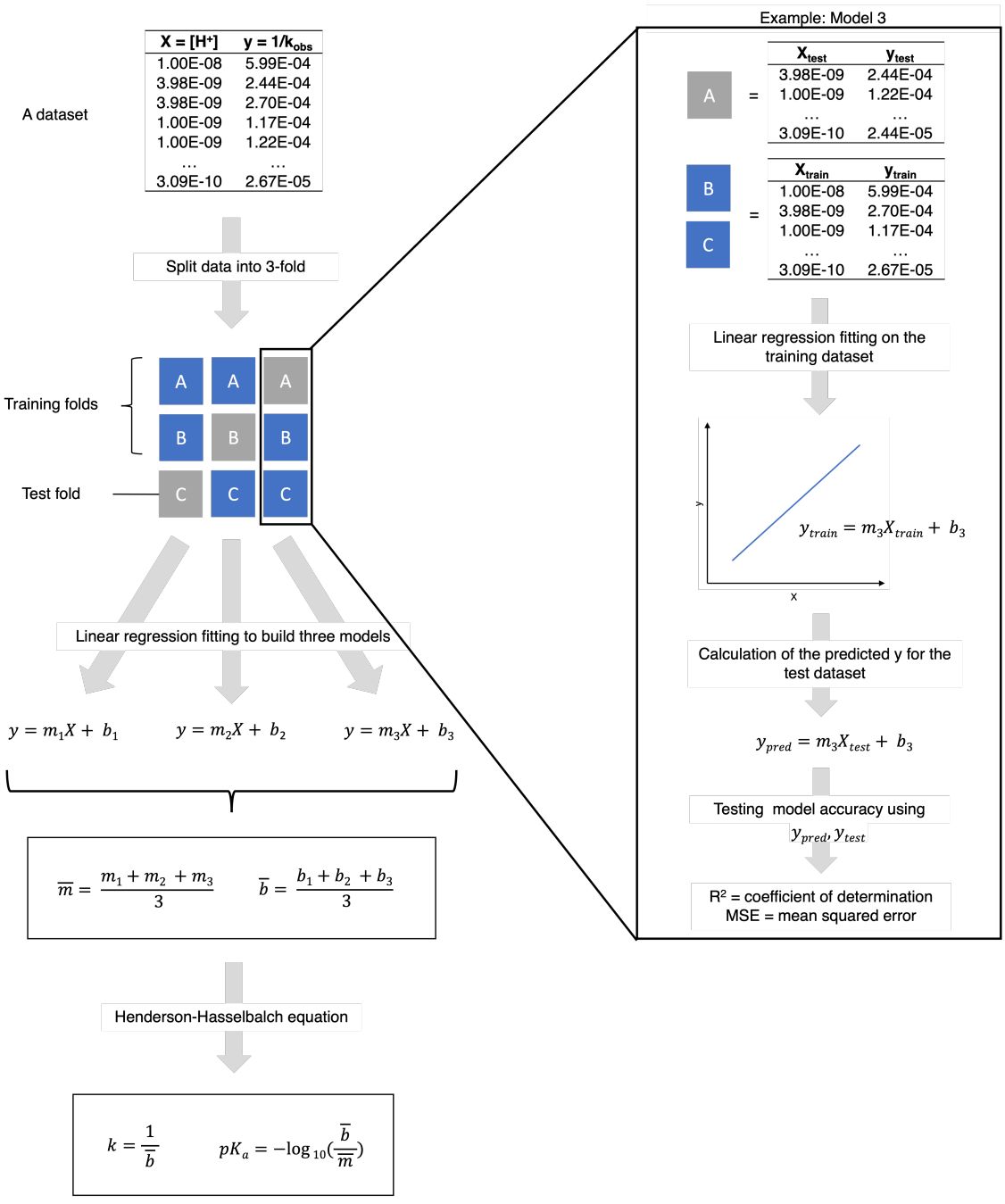


Figure 2.12: Linear regression analysis with 3-fold cross-validation and subsequent calculations using the Henderson-Hasselbalch equation. Reprinted from Reference[14] with permission.

Cysteine position	Gradient (10^5 M s)	y-intercept (10^{-5} M s)	R^2 score[a]	RMSE (10^{-3})[b]
Cys-113	5.4 ± 0.4	36 ± 3	0.89	0.50
Cys-115	0.78 ± 0.04	3.3 ± 0.2	0.76	0.12
Cys-117	1.0 ± 0.1	4.9 ± 0.4	0.92	0.10
Cys-119	0.45 ± 0.01	1.8 ± 0.1	0.81	0.07
Cys-121	0.66 ± 0.04	1.7 ± 0.2	0.92	0.07
Cys-123	11 ± 1	16 ± 1	0.84	1.8

Table 2.1: Summary of gradients, y-intercepts, and accuracy metrics of the linear regression analysis for Step 1. Adapted from Reference[14] with permission.

Cysteine position	Gradient (10^5 M s)	y-intercept (10^{-5} M s)	R^2 score[a]	RMSE (10^{-3})[b]
Cys-113	66 ± 6	602 ± 14	0.85	7
Cys-115	243 ± 14	298 ± 42	0.92	20
Cys-117	87 ± 3	516 ± 32	0.89	10
Cys-123	103 ± 3	518 ± 3	0.89	10

Table 2.2: Summary of gradients, y-intercepts, and accuracy metrics of the linear regression analysis for Step 3. Adapted from Reference[14] with permission.

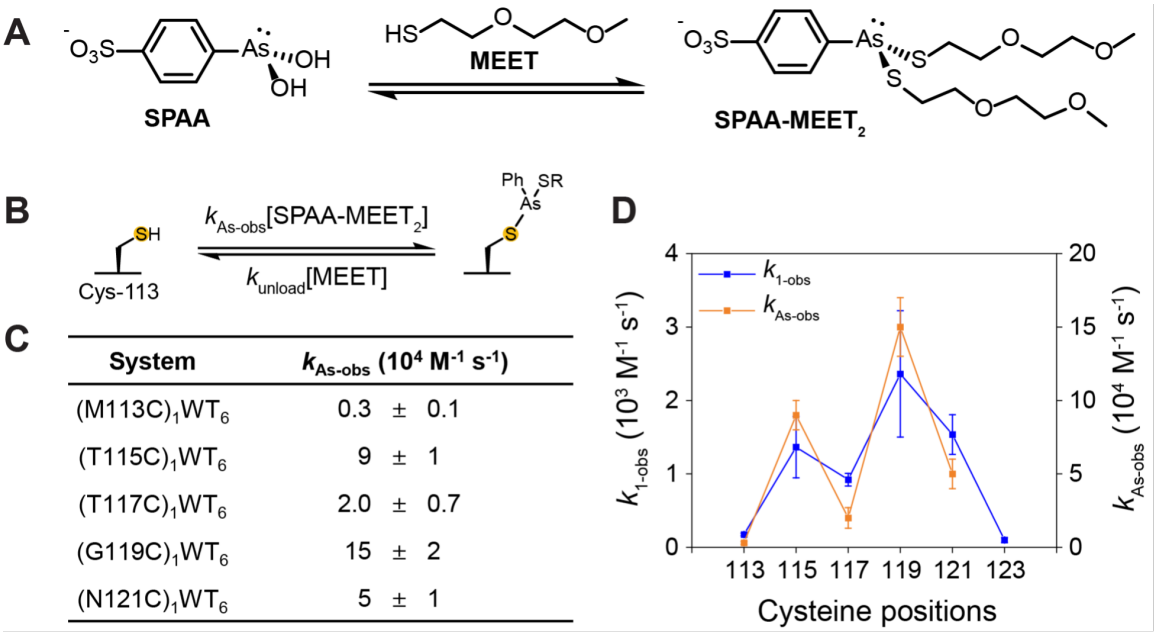
The pK_a values of the six cysteines ranged from 9.17 (Cys-113) to 9.85 (Cys-123) (Figure 2.13). The pH-independent rate constants of the cysteine thiolates for the reaction with DTNB (k_1) (Cys-113, Cys-115, Cys-117, Cys-119, Cys-121, or Cys-123) ranged from $2.76 \pm 0.21 \times 10^3$ M^{-1} s^{-1} to $5.89 \pm 0.52 \times 10^4$ M^{-1} s^{-1} whereas those for intramolecular cyclization of the αHL-DTT adducts (k_3) (Cys-113, Cys-115, Cys-117 or Cys-123) ranged from 174 ± 38 s^{-1} to 335 ± 48 s^{-1}. The errors of the cysteine pK_a values were less than 0.1, suggesting good data reproducibility despite various sources of experimental error (*e.g.*, determinations of pH, concentration, and temperature) and pore-to-pore variation. In comparison to the ensemble pH-independent rate constants measured for cysteines within glutathione (Figure 2.13) with DTNB, the k_1 values obtained from single-molecule experiments at Cys-115, Cys-117, Cys-119 and Cys-121 were comparable ($\sim 10^4$ M^{-1} s^{-1}), while the k_1 values at Cys-113 and Cys-123 were around 10-fold lower.

System	Step 1: thiol-disulfide interchange between cysteine thiolate and DTNB, k_1 (10^3 M^{-1} s^{-1}) [a]	pK_a (cysteine)
$(M113C)_1WT_6$	2.76 ± 0.21	9.17 ± 0.04
$(T115C)_1WT_6$	30.0 ± 1.5	9.37 ± 0.03
$(T117C)_1WT_6$	20.0 ± 2.2	9.31 ± 0.05
$(G119C)_1WT_6$	56.2 ± 9.3	9.41 ± 0.07
$(N121C)_1WT_6$	58.9 ± 5.2	9.59 ± 0.05
$(N123C)_1WT_6$	6.30 ± 0.50	9.85 ± 0.04
System	**Ensemble experiments: thiol-disulfide interchange between thiolate and DTNB (10^3 M^{-1} s^{-1})**	**pK_a (thiol)**
Small-molecule thiols [b]	180 – 2890 [b]	9.33 – 9.80 [b]
Glutathione thiols [c]	35 – 40 [c]	/
System	**Step 3: self-cyclization of αHL-DTT, k_3 (s^{-1}) [a]**	**pK_a (αHL-DTT)**
$(M113C)_1WT_6$	174 ± 38	9.07 ± 0.10
$(T115C)_1WT_6$	335 ± 48	9.91 ± 0.07
$(T117C)_1WT_6$	194 ± 12	9.23 ± 0.03
$(N123C)_1WT_6$	188 ± 15	9.29 ± 0.04

Figure 2.13: pH-independent rate constants and pK_a values of nucleophilic thiols in thiol-disulfide interchange reactions in single-molecule experiments and ensemble experiments. Reprinted from Reference[14] with permission.

These results suggest that the nucleophilicity of cysteine residues in single-cysteine nanoreactors is not linearly correlated with their pK_a values. Previous studies of thiol-disulfide interchange reactions between small-molecule thiols and disulfides revealed a Brønsted relation wherein the pH-independent rate constant of the thiolate correlated positively with the pK_a of the nucleophilic thiol but negatively with the pK_a of the leaving group thiol.[95] From these studies, Brønsted coefficients of the nucleophilic thiolate were determined to be 0.4-0.5. In a protein environment, the expected Brønsted relation may be complicated by factors such as electrostatics, sterics, and solvation.[58,96–98] For example, a study of reactions between disulfides and cysteine

thiols embedded in polypeptide fragments revealed the importance of the net charge of disulfides; Brønsted coefficients for the nucleophilic polypeptide thiolates were 0 and 0.8 for the negatively charged glutathione disulfide and the positively charged cystamine respectively.[99] Given the small dataset and the narrow pK_a range covered in the present work, it is unrealistic to extrapolate a free-energy relationship. Nevertheless, the poorest nucleophiles corresponded to the cysteines with the lowest pK_a (Cys-113) and, surprisingly, the highest pK_a (Cys-123), with k_1 values approximately an order of magnitude lower than the remaining cysteines.



System	$k_{As\text{-}obs}$ (10^4 M^{-1} s^{-1})
$(M113C)_1WT_6$	0.3 ± 0.1
$(T115C)_1WT_6$	9 ± 1
$(T117C)_1WT_6$	2.0 ± 0.7
$(G119C)_1WT_6$	15 ± 2
$(N121C)_1WT_6$	5 ± 1

Figure 2.14: Observed rate constants for reactions between single-cysteine mutants and SPAA-$MEET_2$. (A) Structure of SPAA-$MEET_2$ formed in-situ. (B) The observed rate of thiol-arsenic exchange between a single-cysteine mutant (Cys-113 in $(M113C)_1WT_6$ shown here) and SPAA-$MEET_2$ obeys second-order association rates. The reverse reaction obeys second-order association rates and is directly dependent on the concentration of MEET. Experiments for the arsenic walker were performed at pH 8.0 (10 mM bis-tris propane, 2 M KCl, 100 μM EDTA). (C) Observed rate constants (k_{As-obs}) for single-cysteine mutants are tabulated. (D) The observed rate constants for an attack on DTNB (k_{1-obs}, blue) followed a similar trend to that for an attack on SPAA-MEET2 (k_{As-obs}, orange). Observed rate constants for thiol-disulfide interchange were from experiments conducted at pH 8.0 (2 M KCl, 50 mM HEPBS, 20 μM EDTA). Reprinted from Reference[14] with permission.

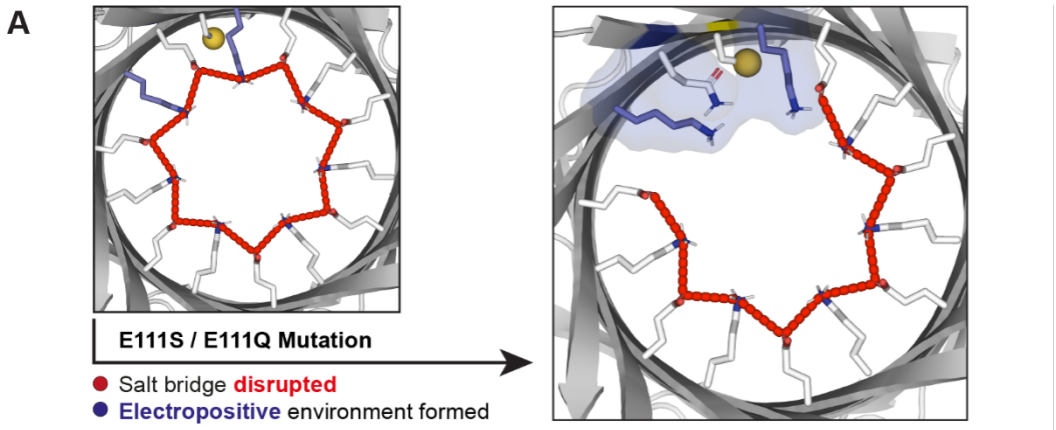
Previously, an arsenic(III)-based molecular walker was loaded onto a cysteine track by a thiol-arsenic interchange reaction between an αHL cysteine and a thiol-arsenic(III) adduct (Figure 2.14A and B).[10] The observed rate constants of walker loading (k_{As-obs})

at Cys-113, Cys-115, Cys-117, Cys-119, and Cys-121 were measured individually at pH 8.0 (Figure 2.14C). A plot of k_{As-obs} showed a similar trend to the observed rate constants for the reaction with DTNB at pH 8.0 (k_{1-obs}), with Cys-113 being the most slow reacting (Figure 2.14D). Thiol-arsenic exchange reactions were shown to occur with stereochemical inversion,[9] suggesting a collinear transition state (S-As-S) at the arsenic centre. This parallels the collinear transition state required for thiol-disulfide interchange.[100] Hence, the factors influencing the trends observed for the rate constants of the thiol-disulfide interchange (*e.g.*, the pK_a values of the cysteine thiols, the sterics of the protein environment) might have a parallel effect on thiol-arsenic exchange.

The applied potential could significantly alter the reaction course within a nanopore in certain cases, such as when it carries extensive charge (*e.g.*, a DNA oligonucleotide).[70] An increase in the applied potential can also accelerate or slow down reactions when the migration of molecules is affected by electroosmosis (β-cyclodextrin) or electrophoresis (DNA).[53] However, for small-molecule reagents such as DTNB and DTT, whose transport in and out of the pore is unhindered, it is unlikely that the applied potential will greatly influence reaction rates. This is because there would not be an 'accumulation' of reagents within the pore, and the overall concentration of these reagents within the pore would be close to that in the external medium even in the presence of an applied potential. Here, I work at an applied potential of +50 mV, which drops across the barrel of the αHL pore at 10 mV nm^{-1}. The energy required to orient a cysteine side chain parallel to the electric field is estimated to be less than kT (T = 293 K).[53] This energy is further reduced by the shielding effects of the high salt concentration in the buffer (*i.e.*, 2 M KCl). Hence, throughout the length of the β-barrel, both freely diffusing reagents and tethered chemical species are expected to only be weakly oriented within the electric field. Neither applied potential nor electric field drop is expected to affect the observed reaction rates.

2.5.3 Enhanced cysteine reactivity by mutating a neighbouring residue

Noting that the Cys-113 thiolate had the lowest k_1, I, together with Zhong Hui Lim, investigated how changing its surrounding protein environment could alter its reactivity. In this section, Zhong performed the experiments and then we performed data analysis together. Two mutant single-cysteine nanoreactors were generated, $(E111S\text{-}M113C)_1WT_6$ and $(E111Q\text{-}M113C)_1WT_6$, containing a Cys-113 and either an E111S or E111Q mutation in one of the seven subunits. A net positive charge from Lys-147 is obtained when the negatively charged Glu-111 is replaced with the neutral Ser-111 or Gln-111. Consequently, the Lys-147 residues (on the mutant subunit and adjacent wild-type subunit) are no longer salt-bridged to the absent Glu-111 residue (Figure 2.15A).



B

Residue 109	111	113	115	System	$k_{1\text{-obs}}$ (M^{-1} s^{-1})			$k_{3\text{-obs}}$ (s^{-1})		
T	E	C	T	$(M113C)_1(WT)_6$	474	±	92	34.1	±	4.9
T	S	C	T	$(E111S\text{-}M113C)_1(WT)_6$	1650	±	230	46.5	±	2.7
T	Q	C	T	$(E111Q\text{-}M113C)_1(WT)_6$	2610	±	440	64.0	±	6.1

Figure 2.15: Rational redesign of an engineered track. (A) Mutations E111S and E111Q are predicted to disrupt the ring of salt bridges (red dashed lines) and liberate lysine residues (coloured in blue). (B) Observed rate constants for nucleophilic attack by Cys-113 on DTNB (k_{1-obs}) and intramolecular cyclization of the DTT adducts (k_{3-obs}). Dwell-time analysis and rate constant estimations were performed by using the maximum interval likelihood algorithm of QuB software.[74] Conditions: 2 M KCl, 50 mM HEPBS, pH 8.5, 20 μM EDTA, 20 ± 1 °C. Adapted from Reference[14] with permission.

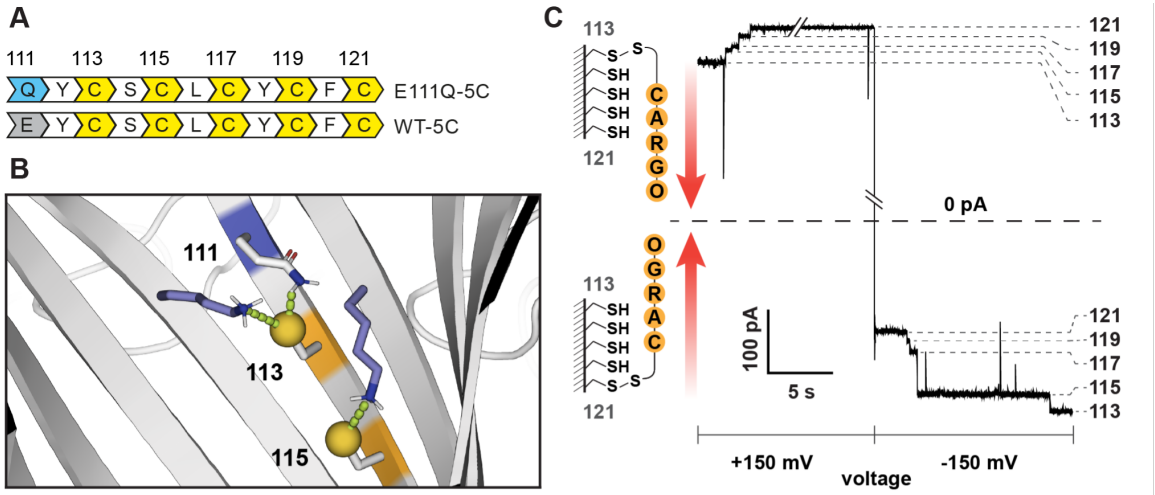
Both (E111S-M113C)$_1$WT$_6$ and (E111Q-M113C)$_1$WT$_6$ gave a stable steady-state current without frequent gating under ±150 mV, which was essential for the observation of chemical stepping of a DNA cargo molecule. Both mutations led to an increase in the observed rate constants of the thiol-disulfide interchange (Step 1, k_{1-obs}) between Cys-113 and DTNB by 3.5-fold (E111S) or 5.5-fold (E111Q) at pH 8.5 (Figure 2.15B). Similarly, the observed rate constant of DTT cyclization (Step 3, k_{3-obs}) at pH 8.5 increased by 1.4-fold (E111S) or 1.9-fold (E111Q) (Figure 2.15B). The electropositive local environment, contributed by the Lys-147 residues, and hydrogen bond donation from Ser-111 or Gln-111 residues were hypothesized to stabilize the Cys-113 thiolate (*i.e.*, lower the pK_a of Cys-113). The increased probability of thiol deprotonation increases k_{1-obs}.[101,102] The greater rate enhancements observed for the E111Q variant implied that a longer side chain enables more favourable interactions for thiol-disulfide interchange rather than steric hindrance. The rate enhancement for DTT cyclization (k_{3-obs}) is attributed to (a) a lower pK_a of the αHL-DTT thiol producing increased thiol deprotonation, and (b) a lower pK_a of the Cys-113 thiol resulting in the thiolate becoming a better leaving group.[95] The rate enhancement for DTT cyclization was lower than that for the preceding Step 1. This could be due to the αHL-DTT adduct extending further into the lumen of the pore, away from the Lys-147 and Ser-111 or Gln-111 residues as compared to the Cys-113 thiol.[102]

2.5.4 Accelerated chemical stepping along engineered tracks

After successfully increasing the reactivity of Cys-113 in the small-molecule reaction, Gln-111 was incorporated into nanoreactors containing a five-cysteine track. In this section, Zhong performed the experiments and then we performed data analysis together. Using PyMOL modelling of mutants based on the X-ray structure of αHL (PDB ID 7AHL),[32] favourable hydrogen-bonding and ionic interactions within the mutant environment that can stabilize the deprotonated form of either Cys-113 (NH-

S ∼3.6 Å between Lys-147 and Cys-113 or NH-S ∼2.2 Å between Gln-111 and Cys-113) or Cys-115 (NH-S ∼3.3 Å between Lys-147 and Cys-115) were identified (Figure 2.16A). These interactions were expected to lower the pK_a values of both Cys-113 and Cys-115 thiols, promoting the formation of nucleophilic thiolate and improving the leaving group ability of the cysteine thiolate. The $(\text{E111Q-5C})_1\text{WT}_6$ nanoreactor containing the engineered protein track E111Q-M113C-T115C-T117C-G119C-N121C was tested with a molecular hopper carrying a multi-adenosine 40-mer cargo (Figure 2.16A).The results were compared to the $(\text{WT-5C})_1\text{WT}_6$ nanoreactor containing the protein track M113C-T115C-T117C-G119C-N121C (Figure 2.16B). On the cysteine track, the hopper moved forward in the direction of Cys-113 to Cys-121 under +150 mV, and from Cys-121 to Cys-113 under −150 mV. Occasional backstepping could occur when the hopper moved in the opposite direction of that determined by the applied electric potential. Both forward and backward steps occur by thiol-disulfide interchange where the final foothold acts as the nucleophile, the initial foothold acts as the leaving group, and the hopper thiol as an electrophile.

Oligonucleotide translocation in $(\text{E111Q-5C})_1\text{WT}_6$ was first investigated at pH 8.5. Faster 113-to-115 stepping were observed under +150 mV ($k_{f113-115} = 0.12 \pm 0.02$ s^{-1} in $(\text{E111Q-5C})_1\text{WT}_6$ compared to 0.044 ± 0.009 s^{-1} in $(\text{WT-5C})_1\text{WT}_6$), while 115-to-113 stepping under −150 mV remained slow ($k_{f115-113} = 0.068 \pm 0.012$ s^{-1} in $(\text{E111Q-5C})_1\text{WT}_6$ compared to 0.054 ± 0.010 s^{-1} in $(\text{WT-5C})_1\text{WT}_6$) (Figure 2.16D). This was attributed to the variations in rate enhancement to differences in DNA conformational lability under opposite potentials. When the DNA is bonded to Cys-113 and oriented towards the *trans* opening of the pore at +150 mV, ∼4 phosphodiester groups lie within the electric field, which for the most part drops along the length of the β-barrel. In contrast, an oligonucleotide bonded to Cys-115 and oriented towards the *cis* opening of the pore at −150 mV would only have ∼1 phosphodiester group within the electric field.[70] Therefore, the stronger electrophoretic force acting on the



D

		(pH 9.5) $(E111Q\text{-}5C)_1WT_6$ k^f (s^{-1})	(pH 8.5) $(E111Q\text{-}5C)_1WT_6$ k^f (s^{-1})	(pH 8.5) $(WT\text{-}5C)_1WT_6$ k^f (s^{-1})	$k_{E111Q,pH\ 9.5}$ / $k_{E111Q,pH\ 8.5}$	$k_{E111Q,pH\ 8.5}$ / $k_{WT,pH\ 8.5}$
+150 mV	113-to-115	0.94 ± 0.10	0.12 ± 0.02	0.044 ± 0.009	8.0	2.7
	115-to-117	1.7 ± 0.2	0.33 ± 0.07	0.18 ± 0.04	5.3	1.9
	117-to-119	5.9 ± 0.6	1.00 ± 0.22	0.87 ± 0.18	5.9	1.2
	119-to-121	4.6 ± 0.3	0.64 ± 0.11	0.45 ± 0.06	7.2	1.4
-150 mV	121-to-119	0.74 ± 0.09	0.32 ± 0.07	0.49 ± 0.10	2.3	0.7
	119-to-117	0.51 ± 0.06	0.17 ± 0.03	0.18 ± 0.04	3.0	0.9
	117-to-115	2.1 ± 0.2	0.75 ± 0.15	0.41 ± 0.09	2.9	1.9
	115-to-113	0.12 ± 0.01	0.068 ± 0.012	0.054 ± 0.010	1.8	1.3
Average stepping rate / s^{-1}		0.57 ± 0.06	0.21 ± 0.04	0.13 ± 0.03	2.8	1.5
$<\tau_{total}>$ / s		14 ± 2	40 ± 9	59 ± 13	-	-

Figure 2.16: Acceleration of hopping rates along a modified cysteine track. (A) Primary sequences of the protein tracks investigated. (B) Rate enhancements of steps involving Cys-115 were attributed to a lowering of the pK_a of the Cys-115 thiol. The Cys-115 thiolate was stabilized by interaction with the Lys-147 residue on the adjacent subunit. (C) On the five-cysteine track, four hopping steps were observed at both +150 mV and −150 mV. The direction of the hops is indicated by the red arrows. Conditions: 2 M KCl, 50 mM HEPBS, pH 8.5, 20 μM EDTA, 20 ± 1 °C. (D) Hopping rates involving Cys-113 and Cys-115 were accelerated under positive (113-to-115, 115-to-117) and negative potentials (117-to-115) on the E111Q-5C track relative to the WT-5C track. Stepping rates for all steps were accelerated under both positive and negative potentials when the pH of the buffer was increased to pH 9.5 (50 mM AMPSO). Dwell-time analysis and rate constant estimations were performed by using the maximum interval likelihood algorithm of QuB software.[74] The average stepping rate along each track was calculated by taking the reciprocal of the average time per step (*i.e.*, average time per step = $< \tau_{total} >$ / 8). Adapted from Reference[14] with permission.

DNA under +150 mV might align the disulfide bond linking the oligonucleotide to Cys-113 with Cys-115; by contrast, the oligonucleotide bonded to Cys-115 might have more degrees of freedom under −150 mV, which reduces occupation of the collinear

conformation required for stepping to Cys-113.[100] Previous computational studies have shown that deviation of the collinear transition state from a S-S-S angle of 172° to 108° leads to an increase in activation energy of 167 kJ mol^{-1}.[103] As predicted by PyMOL modeling, a ∼2-fold increase in the rate for the 115-to-117 step under +150 mV and for the 117-to-115 step under −150 mV in $(E111Q\text{-}5C)_1WT_6$ were also observed(Figure 2.16D). The former is attributed to the Cys-115 thiolate being a better leaving group, and the latter to increased Cys-115 thiol deprotonation. All other steps occurred at a similar rate by comparison to those in $(WT\text{-}5C)_1WT_6$.

Next, oligonucleotide translocation was recorded at pH 9.5 in $(E111Q\text{-}5C)_1WT_6$. The rates of individual hopping steps at pH 9.5 increased by 5 to 8-fold under +150 mV and 2 to 3-fold under −150 mV by comparison with those measured at pH 8.5. Using the Henderson-Hasselbalch equation with the pK_a values derived for individual cysteine residues, the stepping rates were predicted to have a 4 to 6-fold increase based on increased thiol deprotonation. This was in agreement with the observed rates. Smaller rate increases for steps under a potential of −150 mV could again be attributed to the DNA having a higher degree of freedom, thereby reducing the occupation of the linear conformation required for thiol-disulfide interchange. While rate enhancement for the 115-to-113 step under −150 mV was low, the directionality of translocation was unaffected by 115-to-117 backstepping, with the rate of forward stepping ($k_{f115-113} = 0.12 \pm 0.01$) being an order of magnitude greater than backstepping ($k_{b115-117} = 0.012 \pm 0.005$) in $(E111Q\text{-}5C)_1WT_6$. Rates for other backsteps were not calculated due to insufficient numbers of events (*i.e.*, 14 backsteps out of 654 forward steps in $(E111Q\text{-}5C)_1WT_6$ at pH 9.5, 7 of which were 115-to-117 backsteps under −150 mV).

For enzymeless sequencing, the biopolymer has to be moved from one end of the track to the other repeatedly for the measurement of current blockades. The mean total

transit time ($< \tau_{total} >$) is defined as the time taken to complete a full cycle along a given track (*i.e.*, from Cys-113 to Cys-121 under +150 mV followed by Cys-121 to Cys-113 under −150 mV). At pH 8.5, the E111Q mutation reduced $< \tau_{total} >$ from 59 ± 13 s in $(\text{WT-5C})_1\text{WT}_6$ to 40 ± 9 s in $(\text{E111Q-5C})_1\text{WT}_6$. In particular, the E111Q mutation reduced the mean dwell time of the slowest step at pH 8.5 (*i.e.*, 113-to-115 under +150 mV) from 23 ± 5 s to 8.4 ± 1.7 s, accounting for 70% of the overall transit time reduction. If a uniform track with no hold-ups were desired, site-directed mutagenesis provides a means of targeting the slowest steps. A pH change from 8.5 to 9.5 further shortens $< \tau_{total} >$ to 14 ± 2 s in $(\text{E111Q-5C})_1\text{WT}_6$. Rate acceleration of the two slowest steps (*i.e.*, 113-to-115 step under +150 mV and 115-to-113 step under −150 mV) accounted for 56% of the overall transit time reduction. The average stepping rate was 0.13 ± 0.03 s^{-1} in $(\text{WT-5C})_1\text{WT}_6$ at pH 8.5, and 0.21 ± 0.04 s^{-1} and 0.57 ± 0.06 s^{-1} in $(\text{E111Q-5C})_1\text{WT}_6$ at pH 8.5 and pH 9.5, respectively (Figure 2.16D).

2.6 Conclusion

In conclusion, I established a workflow to systematically profile the reactivity of individual cysteine footholds along an engineered protein track using a single-molecule approach. I calculated the pK_a values of cysteine thiols and the pH-independent rate constants of the thiolates with a small-molecule disulfide varied. The pK_a values of the six cysteines ranged from 9.17 (Cys-113) to 9.85 (Cys-123). The pH-independent rate constants of the cysteine thiolates for the reaction with DTNB (k_1) ranged from 2.76 ± 0.21 × 10^3 M^{-1} s^{-1} to 5.89 ± 0.52 × 10^4 M^{-1} s^{-1} whereas those for intramolecular cyclization of the αHL-DTT adducts (k_3) ranged from 174 ± 38 s^{-1} to 335 ± 48 s^{-1}. I also used computational modelling to reveal a strong correlation between the relative spatial blockade and the relative current blockade of covalent adducts using a minimal

pore model.

Guided by the reactivity profile, site-specific mutagenesis was performed, and together with optimization of experimental conditions, the overall stepping rate of a DNA cargo along a five-cysteine track was accelerated. The mean total transit time for DNA cargo translocation was shortened from 59 s to 14 s. Each step in $(E111Q\text{-}5C)_1WT_6$ at pH 9.5 averaged 0.57 s^{-1}, a 4-fold increase in rate from the original system (*i.e.*, $(WT\text{-}5C)_1WT_6$ at pH 8.5). The rate-limiting step in the $(WT\text{-}5C)_1WT_6$ nanoreactor was accelerated 21-fold ($k_{f113-115} = 0.044 \pm 0.009$ s^{-1} in $(WT\text{-}5C)_1WT_6$ at pH 8.5 compared to 0.94 ± 0.10 s^{-1} in $(E111Q\text{-}5C)_1WT_6$ at pH 9.5).

These results from site-directed mutagenesis indicate that favourable changes to cysteine reactivity can be elicited by introducing the E111Q mutation. In this study, a 3.5 to 5.5-fold increase in reactivity was observed towards the small-molecule disulfide DTNB, while a 2 to 3-fold increase in reactivity was observed for selected steps in the cysteine track. All hopping steps within the $(E111Q\text{-}5C)_1WT_6$ nanoreactor were sped up by increasing the pH from 8.5 to 9.5. Without further track engineering, faster translocation could be achieved at a pH above 9.5, where all cysteine footholds are fully deprotonated. Rate enhancements by the E111Q mutation and pH increase under ±150 mV are asymmetric. I speculate that the DNA cargo adopts different conformations under opposite potentials; the collinear transition state for thiol-disulfide interchange will be accessed less often when the DNA cargo has more degrees of freedom.[100]

The kinetics of mobile molecules might also be influenced by increasing the temperature. Previously, protein pores leukocidin, OmpG and various αHL mutants were shown to be stable in single-channel recordings when the temperature was brought to more than 90 °C.[55] By this means, the binding of β-cyclodextrin to the αHL mutant $(M113N)_7$ was shown to follow the linear form of the Van't Hoff equation. Using an activation energy

of 50 kJ mol^{-1},[104] a 50 °C increase in temperature from 25 °C to 75 °C is predicted to produce an 18-fold rate acceleration for thiol-disulfide interchange. Assuming each step is accelerated 18-fold, the mean total transit time in $(E111Q\text{-}5C)_1WT_6$ would be reduced to 0.78 s. Alternatively, selenocysteines might be incorporated into the protein track in place of cysteine. Selenolate (*i.e.*, a deprotonated selenol) would act as the nucleophile and the leaving group during each hopping step. If selenium were incorporated into the biopolymer cargo, selenium could act as the electrophile for nucleophilic attack during each hopping step. Previous studies have suggested that selenols are ∼1 or ∼4 orders of magnitude more reactive than thiols as nucleophiles or electrophiles, respectively.[105] The incorporation of electrophilic selenium as the cargo handle could reduce the mean total transit time to < 1 ms, equivalent to > 1000 complete cycles across the track per second.

This system can be generalized for the translocation of biopolymers other than DNA, such as polypeptides. While the rates in this system do not so far compare to those of the helicase motors used in nanopore DNA/RNA sequencing devices, which make hundreds of steps per second, equivalent enzymes do not exist for other biopolymers, making sequencing by chemical stepping an attractive option. Furthermore, with optimization of conditions (*e.g.*, increased pH and temperature, the use of selenium *etc.*), chemical stepping along cysteine tracks could be improved to provide rates comparable to those seen for sequencing devices for nucleic acids. An ideal track for biopolymer translocation should have the following properties: (1) individual steps are observable, and each foothold can be differentiated by the ionic current; (2) each track can be used for repeated translocation; (3) the track should be easily tunable by site-directed mutagenesis or chemical modification; (4) stepping rates should be similar throughout the track to avoid hold-ups during translocation. The present work presents significant advances in the endeavour to build an ideal track.

Chapter 3

Elucidating the transport activity of synthetic anion transporters through computational modelling

3.1 Abstract

Synthetic supramolecular anion transporters have attracted significant attention with promising therapeutic applications in ion channel diseases. Recently, a family of photo-switchable supramolecular anion transporters has been developed, in which the transmembrane transport of chloride can be switched on and off using different wavelengths of visible light to switch between the *E*-(OFF) and *Z*-(ON) isomers.[106] The enhancement in transport activity of the *Z*-isomers exceeded that expected solely from the corresponding increase in chloride binding affinity upon photo-isomerisation. Here, I investigate the anion transport process via molecular dynamics simulations. These reveal the enhanced activity of the *Z*-isomers results in part from better anion encapsulation and pre-organization at the water/membrane interface, while the

poorer transport capability of the *E*-isomers is due to a combination of unfavourable reorientation at the interface during anion binding and competitive dimerization in the membrane. I also provide a workflow to construct full energy profiles for the transmembrane anion transport process for the computational study and design of synthetic ion transporters. Parts of this work were published in *Org. Biomol. Chem.* in 2021.[107]

3.2 Introduction

3.2.1 Synthetic anion transporters

Channel proteins facilitate the transport of biologically relevant ions across phospholipid bilayers.[21] Their disruption, due to toxic, autoimmune, or genetic reasons, can lead to various disorders,[21] including cystic fibrosis,[22,23] Bartter's syndrome,[24,25] and Dent's disease.[26] Given their crucial role in cellular function, there has been significant interest over the past few decades in developing synthetic ion channels and mobile carriers as potential therapeutics.[27–29]

Most synthetic anion transporters contain motifs that employ hydrogen bond (HB) interactions for anion recognition, usually involving NH donors such as ureas and thioureas,[108–116] squaramides,[117–121] indoles,[114,122–125] tambjamine,[126] carbazole,[127,128] acridinone,[129] isophthalamides and dipicolineamides,[130] and 1,3-bis(benzimidazol-2-yl)pyrimidine[131] (Figure 3.1). Alternative binding motifs utilising non-classical intermolecular interactions have also emerged as an effective means for anion transport, such as CH hydrogen bonding,[132] anion-π,[133] and sigma hole interactions.[134] In this Thesis, I focus on synthetic anion transporters with HB donors.

Pioneered by Crabtree *et al.* in 1999,[135] isophthalamides and dipicolineamides have stood out as effective motifs in anion recognition. These motifs feature two amide

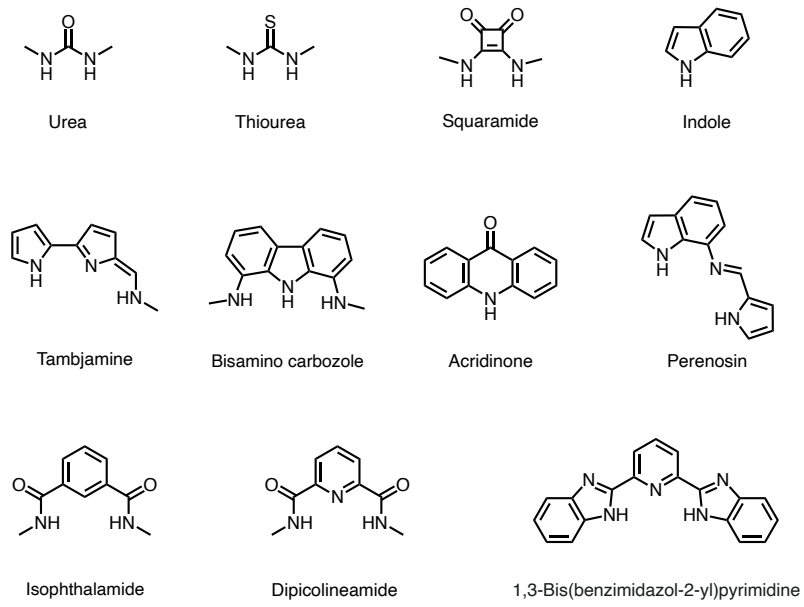


Figure 3.1: Hydrogen bond binding motifs commonly found in synthetic anion transporters.

NH groups that create a cleft where HBs can readily form with the anion. Recently, Caltagirone, Quesada and coworkers investigated the substitution patterns of aromatic functional groups on these systems and found that those containing electron-withdrawing groups were particularly active.[130] Ureas and thioureas have also been extensively investigated for their anion transport properties.[108–116] In 2012, Gale *et al.* first reported the use of squaramides in transmembrane anion transport,[117] demonstrating their superior activity compared to analogous ureas and thioureas. Subsequently, Gale, Jolliffe and coworkers investigated an analogous oxothiosquaramide compound,[118] which exhibited significantly higher acidity compared to regular squaramides, with pK_a values of approximately 6. In 2017, Sessler, Gale, Shin and coworkers provided the first experimental evidence that synthetic squaramide-based ion transporters can disrupt autophagy and induce apoptosis by perturbing cellular chloride concentrations.[119]

One hallmark of biological ion channels and transporters is their precise regulation by biological stimuli,[136,137] such as changes in concentration gradient,[138,139] ligand binding,[140,141] and chemical reactions.[142,143] In order to replicate this biological feature,

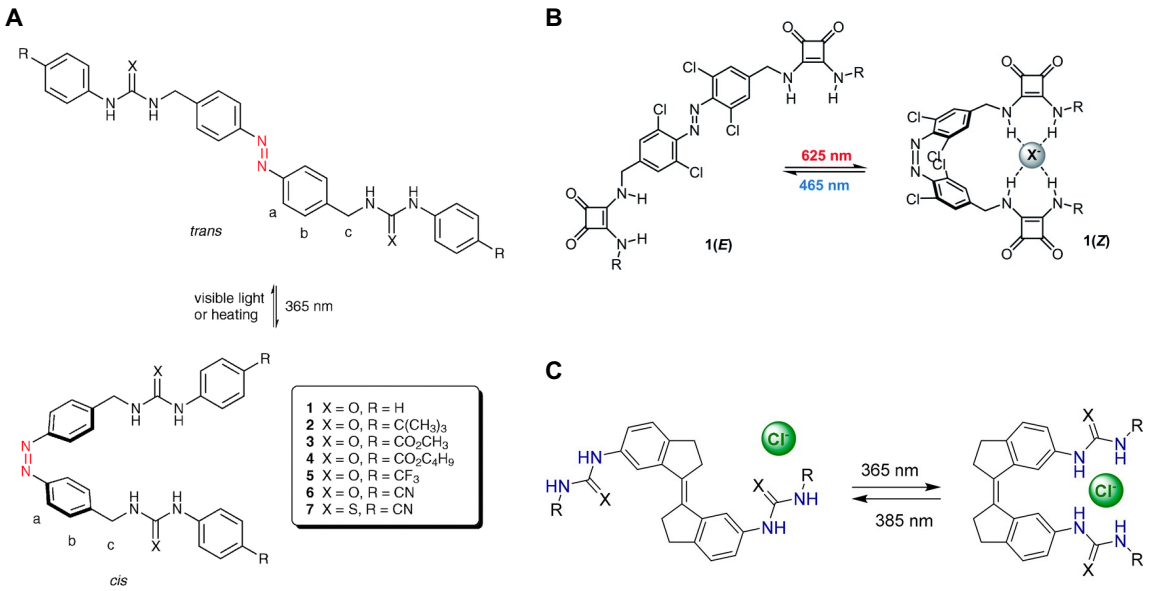


Figure 3.2: Examples of synthetic transporters featuring *E*-/*Z*-switchable scaffolds reported by the (A) Jeong group in 2014,[150] (B) Langton group in 2020,[106] and (C) Gale group in 2022.[153] Reproduced from References[106,150,153] with permissions.

researchers have explored synthetic transporters responsive to a range of stimuli[144] including pH change,[118,121] voltage,[145] chemical or enzymatic reactions,[146–149] and light.[106,107,150–158]

In 2015, Gale *et al.* introduced thiosquaramide-based transporters and demonstrated that some thiosquaramides only became active transporters under acidic conditions.[121] The same group also demonstrated unprecedented voltage-switchable transport properties linked to the lack of uniport activity of a tetraurea macrocycle transporter.[145] Inspired by the concept of procarriers, Park, Namkung, Jeong and coworkers demonstrated in 2016 an enzyme-responsive chloride transporter, where an initially inactive transporter with highly hydrophilic appendages could be activated in the presence of specific enzymes capable of hydrolyzing these appendages, thereby generating an active carrier.[147]

Synthetic transporters based on switchable scaffolds with *E*-/*Z*-isomerization have emerged as an effective control of OFF and ON states for transmembrane trans-

port.[106,107,147,149–154,158] In 2014, Jeong *et al.* first demonstrated photo-switchable chloride transport activity by squaramide-based anion transporters (Figure 3.2A), with the *Z*-isomer being active and the *E*-isomer showing minimal transport activity.[150] In 2020, Langton *et al.* reported a family of squaramide-based anion transporters capable of reversible *E*-/*Z*-switching using bio-compatible visible light (Figure 3.2B),[106] thereby expanding the potential for practical therapeutic applications of synthetic anion transporters. In 2022, similar photo-modulation was realised in stiff-stilbene-based receptors by Wezenberg *et al.* (Figure 3.2C).[153] These examples highlight the *E*-/*Z*-switchable scaffolds as an effective means to control the OFF and ON states for transmembrane transport.

3.2.2 Computational modelling of synthetic anion transporters

The transmembrane anion transport mechanism is a complex process involving several interdependent steps, including anion binding, translocation, anion release, and transporter return.[159] Computational modelling including density functional theory (DFT) and molecular dynamics (MD) methods have been employed to explore one or more of these stages.

In 2017, Dąbrowa *et al.*[151] and Wezenberg *et al.*[158] employed DFT to investigate the binding conformations of *E*-/*Z*- switchable anion transporters. They concluded that the *Z*-isomers were better hosts for oxoanions due to more effective corporative binding via HBs (Figure 3.3A). Additionally, DFT has been employed by Duarte, Langton and coworkers to calculate the electrostatic potential (ESP) surfaces of halogen bonding anion transporters, which revealed the role of charge delocalization and sigma-hole mediated anion recognition.[160] The same authors have developed a simple model, using chloroform as a membrane-mimetic to quantify the energy penalty due to anion dehydration at the water/membrane interface (Figure 3.3A).[161] The transport activity was then determined by the balance between anion dehydration and the transporter's

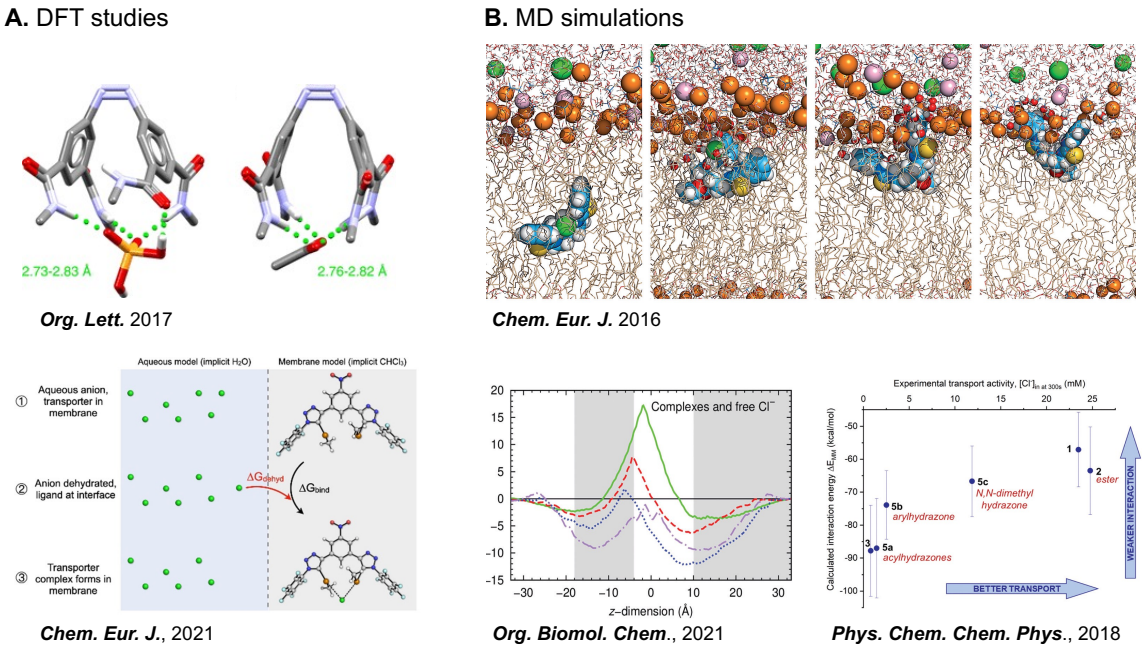


Figure 3.3: (A) DFT[151,160,161] and (B) MD methods[162–165] for studying anion binding and transport activity of synthetic anion transporters.Reproduced from References[151,160–165] with permission.

binding strength. This study demonstrated that energetics related to the transport process can be approximated with the DFT method. However, these studies provided only static information without consideration of explicit solvent or membrane, limiting a comprehensive understanding of the transmembrane anion transport process.

To account for a more realistic representation of the environment and obtain dynamic information on anion transport, all-atom MD simulations in explicit solvent have been employed (Figure 3.3B).[162–165] In 2016, Edwards *et al.* employed conventional MD to simulate the passive diffusion of dithioureidodecalin-based transporters in a membrane, revealing that transport activities of urea-based chloride transporters were inhibited when functionalised with long alkyl chains due to increased transport difficulty.[162] In 2018, Marques *et al.* employed steered MD and umbrella sampling (US) simulations to calculate the potential of mean force (PMF) profiles along the membrane normal of squaramide-based transporters, confirming that assisted chloride transport is thermodynamically favoured.[163] Spooner *et al.* used molecular mechanics

to calculate the interaction energies between tris-thiourea-based transporters and 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC) lipids and found that transport activities depend on van der Waals interactions between the transporters and the membrane.[164] Martínez-Crespo *et al.* computed interaction energies, which revealed the detrimental effects of additional binding sites on the transport activities of acylhydrazone-based transporters.[165] These studies provided a more realistic representation of the environment and obtained dynamic information on transmembrane anion transport. However, these studies mainly focused on unswitchable transporters and investigated individual stages of the transport process. The recently developed photo-responsive switchable systems are significantly more complex and would benefit from a system-specific investigation. A global picture of the full transport process would provide comprehensive insights and a holistic understanding of the entire process of anion transport.

In this study, I present an in-depth computational investigation of the transport process of synthetic transporters, using squaramide-based photo-switchable transporters as a representative example. I employed MD and extended the US simulations to quantitatively characterize the key structural and energetics features contributing to the difference in transport activity in the *E*- and *Z*-isomers of the squaramide-based photo-switchable transporters. Additionally, I provide a workflow to construct full energy profiles for the transmembrane anion transport process, aiding the computational study and design of synthetic ion transporters.

3.3 Theory

I chose MD methods instead of DFT due to the systems' complexity, dynamic nature, and the need to consider solvent effects.

- **System complexity and size**: Synthetic anion transporters in membranes are

relatively large and flexible systems. MD is a more suitable choice for simulating such systems as it allows explicit modelling of solvent-solute interactions. While DFT calculations are, in principle, more accurate, describing electrons explicitly, they are computationally costly for this type of system.

- **Dynamic sampling**: Dynamics is key in transmembrane anion transport. MD simulations easily capture dynamic properties by sampling the system's configurations over time, enabling the observation of fluctuations and ensemble properties. In contrast, single DFT calculations provide a static view of the system.

- **Explicit description of environment**: In DFT, solvent effects are usually addressed using implicit solvent models represented as a polarizable continuum. While this approach is computationally simple and efficient, it does not capture solute-solvent interactions. Explicit inclusion of solvent in DFT significantly increases the computational cost. Furthermore, the number and position of these explicit solvent molecules are not easy to determine *a priori*. However, MD simulations describe solute-solvent/membrane interactions through simplified potentials, providing a more realistic representation of the systems.

In the following section, I introduce key concepts and discuss practical considerations in MD simulations. Further details on the theory behind MD simulations can be found in the Appendix Section C.

3.3.1 Force fields

In classical molecular mechanics (MM), the energy of the system is described by the potential energy function, often referred to as a force field,[166] in terms of bonded (U_{bonded}) and non-bonded interactions ($U_{\text{non-bonded}}$).

$$U = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{3.1}$$

The former can be expressed as a sum of the contributions from bond stretching, angle bending, and torsional interactions. Many different force fields have been developed for different molecular systems, each with unique energy forms and underlying philosophy. These force fields are often designed for specific types of molecular systems, including proteins, RNA/DNA and liquids, and can differ significantly in terms of accuracy and computational cost.

