Combining Bio- and Cheminformatics for Small Data Sets – Microcystins as a Use Case

Doctoral thesis for obtaining the

academic degree

Doctor of Natural Sciences (Dr. rer. nat.)

submitted by

Sabrina Jaeger-Honz

at the

Faculty of Sciences

Department of Computer and Information Science

Konstanz, 2023

Date of the oral examination: 18.01.2024
1. Reviewer: Prof. Dr. Falk Schreiber (University of Konstanz)
2. Reviewer: Prof. Dr. Daniel R. Dietrich (University of Konstanz)
3. Reviewer: Prof. Dr. Ina Koch (Goethe University Frankfurt a. Main)

Acknowledgments

Mein besonderer Dank gilt zunächst meinem Doktorvater Prof. Dr. Falk Schreiber, für die Betreuung und Möglichkeit diese Arbeit anzufertigen. Für seine Ideengebung, Unterstützung sowie den konstruktiven und bereichernden Austausch möchte ich mich herzlich bedanken. Ich danke weiter Prof. Dr. Daniel Dietrich für die wissenschaftliche Betreuung und übernahme des Zweitgutachtens. Ebenfalls möchte ich mich bei Prof. Dr. Christine Peter für die wissenschaftliche Betreuung bedanken. Für die kritischen Auseinandersetzungen mit meinem Thema und das Korrekturlesen in letzter Minute möchte ich mich besonders bei Dr. Karsten Klein bedanken. Ich habe unsere wissenschaftlichen Dialoge stets als Ermutigung empfunden und habe die Betreuung sehr geschätzt. Weiterer Dank gilt der ganzen AG Schreiber für die anregenden Anmerkungen, kritischen Betrachtungen und die unterstützende Zusammenarbeit zur Vollendung meiner Arbeit. Mein Dank an die ganze AG gilt aber auch für die nichtfachliche und zwischenmenschliche Unterstützung in den letzten Jahren, welche die Zeit und die Arbeit kurzweilig gemacht hat. Darüber hinaus möchte ich allen Kooperationspartnern und Ko-Autoren der Paper danken, die alle ihren Teil dazu beigetragen haben, diese Werke zu verbessern. Herzlichen Dank vor allem an Steff, Regina und Marius für die experimentellen Daten zu meinen Simulationen. Vielen Dank vor allem an Regina, für die schönen Kaffeepausen, deine Unterstützung, dein offenes Ohr und die Versorgung mit Süßigkeiten. Vielen Dank an meine Freunde Laura und Leon, für die fachliche und moralische Unterstützung und das Korrekturlesen der Arbeit in den letzten Minuten. Zuletzt möchte ich den wichtigsten Menschen in meinem Leben vielen Dank sagen, meiner Familie. Vielen Dank für die bedingungslose Unterstützung, während des Studiums und bei der Doktorarbeit. Ohne Euch und Eure grenzenlose Unterstützung und Motivation wäre das alles nicht möglich gewesen.

Zusammenfassung

In den Forschungsbereichen der Bio- und Chemieinformatik stellen kleine Datensätze eine besondere Herausforderung dar. Aufgrund einer nur begrenzten Anzahl an verfügbaren Datenpunkten, ist es schwierig computergestützte Modelle und Methoden zu entwickeln, um Eigenschaften wie Aktivität oder Toxizität vorherzusagen. Das Interesse an der Aktivität eines Moleküls geht oft auch mit einem Interesse an seiner Interaktion mit einem Zielprotein einher, um molekulare Prozesse wie die Protein-Ligand Bindung zu verstehen. In diesem Forschungsbereich überschneiden sich die Gebiete der Bio- und Chemieinformatik, da hier keine klare Trennung zwischen kleinen Molekülen und Makromolekülen getroffen werden kann. Zusätzlich gibt es in der Chemieinformatik eine weitere Klasse von Molekülen, sogenannte Makrozyklen. Diese wurden in der Vergangenheit stark vernachlässigt, da sie teuer und schwierig herzustellen waren. Makrozyklen bestehen aus mehr als 12 Atomen in ihrer Ringstruktur und wurden mit einem chemischen Raum assoziiert, der als wenig nützlich für den Bereich der Wirkstoffforschung gesehen wurde. Dadurch sind Datensätze für Makrozyklen oft sehr klein. Aktuell sind Makrozyklen ein aktives Forschungsfeld im Bereich der Chemieinformatik und viele Studien fokussieren sich darauf, ihre Struktur, Konformation und physikalisch-chemischen Eigenschaften zu verstehen und vorherzusagen, vor allem im Bereich der Wirkstoffforschung und Toxikologie. Methoden wie molekulares docking, maschinelles Lernen und Molekulardynamik (MD) Simulationen können dabei verwendet werden, um ihre Aktivität oder Toxizität gegenüber großen Molekülen, wie zum Beispiel Proteinen, vorherzusagen, zu untersuchen und den Mechanismus zu verstehen.

Diese Dissertation stellt verschiedene Ansätze und neue Methoden vor, um die Schwierigkeiten beim Analysieren von kleinen Datensätzen zu überwinden. Dabei liegt der besondere Fokus auf makrozyklischen Strukturen, wobei die Toxizität von Microcystin (MC) Kongeneren als Anwendungsfall dient. MC Kongenere sind eine Klasse von strukturell ähnlichen Makrozyklen, von denen nur wenige getestet und analysiert wurden. Im Verlauf dieser Dissertation wurden zwei Modelle entwickelt, um die Toxizität von MC Kongeneren auf Ser/Thr-Protein Phosphatasen (PPP) 1, PPP2A und PPP5 vorherzusagen. Ein Vorhersagemodell des maschinellen Lernens wurde genutzt, um MC Kongenere anhand ihrer Inhibitionskapazität gegen PPPs in drei Klassen (toxisch, wenig toxisch, nicht toxisch) einzuordnen. Für die Darstellung von MC Kongeneren und PPPs als numerische Vektoren wurden Ansätze aus dem Bereich der Computerlinguistik genutzt. Da die meisten MC Kongenere toxisch sind, sind die Klassen ungleichmäßig verteilt. Deshalb wurde SMOTE („synthetic minority oversampling technique“) genutzt, um mehr Datenpunkte in den Bereichen wenig und nicht toxisch zu erzeugen. Das Resultat war ein Vorhersagemodell mit Konsensprinzip und 80–90 % korrekter Vorhersagen. Dieses Vorhersagemodell ist eine sogenannte „Black Box“ und die Erklärbarkeit, die für die Risikobewertung und Toxikologie wichtig ist, wird damit vernachlässigt. Deshalb wurde ein zweites Vorhersagemodell entwickelt, das auf einer mathematischen Optimierung beruht, dem sogenannten (α, β)-k-Feature Set-Problem. Dabei werden die Dimensionen des Datensatzes signifikant reduziert, basierend auf klasseninternen Ähnlichkeiten und klassenübergreifenden Unterschieden (hier die Anzahl der Dimensionen des Extended Connectivity Fingerprints). Diese werden zu relevanten Merkmalen zusammengefasst, aus denen boolesche Regeln abgeleitet werden konnten. Dieser Ansatz resultierte in Signaturen von Toxizität, die zur Vorhersage genutzt wurden und mit experimentellen Ergebnissen aus der Literatur verifiziert werden konnten. Zusätzlich wurden MD Simulationen von MC Kongeneren aus verschiedenen PPP1-Inhibitionsklassen durchgeführt, um die Interaktionen und Konformationen von MC-Kongeneren und PPP1 zu untersuchen, da neuere Studien ergeben haben, dass die drei-dimensionale Strukturähnlichkeit auss-