In this Thesis, I use AMBER force fields with the GROMACS molecular simulation package. GROMACS supports several AMBER force fields: AMBER94,[167] AMBER96,[168] AMBER99,[169] AMBER99SB,[170] AMBER99SB-ILDN,[81] AMBER03,[171] and AMBERGS.[172] These force fields provide standard parameters for proteins (including standard amino acids and certain post-translational modifications), nucleic acids, ions, and water. It also supports the GAFF force field,[76,77] which is a versatile force field that includes a variety of atom types and is compatible with the AMBER force fields. The energy form of the bonded interactions in the AMBER force field is expressed as:[173]

$$U_{\text{bonded}} = \sum_{\text{bonds}} k_b(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{torsions}} k_\phi \left[1 + \cos(n\phi - \phi_0)\right] \tag{3.2}$$

The first term describes bonds, with r_0 corresponding to the equilibrium bond length and k_b representing the bond force constant derived from Hooke's Law. The second term corresponds to the potential energy due to angle bending, with θ as the current bond angle, θ_0 as the equilibrium bond angle, and k_θ as the angle force constant, which controls the stiffness of the angle spring. The third term describes torsional

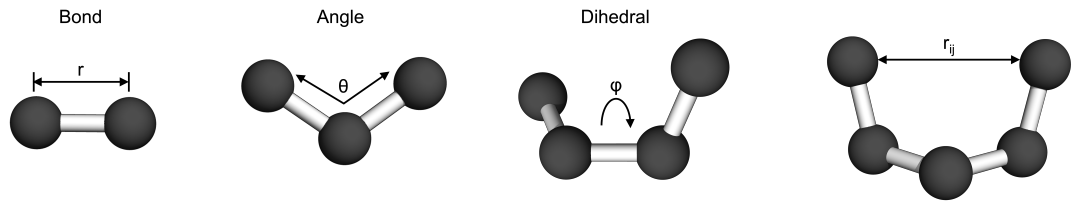


Figure 3.4: Molecular interactions in classical force fields. r is the bond length. θ is the bond angle. ϕ is the torsional angle. r_{ij} represents the distance between atoms i and j.

interactions by a n-fold periodic function, where n represents the periodicity of the torsion, k_ϕ the amplitude and ϕ_0 the phase angle.

The $U_{\text{non-bonded}}$ term describe interactions between atoms or molecules that are not directly bonded, including van der Waals forces and electrostatic interactions that arise from the Coulombic interactions. In molecular simulations, the total potential energy due to non-bonded interactions is given by the sum of Lennard-Jones (U_{LJ}) and Coulomb terms (U_{elec}).

In the AMBER force field, two common versions of this potential exist: the 12-6 and the 12-6-4 Lennard-Jones (LJ) potential. However, only the 12-6 LJ potential is supported in the AMBER implementation available in GROMACS,[80] which is the MD engine used through this Thesis. As a result, I use this form, which is described by:

$$U_{\text{LJ}} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{3.3}$$

where r_{ij} represents the distance between atoms i and j, ϵ_{ij} is the depth of the potential well, and σ_{ij} is the distance at which the LJ potential is zero The sum is taken over all pairs of non-bonded atoms i and j.

The potential energy due to electrostatic interactions is expressed as:

$$U_{\text{elec}} = \sum_{i<j} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{3.4}$$

where ϵ_0 is the vacuum permittivity and q_i and q_j represent the partial charges of atoms i and j. These partial charges are assigned to atoms to represent their relative electronegativities and their contribution to the overall charge distribution. Accurate partial charge assignment is crucial for capturing the correct electrostatic behaviour of the system; however, it is inherently limited by the charge models and the level of theory employed for charge derivation.

3.3.2 Nonstandard parameters

Parameters for the synthetic anion transporters and the POPC membrane are not covered by the AMBER force field. For the synthetic anion transporters, nonstandard parameters were generated using the GAFF force field.[76,77] For the POPC membrane, parameters are obtained from Stockholm lipids,[174] as they are readily available in the GROMACS format and are compatible with the AMBER force field. Additionally, nonstandard parameters were obtained for chloride for the simulations of chloride-transporter complexes. Although the AMBER force field includes chloride parameters from Dang,[175] these parameters are over a decade old and have not been subject to systematic assessment or validation. As a result, their accuracy and applicability to various systems may be limited. To ensure a more accurate simulation of the chloride-transporter complexes, nonstandard parameters for the chloride ion were obtained from Li *et al.*[176]

3.3.3 MD simulation

In MD, the motion of interacting particles in a molecular system is governed by Newton's laws of motion. The net force acting on each atom is obtained from the gradient of the potential energy function with respect to the atom's position. Numerical integration schemes like the leapfrog methods are used to propagate the equation of motion.

To perform MD simulation, several key steps are necessary before propagating Newton's equation, including system preparation, energy minimization, and velocity assignment, after which the system proceeds to equilibration and production (Figure 3.5).

1. **System preparation**: This step includes constructing the starting structure (or preparing it from available structures), solvating, neutralizing, adding ions (if needed), and selecting a force field to generate the topology.

2. **Energy minimization**: Atoms may be initially positioned in close proximity, resulting in a high-energy state during the MD simulations. This step aims to minimise the energy of the system by relaxing it to a local minimum in the potential energy surface (PES). Several methods exist, including conjugate gradient, L-BFGS, and steepest descent. The latter is used in this Thesis.

3. **Velocity assignment**: Random initial velocities are typically assigned to atoms in a way that achieves the desired Maxwell-Boltzmann distribution at the target temperature. Multiple independent simulations of the same system can be run to gather better statistics. In principle, if the system has been sufficiently sampled, the result should not depend on the initial conditions. Further, different initial configurations, such as different conformations of the simulated molecule(s), can be used to immediately introduce distinct behaviour in the system. Restraints can be added in this step as well.

4. **Equilibration**: This step aims to equilibrate the system to reach a stable state representative of the desired thermodynamic conditions, including temperature and pressure before the production stage is run.

5. **Production**: Finally, the data collection phase, commonly referred to as "production", begins. The key distinction between equilibration and production is that in the production simulation, the collected data is intended for analysis and retention.

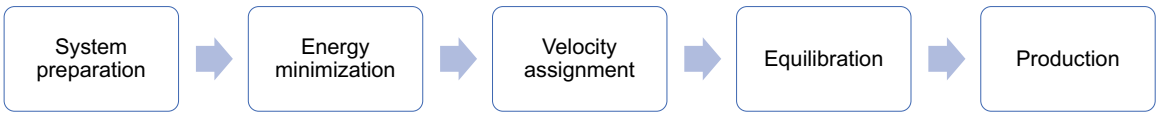


Figure 3.5: Molecular dynamics simulation workflow.

3.3.4 Isobaric-isothermal ensemble

Ensembles are distinct sets of conditions under which MD simulations are performed. These ensembles capture specific thermodynamic properties. Ensembles are crucial because they provide a structured framework for studying and understanding complex molecular systems under varying constraints.

For most chemistry experiments, temperature and pressure are kept constant. The isobaric-isothermal ensemble (NPT) ensemble provides a more realistic representation of experimental conditions. In this ensemble, the number of particles, pressure and temperature are conserved throughout the simulation. Both heat exchange and volume adjustments are allowed, maintaining a constant temperature and pressure. Here, I briefly discuss the widely used thermostats and barostats and my choices in this Thesis. Further details on thermostats and barostats can be found in the Appendix Section C.

The system's temperature is controlled with a thermostat. They work by adjusting the velocities of particles in the simulation to achieve and maintain the desired temperature. Widely used thermostats include the Anderson,[177] Berendsen,[178] Nosé-Hoover,[179,180] and V-rescale[83] thermostats. The Anderson thermostat works by assigning new velocities to randomly selected subsets of atoms from the Maxwell-Boltzmann distribution for the desired temperature.[177] It can correctly sample the canonical ensemble, but the system's dynamics are not physical. The Berendsen thermostat weakly couples the system to an external heat bath and rescales the

velocities periodically.[178] It does not correctly sample the canonical ensemble, making it unsuitable for production runs. The Nosé-Hoover thermostat is an extended system thermostat, which integrates the heat bath as part of the system.[179,180] It samples the correct canonical ensemble, but it may lead to temperature oscillations. The V-rescale thermostat is similar to the Berendsen thermostat, but it performs temperature coupling using velocity rescaling with a stochastic term, which ensures that a correct canonical ensemble is sampled.[83] The V-rescale thermostat can be effectively used for both equilibration and production runs. In this Thesis, I use V-rescale as the thermostat for both equilibration and production runs.

The system's volume fluctuations in response to changes in pressure are controlled with a barostat. Widely used barostats are the Berendsen[178] and the Parrinello Rahman barostats.[84] Similar to the Berendsen thermostat, the Berendsen barostat works by weakly coupling the pressure of the system to a "pressure bath" and rescaling the volume periodically.[178] It does not correctly sample the isobaric-isothermal ensemble. The Parrinello Rahman barostat, similar to the Nosé-Hoover thermostat, is an extended system method and can correctly sample the isobaric-isothermal ensemble.[84] In this Thesis, I used Parrinello-Rahman barostat for both equilibration and production runs.

For membrane simulations in the isobaric-isothermal ensemble, pressure control is often conducted in a semi-isotropic scheme ($NP_zP_{xy}T$), where $P_{xy} = P_x = P_y$ is pressure in the lateral or membrane in-plane directions and P_z is the pressure in the transverse direction. This ensures consistent scaling in the bilayer plane, independent of the out-of-plane dimension.

3.3.5 Umbrella sampling

Conventional MD simulations are valuable for studying molecular systems under equilibrium conditions. However, for this investigation of the transmembrane anion transport process in this Chapter, conventional MD simulations alone are unable to capture the complete energy landscape or explore relevant higher energy states within a reasonable timescale. This is due to the inherently slow kinetics and rare occurrence of the transport events studied here.

To overcome these limitations and gain deeper insights into the anion transport process, I employed umbrella sampling (US), which is an enhanced sampling technique used to explore the potential energy landscape of a system along a specific reaction coordinate (RC).[181,182] In US, the RC is divided into evenly spaced windows, each representing a specific position along the RC. In each window, extensive sampling is achieved by applying an external biasing potential, which acts as a restraining force to encourage the system to explore that particular region of the RC. The purpose of these windows and the biasing potential is to ensure that the samplings in each window overlap with adjacent windows. This overlapping is crucial for accurately reconstructing the unbiased potential of mean force (PMF) when the biasing potential is later removed. The equation for the external biasing potential for window i (u_i) is defined as:

$$u_i = k_i \cdot (r - r_i)^2 \tag{3.5}$$

where k_i is the harmonic force constant for window i, determining the strength of the biasing potential. r is the value of the reaction coordinate at any given time during the simulation. r_i is the reference position for window i, representing the specific value of the RC that is being sampled in that window. In my systems, the reaction coordinate is chosen as the direction perpendicular to the membrane plane, which

effectively represents the reactive stepwise movement in the direction of the membrane normal.

3.4 Materials and methods

3.4.1 Transporter parametrization

An anion transporter complex was optimised in ORCA (version 4.2.1)[183] at the SMD(DMSO)-ωB97X-D3/def2-SVP level of theory[184–186] (ma-def2-SVP on the chloride anion),[187] which includes Grimme's D3 dispersion correction.[188] To speed up the calculations, the resolution-of-identity chain-of-spheres exchange (RIJCOSX) approximation was employed.[189] Force field parameters for the anion transporters were obtained from the GAFF force field.[76,77] The optimised structure was used for particle charges calculations with the Restrained Electrostatic Potential (RESP) method at the HF/6-31G* level using Gaussian16 and Antechamber.[76–79]

3.4.2 MD simulation protocol

MD simulations were performed in the GROMACS (v.2019.2),[80] employing the GAFF force field.[76,77] Parameters for DMSO were taken as described by Caleman *et al.*[190,191] Parameters for POPC were taken from Stockholm lipids.[174] Chloride anion parameters were taken from Li *et al.*[176] These parameters have been found to reproduce the QM hydrogen bond lengths between the chloride anion and squaramide NH protons (2.4 Å).

For the simulations in DMSO, a transporter was inserted in a cubic box and solvated with DMSO molecules. The system was neutralised and minimised using the steepest descent algorithm until the maximum force was below 1000 kJ mol^{-1} nm^{-1}. Random initial velocities were generated according to the Maxwell–Boltzmann distribution

at 298 K (100 ps, 2 fs timestep), followed by equilibration under an NPT ensemble (100 ps, 2 fs timestep). Production simulations were performed in the NPT ensemble during 200 ns (2 fs timestep, 298 K and 1 bar). The temperature of the system was maintained at 298 K using the V-rescale thermostat.[83] The pressure was controlled by the Parrinello-Rahman barostat 1.0 bar,[84] with an isotropic compressibility of 4.5×10^{-5} bar^{-1} for the simulations in DMSO.

For the simulations in POPC, a transporter was placed in the centre of a pre-equilibrated POPC membrane. The system was neutralised and minimised using the steepest descent algorithm until the maximum force was below 1000 kJ mol^{-1} nm^{-1}. Random initial velocities were generated according to the Maxwell–Boltzmann distribution at 298 K in an NPT ensemble (500 ps, 2 fs timestep). Production simulations were performed in the NPT ensemble during 200 ns (2 fs timestep, 298 K and 1 bar). The temperature of the system was maintained at 298 K using the V-rescale thermostat.[83] The pressure was controlled by the Parrinello-Rahman barostat at 1.0 bar,[84] with a semi-isotropic compressibility of 4.5×10^{-5} $bar^{\breve{}1}$.

All simulations were performed with three-dimensional periodic boundary conditions. Long-range electrostatics was described with the Particle Mesh Ewald (PME) algorithm.[85,86] All bond lengths involving hydrogen atoms were constrained using the LINCS algorithm.[87] For each system, the temperature and pressure fluctuations of at least one of the replicas were plotted to check for system equilibration.

3.4.3 Free energy calculations of chloride transport across a POPC membrane

Under the assumption that the energy profile across the entire membrane is symmetrical, energy profiles were generated to span half the membrane from the membrane core to the aqueous phase. Pulling trajectories were generated using steered MD (SMD). A

free transporter or complex was pulled at a constant velocity of 0.5 nm ns^{-1} from the aqueous phase to the membrane core or from the membrane core to the aqueous phase. The potential of mean force (PMF) profile was obtained from umbrella sampling (US) simulations. I performed US simulations at 100 ns per window. The convergence of the PMF profiles was assessed by analysing the PMF profile in sequential different intervals. Convergence is achieved at 50 ns per window for the free transporters and 100 ns per window for the complexes (Figure A13). Therefore, these simulation times were chosen as a suitable protocol for US. From each SMD simulation, the windows were sampled at 0.15 or 0.2 nm. For each system, selected windows from all SMD simulations were combined to ensure good window overlap.

Before analysis, the first 10 ns of each US window was discarded. A PMF profile as a function of anion transporter distance to the membrane centre of mass ($\xi = 0$ nm) is extracted using the Weighted Histogram Analysis Method (WHAM). Statistical uncertainty was estimated with bootstrap analysis.[192] Complete histograms were considered independent data points, and the bootstrap was carried out by assigning random weights to the histograms ("Bayesian bootstrap"). For comparison across all systems of interest, the PMF profiles were normalised to zero in the aqueous phase. Energy barriers associated with each stage were determined from the PMF profiles.

3.5 Results and discussion

3.5.1 Chloride binding studies

I first modelled 1^Z as a representative system in its free form (1^Z) and in a complex with chloride (1^Z-Cl) to explore the enhanced binding affinity of the Z-isomers. For each system, explicitly solvated (DMSO) MD simulations were propagated 5 × 200 ns, resulting in 1 μs of total sampling time. I categorised the conformations observed

during the simulations into three states: an open conformation (d_{RR} = 25 Å), semi-closed (d_{RR}= 15 Å), and closed (d_{RR} = 6 Å), where d_{RR} corresponds to the distance between the *para* carbons of the squaramide phenyl rings (Figure 3.6B, C).

For the free ligand, an open conformation is achieved after only 2 ns, despite the simulation being initiated from a closed conformation, with both squaramide motifs pointing inwards. For the rest of the MD simulation, 1^Z remains in an open or semi-close conformation which maximises hydrogen bond (HB) interactions with DMSO. The 1^Z-Cl predominantly exists in a closed conformation, enabling a two-site binding mode with four HB interactions between the chloride anion and the squaramide NH protons (coordination number, CN = 4). In only 1 of the 5 simulations, chloride is observed to dissociate, leading to a transient one-site binding (CN = 2) state, followed by complete dissociation of the chloride anion.

To explore the role of the linker position and length in chloride binding, the linker position was changed from *para* to *meta* (transporter 12) and linker length from one to two carbon (transporter 13) on the azobenzene core (Figure 3.6A). The Gibbs free energies of binding ($\Delta G_{bind} = -RTlnK^Z_{bind}$) for 1^Z, 12^Z and 13^Z are –3.3, –3.1 and –2.9 kcal mol^{-1} respectively.[107] The differences between the energies are within the computational error of 1 kcal mol^{-1}, so it is unlikely that computational modelling can provide a quantitative correlation within such a small energy range.

To gain a qualitative insight into the structural dynamics of these linker variants, I modelled compounds 12^Z and 13^Z in their free form and complexed with chloride and compared them to the results obtained for 1^Z. In the free form, the *para*-linked compounds, 1^Z and 13^Z, adopt predominantly an open conformation, while for the *meta*-linked compound, 12^Z, a semi-closed or closed conformation dominates (Figure 3.6B). The *meta*-linkage allows a closer distance between the two squaramide, which makes it easier to form π-stacking interactions. In the 1:1 complex with chloride,

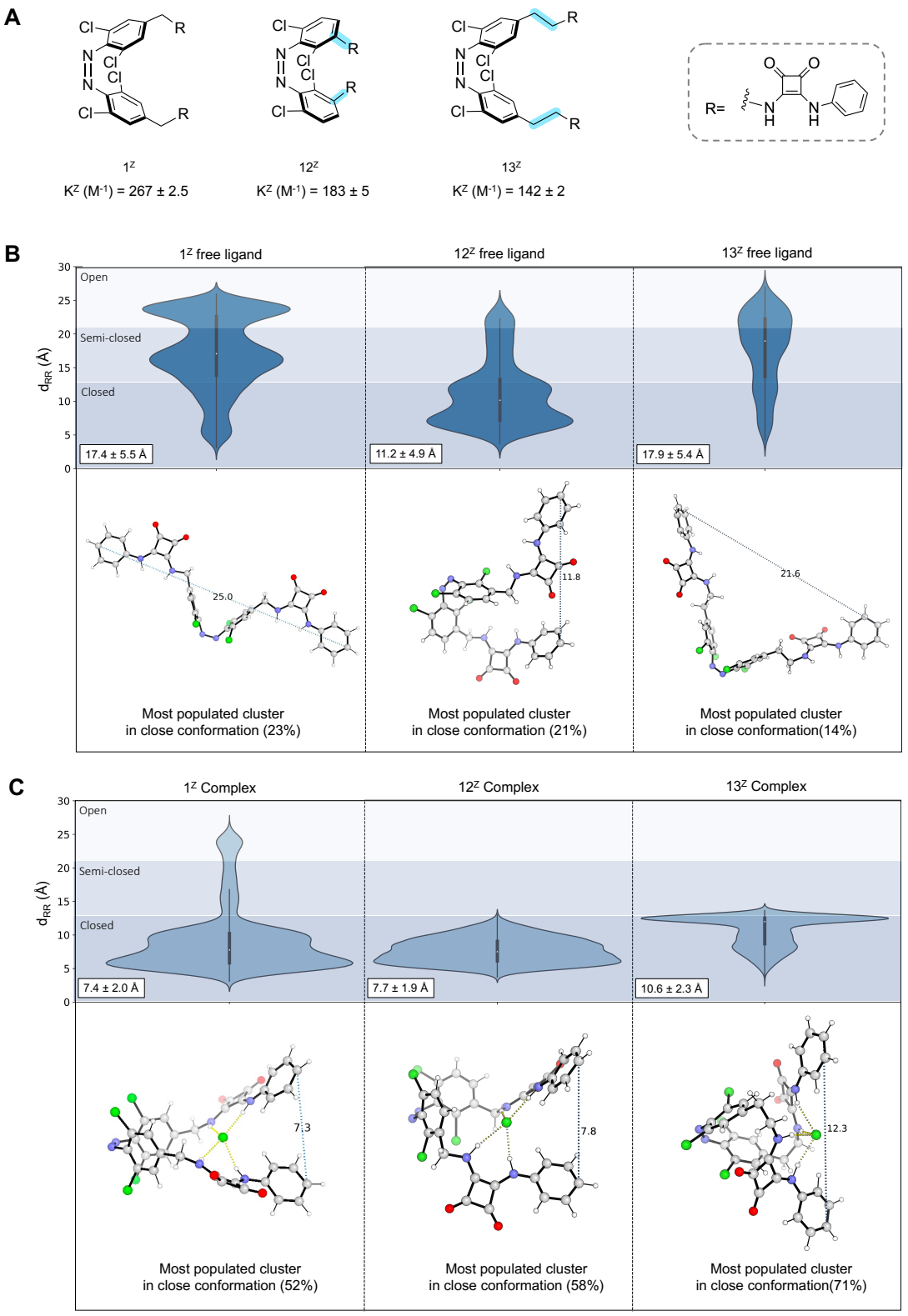


Figure 3.6: Role of linker position and length in chloride binding of the Z-isomers. (A) Structures of 1^Z, 12^Z and 13^Z. (B) End-to-end distance of the transporters as free ligands in DMSO. (C) End-to-end distance of the transporters in complex with chloride in DMSO. Adapted from Reference[107] with permission.

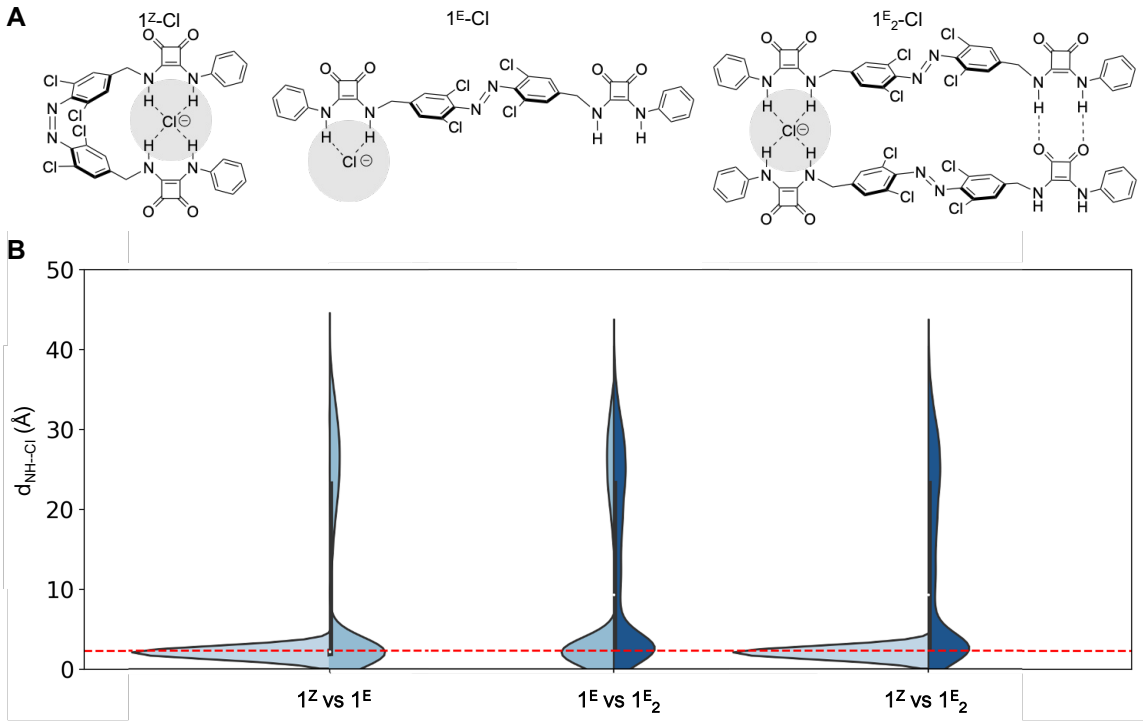


Figure 3.7: Distribution of average HB distance between the squaramide NH protons to the Cl– anion (d_{NH-Cl}) in 1^Z-Cl, 1^E-Cl and 1_2^E-Cl, respectively. The HB distance in a 1^Z-Cl structure, optimised at the [SMD (DMSO)-wB97X-D3/def2-SVP (ma-def2-SVP on anion, Cl)] level of theory, is 2.4 Å (red dashed line)

12^Z-Cl and 13^Z-Cl afford similar binding conformations to 1^Z-Cl, with four HBs to the anion in each case (Figure 3.6C). The end-to-end distance between the *para* carbons of the squaramide phenyl rings is also similar between 1^Z-Cl and 12^Z-Cl. For 13^Z-Cl, which possesses a longer linker, the two squaramide motifs bind from the same side and a more convex geometry around the chloride anion is observed. This conformation makes the anion charge more exposed and less well-encapsulated in 13^Z-Cl. This could explain the smaller experimental binding constant of 142 M^{-1} for 13^Z-Cl, in comparison to 267 M^{-1} and 183 M^{-1} for 1^Z-Cl and 12^Z-Cl respectively. In summary, 12^Z-Cl and 13^Z-Cl afforded similar binding conformations to 1^Z-Cl, with four HB interactions to the anion in each case, which highlights the importance of this four HB cooperative binding mode.

Such optimal interaction with four HBs is not possible for the *E*-isomers, which can at most form two HBs with chloride in a 1:1 complex, 1^E-Cl. This is supported by a lower complex stability of 1^E-Cl in MD simulations (Figure 3.7), with chloride dissociation observed in 3 out of 5 simulations (Figure A10), and a lower binding constant in the chloride binding titration experiments.[106,107] In conclusion, these results demonstrate that stronger binding of the *Z*-isomers in DMSO is due to (1) the formation of four HBs and (2) better charge encapsulation in the Z-complex.

3.5.2 *E*-isomer dimerization

Theoretically, the *E*-isomer could achieve better ion encapsulation by forming four HBs to the chloride anion in a 2:1 complex (1_2^E-Cl). The formation of two additional HBs, while enthalpically favourable, is entropically unfavourable. Indeed, MD simulations of 1_2^E-Cl reveal that the 2:1 complex exhibits no improvement in the complex stability in comparison to the 1:1 complex. I also assessed the dimer stability of 1_2^E in DMSO (Figure 3.8A). The starting dimeric species dissociates within 10 ns and remains independently solvated during the rest of the simulation in 3 out of 5 simulations. One-

site rebinding events are also observed but only as transient states. These results are consistent with the observation of small chemical shift perturbations of the squaramide NH protons in NMR dilution experiments in DMSO with 1^E (<0.05 ppm),[107] suggesting minimal dimerization in this polar media.

Given the high effective concentration of the transporters when confined in the membrane and the lower polarity of the membrane compared to the DMSO solution in which the titrations were conducted, it is likely that the *E*-isomers, in particular, have the propensity to form dimers or higher-order aggregates in the membrane via inter-molecular HBs. Indeed, squaramide derivatives are well known to form strong, bidentate intermolecular HBs, and often have poor solubility in non-polar solvents.[193] MD simulations of 1_2^E in a model POPC membrane reveal the formation of a stable dimer. In the initial position with 1_2^E positioned at the centre of the membrane, the 1_2^E is held together via HB interactions between the squaramide motifs of each monomer (Figure 3.8B). As the dimer moves towards the membrane-water interface, interactions with the phosphate head groups of the POPC lipid replace some of the intermolecular HB interactions leading to a one-site binding state (d_{MM} = 10 Å). These results suggest that, in contrast to when in DMSO, *E*-isomers, and potentially higher-order aggregates, could exist in the membrane. I hypothesise that this dimerisation could compete with anion binding and suppress transport activity.

3.5.3 Passive diffusion in a POPC membrane

To gain insights into the structural dynamics and complex stability of the *Z*- and *E*-isomers in the membrane, I modelled 1^Z-Cl and 1^E-Cl in a model POPC membrane. During the MD simulation, the complexes that were initially positioned at the centre of the membrane migrate towards the water/membrane interface. As they approach the interface, the HBs between the transporter and the chloride are disrupted. The squaramide motifs form new HB interactions with the phospholipid headgroups and

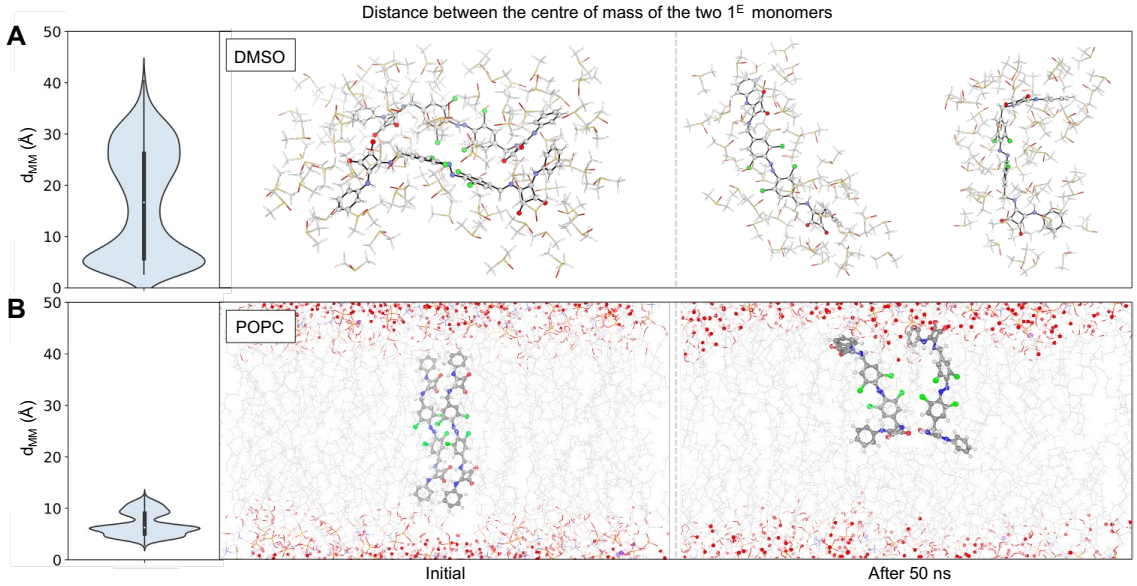


Figure 3.8: Distance distribution between the centre of mass (COM) of the two 1^E monomers, d_{MM}, over 5 replicates of 50 ns MD simulations, resulting in 250 ns of total sampling time; and the corresponding snapshots of the 1aE dimer at the start and end of a 50 ns explicitly solvated MD simulation in DMSO (top) and a POPC membrane (bottom). Adapted from Reference[107] with permission.

water molecules, which in turn facilitates chloride release into the aqueous phase. Interestingly, I observed that chloride dissociation from 1^E-Cl occurred faster than from 1^Z-Cl. This behaviour is consistent with the results observed when modelling each complex in the DMSO solution, which also indicated stronger binding by 1^Z due to more HB interactions between chloride and 1^Z.

Upon chloride release, both 1^Z and 1^E remain at the water/membrane interface throughout the simulations. In the Z-isomer, the squaramide motif points towards the water phase with the azobenzene backbone embedded in the hydrophobic lipid tails. This conformation enables the Z-isomer to readily capture chloride anion. Due to the formation of intramolecular HBs between the two squaramide motifs, 1^Z adopts a closed conformation (Figure 3.9A), similar to the dominant 1^Z-Cl complex conformation in DMSO. This closed conformation appears to pre-organize the transporter and may result in less structural distortion upon chloride binding in POPC than in DMSO (Figure 3.6C). On the other hand, the E-isomer exists in

two distinct orientations: (1) parallel to the membrane normal, with polar interactions forming with one squaramide motif, and (2) perpendicular to the membrane normal, with both squaramide motifs interacting with the phospholipid headgroups simultaneously (Figure 3.9B). The second binding mode brings the *E*-isomer to the water/membrane interface, suggesting an additional kinetic barrier to re-orientation in chloride transport.

3.5.4 Chloride transport across a POPC membrane

The chloride transport process occurs in four steps, as shown in Figure 3.10A: (1) chloride binding by a free transporter; (2) complex translocation across the POPC membrane, (3) chloride release at the other side of the membrane, and (4) recycling of the free transporter for the next transport cycle.[159] Four anion transporters with different functional R groups on the squaramide motifs, covering wide range of transport activities (EC_{50}) but similar binding affinities (K_{bind}), were investigated (Figure 3.10B): 1^Z (R = phenyl, high transport activity, K_{bind} = 267 M^{-1}, EC_{50} = 22 nM), 2^Z (R = ethyl, K_{bind} = 142 M^{-1}, EC_{50} = 432 nM), 3^Z (R = hexyl, K_{bind} = 204 M^{-1}, EC_{50} = 28 nM) and 4^Z (R = octyl, K_{bind} = not determined, EC_{50} > 20000 nM). Comparison between transporter 1^Z and 2^Z – 4^Z energetically assesses how the difference between aryl and alkyl substituents affect chloride transport, while the effect of increasing alkyl chain length was evaluated with transporters 2^Z – 4^Z. The potential of mean force (PMF) profiles of chloride dissociation from the complexes and PMF profiles of the free transporters and their chloride complexes along the membrane normal were determined using steered MD methods and umbrella sampling (US) simulations.

The lipophilicity of the transporters was assessed through the determination of energy barriers ($\Delta G^{\ddagger}_{\text{popc-to-wat}}$) for their transition from the membrane to the water phase (Figure 3.10C). The smaller values of $\Delta G^{\ddagger}_{\text{popc-to-wat}}$, complex suggest that the complex,

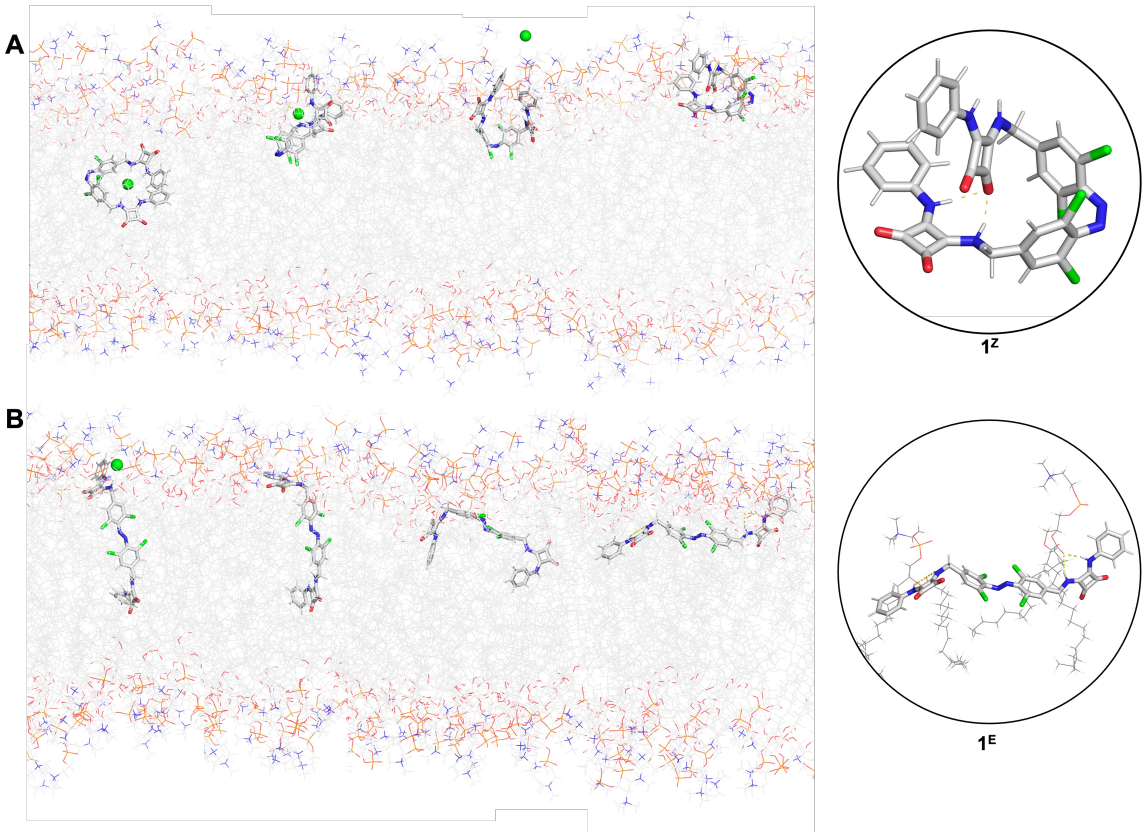


Figure 3.9: Passive diffusion in a POPC membrane. MD snapshots depicting the time evolution of the orientations and interactions of (A) 1^Z-Cl and (B) 1^E-Cl in a POPC membrane.

which carries a negatively charged chloride, disrupts the hydrophobic membrane environment, and is less lipophilic. Thus, it is less favourable to be in the membrane. From the PMF profiles, I observed that the free transporters surmount higher $\Delta G^{\ddagger}_{\text{popc-to-wat}}$ (10.7 ± 0.2, 16.8 ± 0.6, 16.7 ± 0.7 and 17.4 ± 0.3 kcal mol^{-1} for 1^Z, 2^Z, 3^Z and 4^Z respectively) than their corresponding complexes (8.3 ± 0.5, 6.9 ± 0.2, 12.2 ± 0.5 and 16.1 ± 0.6 kcal mol^{-1} in the same order). For 2^Z, 3^Z and 4^Z, the difference of $\Delta G^{\ddagger}_{\text{popc-to-wat}}$ between the complex and the free transporter ($\Delta\Delta G^{\ddagger}_{\text{popc-to-wat}}$) decreases are 9.9, 4.6, 1.3 kcal mol^{-1} respectively, which suggests better charge shielding by the transporter as the alkyl chain length increases.

Full energy profiles were subsequently constructed from the PMF profiles (Table 3.1, Figure 3.10D). Chloride binding occurs at the water/membrane interface with similar energy barriers ($\Delta G^{\ddagger}_{\text{bind}}$). This energy cost is attributed to the unfavourable elec-

trostatic repulsion between the negatively charged chloride anion and the negatively charged phosphate headgroup region of the POPC membrane, which is independent of the anion transporter. Upon complexation, the chloride anion is stabilised by four HBs with the squaramide NHs. The process of interfacial binding is endothermic, costing 1.3, 1.7, 1.5 and 1.3 kcal mol^{-1} for 1^Z, 2^Z, 3^Z and 4^Z respectively. Complex translocation across the membrane core towards the opposite leaflet faces the highest energy barriers of the translocation process ($\Delta G^{\ddagger}_{\text{trans}}$), costing $>$ 7.5 kcal mol^{-1} for all transporters. Return of free transporters across the membrane surmount lower energy barriers ($\Delta G^{\ddagger}_{\text{return}}$) of 7.0, 7.1, 5.9 and 6.1 kcal mol^{-1} for 1^Z, 2^Z, 3^Z and 4^Z respectively.

From the full energy profiles, the overall energy barrier for the entire transport process (*i.e.*, the chloride complex translocation process, $\Delta G^{\ddagger}_{\text{total}}$) are 12.4, 11.2, 9.9, and 8.8 kcal mol^{-1} for transporter 1^Z, 2^Z, 3^Z and 4^Z respectively. These all correspond to the complex translocation step, which suggests that complex translocation is the rate-limiting step of the transport process for these squaramide-based anion transporters. The calculated energy barrier, $\Delta G^{\ddagger}_{\text{total}}$, shows some correlation with experimentally determined log(k) values for 1^Z, 2^Z, and 3^Z ($R^2 = 0.7$, Figure 3.10E), where k is the initial rate of chloride transport calculated at 125 nM transporter concentration 206 seconds after transporter addition in the HPTS assay.

In contrast, 4^Z, which has the smallest $\Delta G^{\ddagger}_{\text{total}}$, shows no transport activity experimentally. However, this computational value of the translocation barrier, $\Delta G^{\ddagger}_{\text{total}}$, only represents the thermodynamic driving force for transfer from the centre of the membrane to water and does not account for other kinetic factors such as mobility, or membrane-uptake that will also contribute to the overall observed transport activity. Experimentally, in the transport assay, the transporters were added as DMSO solutions to pre-formed lipid vesicles; low solubility of 4^Z is likely responsible for poor membrane

uptake and negligible activity.

Table 3.1: Energy associated with the chloride transport process by the transporters 1^Z, 2^Z, 3^Z and 4^Z calculated from US simulations.

	$\Delta G^{\ddagger}_{\text{bind}}$ (kcal mol^{-1})	$\Delta G^{\ddagger}_{\text{trans}}$ (kcal mol^{-1})	$\Delta G^{\ddagger}_{\text{release}}$ (kcal mol^{-1})	$\Delta G^{\ddagger}_{\text{return}}$ (kcal mol^{-1})	$\Delta G^{\ddagger}_{\text{total}}$ (kcal mol^{-1})
1^Z	2.52 ± 0.51	9.92 ± 0.28	1.21 ± 0.37	7.01 ± 0.10	11.23
2^Z	2.13 ± 0.22	10.73 ± 0.13	0.45 ± 0.12	7.10 ± 0.33	12.41
3^Z	2.17 ± 0.18	8.39 ± 0.18	0.65 ± 0.18	5.94 ± 0.32	9.91
4^Z	2.76 ± 0.54	7.49 ± 0.39	1.44 ± 0.38	6.09 ± 0.20	8.81

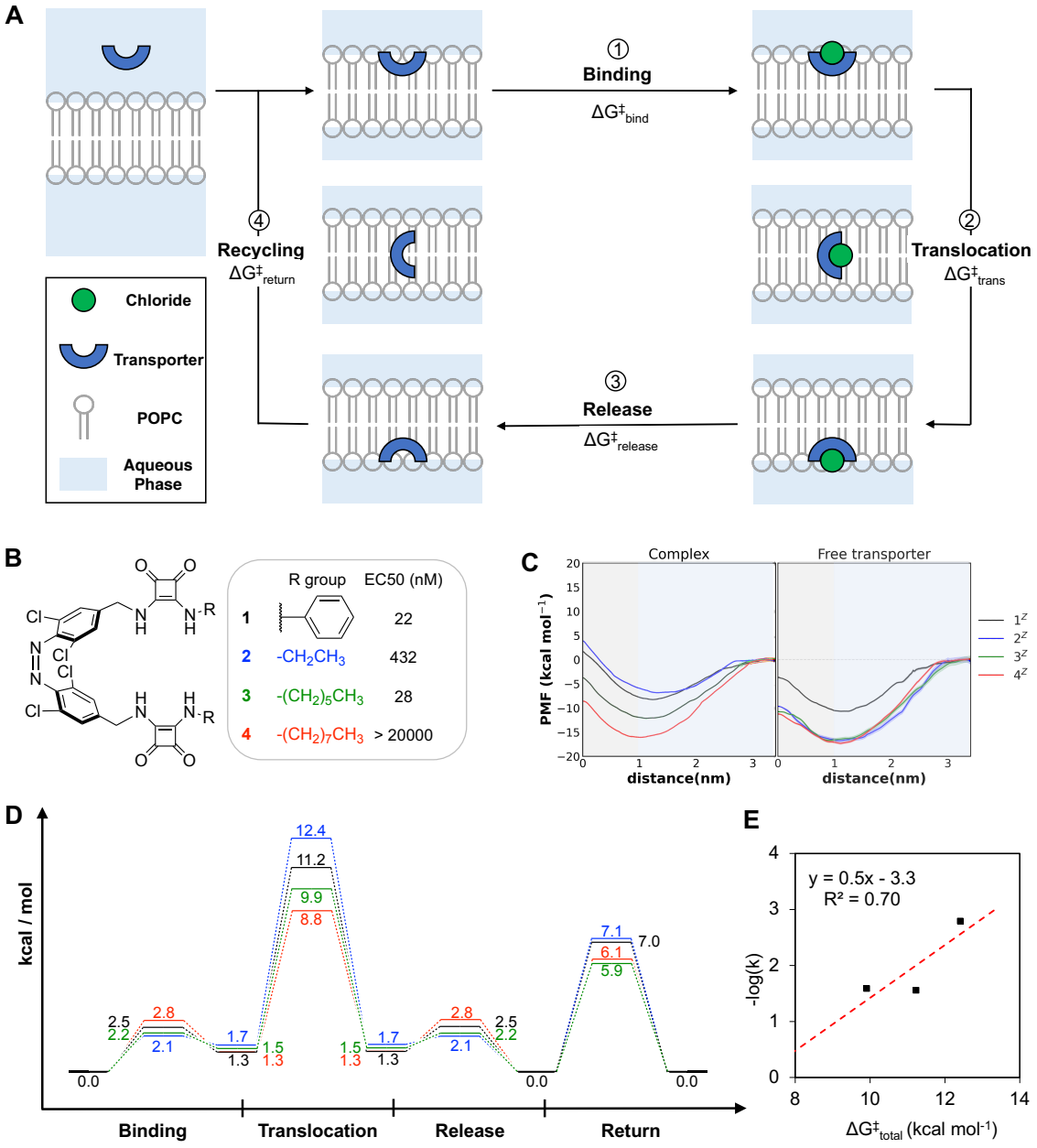


Figure 3.10: Chloride transport across a POPC membrane. (A) Chloride transport mechanism. (B) Chemical structures and corresponding experimental transport activities of transporters 1^Z, 2^Z, 3^Z and 4^Z. (C) PMF as a function of the distance between the transporters and the centre of mass of the POPC membrane (denoted as distance = 0 nm). The water phase is coloured in blue, and the membrane phase is coloured in grey. (D) Energy profiles associated with the chloride transport process by the transporters 1^Z, 2^Z, 3^Z and 4^Z. (E) Correlation between the total energy barrier of chloride complex translocation $\Delta G^{\ddagger}_{\text{total}}$, and the initial rate of chloride efflux from vesicles calculated at 125 nM transporter concentration 206 s after transporter addition in the HPTS assay

3.6 Conclusion

I investigated the conformations and dynamic behaviours of squaramide-based anion transporters using MD simulations. The enhanced anion binding in the Z-isomer is attributed to better complex stability due to better chloride encapsulation and increased HB interactions. This difference in binding mode is also found in a POPC membrane. Further, 1^Z adopts a closed conformation at the water/membrane interface, which pre-organizes the transporter for chloride capture. In contrast, the re-orientation of E-isomer introduces an additional barrier to chloride translocation. Competitive dimerization of the E-isomer further diminishes its transport activity. Free energy calculations show that complex translocation is the rate-limiting step of the transport process. I note that there is a weak linear relationship between the $\Delta G^{\ddagger}_{\text{total}}$ and log(k) for 1^Z, 2^Z and3^Z. The lack of measurable transport activity of 4^Z is attributed to its poor solubility. For the future design of E-/Z-switchable anion transporter, a higher degree of pre-organization of the Z-isomer could lead to better transport activity of the Z-isomer in comparison to the E-isomer and improve the difference in activity between the ON (Z-isomer) and OFF (E-isomer) state. To improve transport activity, the transport activity depends on a fine balance between the energy barrier of transport and the solubility of the transporter.

Chapter 4

Predicting chloride transport activity of synthetic anion transporters with machine learning

4.1 Abstract

The transport of chloride across membranes is crucial for maintaining concentration gradients vital to biological processes, and its dysregulation has been linked to various medical conditions. While synthetic anion transporters hold potential as therapeutic agents, the development of efficient design strategies remains challenging. Additionally, the intricate nature of transport activity, dependent on numerous variables, poses difficulties for traditional molecular modelling. Here, I showcase the use of machine learning (ML) as a predictive tool for assessing the chloride transport ability of synthetic transporters. Employing both classification and regression frameworks, the trained models exhibit remarkable performance across a diverse range of systems. Moreover, they offer insights relevant to future design efforts. For example, the ML

models identify key structural features and experimental conditions that influence the observed transport activity. This work bridges data-driven approaches and molecular design, providing interpretable structure-activity relationships that will aid hypothesis generation and contribute to synthetic advances in this field.

4.2 Introduction

The transport of chloride across membranes is crucial for maintaining concentration gradients vital to biological processes.[21] This process relies on chloride channel proteins such as the voltage-sensitive chloride ion channel (CIC) subfamily,[194–196] the cystic fibrosis transmembrane conductance regulator (CFTR)[22] and calcium-activated chloride channels (CaCCs).[197,198] Malfunction of these transport proteins can result in disorders,[21] such as cystic fibrosis,[22,23] Bartter's syndrome,[24,25] and Dent's disease.[26]

Synthetic anion transporters have emerged as valuable models for studying transmembrane transport processes with potential applications as targeted therapeutics for ion channel-related diseases. Increasing progress has been reported in recent years,[27–29] including the development of anionophores with hydrogen bonding motifs[108–131,199–205] and with non-classical intermolecular binding motifs like halogen bonding, pioneered by Matile and co-workers.[134] Despite these advances, designing effective synthetic anion transporters remains challenging and often relies on trial-and-error approaches or known active transporters, which limits the structural diversity of new designs.

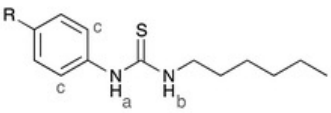
Computational modelling has provided key insights into the binding and transport mechanisms of synthetic anion transporters.[107,151–153,158,160–165] For example, DFT studies on chalcogen bonding anion complexes have revealed the role of charge delocalisation, sigma-hole interactions and desolvation on recognition and transport.[160,161] MD simulations have also been used to investigate the dynamic behaviour of synthetic

anion transporters in explicit membrane environments. For example, Spooner *et al.* used MD to calculate the interaction energies between tris-thiourea-based transporters and POPC lipids and found that transport activities depend on van der Waals interactions between the transporters and the membrane.[164] I have also employed MD to study a photoswitchable phenyl-squaramide functionalised azobenzene transporter, revealing the cooperative binding of the *Z*-isomer and the negative effect of transporter dimerization in the membrane.[107]

However, despite these key insights, several limitations remain. DFT methods often fail to account for structural flexibility/conformations and explicit solvent effects, which restrict a comprehensive understanding of the dynamic transmembrane transport process. MD simulations can provide dynamic insights but lack key effects such as charge transfer and polarisation, which are key in anion binding and charge shielding during transmembrane transport. Moreover, studying ion transport in the presence of the membrane remains computationally costly and their set-up requires expert knowledge, posing challenges for non-experts and generating a gap between synthetic and computational efforts in the field. Given these challenges, the ability to directly predict transport activity from molecular structure becomes highly valuable. This could accelerate the evaluation and design of potential anion transporters, which circumvents the computational demands and expertise required for DFT and MD simulations.