chlaggebender für eine ähnliche Bioaktivität ist, als die zweidimensionale Strukturähnlichkeit. Es wurden gemeinsame und unterschiedliche Konformationen für die MC Kongenere identifiziert, die teilweise auch für verschiedene Toxizitätsklassen identisch zu den toxischen waren. Zusätzlich wurden die MD Simulationsdatensätze genutzt, um eine Methode zur Erstellung von „Interaktions-Fingerprints“ (IFP) abzuleiten. IFPs sind numerische Vektoren, die eine Interaktion zwischen zwei Molekülen (hier MC Kongenere und PPP1) als vorhanden (1) oder nicht vorhanden (0) kodieren. Im Rahmen dieser Dissertation wurde eine Methode entwickelt, um IFP automatisiert aus einer MD Simulation für jeden Zeitschritt zu berechnen und abzuleiten. Weil das Standardverfahren das Aufaddieren von allen IFP auf einen sogenannten aggregierten IFP ist, wurde eine Methode entwickelt, um IFP im Detail zu analysieren und ihre Dynamik und Veränderungen über die Zeit zu berücksichtigen. Mit diesem Verfahren können individuelle Unterschiede und Gemeinsamkeiten in den Interaktionen der einzelnen MC Kongenere mit PPP1 analysiert werden. Die hier entwickelte Analyse für IFP ist nicht nur auf dieses System oder diese Makrozyklen anwendbar, sondern auf alle Systeme, die Interaktionen zwischen Molekülen untersuchen. Deshalb ist diese Methode ein wertvoller Beitrag zum Forschungsgebiet der Bio- und Chemieinformatik und ermöglicht eine einfachere und schnellere Analyse von Interaktionen zwischen Molekülen in MD Simulationen.

In dieser Dissertation wurden neue Methoden und Ansätze vorgestellt, die speziell auf kleinen Datensätzen angewendet werden können. Am Beispiel der Toxizität von MC Kongeneren wurde gezeigt, dass biologische Erkenntnisse aus computergestützten Methoden abgeleitet werden können. Da diese Methoden auf allgemeinen Darstellungen von Molekülen basieren, sind die hier vorgestellten Methoden für eine Vielzahl von Anwendungsfällen in der Bio- und Chemieinformatik nutzbar.

Abstract

In the research fields of bio- and cheminformatics, small data sets are a particular challenge. Due to the limited number of data points available, it is often difficult to draw conclusions or to develop computational models and methods to predict endpoints of interest, e. g. activity or toxicity. Methods such as molecular docking, machine learning, or molecular dynamics (MD) simulation can be used to predict or analyse potential activity or toxicity of small data sets, and help to accelerate and understand mechanisms of bioactivity. The interest in activity of molecules and their targets often involves understanding biomolecular processes, e. g. protein-ligand binding. For investigations of protein-ligand binding, which is the major interest of this thesis, the research fields of bio- and cheminformatics naturally overlap.

Macrocycles are a class of molecules that have twelve or more atoms in their central ring structure. They have been highly neglected in the past, with most research focusing on small molecules or acyclic structures, because macrocycles are expensive and time-consuming to synthesise and test experimentally. Macrocycles were not considered useful for drug discovery, due to the chemical space associated with them. For these reasons, often only small data sets of macrocyclic structures are available. Today, macrocycles are actively researched in cheminformatics, focusing on the study of their structure, conformation, and physicochemical properties in the context of toxicology and drug discovery.

This thesis presents interconnected approaches aiming to overcome the difficulties related with studying toxicity on small data sets, particularly in the context of macrocycles. Microcystin (MC) congener toxicity is utilised as a use case. MC congeners are a class of structurally similar macrocycles, of which only a few have been tested and analysed yet.

In the course of this thesis, two prediction models for the toxicity/inhibition of MC congeners on ser/thr-protein phosphatases (PPP) 1, PPP2A and PPP5 are developed. The first model built is a machine learning model used to classify MC congeners inhibition capacity towards PPPs into three toxicity classes (toxic, less toxic, non-toxic). Feature representations from natural language processing are used to represent MC congeners and PPPs as numerical vectors. The data set is imbalanced towards the toxic class, since most MC congeners are highly toxic. To increase the number of data points for the non-toxic class, synthetic minority oversampling is applied. This results in a consensus model with 80-90 % correct predictions. Nevertheless, this machine learning model is a black box and explainability is preferred in risk assessment for estimating toxicity. For this reason, the second model built is based on a mathematical optimisation method, the so-called (α, β)-k-Feature Set-Problem, to reduce the dimensionality of the data set based on inter-class similarities and intra-class differences. This significantly reduces the number of features (i.e., dimensions in the extended connectivity fingerprint) while retaining meaningful features. Finally, Boolean rules are derived from the feature set to obtain signatures explaining the toxicity of MC congeners. Our findings are verified by experimental data which is available in the literature.

The machine learning approaches are build based on two-dimensional similarity approaches. Recent studies show that three-dimensional conformational similarity could explain similar bioactivity better than two-dimensional similarity. Therefore, MD simulations are run on different MC congeners with a wide range of PPP1 inhibition capacity to investigate the interactions and conformations of MC congeners. Our results indicate that the toxic MC congener MC-LR has two conformational backbone clusters in solvent and that the other toxic MC congener MC-LF has a similar conformational structure. In contrast, the less toxic and non-toxic MC congeners differ from the toxic MC congeners in their backbone conformation and

have less well defined conformational clusters. A second MD simulation data set of 12 MC congeners and their conjugates shows slightly different conformations for the MC congeners previously simulated, and a greater variety of backbone conformations for the conjugate structures. Some non-toxic MC congeners have the same backbone conformation as toxic congeners, supporting the assumption that the Adda side chain is indeed an important factor in binding. For [Enantio-Adda5]-MC-LF, however, a completely different backbone to MC-LF is observed, suggesting that modification of Adda can, but does not necessarily have to, alter the backbone conformation.

In addition, the two MD simulation data sets are used to derive Interaction Fingerprints (IFP). IFPs are numerical vectors that encode an interaction between two molecules, in our case MC congeners and PPP1, as present (1) or absent (0). A method is developed to automatically derive IFP from MD simulation data for each time step. In contrast to the reported procedure, where all IFP of each individual time step are summarised into one so-called aggregated IFP, we develop a method to analyse IFPs in more detail. We consider IFP dynamics and changes, in order to analyse and compare individual changes in interactions between individual MC congeners over time. The development of IFP analysis is not limited to macrocycles, but is applicable to any system where interactions between two molecules are studied. The approach for aggregation and visualisation developed here is therefore a valuable contribution to the field and enables easier and faster analysis of interactions in MD simulations.

In summary, this thesis presents and discusses interconnected approaches and methods that have been developed to deal with small data sets, and have demonstrated their applicability to the use case of MC congeners. These approaches and methods helped to derive biologically meaningful insights with computational methods, applicable to a wide range of other use cases in the area of bio- or cheminformatics.

Table of Contents

List of Abbreviations

ACE	acetyl group
Adda	(2S,3S,8S,9S,4E,6E)-3-amino-9-methoxy-2,6,8-trimethyl-10-phenyl-4,6-decadienoic acid
ADMET	adsorption, distribution, metabolism, excretion, toxicity
AF	allocation factor
AI	artificial intelligence
Amba	(2S,3S)-3-amino-2-methylbutanoic acid
Anda	(2S,3S,4E,6E)-3-amino-2-methylnona-4,6-dienoic acid
ANN	artificial neural network
ANP	atrial natriuretic peptide
Apda5	(2S,3S)-3-amino-2-methyl-10-phenyldecanoic acid
Apha5	(2S,3S)-3-amino-2-methyl-7-phenylheptanoic acid
bw	body weight
CIP	Cahn-Ingold-Prelog-Convention
CPU	central processing unit
CV	cross validation
cryo-EM	cryogenic electron microscopy
Dhb	(E)-2-amino-2-butenoic acid
dof	degrees of freedom
DMSO	dimenthyl sulfoxide
EG	European guideline
FAIR	Findability, Accessibility, Interoperability and Reuse of digital assets
FN	False negative
FP	False positive
GAFF	general amber force field
GBM	gradient boosting machines
GDT	global distance test
GPU	graphics processing unit
GROMACS	Groningen Machine for Chemical Simulations
GSH	glutathione
HPLC	high pressure liquid chromatography
IC50	inhibitory concentration
ID	identity
IFP	Interaction Fingerprint
IQR	interquartile range
IUPAC	International Union of Pure and Applied Chemistry
LD	lethal dose
LINCS	linear constraints solver
MD	Molecular dynamics
MC	Microcystin
ML	Machine learning
MRP2	Multidrug Resistance-associated Protein 2

MSecPh	N-methyl-Se-phenyl-L-selenocysteine
n.d.	not determined
NLP	Natural Language Processing
NMDA	N-methyl-D-Aspartate Receptor
NME	N-terminal Methyl Group
NMR	Nuclear Magnetic Resonance
NPT	constant Pressure, constant Temperature
NVT	constant Volume, constant Temperature
OATP	Organic Anion Transporting Polypeptide
OECD	Organisation for Economic Co-operation and Development
PBC	Periodic Boundary Conditions
PC	Principal Component
PCA	Principal Component Analysis
PDB	Protein Data Bank
PME	Particle Mesh Ewald
PPP	Ser/thr Protein Phosphatase
Prg	Propargyl
ProLIF	Protein-Ligand Interaction Fingerprint
QSAR	Quantitative Structure-Activity Relationship
RCSB	Research Collaboratory for Structural Bioinformatics
RF	Random Forest
RMS	Root Means Square
RMSD	Root-Mean-Square Deviation
RMSF	Root-Mean-Square Fluctuation
ROC	Receiver Operator Curve
RQ	Research Question
SASA	Solvent-Accessible Surface Area
SAR	Structure-Activity Relationship
SI	Supplementary Information
SIFt	Structural Interaction Fingerprint
SMARTS	SMILES Arbitrary Target Specification
SMILES	Simplified Molecular Input Line Entry System
SMOTE	Synthetic Minority Oversampling Technique
SPLIF	Structural Protein-Ligand Interaction Fingerprints
TDI	Tolerable Daily Intake
TN	True Negative
TP	True Positive
VdW	Van der Waals
WHO	World Health Organization
wwPDB	worldwide Protein Data Bank
XGBoost	Extreme Gradient Boosting Machines

Chapter 1

Introduction

1.1 Motivation

Cheminformatics is a field of research at the interface of computer science and chemistry. The focus of cheminformatics is to solve chemistry-related problems, e. g. in drug discovery, fundamental biochemical studies or toxicology, and through the application and development of computational methods [1, 2]. The focus of the field is not only on prediction of e. g. bioactivity or properties of molecules, but also on data curation and databases, as well as visualisation and analysis of chemical space. This leads to a wide range of applications, with many approaches and methods being developed to solve these tasks [2, 3].

When bioactivity is considered, it often includes understanding of biomolecular processes, e. g. protein-ligand binding [1]. This is where the research fields of bio- and cheminformatics naturally overlap, as there is no clear separation between small molecules, which is the focus of cheminformatics, and macromolecules, which is one aspect of bioinformatics [1, 4, 5]. More specifically, macromolecules, such as proteins or nucleic acids, are analysed and modelled in the field of structural bioinformatics, which focuses on analysing and predicting the three-dimensional (3D) structures of large biological molecules, modelling of interactions, and visualising, analysing and comparing 3D structural features [6]. In both research fields it is crucial to prioritise, model, and understand molecular structures and their interactions, in order to identify molecules with specific properties [2, 6].

Therefore, this thesis relies on both research fields and combines them. In both research fields, small data sets of molecular structures are a particular challenge. Due to the limited amount of available data points, it is difficult to develop and test computational models and methods to predict endpoints of interest, e. g. activity or toxicity, and to draw conclusions from small data sets [7, 8]. Especially for macrocycles, mostly small data sets are available because they have been highly neglected in the past, with most investigations focusing on small molecules or acyclic structures [7]. Macrocycles are defined as cyclic structures with at least twelve atoms in their central ring structure [3, 7, 9–11]. They are not only difficult and expensive to synthesise, but it is also difficult to study their pharmacological, pharmacokinetic and toxicological profiles [3]. In addition, they have been associated with a chemical space that was traditionally not considered useful in drug discovery, since have been associated with reduced cell permeability and bioavailability [11, 12]. This resulted in small data sets that are difficult to analyse due to the lack of data points. Recently, there has been a growing interest in macrocycles, as they are structurally interesting since they cover a broad surface area in a restricted conformation, leading towards higher target selectivity. This is caused by a lower

number of rotatable bonds and therefore reduced degrees of freedom [11]. Today, macrocycles are one of the most active research areas in cheminformatics, where the focus is on studying their structure, conformation and physicochemical properties in the context of toxicology and drug discovery [3]. Methods such as molecular docking, quantitative structure-activity relationship (QSAR) models, machine learning (ML), or molecular dynamics (MD) simulation can be used to predict or analyse potential activity or toxicity of macromolecules and help speeding up the process and aid in understanding mechanisms of bioactivity [13, 14].
The use case analysed in this thesis is Microcystin (MC) congener toxicity. MC congeners are cyanotoxins and a group of structurally similar macrocycles (see Figure 2.1). There are currently 279 known MC congeners [15], but only a few of them have been tested and analysed, so there is limited data available [16]. They are released during cyanobacterial blooms in water bodies worldwide [17] and are critical to animal and human consumption [18]. Most of the MC congeners are known to be highly toxic, and exposure to them often occurs by drinking, bathing, ingestion or inhalation of contaminated water [19–22]. Long-term damage to multiple organs and, in extreme cases, death has been associated with a high intake of MC congeners [23–25]. The current guideline of consumption of MC congeners by the World Health Organization (WHO) [26] focuses on only one MC congener, namely MC-LR. The provisional guideline is based on one 13-week toxicity study in mice [27] and requires the presence of 0.001 mg/l of MC-LR, before bathing water or drinking water supply are shut down. All other MC congeners are summed up to an equivalent of MC-LR, which is scientifically highly questionable, as they differ in toxicokinetics and dynamics, while experiments have shown that MC-LR is not the most toxic MC congener [16]. Their structural modifications lead to differences in the uptake of MC congeners via organic anion polypeptide transporters (OATP) into the cell [28–30]. In addition, they have different toxicodynamics through interaction with ser/thr-protein phosphatases (PPPs, i.e., PPP1, PPP2A, PPP5, PPP6) [31–33], as well as conjugation via Glutathione (GSH)-Transferase [34, 35]. They are also exported differently via the multidrug resistance-associated protein 2 (MRP2) and a still unknown exporter [36].

Computational predictions and modelling are difficult because MC congeners are macrocycles with very different bioactivities [16], although they are structurally closely related [15]. Moreover, the available data is limited. [16, 33, 37] Therefore, new methods and workflows need to be developed to investigate macrocyclic behaviour and bioactivity.

1.2 Research Questions

The aim of this thesis is to develop methods and workflows to investigate molecular toxicity, where only a limited number of data points are available. The use case studied here is based on MC congeners toxicity. MC congeners are macrocycles that are particularly challenging and difficult to analyse compared to small molecules [3, 14, 38, 39]. However, the methods developed in the course of this thesis, are also applicable to small molecules and, more generally, to other use cases or biological systems, as the methods are applicable to molecular structures in general. The research questions (RQs) are described in the following:

RQ1: What new methods can help to overcome the shortcomings of small molecular data sets?

Due to the small amount of experimentally validated data that is available, it is difficult to estimate toxicity and interactions for all other MC congeners. Therefore, the focus of this thesis was to develop and evaluate methods that are suitable for a small number of data points. In addition, it was investigated how the knowledge derived from a small data set can be transferred to learn about a larger data set, i.e., MC congeners in general.