A widely adopted approach in medicinal chemistry for connecting molecular structures and activity involves the use of Quantitative Structure-Activity Relationship (QSAR) methods.[208,209] These approaches utilize statistical models to explain and even predict variation in biological activity due to changes in structural, chemical, statistical, and physical properties of the molecules in the dataset.[210] In 2013, Gale *et al.* pioneered the use of QSAR approaches to predict anion transport activity.[206,207]

A. QSAR analysis of substitution patterns on thiourea-based transporters

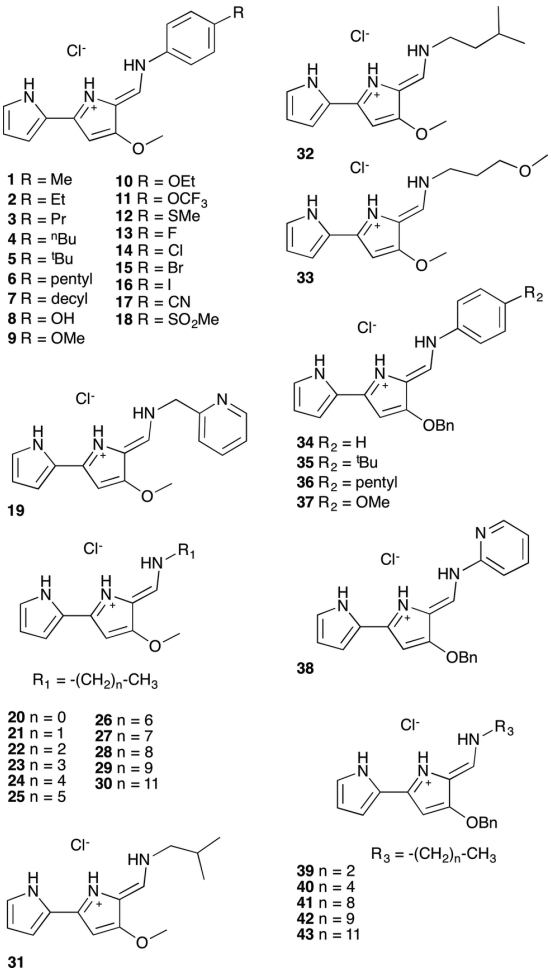


1 R = Br
2 R = CF_3
3 R = Cl
4 R = CN
5 R = $COCF_3$
6 R = COMe
7 R = COOMe
8 R = F
9 R = H
10 R = I
11 R = NO_2
12 R = O(CO)Me
13 R = OCF_3
14 R = OEt
15 R = OMe
16 R = SMe
17 R = SO_2Me
18 R = CH_3
19 R = CH_2CH_3
20 R = $(CH_2)_2CH_3$
21 R = $(CH_2)_3CH_3$
22 R = $(CH_2)_4CH_3$

$$\log K_a(Cl^-) = 0.55(\pm 0.03)\sigma_p + 1.17(\pm 0.01),\ N = 18,\ R^2 = 0.96,\ R_{adj}{}^2 = 0.96,\ \mathrm{RMSE} = 0.04,\ F = 424$$

$$\log(1/EC_{50}) = -0.68(\pm 0.06)n - 1.25(\pm 0.08)\log(1/k_{ini})/n + 1.3(\pm 0.1),\ N = 18,\ R^2 = 0.95,\ R_{adj}{}^2 = 0.94,\ \mathrm{RMSE} = 0.14,\ F = 145$$

B. QSAR analysis of substitution patterns tambjamine-based transporters



1 R = Me
2 R = Et
3 R = Pr
4 R = ^{n}Bu
5 R = ^{t}Bu
6 R = pentyl
7 R = decyl
8 R = OH
9 R = OMe
10 R = OEt
11 R = OCF_3
12 R = SMe
13 R = F
14 R = Cl
15 R = Br
16 R = I
17 R = CN
18 R = SO_2Me

19

R_1 = $-(CH_2)_n-CH_3$

20 n = 0
21 n = 1
22 n = 2
23 n = 3
24 n = 4
25 n = 5
26 n = 6
27 n = 7
28 n = 8
29 n = 9
30 n = 11

31

32

33

34 R_2 = H
35 R_2 = ^{t}Bu
36 R_2 = pentyl
37 R_2 = OMe

38

R_3 = $-(CH_2)_n-CH_3$

39 n = 2
40 n = 4
41 n = 8
42 n = 9
43 n = 11

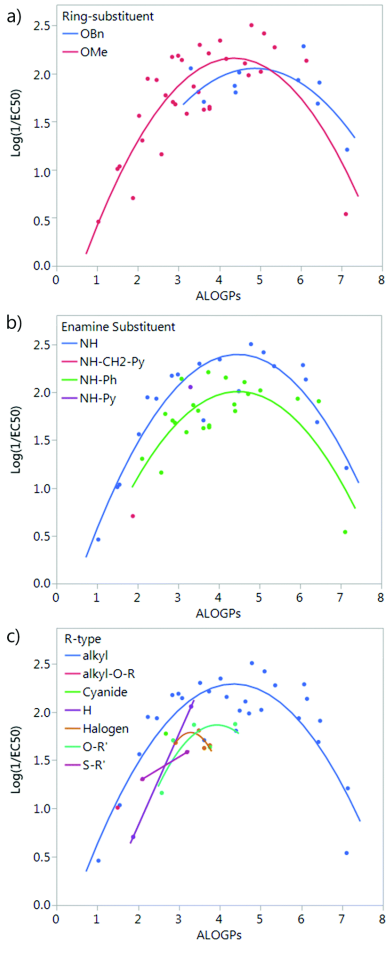


Figure 4.1: QSAR approaches for predicting anion transport activity of (A) squaramide-based[206] and (B) tambjamine-based anion transporters.[207] Reproduced from References[206,207] with permission from the Royal Society of Chemistry.

Their first QSAR model evaluated the impact of functional group variations on transport properties for a set of 22 fully characterised thiourea-based transporters (Figure 4.1A).[206] This model revealed that the transport ability of these compounds was primarily governed by the hydrogen bond (HB) acidity of the thiourea scaffold and octanol–water partition coefficients (log P values), which are often used as a quantitative measure for lipophilicity. Subsequently, in 2016 they generated a QSAR model for 43 tambjamine-based anion transporters for which a parabolic correlation between transport activity and log P was identified (Figure 4.1B).[207] While promising, these QSAR studies were restricted to relatively small datasets, each of them based on a specific family of transporters.

Recent years have witnessed the emergence of machine learning (ML) as a compelling alternative to traditional QSAR methods and offers an adaptable, data-driven, and flexible approach to handling high-dimensional datasets.[211] The efficiency of current ML methodologies enables the use of diverse and large datasets, uncovering hidden patterns, correlations, and causality within the data without requiring preconceived assumptions.[212] ML has found applications in various areas of chemistry, including molecular property prediction, inhibitor design and synthesis predictions.[213,214] However, to the best of my knowledge, no application of such approaches has been reported for synthetic ion transporters.

In this Thesis, I employed ML to predict the chloride transport activity of synthetic anion transporters. I built both classification and regression models and demonstrated remarkable performance across a diverse range of systems. Moreover, these models are interpretable, offering insights relevant to future design efforts.

4.3 Theory

The fundamental goal of ML models is to make fast and accurate predictions, estimate properties, or generate insights for new, unseen data based on patterns and relationships learned from the training data.[212] Depending on the scientific question to be addressed, learning algorithms can be broadly categorized into two classes: supervised learning and unsupervised learning (Figure 4.2).[212] The key distinction lies in the presence or absence of labelled data during the learning process. In supervised learning, the training dataset consists of input data along with their corresponding desired output or labels. The goal is to learn a mapping between the input features and the known output labels. In unsupervised learning, the training dataset consists of only input data, with no corresponding labels or desired outputs. The algorithm learns patterns or structures in the data without any prior knowledge of the expected outputs. The model construction workflow involves the following steps:[215] (1) data collection and preprocessing, (2) feature representation, (3) learning algorithm selection, (4) model validation and evaluation and (5) model interpretability and explainability.

In this Thesis, I use supervised learning, specifically classification and regression tasks. In the following section, I introduce the key steps and considerations in building an ML model, ML algorithms, and model evaluation metrics.

4.3.1 Data collection and preprocessing

In this Thesis, I collected data from the literature. First, relevant literature sources are identified using criteria such as publication date, research methodology, or the specific chemical property of interest. The criteria are well-defined to ensure consistency and minimize bias during the selection process. Once the relevant articles are identified, information is extracted. Errors, omissions, and outliers are common in raw datasets.[215] The collected data are cleaned by removing duplicates, handling missing values, and

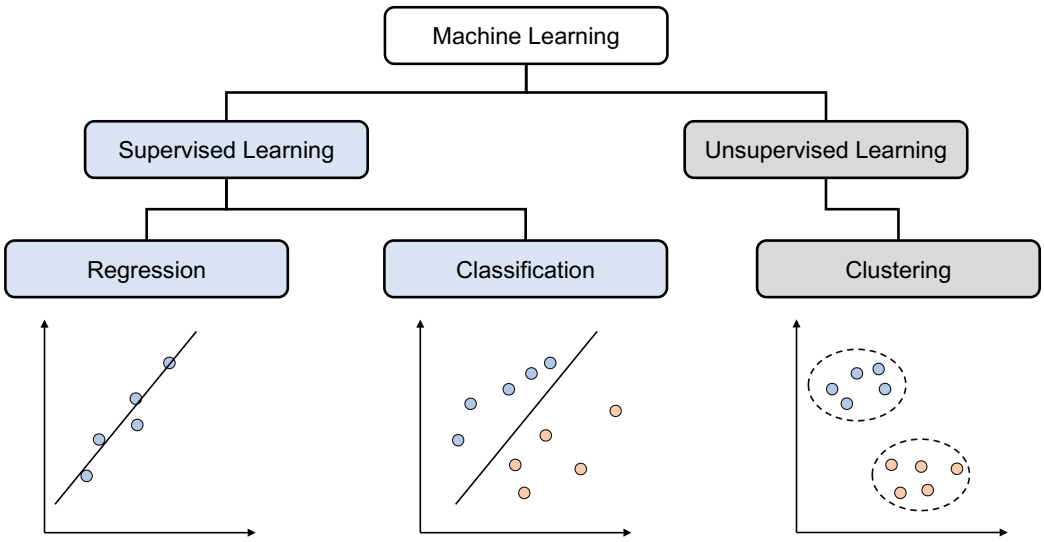


Figure 4.2: Overview of machine learning algorithms.

correcting errors to ensure data quality.

The dataset is split into training and test sets. This split allows for an unbiased evaluation of the model's performance.[215] By withholding a portion of the data as the test set, it provides an independent dataset that has not been seen or used during the model training process. This ensures that the evaluation metrics obtained from the test set reflect the model's true ability to generalize and make predictions on unseen data.

4.3.2 Feature representation

Feature representation refers to the process of converting raw data into a suitable format or representation that can be effectively processed by ML algorithms.[216] There are several ways to represent data. In this Thesis, I used molecular fingerprints and Mordred descriptors.

Molecular fingerprints Molecular fingerprints are binary or binary-like representations that capture the structural characteristics of molecules.[212] They encode information about the presence or absence of specific structural fragments, substructures, or

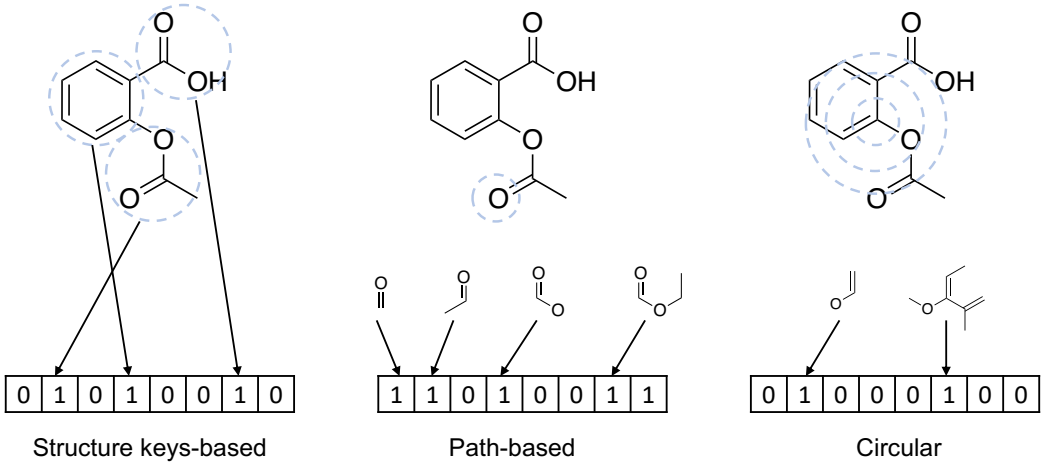


Figure 4.3: Different types of molecular fingerprints.

molecular properties. Molecular fingerprints come in various types, depending on how the molecular representation is transformed into a bit string. The main approaches include substructure keys-based fingerprints, topological or path-based fingerprints, and circular fingerprints (Figure 4.3).

- Substructure keys-based fingerprints, such as MACCS keys fingerprint,[217] set bits based on the presence of specific substructures from a given list of structural keys. These fingerprints are most effective for molecules containing the specified keys but may miss features in other molecules. The number of bits is determined by the number of keys. Each bit represents the presence or absence of a single key.

- Topological or path-based fingerprints analyze all molecule fragments along linear paths up to a certain bond limit, creating a fingerprint. Typical examples include RDKit's atom pair[218] and topological torsion fingerprint.[219] They can represent any molecule and support fast substructure searching and filtering. They are hashed fingerprints, meaning a single bit can't be traced to a specific feature, leading to potential "bit collisions."

- Circular fingerprints are generated by considering the "circular" environment of

each atom up to a certain radius.[220,221] They excel at full structure similarity searching but are not suitable for precise substructure queries due to potential differences in atom environments within the same fragment.

In this Thesis, I tested different fingerprints and found Morgan fingerprints,[220] which is a type of circular fingerprint,[221] to be the best performing. I generated the Morgan fingerprint using RDKit.[222] The Morgan fingerprint generation process in RDKit involves parsing the input molecule to extract atom and bond information. The fingerprint parameters, including the fingerprint radius and bit length, are set. The algorithm then iteratively extends atom environments around each atom, generating substructures at different radii. RDKit hashes these substructures and maps them to specific bits in the fingerprint. The resulting binary fingerprint represents the presence or absence of unique substructures in the molecule's environment. Morgan fingerprint is invariant to atom permutation and can capture local substructures and molecular fragments effectively. The choice of radius determines the level of structural detail captured, whereas larger radii capture more global structural information.[220,221]

Molecular descriptors Molecular descriptors are numerical representations that capture various properties of molecules, which provides a detailed characterization beyond the structural information captured by fingerprints. Mordred is a Python library that offers over 1800 descriptors, including 1D, 2D, and 3D descriptors.[223] 1D descriptors provide information based on atom counts, bond counts, and other simple metrics, such as number of rotatable bonds or the number of hydrogen bond donors/acceptors. 2D descriptors capture the structural properties of molecules, including connectivity, topology, and substructure information, such as counts of specific substructures or functional groups. 3D descriptors take into account the three-dimensional arrangement of atoms in a molecule, providing insights into shape, conformation, and spatial properties, such as molecular surface area, volume, moments of inertia, or measures of

molecular flexibility. Mordred also includes a range of physicochemical descriptors, such as molecular weight, log P, solubility, or polarizability, which captures information about the chemical and physical properties of molecules.

Feature selection Having numerous features in a dataset can lead to several challenges:[212,224]

- The curse of dimensionality emerges. High dimensionality requires exponentially more data, which results in sparsity and difficulties in capturing underlying patterns effectively.
- Dealing with a larger number of features increases the complexity of the model training process, which demands additional computational resources and time.
- The risk of overfitting rises, as random correlations among numerous features can lead to reduced model generalization on new data.
- As model complexity grows, the interpretability of the model diminishes, making it harder to understand the relationships between variables.
- Managing and storing large datasets with numerous features becomes challenging, which creates data collection and storage overheads.

In light of these challenges, feature selection and engineering are crucial to building effective and efficient ML models.[224] There are various techniques for feature selection, ranging from simple methods such as correlation analysis to more advanced techniques such as recursive feature elimination.[224] The choice of feature selection method depends on the specific problem, dataset size, and the complexity of the model being used.

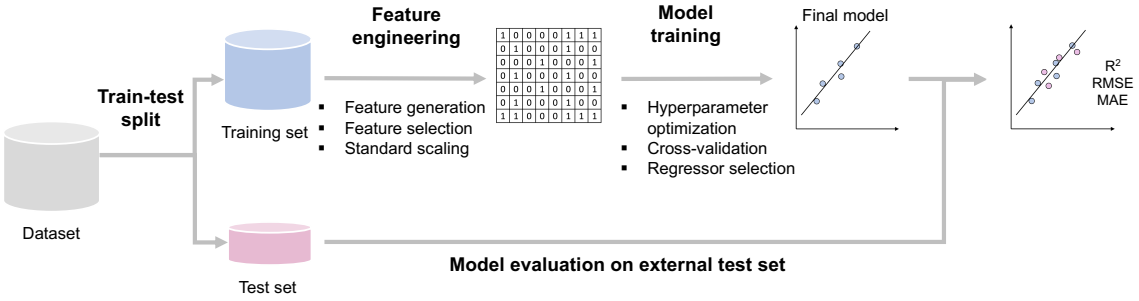


Figure 4.4: Typical supervised learning training and testing workflow for a regression task.

4.3.3 Model training

For a supervised learning task, the goal is to find a relationship between input data and corresponding target labels, as controlled by the model's parameters.[212] In the training process, the data is used to evaluate various choices of parameters to determine the best-performing parameters. This evaluation is achieved through a loss function, which measures how well the model's predictions match the data.[216] The main objective of the learning task is to minimize the chosen loss function.[212] This optimization is typically carried out iteratively, starting with an initial choice of parameters and then using an optimization routine to update the parameters based on the gradient of the loss function (Figure 4.4). After training is completed and the final model parameters are selected, the model is tested on test data, which consists of observations not included in the training set.

4.3.3.1 Bias-variance trade-off

Bias and variance are fundamental concepts in ML. Bias refers to the systematic error introduced by a model's assumptions or simplifications, indicating how far the model's predictions deviate from the true values.[216] Variance quantifies the model's sensitivity to small fluctuations in the training data, measuring the variability in its predictions.[216] Underfitting occurs when a model is too simplistic and fails to capture the underlying patterns in the data.[216] It exhibits high bias and low variance, resulting

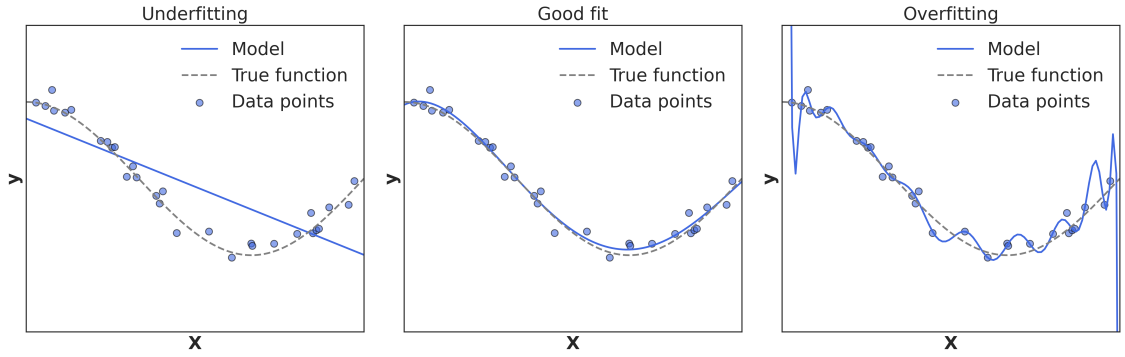


Figure 4.5: Examples of underfitting, good fit, and overfitting for a regression task. The plot showcases three different models applied to a dataset. The true function (grey dashed line) represents the underlying relationship between the input variable (X) and the output variable (y). The data points (blue solid line) are scattered around the true function with some added noise.

in poor performance on both training and test/validation data. In contrast, overfitting arises when a model becomes overly complex and fits the training data too closely.[216] It has low bias but high variance, which leads to excellent performance on the training data but poor generalization to new, unseen data. Examples of underfitting, good fit, and overfitting are shown in Figure 4.5.

The goal of model training is to find the right balance between overfitting and underfitting, known as the bias-variance trade-off.[212] The aim is to develop a model that generalizes well to new data by capturing the underlying patterns without overfitting to noise or underfitting due to excessive simplicity. Underfitting can be addressed by increasing model complexity, incorporating more relevant features, or utilizing more sophisticated algorithms.[212] To address overfitting, techniques such as regularization or early stopping can be employed.[225]

4.3.3.2 Cross-validation

Perfectly reproducing finite training data does not guarantee good generalization on new data.[212] Validation data is crucial in addressing this issue. By further dividing the training data into a set for model training and a validation set, one can gauge how well the model would generalize to new, unseen data. Cross-validation (CV) is

commonly employed to reduce dependence on the specific choice of validation sets during training.[216] It involves dividing the available dataset into multiple subsets or folds. The model is then trained and evaluated multiple times, each time using a different subset as the validation set and the remaining subsets as the training set. One of the most common forms of cross-validation is k-fold cross-validation, where the dataset is divided into k equally sized subsets or folds.[226–228] The model is trained and evaluated k times, each time using a different subset as the validation set and the remaining subsets as the training set. The performance is averaged across the k folds.

4.3.3.3 Class imbalance

Class imbalance occurs when the classes are not equally represented in the training data. When dealing with imbalanced datasets, standard model evaluation metrics like accuracy can be misleading.[229,230] Since the majority class dominates the dataset, a model that predicts only the majority class could achieve high accuracy, despite providing limited practical value.

To address the class imbalance problem during model validation and evaluation, several strategies can be employed. Metrics such as precision, recall, F1-score, and area under the ROC curve are more informative for imbalanced datasets.[229,230] These metrics focus on the performance of the minority class and provide a clearer understanding of the model's predictive ability.

Resampling can be used to oversample the minority class or undersample the majority class. Synthetic Minority Over-sampling Technique (SMOTE)[231] can be used to generate synthetic samples for the minority class by interpolating between existing data points, which helps to increase the representation of the minority class without exact duplication. Giving more weight to misclassified instances of the minority class can also help to address class imbalance.[229,230]

4.3.4 Supervised learning algorithms

In supervised learning, the goal is to learn a mapping between input attributes and corresponding target variables based on labelled training data.[212] The training data consists of a set of observations, denoted as $\mathbf{X}$, where each observation is associated with a target variable, denoted as $\mathbf{y}$. The training data is organized as follows: $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]$, where $\mathbf{x}_i$ represents the feature vector of the i-th observation, and $\mathbf{y} = [y_1, y_2, ..., y_n]$ represents the corresponding target values. Supervised learning encompasses both classification and regression tasks.

4.3.4.1 Classification tasks

In classification, the output variable is categorical, representing distinct classes or labels.[216] The objective is to assign input data points to predefined categories or classes. It aims to identify the class or label that best represents the input based on the available features.

Logistic regression Logistic regression is a widely used linear model that estimates the probability of an event occurring based on input features.[232] It models the relationship between the linear combination of features and the probability of a binary outcome using the logistic (sigmoid) function.

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)}} \tag{4.1}$$

where $P(y = 1|x)$ represents the probability of the event occurring given the input features x, and $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the coefficients or weights associated with each input features $x_1, x_2, ..., x_n$. The coefficients are estimated using the method of maximum likelihood, minimizing the logistic loss function. Optimization algorithms like gradient descent update the coefficients iteratively. The model can predict new data by setting

a threshold on the predicted probability.

Support vector machines Support Vector Machines (SVM) seek to find the hyperplane that maximizes the margin between the support vectors (data points closest to the decision boundary) of different classes.[233] The decision boundary is defined by a linear function. SVM uses the hinge loss function to penalize misclassifications and aims to find the optimal hyperplane by solving a convex quadratic optimization problem. The hinge loss function for the data point x with true label y is defined as:

$$L(w^T x, y) = \max(0, 1 - y \cdot (w^T x)) \tag{4.2}$$

where w is the weight vector (coefficients) of the linear function that defines the decision boundary, and $w^T x$ represents the linear combination of the input features and weights. For binary classification, y is either +1 or -1, depending on the class label. The hinge loss function calculates the difference between the predicted score (the linear combination $w^T x$) and the true score (the true label y). If the data point is correctly classified and falls within the margin, the hinge loss is zero. However, if the data point is misclassified or lies on the wrong side of the margin, the hinge loss is positive and proportional to the distance from the correct side of the margin. By minimizing the hinge loss, SVM is encouraged to find a decision boundary that separates the classes with the maximum margin. SVM can also handle non-linearly separable data by using the kernel trick, which maps the data into a higher-dimensional space where the data points might become linearly separable.

Decision trees Decision trees employ a hierarchical structure of rules to classify the data.[234] By iteratively splitting the data based on different features, decision trees create a tree-like structure that assigns data points to specific classes (Figure 4.6A).

A decision tree consists of nodes, edges, and leaves. Internal nodes represent features, and edges represent possible feature values. Leaves represent predicted classes. The decision tree algorithm builds the tree recursively by selecting the best features to split the data based on criteria like Gini impurity or entropy. Gini impurity is the probability of incorrectly classifying a randomly chosen element in the dataset if it were randomly labelled according to the class distribution in the dataset.[235] The term 'entropy' stems from physics and is a measure of the disorder of a system.[236] In ML, entropy is employed as a metric of impurity or disorder within datasets.[216] The objective of decision trees is to minimize impurity or entropy.[234] This process enhances the creation of a decision tree that effectively classifies the data. The algorithm continues until a stopping criterion is met. The decision tree makes predictions on new, unseen data by following the decision path based on the feature values.[234] The leaf reached at the end of the path determines the predicted class.

Decision trees are interpretable, handle categorical and numerical features, and capture non-linear relationships, but they can suffer from overfitting, especially when the tree becomes too deep or complex.[237] Techniques like pruning and ensemble methods such as random forests mitigate overfitting.[235,238]

Neighbor-based methods Neighbour-based algorithms, such as k-nearest neighbours (kNN), classify data points based on their proximity to labelled examples (Figure 4.6B).[239] In the kNN, the feature vectors and their corresponding class labels from the training dataset are stored. When making predictions for a new, unseen data point, the kNN algorithm calculates the distance between the new data point and all the training data points using a chosen distance metric, such as Euclidean. The algorithm selects the k nearest neighbours to the new data point based on the calculated distances. The value of k is a critical factor in the algorithm's performance.[240] A small k value may lead to overfitting and increased sensitivity to noise,

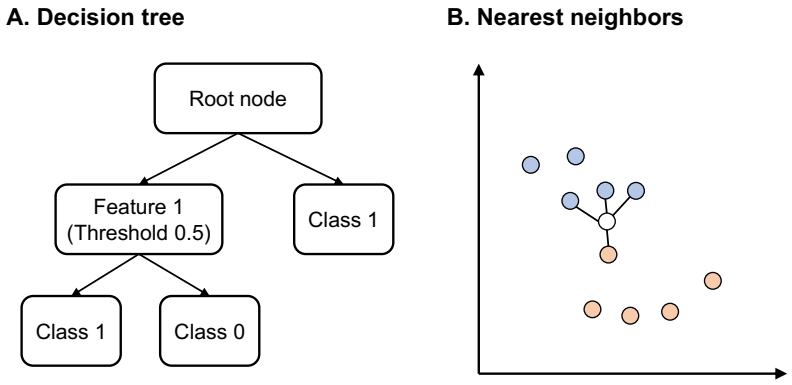


Figure 4.6: Decision tree and nearest neighbour algorithms. (A)Example of a decision tree for a binary classification task. In this example, the root node is the starting point. It leads to a decision node representing "Feature 1" with a threshold of 0.5. If a data point's "Feature 1" value exceeds 0.5, it follows the True path to a terminal node predicting "Class 1". Otherwise, it follows the False path to another terminal node predicting "Class 0". (B) Graphical representation of kNN algorithm the new data point (white) is classified according to its neighbors. The two classes are represented in blue and orange.

while a large k value may cause underfitting and difficulties in capturing local patterns. Once the k nearest neighbours are identified, the algorithm uses a voting scheme to determine the predicted class of the new data point. In binary classification, the class that appears more frequently among the k neighbours is assigned as the predicted class. For multi-class classification, various voting schemes can be employed, such as majority voting or weighted voting.

Ensemble methods Bagging and boosting are two popular ensemble methods used to improve the performance of ML models.[241] Bagging methods create an ensemble by training models on different subsets of the data with replacement.[216] Random Forest[235] and Extra Trees[242] are ensemble learning methods that utilize the bagging technique to improve model performance. Random Forest is an ensemble of decision trees, where each tree is trained independently on a random subset of the training data.[235] The final prediction is determined by aggregating the predictions of all the trees, either through majority voting for classification or averaging for regression. Random Forest reduces the risk of overfitting, making it more robust and less prone to variance compared to individual decision trees. Extra Trees (Extremely Randomized Trees) is also an

ensemble of decision trees trained on bootstrap samples of the data.[242] In Extra Trees, each tree is constructed with additional randomness. Instead of searching for the best split, Extra Trees randomly selects splitting thresholds, resulting in more diverse and less correlated trees. This added randomness increases the diversity among the trees, reducing overfitting and enhancing the model's ability to generalize well to unseen data.

Boosting methods train models iteratively, focusing on correcting mistakes made by previous models.[216] It gives more weight to challenging data points and combines weak models into a strong learner. Adaboost (Adaptive Boosting) is a boosting algorithm used for ensemble learning, which combines multiple weak learners, often decision trees, to create a strong learner.[243] The weak learners are trained sequentially. Weights are assigned to each weak learner based on their performance. Higher weights are assigned to misclassified data points, allowing subsequent models to prioritize those data points and learn from their errors. The final prediction is obtained by combining the predictions of all weak learners using weighted voting. Adaboost handles difficult data points by emphasizing misclassified samples during training and adapts to the complexity of the data by progressively improving its performance. However, Adaboost can be sensitive to noisy data and outliers.[244,245] In summary, bagging reduces variance and handles high-dimensional data, while boosting focuses on challenging data points and improves accuracy.[241]

4.3.4.2 Regression tasks

In regression, the output variable is continuous, representing a numerical value or a range of values.[216] The objective is to predict a quantitative output based on the input features.

Linear regression Linear regression assumes a linear relationship between the inputs and the target variable, with the objective of determining the line or hyperplane of best fit that minimizes the differences between the predicted and actual values. In simple linear regression, there is a single input feature, denoted as x, and a corresponding continuous output variable, denoted as y. The relationship between x and y can be represented by the equation of a straight line, where β_0 represents the y-intercept (the value of y when $x = 0$), and β_1 represents the slope of the line.

$$y = \beta_0 + \beta_1 x \tag{4.3}$$

In higher dimensions where there are more than one independent variables x, the relationship can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n \tag{4.4}$$

Ordinary least squares regression The ordinary least squares (OLS) method works by minimizing the sum of squared differences between the predicted and actual values of the target variable. It assumes several assumptions, such as linearity, independence of errors, constant variance of errors, and normality of error terms.

For a dataset with n observations where each observation consists of an input feature x_i and the corresponding continuous output variable y_i, the relationship between the inputs and the target variable is assumed to be linear.

$$y_i = \beta_0 + \beta_1 x_i \tag{4.5}$$

To estimate the values of the coefficients β_0 and β_1 that best fit the given data, OLS minimizes the sum of squared differences between the predicted and actual values of

the target variable, or the residual sum of squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{4.6}$$

The RSS is differentiated with respect to β_0 and β_1 and set the derivatives to 0. Solving these equations simultaneously provides the estimates for β_0 and β_1, denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The model can then be represented as Eq 4.7 and used to make predictions on new, unseen data by substituting the values of x into the equation.

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{4.7}$$

OLS provides unbiased estimates of the regression coefficients but may be susceptible to overfitting when the number of predictors is large compared to the number of observations. Regularization techniques can be used in ML to prevent overfitting and improve model performance.

LASSO regression LASSO (Least Absolute Shrinkage and Selection Operator) introduces an L_1 regularization term that adds the sum of the absolute values of the coefficients to the loss function.[246] This penalty encourages sparsity in the coefficient values, effectively performing feature selection. The loss function for LASSO regression is defined as:

$$\text{minimize} \left\{ RSS + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.8}$$

where RSS represents the residual sum of squares, β_j are the coefficients, and p is the number of input features. λ controls the amount of regularization applied in LASSO regression. A higher λ value increases the level of shrinkage and sparsity in the model, reducing the impact of less influential features. Conversely, a lower λ value results in

less shrinkage and allows more features to contribute to the model. The estimated coefficient $\hat{\beta}_{\text{lasso}}$ is:

$$\hat{\beta}_{\text{lasso}} = \text{argmin} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \tag{4.9}$$

In the L_1 regularization, the penalty surface exhibits sharp edges forming a diamond shape. This indicates that some coefficient values can become exactly 0, resulting in a model with fewer features and providing interpretable and simpler models. This is particularly useful when dealing with high-dimensional datasets containing many potentially redundant or irrelevant features.

Ridge Regression Ridge regression introduces a L_2 penalty term to the OLS loss function, which adds the sum of the squared coefficients to the loss function.[247] This penalty promotes smaller and more distributed coefficient values. The loss function for Ridge regression is defined as:

$$\text{minimize} \left\{ RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{4.10}$$

The estimated coefficient $\hat{\beta}_{\text{ridge}}$ is:

$$\hat{\beta}_{\text{ridge}} = \text{argmin} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{4.11}$$

By adding the squared magnitudes of the coefficients to the loss function, Ridge regression distributes the impact of correlated features more evenly, preventing any single feature from dominating the model.

L_1 vs L_2 regularization The key distinction between L_1 and L_2 regularization lies in their impact on coefficient shrinkage and feature selection.[212] LASSO regression tends to drive some coefficients to 0, effectively performing automatic feature selection. In contrast, Ridge regression reduces the magnitude of all coefficients without reducing them to 0, leading to less feature sparsity. Therefore, LASSO regression is particularly useful when dealing with high-dimensional datasets and identifying the most relevant features, while Ridge regression can help reduce the impact of multicollinearity and stabilize the model.

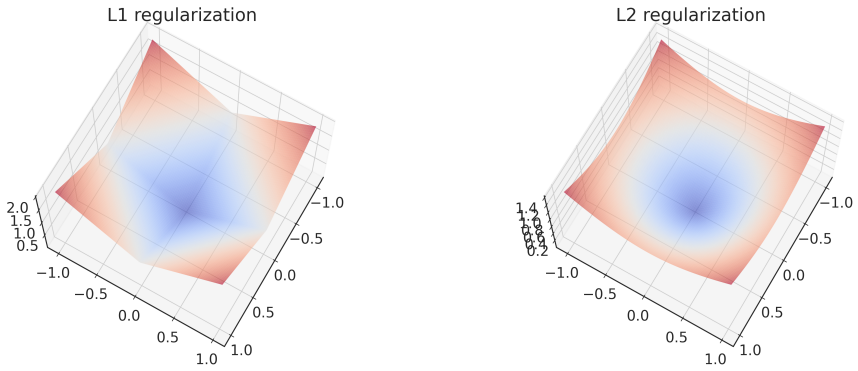


Figure 4.7: L1 and L2 regularization. The L_1 and L_2 penalty functions are shown in 3D. The penalty in both cases is 0 at the centre, indicating that the weights are also 0, rendering the model ineffective. As the weights attempt to minimize the function, they move away from the centre and towards the outer regions. In the L_1 regularization (left), the penalty surface exhibits sharp edges forming a diamond shape. The corners represent instances where one of the axes or features becomes 0, resulting in sparse matrices. In the L_2 regularization (right), the penalty surface exhibits a rounded bowl-like shape, indicating that all coefficients contribute to the model.

4.3.5 Model evaluation

4.3.5.1 Evaluation metrics for classification tasks

For classification, there are four prediction outcomes. True positive (TP) is the number of correctly predicted positive data points. True negative (TN) is the number of correctly predicted negative data points. False positive (FP) is the number of incorrectly predicted positive data points. False negative (FN) is The number of incorrectly predicted negative data points.

A confusion matrix is typically plotted as a summary of the performance of a classification model (Figure 4.8). By analyzing the values in the matrix, various metrics can be calculated to assess the model's accuracy and error rates:

- Accuracy (ACC) measures the proportion of correctly classified data points over the total number of data points. A higher accuracy indicates a better-performing model in terms of correctly predicting class labels.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{4.12}$$

- Precision (P) measures the proportion of true positive predictions over the total predicted positive data points. It focuses on the accuracy of positive predictions. Higher precision indicates fewer false positives.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4.13}$$

- Recall (R) or Sensitivity or True Positive Rate (TPR) measures the proportion of true positive predictions over the total actual positive data points. It focuses on the ability of the model to identify positive data points correctly. Higher recall indicates fewer false negatives.

$$\mathrm{R} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{4.14}$$

- Specificity or True Negative Rate (TNR) measures the proportion of true negative predictions over the total actual negative data points. It focuses on the ability of the model to identify negative data points correctly. Higher specificity indicates fewer false positives.

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{4.15}$$

- F1-Score is the harmonic mean of precision and recall, providing a balanced measure of the two. It is particularly useful when a class imbalance exists. A higher F1-Score indicates better overall performance.

$$\text{F1-Score} = 2 \times \frac{\text{P} \times \text{R}}{\text{P} + \text{R}} \tag{4.16}$$

A Receiver Operating Characteristic (ROC) curve is also typically plotted to evaluate the performance of a binary classification model. It is created by plotting the TPR (sensitivity) against the FPR (1-specificity) at various classification thresholds (Figure 4.8). The area under the ROC curve (AUC) summarizes the overall performance of the model. The AUC ranges from 0 to 1, where a value of 0.5 indicates a random classifier and a value of 1 represents a perfect classifier.

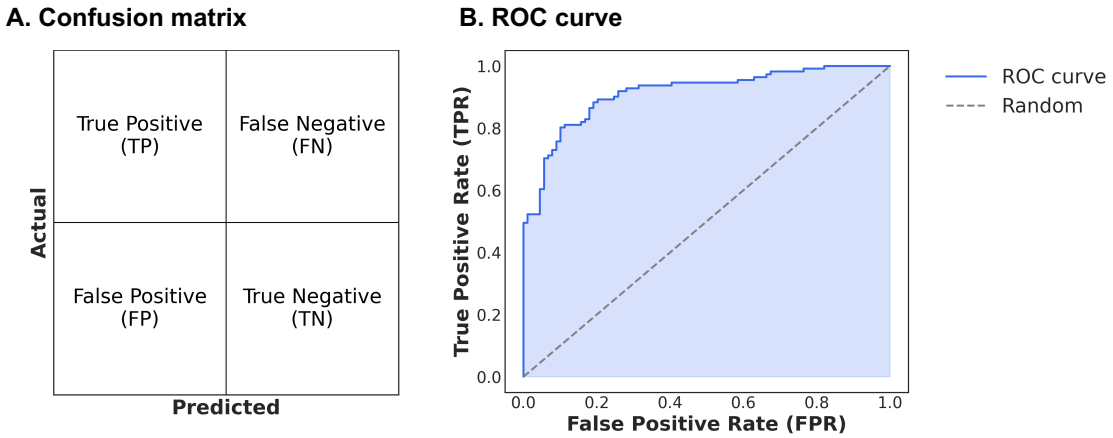


Figure 4.8: Evaluation metrics for classifiers. (A) Confusion matrix for a classification model. The confusion matrix presents the performance evaluation of a classification model, showing the predicted and actual class labels. The matrix consists of four regions representing True Positive (TP), False Negative (FN), False Positive (FP), and True Negative (TN) outcomes. (B) Example of a ROC curve. The ROC curve (blue solid line) illustrates the trade-off between the True Positive Rate (TPR) and False Positive Rate (FPR) as the classification threshold is varied. The diagonal dashed line represents a random classifier, while the plotted curve demonstrates the model's superior performance. The area under the ROC curve (AUC, blue-shaded region) indicates the model's ability to distinguish between positive and negative classes.

4.3.5.2 Evaluation metrics for regression tasks

For regression, the common evaluation metrics are:

- Mean Squared Error (MSE) measures the average of the squared differences between the predicted and actual values. Lower MSE values indicate better accuracy in predicting continuous numerical values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.17}$$

- Root Mean Squared Error (RMSE) is the square root of MSE, providing a measure in the same units as the target variable. It represents the average magnitude of the prediction errors. Lower RMSE values indicate better accuracy.

$$\text{RMSE} = \sqrt{\text{MSE}} \tag{4.18}$$

- Mean Absolute Error (MAE) measures the average of the absolute differences between the predicted and actual values. It represents the average magnitude of the errors, regardless of their direction. Lower MAE values indicate better accuracy.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4.19}$$

- R-squared (R^2) measures the proportion of the variance in the target variable that is explained by the model, with values ranging from 0 to 1. A higher R^2 indicates a better model performance.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{4.20}$$

where $\hat{y}_i$ is the predicted value, y_i is the actual value, and $\bar{y}$ is the mean of the actual values.

4.4 Materials and method

All ML models were trained using *PythiaCHEM* (`https://github.com/duartegroup/PythiaCHEM`). *PythiaCHEM* is an open-source ML toolkit, primarily written by Stamatia Zavitsanou in the Duarte group. I contributed to code development in Tanimoto matrix visualization, automatic categorical feature identification, permutation feature importance, and fingerprint bit importance analysis and visualization. I also contributed to the testing and deployment of *PythiaCHEM* .

PythiaCHEM is organized into six modules (Figure 4.9), each containing relevant functions accessible from Jupyter Notebooks.[248] These modules are documented and structured based on their intended purpose, ensuring ease of use for both experienced and novice users. This approach keeps the Jupyter Notebooks uncluttered while providing users with the flexibility to add their functions to the appropriate module. Notebooks are organized and named according to their aim, offering standard workflows for regression and classification tasks (Figure 4.9). The workflows can be customized by adding or removing steps to tailor them to the specific research question. This structured approach provides a flexible framework for implementing various modules and functions included in the toolkit.

In the following section, I discuss in detail the Jupyter Notebooks used in this project as part of the *PythiaCHEM* package.

Notebook 1: Dataset Analysis This notebook provides a comprehensive exploration of the dataset through various analytical and visualization techniques. The goal is to gain insights into target distribution, chemical diversity, and the potential influence of binding motifs on target and class distributions.

First, the necessary modules (scaling, fingerprints, and molecules and images) are imported. After this, the dataset is imported as a .csv file containing the SMILES

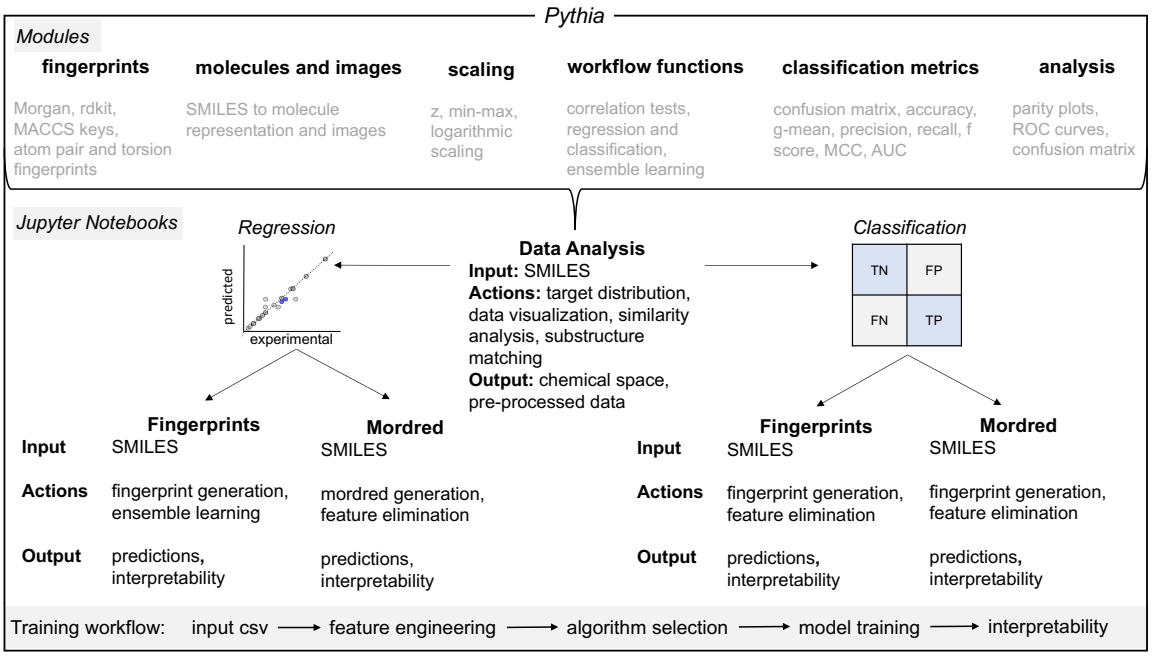


Figure 4.9: Pythia workflow. Pythia is organised into six modules, each containing relevant functions. The modules are used by the Jupyter Notebooks to perform classification and regression tasks. In every case, the training workflow is the same. From a CSV file, the appropriate features and algorithms are selected, and the models are trained and explained.

entries for the molecules under study, along with any other information The standard analysis comprises investigating the distribution of the target variable, assessing class imbalances, visualizing 2D chemical structures, and computing the Tanimoto similarity matrix to gauge structural diversity. Chemical space visualization is also supported, which uses techniques like PCA,[249] t-SNE,[250] and UMAP[251] to gain insights into the underlying structure and relationships between the molecules in the dataset. It can help visualize molecular similarities and identify potential clusters or trends. This is especially useful when exploring the chemical space or assessing the diversity of a compound library.

Furthermore, it introduces an advanced analysis section, optional for users, focusing on recognized binding motifs for anion transporters. This section covers substructure matching against binding motifs, exploring overlap between motif-specific datasets, and assessing Tanimoto similarity within substructure sets. Additionally, it delves

into the target and class distributions of each substructure dataset, shedding light on potential patterns and imbalances.

This comprehensive notebook combines standard exploratory analyses with optional advanced analyses that delve into substructure-specific insights. It enables users to gain a deep understanding of the dataset's properties and the potential impact of binding motifs on various aspects of the data.

Notebook 2: Classification-Fingerprints This notebook is specifically designed for cases where target values are discrete, making it a suitable tool for binary classification tasks. It walks users through techniques for exploring features tailored to analyze categorical data, with a particular focus on binary classification scenarios.

- **Fingerprinting generation**: *PythiaCHEM* supports various fingerprints options, including Morgan fingerprints,[220] RDKit fingerprints,[222] MACCS keys,[217] atom pair fingerprints,[218] and torsion fingerprints.[219] For this example, the usage of Morgan fingerprints is illustrated. These circular fingerprints encode the presence of substructures within molecules. An initial preprocessing step filters out fingerprint bits with values of 0 across all molecules.
- **Custom features addition (optional)**: Users have the choice to introduce custom features. Specifically, to ensure the inclusion of chloride recognition sites, a one-hot encoding approach is used to indicate the presence (1) or absence (0) of specific motifs. Moreover, experimental conditions like internal chloride concentration (in mM) and the time interval between transporter initiation and detergent addition (in seconds) are incorporated as additional features. This step is optional and can be skipped if custom features are not available.
- **Handling imbalanced classes**: Recognizing potential class imbalances, the Synthetic Minority Over-sampling Technique for Nominal and Continuous Fea-

tures (SMOTEN) is implemented to generate synthetic data points.[231] This process aids in achieving a balanced representation of classes, which is pivotal for accurate model training.

- **Model training**: Regardless of the chosen descriptors, a consistent methodology is applied for model training and evaluation. Two primary approaches are adopted:

    1. **Cross-validation approach**: Utilizes k-fold test classifiers with optimization functions from the workflow functions module. The dataset is divided into k folds, where the model is trained on k-1 folds and evaluated on the remaining fold. This process is repeated k times, with each fold serving as the validation set once. Model performance is averaged across folds, yielding a robust estimation of the model's capabilities.

    2. **Train-test split approach**: Employs split test classifiers with optimization functions. The dataset is divided into training and test subsets. The model is trained on the training subset and evaluated on the test subset. Unlike the cross-validation approach, the training set is not further divided into multiple folds for validation. This approach offers simplicity and computational efficiency.

- **Model evaluation**: Model performance evaluation employs multiple metrics, including Accuracy, Sensitivity (Recall), Specificity, Matthew's Correlation Coefficient (MCC), Precision, and G-mean. Confusion matrics, Receiver Operating Characteristic (ROC) curves, and precision/recall curves can also be plotted.

- **Model interpretation**: Model interpretation is facilitated by extracting feature importance using sklearn's capabilities. Additionally, permutation feature importance and SHAP (SHapley Additive exPlanations) analysis are conducted.[252] SHAP summary plots are generated to visualize the influence of each feature on

model predictions. Important fingerprint bit visualization is also supported.

This notebook equips users to perform binary classification, efficiently evaluate models, and interpret their features. It provides valuable insights into the classification process, enabling a comprehensive understanding of binary classification scenarios.

Notebook 3: Classification-Mordred This notebook primarily revolves around the process of generating features using the Mordred calculator to create descriptors for molecules.[223] Specifically, the emphasis is on 2D Mordred descriptors. The Mordred calculator is capable of generating an extensive range of descriptors, and for this analysis, I computed 1613 2D Mordred descriptors.

After feature generation, data cleaning is performed to enhance data quality and relevance. This is achieved through the following steps:

1. Dropping descriptors (columns) with more than 90% missing values. In cases where the Mordred calculator fails to generate descriptors for all molecules, missing values are introduced.

2. Removing columns with a standard deviation less than 0.5. Columns with limited variability contribute less meaningful information to the analysis.

An essential subsequent step is identifying descriptors relevant to predicting the target. Feature selection is conducted using correlation analysis, which tests each feature individually for its correlation with the target property and with other features. Features that fail specific statistical criteria are discarded. This is accomplished by utilizing the *find_correlating_features* function from the *workflow_functions* module. This function takes the computed Mordred descriptors as input and returns features that exhibit Pearson or Spearman correlation coefficients with target values surpassing a user-defined threshold.

Pearson correlation measures the linear relationship between two continuous variables.

It quantifies the strength and direction of the linear association between each feature X and the target variable Y. Pearson correlation, ρ, is calculated as:

$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{4.21}$$

where $\text{cov}(X, Y)$ represents the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively. ρ ranges from -1 to 1, where a value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. It ranges from -1 to 1, where a value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship.

Spearman correlation, also known as rank correlation, assesses the monotonic relationship between two variables. Spearman correlation is calculated by first ranking the values of each feature and the target variable, and then calculating the Pearson correlation between the ranked values. Pearson correlation is suitable for continuous data, while Spearman correlation is preferable for categorical data. The function can also handle numerical features with missing values, filling them with column means. In this example, Spearman correlation coefficients are computed to identify descriptors that correlate with the target value.

Similar to Notebook 2, custom features are added. A balanced dataset is generated by employing additional sampling techniques to address class imbalance. Subsequently, the notebook proceeds with model training, evaluation, and interpretation, adhering to the approaches outlined in Notebook 2.

Notebook 4: Regression-Fingerprints This notebook focuses on regression tasks where the target values are continuous and offers a rapid approach to building a baseline model and assessing its predictive capabilities.

The notebook uses as input (1) the dataset created by Notebook 1 or (2) provided by the user. For this example, the usage of Morgan fingerprints is illustrated. Users have the choice to introduce custom features. The custom features added here are the same as the ones in the previous Notebook 2. Regardless of the chosen descriptors, a consistent methodology is applied for model training and evaluation. Two training approaches adopted here are the same as Notebook 2. Model performance evaluation employs multiple metrics, including (1) Mean Squared Error (MSE), which measures the average of the squared differences between the predicted and actual values. Lower MSE values indicate better accuracy in predicting continuous numerical values. (2) Mean Absolute Error (MAE), which measures the average of the absolute differences between the predicted and actual values. It represents the average magnitude of the errors, regardless of their direction. Lower MAE values indicate better accuracy. (3) R-squared (R^2), which measures the proportion of the variance in the target variable that is explained by the model, with values ranging from 0 to 1. A higher R^2 indicates a better model performance. Model interpretation is facilitated by extracting feature importance using sklearn's capabilities. Additionally, permutation feature importance and SHAP (SHapley Additive exPlanations) analysis are conducted.[252] SHAP summary plots are generated to visualize the influence of each feature on model predictions. Important fingerprint bit visualization is also supported. This notebook equips users to perform regression tasks, efficiently evaluate models, and interpret their features.

Notebook 5: Regression-Mordred Similarly to Notebook 5, this notebook is also used for regression tasks; however, it uses Mordred descriptors. As there are more than 1800 Mordred descriptors that can be calculated, this notebook, similarly to Notebook 3, offers a feature elimination method to retain only relevant descriptors, reducing the noise in the models.

4.5 Dataset

The transport activity is experimentally measured by EC_{50}, which indicates the concentration of transporter required to obtain 50% chloride efflux (in mol% transporter to lipid). Different experimental assays exist, such as the chloride ion-selective electrode (ISE), which directly measures changes in ion concentrations, and the HPTS assays, which monitor the rate of transporter-induced pH change inside vesicles using the fluorescent pH indicator 8-hydroxypyrene-1,3,6-trisulfonic acid (HPTS).[253,254] Due to these variations in the underlying mechanisms of each assay, the transport activity values obtained from different assays are not directly comparable. To ensure consistency in the dataset, here I solely used transport activity data available in the literature that have been measured using the ISE assay. This approach allows for better comparability between the transport activity values as it uses data with consistent experimental methods and conditions. Further, at the time of data collection, transport activity data for the ISE assay was more readily available.

The dataset comprises 304 transport activity values described by $log(EC_{50})$ values, obtained using the ISE assay from 31 scientific articles published in the period 2011-2021.[108–131,199–205] Alongside the transport activity values, I collected relevant experimental conditions, including chloride concentration in the vesicles (in mM) and the time taken from adding the transporter to lysing the vesicles (in seconds). The target values ranged from -3.70 to 1.68 (Figure 4.10A). For the train-test split, I organized the dataset chronologically based on publication dates and adopted a temporal split strategy, as the field continuously evolves to design novel and better transporters. I assessed the chemical space coverage of the entire dataset using the Tanimoto index (Figure 4.10B).[255] The dataset is diverse, as evidenced by the average Tanimoto similarity value of 0.2. The five most common structures are transporters possessing urea, thiourea, tambjamine, squaramide and indole (Figure 4.10C).