RQ2: What new methods can help to understand macrocyclic influence and dynamics, and hence bioactivity?

It is known that macrocycles can lock in and lock out certain conformations, which influences binding and hence bioactivity, making it difficult to predict the 3D conformation of macrocycles [40]. It has recently been shown that similarity-based calculations on the 3D structure of small molecules perform better than similarity-based calculations on two-dimensional structures [41]. Considering this finding, the 3D conformation of macrocycles was modelled and investigated to infer about their bioactivity and to extrapolate the findings to other MC congeners. One important approach to study 3D conformations are MD simulations, and thus a further research question was:

RQ3: How can we aggregate and compare interaction fingerprints based on Molecular Dynamics simulations?

Binding to a target and therefore toxicity is not only influenced by conformation, but also by interactions. For this reason, the interactions of MC congeners with PPP1 were investigated to systematically evaluate which interactions influence toxicity. Therefore, a method was developed to automatically derive present and absent interactions, which resulted in a large amount of interaction fingerprints over time. The interaction fingerprints were then aggregated and analysed in a newly developed approach, to evaluate the interaction differences between MC congeners and PPP1.

1.3 Thesis Outline and Contribution

This section outlines the structure of the thesis and contributions. Chapter 1 motivates the research of this thesis. Chapter 2 provides the theoretical background. Chapter 3 discusses approaches and methods to predict and analyse the toxicodynamics of MC congener toxicodynamics, while Chapter 4 describes molecular modelling approaches to study interactions and conformations. Chapter 5 describes the development of automated generation and analysis of interaction fingerprints. Chapter 6 summarises the contributions of this thesis and highlights open research challenges.

Chapter 1 motivates the research, outlines the research objectives, and gives a general overview of this thesis.

Chapter 2 summarises the theoretical background of different research areas which are relevant to this thesis. First, the areas of bio- and cheminformatics, macrocycles in cheminformatics, representation and encoding of molecular structures, and molecular modelling are discussed. Subsequently, aspects of computer science are discussed in more detail, i.e., methods for feature space reduction and specific aspects of machine learning. Then, the area of biology is discussed with information on the biological use case, i.e., MC congener structure and toxicity, as well as one of their targets, ser/thr protein phosphatases (PPPs)).

Chapter 3 discusses the structural analysis and prediction of MC congener inhibition of PPPs and is divided into two parts. Both parts are based on a data set with half maximal inhibitory concentration (IC_{50}) values of 17 MC congeners on PPP1, PPP2A and PPP5.

Machine Learning: For the first approach, a consensus machine learning model is built to classify IC_{50} values in three toxicity classes. MC congeners and PPPs are encoded with features based on natural language processing and syntethic minority oversampling technique is applied to upsample the small data set. This work has been published in:

- Stefan Altaner*, **Sabrina Jaeger***, Regina Fotler, Ivan Zemskov, Valentin Wittman, Falk Schreiber and Daniel R. Dietrich: 'Machine learning prediction of cyanobacterial toxin (microcystin) toxicodynamics in humans' Alternatives to Animal Experimentation 37 (1): pp. 24-36, 2020. doi: `10.14573/altex.1904031`. * Equal contribution.

Contribution: Collaboration with S. Altaner, D. Dietrich and F. Schreiber, R. Fotler, I. Zemskov and V. Wittman. S. Altaner and R. Fotler did the experimental studies and measured the IC_{50} values for PPP1, PPP2A (S. Altaner) and PPP5 (R. Fotler). The synthetic variants were synthesised by I. Zemskov. We (S. Altaner, D. Dietrich, F. Schreiber and me) came up with the idea to develop a machine learning model, encoded MC congeners and PPPs as feature vector and built and evaluated the prediction models. S. Altaner and I interpreted the results of the prediction models in a biological context. The manuscript was written by S. Altaner and all parts regarding the machine learning approach were written by me. All authors read and approved the final manuscript.

Mathematical Optimisation: Since the machine learning model built is a black box, mathematical optimisation, i.e., the (α, β)-k-Feature Set Problem is used as a complementary approach. This approach is used for the first time on molecular data by representing MC congeners as extended connectivity fingerprint representation. This method reduces dimensionality

by optimising for a minimum number of feature sets (k) with at least α different features of samples of different classes and β common features of samples of the same target class. Biologically meaningful toxicity signatures can be identified, and Boolean rules derived to classify toxicity by mapping features on chemical substructures, which gives insights into important substructures that explain toxicity. This work has been published in:

- Pablo Moscato, **Sabrina Jaeger-Honz**, Mohammad N. Haque, and Falk Schreiber: 'The (α,β)-k Boolean signatures of molecular toxicity: Microcystin as a case study', bioRxiv, 2024. doi:10.1101/2024.12.29.630644.

Contribution: Collaboration with P. Moscato, M. N. Haque and F. Schreiber. P. Moscato and F. Schreiber had the idea to apply the (α, β)-k-Feature Set Problem to MC congener toxicity data. I came up with the idea to encoded MC congeners as extended connectivity fingerprint. M. N. Haque implemented the optimisation, P. Moscato derived the Boolean Rules. I interpreted the results of the prediction models, generated all images with substructures mapped to the MC congener structure, interpreted the biological results and collected a new data set. I derived the feature sets for the new data set based on the instruction of P. Moscato. The manuscript was partly written by P. Moscato and all parts regarding the encoding of MC congeners, parts of the introduction, and results and discussion section with biological interpretations were written by me. All authors read and approved the final manuscript.

Chapter 4 discusses the molecular modelling of macrocyclic structures and is divided into two parts. The approaches and methods used in Chapter 3 help to estimate MC congener toxicity, but do not provide information on the interaction of molecular structures to understand toxicity in detail. As the 3D structure and conformation is relevant, MD simulations were performed on MC congeners and PPP1 to study and investigate dynamics and interactions of the binding process.

Molecular Dynamics Simulation of 4 MC Congeners: Since MD simulation is tedious and time-consuming, the initial data set built was small but covered all three toxicity classes (toxic, less toxic and non-toxic). Parameters for MC congeners are derived manually to run the simulation, which was a cumbersome process on this data set. This work has been published in:

- **Sabrina Jaeger-Honz**, Jahn Nitschke, Stefan Altaner, Karsten Klein, Daniel R. Dietrich and Falk Schreiber: 'Investigation of microcystin conformation and binding towards PPP1 by molecular dynamics simulation' Chemico-Biological Interactions 351: 15 pages, 2022. doi: 10.1016/j.cbi.2021.109766.

Contribution: Collaboration with J. Nitschke, S. Altaner, K. Klein, D. Dietrich and F. Schreiber. The study was designed by me, S. Altaner, K. Klein, D. Dietrich and F. Schreiber. J. Nitschke and S. Altaner prepared the structures and were responsible for docking. I parameterised all structures, and did the MD simulations. I did the analysis, scripting and interpretation. The initial interpretations were discussed with J. Nitschke and K. Klein. J. Nitschke wrote the initial draft for the introduction and docking of structures for the manuscript. D. Dietrich and I wrote the draft for the rest of the manuscript. All authors read and approved the final manuscript.

Extending Molecular Dynamics Simulation with PPP1: With the first data set, it was possible to identify meaningful differences in MC congener conformation and interaction dependent on

the toxicity classes. Therefore, a second data set was built with 10 MC congeners and two of their respective conjugates (cysteine and glutathione) in two different configurations (R and S), i.e., different orientation in 3D space. As this data set was much larger than the other one, different parametrisation methods were used to accelerate MD simulation setup and to make the approach feasible. This work has been published in:

- **Sabrina Jaeger-Honz**, Raymund Hackett, Regina Fotler, Daniel R. Dietrich and Falk Schreiber: 'Conformation and binding of 12 Microcystin (MC) congeners to PPP1 using molecular dynamics simulations: A potential approach in support of an improved MC risk assessment.', Chemico-Biological Interactions 407: 15 pages, 2025. doi: `10.1016/j.cbi.2025.111372`.

Contribution: Collaboration with R. Hackett, R. Fotler, D. Dietrich and F. Schreiber. The study was designed by me, D. Dietrich and F. Schreiber. R. Hackett was supervised by me during an internship and assistant job. R. Hackett and I prepared the structures, were responsible for docking, parameterised all structures, and did the MD Simulations. R. Hackett provided scripts for parts of the analysis based on my previous analysis, and implemented 3D density mesh grid visualisation for PCA analysis. I carried out the analysis and interpretation of the results and discussed their biological relevance with R. Fotler. R. Hackett wrote the initial draft for the methods part. The manuscript was written by me and I prepared figures and tables. All authors read and approved the final manuscript.

Chapter 5 discusses interaction fingerprints (IFPs) for MD simulations. Since interactions are crucial for toxicity, and interaction analysis in MD simulations is tedious and time-consuming, a method was developed to automatically derive IFP from MD simulation data for each time step. This returns many vectors describing interactions between MC congener and PPP1 residues (i.e., IFPs). Since it is unfeasible to analyse all IFP, methods to aggregate, visualise and compare them are developed. This work has been published in:

- **Sabrina Jaeger-Honz**, Karsten Klein and Falk Schreiber: 'Systematic analysis, aggregation and visualisation of interaction fingerprints for molecular dynamics simulation data', Journal of Cheminformatics, 16 (28): 15 pages, 2024. doi:10.1186/s13321-024-00822-3.

Contribution: Collaboration with K. Klein and F. Schreiber. I had the initial idea to analyse IFPs from MD simulation and developed a method to automatically derive them from MD simulation data and had the idea to aggregate, visualise, and compare the IFP since the standard method is to only consider one aggregated frame. Different approaches for visualisation and aggregation, were intensely discussed with and suggested by K. Klein and F. Schreiber before resulting in the final methodology. I implemented a library to automatically do the post-processing, analysis and visualisation.

Figure 1.1 summarises the different concepts discussed in Chapters 3 to 5 and shows how the different concepts interrelate.

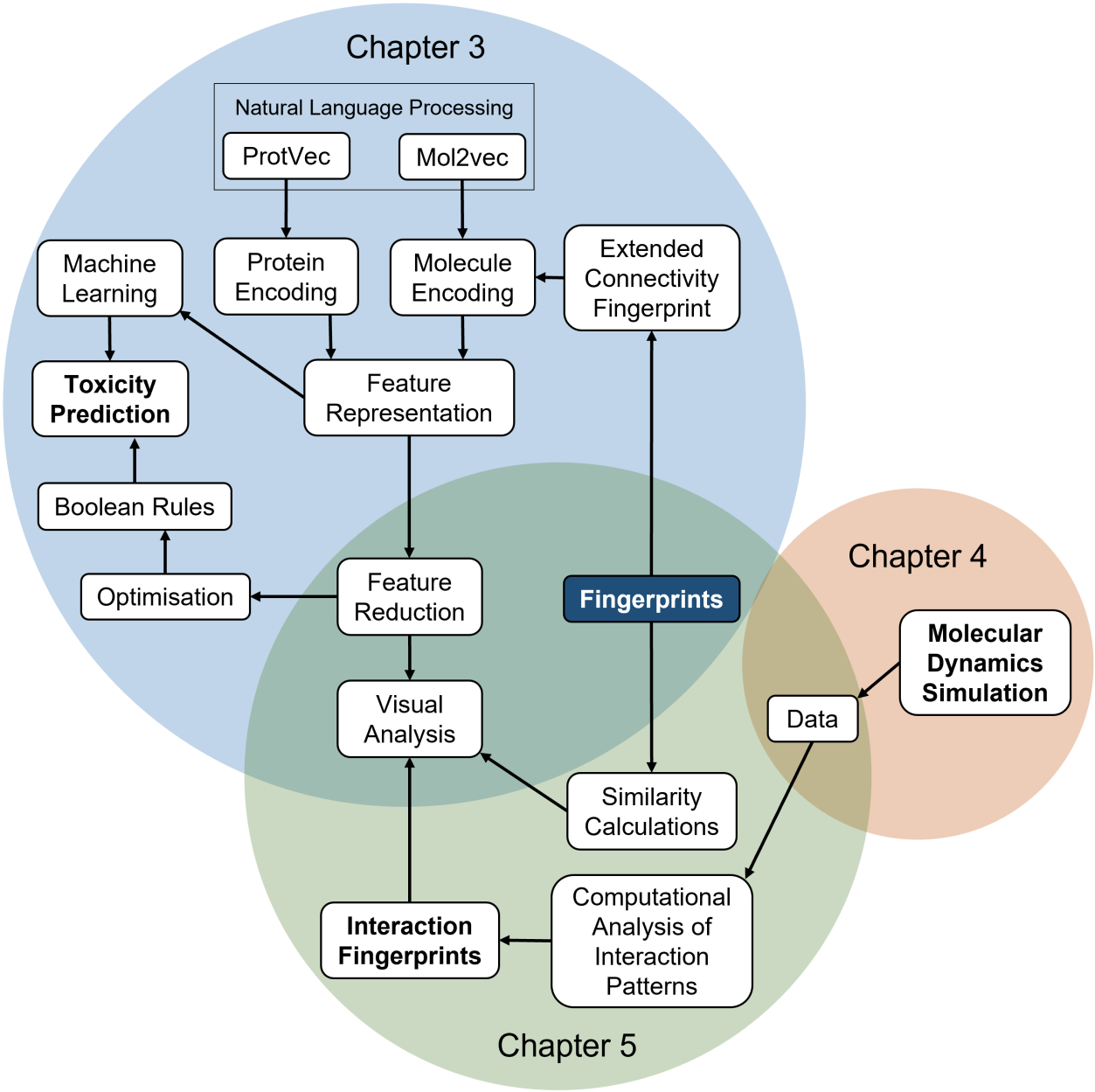


Figure 1.1: Illustration of the interrelated content, concepts and structure of the thesis. The size of the circles does not indicate the importance or length of the individual chapters, but has been adapted to cover the most important bio- and cheminformatics concepts discussed in this thesis. The main theme of each chapter is written in bold letters, and the overall theme of fingerprints, linking chapters is marked blue.

Chapter 6 summarises the findings of this thesis. In addition, an outlook on open research questions and future developments is given.

Parts of this thesis have been published in the journals as listed above. In order to keep the thesis readable and understandable, the publications have been rewritten, but ideas, figures and tables have been reused in the respective chapters. If a chapter or part of a chapter has been published, this is indicated at the beginning (summary) of each chapter.

1.4 Open Science

In this section, the concept of Open Science is briefly discussed, as this is an important topic for the scientific community and is also relevant to this work, as part of it has been published according to the principles of open science. The amount of data produced in recent years continues to grow, and we are collecting more, larger and more diverse data sets. Even though there have been great improvements in generating data, reproducibility is often an issue and many results and findings cannot be reproduced [42]. In some scientific fields it is good practice to share data or software, such as in bioinformatics [43].

Open science as a concept of sharing data, information, and knowledge openly has gained popularity over the last few years, even though the concept is not new. Different global scientific bodies, such as the European Commission, the US Office of Science and Technology Policy, and the Global Research Council, try to encourage the scientific community to make research data, software code and experimental methods publicly available. The aim is to address the problem of reproducibility and to promote scientific progress by making data accessible. Hence, the aim of open science is to enhance efficient sharing of scientific data and knowledge within and across the scientific community, to accelerate research and understanding [44].