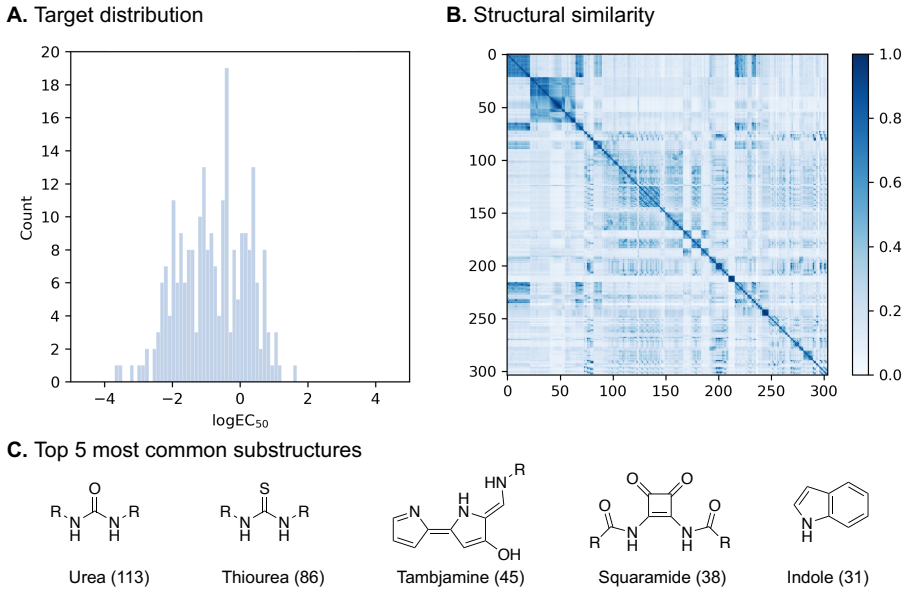


Figure 4.10: Dataset of synthetic anion transporters. (A) Distribution of the transport activities in $log(EC_{50})$. (B) Structural similarity within the dataset. (C) The five most common substructures and their corresponding counts in the dataset.

For classification, I adopted a binary approach, with class 1 representing active transporters with $log(EC_{50}) < 0$ and class 0 representing inactive transporters with $log(EC_{50}) \geq 0$. In this dataset, 178 transporters fell into class 1 and 126 into class 0. Data from years earlier than 2021 formed the training set (260 data points), while data from 2021 comprised the testing set (44 data points) (Figure B5A). Both the training and test sets exhibited a remarkable degree of structural diversity (Figure B5B) and share low similarity between them, indicating that the molecules in the test set are sufficiently distinct from those in the training set. This ensures that the models are evaluated based on unseen data with unique structural characteristics. The robust structural diversity within the training and test sets, combined with the low similarity between them, suggests that the model will generalize well to new and diverse chemical structures, making it a reliable predictor of transport activities for molecules outside the training set. To address class imbalance in the training set, additional sampling points were generated for the minority class (class 0) using SMOTEN.[231]

For regression, in the collected dataset of 304 data points, some synthetic transporters

were not active enough to perform Hill analysis and were reported as greater than a certain cutoff. These arbitrary target values can be used for the classification tasks as they belong to the inactive class (class 0). However, for regression tasks, these arbitrary values are problematic. These artificially assigned values can introduce considerable noise and distortion to the dataset, undermining the accuracy and reliability of any regression analysis. Since regression seeks to establish a quantitative relationship between variables, the inclusion of such non-informative data points can lead to erroneous interpretations, unreliable predictions, and compromised model performance. To address this issue, I exclude these data points, enabling a more meaningful regression analysis that accurately reflects the underlying relationships within the transport activity dataset. After removing arbitrary values, 241 transport activity values remained. I again applied a temporal split. Data from years earlier than 2021 formed the training set (216 data points), while data from 2021 comprised the testing set (25 data points) (Figure B6A).

4.6 Feature engineering

I explored two types of descriptors to represent the anion transporters: Morgan fingerprints[220] and Mordred descriptors.[223]

- **Morgan fingerprints:** I generated Morgan fingerprints as bit vectors with a radius of 2 and size of 1024 bits. After removing features represented by 0 along all fingerprint bits, 353 features remained to describe each transporter.
- **Mordred descriptors:** I first calculated 1613 2D Mordred descriptors. To select the most relevant features, I conducted a correlation analysis, individually testing each feature's correlation with the target property. I used Spearman correlation, which is suitable for both continuous and ordinal variables and captures linear and non-linear associations. Features with correlation coefficients

of 0.2 or higher were retained. To further assess the significance of these features, I performed a two-tailed p-test using the Spearman correlation coefficient as the test statistic. I conducted this test over 5000 random sample permutations. After feature generation and selection, I applied one-hot encoding for categorical features and min-max scaling for continuous features.

To ensure the inclusion of chloride recognition sites, I applied one hot encoding for the presence (1) and absence (0) of urea, thiourea, tambjamine, squaramide, indole, carbazole, acridinone, perenosin, 1,3-bis(benzimidazol-2-yl)pyrimidine, and isophthalamides/dipicolineamides. Additionally, I included the key following experimental conditions as features: internal chloride concentration (in mM) and the time between the transporter to initiate chloride transport addition detergent addition to lyse the vesicles (in seconds).

4.7 Model training and evaluation

Model training follows a consistent methodology, irrespective of the chosen descriptors. Two primary approaches are available for model training and evaluation.

1. **Cross-validation**. This approach employs k-fold test regressor with optimization and $k$fold test classifiers with optimization functions from the workflow functions module. It involves partitioning the data into k-folds of equal size, where the model is trained on k-1 folds and assessed on the remaining fold. The process is repeated k times, with each fold serving as the validation set once. The model's performance is obtained by averaging the performance scores from each fold. This approach leverages the entire dataset for both training and evaluation, providing a robust estimation of the model's performance and generalization capabilities.

2. **Train-test split**. This approach uses the split test regressors with optimization and split test classifiers with optimization functions from the workflow functions module. The dataset is divided into two distinct subsets: a training set and a test set. The model is trained on the training set, while the test set remains separate and is employed to evaluate the model's performance. Unlike approach (1), the training set is not further divided into multiple folds for validation. This approach offers simplicity and computational efficiency.

In both approaches, grid search optimization is executed to determine the optimal hyperparameters for each algorithm, enhancing its performance.

Class imbalance is a common challenge in classification tasks where one class has significantly fewer instances than the other(s). This imbalance can lead to biased model performance, as classifiers tend to favour the majority class due to its larger representation in the training data. As a result, minority class instances are often misclassified, leading to poor generalization and low predictive accuracy. To address the class imbalance issue, I tested two methods: class weights and synthetic sampling with SMOTE.[231] I found that class weights did not improve model performance while using SMOTE to oversample the minority class (class 0) resulted in a significant enhancement in model performance. Thus, here I discuss the results trained on balanced datasets, generated by additional sampling for the minority class (class 0).[231] Details on the model performance with class weights can be found in Appendix Section B.2.1.

4.7.1 Classification models

I trained ML models on Morgan fingerprints and Mordred descriptors separately. Several classifiers were tested, including Ada Boost,[243] Extra Trees (ET),[242] logistic regression (LR),[232] decision tree (DT),[234] Gaussian process (GP),[256] and k-nearest

neighbours (kNN).[239]

To evaluate model performance, I use multiple evaluation metrics, including (1) Accuracy, which measures the ratio of correctly classified instances to the total number of instances in the dataset, (2) Sensitivity (True Positive Rate or Recall, which measures the proportion of true positive instances that are correctly identified by the model, (3) Specificity (True Negative Rate), which measures the proportion of true negative instances that are correctly identified by the model, (4) Matthew's Correlation Coefficient (MCC), which is a correlation coefficient between the observed and predicted binary classification, (5) Precision (Positive Predictive Value), which measures the proportion of true positive instances among all instances predicted as positive by the mode, and (6) G-mean (Geometric Mean), which is the geometric mean of sensitivity and specificity. Details on how these metrics are calculated can be found in Section 4.3.5.

In the following section, I discuss the model performances of both the cross-validation approach and the train-test split approach. I also explore model interpretability in the train-test split approach.

4.7.1.1 Model performance on Morgan fingerprints

Approach 1: cross-validation on the entire dataset For the cross-validated approach on the entire dataset, GP and ET emerged as the top-performing classifiers (Table 4.1). GP exhibited the highest Accuracy among all classifiers of 0.88, indicating its adeptness in making accurate predictions across both classes. This was complemented by a robust Sensitivity of 0.86 and a high Specificity of 0.89, underscoring its proficiency in correctly identifying positive and negative instances alike. The achieved MCC of 0.75 further validated its balanced performance and predictive precision. Additionally, GP's Precision of 0.89 and G-mean of 0.88 demonstrated its capacity for accurate predictions.

Similarly, ET displayed competitive results, with an Accuracy of 0.86, Sensitivity of 0.86, and Specificity of 0.86. This balanced performance was mirrored in its MCC of 0.72, Precision of 0.86, and G-mean of 0.86. These metrics collectively demonstrate ET's capability to achieve accurate predictions while maintaining an equilibrium between sensitivity and specificity.

	Accuracy	Sensitivity	Specificity	MCC	Precision	G-mean
Nearest neighbors	0.79	0.75	0.84	0.59	0.83	0.79
Gaussian process	0.88	0.86	0.89	0.75	0.89	0.88
Decision tree	0.82	0.82	0.81	0.63	0.82	0.82
Extra trees	0.86	0.86	0.86	0.72	0.86	0.86
Ada boost	0.81	0.80	0.82	0.62	0.82	0.81
Logistic regression	0.82	0.84	0.84	0.64	0.81	0.82

Table 4.1: Model performance cross-validated on the entire dataset. The anion transporters are represented by Morgan fingerprints. MCC: Matthew's correlation coefficient.

Analysis of the confusion matrices shows GP and ET has performance in differentiating between positive and negative instances (Figure 4.11). GP predicts more false negatives (FN = 25) than false positives (FP = 19). ET has a more balanced performance in distinguishing between positive and negative instances (FN = 25, FP = 25). For designing new transporters, ET is preferred over GP, as it avoids prematurely discarding potential molecular candidates. Generally, for molecular design, a bias towards FP is strategic. This approach, while potentially leading to additional experimentation, safeguards against overlooking potentially promising molecular candidates. By choosing a model that predicts higher FP than FN, the risk of prematurely discarding potential molecular candidates is minimized.

The Receiver Operating Characteristic (ROC) curves for the GP and ET models are also analyzed. The ROC curve is a graphical representation that illustrates the trade-off between the True Positive Rate (Sensitivity) and the False Positive Rate (1 - Specificity) at various classification thresholds. Analysis of ROC curves indicates that both GP and ET exhibit good classification performance. The Area

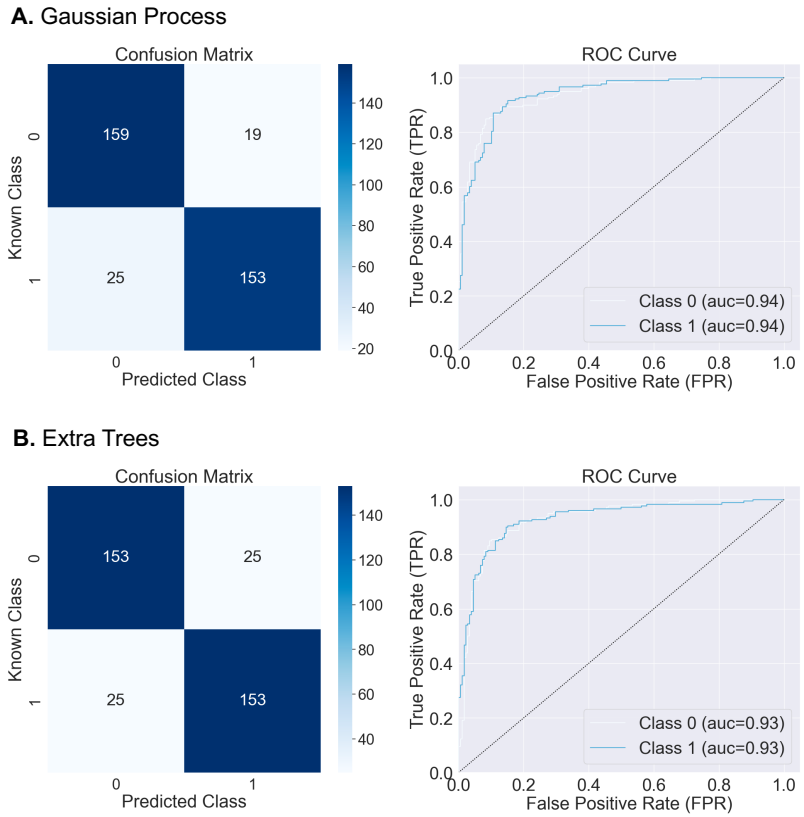


Figure 4.11: Model performance cross-validated on the entire dataset. The anion transporters are represented by Morgan fingerprints. The confusion matrices and the Receiver Operating Characteristic (ROC) curves for the top two classifiers, (A) GP and (B) ET models.

Under the ROC Curve (AUC) provides a single measure of the model's ability to discriminate between positive and negative instances. A higher AUC value indicates better discrimination and better model performance. Both GP and ET achieve high AUC values of 0.94 and 0.93 respectively. This demonstrates that both models possess high discriminatory abilities and rapidly improve their classification skills as the decision threshold varies.

Approach 2: train-test split For the train-test split training approach, GP and ET demonstrated the highest overall performance, achieving an Accuracy of 0.98 and a perfect Sensitivity of 1.00 (Table 4.2), indicating accurate identification of positive cases without any false negatives. Additionally, GP and ET exhibited a high Specificity of 0.95, effectively minimizing false positives. The MCC values of 0.95 for both GP and ET further highlighted their balanced accuracy considering both true positive and true negative rates. LR and Ada Boost also performed well with similar Precision values of approximately 0.96. In contrast, DT showed the lowest overall performance with an Accuracy of 0.73 and a Sensitivity of 0.60, indicating higher false negatives. These findings suggest that GP and ET are the most promising classifiers for this predictive task.

Examination of the confusion matrices indicates that GP and ET display comparable efficacy in discerning between positive and negative instances. Analysis of ROC curves indicates that both GP and ET exhibit exceptional classification performance (Figure 4.12A, B). GP achieves a perfect AUC of 1.00, while ET closely follows with an AUC of 0.99. This demonstrates that both models possess high discriminatory abilities and rapidly improve their classification skills as the decision threshold varies. Both models are well-suited for the binary classification task, and the small difference in AUC values highlights their comparable performance.

	Accuracy	Sensitivity	Specificity	MCC	Precision	G-mean
Nearest neighbors	0.91	0.88	0.95	0.82	0.96	0.91
Gaussian process	0.98	1.00	0.95	0.95	0.96	0.97
Decision tree	0.91	0.88	0.95	0.82	0.96	0.91
Extra trees	0.98	1.00	0.95	0.95	0.96	0.97
Ada boost	0.84	0.80	0.89	0.69	0.91	0.85
Logistic regression	0.91	0.92	0.89	0.81	0.92	0.91

Table 4.2: Model performance on the test set. The anion transporters are represented by Morgan fingerprints. MCC: Matthew's correlation coefficient.

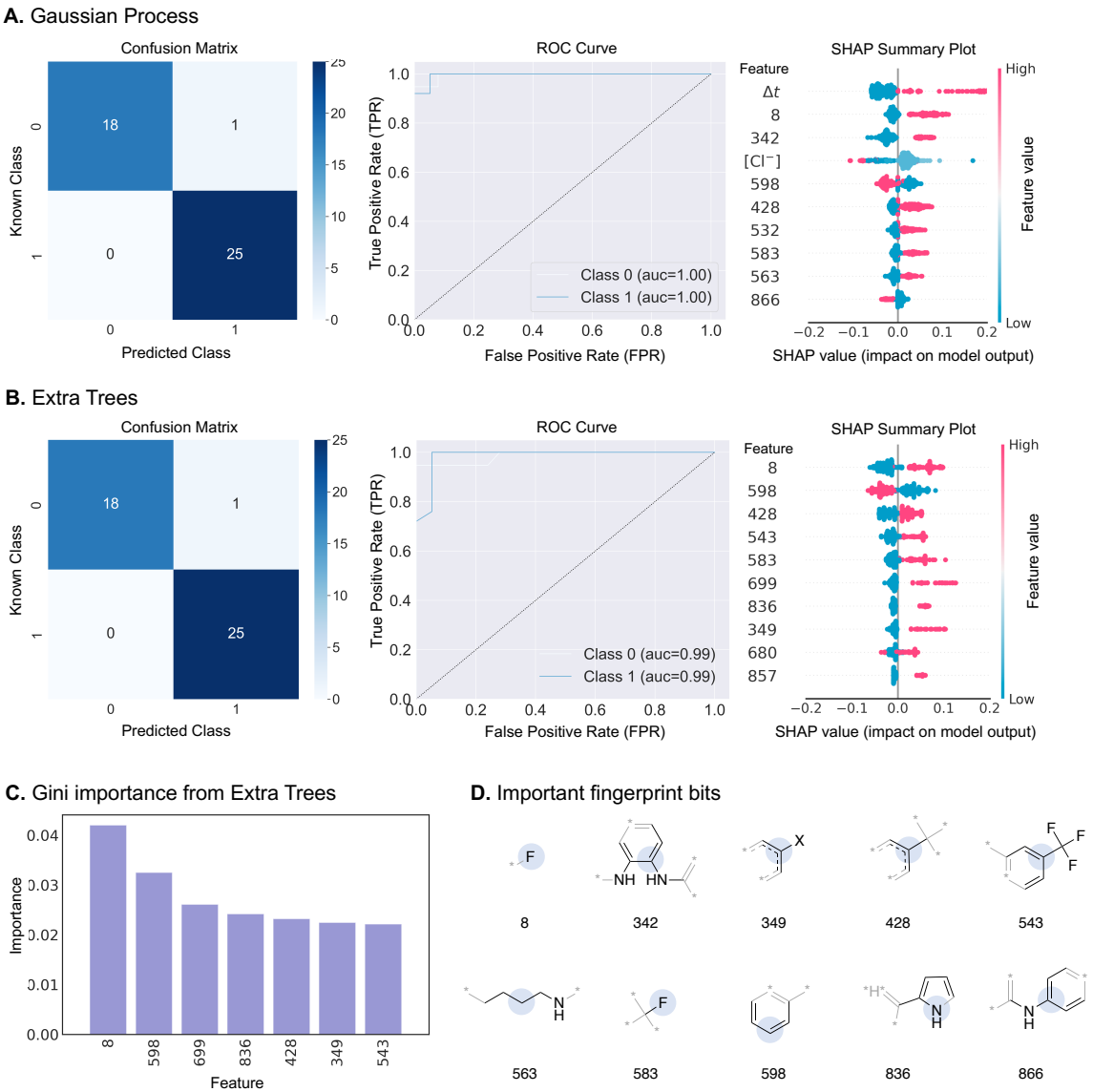


Figure 4.12: Model performance and interpretation on the test set. The anion transporters are represented by Morgan fingerprints. The confusion matrices, the Receiver Operating Characteristic (ROC) curves and SHAP summary plots for the top two classifiers, (A) GP and (B) ET models. For the SHAP summary plots, the x-axis represents the SHAP values, which are centred around 0, and the y-axis represents the feature index. Red hues represent higher feature values, while blue hues represent lower feature values. (C) Gini importance extracted from the ET model. (D) The most important Morgan fingerprint bits.

4.7.1.2 Model interpretability on Morgan fingerprints

Given the comparable predictive performance of GP and ET, I employed SHapley Additive exPlanations (SHAP) to assign importance scores to each input feature[252] (Figure 4.12A, B). The SHAP summary plot visualizes the relationship between feature values and their impact on the prediction of class 1. These scores are centred around 0, with positive SHAP values indicating a higher association with class 1, suggesting that these features contribute to higher predicted probabilities for this class. Conversely, negative SHAP values are indicative of a stronger association with class 0, as they contribute to lower predicted probabilities. The colour spectrum encodes the feature values, with red hues representing higher feature values while blue hues representing lower feature values, which aids in identifying the range and distribution of feature values within the dataset.

The comparison of the SHAP summary plots of the GP and ET models reveals different features as the most influential for each model. The GP model highlighted experimental conditions, with the two most important features corresponding to the time between transporter addition and vesicle lysis (Δt) and chloride concentration inside the vesicles ($[Cl^-]$). The SHAP summary plots suggest that longer Δt leads to a higher predicted probability for class 1 (active transporter). The relationship between $[Cl^-]$ and the predictions are more variable; $[Cl^-]$ that is too large or too small are both observed to correspond to a higher predicted probability for class 0 (inactive transporter). These observations align with the properties of the experimental chloride/nitrate ISE assay. In brief, in the chloride/nitrate ISE assay, the vesicles with $[Cl^-]$ are suspended in an external medium with the same concentration of nitrate. Chloride efflux is monitored using a chloride-sensitive electrode at between transporter addition and vesicle lysis (Δt) and is subsequently fitted to the Hill equation to obtain EC_{50} values. The EC_{50} values obtained in the chloride/nitrate transport assay can

be influenced by both Δt and [Cl^-]. A longer Δt would result in a lower $log(EC_{50})$ as maximum chloride efflux (%) would be higher. The effect of [Cl^-] is more variable and is case-dependent on the transporters. The GP model correctly captures the critical properties of the experimental assay. It is important to note that these experimental conditions affect how the $log(EC_{50})$ values are calculated, but not the transport ability of the transporters themselves. When using these models for making predictions and comparing potential transporters, the Δt and [Cl^-] should be the same across the molecules tested to ensure valid comparison.

In contrast, the ET model places higher importance on the specific molecular features of the transporters, including H-bond donor motifs NH atoms at the binding motifs and atoms near the binding motifs. This suggests that the ET model has learned to identify and prioritize specific fingerprint bits closely associated with the anion-binding sites of the transporters. The model's ability to highlight these critical features makes intuitive sense from a chemistry perspective. The difference in feature importance implies that GP and ET are identifying different underlying patterns in the data. The GP model's probabilistic nature allows it to capture complex, non-linear relationships between features and targets, while ET's decision tree-based approach may be better suited for capturing categorical or discrete interactions between features. The SHAP analysis reveals how each model leverages its inherent characteristics to derive predictions.

As ET is a decision-tree-based model, I also calculated the Gini importance for the features (Figure 4.12C). Gini importance measures the reduction in Gini impurity achieved by each feature across all decision tree nodes in the ensemble. Higher Gini importance indicates that a feature is more critical in splitting the data effectively during the tree-building process, leading to improved overall performance of the decision tree ensemble. It helps to identify the most important features of the model's global performance. While the relative ranking of features differs, there is considerable

overlap between the top important features ranked by Gini and by SHAP.

I visualized the top Morgan fingerprint bits ranked by SHAP and Gini feature importances in both the GP and ET models (Figure 4.12D). These bits correspond to important structural features of the transporters, such as the binding motifs (feature indices 342, 563, 836 and 866). The positive effect of fluorination (feature indices 8, 349, 543 and 583) has also been identified. This is consistent with previous experimental observation by Gale, where fluorinated tren-based transporters were more effective than their parent analogues at transmembrane chloride transport.[115]

4.7.1.3 Model performance on Mordred descriptors

Approach 1: cross-validation on the entire dataset When considering the entire dataset, 128 Mordred descriptors are shown to be statistically significant and selected to describe each transporter. A full list of selected Mordred descriptors can be found in Appendix Section B.2.3. After the discretization of categorical features and the addition of substructure matches and experimental data, there are a total of 141 features for each data point.

Among the classifiers, GP stands out with the highest Accuracy of 0.80 (Table 4.3), indicating its proficiency in making accurate predictions. GP exhibited balanced Sensitivity (0.87) and Specificity (0.70), emphasizing its capability to identify both positive and negative instances. The Precision (0.80) and G-mean (0.78) further validate its capacity for accurate and well-rounded predictions. GP also achieved the highest AUC of 0.86 (Figure 4.13A), showcasing its ability to discriminate between positive and negative instances effectively. ET also displayed noteworthy performance achieving the highest AUC of 0.87 (Figure 4.13B). Other classifiers, including kNN, DT, Ada Boost, and LR, also demonstrated varying degrees of performance.

	Accuracy	Sensitivity	Specificity	MCC	Precision	G-mean
Nearest neighbors	0.75	0.79	0.70	0.49	0.79	0.74
Gaussian process	0.80	0.87	0.70	0.58	0.80	0.78
Decision tree	0.76	0.84	0.66	0.51	0.78	0.74
Extra trees	0.77	0.81	0.72	0.53	0.80	0.76
Ada boost	0.77	0.84	0.67	0.51	0.78	0.75
Logistic regression	0.75	0.81	0.66	0.47	0.77	0.73

Table 4.3: Model performance cross-validated on the entire dataset. The anion transporters are represented by Mordred descriptors. MCC: Matthew's correlation coefficient.

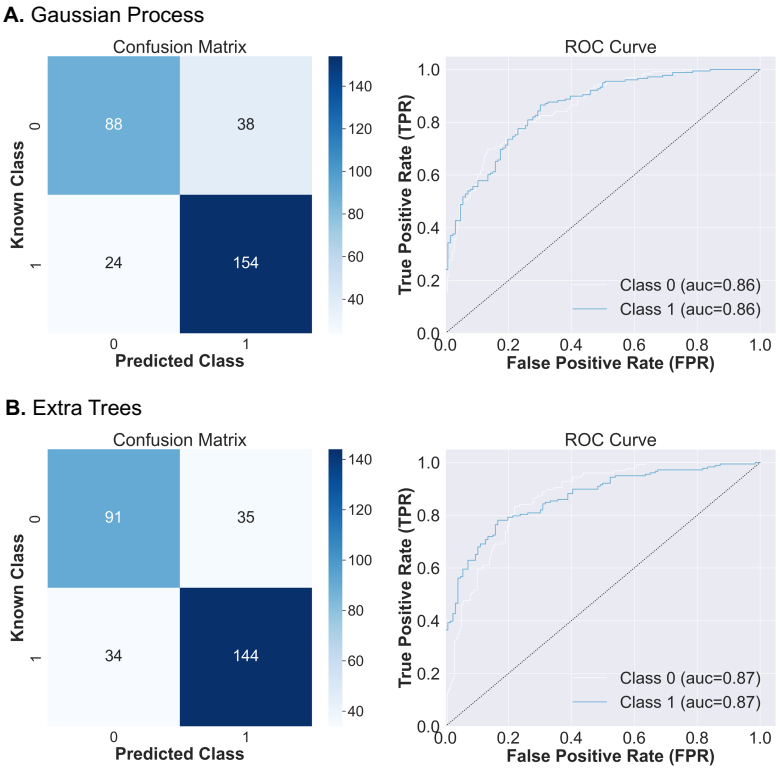


Figure 4.13: Model performance cross-validated on the entire dataset. The anion transporters are represented by Mordred descriptors. The confusion matrices and the Receiver Operating Characteristic (ROC) curves for the top two classifiers, (A) GP and (B) ET models.

Approach 2: train-test split When considering the training set, 86 Mordred descriptors are shown to be statistically significant and selected to describe each transporter. A full list of selected Mordred descriptors can be found in Appendix Section B.2.3. After the addition of substructure matches and experimental data, there are a total of 144 features for each data point.

Similar to the cross-validation approach, GP and ET are the best-performing models with high accuracies of 0.91 and 0.89 respectively (Table 4.4A, B). GP achieved high Sensitivity (0.84) and perfect Specificity (1.00), demonstrating its capacity to accurately identify both positive and negative instances. GP also achieved a perfect Precision of 1.00, showing high predictive reliability. GP also showed a high G-means of 0.92, indicating that the classifier performs well in terms of both positive and negative class predictions. ET also achieved a high Sensitivity (0.92) and Precision (0.88). In contrast, Ada Boost demonstrated challenges in achieving balanced performance on the test set, particularly in terms of Specificity (0.05).

Analysis of the ROC curves reveals that GP has slightly better discriminatory ability than ET, with high AUC values of 0.98 and 0.96 respectively (Figure 4.14A, B). In the confusion matrices, GP yields no FP and four FN predictions. This suggests that GP is highly accurate in correctly identifying instances with positive outcomes and rarely misclassifies negative instances as positive. In comparison, ET demonstrated a balanced performance in distinguishing between positive and negative instances. For designing new transporters, ET is preferred over GP, as it avoids prematurely discarding potential molecular candidates. For designing new transporters, ET is preferred over GP, as it avoids prematurely discarding potential molecular candidates.

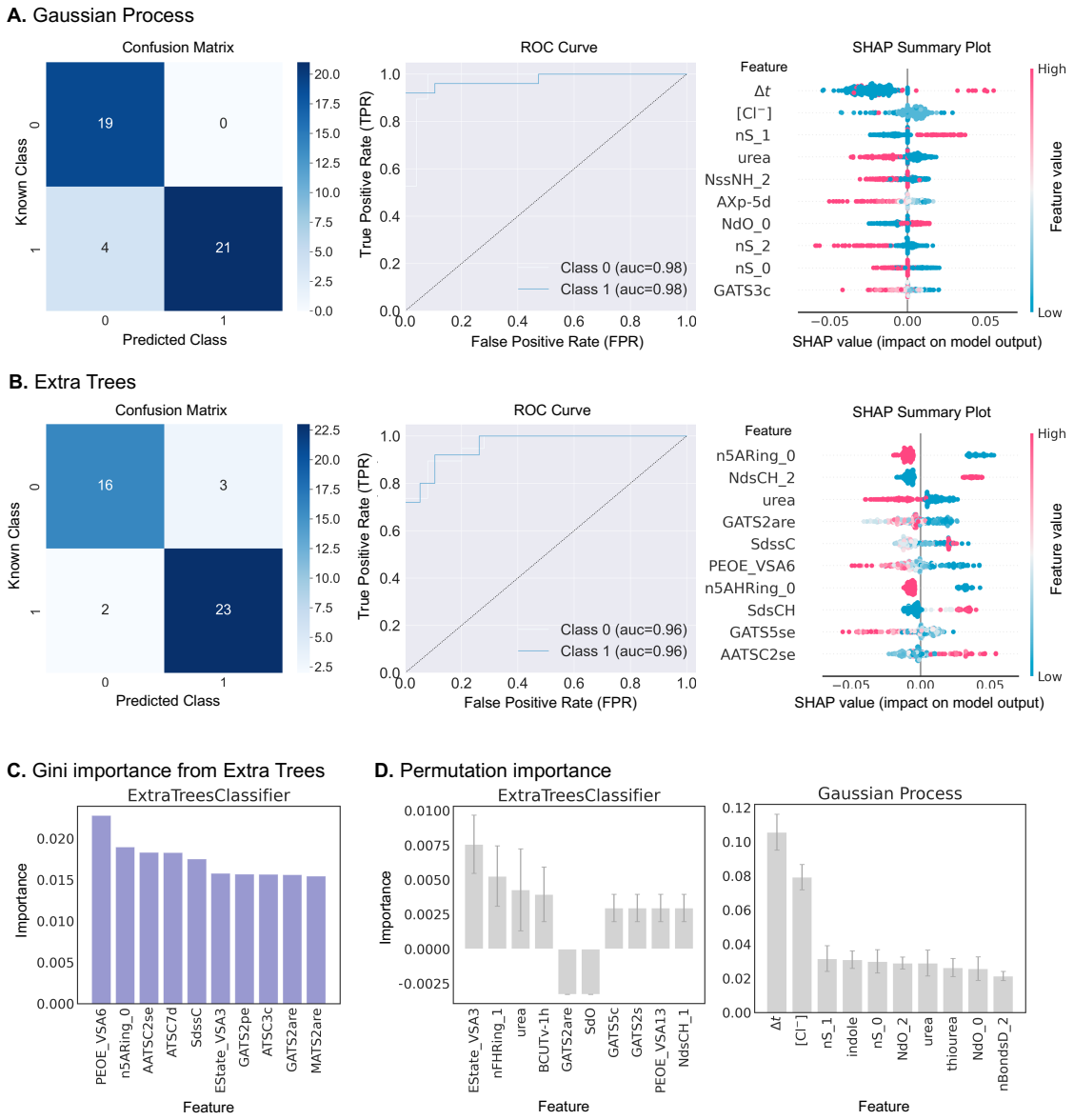


Figure 4.14: Model performance and interpretation. (a) The confusion matrices, the Receiver Operating Characteristic (ROC) curves and SHAP summary plots for the top two classifiers, (a) GP and (b) ET models. For the SHAP summary plots, the x-axis represents the SHAP values, which are centred around 0, and the y-axis represents the feature index. Red hues represent higher feature values, while blue hues represent lower feature values. (c) Gini importance extracted from the ET model. (d) Permutation importance was calculated for the ET and GP models.

	Accuracy	Sensitivity	Specificity	MCC	Precision	G-mean
Nearest neighbors	0.73	0.76	0.68	0.44	0.76	0.72
Gaussian process	0.91	0.84	1.00	0.83	1.00	0.92
Decision tree	0.61	0.48	0.79	0.28	0.75	0.62
Extra trees	0.89	0.92	0.84	0.77	0.88	0.88
Ada boost	0.57	0.96	0.05	0.03	0.57	0.22
Logistic regression	0.70	0.64	0.79	0.43	0.80	0.71

Table 4.4: Model performance on the test set. The anion transporters are represented by Mordred descriptors. MCC: Matthew's correlation coefficient.

4.7.1.4 Model interpretability on Mordred descriptors

Given the comparable predictive performance of GP and ET, I explored model interpretability of both models. I used three methods of model interpretation: (1) I employed SHAP to assign importance scores to each input feature[252] (Figure 4.14A, B). The SHAP summary plot visualizes the relationship between feature values and their impact on the prediction of class 1. These scores are centred around 0, with positive SHAP values indicating a higher association with class 1, suggesting that these features contribute to higher predicted probabilities for this class. Conversely, negative SHAP values are indicative of a stronger association with class 0, as they contribute to lower predicted probabilities. The colour spectrum encodes the feature values, with red hues representing higher feature values while blue hues representing lower feature values, which aids in identifying the range and distribution of feature values within the dataset. (2) As ET is a decision-tree-based model, I also calculated the Gini importance for the features (Figure 4.14D). Gini importance measures the reduction in Gini impurity achieved by each feature across all decision tree nodes in the ensemble. Higher Gini importance indicates that a feature is more critical in splitting the data effectively during the tree-building process, leading to improved overall performance of the decision tree ensemble. It helps to identify the most important features of the model's global performance. (3) I also computed permutation feature importance for both models (Figure 4.14C), which works by shuffling the values of a single feature while keeping the others unchanged and then observing how this shuffling affects the model's performance. If the feature is important, shuffling its values will cause a significant drop in model performance, as the feature's original relationship with the target has been disrupted.

Given the comparable predictive performance of GP and ET, I employed SHAP to assign importance scores to each input feature.[252] As ET is a decision-tree-based

model, I also calculated the Gini importance for the features, which measures the reduction in Gini impurity achieved by each feature across all decision-tree nodes in the ensemble.

A comparison of the SHAP summary plots of the GP and ET models reveals different features as the most influential for each model. The SHAP summary plot of the GP model, once again, highlights key experimental conditions, Δt and [Cl^-] (Figure 4.14A). The number of Sulfur atoms is also found to be important in the GP model. Interestingly, the presence of one Sulfur atom (nS_1) contributes to a higher predicted probability for class 1 (active transporter), while the presence of zero (nS_0) or two Sulfur atoms (nS_2) contributes to a higher predicted probability for class 0 (inactive transporter).

For the ET model, while the relative ranking of features differs, there is considerable overlap between the top important features ranked by SHAP, Gini and permutation feature importances (Figure 4.14B, C and D). SHAP suggests that a higher count of aromatic rings ($n5ARing_0$) might correlate with increased transport ability, while the number of non-hydrogen atoms bonded to methylene groups ($NdsCH_2$) could have a similar impact. The feature with the highest Gini importance is $PEOE_VSA6$, which is THE sum of van der Waals surface area atomic contribution (VSA)[257] for atoms that present Gasteiger charges (PEOE)[258] between 0.15 and 0.20. It intends to capture the direct electrostatic interactions. $PEOE_VSA6$ is also highlighted as a top feature by SHAP, which found a higher value corresponds to a higher predicted probability for class 0 (inactive transporter).

4.7.2 Regression models

I tested several regressors, including LASSO with cross-validation (LassoCV),[246] k-Nearest Neighbors (kNN), Bayesian regressor,[259] Decision Trees (DT),[234] Random Forest (RF),[235] Extra Trees (ET),[242] Ada boost,[243] and XGBoost regressor.[260] To evaluate model performance, I used multiple evaluation metrics, including (1) Mean Squared Error (MSE), which measures the average of the squared differences between the predicted and actual values. Lower MSE values indicate better accuracy in predicting continuous numerical values. (2) Mean Absolute Error (MAE), which measures the average of the absolute differences between the predicted and actual values. It represents the average magnitude of the errors, regardless of their direction. Lower MAE values indicate better accuracy. (3) R-squared (R^2), which measures the proportion of the variance in the target variable that is explained by the model, with values ranging from 0 to 1. A higher R^2 indicates a better model performance. Details on how these metrics are calculated can be found in Section 4.3.5. In the classification section, I demonstrated the purpose and effectiveness of the two model training approaches. In the following section, I discuss the model performances of the train-test split approach for ML.

I initially tested Morgan fingerprints[220] and Mordred descriptor,[223] given their promising performance in the classification model. However, the performance of Mordred descriptors as input features for regression models was unsatisfactory (Appendix Section B.3.3). On the other hand, Morgan fingerprints exhibited reasonable performance in the regression models (discussed later in this chapter). Nonetheless, recognizing the potential for further enhancement, I embarked on an exploration of alternative fingerprint representations. To identify the other suitable fingerprint representations, I undertook a comparative analysis involving atom pair fingerprints,,[218] torsion fingerprints,,[219] Morgan fingerprints,[220] Avalon fingerprints,,[261] and RDKit

fingerprints.[222] This assessment was conducted through a 5-fold cross-validation approach utilizing a ridge regression model without hyperparameter optimization (Figure 4.15). The evaluation criteria employed was the Mean Absolute Error (MAE). Performance increases as the fingerprint length increases and converges around 1000 bits. Morgan and torsion fingerprints showcased the lowest averaged MAE across all folds. As a result, in addition to Morgan fingerprints, I also tested torsion fingerprints as input features for model training. Specifically, I generated both fingerprints with 1024 bits. After removing features represented by 0 along all fingerprint bits, the number of features remaining to describe each transporter are 352 and 357 for Morgan and torsion fingerprints respectively. To ensure the inclusion of chloride recognition sites, I applied one hot encoding for the presence (1) and absence (0) of urea, thiourea, tambjamine, squaramide, indole, carbazole, acridinone, perenosin, 1,3-bis(benzimidazol-2-yl)pyrimidine, and isophthalamides/dipicolineamides. Additionally, I included the key following experimental conditions as features: internal chloride concentration (in mM) and the time between transporter addition to initiate chloride transport and detergent addition to lyse the vesicles (in seconds).

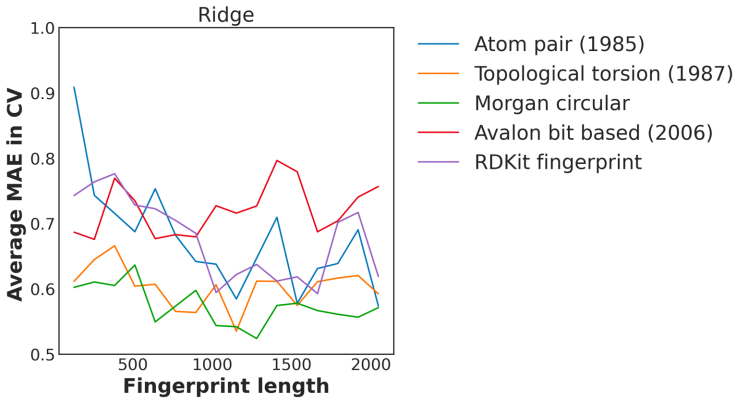


Figure 4.15: Comparison of fingerprints using 5-fold cross-validation with a ridge regression model.

In the following section, I first discuss the model performance on Morgan fingerprints, and then the model performance on torsion fingerprints. I will explain the choice of the final model and then explore the interpretability of the final model.

4.7.2.1 Model performance on Morgan fingerprints

On the training set, the DT, ET and XGBoost regressors achieved exceptionally high R^2 values of 0.96 (Table 4.5), indicating a strong fit to the training data. The RF regressor and LassoCV also exhibited competitive performance with R^2 values of 0.92 and 0.87, respectively. On the test set, the models largely retained their performance trends from the training set (Table 4.6, Figure B18). The DT, ET and XGBoost regressors maintained their high R^2 values of 0.88, demonstrating their robustness in generalization. The LassoCV and RF regressor also showcased consistent performance with R^2 values of 0.77 and 0.85, respectively. To evaluate the models' robustness and generalization under a null hypothesis, I also investigated the impact of shuffling target values in the test set on the predictive models' performance to evaluate the models' robustness and generalization under a null hypothesis (Appendix Section B.3.1). Shuffling the test set target values resulted in a significant performance drop (R^2 <0), suggesting that these models were indeed capturing meaningful feature-target relationships during training.

The convergence of performance between the simple DT and the more complex ET and XGBoost models implies that the inherent relationships within the dataset are well captured by the DT's structure. Occam's Razor principle suggests that when multiple models achieve similar performance, the simpler model is preferred.[216] DT's inherent interpretability offers a significant advantage, enabling transparent insights into the decision-making process. The practical advantages of employing the DT regressor extend beyond interpretability. With a small dataset, the risk of overfitting is a primary concern, and the DT's inherent simplicity makes it less prone to overfitting in comparison to more complex models. Additionally, the computational efficiency of the DT model aligns well with the constraints of the dataset size, eliminating the need for unnecessary computational resources.

In summary, these results suggest that the use of Morgan fingerprints as input features holds promise for accurate transport activity prediction. The convergence of performance metrics among the DT, ET, and XGBoost regressors underscores the effectiveness of the DT in encapsulating the underlying patterns of the dataset.

Table 4.5: Regression model performance on the training set with Morgan fingerprints as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	0.87	0.26	0.13	0.36	0.87
k-Nearest Neighbors Regressor	0.65	0.45	0.35	0.59	0.65
Bayesian Regressor	0.81	0.33	0.19	0.44	0.81
Decision Tree Regressor	0.96	0.07	0.04	0.20	0.96
Random Forest Regressor	0.92	0.21	0.08	0.28	0.92
Extra Trees Regressor	0.96	0.07	0.04	0.20	0.96
Ada Boost Regressor	0.64	0.52	0.36	0.60	0.64
XGBoost Regressor	0.96	0.09	0.04	0.20	0.96

Table 4.6: Regression model performance on the test set with Morgan fingerprints as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	0.77	0.35	0.20	0.45	0.77
k-Nearest Neighbors Regressor	0.75	0.38	0.22	0.47	0.74
Bayesian Regressor	0.76	0.40	0.21	0.46	0.76
Decision Tree Regressor	0.88	0.17	0.10	0.32	0.88
Random Forest Regressor	0.85	0.27	0.13	0.36	0.85
Extra Trees Regressor	0.88	0.17	0.10	0.32	0.88
Ada Boost Regressor	0.64	0.49	0.31	0.56	0.64
XGBoost Regressor	0.88	0.19	0.10	0.32	0.88

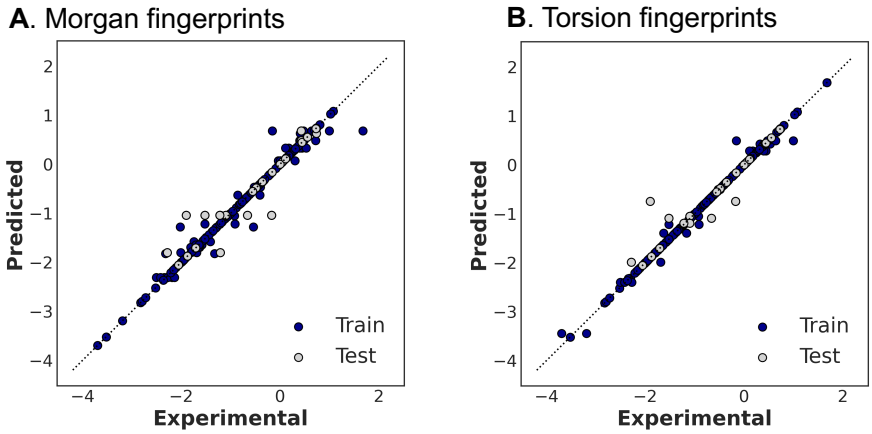


Figure 4.16: Regression model performance on (A) Morgan and (B) torsion fingerprints. Regression model performance on the training (dark blue) and test set (light grey). The predicted $log(EC_{50})$ values are on the y-axis and the experimental $log(EC_{50})$ values are on the x-axis.

4.7.2.2 Model training and evaluation on torsion fingerprints

In comparison to model performance on Morgan fingerprints, torsion fingerprints yield slightly higher predictive performances (Table 4.7 and Table 4.8). Despite this advantage, I opted to utilize Morgan fingerprints due to their better interpretability.

Morgan fingerprints provide a comprehensible representation of a molecule's local structural features. Each bit in a Morgan fingerprint corresponds to the presence or absence of a specific substructure within a defined radius around a central atom. This allows us to directly associate the set bits with particular molecular motifs or functional groups. Such clear identification enables chemists and researchers to infer the presence of significant chemical features, facilitating meaningful insights into a molecule's potential properties and behaviours. Additionally, the iterative nature of generating Morgan fingerprints allows researchers to customize the level of detail captured by varying the radius parameter. This adaptability is particularly beneficial when analyzing molecular similarities at different scales. In contrast, while torsion fingerprints may offer insights into molecular conformations and structural relationships, their complexity arises from the intricate interplay between torsion angles

and the entire molecular conformation. This complexity can hinder straightforward interpretation, making them less intuitive for research objectives.

Interpretable models are essential for guiding experimental chemists in the development of novel molecules, streamlining collaboration, and optimizing resource allocation. These models offer direct insights into the relationship between molecular features and properties, aiding compound design, identifying structure-activity relationships, and facilitating data-driven hypothesis generation. By selecting Morgan fingerprints, I prioritize a representation that aligns with the goals of intuitive interpretation, collaboration with experimental chemists, and computational efficiency, ultimately enhancing the ability to identify novel molecules with desired attributes.

Table 4.7: Regression model performance on the training set with torsion fingerprints as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	0.93	0.20	0.07	0.26	0.93
k-Nearest Neighbors Regressor	0.73	0.38	0.28	0.53	0.72
Bayesian Regressor	0.89	0.25	0.11	0.33	0.89
Decision Tree Regressor	0.98	0.04	0.02	0.14	0.98
Random Forest Regressor	0.94	0.18	0.06	0.24	0.94
Extra Trees Regressor	0.98	0.04	0.02	0.14	0.98
Ada Boost Regressor	0.71	0.48	0.29	0.54	0.71
XGBoost Regressor	0.98	0.05	0.02	0.14	0.98

Table 4.8: Regression model performance on the test set with torsion fingerprints as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	0.82	0.31	0.15	0.39	0.82
k-Nearest Neighbors Regressor	0.69	0.40	0.27	0.52	0.69
Bayesian Regressor	0.78	0.36	0.19	0.44	0.78
Decision Tree Regressor	0.90	0.12	0.09	0.30	0.90
Random Forest Regressor	0.87	0.23	0.11	0.33	0.87
Extra Trees Regressor	0.90	0.12	0.09	0.30	0.90
Ada Boost Regressor	0.68	0.46	0.28	0.53	0.68
XGBoost Regressor	0.90	0.13	0.09	0.30	0.90

4.7.2.3 Model interpretability

Considering interpretability, computational efficiency, and dataset size, the DT regression model is chosen as the final model (Figure B18). From the DT regression model, I computed the Gini importance for the features. Similar to the results from the classification models, the important fingerprint bits correspond to important structural features of the transporters, such as the binding motifs (feature indices 632, 641 and 926). The positive effect of fluorination (feature indices 8) has again been identified, consistent with previous experimental observations.[115]

As the DT regressor is a simple model, one of its notable advantages is the ability to trace the decision path, enabling a clear understanding of its prediction process. An illustrative example of this can be observed in Figure 4.18, which showcases the decision path based on binary-encoded Morgan fingerprint bits for test molecule 11. Following a sequence of conditional checks of the feature values again at a certain threshold, the model arrives at its final prediction. Similar analysis can be performed on all the training and test molecules. This transparency not only offers insights into the specific choices made by the model but also facilitates the understanding of the critical role individual features play in predictions.

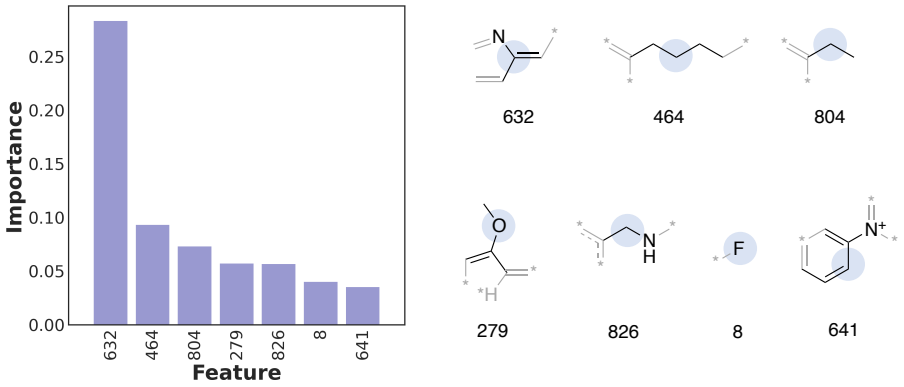


Figure 4.17: Decision tree model interpretation. Left: Gini importance extracted from the DT model. Right: The most important Morgan fingerprint bits.

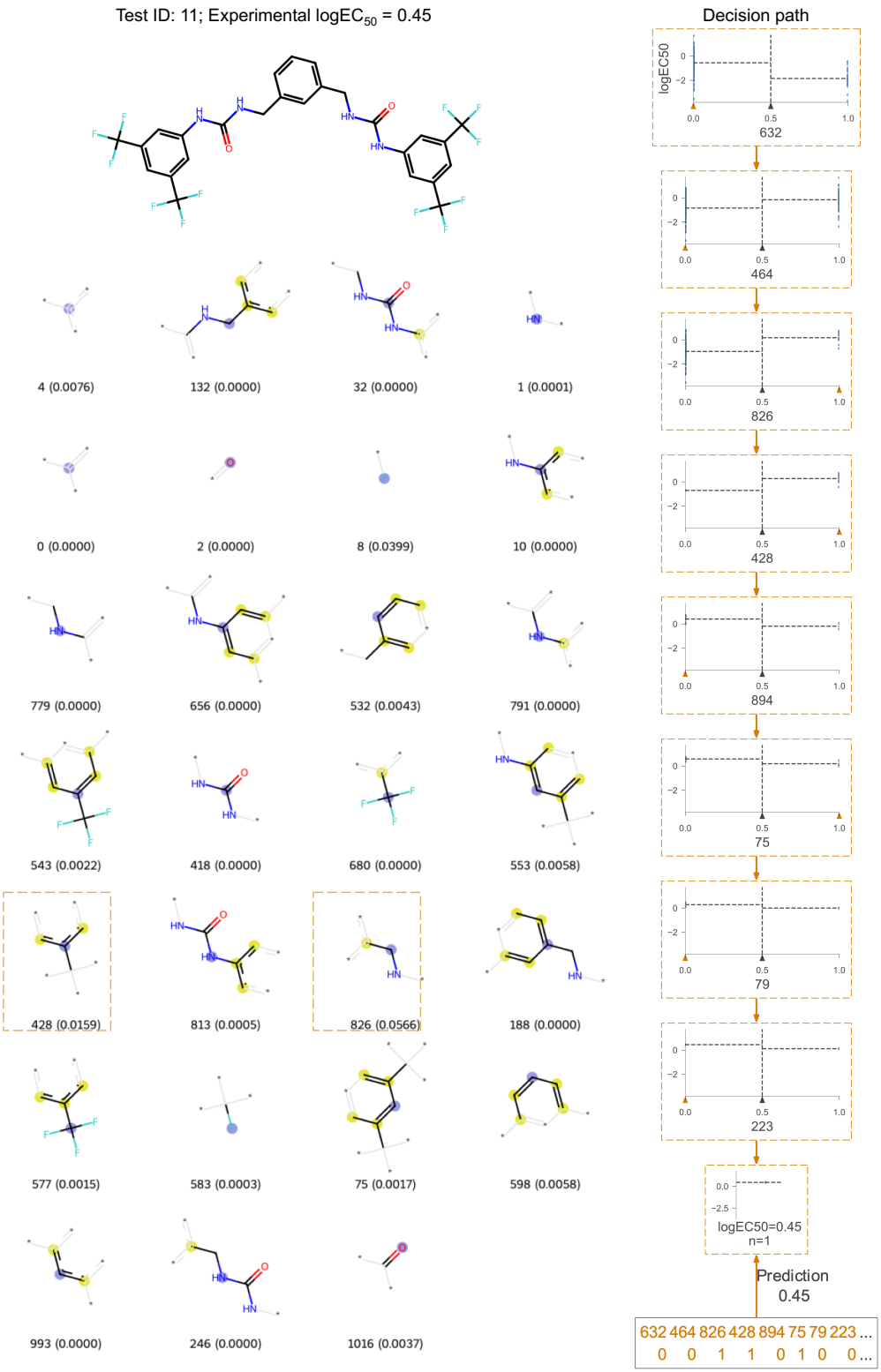


Figure 4.18: Decision path of test molecule 11.