The key concepts of open science are open data, open source software, and open journal access [42, 45]. Open data is about sharing data as a product of scientific knowledge. Data is essential to understand the insights and results obtained, and open science tries to make data more openly available and accessible to reuse. To enable efficient data storage and sharing, specific repositories have been developed to deposit data to allow open source access [42]. One example is Zenodo, which is part of the Europen OpenAIRE program and run by CERN [46]. Any kind of data, such as research papers, data sets, software, and reports, can be uploaded to Zenodo with a persistent digital object identifier (DOI) making these items easily accessible and citable, which can attract attention of other researchers or even the media. Zenodo was established in 2013, and in 2021 Zenodo already consisted of 2 million records with the first Petabyte of data [46]. This is a massive amount of data, and it continues to grow even faster. As a result, it is more and more difficult to keep an overview over the massive amount of data and computational support is necessary to deal with it.

Open source software is important for discovering knowledge and to gain insight from scientific data. Software is often a computer program, an application, or source code that helps or assists in the analysis of data or supports infrastructure to manage data. In the spirit of open science, software must be made publicly available together with a software licence that allows anyone to examine, use, change, and distribute source code for any purpose. Open source software benefits the scientific community, as it makes software reusable, reduces duplication, enhances use of data and ensures longevity of the code. In addition, it ensures that scientific results are reproducible and transparent [42].

To communicate research and science, publications in journals are essential to share results with the community. Nevertheless, access to those articles can be expensive and therefore limited, and copyright restrictions can make it difficult to share literature. In the last few years, open access journals have got more and more popular, as they provide the articles digitally available online, free of charge and with most copyright restrictions removed [42].

To make the research of this thesis accessible to the community, part of this work has been published in an open source journal:

- Stefan Altaner*, **Sabrina Jaeger***, Regina Fotler, Ivan Zemskov, Valentin Wittman, Falk Schreiber and Daniel R. Dietrich: 'Machine learning prediction of cyanobacterial toxin (microcystin) toxicodynamics in humans' Alternatives to Animal Experimentation 37 (1): pp. 24-36, 2020. doi: `10.14573/altex.1904031`.

- **Sabrina Jaeger-Honz**, Karsten Klein and Falk Schreiber: 'Systematic Analysis, Aggregation and Visualisation of Interaction Fingerprints for Molecular Dynamics Simulation Data', Journal of Cheminformatics, 16 (28): 15 pages, 2024. doi:10.1186/s13321-024-00822-3.

- **Sabrina Jaeger-Honz**, Raymund Hackett, Regina Fotler, Daniel R. Dietrich and Falk Schreiber: 'Conformation and binding of 12 Microcystin (MC) congeners to PPP1 using molecular dynamics simulations: A potential approach in support of an improved MC risk assessment.', Chemico-Biological Interactions 407: 15 pages, 2025. doi: `10.1016/j.cbi.2025.111372`.

In order to ensure to gain maximum value out of published data, workflows, tools and algorithms, the FAIR principles have been developed in 2016 [47]. FAIR stands for Findability, Accessibility, Interoperability, and Reuse of digital assets. Findability includes finding data for reuse, assigning a unique and persistent identifier, providing metadata, and describing and indexing the data set in a searchable resource. Accessibility should ensure that the user can access the data, either with or without authentication, in a standardised communication protocol. Interoperability should enable interoperating data in applications or workflows for analysis, storage, and processing. Reusability is the ultimate goal and aims to produce well described data so that replication or combination in different settings is possible [47].

Following the FAIR principles, data sets and scripts which are part of Section 4.3 and Section 5.4 have been published open-access in the Zenodo repository:

1. https://doi.org/10.5281/zenodo.5017745 [48]
2. https://doi.org/10.5281/zenodo.5017839 [49]
3. https://doi.org/10.5281/zenodo.5017851 [50]
4. https://doi.org/10.5281/zenodo.10423389 [51]
5. https://doi.org/10.5281/zenodo.10424417 [52]
6. https://doi.org/10.5281/zenodo.14501700 [53]

Chapter 2

Theoretical Foundations

In this chapter, the foundations to understand the thesis are described in detail. First, the fields of chem- and bioinformatics are explained to provide background knowledge on the development of the field. Then, macrocycles are described, followed by methods to represent and encode molecular structures. Afterwards, methods in the field of molecular modelling are described in greater detail. Next, methods for feature space reduction and machine learning are explained. This is followed by a description of the use case of this thesis, a class of structurally similar molecules called Microcystins (MC).

2.1 Bio- and Cheminformatics

Cheminformatics (or chemoinformatics) is a rather new research area at the interface between chemistry and informatics [1, 2]. During the second half of the 20$^{\text{th}}$ century, a huge amount of data and information on chemical structures became available. It was soon clear that this data would have to be stored and managed electronically in order to process the quantity. In the 1960s many methods and approaches were developed to store, manipulate and process chemical information to identify relationships between chemical structure and properties [1]. While in the beginning many researchers considered themselves as working with "chemical information" or working in "computational chemistry", cheminformatics as a term for a new research area was introduced in the late 1990s. Several definitions of cheminformatics have been proposed since then, which was also accompanied by a discussion on whether cheminformatics is a new research area or merely a new term for the combination of two previously existing research fields [1, 54]. Frank Brown defined cheminformatics in 1998 as follows: "The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organisation." [55]. In 2006, the term cheminformatics was broadened by Gasteiger and Engel and defined as: "The application of informatics methods to solve chemical problems." [56].

Today, cheminformatics has developed into its own research area and has a broad application in e. g. drug discovery, fundamental biochemical studies, and toxicology [1, 9, 57, 58]. Methods and approaches developed in cheminformatics help to solve chemistry-related tasks [9] and to generate and extract knowledge from data [2]. Cheminformatics includes different aspects such as representations of chemical data, chemical databases and data sources, visualisation and analysis of chemical space, as well as different methods for data analysis, e. g. search methods,

calculation of descriptors, physical and chemical data, prediction models and more [1, 2].

However, if we try to enhance our understanding of biomolecular processes, i.e., in understanding protein-ligand binding [1], which is also the aim of this thesis, the areas of cheminformatics and bioinformatics do overlap. In this thesis, there is no clear separation between small molecules (ligands), which is the main focus of cheminformatics, and proteins, which is one aspect and focus (among genes and larger chemical compounds) of bioinformatics. For this reason, both research areas do have a natural overlap, which is not limited to protein-ligand interactions, but also includes other research areas, e. g. databases, database search and comparison, as well as prediction of properties [1, 4, 5].

Bioinformatics, like cheminformatics, is a very broad field. Therefore, the focus of the following paragraphs is on structural bioinformatics, as this is the focus of this thesis. Structural bioinformatics is a branch of bioinformatics which analyses and predicts the three-dimensional (3D) structure of large biological molecules such as proteins, RNA and DNA in order to generate new knowledge related to biology. The aim of structural bioinformatics is to visualise, analyse and compare three-dimensional structural features, to predict the structure or function of a macromolecule (e. g. a protein), or to model interactions, e. g. with Docking or Molecular Dynamics (MD) Simulations [6, 13]. The central dogma of structural bioinformatics is, that the structure of a protein leads to its function. The first protein structure was resolved and described in 1958 by Kendrew et al. [59], who showed that proteins have complex 3D structures, which are based on the 3D arrangement of secondary structural elements (i.e., α-helices and β-sheets) [59].

The development of a collection of 3D protein structures and a molecular graphics display to visualise them [60] led to the establishment of the Protein Data Bank (PDB) [61] in 1971. In the same year, Edgar Meyer developed SEARCH, which allowed researchers to access information in the PDB remotely and offline to study protein structure [62]. The PDB was a joint venture between the Cambridge Crystallographic Data Centre (UK) and the Brookhaven National Laboratory (USA) and originally contained seven structures. At the end of the 1990s, the Research Collaboratory for Structural Bioinformatics (RCSB) took over the management of the PDB, and gave it the name it still has today: RCSB PDB. The PDB developed into the Worldwide Protein Data Bank (wwPDB) which was established in 2003 to make the PDB publicly available worldwide. It was the first open access digital data resource in biology and in medicine and is a single archive that is freely and publicly available worldwide [61, 63]. Since the founding in 1971, the amount of data in the PDB has grown massively and by the end of 2022, the RCSB PDB contained more than 200,000 structures and recently also integrated more than 1 million computed structure models [64].

However, protein structures are not static, as they are closely linked to protein function. Advances in biophysics in the 1950s led to the rapid developments in the field of Molecular Dynamics (MD) simulation, which helps to gain insight into the dynamics of biological systems [65, 66]. In the following, the progress in the field of MD simulation is briefly described, as this is a crucial technique used in this work. The first system studied with MD simulation was a collision among hard spheres described by Alder and Wainwright in 1957 [67]. This model was simple, yet not realistic. The first realistic model of an MD simulation of liquid Argon was published by Rahman in 1964 [68], and in 1971 the first MD simulation of a molecule (liquid water) and not just atoms was published by Rahman and Stillinger [69]. Soon after, in 1977, the first MD simulation of a biomolecule was published by McCammon, Gelin and Karplus [70], which was the Bovine Pancreatic Trypsin Inhibitor simulated for 9.2 ps in vacuum with a simple crude molecular mechanics potential [70]. This achievement was the starting point for a rapid

and great success story. In the 1980s, internal motions of proteins were investigated and interaction studies of proteins and small molecules followed [71]. At the end of the 1990s, advances in MD simulation led to the first 1 μs simulation of a protein in solvent while observing its folding, which took two months of central processing unit (CPU) time [72]. State-of-the-art simulations of that time were two orders of magnitude shorter, and the achievement was only made possible due to massively parallel supercomputers and efficient code [72]. Since then, there has been a massive increase in computing power as graphics processing units (GPUs) became available. This development has greatly reduced the waiting time for simulations, and we are now able to simulate huge systems [13]. In 2006, Freddolino and colleagues published the first MD simulation of a cell — the Satellite Tobacco Mosaic Virus with 1 million atoms were simulated for 50 ns [73]. One of the greatest achievements in the field followed in 2013, when the Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt, and Arieh Warshel for developing methods to simulate the behaviour of molecules by using classical and quantum mechanical theory [74].

In summary, in the era of big data, computers, and software are an essential tool to analysing and understanding big data sets, which can take up to several gigabytes on a hard drive. With advances in the field of structural bioinformatics (and other disciplines of bioinformatics) and cheminformatics, the mass production of data continues and specialised software and algorithms are required to analyse these big data sets, making computational methods more relevant than ever [75].

2.1.1 Macrocycles in Cheminformatics

Today, the investigation of macrocyclic structures, their physicochemical properties, and conformation is among the most active research areas in cheminformatics [3]. In the past, they have been severely neglected and research and data sets focused on small molecules or acyclic structures. Compared to small molecules, macrocycles are difficult and costly to synthesise, identify and analyse, i.e., to determine their mode of action, as well as their pharmacological, pharmacokinetic and toxicological profile [3]. In addition, they were associated with a chemical space (beyond the rule of 5 [76]), which was traditionally said to have reduced cell permeability and oral bioavailability, which are crucial in drug discovery and therefore lead to a neglect of those molecules [11, 12]. This often results in small data sets that are more difficult to analyse due to the lack of data points [7, 8]. Recently, however, there is a growing interest in macrocycles due to their interesting properties and differences in behaviour compared to small molecules [40].

Macrocycles are defined as chemical structures that are made up of at least twelve atoms in the cyclic structure [3, 7, 9–11]. The most relevant classes of macrocycles are natural products (peptidic and non-peptidic), or non-natural/synthetic macrocycles or peptides [77]. Due to their size, they have a higher molecular weight and are more hydrophobic than small molecules. Macrocycles have a complex 3D structure, due to higher complexity in stereochemistry with usually many chiral centres, a high diversity of ring systems, and ring complexity [14, 38, 39].

In comparison to small molecules, macrocycles are conformationally restricted and are more likely to adopt disk- and sphere-like conformations [78, 79]. The reduced flexibility influences the overall molecular shape, leading to modified biological and physicochemical properties, which can be pharmaceutically relevant [77, 80]. Macrocyclisation can lock in and lock out conformations in the conformational space accessible to the molecule, potentially making them more affine towards target molecules [40].

The lower number of rotatable bonds, and therefore fewer degrees of freedom, lead to a high pre-organisation in the cycle compared to acyclic molecules with similar weight [11]. This results in a lower number of conformations that are available to the unbound macrocycle [77]. To successfully bind to another biomolecule, a bioactive conformation must be adopted. Due to the high structural restrictions, the entropic cost upon molecular binding is lower, and consequently a bioactive conformation is entropically favoured [7, 77]. Macrocycles require large surface contacts [80] and form multiple, complex protein-ligand interactions [10] with large, flat or groove-shaped binding sitesdue to their conformation and size [78, 81]. Moreover, the macrocycle backbone provides a scaffold for side chains that can further facilitate potency and selectivity towards a protein target, making macrocycles also interesting for drug development [7, 82] against difficult-to-drug targets [83].

Cheminformatics and bioinformatics play a crucial role in prioritising, modelling, and understanding molecular structures and their interactions to identify macrocycles with specific properties. Potential activity or toxicity of macrocycles can be predicted using methods, such as docking, quantitative structure-activity relationship models, machine learning, or Molecular Dynamics (MD) simulation, and thereby highly speed up the process and help in understanding mechanisms of bioactivity [3, 5, 9, 13]. Macrocycles have complex structures that pose fundamental challenges to researchers, and new algorithms and methods need to be developed to improve the understanding of their conformation and binding mechanisms in the application areas of toxicology and drug discovery [3, 10, 79, 84].

2.1.2 Representation and Encoding of Molecular Structures

Many different methods and ways exist to represent molecular structures. If we want to encode and represent a molecular structure to a computer, a machine-readable encoding is necessary to access, query, and search digitally stored data [2, 85]. Different classes of notations and representations exist, encoding information on a different level. The simplest ones are linear representations (one-dimensional — 1D), two-dimensional (2D) representations and 3D representations for specialised programs and software [2].

2.1.2.1 Molecules

From a chemists point of view, a molecular structure is a diagram drawn with bonds and atoms. This approach of looking at a molecular structure as a graph is borrowed from graph theory in computer science. Molecular structures can be represented as a graph, where the nodes are a set of atoms, labelled with the element name (letters) and the edges are a set of (undirected) interactions. To represent interactions, an adjacency, or connectivity matrix can be used, which does not specify the type of bond. A graph is defined as a 2D object and formally has no spatial position, only pairwise relationships. Therefore, positions or other 3D information such as bond angles or chirality need to be encoded as attributes [86]. When a molecule is described by a property i.e., numerical values, it is called a molecular descriptor, which can be used to manipulate and analyse chemical structures. Many different molecular descriptors exist and they vary in their complexity [2, 87]. In the following, different representation of molecules relevant to this thesis will be discussed. Many other methods have been developed for the representation of molecular data, and for a recent and detailed review the reader is referred to Wigh et al. [88].