4.8 Conclusion

In conclusion, I demonstrated the potential of ML as a tool in predicting the chloride transport ability of synthetic transporters. Both classification and regression models achieve high performance, indicating that Morgan fingerprints are a suitable molecular representation method for synthetic chloride transporters for accurate transport activity prediction. Interpretable models are essential for guiding experimental chemists in the development of novel molecules, streamlining collaboration, and optimizing resource allocation. These models offer direct insights into the relationship between molecular features and properties, aiding compound design, identifying structure-activity relationships, and facilitating data-driven hypothesis generation. I demonstrated model interpretability with chemically relevant insights for both classification and regression models. From the classification models, analysis with SHAP and feature importance extraction revealed that GP emphasizes experimental conditions, while ET prioritizes specific molecular features, including the binding motifs. From the regression models, the convergence of performance metrics among the DT, ET, and XGBoost regressors underscores the effectiveness of the DT in encapsulating the underlying patterns of the dataset. Through an example decision path analysis, I demonstrated the Decision Tree regressor's capacity to unravel complex relationships using simple yet informative input features. The decision path offers a clear and concise depiction of the model's decision logic, contributing to the development of interpretable ML methods in chemistry research. This work bridges data-driven approaches and molecular design, providing interpretable structure-activity relationships that will aid hypothesis generation and contribute to synthetic advances in this field.

Chapter 5

Conclusions

5.1 Thesis summary

In this Thesis, I used both experimental and computational methods to investigate two distinct membrane-related systems that serve as platforms for practical applications, including single-molecule chemistry and anion transport. Herein, I summarise the major findings for each of them.

Chapter 2 explores the use of engineered biological nanopores for sequencing and as nanoreactors for single-molecule chemistry. I established a nanopore-based approach to systematically profile the reactivity of individual cysteine footholds along a protein track at the single-molecule level. In this approach, a single-cysteine nanoreactor was inserted into a planar lipid bilayer. Small-molecule disulfide and thiol reagents were spatially separated by the lipid bilayer. Thiol-disulfide interchange reactions occurred at the cysteine site in a three-step reaction cycle. Changes in the ionic current passing through the αHL nanoreactor under an applied transmembrane electric potential were utilized to monitor the reaction cycle. I obtained the observed rate constants (k_{i-obs}) at different pH values. These k_{i-obs} values were then used to calculate the pK_a

values of the cysteine thiols and the pH-independent rate constants of the thiolates for the attack on the disulfide. The pK_a values of the six cysteines ranged from 9.17 to 9.85. The pH-independent rate constants of the cysteine thiolates for the reaction with DTNB ranged from 2.76 ± 0.21 × 10^3 M^{-1} s^{-1} to 5.89 ± 0.52 × 10^4 M^{-1} s^{-1}. This approach can be applied to profile the reactivity of reactive site(s) on protein tracks built in nanopores in general. Guided by this reactivity profile, site-specific mutagenesis was performed. Together with the optimization of experimental conditions, the overall stepping rate of a DNA cargo along a five-cysteine track was accelerated. This acceleration reduced the mean total transit time for DNA cargo translocation from 59 s to 14 s, with the rate-limiting step in the original track being accelerated 21-fold. This work advances the practical application of an enzymeless system for a generalized sequencing method for biopolymers beyond DNA.

In Chapter 3, I investigated the conformations and dynamic behaviours of *E*-/*Z*-switchable squaramide-based anion transporters using MD simulations in DMSO solution and within a POPC membrane. I attributed the enhanced anion binding of the *Z*-isomer in DMSO to better chloride encapsulation and increased HB interactions compared to the *E*-isomer. This difference in binding mode persisted within the membrane, where the *Z*-isomer adopted a closed conformation at the water/membrane interface, facilitating pre-organization for efficient chloride capture. In contrast, the *E*-isomer needed to reorient to achieve an optimal conformation, introducing an additional barrier to chloride translocation. Moreover, competitive dimerization of the *E*-isomer further diminished its transport activity. Results from free energy calculations indicated that translocation was the rate-limiting step of the transport process. This work provides a computational workflow to quantitatively characterise the transmembrane anion transport as well as to gain further insights for the future design of *E*-/*Z*-switchable anion transporters. To improve the ON and OFF control in these systems, I suggest incorporating a higher degree of pre-organization of the

Z-isomer, while maintaining a fine balance between the energy barrier of transport and the solubility of the transporter.

In Chapter 4, I demonstrated the potential of ML as a tool for predicting the chloride transport activity of synthetic transporters. The ML models, employing both classification and regression frameworks, exhibited remarkable performance across a diverse range of systems. I identified Morgan fingerprints as a suitable molecular representation for this task. Moreover, these models offered chemical insights, aiding experimental chemists to interpret the model and generate improved designs, streamlining collaboration and optimizing resource allocation. For example, the ML models identified key structural features (*e.g.*, the chloride binding motifs) and experimental conditions (*e.g.*, the time between transporter addition and vesicle lysis) that influence the observed transport activity. This work bridges data-driven approaches and molecular design, providing interpretable structure-activity relationships that will aid hypothesis generation and contribute to synthetic advances in this field.

5.2 Perspectives and future directions

In this section, I will discuss several points of interest that have not yet been explored and call for future study.

5.2.1 Towards an ideal track for biopolymer translocation

The system reported in Chapter 2 can be generalized for the translocation of biopolymers beyond DNA, such as polypeptides. While the rates in this system do not so far compare to those of the helicase motors used in nanopore DNA/RNA sequencing devices, which make hundreds of steps per second, equivalent enzymes do not exist for other biopolymers, making sequencing by chemical stepping an attractive option. I believe an ideal track for biopolymer translocation should have the following properties: (1) individual steps should be observable, and each foothold can be differentiated by the ionic current; (2) each track should be usable for repeated translocation; (3) the track should be easily tunable by site-directed mutagenesis or chemical modification; and (4) stepping rates should be similar throughout the track to avoid hold-ups during translocation.

5.2.1.1 Faster stepping rate on cysteine tracks

With further optimization, I believe chemical stepping along cysteine tracks could be improved to provide rates comparable to those seen for sequencing devices for nucleic acids. Currently, the results from site-directed mutagenesis showed that favourable changes to cysteine reactivity can be elicited by introducing the E111Q mutation, leading to a 3-fold increase in reactivity for selected steps in the cysteine track. All steps were sped up by increasing the pH from 8.5 to 9.5. Without further track engineering, faster translocation could be achieved at a pH above 9.5, where all cysteine footholds are fully deprotonated. Increasing temperature may also improve the

stepping kinetics. αHL mutants were shown to be stable in single-channel recordings when the temperature was brought to more than 90 °C.[55] Using activation energy of 50 kJ mol^{-1},[104] a 50 °C increase in temperature from 25 °C to 75 °C is predicted to produce an 18-fold rate acceleration for thiol-disulfide interchange.

5.2.1.2 Selenocysteine tracks

One can even move away from the cysteine footholds and replace them with selenocysteines. On a selenocysteine track, selenolate (*i.e.*, a deprotonated selenol) would act as the nucleophile and the leaving group during each hopping step. If selenium were incorporated into the biopolymer cargo, selenium could act as the electrophile for nucleophilic attack during each hopping step. Previous studies have suggested that selenols are ∼1 order of magnitude more reactive than thiols as nucleophiles and ∼4 orders of magnitude more reactive than thiols as electrophiles.[105] Incorporation of electrophilic selenium as the cargo handle could reduce the mean total transit time to <1 ms, equivalent to >1000 complete cycles across the track per second.

5.2.1.3 Longer tracks

For the enzymeless system to find practical application in a generalized sequencing method for biopolymers beyond DNA, in addition to fast stepping rates, it will be necessary to extend the length of the track as well. The β-barrel in the αHL nanopore has a length of 5 nm, which can accommodate up to six footholds along a single strand. Other β-barrel nanopores may be explored to construct longer protein tracks. One option is aerolysin.[262] Aerolysin has a β-barrel that is around 10 nm long, which is twice the length of αHL's (Figure 5.1). I hypothesize that up to twelve footholds may be built on a single strand in aerolysin. Furthermore, aerolysin has already been used for sensing and sequencing peptides and nucleic acids.[263–266] To construct a cysteine track within the aerolysin nanopore, I propose that a heteroheptamer generation

strategy similar to the one used here may be applied.

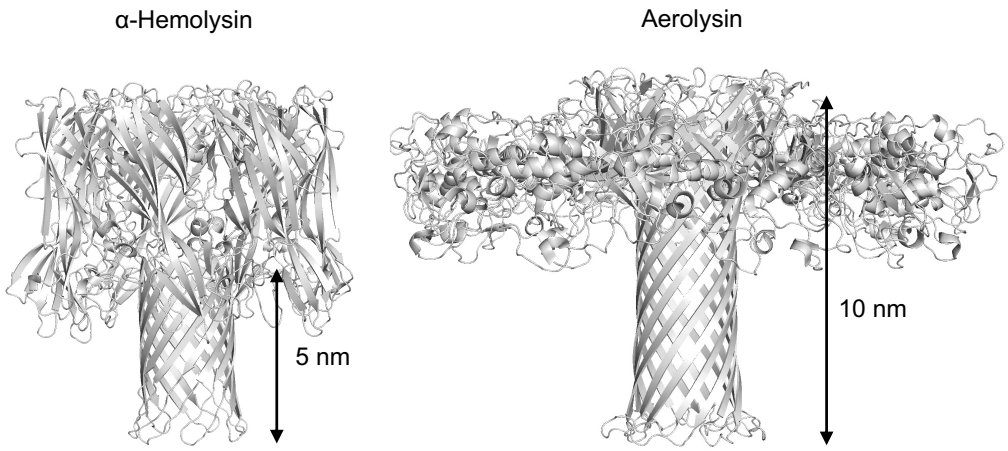


Figure 5.1: Comparison of the β-barrel length of αHL and aerolysin.

5.2.2 Computer-aided design of synthetic anion transporters

I envision the MD and ML workflows developed in this Thesis to be applied to the future design of synthetic ion transporters. In Chapter 3, I introduced a quantitative approach based on MD/US simulations to fully characterise transmembrane chloride transport. This methodology offers a holistic perspective of the entire transport process, identifying the rate-limiting step, which I plan to extend for the study of new small-molecule scaffolds as potential transmembrane ion transporters.

The ML models developed in Chapter 4 are expected to be readily deployed by experimental chemists simplifying the assessment of new designs of synthetic chloride transporters. These models employ highly automated and user-friendly workflows via Jupyter Notebooks, requiring only SMILES strings of the synthetic chloride transporters as input. The notebook then streamlines feature generation and selection as well as model training, which delivers either an active/inactive classification or a numerical $log(EC_{50})$ value from the regression model. I expect that such a user-friendly setup will facilitate the use of the ML models by the wider supramolecular community as a preliminary screening tool, significantly reducing the time and resources invested

in experimental design.

The ML models developed in Chapter 4 were trained on experimental transport activities collected from the chloride/nitrate ISE assay exclusively. To advance towards a more generalized ML model for anion transport activity prediction, I propose expanding the models to include chloride transport activities collected from other assays such as the HPTS assays.[253,254] This information could be one-hot encoded as an input feature. Furthermore, the models could be expanded beyond chloride to encompass various anion types for which selectivity remains an issue, such as chloride, bromide, iodide, and hydroxides. The anion types may be one-hot encoded as an input feature. The models will still be predicting transport activities. Alternatively, the anion types could be the target for a multi-label classification model (Figure 5.2), where a single molecule can belong to multiple classes at the same time.

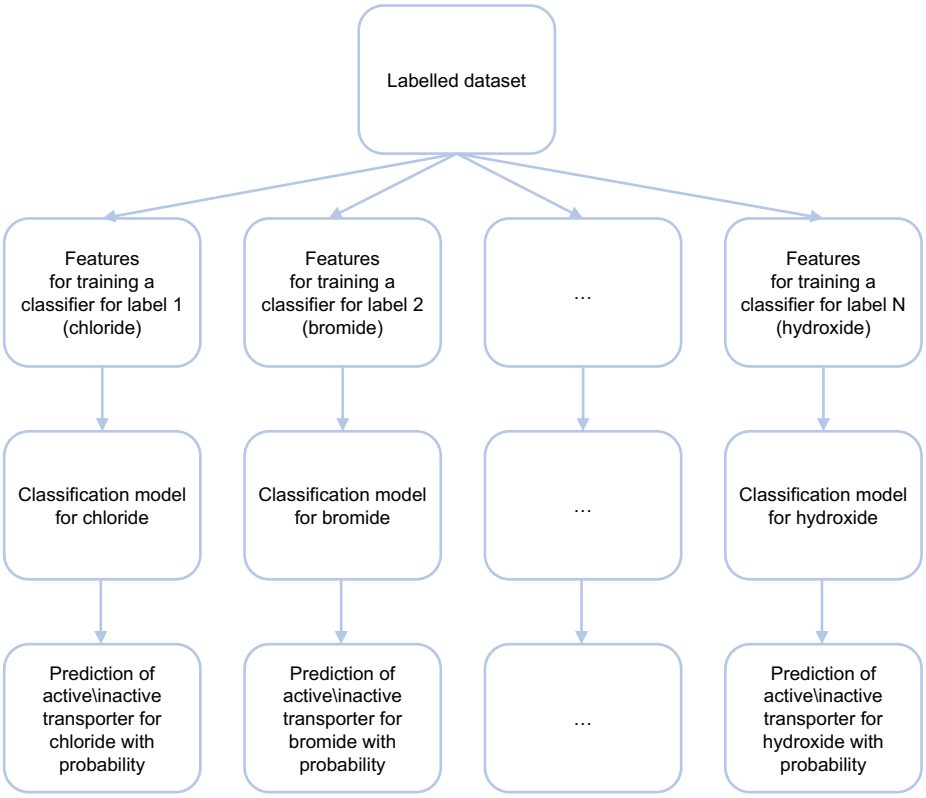


Figure 5.2: A multi-label classification model for predicting transport activity of multiple anions.

To train such a model, I propose the following steps as an example.

1. **Dataset preparation:** A labelled dataset will be curated to consist of SMILES representations for synthetic anion transporters and their corresponding binary-encoded target labels: 1 for active transporter and 0 for inactive transporter.

2. **Binary classifiers training:** Separate binary classifiers for each anion (Chloride, Bromide, Iodide, Hydroxide) will be trained. For each binary classifier, the SMILES representations of the molecules will be used as features and the corresponding binary-encoded label for the specific anion will be used as the target. appropriate ML algorithms (e.g., logistic regression, random forest, neural networks) will be selected for each binary classifier.

3. **Model training:** Each binary classifier will be trained on the prepared dataset by optimizing the chosen algorithm's parameters for best performance.

4. **Prediction:** Predictions on test molecules will be performed using the trained binary classifiers. For each binary classifier, a binary output (1 for active, 0 for inactive) indicating whether the test molecule is an active transporter for that specific anion will be calculated.

5. **Multi-label prediction:** The binary outputs from each classifier to generate a multi-label prediction for the test molecule, indicating which anions the test molecule is an active transporter for, will be interpreted.

6. **Thresholding (optional):** A thresholding mechanism will be introduced to determine when a molecule is considered an active transporter for a specific anion. This is done by setting a minimum probability score or confidence level for a classifier's output before labelling the molecule as an active transporter.

The benefits of including other assays and anions in the ML models for predicting transport activities can be significant. Different assays have variations in experimental

conditions, and including these conditions as input features can make the models more robust to assay-specific variations. It allows the models to adapt to the specific context of each assay. Including other anions provides a more comprehensive understanding of the transporter's behaviour. This can lead to more accurate predictions as the models learn to capture subtle differences in transport activity for different anions.

With a broader dataset that includes multiple anions and assays, one can perform more comprehensive analyses. For example, correlations between transport activities for different anions can be explored. In addition, how experimental conditions impact transport behaviour across various assays can be assessed. This could also lead to new scientific insights and uncover patterns or relationships that were not evident when focusing solely on a single anion or assay type. The models can be applied to a broader range of research scenarios. Researchers studying chloride transport might benefit from the same models when investigating bromide or iodide transport, reducing the need to develop new models for each anion.

Introducing multi-label classification for anion types allows the prediction of transport activities for multiple anions simultaneously. This is particularly valuable when a molecule can transport more than one type of anion. It provides a holistic view of a molecule's transport capabilities. I expect these models to be used by researchers in the wider supramolecular community as a preliminary screening tool to significantly reduce the time and resources invested in experimental design.

Bibliography

[1] Jeremy M Berg, John L Tymoczko, and Lubert Stryer. *Biochemistry*. Macmillan, 2007.

[2] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. Garland Science, 2002.

[3] Hagan Bayley and Paul S. Cremer. Stochastic sensors inspired by biology. *Nature*, 413(6852):226–230, 2001.

[4] Yi-Lun Ying, Chan Cao, and Yi-Tao Long. Single molecule analysis by biological nanopore sensors. *Analyst*, 139(16):3826–3835, 2014.

[5] Dominic F Cairns-Gibson and Scott L Cockroft. Functionalised nanopores: chemical and biological modifications. *Chem. Sci.*, 13(7):1869–1882, 2022.

[6] Hagan Bayley. Nanopore Sequencing: From Imagination to Reality. *Clinical Chemistry*, 61(1):25–31, 01 2015.

[7] David Deamer, Mark Akeson, and Daniel Branton. Three decades of nanopore sequencing. *Nature Biotechnology*, 34(5):518–524, may 2016.

[8] Tudor Luchian, Seong-Ho Shin, and Hagan Bayley. Single-molecule covalent chemistry with spatially separated reactants. *Angewandte Chemie International Edition*, 42(32):3766–3771, 2003.

[9] Mackay B. Steffensen, Dvir Rotem, and Hagan Bayley. Single-molecule analysis of chirality in a multicomponent reaction network. *Nature Chemistry*, 6(7):603–607, July 2014.

[10] Gökçe Su Pulcu, Ellina Mikhailova, Lai-Sheung Choi, and Hagan Bayley. Continuous observation of the stochastic motion of an individual small-molecule walker. *Nature nanotechnology*, 10(1):76—83, January 2015.

[11] Yujia Qing, Gökçe Su Pulcu, Nicholas A. W. Bell, and Hagan Bayley. Bioorthogonal cycloadditions with sub-millisecond intermediates. *Angewandte Chemie International Edition*, 57(5):1218–1221, 2018.

[12] William J. Ramsay and Hagan Bayley. Single-molecule determination of the isomers of d-glucose and d-fructose that bind to boronic acids. *Angewandte Chemie International Edition*, 57(11):2841–2845, 2018.

[13] Gökçe Su Pulcu, Nicole Stéphanie Galenkamp, Yujia Qing, Giulio Gasparini, Ellina Mikhailova, Stefan Matile, and Hagan Bayley. Single-molecule kinetics of growth and degradation of cell-penetrating poly(disulfide)s. *Journal of the American Chemical Society*, 141(32):12444–12447, 2019.

[14] Zonghua Bo, Zhong Hui Lim, Fernanda Duarte, Hagan Bayley, and Yujia Qing. Mobile molecules: Reactivity profiling guides faster movement on a cysteine track. *Angewandte Chemie International Edition*, 62(21):e202300890, 2023.

[15] Hagan Bayley. Designed membrane channels and pores. *Current Opinion in Biotechnology*, 10(1):94–103, 1999.

[16] Hagan Bayley and Lakmal Jayasinghe. Functional engineered channels and pores (Review). *Molecular Membrane Biology*, 21(4):209–220, 2004.

[17] Mariam Ayub and Hagan Bayley. Engineered transmembrane pores. *Current Opinion in Chemical Biology*, 34:117–126, oct 2016.

[18] I.D. Ryabova, G.A. Gorneva, and Yu A. Ovchinnikov. Effect of valinomycin on ion transport in bacterial cells and on bacterial growth. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 401(1):109–118, 1975.

[19] F.M. Harold and J.R. Baarda. Gramicidin, valinomycin, and cation permeability of streptococcus faecalis. *Journal of Bacteriology*, 94(1):53–60, 1967.

[20] Devaki A Kelkar and Amitabha Chattopadhyay. The gramicidin ion channel: A model membrane protein. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1768(9):2011–2025, 2007.

[21] Gastone G. Celesia. Disorders of membrane channels or channelopathies, jan 2001.

[22] Christine E. Bear, Canhui Li, Norbert Kartner, Robert J. Bridges, Tim J. Jensen, Mohabir Ramjeesingh, and John R. Riordan. Purification and functional reconstitution of the cystic fibrosis transmembrane conductance regulator (cftr). *Cell*, 68(4):809–818, 1992.

[23] Jessica LaRusch, Jinsei Jung, Ignacio J General, Michele D Lewis, Hyun Woo Park, Randall E Brand, Andres Gelrud, Michelle A Anderson, Peter A Banks, Darwin Conwell, et al. Mechanisms of CFTR Functional Variants That Impair Regulated Bicarbonate Permeation and Increase Risk for Pancreatitis but Not for Cystic Fibrosis. *PLoS Genetics*, 10(7), 2014.

[24] R.A. Sutton, V. Mavichak, A. Halabe, and G.E. Wilkins. Bartter's syndrome: evidence suggesting a distal tubular defect in a hypocalciuric variant of the syndrome. *Mineral and electrolyte metabolism*, 18(1):43–51, 1992.

[25] David B. Simon, Ranjit S. Bindra, Traci A. Mansfield, Carol Nelson-Williams, Erica Mendonca, Rosário Stone, Scott Schurman, Ahmet Nayir, Harika Alpay, Aysin Bakkaloglu, Juan Rodriguez-Soriano, Jose M. Morales, Sami A. Sanjad,

C. Mark Taylor, Daniela Pilz, Andrew Brem, Howard Trachtman, William Griswold, George A. Richard, Eunice John, and Richard P. Lifton. Mutations in the chloride channel gene, clcnkb, cause bartter's syndrome type iii. *Nature Genetics*, 17(2):171–178, 1997.

[26] Willy Günther, Anke Lüchow, Françoise Cluzeaud, Alain Vandewalle, and Thomas J. Jentsch. Clc-5, the chloride channel mutated in dent's disease, colocalizes with the proton pump in endocytotically active kidney cells. *Proceedings of the National Academy of Sciences*, 95(14):8075–8080, 1998.

[27] Philip A. Gale, Jeffery T. Davis, and Roberto Quesada. Anion transport and supramolecular medicinal chemistry. *Chem. Soc. Rev.*, 46:2497–2519, 2017.

[28] Jeffery T. Davis, Philip A. Gale, and Roberto Quesada. Advances in anion transport and supramolecular medicinal chemistry. *Chem. Soc. Rev.*, 49(16):6056–6086, 2020.

[29] Laura E. Bickerton, Toby G. Johnson, Aidan Kerckhoffs, and Matthew J. Langton. Supramolecular chemistry in lipid bilayer membranes. *Chem. Sci.*, 12:11252–11274, 2021.

[30] Shao-Ping Zheng, Li-Bo Huang, Zhanhu Sun, and Mihail Barboiu. Self-assembled artificial ion-channels toward natural selection of functions. *Angewandte Chemie International Edition*, 60(2):566–597, 2021.

[31] S Bhakdi and Jogdep Tranum-Jensen. Alpha-toxin of staphylococcus aureus. *Microbiological reviews*, 55(4):733–751, 1991.

[32] Langzhou Song, Michael R. Hobaugh, Christopher Shustak, Stephen Cheley, Hagan Bayley, and J. Eric Gouaux. Structure of staphylococcal α-hemolysin, a heptameric transmembrane pore. *Science*, 274(5294):1859–1865, 1996.

[33] Erwin Neher and Bert Sakmann. Single-channel currents recorded from membrane of denervated frog muscle fibres. *Nature*, 260(5554):799–802, 1976.

[34] Orit Braha, Barbara Walker, Stephen Cheley, John J. Kasianowicz, Langzhou Song, J.Eric Gouaux, and Hagan Bayley. Designed protein pores as components for biosensors. *Chemistry Biology*, 4(7):497–505, 1997.

[35] Orit Braha, Li-Qun Gu, Li Zhou, Xiaofeng Lu, Stephen Cheley, and Hagan Bayley. Simultaneous stochastic sensing of divalent metal ions. *Nature Biotechnology*, 18(9):1005–1007, 2000.

[36] Li-Qun Gu, Orit Braha, Sean Conlan, Stephen Cheley, and Hagan Bayley. Stochastic sensing of organic analytes by a pore-forming protein containing a molecular adapter. *Nature*, 398(6729):686–690, 1999.

[37] Li-Qun Gu, Stephen Cheley, and Hagan Bayley. Capture of a single molecule in a nanocavity. *Science*, 291(5504):636–640, 2023/09/01 2001.

[38] James Clarke, Hai-Chen Wu, Lakmal Jayasinghe, Alpesh Patel, Stuart Reid, and Hagan Bayley. Continuous base identification for single-molecule nanopore dna sequencing. *Nature nanotechnology*, 4(4):265–270, 2009.

[39] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of dna in a nanopore at 5-å precision. *Nature biotechnology*, 30(4):344–348, 2012.

[40] Elizabeth A Manrao, Ian M Derrington, Andrew H Laszlo, Kyle W Langford, Matthew K Hopper, Nathaniel Gillgren, Mikhail Pavlenok, Michael Niederweis, and Jens H Gundlach. Reading dna at single-nucleotide resolution with a mutant mspa nanopore and phi29 dna polymerase. *Nature biotechnology*, 30(4):349–353, 2012.

[41] Chan Cao, Yi-Lun Ying, Zheng-Li Hu, Dong-Fang Liao, He Tian, and Yi-Tao Long. Discrimination of oligonucleotides of different lengths with a wild-type aerolysin nanopore. *Nature nanotechnology*, 11(8):713–718, 2016.

[42] Todd C Sutherland, Yi-Tao Long, Radu-Ioan Stefureac, Irene Bediako-Amoa, Heinz-Bernhard Kraatz, and Jeremy S Lee. Structure of peptides investigated by nanopore analysis. *Nano letters*, 4(7):1273–1277, 2004.

[43] Liviu Movileanu, Jason P Schmittschmitt, J Martin Scholtz, and Hagan Bayley. Interactions of peptides with a protein pore. *Biophysical journal*, 89(2):1030–1045, 2005.

[44] Liviu Movileanu, Stefan Howorka, Orit Braha, and Hagan Bayley. Detecting protein analytes that modulate transmembrane movement of a polymer chain within a single protein pore. *Nature biotechnology*, 18(10):1091–1095, 2000.

[45] Iftach Nir, Diana Huttner, and Amit Meller. Direct sensing and discrimination among ubiquitin and ubiquitin chains using solid-state nanopores. *Biophysical journal*, 108(9):2340–2349, 2015.

[46] Sergey M. Bezrukov and John J. Kasianowicz. Current noise reveals protonation kinetics and number of ionizable sites in an open protein ion channel. *Phys. Rev. Lett.*, 70:2352–2355, Apr 1993.

[47] John J Kasianowicz, Eric Brandin, Daniel Branton, and David W Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[48] Nurit Ashkenasy, Jorge Sánchez-Quesada, Hagan Bayley, and M Reza Ghadiri. Recognizing a single base in an individual dna strand: a step toward dna sequencing in nanopores. *Angewandte Chemie International Edition*, 44(9):1401–1404, 2005.

[49] Kate R Lieberman, Gerald M Cherf, Michael J Doody, Felix Olasagasti, Yvette Kolodji, and Mark Akeson. Processive Replication of Single DNA Molecules in a Nanopore Catalyzed by phi29 DNA Polymerase. *Journal of the American Chemical Society*, 132(50):17961–17972, 2010.

[50] Yujia Qing and Hagan Bayley. Enzymeless dna base identification by chemical stepping in a nanopore. *Journal of the American Chemical Society*, 143(43):18181–18187, 2021.

[51] Gerald M Cherf, Kate R Lieberman, Hytham Rashid, Christopher E Lam, Kevin Karplus, and Mark Akeson. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nature Biotechnology*, 30(4):344–348, 2012.

[52] Giovanni Maglia, Andrew J. Heron, David Stoddart, Deanpen Japrung, and Hagan Bayley. Chapter 22 - analysis of single nucleic acid molecules with protein nanopores. In Nils G. Walter, editor, *Single Molecule Tools, Part B:Super-Resolution, Particle Tracking, Multiparameter, and Force Based Methods*, volume 475 of *Methods in Enzymology*, pages 591–623. Academic Press, 2010.

[53] Hagan Bayley, Tudor Luchian, Seong-Ho Shin, and Mackay B. Steffensen. *Single-Molecule Covalent Chemistry in a Protein Nanoreactor*, pages 251–277. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[54] Li-Qun Gu and Hagan Bayley. Interaction of the noncovalent molecular adapter, β-cyclodextrin, with the staphylococcal α-hemolysin pore. *Biophysical Journal*, 79(4):1967–1975, 2000.

[55] Xiao-feng Kang, Li-Qun Gu, Stephen Cheley, and Hagan Bayley. Single protein pores containing molecular adapters at high temperatures. *Angewandte Chemie International Edition*, 44(10):1495–1499, 2005.

[56] Seong-Ho Shin and Hagan Bayley. Stepwise growth of a single polymer chain. *Journal of the American Chemical Society*, 127(30):10462–10463, 08 2005.

[57] Mackay B Steffensen, Dvir Rotem, and Hagan Bayley. Single-molecule analysis of chirality in a multicomponent reaction network. *Nature chemistry*, 6(7):603–607, 2014.

[58] Leslie B Poole. The basics of thiols and cysteines in redox biology and chemistry. *Free Radical Biology and Medicine*, 80:148–157, 2015.

[59] Seong-Ho Shin, Tudor Luchian, Stephen Cheley, Orit Braha, and Hagan Bayley. Kinetics of a reversible covalent-bond-forming reaction observed at the single-molecule level. *Angewandte Chemie International Edition*, 41(19):3707–3709, 2002.

[60] Sandra Loudwig and Hagan Bayley. Photoisomerization of an individual azobenzene molecule in water: An onoff switch triggered by light at a fixed wavelength. *Journal of the American Chemical Society*, 128(38):12404–12405, 09 2006.

[61] Anne.F. Hammerstein, Seong-Ho Shin, and Hagan Bayley. Single-molecule kinetics of two-step divalent cation chelation. *Angewandte Chemie International Edition*, 49(30):5085–5090, 2010.

[62] Joongoo Lee and Hagan Bayley. Semisynthetic protein nanoreactor for single-molecule chemistry. *Proceedings of the National Academy of Sciences*, 112(45):13768–13773, 2015.

[63] Joongoo Lee, Arnold J. Boersma, Marc A. Boudreau, Stephen Cheley, Oliver Daltrop, Jianwei Li, Hiroko Tamagaki, and Hagan Bayley. Semisynthetic nanoreactor for reversible single-molecule covalent chemistry. *ACS Nano*, 10(9):8843–8850, 09 2016.

[64] Hongzhou Gu, Jie Chao, Shou-Jun Xiao, and Nadrian C Seeman. A proximity-based programmable dna nanoscale assembly line. *Nature*, 465(7295):202–205,

2010.

[65] Kyle Lund, Anthony J Manzo, Nadine Dabby, Nicole Michelotti, Alexander Johnson-Buck, Jeanette Nangreave, Steven Taylor, Renjun Pei, Milan N Stojanovic, Nils G Walter, et al. Molecular robots guided by prescriptive landscapes. *Nature*, 465(7295):206–210, 2010.

[66] S.J. Green, Jonathan Bath, and A.J. Turberfield. Coordinated chemomechanical cycles: a mechanism for autonomous molecular motion. *Physical review letters*, 101(23):238101, 2008.

[67] Max Von Delius, Edzard M Geertsema, and David A Leigh. A synthetic small molecule that can walk down a track. *Nature chemistry*, 2(2):96, 2010.

[68] Miriam R Wilson, Jordi Solà, Armando Carlone, Stephen M Goldup, Nathalie Lebrasseur, and David A Leigh. An autonomous chemically fuelled small-molecule motor. *Nature*, 534(7606):235–240, 2016.

[69] Wesley R Browne and Ben L Feringa. Making molecular machines work. *Nature nanotechnology*, 1(1):25–35, 2006.

[70] Yujia Qing, Sandra A. Ionescu, Gökçe Su Pulcu, and Hagan Bayley. Directional control of a processive molecular hopper. *Science*, 361(6405):908–912, 2018.

[71] W.D. Stein and Thomas Litman. Channels, carriers, and pumps: An introduction to membrane transport. *Channels, Carriers, and Pumps: An Introduction to Membrane Transport*, pages 1–406, 01 2014.

[72] Zhengnan Yuan, Yueqin Zheng, Bingchen Yu, Siming Wang, Xiaoxiao Yang, and Binghe Wang. Esterase-sensitive glutathione persulfide donor. *Organic Letters*, 20(20):6364–6367, 2018.

[73] Christopher Nicolai and Frederick Sachs. Solving ion channel kinetics with the qub software. *Biophysical Reviews and Letters*, 08(03n04):191–211, 2013.

[74] F. Qin, A. Auerbach, and F. Sachs. Estimating single-channel kinetic parameters from idealized patch-clamp data containing missed events. *Biophysical Journal*, 70(1):264–280, 1996.

[75] Peter J. Bond, Andrew T. Guy, Andrew J. Heron, Hagan Bayley, and Syma Khalid. Molecular dynamics simulations of dna within a nanopore: Arginine-phosphate tethering and a binding/sliding mechanism for translocation. *Biochemistry*, 50(18):3777–3783, 2011.

[76] Junmei Wang, Romain M. Wolf, James W. Caldwell, Peter A. Kollman, and David A. Case. Development and testing of a general amber force field. *Journal of Computational Chemistry*, 25(9):1157–1174, 2004.

[77] Junmei Wang, Wei Wang, Peter A. Kollman, and David A. Case. Automatic atom type and bond type perception in molecular mechanical calculations. *Journal of Molecular Graphics and Modelling*, 25(2):247–260, 2006.

[78] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene,

C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian 16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.

[79] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 10 1993.

[80] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[81] Kresten Lindorff-Larsen, Stefano Piana, Kim Palmo, Paul Maragakis, John L. Klepeis, Ron O. Dror, and David E. Shaw. Improved side-chain torsion potentials for the amber ff99sb protein force field. *Proteins: Structure, Function, and Bioinformatics*, 78(8):1950–1958, 2010.

[82] Daniel J. Price and III Brooks, Charles L. A modified TIP3P water potential for simulation with Ewald summation. *The Journal of Chemical Physics*, 121(20):10096–10103, 11 2004.

[83] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1), 01 2007.

[84] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 12 1981.

[85] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N log(N) method for Ewald sums in large systems. *The Journal of Chemical Physics*,

98(12):10089–10092, 06 1993.

[86] Ulrich Essmann, Lalith Perera, Max L. Berkowitz, Tom Darden, Hsing Lee, and Lee G. Pedersen. A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, 103(19):8577–8593, 11 1995.

[87] Berk Hess, Henk Bekker, Herman J. C. Berendsen, and Johannes G. E. M. Fraaije. Lincs: A linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, 18(12):1463–1472, 1997.

[88] O.S. Smart, J.M. Goodfellow, and B.A. Wallace. The pore dimensions of gramicidin A. *Biophysical Journal*, 65(6):2455–2460, 1993.

[89] Oliver S Smart, Joseph G Neduvelil, Xiaonan Wang, B A Wallace, and Mark S P Sansom. HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *Journal of Molecular Graphics*, 14(6):354–360, 1996.

[90] Lukas S Stelzl, Philip W Fowler, Mark S P Sansom, and Oliver Beckstein. Flexible Gates Generate Occluded Intermediates in the Transport Cycle of LacY. *Journal of Molecular Biology*, 426(3):735–751, 2014.

[91] S.W. Englander, N.W. Downer, and H. Teitelbaum. Hydrogen exchange. *Annual Review of Biochemistry*, 41(1):903–924, 1972.

[92] Edvards Liepinsh and Gottfried Otting. Proton exchange rates from amino acid side chains— implications for image contrast. *Magnetic Resonance in Medicine*, 35(1):30–42, 1996.

[93] Guanghua Gao, Eugene F. DeRose, Thomas W. Kirby, and Robert E. London. Nmr determination of lysine pka values in the pol λ lyase domain: mechanistic implications. *Biochemistry*, 45(6):1785–1794, 2006.

[94] Jennifer H. Tomlinson, Saif Ullah, Poul Erik Hansen, and Mike P. Williamson. Characterization of salt bridges to lysines in the protein g b1 domain. *Journal of the American Chemical Society*, 131(13):4674–4684, 2009.

[95] R. Singh and G. M. Whitesides. *Thiol-Disulfide Interchange*, volume Supplement S, pages 633–658. J. Wiley and Sons, Ltd., London, 1993. 369.

[96] Kristine Steen Jensen, Jeppe T Pedersen, Jakob R Winther, and Kaare Teilum. The pka value and accessibility of cysteine residues are key determinants for protein substrate discrimination by glutaredoxin. *Biochemistry*, 53(15):2533–2540, 2014.

[97] Richard M LoPachin and Terrence Gavin. Reactions of electrophiles with nucleophilic thiolate sites: relevance to pathophysiological mechanisms and remediation. *Free radical research*, 50(2):195–205, 2016.

[98] Eric V Anslyn and Dennis A Dougherty. *Modern physical organic chemistry*. University science books, 2006.

[99] Grzegorz Bulaj, Tanja Kortemme, and David P. Goldenberg. Ionization-reactivity relationships for cysteine thiols in polypeptides. *Biochemistry*, 37(25):8965–8972, 1998.

[100] Pedro Alexandrino Fernandes and Maria João Ramos. Theoretical insights into the mechanism for thiol/disulfide exchange. *Chemistry – A European Journal*, 10(1):257–266, 2004.

[101] Pan-Fen Wang, Michael J. McLeish, Malea M. Kneen, Gavin Lee, and George L. Kenyon. An unusually low pka for cys282 in the active site of human muscle creatine kinase. *Biochemistry*, 40(39):11698–11705, 2001.

[102] Rosa E. Hansen, Henrik Østergaard, and Jakob R. Winther. Increasing the reactivity of an artificial dithiol-disulfide pair through modification of the electrostatic milieu. *Biochemistry*, 44(15):5899–5906, 2005.

[103] Rui P. P. Neves, Pedro Alexandrino Fernandes, and Maria João Ramos. Mechanistic insights on the reduction of glutathione disulfide by protein disulfide isomerase. *Proceedings of the National Academy of Sciences*, 114(24):E4724–E4733, 2017.

[104] Saradha Chandrasekhar, Balakrishnan S. Moorthy, Ruichao Xie, and Elizabeth M. Topp. Thiol-disulfide exchange in human growth hormone. *Pharmaceutical Research*, 33(6):1370–1382, February 2016.

[105] Daniel Steinmann, Thomas Nauser, and Willem H. Koppenol. Selenium and sulfur in exchange reactions: A comparative study. *The Journal of Organic Chemistry*, 75(19):6696–6699, 2010.

[106] Aidan Kerckhoffs and Matthew J. Langton. Reversible photo-control over transmembrane anion transport using visible-light responsive supramolecular carriers. *Chem. Sci.*, 11:6325–6331, 2020.

[107] Aidan Kerckhoffs, Zonghua Bo, Samuel E. Penty, Fernanda Duarte, and Matthew J. Langton. Red-shifted tetra-ortho-halo-azobenzenes for photo-regulated transmembrane anion transport. *Org. Biomol. Chem.*, 19:9058–9067, 2021.

[108] Nathalie Busschaert, Samuel J. Bradberry, Marco Wenzel, Cally J. E. Haynes, Jennifer R. Hiscock, Isabelle L. Kirby, Louise E. Karagiannidis, Stephen J. Moore, Neil J. Wells, Julie Herniman, G. John Langley, Peter N. Horton, Mark E. Light, Igor Marques, Paulo J. Costa, Vítor Félix, Jeremy G. Frey, and Philip A. Gale. Towards predictable transmembrane transport: Qsar analysis of anion binding and transport. *Chem. Sci.*, 4:3036–3045, 2013.

[109] Hennie Valkenier, Cally J. E. Haynes, Julie Herniman, Philip A. Gale, and Anthony P. Davis. Lipophilic balance – a new design principle for transmembrane anion carriers. *Chem. Sci.*, 5:1128–1134, 2014.

[110] Cristina Moiteiro, Igor Marques, William G. Ryder, Vasco Cachatra, Sílvia

Carvalho, Li-Jun Chen, Brian J. Goodfellow, Philip A. Gale, and Vítor Félix. Binding and transport properties of a benzo[b]thiophene-based mono-(thio)urea library. *European Journal of Organic Chemistry*, 2022(3):e202101484, 2022.

[111] Stuart N. Berry, Vanessa Soto-Cerrato, Ethan N. W. Howe, Harriet J. Clarke, Ishna Mistry, Ali Tavassoli, Young-Tae Chang, Ricardo Pérez-Tomás, and Philip A. Gale. Fluorescent transmembrane anion transporters: shedding light on anionophoric activity in cells. *Chem. Sci.*, 7:5069–5077, 2016.

[112] Martina Olivari, Riccardo Montis, Louise E. Karagiannidis, Peter N. Horton, Lucy K. Mapp, Simon J. Coles, Mark E. Light, Philip A. Gale, and Claudia Caltagirone. Anion complexation, transport and structural studies of a series of bis-methylurea compounds. *Dalton Trans.*, 44:2138–2149, 2015.

[113] Cally J. E. Haynes, Nathalie Busschaert, Isabelle L. Kirby, Julie Herniman, Mark E. Light, Neil J. Wells, Igor Marques, Vítor Félix, and Philip A. Gale. Acylthioureas as anion transporters: the effect of intramolecular hydrogen bonding. *Org. Biomol. Chem.*, 12:62–72, 2014.

[114] Natalie J. Andrews, Cally J. E. Haynes, Mark E. Light, Stephen J. Moore, Christine C. Tong, Jeffery T. Davis, William A. Harrell Jr., and Philip A. Gale. Structurally simple lipid bilayer transport agents for chloride and bicarbonate. *Chem. Sci.*, 2:256–260, 2011.

[115] Nathalie Busschaert, Marco Wenzel, Mark E. Light, Paulina Iglesias-Hernández, Ricardo Pérez-Tomás, and Philip A. Gale. Structure–activity relationships in tripodal transmembrane anion transporters: The effect of fluorination. *Journal of the American Chemical Society*, 133(35):14136–14148, 2011. PMID: 21846096.

[116] Ye Rin Choi, Gyu Chan Kim, Hae-Geun Jeon, Jinhong Park, Wan Namkung, and Kyu-Sung Jeong. Azobenzene-based chloride transporters with light controllable

activities. *Chem. Commun.*, 50:15305–15308, 2014.

[117] Nathalie Busschaert, Isabelle L. Kirby, Sarah Young, Simon J. Coles, Peter N. Horton, Mark E. Light, and Philip A. Gale. Squaramides as potent transmembrane anion transporters. *Angewandte Chemie International Edition*, 51(18):4426–4430, 2012.

[118] Robert B. P. Elmes, Nathalie Busschaert, Dawid D. Czech, Philip A. Gale, and Katrina A. Jolliffe. ph switchable anion transport by an oxothiosquaramide. *Chem. Commun.*, 51:10107–10110, 2015.

[119] Nathalie Busschaert, Seong-Hyun Park, Kyung-Hwa Baek, Yoon Pyo Choi, Jinhong Park, Ethan N. W. Howe, Jennifer R. Hiscock, Louise E. Karagiannidis, Igor Marques, Vítor Félix, Wan Namkung, Jonathan L. Sessler, Philip A. Gale, and Injae Shin. A synthetic ion transporter that disrupts autophagy and induces apoptosis by perturbing cellular chloride concentrations. *Nature Chemistry*, 9(7):667–675, 2017.

[120] Xiaoping Bao, Xin Wu, Stuart N. Berry, Ethan N. W. Howe, Young-Tae Chang, and Philip A. Gale. Fluorescent squaramides as anion receptors and transmembrane anion transporters. *Chem. Commun.*, 54:1363–1366, 2018.

[121] Nathalie Busschaert, Robert B. P. Elmes, Dawid D. Czech, Xin Wu, Isabelle L. Kirby, Evan M. Peck, Kevin D. Hendzel, Scott K. Shaw, Bun Chan, Bradley D. Smith, Katrina A. Jolliffe, and Philip A. Gale. Thiosquaramides: ph switchable anion transporters. *Chem. Sci.*, 5:3617–3626, 2014.

[122] Stephen J. Moore, Marco Wenzel, Mark E. Light, Rebeka Morley, Samuel J. Bradberry, Patricia Gómez-Iglesias, Vanessa Soto-Cerrato, Ricardo Pérez-Tomás, and Philip A. Gale. Towards "drug-like" indole-based transmembrane anion transporters. *Chem. Sci.*, 3:2501–2509, 2012.

[123] Cally J. E. Haynes, Stephen J. Moore, Jennifer R. Hiscock, Igor Marques, Paulo J. Costa, Vítor Félix, and Philip A. Gale. Tunable transmembrane chloride transport by bis-indolylureas. *Chem. Sci.*, 3:1436–1444, 2012.

[124] Laura A. Jowett, Ethan N. W. Howe, Vanessa Soto-Cerrato, Wim Van Rossom, Ricardo Pérez-Tomás, and Philip A. Gale. Indole-based perenosins as highly potent hcl transporters and potential anti-cancer agents. *Scientific Reports*, 7(1):9397, 2017.

[125] Wim Van Rossom, Daniel J. Asby, Ali Tavassoli, and Philip A. Gale. Perenosins: a new class of anion transporter with anti-cancer activity. *Org. Biomol. Chem.*, 14:2645–2650, 2016.

[126] Nicola J. Knight, Elsa Hernando, Cally J. E. Haynes, Nathalie Busschaert, Harriet J. Clarke, Koji Takimoto, María García-Valverde, Jeremy G. Frey, Roberto Quesada, and Philip A. Gale. Qsar analysis of substituent effects on tambjamine anion transporters. *Chem. Sci.*, 7:1600–1608, 2016.

[127] Patrick Wang, Xin Wu, and Philip A. Gale. Carbazole-based bis-ureas and thioureas as electroneutral anion transporters. *Supramolecular Chemistry*, 33(4):143–149, 2021.

[128] Krzysztof M. Bąk, Krzysztof Chabuda, Helena Montes, Roberto Quesada, and Michał J. Chmielewski. 1,8-diamidocarbazoles: an easily tuneable family of fluorescent anion sensors and transporters. *Org. Biomol. Chem.*, 16:5188–5196, 2018.

[129] Daniel A. McNaughton, Lauren K. Macreadie, and Philip A. Gale. Acridinone-based anion transporters. *Org. Biomol. Chem.*, 19:9659–9674, 2021.

[130] Giacomo Picci, Israel Carreira-Barral, Daniel Alonso-Carrillo, David Sanz-González, Pablo Fernández-López, María García-Valverde, Claudia Caltagirone,

and Roberto Quesada. Simple isophthalamides/dipicolineamides as active transmembrane anion transporters. *Supramolecular Chemistry*, 32(2):112–118, 2020.

[131] Mohamed Fares, Xin Wu, Deepthi Ramesh, William Lewis, Paul A. Keller, Ethan N. W. Howe, Ricardo Pérez-Tomás, and Philip A. Gale. Stimuli-responsive cycloaurated "off-on" switchable anion transporters. *Angewandte Chemie International Edition*, 59(40):17614–17621, 2020.

[132] Micke Lisbjerg, Hennie Valkenier, Bo M. Jessen, Hana Al-Kerdi, Anthony P. Davis, and Michael Pittelkow. Biotin[6]uril esters: Chloride-selective transmembrane anion carriers employing c—h···anion interactions. *Journal of the American Chemical Society*, 137(15):4948–4951, 04 2015.

[133] Andreas Vargas Jentzsch, Daniel Emery, Jiri Mareda, Pierangelo Metrangolo, Giuseppe Resnati, and Stefan Matile. Ditopic ion transport systems: Anion–interactions and halogen bonds at work. *Angewandte Chemie International Edition*, 50(49):11675–11678, 2011.

[134] Andreas Vargas Jentzsch, Andreas Hennig, Jiri Mareda, and Stefan Matile. Synthetic ion transporters that work with anionπ interactions, halogen bonds, and anion–macrodipole interactions. *Accounts of Chemical Research*, 46(12):2791–2800, 12 2013.

[135] Konstantinos Kavallieratos, Christina M. Bertao, and Robert H. Crabtree. Hydrogen bonding in anion recognition: A family of versatile, nonpreorganized neutral and acyclic receptors. *The Journal of Organic Chemistry*, 64(5):1675–1683, 03 1999.

[136] D. Purves. *Neuroscience*. National Center for Biotechnology Information., 2000.

[137] Alan S. Verkman and Luis J. V. Galietta. Chloride channels as drug targets. *Nature Reviews Drug Discovery*, 8(2):153–171, 2009.

[138] Manuel de Lera Ruiz and Richard L. Kraus. Voltage-gated sodium channels: Structure, function, pharmacology, and clinical indications. *Journal of Medicinal Chemistry*, 58(18):7093–7118, 09 2015.

[139] Christoph Fahlke, Henry T. Yu, Carol L. Beck, Thomas H. Rhodes, and Alfred L. George. Pore-forming segments in voltage-gated chloride channels. *Nature*, 390(6659):529–532, 1997.

[140] Mary Chebib and Graham A. R. Johnston. Gaba-activated ligand gated ion channels: Medicinal chemistry and molecular biology. *Journal of Medicinal Chemistry*, 43(8):1427–1447, 04 2000.

[141] Niels Ringstad, Namiko Abe, and H. Robert Horvitz. Ligand-gated chloride channels are receptors for biogenic amines in c. elegans. *Science*, 325(5936):96–100, 2009.

[142] Kristopher T. Kahle, Tarek Z. Deeb, Martin Puskarjov, Liliya Silayeva, Bo Liang, Kai Kaila, and Stephen J. Moss. Modulation of neuronal activity by phosphorylation of the k–cl cotransporter kcc2. *Trends in Neurosciences*, 36(12):726–737, 2013.

[143] Motoyuki Hattori and Eric Gouaux. Molecular mechanism of atp binding and ion channel activation in p2x receptors. *Nature*, 485(7397):207–212, 2012.

[144] Matthew J. Langton. Engineering of stimuli-responsive lipid-bilayer membranes using supramolecular systems. *Nature Reviews Chemistry*, 5(1):46–61, 2021.

[145] Xin Wu, Jennifer R. Small, Alessio Cataldo, Anne M. Withecombe, Peter Turner, and Philip A. Gale. Voltage-switchable hcl transport enabled by lipid headgroup–transporter interactions. *Angewandte Chemie International Edition*, 58(42):15142–15147, 2019.

[146] Ye Rin Choi, Bom Lee, Jinhong Park, Wan Namkung, and Kyu-Sung Jeong. Enzyme-responsive procarriers capable of transporting chloride ions across lipid and cellular membranes. *Journal of the American Chemical Society*, 138(47):15319–15322, 11 2016.

[147] Eun Bit Park and Kyu-Sung Jeong. Chloride transport activities of trans- and cis-amide-linked bisureas. *Chem. Commun.*, 51:9197–9200, 2015.

[148] Andrew Docker, Toby G. Johnson, Heike Kuhn, Zongyao Zhang, and Matthew J. Langton. Multistate redox-switchable ion transport using chalcogen-bonding anionophores. *Journal of the American Chemical Society*, 145(4):2661–2668, 02 2023.

[149] Mohamed Fares, Xin Wu, Deepthi Ramesh, William Lewis, Paul A. Keller, Ethan N. W. Howe, Ricardo Pérez-Tomás, and Philip A. Gale. Stimuli-responsive cycloaurated "off-on" switchable anion transporters. *Angewandte Chemie International Edition*, 59(40):17614–17621, 2020.

[150] Ye Rin Choi, Gyu Chan Kim, Hae-Geun Jeon, Jinhong Park, Wan Namkung, and Kyu-Sung Jeong. Azobenzene-based chloride transporters with light-controllable activities. *Chem. Commun.*, 50:15305–15308, 2014.

[151] Kajetan Dąbrowa and Janusz Jurczak. Tetra-(meta-butylcarbamoyl)azobenzene: A rationally designed photoswitch with binding affinity for oxoanions in a long-lived z-state. *Organic Letters*, 19(6):1378–1381, 03 2017.

[152] David Villarón, Maxime A. Siegler, and Sander J. Wezenberg. A photoswitchable strapped calix[4]pyrrole receptor: highly effective chloride binding and release. *Chemical Science*, 12(9):3188–3193, mar 2021.

[153] Sander J. Wezenberg, Li-Jun Chen, Jasper E. Bos, Ben L. Feringa, Ethan N. W. Howe, Xin Wu, Maxime A. Siegler, and Philip A. Gale. Photomodulation

of transmembrane transport and potential by stiff-stilbene based bis(thio)ureas. *Journal of the American Chemical Society*, 144(1):331–338, 01 2022.

[154] Manzoor Ahmad, Surajit Metya, Aloke Das, and Pinaki Talukdar. A sandwich azobenzene–diamide dimer for photoregulated chloride transport. *Chemistry – A European Journal*, 26(40):8703–8708, 2020.

[155] Ira M. Bennett, Hebe M. Vanegas Farfano, Federica Bogani, Alex Primak, Paul A. Liddell, Luis Otero, Leonides Sereno, Juana J. Silber, Ana L. Moore, Thomas A. Moore, and Devens Gust. Active transport of ca2+ by an artificial photosynthetic membrane. *Nature*, 420(6914):398–401, 2002.

[156] Swati Bansi Salunke, Javid Ahmad Malla, and Pinaki Talukdar. Phototriggered release of a transmembrane chloride carrier from an o-nitrobenzyl-linked procarrier. *Angewandte Chemie International Edition*, 58(16):5354–5358, 2019.

[157] Laura E. Bickerton and Matthew J. Langton. Controlling transmembrane ion transport via photo-regulated carrier mobility. *Chem. Sci.*, 13:9531–9536, 2022.

[158] Sander J. Wezenberg and Ben L. Feringa. Photocontrol of anion binding affinity to a bis-urea receptor derived from stiff-stilbene. *Organic Letters*, 19(2):324–327, 01 2017.

[159] Peter Läuger. Mechanisms of biological ion transport — carriers, channels, and pumps in artificial lipid membranes. *Angewandte Chemie International Edition in English*, 24(11):905–923, 1985.

[160] Laura E. Bickerton, Alistair J. Sterling, Paul D. Beer, Fernanda Duarte, and Matthew J. Langton. Transmembrane anion transport mediated by halogen bonding and hydrogen bonding triazole anionophores. *Chemical Science*, 11(18):4722–4729, may 2020.

[161] Laura E. Bickerton, Andrew Docker, Alistair J. Sterling, Heike Kuhn, Fernanda Duarte, Paul D. Beer, and Matthew J. Langton. Highly active halogen bonding and chalcogen bonding chloride transporters with non-protonophoric activity. *Chemistry – A European Journal*, 27(45):11738–11745, aug 2021.

[162] Sophie J. Edwards, Igor Marques, Christopher M. Dias, Robert A. Tromans, Nicholas R. Lees, Vítor Félix, Hennie Valkenier, and Anthony P. Davis. Tilting and tumbling in transmembrane anion carriers: activity tuning through n-alkyl substitution. *Chemistry - A European Journal*, 22(6):2004–2011, feb 2016.

[163] Igor Marques, Pedro M.R. R. Costa, Margarida Q. Miranda, Nathalie Busschaert, Ethan N.W. W. Howe, Harriet J. Clarke, Cally J.E. E. Haynes, Isabelle L. Kirby, Ananda M. Rodilla, Ricardo Pérez-Tomás, Philip A. Gale, Vítor Félix, Margarida Q. Miranda, Nathalie Busschaert, Ethan N.W. W. Howe, Harriet J. Clarke, Cally J.E. E. Haynes, Isabelle L. Kirby, Ananda M. Rodilla, Ricardo Pérez-Tomás, Philip A. Gale, Vítor Félix, Margarida Q. Miranda, Nathalie Busschaert, Ethan N.W. W. Howe, Harriet J. Clarke, Cally J.E. E. Haynes, Isabelle L. Kirby, Ananda M. Rodilla, Ricardo Pérez-Tomás, Philip A. Gale, and Vítor Félix. Full elucidation of the transmembrane anion transport mechanism of squaramides using in silico investigations. *Physical Chemistry Chemical Physics*, 20(32):20796–20811, aug 2018.

[164] Michael J. Spooner, Hongyu Li, Igor Marques, Pedro M.R. Costa, Xin Wu, Ethan N.W. Howe, Nathalie Busschaert, Stephen J. Moore, Mark E. Light, David N. Sheppard, Vítor Félix, and Philip A. Gale. Fluorinated synthetic anion carriers: experimental and computational insights into transmembrane chloride transport. *Chemical Science*, 10(7):1976–1985, feb 2019.

[165] Luis Martínez-Crespo, Lau Halgreen, Márcio Soares, Igor Marques, Vítor Félix, and Hennie Valkenier. Hydrazones in anion transporters: the detrimental effect of

a second binding site. *Organic Biomolecular Chemistry*, 19(38):8324–8337, oct 2021.

[166] Frank Jensen. *Introduction to Computational Chemistry*. Wiley, 2017.

[167] Wendy D. Cornell, Piotr Cieplak, Christopher I. Bayly, Ian R. Gould, Kenneth M. Merz, David M. Ferguson, David C. Spellmeyer, Thomas Fox, James W. Caldwell, and Peter A. Kollman. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *Journal of the American Chemical Society*, 117(19):5179–5197, 1995.

[168] Peter A. Kollman. Advances and continuing challenges in achieving realistic and predictive simulations of the properties of organic and biological molecules. *Accounts of Chemical Research*, 29(10):461–469, 1996.

[169] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. How well does a restrained electrostatic potential (resp) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry*, 21(12):1049–1074, 2000.

[170] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.

[171] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, 2003.

[172] Angel E. García and Kevin Y. Sanbonmatsu. α-helical stabilization by side

chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences*, 99(5):2782–2787, 2002.

[173] Jay W. Ponder and David A. Case. Force fields for protein simulations. In *Protein Simulations*, volume 66 of *Advances in Protein Chemistry*, pages 27–85. Academic Press, 2003.

[174] Joakim P. M. Jämbeck and Alexander P. Lyubartsev. An extension and further validation of an all-atomistic force field for biological membranes. *Journal of Chemical Theory and Computation*, 8(8):2938–2948, 08 2012.

[175] Liem X. Dang. Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-crown-6 ether: A molecular dynamics study. *Journal of the American Chemical Society*, 117(26):6954–6960, 1995.

[176] Pengfei Li, Lin Frank Song, and Kenneth M. Jr. Merz. Systematic parameterization of monovalent ions employing the nonbonded model. *Journal of Chemical Theory and Computation*, 11(4):1645–1657, 2015. PMID: 26574374.

[177] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics*, 72(4):2384–2393, 1980.

[178] Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[179] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, 52(2):255–268, 1984.

[180] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[181] Glenn M. Torrie and John P. Valleau. Monte carlo free energy estimates using non-boltzmann sampling: Application to the sub-critical lennard-jones fluid. *Chemical Physics Letters*, 28(4):578–581, 1974.

[182] G.M. Torrie and J.P. Valleau. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23(2):187–199, 1977.

[183] Frank Neese. Software update: the orca program system, version 4.0. *WIREs Computational Molecular Science*, 8(1):e1327, 2018.

[184] Aleksandr V. Marenich, Christopher J. Cramer, and Donald G. Truhlar. 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. *The Journal of Physical Chemistry B*, 113(18):6378–6396, 2009. PMID: 19366259.

[185] Jeng-Da Chai and Martin Head-Gordon. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.*, 10:6615–6620, 2008.

[186] Florian Weigend and Reinhart Ahlrichs. 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.*, 7:3297–3305, 2005.

[187] Jingjing Zheng, Xuefei Xu, and Donald G. Truhlar. Minimally augmented karlsruhe basis sets. *Theoretical Chemistry Accounts*, 128(3):295–305, 2011.

[188] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry*, 32(7):1456–1465, 2011.

[189] Róbert Izsák and Frank Neese. An overlap fitted chain of spheres exchange method. *The Journal of Chemical Physics*, 135(14):144105, 10 2011.

[190] Carl Caleman, Paul J. van Maaren, Minyan Hong, Jochen S. Hub, Luciano T. Costa, and David van der Spoel. Force field benchmark of organic liquids: Density, enthalpy of vaporization, heat capacities, surface tension, isothermal compressibility, volumetric expansion coefficient, and dielectric constant. *Journal of Chemical Theory and Computation*, 8(1):61–74, 01 2012.

[191] David van der Spoel, Paul J. van Maaren, and Carl Caleman. GROMACS molecule & liquid database. *Bioinformatics*, 28(5):752–753, 01 2012.

[192] Shankar Kumar, John M. Rosenberg, Djamal Bouzida, Robert H. Swendsen, and Peter A. Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of Computational Chemistry*, 13(8):1011–1021, 1992.

[193] R. Ian Storer, Caroline Aciro, and Lyn H. Jones. Squaramides: physical properties, synthesis and applications. *Chem. Soc. Rev.*, 40:2330–2346, 2011.

[194] Alessandra Picollo and Michael Pusch. Chloride/proton antiporter activity of mammalian clc proteins clc-4 and clc-5. *Nature*, 436(7049):420–423, 2005.

[195] Austin R. Graves, Patricia K. Curran, Carolyn L. Smith, and Joseph A. Mindell. The cl-/h+ antiporter clc-7 is the primary chloride permeation pathway in lysosomes. *Nature*, 453(7196):788–792, 2008.

[196] Raúl Estévez, Thomas Boettger, Valentin Stein, Ralf Birkenhäger, Edgar Otto, Friedhelm Hildebrandt, and Thomas J. Jentsch. Barttin is a cl- channel -subunit crucial for renal cl- reabsorption and inner ear k+ secretion. *Nature*, 414(6863):558–561, 2001.

[197] Björn Christian Schroeder, Tong Cheng, Yuh Nung Jan, and Lily Yeh Jan. Expression cloning of tmem16a as a calcium-activated chloride channel subunit. *Cell*, 134(6):1019–1029, 2008.

[198] Antonella Caputo, Emanuela Caci, Loretta Ferrera, Nicoletta Pedemonte, Cristina Barsanti, Elvira Sondo, Ulrich Pfeffer, Roberto Ravazzolo, Olga Zegarra-Moran, and Luis J. V. Galietta. Tmem16a, a membrane protein associated with calcium-dependent chloride channel activity. *Science*, 322(5901):590–594, 2008.

[199] Stephen J. Moore, Cally J. E. Haynes, Jorge González, Jennifer L. Sutton, Simon J. Brooks, Mark E. Light, Julie Herniman, G. John Langley, Vanessa Soto-Cerrato, Ricardo Pérez-Tomás, Igor Marques, Paulo J. Costa, Vítor Félix, and Philip A. Gale. Chloride, carboxylate and carbonate transport by ortho-phenylenediamine-based bisureas. *Chem. Sci.*, 4:103–117, 2013.

[200] Louise E. Karagiannidis, Cally J. E. Haynes, Katie J. Holder, Isabelle L. Kirby, Stephen J. Moore, Neil J. Wells, and Philip A. Gale. Highly effective yet simple transmembrane anion transporters based upon ortho-phenylenediamine bis-ureas. *Chem. Commun.*, 50:12050–12053, 2014.

[201] Paulo Vieira, Margarida Q. Miranda, Igor Marques, Sílvia Carvalho, Li-Jun Chen, Ethan N. W. Howe, Carl Zhen, Claudia Y. Leung, Michael J. Spooner, Bárbara Morgado, Odete A. B. da Cruz e Silva, Cristina Moiteiro, Philip A. Gale, and Vítor Félix. Development of a library of thiophene-based drug-like lego molecules: Evaluation of their anion binding, transport properties, and cytotoxicity. *Chemistry – A European Journal*, 26(4):888–899, 2020.

[202] Martina Olivari, Riccardo Montis, Stuart N. Berry, Louise E. Karagiannidis, Simon J. Coles, Peter N. Horton, Lucy K. Mapp, Philip A. Gale, and Claudia Caltagirone. Tris-ureas as transmembrane anion transporters. *Dalton Trans.*, 45:11892–11897, 2016.

[203] Tamara Merckx, Cally J. E. Haynes, Louise E. Karagiannidis, Harriet J. Clarke, Katie Holder, Alexandra Kelly, Graham J. Tizzard, Simon J. Coles, Peter Verwilst,

Philip A. Gale, and Wim Dehaen. Anion binding and transport properties of cyclic 2,6-bis(1,2,3-triazol-1-yl)pyridines. *Org. Biomol. Chem.*, 13:1654–1661, 2015.

[204] Christopher M. Dias, Hennie Valkenier, and Anthony P. Davis. Anthracene bisureas as powerful and accessible anion carriers. *Chemistry – A European Journal*, 24(23):6262–6268, 2018.

[205] Yufeng Yang, Xin Wu, Nathalie Busschaert, Hiroyuki Furuta, and Philip A. Gale. Dissecting the chloride–nitrate anion transport assay. *Chem. Commun.*, 53:9230–9233, 2017.

[206] Nathalie Busschaert, Samuel J. Bradberry, Marco Wenzel, Cally J. E. Haynes, Jennifer R. Hiscock, Isabelle L. Kirby, Louise E. Karagiannidis, Stephen J. Moore, Neil J. Wells, Julie Herniman, G. John Langley, Peter N. Horton, Mark E. Light, Igor Marques, Paulo J. Costa, Vítor Félix, Jeremy G. Frey, and Philip A. Gale. Towards predictable transmembrane transport: Qsar analysis of anion binding and transport. *Chem. Sci.*, 4:3036–3045, 2013.

[207] Nicola J. Knight, Elsa Hernando, Cally J. E. Haynes, Nathalie Busschaert, Harriet J. Clarke, Koji Takimoto, María García-Valverde, Jeremy G. Frey, Roberto Quesada, and Philip A. Gale. Qsar analysis of substituent effects on tambjamine anion transporters. *Chem. Sci.*, 7:1600–1608, 2016.

[208] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. Qsar modeling: Where have you been? where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, 2014. PMID: 24351051.

[209] Christian Tyrchan, Eva Nittinger, Dea Gogishvili, Atanas Patronov, and Thierry Kogej. Chapter 4 - approaches using ai in medicinal chemistry. In Takashiro Akitsu, editor, *Computational and Data-Driven Chemistry Using Artificial Intelligence*, pages 111–159. Elsevier, 2022.

[210] Alexander Tropsha. Best practices for qsar model development, validation, and exploitation. *Molecular Informatics*, 29(6-7):476–488, 2010.

[211] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, and Mohamed Jawed Ahsan. Machine learning in drug discovery: A review. *Artificial Intelligence Review*, 55(3):1947–1999, 2022.

[212] Jon Paul Janet and Heather J. Kulik. *Machine Learning in Chemistry*. American Chemical Society, May 2020.

[213] Adam C. Mater and Michelle L. Coote. Deep learning in chemistry. *Journal of Chemical Information and Modeling*, 59(6):2545–2559, 2019.

[214] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, July 2018.

[215] Nongnuch Artrith, Keith T Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nature Chemistry*, 13:505–508, 2021.

[216] Geoffrey I. Webb. *Occam's Razor*, pages 735–735. Springer US, Boston, MA, 2010.

[217] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences*, 42(6):1273–1280, 11 2002.

[218] Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, 05 1985.

[219] Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. Topological torsion: a new molecular descriptor for sar applications. comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85, 05 1987.

[220] Alberto Gobbi and Dieter Poppinger. Genetic optimization of combinatorial libraries. *Biotechnology and Bioengineering*, 61(1):47–54, 1998.

[221] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. PMID: 20426451.

[222] Rdkit: Open-source cheminformatics. `http://www.rdkit.org`.

[223] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1):1–14, 2018.

[224] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. *Feature selection for high-dimensional data*. Springer, 2015.

[225] Xue Ying. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing, 2019.

[226] Davide Anguita, Alessandro Ghio, Sandro Ridella, and Dario Sterpi. K-fold cross validation for error rate estimate in support vector machines. In *DMIN*, pages 291–297, 2009.

[227] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[228] Isabelle Guyon, Amir Saffari, Gideon Dror, and Gavin Cawley. Model selection: beyond the bayesian/frequentist divide. *Journal of Machine Learning Research*, 11(1), 2010.

[229] Mohamed Bekkar, Hassiba Kheliouane Djemaa, and Taklit Akrouf Alitouche. Evaluation measures for models assessment over imbalanced data sets. *J Inf Eng Appl*, 3(10), 2013.

[230] Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*, volume 10. Springer, 2018.

[231] Nitesh Chawla, Kevin Bowyer, Lawrence Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res. (JAIR)*, 16:321–357, 06 2002.

[232] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[233] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[234] J.R. Quinlan. Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2):339–346, 1990.

[235] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[236] P.W. Atkins, J. De Paula, and J. Keeler. *Atkins' Physical Chemistry: Quantum Chemistry, Spectroscopy, and Statistical Thermodynamics*. Oxford University Press, 2018.

[237] Vinícius G Costa and Carlos E Pedreira. Recent advances in decision trees: An updated survey. *Artificial Intelligence Review*, 56(5):4765–4800, 2023.

[238] Michele Fratello and Roberto Tagliaferri. Decision trees and random forests. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 374, 2018.

[239] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[240] Gongde Guo, Hui Wang, David Bell, Yaxin Bi, and Kieran Greer. Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings*, pages 986–996. Springer, 2003.

[241] Sergio González, Salvador García, Javier Del Ser, Lior Rokach, and Francisco Herrera. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, 64:205–237, 2020.

[242] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[243] Robert E Schapire. Explaining adaboost. In *Empirical inference*, pages 37–52. Springer, 2013.

[244] Gunnar Rätsch, Takashi Onoda, and K-R Müller. Soft margins for adaboost. *Machine learning*, 42:287–320, 2001.

[245] Thomas G Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40:139–157, 2000.

[246] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[247] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[248] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

[249] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[250] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[251] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[252] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[253] Christoph A Schalley. *Analytical methods in supramolecular chemistry*, volume 1. John Wiley & Sons, 2012.

[254] Xin Wu and Philip A Gale. Measuring anion transport selectivity: a cautionary tale. *Chem. Commun.*, 57(33):3979–3982, 2021.

[255] Darren R Flower. On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Computer Sciences*, 38(3):379–386, may 1998.

[256] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[257] David T Stanton and Peter C Jurs. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. *Analytical Chemistry*, 62(21):2323–2329, 1990.

[258] Johann Gasteiger and Mario Marsili. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.

[259] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. Bayesian data analysis, third edition. 2013.

[260] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA, 2016. Association for Computing Machinery.

[261] Peter Gedeck, Bernhard Rohde, and Christian Bartels. Qsar how good is it in practice? comparison of descriptor sets on an unbiased cross section of corporate data sets. *Journal of Chemical Information and Modeling*, 46(5):1924–1936, 09 2006.

[262] Ioan Iacovache, Sacha De Carlo, Nuria Cirauqui, Matteo Dal Peraro, F. Gisou van der Goot, and Benoît Zuber. Cryo-em structure of aerolysin variants reveals a novel protein fold and the pore-formation process. *Nature Communications*, 7(1):12062, 2016.

[263] Chan Cao, Yi-Lun Ying, Zheng-Li Hu, Dong-Fang Liao, He Tian, and Yi-Tao Long. Discrimination of oligonucleotides of different lengths with a wild-type aerolysin nanopore. *Nature Nanotechnology*, 11(8):713–718, 2016.

[264] Fabien Piguet, Hadjer Ouldali, Manuela Pastoriza-Gallego, Philippe Manivet, Juan Pelta, and Abdelghani Oukhaled. Identification of single amino acid differences in uniformly charged homopolymeric peptides with aerolysin nanopore. *Nature Communications*, 9(1):966, 2018.

[265] Chan Cao, Nuria Cirauqui, Maria Jose Marcaida, Elena Buglakova, Alice Duperrex, Aleksandra Radenovic, and Matteo Dal Peraro. Single-molecule sensing of peptides and nucleic acids by engineered aerolysin nanopores. *Nature Communications*, 10(1):4918, 2019.

[266] Si-Min Lu, Xue-Yuan Wu, Meng-Yin Li, Yi-Lun Ying, and Yi-Tao Long. Diversified exploitation of aerolysin nanopore in single-molecule sensing and protein sequencing. *VIEW*, 1(4):20200006, 2020.

Appendices

Appendix A

Supplementary information for Chapter 3

A.1 MD simulations of 1^Z in DMSO

Two systems were simulated in explicit DMSO: as a free ligand (1^Z) and in a 1:1 complex with chloride (1^Z-Cl).

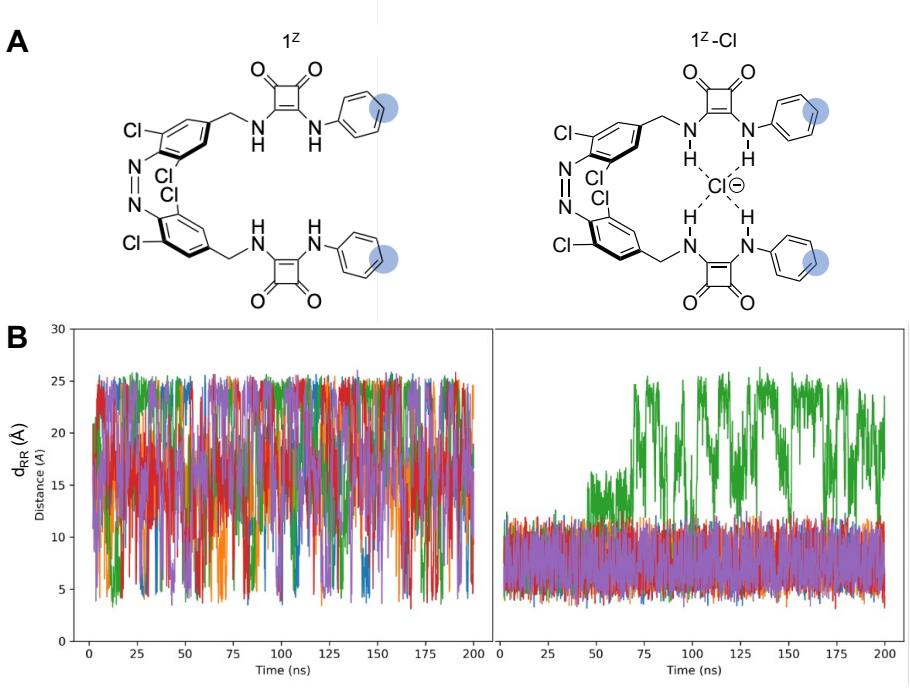


Figure A1: Time evolution of the end-to-end distance (d_{RR}) of 1^Z and 1^Z-Cl over 5 replicates of 200 ns simulation in DMSO. d_{RR} is measured between the carbon atoms circled in blue.

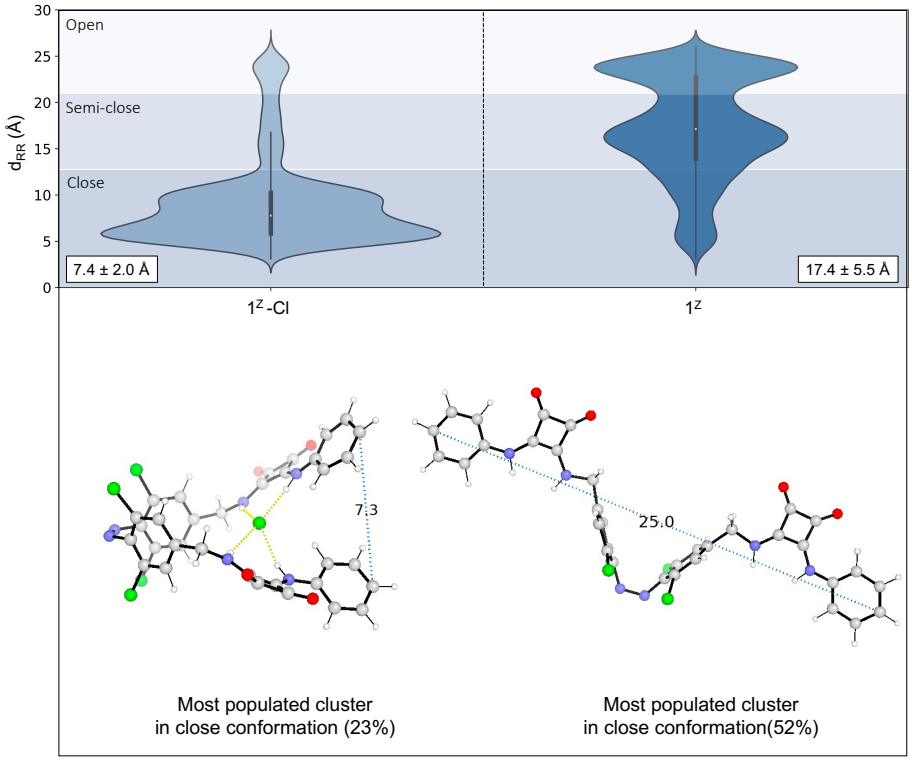


Figure A2: Model Z-isomer in DMSO. (A) 1^Z as a free ligand (left) and as a 1:1 complex with a chloride anion (right). The end-to-end distance is measured between the carbon atoms circled in blue and labelled as d_{RR}. (B) Time evolution of the end-to-end distance of 1^Z and 1^Z-Cl over 5 replicates of 200 ns simulation in DMSO. (C) Top: Distance histograms of the end-to-end distance in (B). Bottom: The structure and percentage population of the most populated clusters, with their corresponding d_{RR} shown as blue dash lines. Cluster analysis was performed and identified 60 clusters with a cut-off of 2 Å. Adapted from Reference[107] with permission.

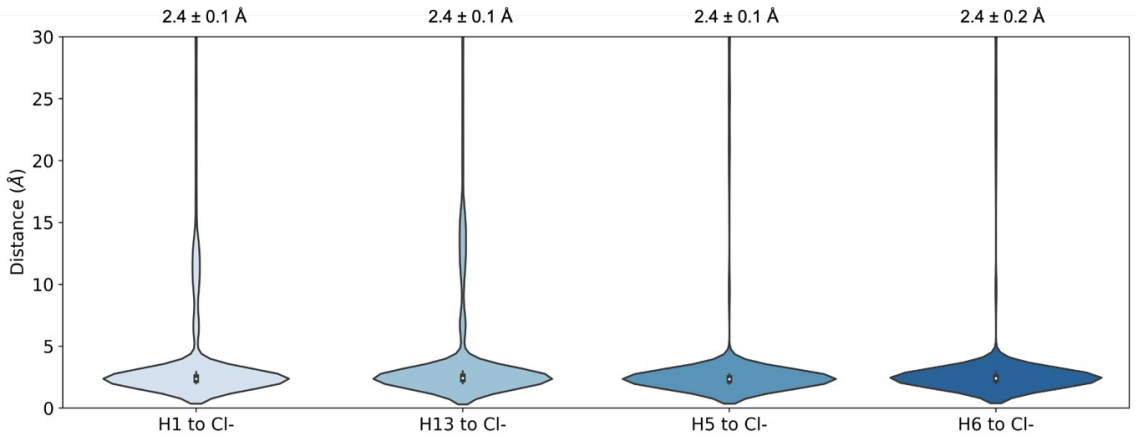


Figure A3: Hydrogen bond distance between the squaramide NH protons to the chloride anion (d_{NH-Cl}) in a 1^Z-Cl in DMSO over 5 replicates of 200 ns explicitly solvated MD simulations.

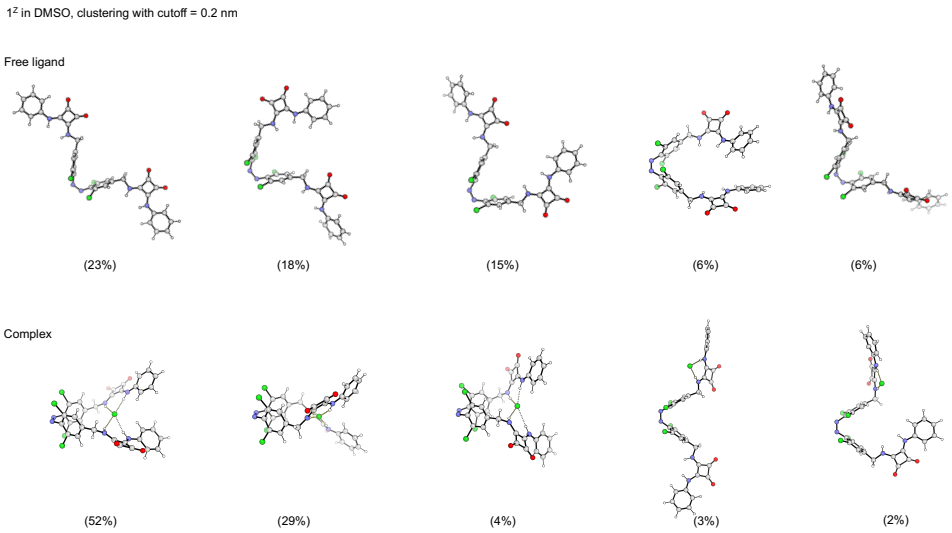


Figure A4: Structure and population of the most populated clusters of 1^{Z}-Cl in DMSO over 5 replicates of 200 ns explicitly solvated MD simulations.

A.2 MD simulations of 12^Z and 13^Z in DMSO

To explore the role of the linker position and length in chloride binding, we varied the linker position from para to meta (transporter 12) and linker length from one to two carbon (transporter 13) on the azobenzene core. We modelled compounds 12^Z and 13^Z in their free form and complexed with chloride in an explicitly solvated (DMSO) MD simulation for 200 ns.

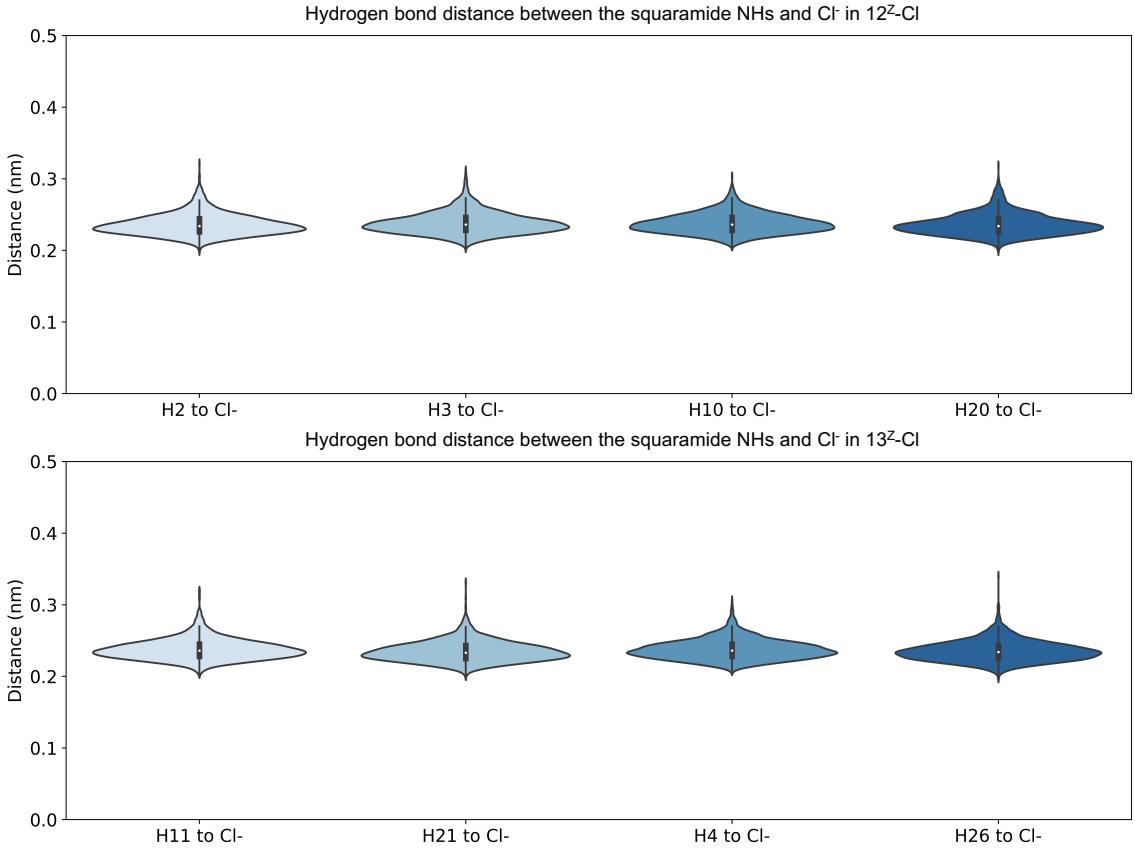


Figure A5: HB distances in 12^Z-Cl and 13^Z-Cl in DMSO over a 200 ns explicitly solvated MD simulation.

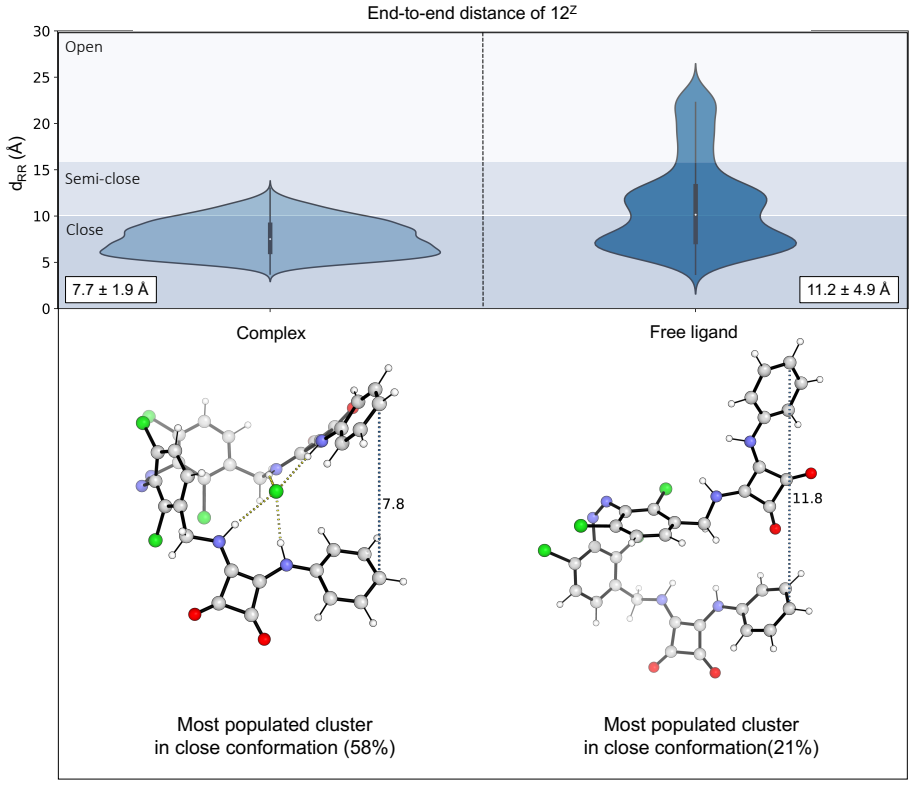


Figure A6: End-to-end distance of 12^{Z}-Cl in DMSO over a 200 ns explicitly solvated MD simulation.

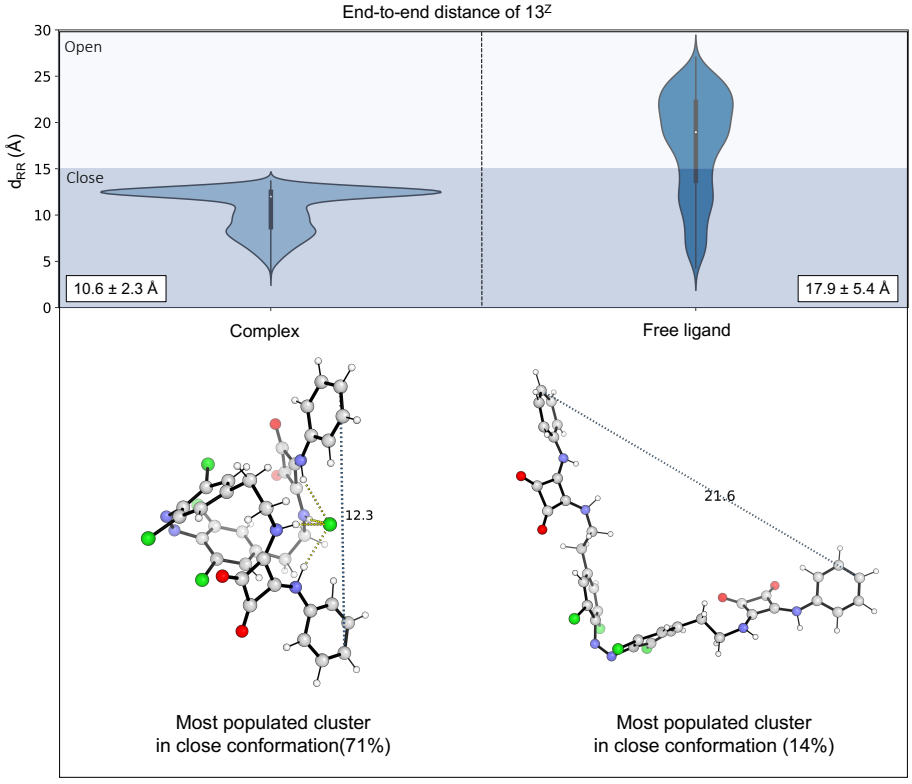


Figure A7: End-to-end distance of 13^{Z}-Cl in DMSO over a 200 ns explicitly solvated MD simulation.

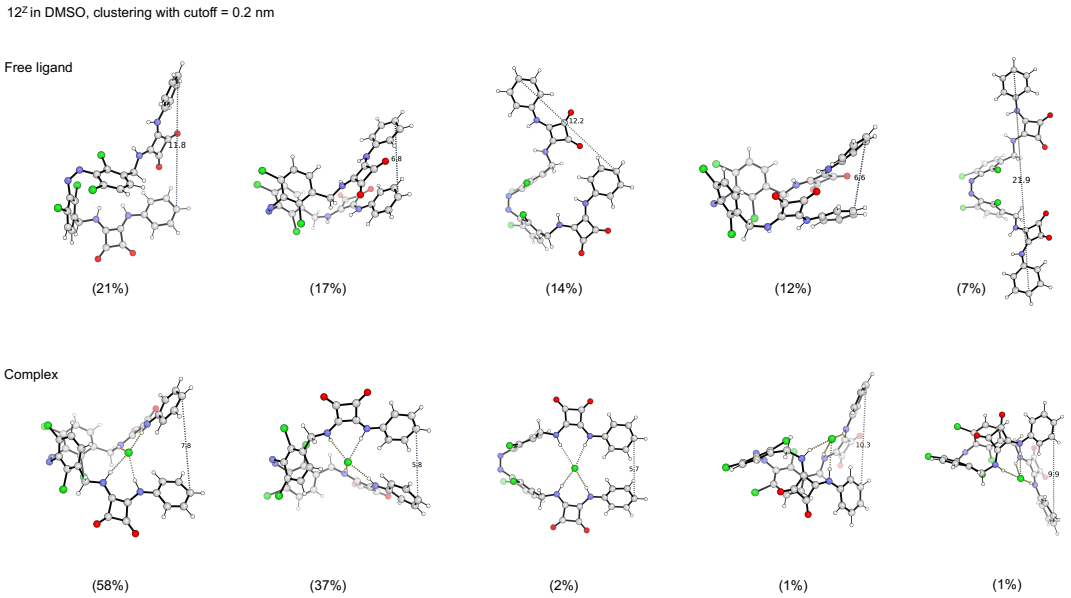


Figure A8: Structure and population of the most populated clusters of 12^Z-Cl in DMSO over a 200 ns explicitly solvated MD simulation. Clustering was performed with a cut-off of 0.2 nm on 1981 conformers.

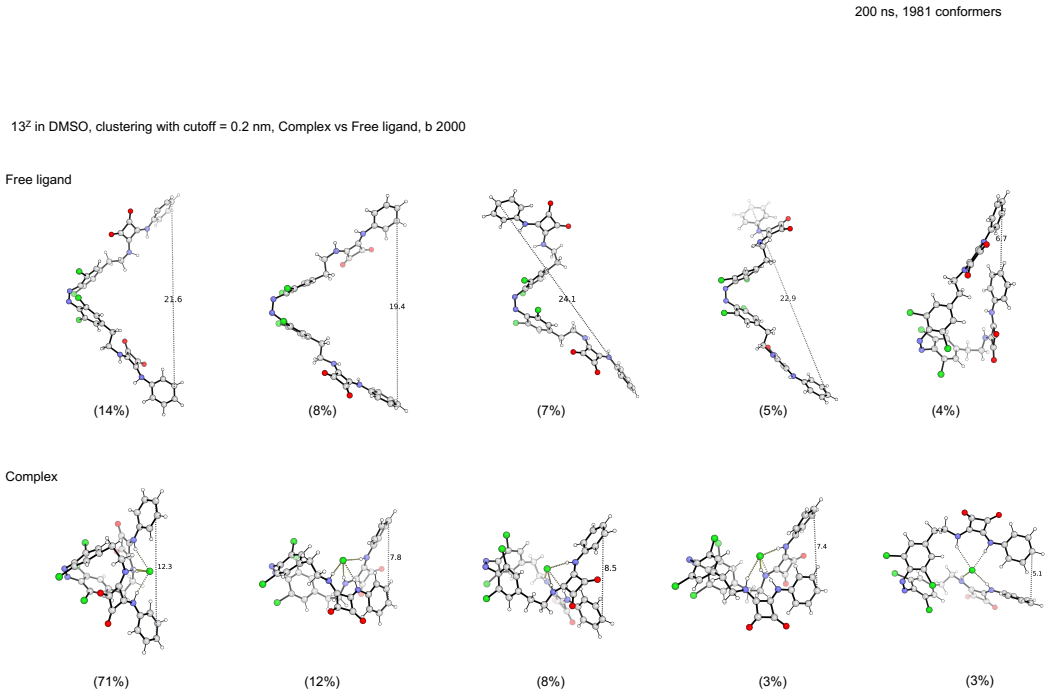


Figure A9: Structure and population of the most populated clusters of 13^Z-Cl in DMSO over a 200 ns explicitly solvated MD simulation. Clustering was performed with a cut-off of 0.2 nm on 1981 conformers.

A.3 MD simulations of 1^{E} in DMSO

To explore the difference between the *E*- and *Z*-isomers. We modelled the model *E*-isomer both as a free ligand (1^{E}) and as a 1:1 complex with a chloride anion (1^{E}-Cl) in DMSO. Because squaramide derivatives are well known to form strong, bidentate intermolecular HBs,[193] we hypothesized that the *E*-isomer could achieve better ion encapsulation by forming 4 HBs to the chloride anion in a 2:1 complex (1_2^{E}-Cl). For each system, explicitly solvated MD simulations were propagated over 5 replicates of 200 ns.

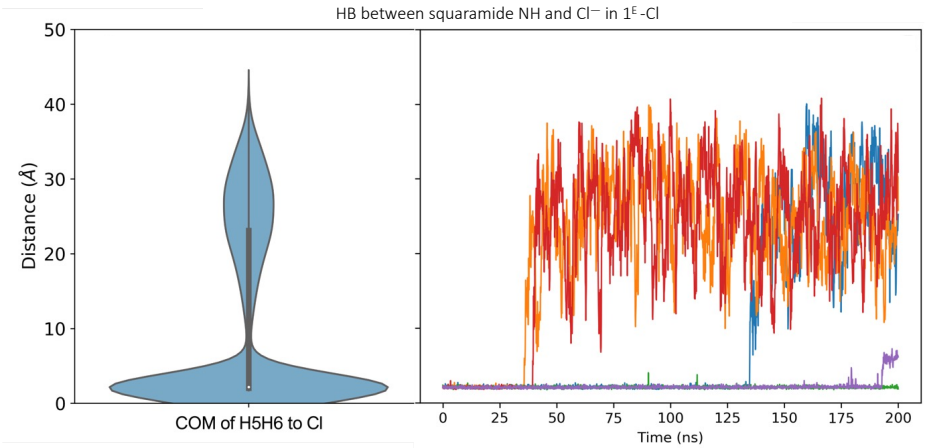


Figure A10: Distribution (left) and time evolution (right) of hydrogen bond distance between the squaramide NH protons to the chloride anion in 1^{E}-Cl in DMSO over 5 replicates of 200 ns explicitly solvated MD simulations.

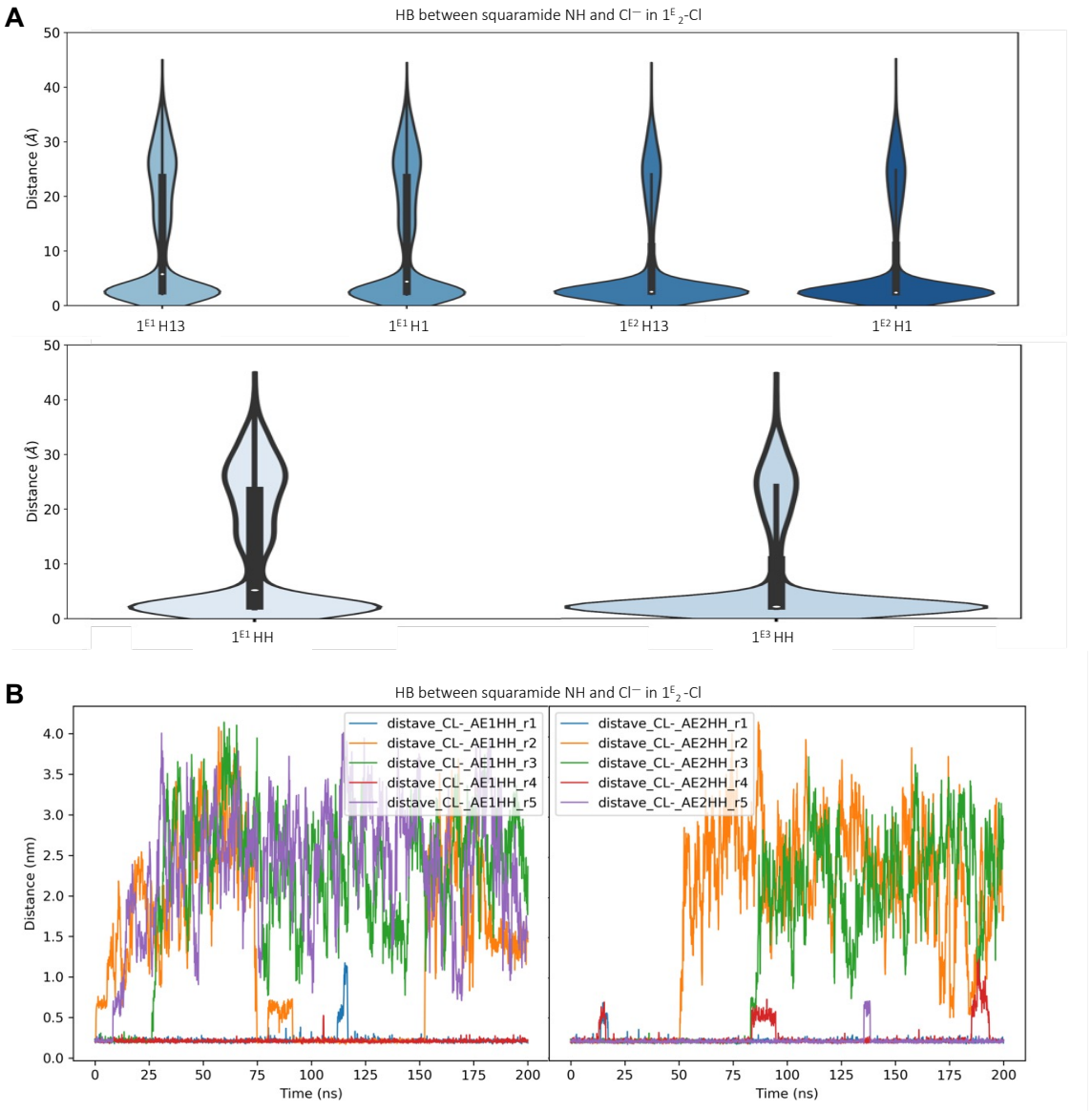


Figure A11: (A) Distribution and (B) time evolution of hydrogen bond distance between the squaramide NH protons to the chloride anion in 1_2^E-Cl in DMSO over 5 replicates of 200 ns explicitly solvated MD simulations.

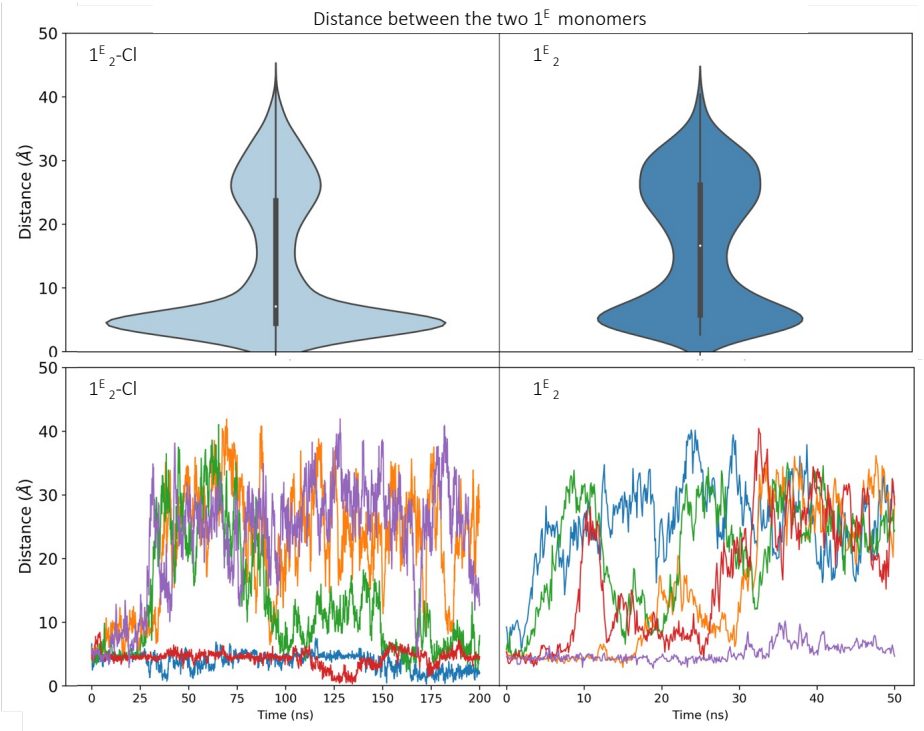


Figure A12: Distribution (top) and time evolution (bottom) of the distance between the centre of mass of the two 1^E monomers in 1_2^E-Cl (left) and 1_2^E (right) in DMSO.

A.4 Free energy calculations of chloride transport across a POPC membrane

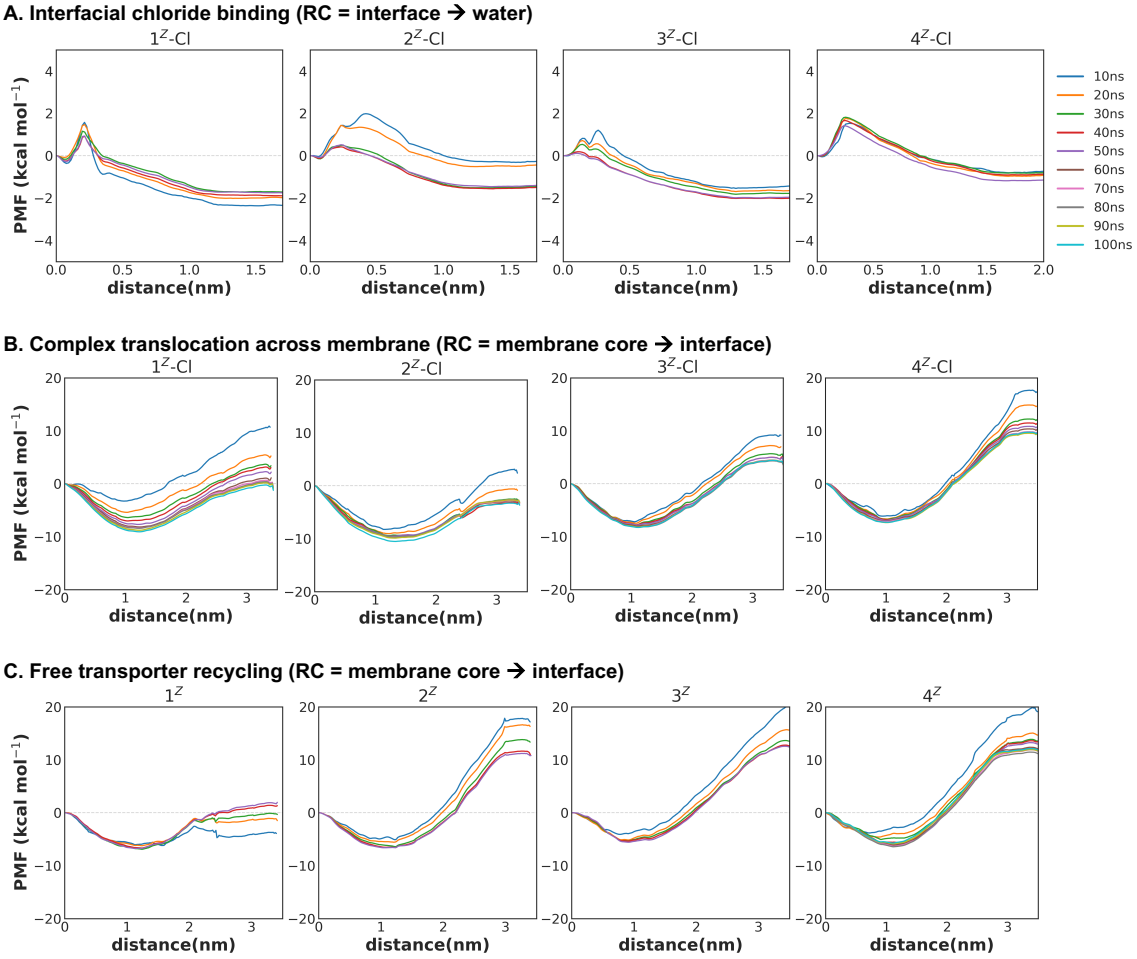


Figure A13: Convergence of the umbrella sampling simulation windows.

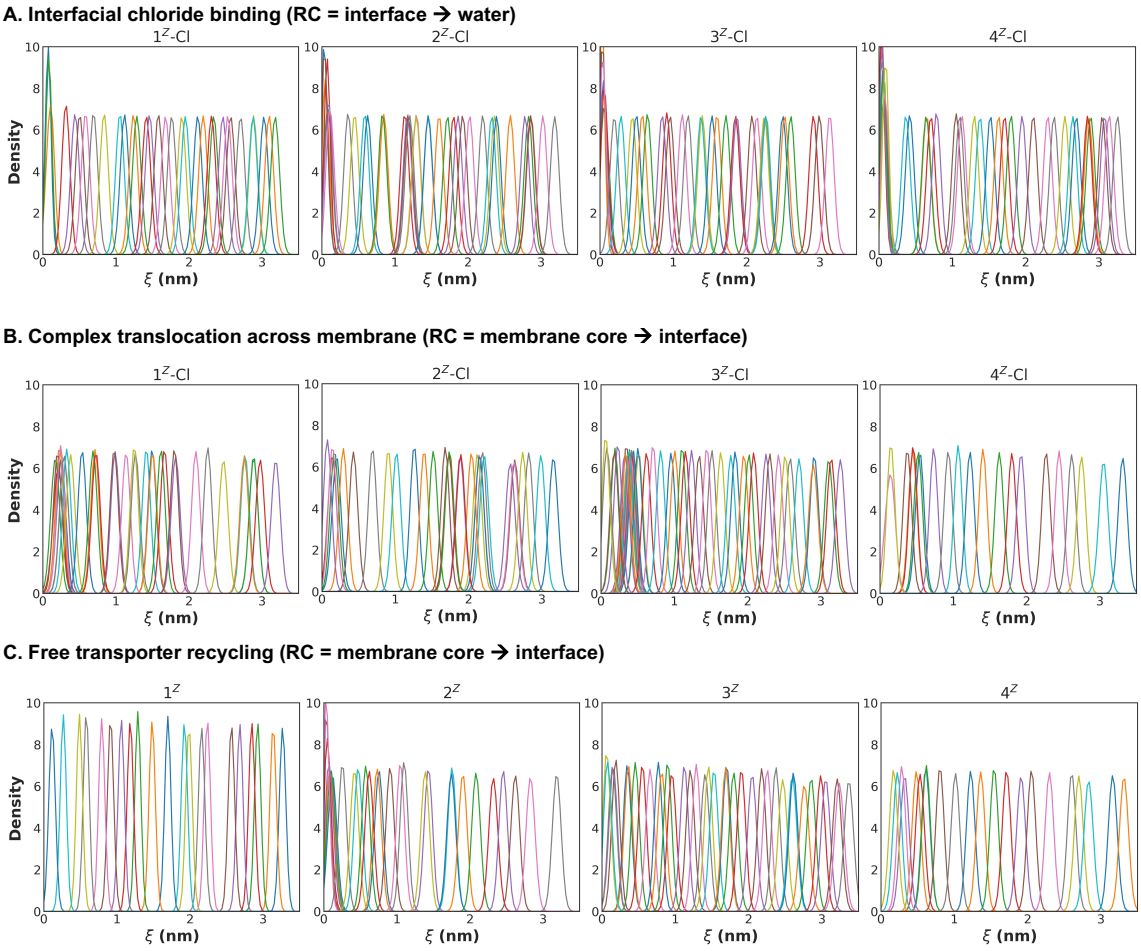


Figure A14: Overlap of the umbrella sampling simulation windows.

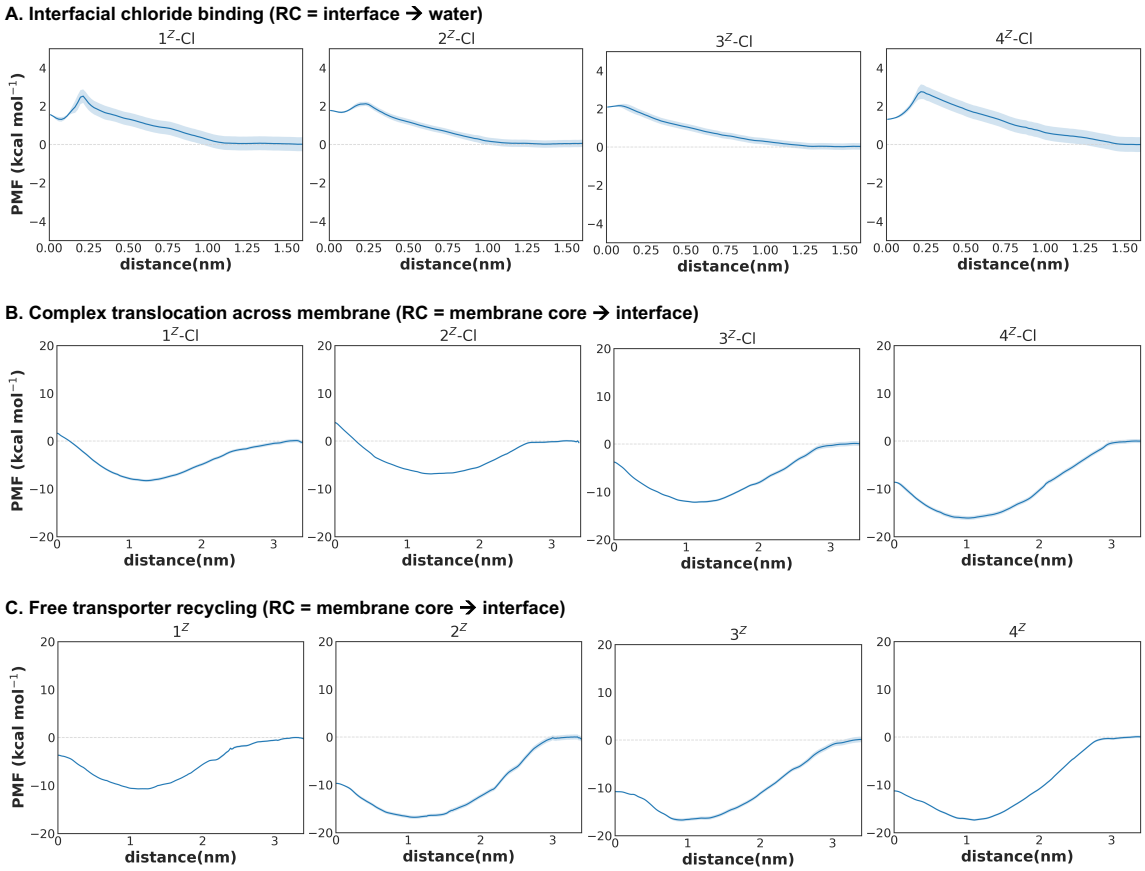


Figure A15: Potential of mean force profiles with 200 bootstraps.

Appendix B

Supplementary information for Chapter 4

B.1 Dataset analysis

We analyzed the substructure composition within the dataset by substructure matching to the SMARTS of common binding motifs, including thiourea, urea, squaramide, indole, tambjamine, carbazole, acridinone, perenosin, bisbzimpy, and isophthalamides/dipicolineamides (Figure B1). Thiourea and urea exhibit the highest occurrences, with 86 and 113 molecules respectively (Table B1. We explored the distribution of various substructures within each year (Figure B2). Thiourea and urea substructures exhibit varying occurrences across the years but are present in both the training (2011-2020) and test set (2021). Overlaps between different substructure datasets demonstrate the consideration of hybrid/shared binding modes and molecular complexity of the dataset (Figure B3). We also assessed the Tanimoto similarity between molecules within each substructure set and found good structural variety within each of them (Figure B4).

Substructure	Count
Urea	113
Thiourea	86
Tambjamine	45
Indole	41
Squaramide	30
Perenosin	15
Carbazole	12
Acridinone	10
Isophthalamides/dipicolineamides	8
BisBzImPy	5

Table B1: Number of molecules containing common binding motifs.

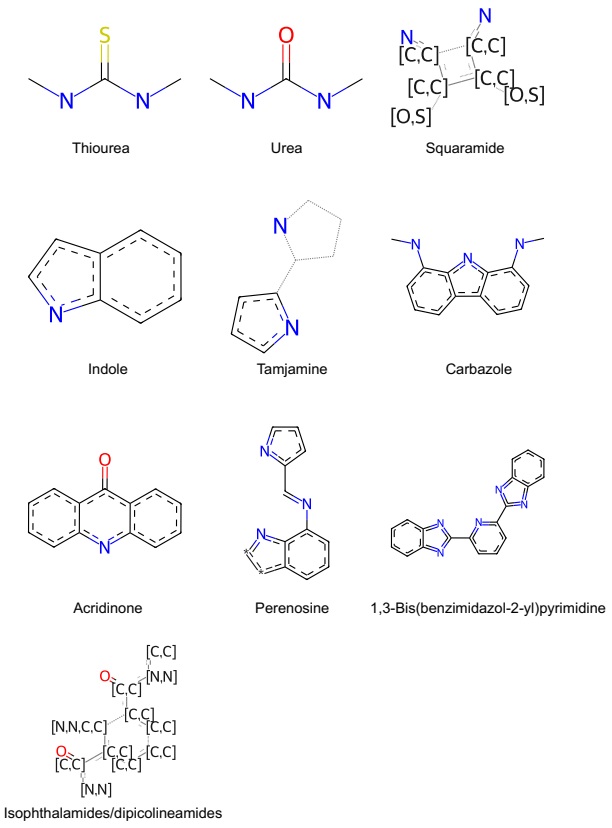


Figure B1: Substructure smarts

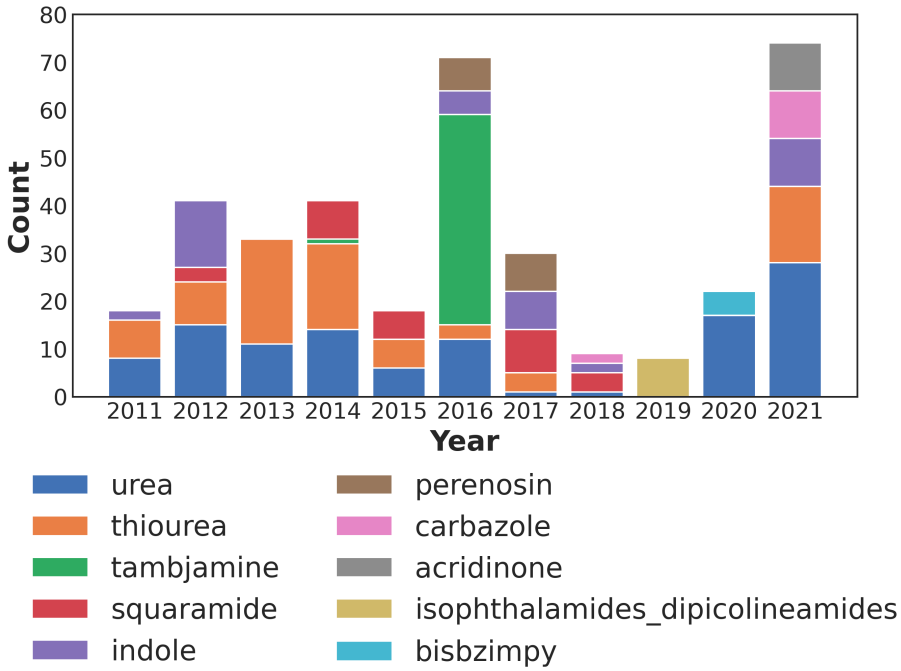


Figure B2: Substructures within each year

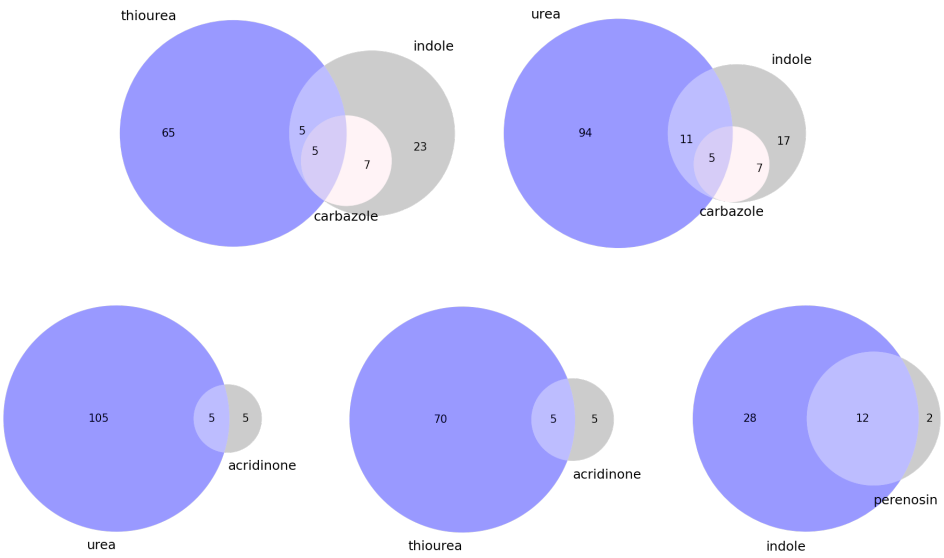


Figure B3: Venn diagrams of substructure composition in the entire dataset.

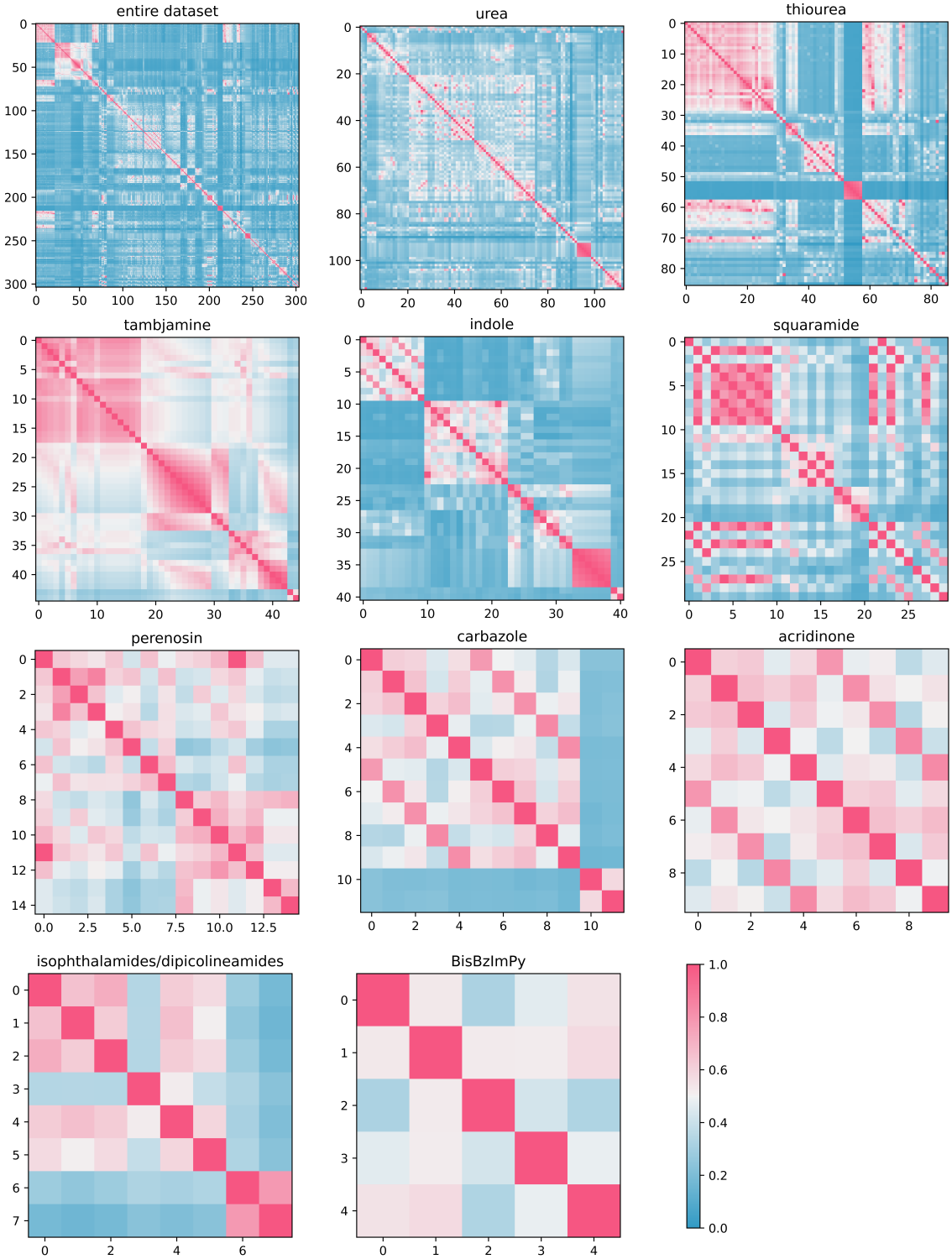


Figure B4: Tanimoto similarity within substructure datasets.

B.1.1 Train-test split

For the train-test split, we organized the dataset chronologically based on publication dates and adopted a temporal split strategy. Data from years earlier than 2021 formed the training set, while data from 2021 comprised the testing set.

For classification tasks, 260 data points fall into the training set and 44 data points fall into the test set (Figure B5A). Both the training and test sets exhibited a remarkable degree of structural diversity (Figure B5B) and share low similarity between them, indicating that the molecules in the test set are sufficiently distinct from those in the training set. This ensures that our model's evaluation is based on unseen data with unique structural characteristics. The robust structural diversity within the training and test sets, combined with the low similarity between them, suggests that our model will generalize well to new and diverse chemical structures, making it a reliable predictor of transport activities for molecules outside the training set.

For regression tasks, after removing arbitrary values, 241 transport activity values remained. 260 data points fall into the training set and 25 data points fall into the test set (Figure B6A). Both the training and test sets exhibited a remarkable degree of structural diversity with an average Tanimoto index of 0.2 and 0.3 respectively (Figure B6B). The training and test set also share low similarity between them, suggesting that the molecules in the test set are sufficiently distinct from those in the training set.

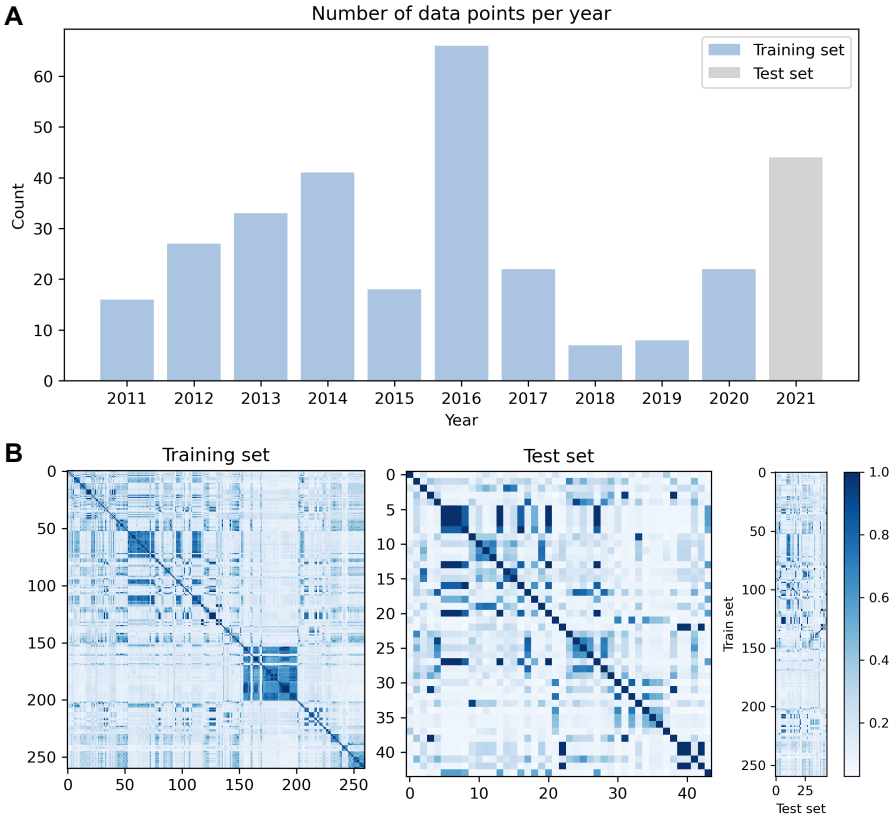


Figure B5: Temporal split of the dataset for classification tasks. (A) A temporal split of the dataset was performed. Data from years earlier than 2021 formed the training set, while data from 2021 comprised the testing set. (B) Structural similarity within and between the training and test sets.

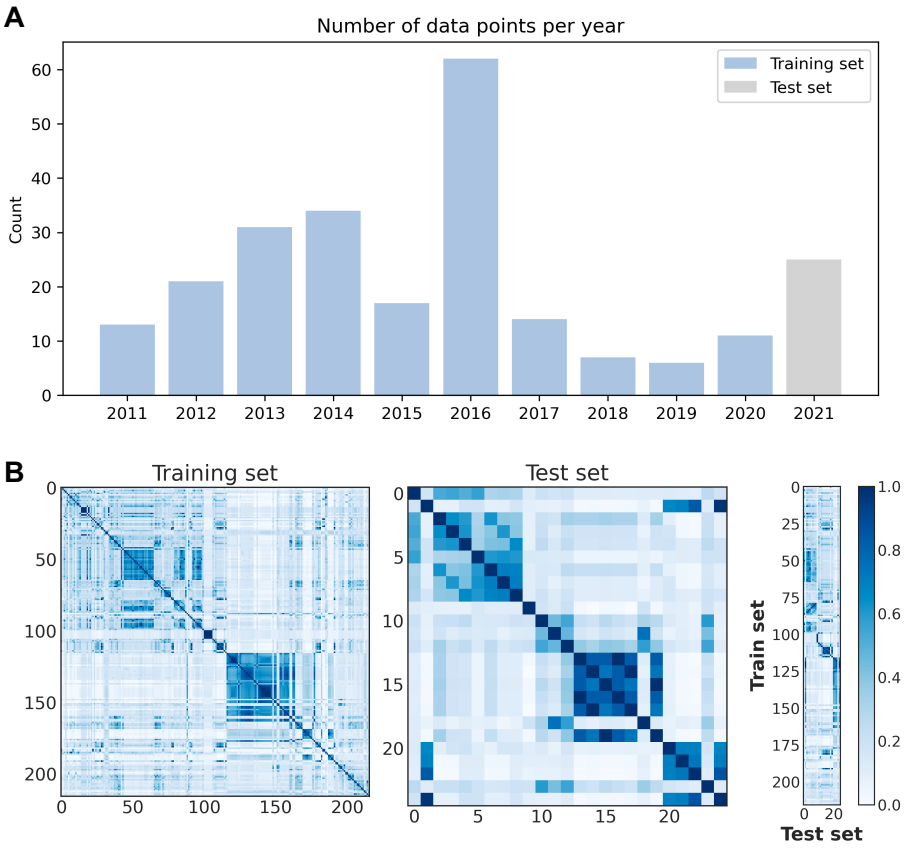


Figure B6: Dataset of synthetic anion transporters for regression. (A) A temporal split of the dataset was performed. Data from years earlier than 2021 formed the training set, while data from 2021 comprised the testing set. (B) Structural similarity within and between the training and test sets.

B.2 Classification models

B.2.1 Addressing class imbalance with class weights

Class imbalance is a common challenge in classification tasks where one class has significantly fewer instances than the other(s). This imbalance can lead to biased model performance, as classifiers tend to favour the majority class due to its larger representation in the training data. As a result, minority class instances are often misclassified, leading to poor generalization and low predictive accuracy. In Chapter 4, we used synthetic sampling to oversample the minority class to obtain a balanced dataset. Class weights is another method to tackle class imbalance.

In the class weights method, different weights are assigned to the classes during the training process. By assigning higher weights to the minority class and lower weights to the majority class, the algorithm pays more attention to correctly classifying minority class instances. This encourages the model to focus on improving accuracy for the underrepresented class, effectively mitigating the impact of class imbalance. The class weights are calculated as follows:

$$w_j = \frac{n_{\text{samples}}}{n_{\text{classes}} \times n_{\text{samples}_j}}$$

where w_j represents the weight assigned to each class j. n_{samples} denotes the total number of samples or rows present in the dataset. n_{classes} indicates the total number of distinct classes within the target variable. n_{samples_j} corresponds to the total number of rows associated with class j.

In our dataset, there are 126 and 178 data points in class 0 and class 1 respectively. Using the above equation, the balanced class weights for class 0 and class 1 are 1.20634921 and 0.85393258 respectively. The higher weight for class 0 indicates that

the algorithm will focus more on this class to mitigate its underrepresentation, while the lower weight for class 1 signifies that the algorithm will pay slightly less emphasis to this class due to its higher representation. By incorporating these class weights into the training process, we aim to address the class imbalance issue and improve the model's ability to make accurate predictions for both classes.

We tested the calculated class weights with Morgan fingerprints as input features on three classifiers: decision tree (DT),[234] Extra Trees (ET),[242] and logistic regression (LR).[232] Surprisingly, compared to the baseline model trained on the imbalanced dataset (Table B2, Figure B7), there were no improvements in the model performance when utilising the balanced class weights (Table B3, Figure B8). This outcome suggests that the class weight adjustments alone might not have been sufficiently effective in mitigating the adverse effects of class imbalance for these classifiers. In comparison, using SMOTE to oversample the minority class (class 0) resulted in a significant enhancement in model performance, as shown in Chapter 4.

This result underscores the importance of carefully selecting and experimenting with appropriate techniques tailored to the dataset. While class weights are designed to adapt the learning process according to class distribution, the success of SMOTE in improving model performance suggests that the creation of synthetic instances for the minority class helped enrich the training data, enabling the classifiers to capture previously overlooked patterns.

	Accuracy	Sensitivity	Specificity	MCC	Precision	G-mean
Decision tree	0.77	0.83	0.69	0.53	0.79	0.76
Extra trees	0.85	0.90	0.78	0.69	0.85	0.84
Logistic regression	0.79	0.85	0.70	0.56	0.80	0.77

Table B2: Model performance on the imbalanced dataset without class weights. The anion transporters are represented by Morgan fingerprints.

	Accuracy	Sensitivity	Specificity	MCC	Precision	G-mean
Decision tree	0.76	0.82	0.68	0.51	0.78	0.75
Extra trees	0.84	0.85	0.83	0.68	0.87	0.84
Logistic regression	0.78	0.78	0.77	0.55	0.83	0.78

Table B3: Model performance on an imbalanced dataset with balanced class weights. The anion transporters are represented by Morgan fingerprints.

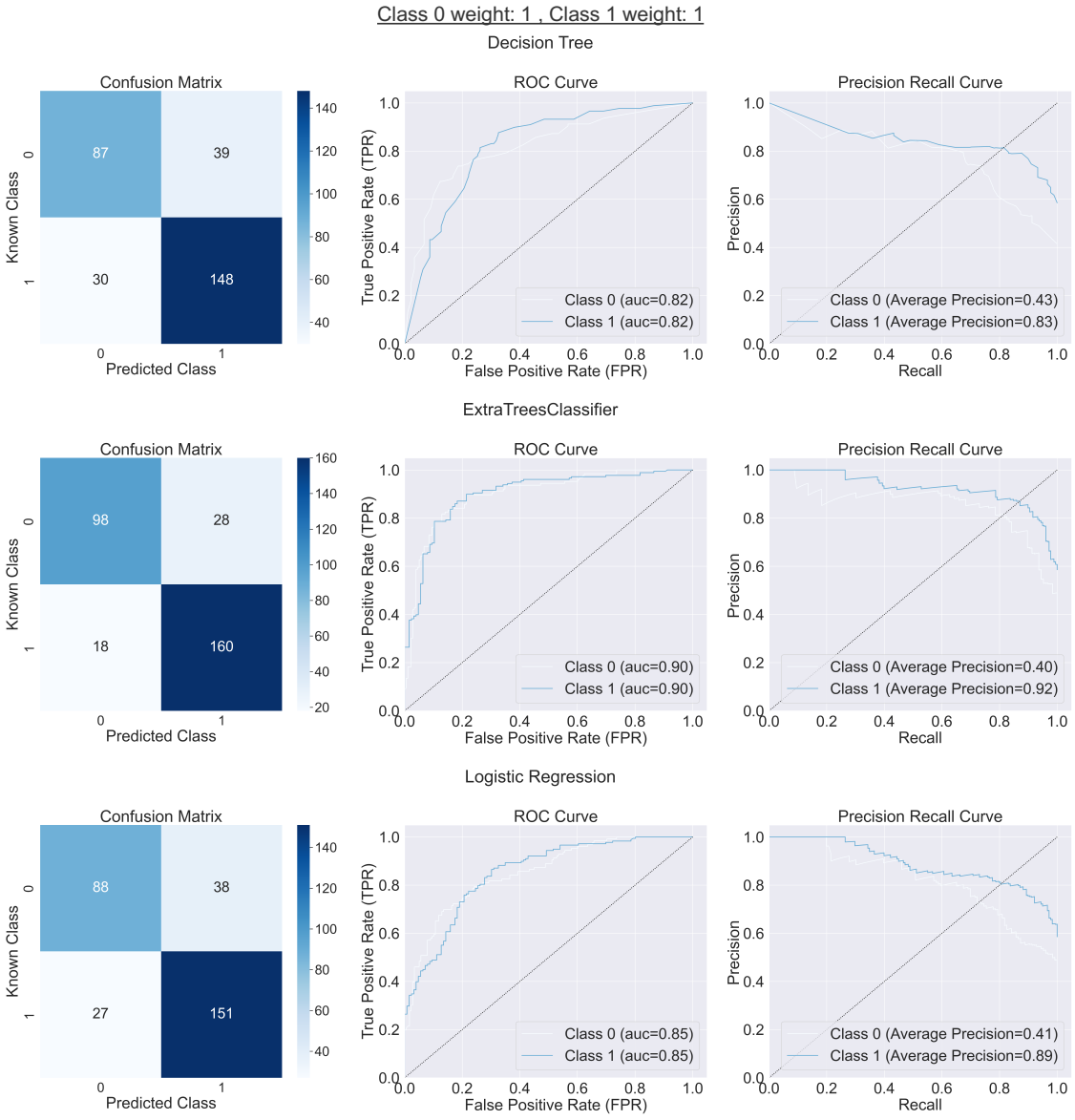


Figure B7: Model performance on the imbalanced dataset without class weights. The anion transporters are represented by Morgan fingerprints.

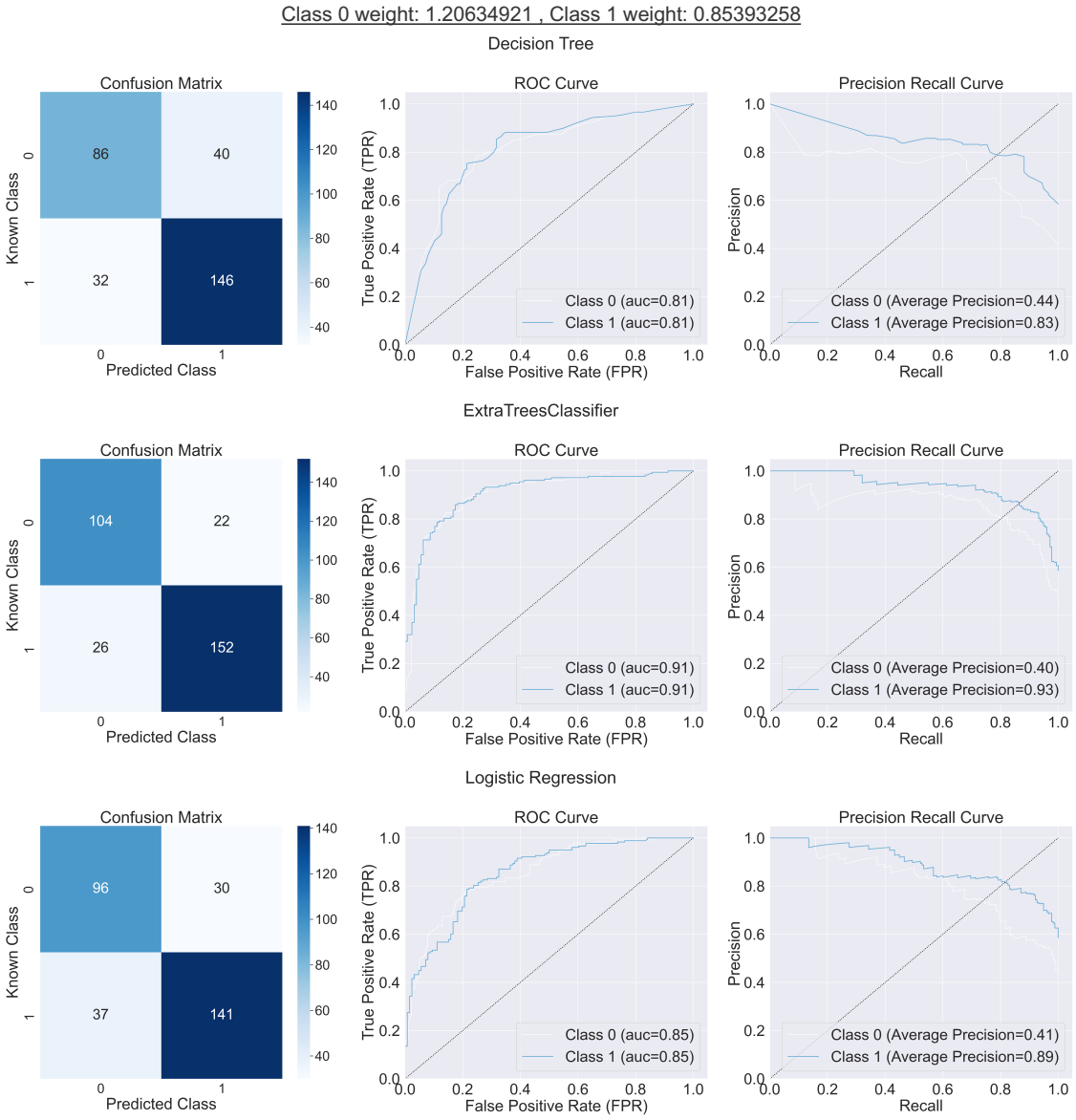


Figure B8: Model performance on an imbalanced dataset with balanced class weights. The anion transporters are represented by Morgan fingerprints.

B.2.2 Model performance on Morgan fingerprints

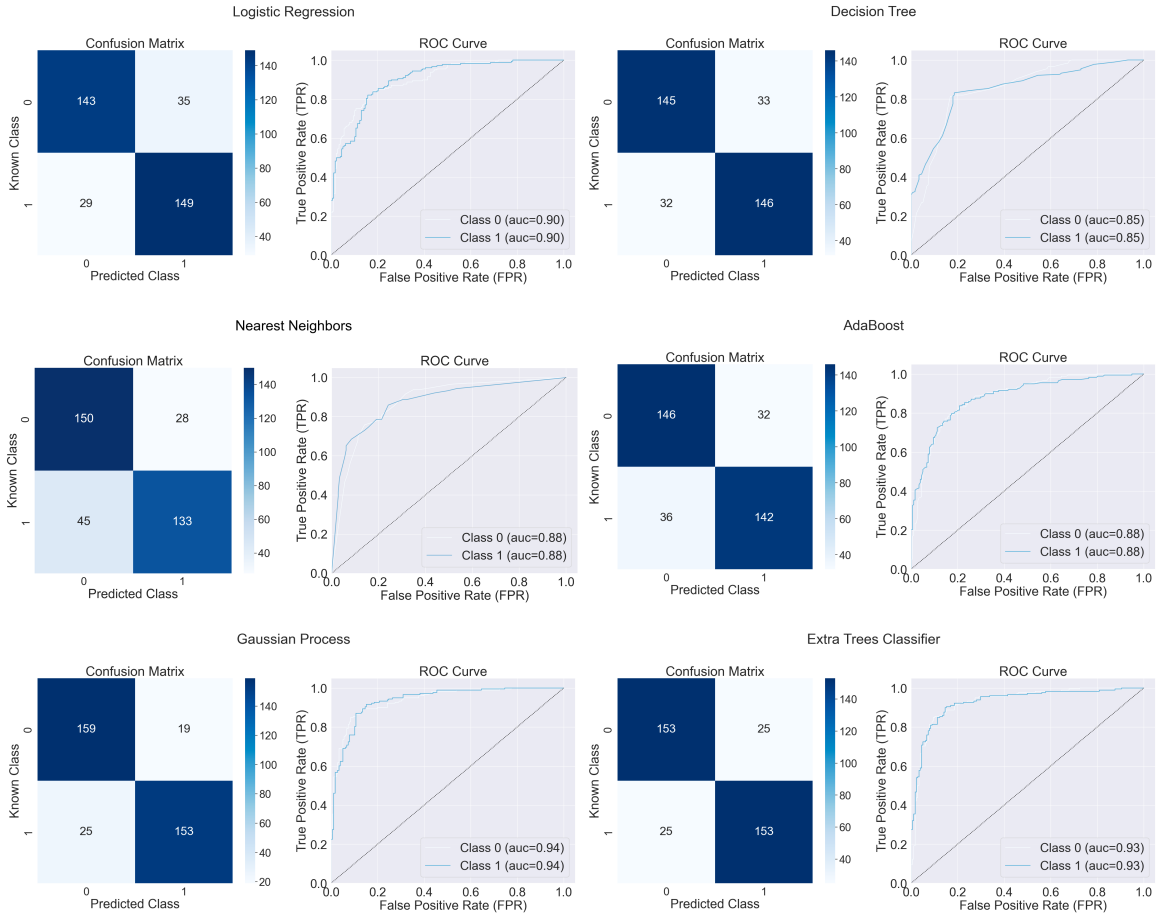


Figure B9: Classification model performance on Morgan fingerprints cross-validated on the entire dataset.

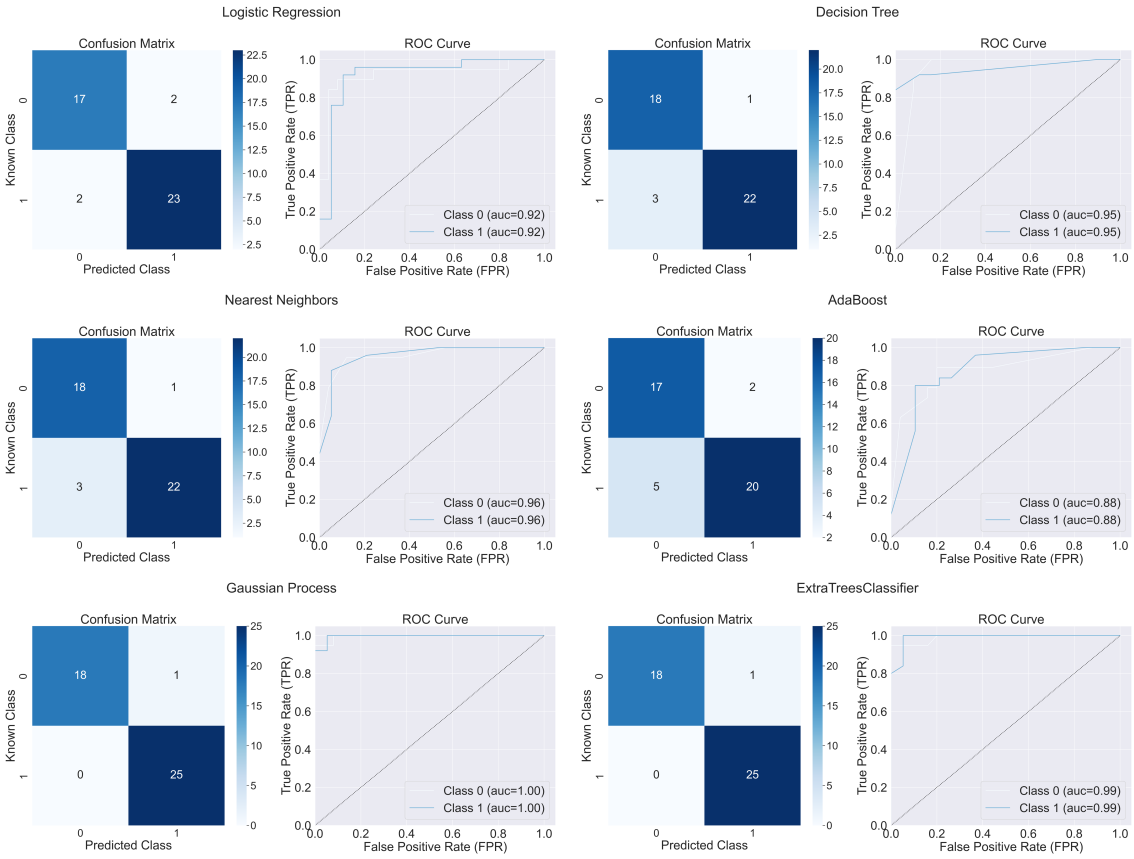


Figure B10: Classification model performance on Morgan fingerprints on the test set.

B.2.3 Model performance on Mordred descriptors

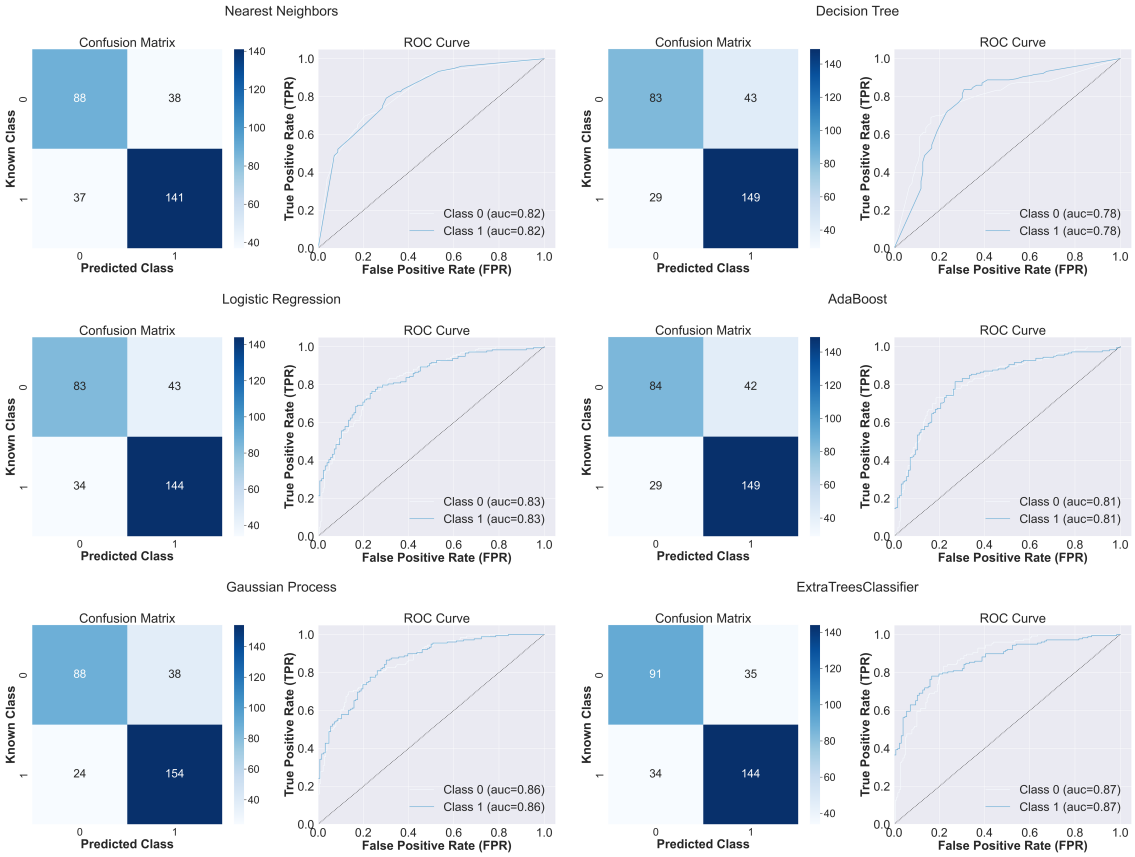


Figure B11: Classification model performance on Mordred descriptors cross-validated on the entire dataset.

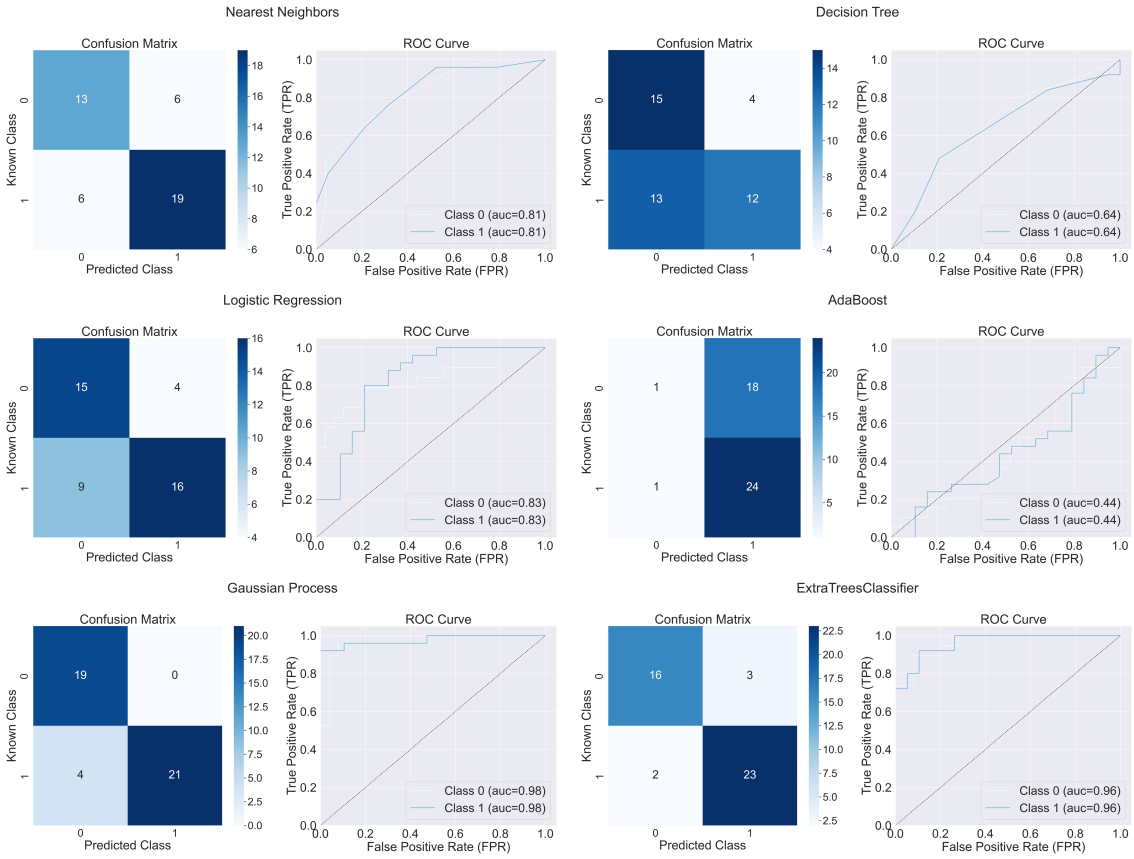


Figure B12: Classification model performance on Mordred descriptors on the test set.

B.2.3.1 Selected Mordred descriptors

For the Mordred descriptors, we calculated 1613 2D molecular descriptors from SMILES. To select the most relevant features, we conducted correlation analysis, individually testing each feature's correlation with the target property. We used Spearman correlation, which is suitable for both continuous and ordinal variables and captures linear and non-linear associations. Features with correlation coefficients of 0.2 or higher were retained. To further assess the significance of these features, we performed a two-tailed p-test using the Spearman correlation coefficient as the test statistic. We conducted this test over 5000 random sample permutations. After feature generation and selection, we applied one-hot encoding for categorical features and min-max scaling for continuous features.

When considering the entire dataset, 128 features remained to describe each transporter. These descriptors are:

1. $nBase$
2. nF
3. nX
4. $AATS1se$
5. $AATS2se$
6. $AATS1pe$
7. $AATS1are$
8. $AATS3are$
9. $AATS0i$
10. $AATS1i$

11. $AATS2i$
12. $AATS8i$
13. $ATSC0c$
14. $ATSC3c$
15. $ATSC4c$
16. $ATSC3s$
17. $ATSC0se$
18. $ATSC0pe$
19. $ATSC0are$
20. $ATSC6are$
21. $ATSC0i$
22. $ATSC1i$
23. $ATSC6i$
24. $AATSC0c$
25. $AATSC2c$
26. $AATSC3c$
27. $AATSC4c$
28. $AATSC3s$
29. $AATSC0v$
30. $AATSC6se$
31. $AATSC6pe$

32. $AATSC6are$

33. $AATSC1i$

34. $AATSC6i$

35. $MATS1c$

36. $MATS2c$

37. $MATS3c$

38. $MATS4c$

39. $MATS6s$

40. $MATS3se$

41. $MATS6se$

42. $MATS6pe$

43. $MATS3are$

44. $MATS6are$

45. $MATS6i$

46. $GATS2c$

47. $GATS3c$

48. $GATS5c$

49. $GATS8c$

50. $GATS4d$

51. $GATS5s$

52. $GATS6s$

53. $GATS3v$

54. $GATS7v$

55. $GATS3se$

56. $GATS5se$

57. $GATS6se$

58. $GATS3pe$

59. $GATS5pe$

60. $GATS6pe$

61. $GATS3are$

62. $GATS5are$

63. $GATS6are$

64. $GATS6i$

65. $BCUTc\text{-}1h$

66. $BCUTc\text{-}1l$

67. $BCUTi\text{-}1h$

68. $BCUTi\text{-}1l$

69. $SM1_Dzv$

70. $SM1_Dzare$

71. $SM1_Dzp$

72. $SM1_Dzi$

73. $nBondsD$

74. $AXp\text{-}5d$
75. $AXp\text{-}4dv$
76. $AXp\text{-}5dv$
77. $AXp\text{-}6dv$
78. $AXp\text{-}7dv$
79. Mi
80. $NdsCH$
81. $NdssC$
82. $NaaaC$
83. NsF
84. $NaaS$
85. $SdsCH$
86. $SaaCH$
87. $SaaaC$
88. $SaaNH$
89. $SssO$
90. $SaaS$
91. $AETA_alpha$
92. ETA_shape_p
93. ETA_psi_1
94. ETA_dPsi_A

95. $PEOE_VSA4$
96. $PEOE_VSA5$
97. $PEOE_VSA6$
98. $PEOE_VSA10$
99. $PEOE_VSA11$
100. $PEOE_VSA12$
101. $PEOE_VSA13$
102. SMR_VSA1
103. $EState_VSA1$
104. $EState_VSA7$
105. $VSA_EState1$
106. $VSA_EState3$
107. MID_X
108. $AMID_X$
109. $nARing$
110. $n5ARing$
111. $nAHRing$
112. $n5AHRing$
113. $nFRing$
114. $n9FRing$
115. $nFHRing$

116. $n9FHRing$

117. $nFaRing$

118. $n9FaRing$

119. $nFaHRing$

120. $n9FaHRing$

121. $GGI8$

122. $GGI9$

123. $JGI1$

124. $JGI2$

125. $JGI4$

126. $JGI8$

127. $JGI9$

128. $JGT10$

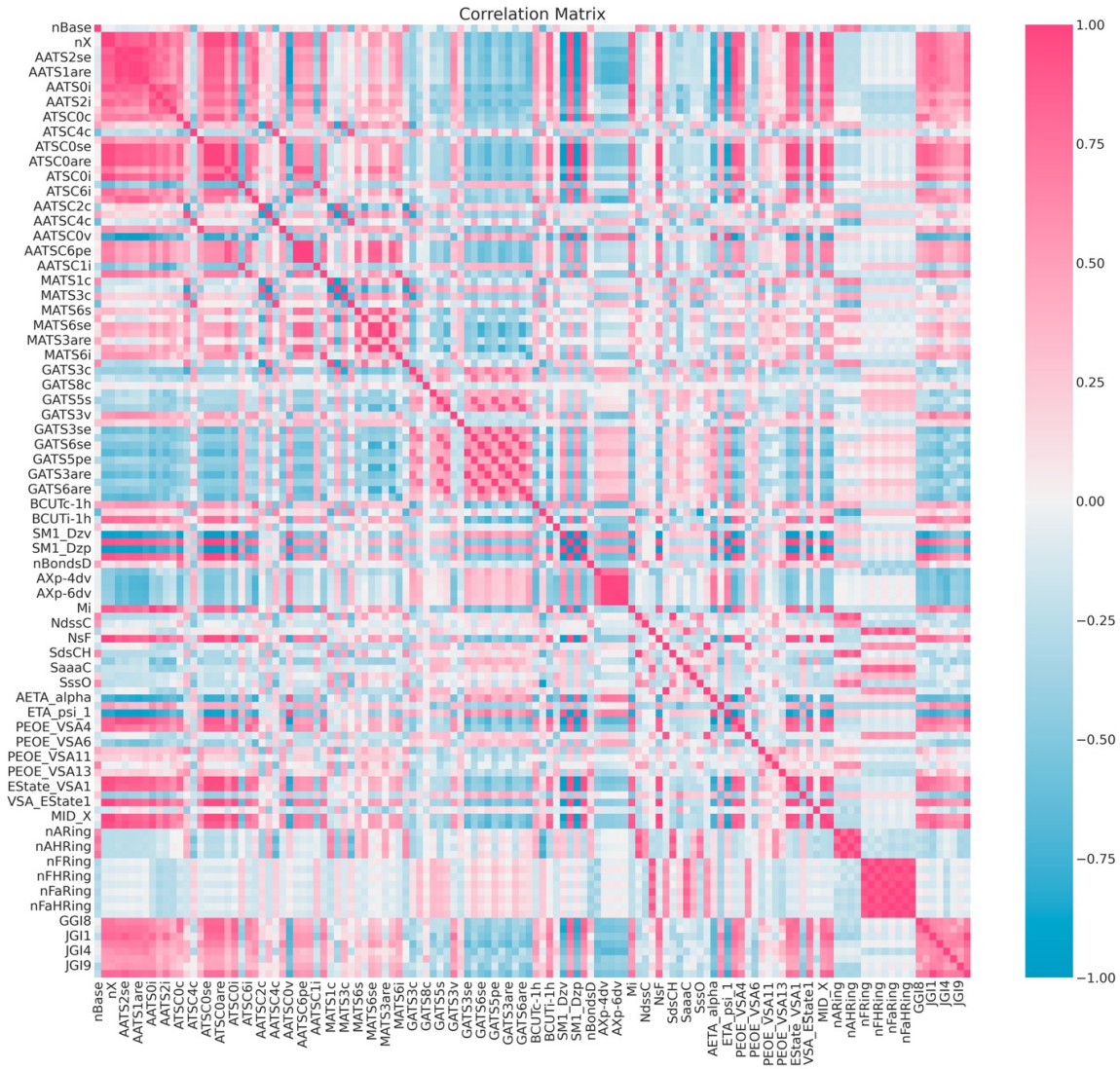


Figure B13: Correlation matrix of selected Mordred descriptors. Feature generation and selection considered the entire dataset.

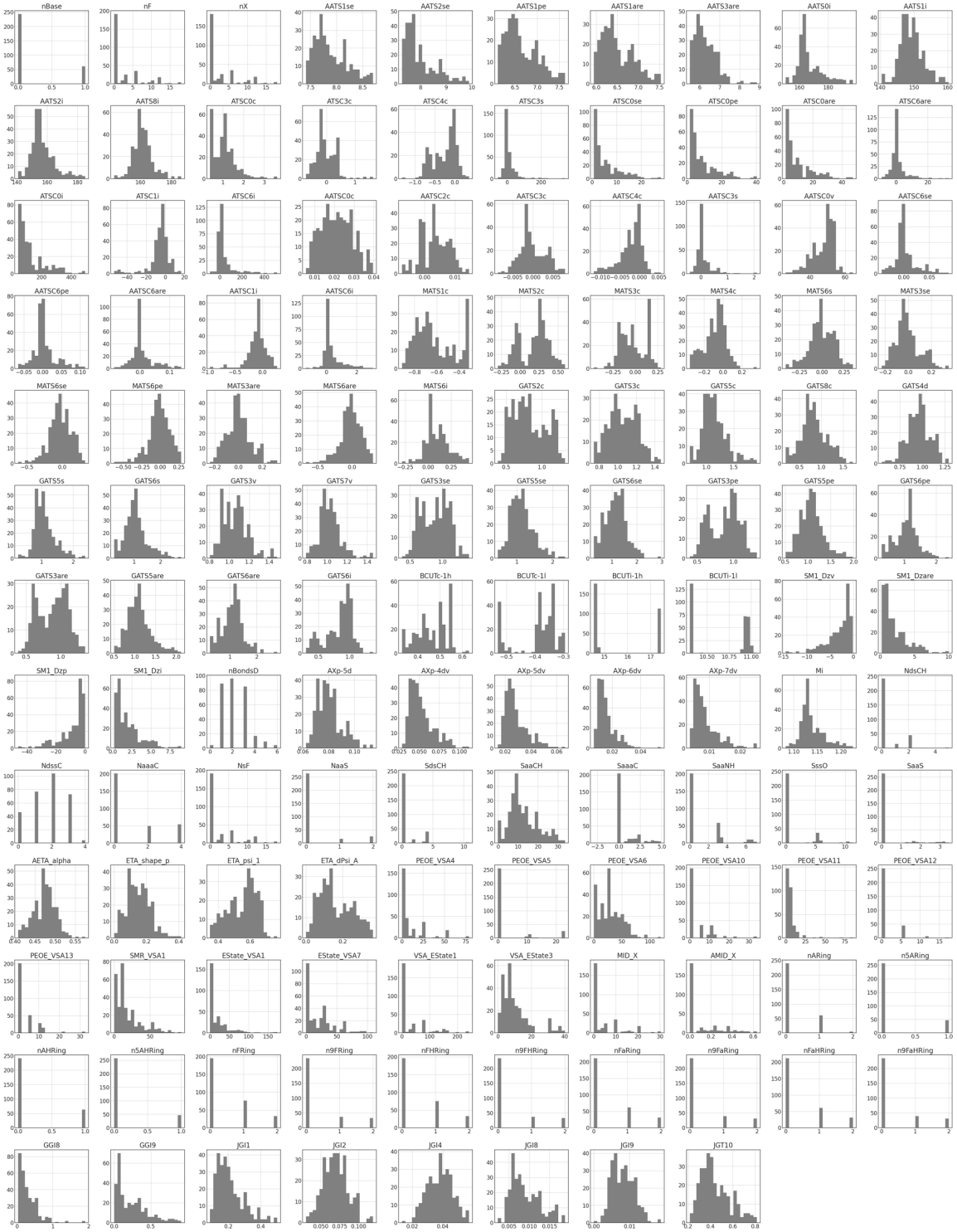


Figure B14: Distribution of selected Mordred descriptors across all classes. Feature generation and selection considered the entire dataset.



Figure B15: Distribution of selected Mordred descriptors for class 0. Feature generation and selection considered the entire dataset.

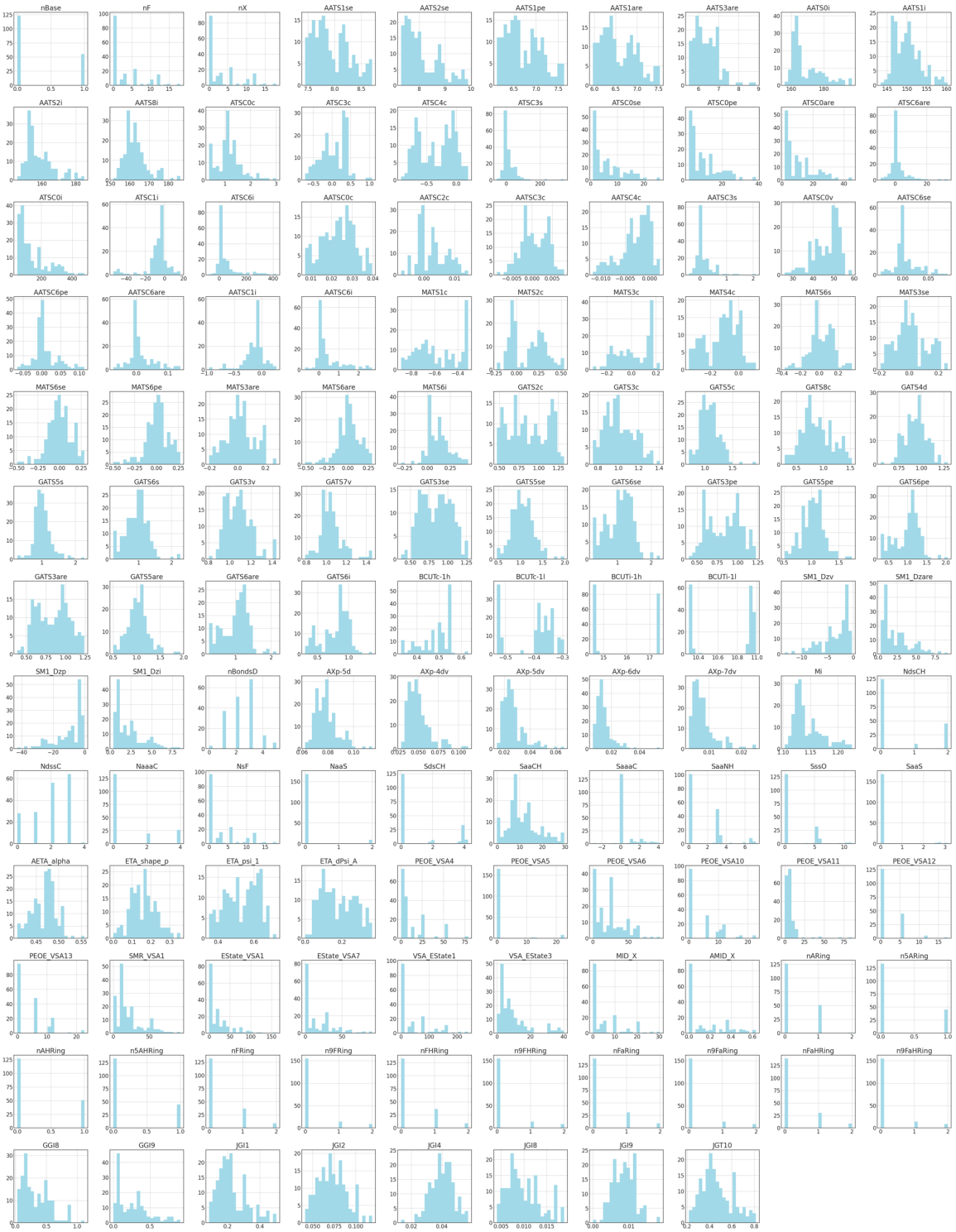


Figure B16: Distribution of selected Mordred descriptors for class 1. Feature generation and selection considered the entire dataset.

When considering the training set only, 86 features remained to describe each transporter. These descriptors are listed below:

1. $nBase$
2. $SpMax_A$
3. nS
4. $ATSC0c$
5. $ATSC2c$
6. $ATSC3c$
7. $ATSC4c$
8. $ATSC5c$
9. $ATSC6d$
10. $ATSC7d$
11. $ATSC7v$
12. $ATSC2se$
13. $ATSC2pe$
14. $ATSC2are$
15. $AATSC0c$
16. $AATSC2c$
17. $AATSC3c$
18. $AATSC4c$
19. $AATSC5c$

20. $AATSC6d$

21. $AATSC7v$

22. $AATSC2se$

23. $AATSC2p$

24. $MATS1c$

25. $MATS2c$

26. $MATS3c$

27. $MATS4c$

28. $MATS6d$

29. $MATS6s$

30. $MATS7v$

31. $MATS2se$

32. $MATS3se$

33. $MATS2pe$

34. $MATS2are$

35. $MATS3are$

36. $GATS1c$

37. $GATS2c$

38. $GATS3c$

39. $GATS5c$

40. $GATS1s$

41. $GATS2s$

42. $GATS4v$

43. $GATS2se$

44. $GATS5se$

45. $GATS2pe$

46. $GATS2are$

47. $BCUTc-1h$

48. $BCUTc-1l$

49. $BCUTd-1l$

50. $BCUTv-1h$

51. $BCUTp-1h$

52. $BCUTi-1l$

53. $nBondsD$

54. $AXp-5d$

55. $NdsCH$

56. $NdssC$

57. $NssNH$

58. $NaaNH$

59. NdO

60. $NssO$

61. $SdsCH$

62. $SdssC$

63. $SssNH$

64. $SaaNH$

65. SdO

66. $SssO$

67. $PEOE_VSA6$

68. $PEOE_VSA10$

69. $PEOE_VSA11$

70. $PEOE_VSA12$

71. $PEOE_VSA13$

72. $EState_VSA3$

73. $VSA_EState1$

74. $VSA_EState3$

75. $VSA_EState5$

76. $VSA_EState9$

77. $n5Ring$

78. $n5HRing$

79. $nARing$

80. $n5ARing$

81. $nAHRing$

82. $n5AHRing$

83. $nFRing$

84. $nFHRing$

85. $TopoPSA$

86. $SRW05$

B.3 Regression models

B.3.1 Permutation test on the test set

In this section, we explore the effects of shuffling target values in the test set on the performance of our predictive models. This technique, often referred to as a "permutation test" or "permutation analysis," provides valuable insights into the models' ability to capture meaningful relationships between input features and the target variable, and their generalization capabilities. The purpose of shuffling the target values is to assess the robustness and generalization of our predictive models under a null hypothesis scenario. By randomizing the target values, we effectively break any genuine relationships that the models might have learned during training. This allows us to evaluate how well the models perform when there is no true predictive power in the data, helping us determine whether the models are merely memorizing the training data or truly learning useful patterns.

When the target values in the test set are shuffled, we observe a significant drop in performance after shuffling (Table B4, Figure B17). This indicates that our models were indeed capturing meaningful relationships between the input features and the target variable during training. In other words, the models were learning patterns that enable accurate predictions. This drop in performance reinforces the notion that the models were not just overfitting to the training data but were learning information that generalizes to unseen examples.

Table B4: Permutation test on the test set using the regression models with Morgan fingerprints as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	-1.4	1.2	2.07	1.44	-1.4
k-Nearest Neighbors Regressor	-1.17	1.14	1.88	1.37	-1.18
Bayesian Regressor	-1.27	1.16	1.95	1.40	-1.27
Decision Tree Regressor	-1.24	1.18	1.93	1.39	-1.24
Random Forest Regressor	-1.15	1.15	1.84	1.36	-1.15
Extra Trees Regressor	-1.24	1.18	1.93	1.39	-1.24
Ada Boost Regressor	-0.92	1.07	1.65	1.28	-0.92
XGBoost Regressor	-1.24	1.18	1.93	1.39	-1.24

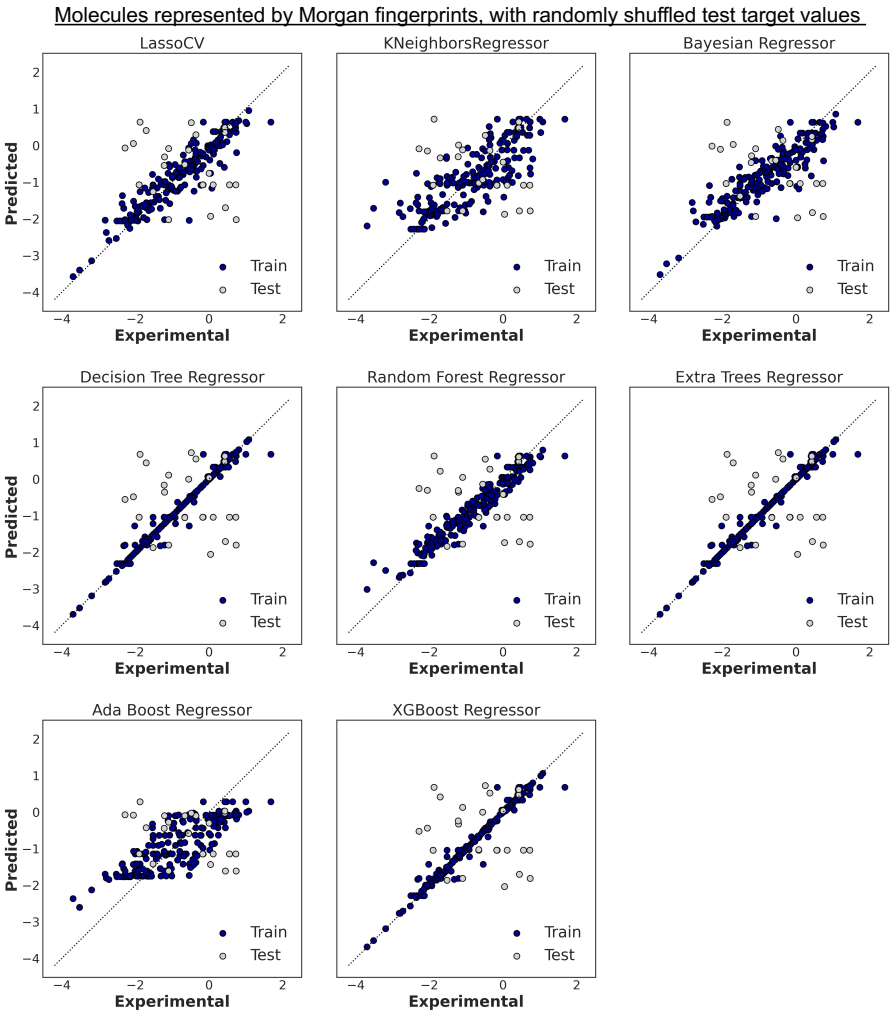


Figure B17: Regression model performance on the training (dark blue) and test set (light grey). The molecules are represented by Morgan fingerprints. The randomised target values in the test set. The predicted $log(EC_{50})$ values are on the y-axis and the experimental $log(EC_{50})$ values are on the x-axis.

B.3.2 Model performance on fingerprints

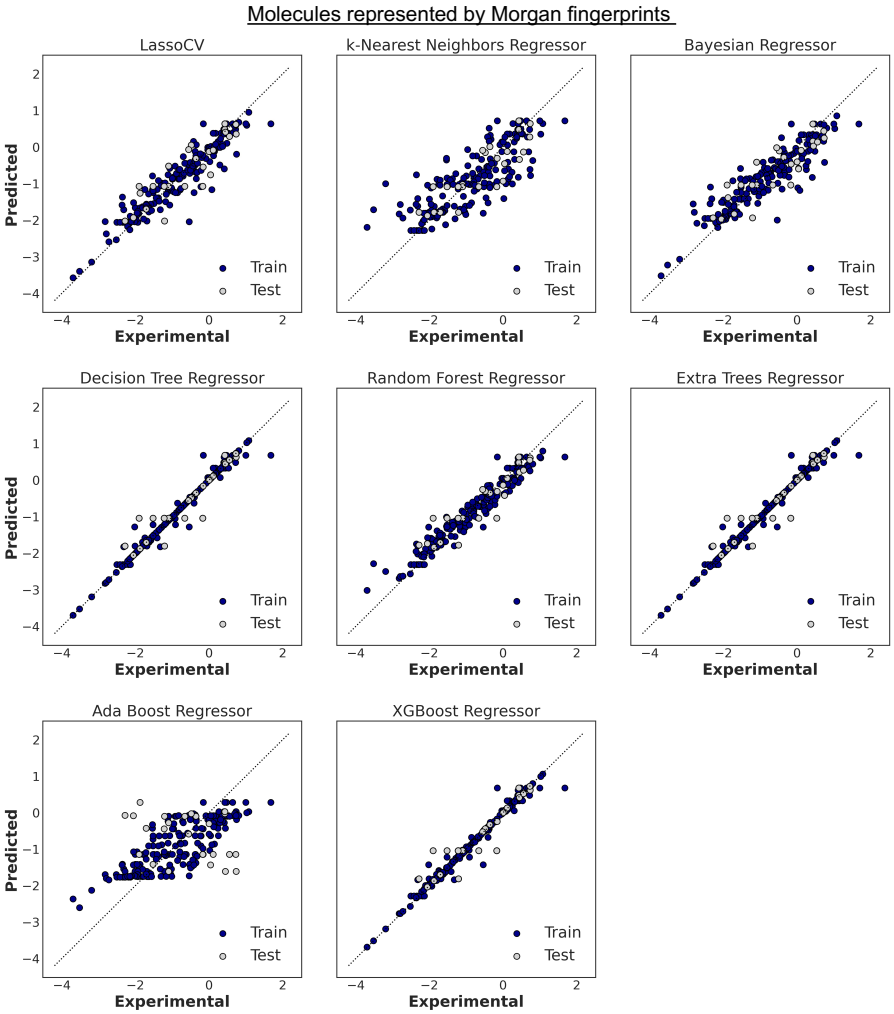


Figure B18: Regression model performance on Morgan fingerprints. Regression model performance on the training (dark blue) and test set (light grey). The predicted $log(EC_{50})$ values are on the y-axis and the experimental $log(EC_{50})$ values are on the x-axis.

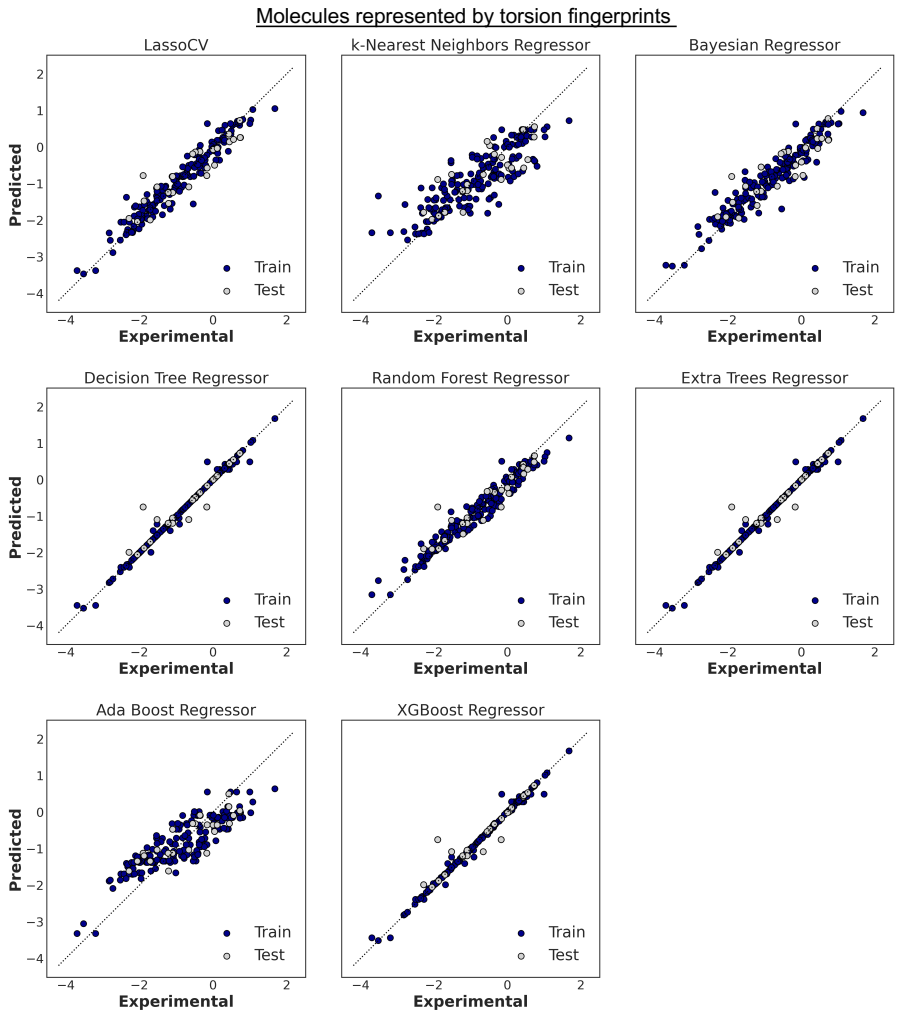


Figure B19: Regression model performance on torsion fingerprints. Regression model performance on the training (dark blue) and test set (light grey). The predicted $log(EC_{50})$ values are on the y-axis and the experimental $log(EC_{50})$ values are on the x-axis.

B.3.3 Model training and evaluation on Mordred descriptors

We first calculated 1613 2D Mordred descriptors. To select the most relevant features, we conducted correlation analysis, individually testing each feature's correlation with the target property. We used Spearman correlation, which is suitable for both continuous and ordinal variables and captures linear and non-linear associations. Features with correlation coefficients of 0.2 or higher were retained. To further assess the significance of these features, we performed a two-tailed p-test using the Pearson correlation coefficient as the test statistic. We conducted this test over 5000 random sample permutations. In total, 150 descriptors are selected. We then applied one-hot encoding for categorical features and min-max scaling for continuous features. To ensure the inclusion of chloride recognition sites, we applied one hot encoding for the presence (1) and absence (0) of urea, thiourea, tambjamine, squaramide, indole, carbazole, acridinone, perenosin, 1,3-bis(benzimidazol-2-yl)pyrimidine, and isophthalamides/dipicolineamides. Additionally, we included the key following experimental conditions as features: internal chloride concentration (in mM) and the time between the transporter to initiate chloride transport addition detergent addition to lyse the vesicles (in seconds).

On the training set, the DT regressor and ET regressor achieve exceptionally high R^2 values of 0.99 (Table B5). Similarly, the RF regressor performs remarkably well with an R^2 value of 0.94. However, the LassoCV and kNN regressor demonstrate comparatively lower R^2 values of 0.54 and 0.70, respectively.

The test set performance unveils further insights into the generalization capabilities of the regression models. The DT regressor's R^2 value falls into the negative range (Table B6) suggesting an inadequate fit to the test data and possible overfitting to the training set. The ET regressor's R^2 value also falls significantly. LassoCV and

kNN Regressor continue to exhibit modest performance, albeit with slight drops in R^2 values compared to the training set.

In conclusion, the performance of regression models employing Mordred descriptors reveals varying levels of generalization capability. The differences between the training and test set performances highlight the potential overfitting of the DT regressor and ET regressor to the training data. These results emphasize the importance of selecting models that strike a balance between capturing underlying patterns and avoiding overfitting. Moreover, it underscores the need for comprehensive model evaluation and validation to ensure reliable predictive performance across diverse datasets.

Table B5: Regression model performance on the training set with Mordred descriptors as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	0.54	0.55	0.47	0.69	0.54
k-Nearest Neighbors Regressor	0.70	0.40	0.31	0.56	0.70
Bayesian Regressor	0.73	0.42	0.27	0.52	0.73
Decision Tree Regressor	0.99	0.02	0.01	0.10	0.99
Random Forest Regressor	0.94	0.18	0.06	0.24	0.94
Extra Trees Regressor	0.99	0.02	0.01	0.10	0.99
Ada Boost Regressor	0.79	0.41	0.21	0.46	0.79
XGBoost Regressor	0.99	0.02	0.01	0.10	0.99

Table B6: Regression model performance on the test set with Mordred descriptors as input features.

	Variance	MAE	MSE	RMSE	R^2
LassoCV	0.52	0.58	0.52	0.72	0.39
k-Nearest NeighborsRegressor	0.44	0.53	0.50	0.71	0.42
Bayesian Regressor	0.55	0.51	0.39	0.62	0.54
Decision Tree Regressor	-0.59	1.22	2.39	1.55	-1.78
Random Forest Regressor	0.48	0.57	0.54	0.73	0.37
Extra Trees Regressor	0.69	0.42	0.29	0.54	0.66
Ada Boost Regressor	0.49	0.56	0.47	0.69	0.45
XGBoost Regressor	0.45	0.60	0.54	0.74	0.37

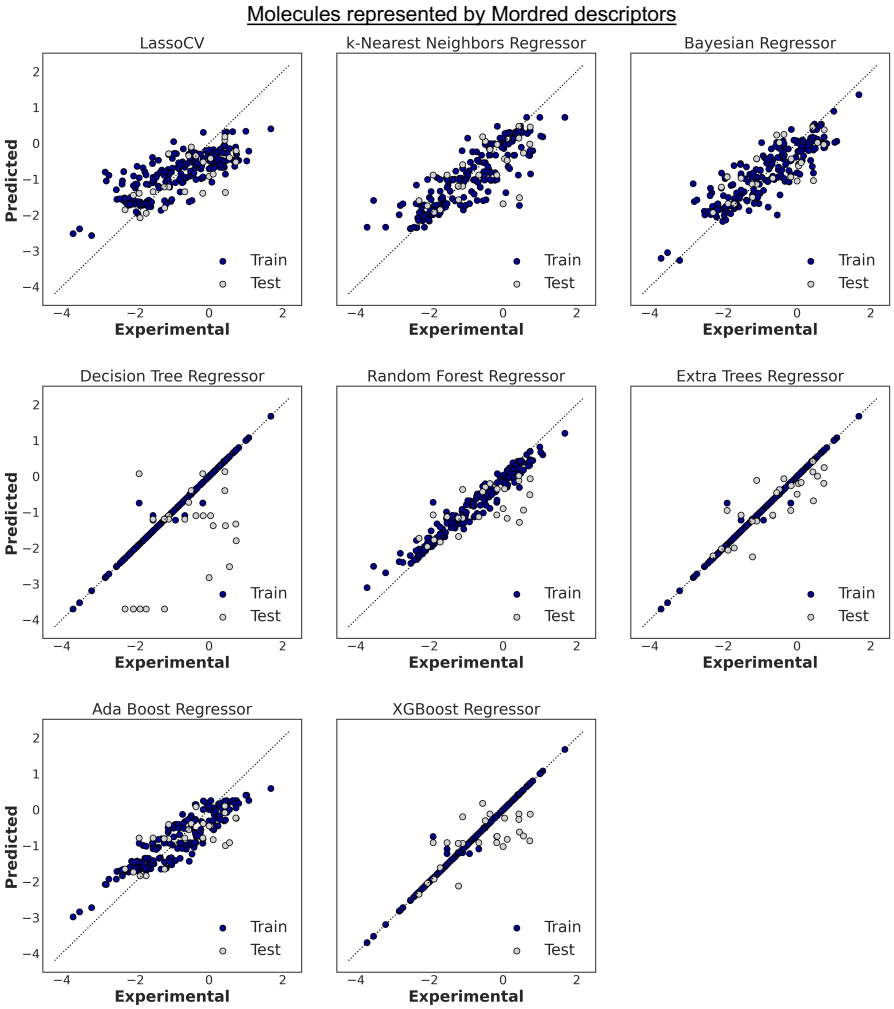


Figure B20: Regression model performance on the training (dark blue) and test set (light grey). The molecules are represented by Mordred descriptors. The predicted $log(EC_{50})$ values are on the y-axis and the experimental $log(EC_{50})$ values are on the x-axis.

B.3.3.1 Selected Mordred descriptors

The 150 selected descriptors are:

1. $nBase$
2. $SpMax_A$
3. $nBridgehead$
4. nS
5. $AATS2d$
6. $AATS3d$
7. $AATS4d$
8. $AATS5d$
9. $AATS6d$
10. $ATSC0c$
11. $ATSC3c$
12. $ATSC4c$
13. $ATSC3d$
14. $ATSC1v$
15. $ATSC2v$
16. $ATSC3v$
17. $ATSC6v$
18. $ATSC7v$
19. $ATSC2p$

20. $ATSC3p$

21. $ATSC6p$

22. $AATSC0c$

23. $AATSC3c$

24. $AATSC4c$

25. $AATSC5c$

26. $AATSC1v$

27. $AATSC2v$

28. $AATSC3v$

29. $AATSC7v$

30. $AATSC2p$

31. $AATSC3p$

32. $MATS1c$

33. $MATS2c$

34. $MATS3c$

35. $MATS4c$

36. $MATS5c$

37. $MATS3d$

38. $MATS1v$

39. $MATS2v$

40. $MATS3v$

41. $MATS7v$

42. $MATS2se$

43. $MATS3se$

44. $MATS2p$

45. $MATS3p$

46. $MATS6p$

47. $GATS1c$

48. $GATS2c$

49. $GATS3c$

50. $GATS4c$

51. $GATS3d$

52. $GATS6d$

53. $GATS2s$

54. $GATS6Z$

55. $GATS6m$

56. $GATS2v$

57. $GATS3v$

58. $GATS2se$

59. $GATS1p$

60. $GATS3p$

61. $GATS6p$

62. $BCUTc$-1h

63. $BCUTc$-1l

64. $BCUTv$-1h

65. $BCUTp$-1h

66. $BCUTi$-1l

67. $nBondsD$

68. Xch-5d

69. Xch-7d

70. Xch-5dv

71. Xch-7dv

72. Xc-5dv

73. Xpc-4dv

74. Xpc-5dv

75. AXp-1d

76. AXp-2d

77. AXp-3d

78. AXp-4d

79. AXp-5d

80. AXp-6d

81. AXp-7d

82. AXp-3dv

83. $AXp\text{-}4dv$
84. $AXp\text{-}5dv$
85. $AXp\text{-}6dv$
86. $AXp\text{-}7dv$
87. $NssCH2$
88. $NdsCH$
89. $NdssC$
90. $NssNH$
91. $NaaNH$
92. $NaaN$
93. $NaasN$
94. $NssO$
95. NdS
96. $SdsCH$
97. $SdssC$
98. $SssNH$
99. $SaaNH$
100. $SaaN$
101. $SaasN$
102. $SssO$
103. SdS

104. ETA_shape_y

105. $AETA_beta_s$

106. $PEOE_VSA9$

107. $PEOE_VSA12$

108. $PEOE_VSA13$

109. SMR_VSA3

110. SMR_VSA7

111. $SlogP_VSA6$

112. $EState_VSA3$

113. $MDEN_22$

114. $piPC4$

115. $piPC6$

116. $piPC7$

117. $piPC8$

118. $piPC9$

119. $piPC10$

120. $TpiPC10$

121. $n5Ring$

122. $nG12Ring$

123. $n5HRing$

124. $n6HRing$

125. $nG12HRing$

126. $n6aHRing$

127. $nARing$

128. $n5ARing$

129. $nG12ARing$

130. $nAHRing$

131. $n5AHRing$

132. $nG12AHRing$

133. $nFHRing$

134. $nG12FHRing$

135. $nFARing$

136. $nG12FARing$

137. $nFAHRing$

138. $nG12FAHRing$

139. $JGI4$

140. $JGI8$

141. $PetitjeanIndex$

142. $MWC06$

143. $MWC07$

144. $MWC08$

145. $MWC09$

146. $MWC10$

147. $SRW05$

148. $SRW07$

149. $SRW09$

150. $TSRW10$

Appendix C

Molecular dynamics theory

C.1 Molecular interactions

In classical molecular mechanics (MM), modelling energy is described by the potential energy function, often referred to as a force field. It describes the energetic contributions of various interactions in the system.[166] For a molecular system, it is calculated by summing up the energy terms that describe bonded interactions (U_{bonded}) and non-bonded interactions ($U_{\text{non-bonded}}$).

$$U = U_{\text{bonded}} + U_{\text{non-bonded}} \tag{C1}$$

C.1.1 Bonded interactions

Chemical bonds formed through electron hybridization require explicit treatment of electrons. In molecular mechanics (MM) methods, the bond, angle, and dihedral or torsional potentials are used to mimic the local chemical environment. These potentials approximate the electronic effects associated with bond stretching, angle bending, and torsional rotations. By incorporating these potentials, molecular simulations can

capture the influence of the local chemical environment on the behaviour of chemical bonds, angles, and dihedrals, providing a reasonable approximation of the electronic effects without explicitly considering the electronic wavefunctions or electron-electron interactions.

In MM, the total potential energy due to bonded interactions can be expressed as a sum of the contributions from bond stretching, angle bending, and torsional interactions:

$$U_{\text{bonded}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{torsion}} \tag{C2}$$

C.1.1.1 Bond stretching

The potential energy due to bond stretching (U_{bond}) is commonly described by a harmonic potential.

$$U_{\text{bond}} = \sum_{\text{bonds}} k_b (r - r_0)^2 \tag{C3}$$

where r is the current bond length and k_b is the force constant associated with the bond. k_b is derived from Hooke's Law, which states that the force required to stretch or compress a bond is proportional to the displacement from its equilibrium length. This method, however, falls short of fully describing a molecule's behaviour because the energy function doesn't converge to the dissociation energy value as bond lengths extend to infinity. For a more accurate representation of bond dissociation, the Morse potential (U_{morse}) is commonly used:

$$U_{\text{morse}} = D_e \left[1 - e^{-\alpha(r - r_0)}\right]^2 \tag{C4}$$

where D_e is the dissociation energy of the bond and α is a parameter linked to the force constant:

$$\alpha = \sqrt{\frac{k_b}{2D_e}} \tag{C5}$$

However, the Morse potential is often not used as it introduces additional parameters that increase computational cost.

C.1.1.2 Bond bending

The potential energy due to angle bending is also described by the harmonic potential.

$$U_{\text{angle}} = \sum_{\text{angles}} k_\theta (\theta - \theta_0)^2 \tag{C6}$$

θ is the current bond angle. θ_0 is the equilibrium bond angle. k_θ is the force constant associated with the angle that controls the stiffness of the angle spring.

C.1.1.3 Torsional interaction

Torsional interactions account for the rotation around a bond. In the AMBER force field, the torsional potential is typically described by a n-fold periodic function.

$$U_{\text{torsion}} = \sum_{\text{torsions}} k_\phi \left[1 + \cos(n\phi - \phi_0)\right] \tag{C7}$$

ϕ is the torsional angle. n represents the periodicity of the torsion. k_ϕ is the amplitude or barrier height of the torsion potential. ϕ_0 is the phase angle. n represents the periodicity of the torsional barrier, reflecting the intrinsic symmetry in the dihedral angle for the four bonded atoms.

C.1.2 Non-bonded interactions

Non-bonded interactions in molecular systems arise from interactions between atoms or molecules that are not directly bonded to each other. These interactions include

van der Waals forces and electrostatic interactions. In molecular simulations, the total potential energy due to non-bonded interactions is given by the sum of the Lennard-Jones potential (U_{LJ}) and the electrostatic potential (U_{elec}):

$$U_{\text{non-bonded}} = U_{\mathrm{LJ}} + U_{\mathrm{elec}} \tag{C8}$$

C.1.2.1 Van der Waals interaction

Van der Waals interactions can be separated into two phases based on their nature. The repulsive force is due to Pauli repulsion and the attractive force is due to the dipole-induced instantaneous dipole moment. The repulsive force arises from the Pauli exclusion principle, which prevents atoms from occupying the same space simultaneously. As atoms get closer, their electron clouds start to overlap, resulting in repulsion. This repulsion force decays exponentially with respect to the distance between the atoms. The attractive force arises from the temporary fluctuations in electron distribution, resulting in instantaneous dipole moments. These temporary dipoles induce opposite dipoles in nearby atoms, leading to attractive interactions. The attractive force decays with respect to the distance between the atoms and can be fitted with a power of -6.

The Lennard-Jones (LJ) potential is commonly used to approximate van der Waals interactions.

$$U_{\mathrm{LJ}} = \sum_{i<j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{C9}$$

where r_{ij} represents the distance between atoms i and j, ϵ_{ij} is the depth of the potential well, and σ_{ij} is the distance at which the LJ potential is 0. The sum is taken over all pairs of non-bonded atoms i and j.

In this equation, the repulsive force term, with a σ^{12} dependence, is computed as the square of the attractive force term, with a σ^{6} dependence. This is an approximation

used to improve computational efficiency. It is based on the observation that the repulsion dominates at short distances, while the attraction dominates at longer distances. By squaring the attractive term, the computation can be simplified without significantly affecting the overall behaviour of the system. This approximation allows for faster calculations while still capturing the overall characteristics of van der Waals interactions.

Alternative functions, such as the Buckingham and Morse potentials, can also represent van der Waals interactions. These alternatives diverge from the LJ potential primarily due to two reasons: (1) The equation used to describe repulsive forces often incorporates exponential functions, which are slower to compute. (2) The required number of parameters is higher. While the LJ potential relies on two parameters, the other potentials involve three. Thus, the LJ potential is faster to compute.

C.1.2.2 Electrostatic interaction

Electrostatic interactions arise from the Coulombic interactions between charged atoms or molecules. These interactions are governed by Coulomb's law, which states that the force between two charged particles is proportional to the product of their charges and inversely proportional to the square of the distance between them. The potential energy due to electrostatic interactions can be expressed as:

$$U_{\text{elec}} = \sum_{i<j} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} \tag{C10}$$

where ϵ_0 is the vacuum permittivity and q_i and q_j represent the partial charges of atoms i and j. These partial charges are assigned to the atoms to represent their relative electronegativities and their contribution to the overall charge distribution. As the accuracy of the electrostatic interactions is inherently limited by the charge models used and the level of theory employed for charge derivation, accurate partial

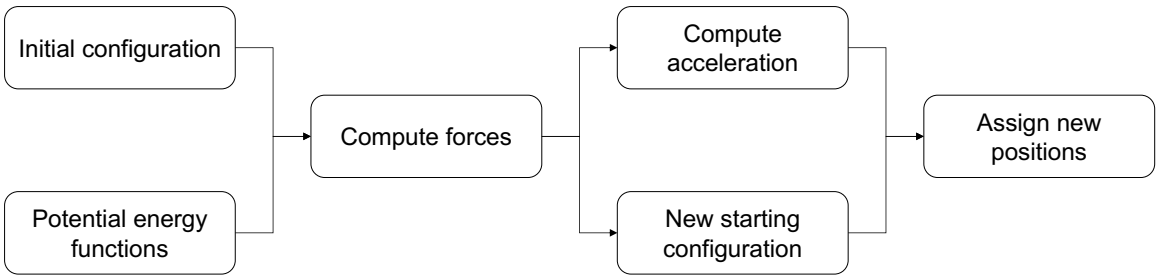


Figure C1: Schematic representation of the MD algorithm.

charge assignment is crucial for capturing the correct electrostatic behaviour of molecules.

C.2 Molecular dynamics algorithm

A schematic representation of the molecular dynamics (MD) algorithm is denoted in Figure C1

C.2.1 Equations of motion for molecular systems

In MD, the motion of interacting particles in a molecular system is governed by Newton's laws of motion. For a molecular system, where the particles are atoms, the equations of motion for each atom in the system can be expressed as:

$$m_i \frac{d^2 \mathbf{r}_i}{dt^2} = \mathbf{F}_i \tag{C11}$$

where m_i is the mass of atom i, $\mathbf{r}_i$ is its position vector, and $\mathbf{F}_i$ is the net force acting on atom i, which is obtained from the gradient of the potential energy function with respect to the atom's position.

$$\mathbf{F}_i = -\nabla U(\mathbf{r}_i) \tag{C12}$$

where $\mathbf{r}_i$ represents the position vector of particle i, and $U_i(\mathbf{r}_i)$ is the potential energy associated with particle i, and $U(\mathbf{r}_i)$ is the potential energy associated with the interactions of atom i.

C.2.2 Integration of equations of motion

In many-body systems, such as molecular systems, analytical solutions to the Lagrangian equations of motion are often unavailable. Numerical integration schemes like the leapfrog methods are used. The leapfrog integrator operates by splitting the time evolution into discrete time steps.[7] The positions and velocities of the atoms are updated alternately at half-step intervals, resulting in a leapfrog-like pattern (Figure C2).

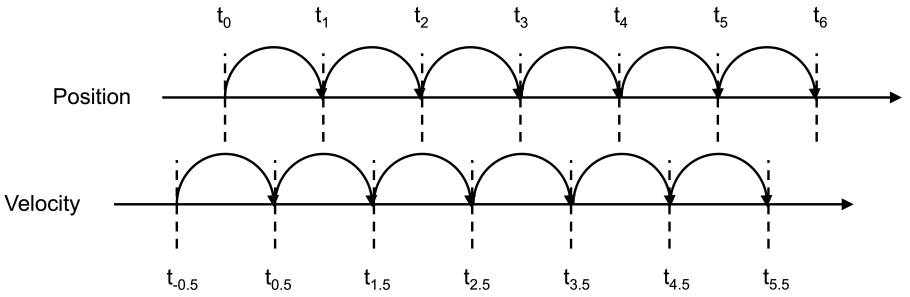


Figure C2: The leapfrog integration method. The positions and velocities of the atoms leap over each other at half-step intervals like frogs.

The integration steps can be summarized as:

1. Initial positions ($\mathbf{r}$) and velocities ($\mathbf{v}$) are assigned to the atoms in the system. Initial forces ($\mathbf{F}$) on each atom based on the potential energy function.

2. The velocities of the atoms are updated by half a time step using the current forces.

$$\mathbf{v}(t + \frac{\Delta t}{2}) = \mathbf{v}(t) + \frac{\mathbf{F}(t)}{2m}\Delta t \tag{C13}$$

where Δt represents the time step size, and m is the mass of the atom.

3. The positions of the atoms are updated by a full time step using the updated velocities.

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t + \frac{\Delta t}{2})\Delta t \tag{C14}$$

4. The forces are calculated on each atom based on the updated positions. The velocities of the atoms are then updated by another half step using the updated forces:

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t + \frac{\Delta t}{2}) + \frac{\mathbf{F}(t + \Delta t)}{2m}\Delta t \tag{C15}$$

C.3 Timestep and constrains

The timestep (Δt) in MD simulations is the discrete interval at which the positions, velocities, and forces of atoms are updated. It determines the temporal resolution of the simulation. It is desirable to choose a longer time step, which would allow longer simulations with fewer computational resources. However, Δt is limited by the fastest motion in the simulated system. For a molecular system where all atoms are treated explicitly, the fastest motion typically corresponds to the bond vibrations involving hydrogen atoms. Because the typical vibration frequency of these hydrogen bonds is around 10 fs, a timestep that is one-tenth of this vibration period or 1 fs is often necessary to achieve stable integration.

However, the average bond length between hydrogen and non-hydrogen atoms usually remains relatively constant throughout the simulations, regardless of the surrounding environment. Therefore, these bonds can be constrained using the simulations to eliminate these high-frequency vibrations, and a timestep of 2 fs can be employed while still achieving stable integration.

C.4 Periodic boundary condition

A molecule is typically solvated by solvent molecules or inserted into lipid membranes, leaving some atoms facing the vacuum. To prevent the outer atoms from boiling off into space and minimizing surface effects, periodic boundary conditions (PBCs) are typically employed. PBCs extend the simulation system periodically in all three directions, creating a pseudo-infinite system and effectively removing the vacuum.

The minimum image convention (MIC) is a technique used within the context of PBCs to calculate distances and interactions between particles in a periodic system. The MIC states that when calculating distances or interactions between particles, one should consider the image that is the closest in space to the reference particle. By applying the MIC, potential artefacts such as long-range interactions or erroneous distances caused by considering all images are minimized.

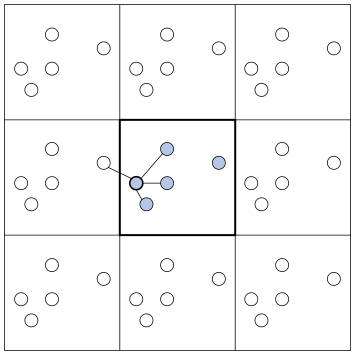


Figure C3: Periodic boundary conditions with minimum image convention. The simulated box shown in the middle is replicated in all directions. Non-bonded interactions are calculated only for the closest atoms or their images.

C.5 Ensembles

Ensembles are distinct sets of conditions under which MD simulations are performed. These ensembles capture specific thermodynamic properties. Ensembles are crucial because they provide a structured framework for studying and understanding complex molecular systems under varying constraints. Common ensembles include:

- **NVE ensemble (microcanonical ensemble):** It represents an isolated system with constant particle number, volume, and total energy. Energy fluctuations are allowed within the system, making it suitable for studying energy conservation and internal dynamics.
- **NVT ensemble (canonical ensemble):** Number of particles, volume and temperature are conserved throughout the simulation. The system is allowed to exchange heat with its surroundings while keeping particle number and volume constant.
- **NPT ensemble (isothermal-isobaric ensemble):** Number of particles, pressure and temperature are conserved throughout the simulation. In this ensemble, both heat exchange and volume adjustments are allowed, maintaining a constant temperature and pressure.
- **μVT ensemble (grand canonical ensemble):** Chemical potential, volume and temperature are conserved throughout the simulation. Unlike the previous ensembles, the μVT ensemble allows for fluctuations in the number of particles in addition to heat exchange and volume adjustments. It models open systems in equilibrium with a particle reservoir.

C.6 Thermostats

The system's temperature is controlled with a thermostat. They work by adjusting the velocities of particles in the simulation to achieve and maintain the desired temperature. Widely used thermostats include the Anderson thermostat,[177] Berendsen thermostat,[178] Nosé-Hoover thermostat,[179,180] V-rescale thermostat.[83]

The Anderson thermostat, a stochastic collision method, works by assigning new velocities to randomly selected subsets of atoms from the Maxwell-Boltzmann distribution

for the desired temperature.[177] The Anderson thermostat can correctly sample the canonical ensemble, but the velocity randomization can decorrelate velocities, and the system's dynamics are not physical.

The Berendsen[178] stems from the velocity rescaling method,[13] which rescales the velocities at each time step to keep a fixed temperature. The velocity rescaling method does not allow temperature fluctuation, which is present in the canonical ensemble. The Berendsen thermostat introduces a weak coupling method to an external heat bath and rescales the velocities periodically.[178] Although the Berendsen thermostat allows temperature fluctuations, it still suffers from the same problem as the velocity rescaling method and cannot be mapped onto a specific thermodynamic ensemble. Thus, the Berendsen thermostat is not suitable for production runs, but useful for equilibration due to its high efficiency.

The Nosé-Hoover thermostat is an extended system thermostat, which integrates the heat bath as part of the system.[179,180] The Nosé-Hoover thermostat regulates temperature without relying on random numbers, so it does not impair correlated motions and provides a more accurate description of kinetics and dynamics. It samples the correct canonical ensemble, but it may lead to temperature oscillations.

The V-rescale thermostat performs temperature coupling using velocity rescaling with a stochastic term.[83] Similar to the Berendsen thermostat, there are no temperature oscillations with the V-rescale thermostat, but the stochastic term ensures that a correct canonical ensemble is sampled.[83] Thus, the V-rescale thermostat can be effectively used for both equilibration and NVT production runs.[14]

C.7 Barostats

The system's volume fluctuations in response to changes in pressure are controlled with a barostat. Widely used barostats are the Berendsen barostat[178] and the Parrinello Rahman barostat.[84] The Berendsen barostat, similar to the Berendsen thermostat, works by weakly coupling the pressure of the system to a "pressure bath" and rescaling the volume periodically.[178] The Berendsen barostat does not correctly sample the isobaric-isothermal ensemble, but it is useful for equilibration due to its high efficiency. The Parrinello Rahman barostat, similar to the Nosé-Hoover thermostat, is an extended system method and can correctly sample the isobaric-isothermal ensemble.[84]

For membrane simulations in the isobaric-isothermal ensemble, pressure control is often conducted in a semi-isotropic scheme ($NP_zP_{xy}T$), where $P_{xy} = P_x = P_y$ is pressure in the lateral or membrane in-plane directions and P_z is the pressure in the transverse direction. This ensures consistent scaling in the bilayer plane, independent of the out-of-plane dimension.

Appendix bibliography

[1] Aidan Kerckhoffs, Zonghua Bo, Samuel E. Penty, Fernanda Duarte, and Matthew J. Langton. Red-shifted tetra-ortho-halo-azobenzenes for photo-regulated transmembrane anion transport. *Org. Biomol. Chem.*, 19:9058–9067, 2021.

[2] R. Ian Storer, Caroline Aciro, and Lyn H. Jones. Squaramides: physical properties, synthesis and applications. *Chem. Soc. Rev.*, 40:2330–2346, 2011.

[3] J.R. Quinlan. Decision trees and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics*, 20(2):339–346, 1990.

[4] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[5] David R Cox. The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232, 1958.

[6] Frank Jensen. *Introduction to Computational Chemistry.* Wiley, 2017.

[7] Quiet high-resolution computer models of a plasma. *Journal of Computational Physics*, 14(2):148–158, 1974.

[8] Hans C Andersen. Molecular dynamics simulations at constant pressure and/or temperature. *The Journal of chemical physics*, 72(4):2384–2393, 1980.

[9] Herman JC Berendsen, JPM van Postma, Wilfred F Van Gunsteren, ARHJ DiNola, and Jan R Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984.

[10] Shūichi Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Molecular physics*, 52(2):255–268, 1984.

[11] William G Hoover. Canonical dynamics: Equilibrium phase-space distributions. *Physical review A*, 31(3):1695, 1985.

[12] Giovanni Bussi, Davide Donadio, and Michele Parrinello. Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, 126(1), 01 2007.

[13] L V Woodcock. Isothermal molecular dynamics calculations for liquid salts. *Chemical Physics Letters*, 10(3):257–261, 1971.

[14] Qia Ke, Xiaoting Gong, Shouwei Liao, Chongxiong Duan, and Libo Li. Effects of thermostats/barostats on physical properties of liquids by molecular dynamics simulations. *Journal of Molecular Liquids*, 365:120116, 2022.

[15] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics*, 52(12):7182–7190, 12 1981.