1D Representation The simplest linear representation of a molecular structure is a string, which is easy to manipulate, as well as compact and memory efficient to store. A common representation of a molecule as a string is for example SMILES [89]. SMILES stands for Simplified Molecular Input Line Entry System and was developed in 1988 by Weininger [89]. It is one of the most popular notations today because only a few rules are necessary to understand SMILES strings. The molecule must be traversed until every atom has been visited once, to generate a SMILES string. If there is a ring structure in the molecule, the ring is 'broken', and branching points in the molecule introduced. Where several options exist for how to proceed, they are marked with an opening parenthesis. If all atoms in this branch have been visited, a closing parenthesis is made. Each atom is assigned to its atomic symbol, with upper case letters representing aliphatic atoms and lower case letters representing aromatic atoms. Double or triple bonds are annotated by "=" and "#", respectively. Chirality is indicated with the "@" symbol, and hydrogens are usually not added to the SMILES string unless they are necessary to represent a centre of chirality. E/Z or cis-trans (geometric isomerism) of the molecular structure is indicated by slashes (/ or \), respectively [89]. SMILES are non-unique and unambiguous, and multiple atom numberings for the same molecule can result in different SMILES, depending on the graph traversal algorithm and starting atom selected. To ensure that the same SMILES is generated for a given molecule, a unique SMILES can be generated by a canonicalisation method. The canonical SMILES is therefore a unique ordering of atoms of a given molecule. Many different algorithms have been developed to determine the canonical SMILES of a molecule (e. g. proposed by Morgan [90], O'Boyle [91] or Schneider et al. [92]). The Morgan algorithm was developed to determine the canonical order of atoms in a molecular structure [90]. It is an iterative algorithm that assigns each atom a connectivity value equal to the number of connected atoms. During the next and subsequent iterations, the connectivity values of the neighbours are summed up, and the sum is the new connectivity value assigned. This is repeated until the different connectivity values reach a maximum. The maximum is then chosen as the first atom, then its neighbours are listed in order of their connectivity. Then their neighbours, etc. If neighbours have the same connectivity, additional properties such as atomic number and bond order are calculated and considered to resolve the conflict [90]. Nevertheless, SMILES is limited by the fact that there are many algorithms resulting in different SMILES for the same molecular structures and handling stereochemistry is limited and dependent on the approach [91].

Later, an extension of SMILES called SMILES Arbitrary Target Specification (SMARTS) was proposed, which allows substructure search. SMARTS allows not only information on atoms and bonds, but also general specifications of the molecular graphs, such as regular expressions and logical operators used in computer science [86, 93].

2D Representation of molecules can be calculated based on their properties, e. g. structural, physicochemical, topological, or electronic properties [2, 87].

Simple ones with very low computational cost include simple counts, e. g. molecular weight, number of hydrogen bond donors and acceptors, ring system, rotatable bonds and others. Often they are not sufficient to distinguish molecules based on their properties, so they are used in combination with other descriptors. For physicochemical properties, one of the most commonly used ones is the log P value (logarithm of partition coefficient between n-octanol and water) which is an indicator of the hydrophobicity of a compound and can be verified experimentally [2]. Using descriptors of a molecule as fingerprint (i.e., a numerical vector representing a molecule) is called chemical fingerprint. They provide a quick and direct mapping from a molecular

structure (or graph) to a representation as a vector. The vector is a flexible representation which is used as input for numerical calculations [86].

In addition, 2D fingerprints also exist and are a widely used representation. Originally they were developed for substructure and similarity searching, but today they are widely used as molecular descriptors in many application cases. In fact, they perform overall very well, probably because the fingerprint encodes properties important for biological activity. There are four different classes of 2D fingerprints: substructure key-based fingerprints, topological or path-based fingerprints, pharmacophore fingerprints or circular fingerprints [94, 95].

The most widely used circular fingerprint is the Extended Connectivity Fingerprint (ECFP) [96]. It is a variant of the Morgan algorithm [90] and therefore also known as Morgan fingerprint. It is a local operation to generate an initial identifier for each heavy atom in a molecular structure. In contrast to the Morgan algorithm, the generation of identifiers terminates after a fixed number of steps, instead of creating unique identifiers. The initial identifiers generated are based on Daylight atomic invariants [97]. The atomic invariants depend on the number of heavy atom neighbours, the valence of an atom minus the number of hydrogens, the atomic number, the mass and charge, and the number of attached hydrogens, plus whether the heavy atom is part of at least one ring. This information is then hashed into a single 32-bit integer value, that is the initial identifier. Dependent on the radius selected, this is repeated for each heavy atom as the radius increases. For example, at radius 0, only the heavy atom and its bonds are considered. For radius 1 and greater, the neighbouring atom information of the current heavy atom is also taken into account. At some point, no new information is added to the identifier set, when iteratively increasing the radius. The final set contains all generated identifiers of the different iterations of each heavy atom. To avoid redundancy in the ECFP structural duplicates (i.e., identical identifiers) are removed. Then, a hash function maps all the generated identifiers randomly and without uniform coverage into a vector of fixed size [96], which is usually between 1024 and 16384 bits long [98]. If a feature (or substructure) is present, the bit is set to 1 (on); if a feature (or substructure) is absent, the bit is set to 0 (off). The resulting vector is sparse, as most bits will be set to 0, so many features are usually absent. In general, this procedure condenses the given information, but can lead to collision when several identifiers are hashed on the same bit. However, this is very rarely the case. The calculation of the fingerprints is really fast, reduces memory consumption and can also represent novel substructures [96]. They can be used for various application cases, such as similarity searching, clustering, virtual screening or machine learning [94–96].

3D Representation of a 3D molecular structure as a vector is of vital interest for many applications, since molecular recognition depends on the 3D structure. Therefore, the 3D molecular structure must be generated, which can be difficult without experimental data [2, 99]. Conformational flexibility can be problematic, especially for macrocyclic structures, since generation is very time-consuming for many molecules. Different encodings were developed, as e.g. 3D Fragment Screens , Pharmacophore Keys , 3D topographical indices, geometric atom pairs, or others [2]. As 3D representations of molecular structures as vectors are not relevant to this thesis, they will not be described in further detail.

2.1.2.2 Natural Language Processing Based Representations

New representations of molecular structures have been developed with the breakthrough of different methods in the area of natural language processing (NLP). NLP is a research area

aiming to learn the meaning of text and to transform language to vectors [100]. In 2013 Google researchers published Word2vec [101], which was a major breakthrough in the NLP area. Word2vec is an unsupervised machine learning approach based on the idea that words which are close in semantic meaning are also close in vector space. Word2vec is a neural network trained on a large corpus of documents (i.e., words) to generate a dense vector representation, a so-called embedding. The major advantage of this approach is that it can be trained once on a huge corpus to learn dense vector representations, and later simply applied to new words. It successfully learned word representations that could capture semantic meaning. It was possible to reconstruct linguistic meaning with simple vector operations: e.g. King - Man + Woman = Queen [101, 102].

In 2015, this approach was transferred to proteins, the so-called ProtVec approach [103]. A corpus of about half a million protein sequences has been collected from UniProt. Each protein sequence (comprised of amino acids) was used as input "sentences" for the approach. For each sequence, all possible 3-grams were generated, since it is unknown where a "protein word" begins and ends, all possible 3-grams were generated by sliding through the sequence. In the end, three sentences per protein were generated. This corpus of protein sequences was then used to train a neural network to learn distributed representations, i.e., n-dimensional vector representation of 3-grams. To represent a new protein structure, all possible 3-grams are generated in three sentences, and all individual 3-gram vectors summed up to represent the protein as one n-dimensional vector. It has been shown that this approach can be successfully used to visualise protein data, to classify protein sequences into their families, and serve as protein representations in machine learning approaches [103].

In 2018 this approach was transferred to molecular structures in an approach called Mol2vec [104]. Here, identifiers generated with the Morgan algorithm are used as "words". Instead of hashing them back into a bit vector, as in the ECFP approach (see Section 2.1.2.1), the identifiers were extracted and arranged as a "molecular sentence" by alternating identifiers of different radii. The order of the identifiers was sorted based on their canonical SMILES, so that a fixed size vector for a molecule was achieved. To train the model, a molecular corpus was collected from ChEMBL [105] and Zinc [106], and filtered according to several criteria, resulting in 19.9 million compounds represented. Researchers showed that this approach can capture chemical relationships between substructures by encoding amino acids as molecules. In addition, Mol2vec showed promising results for classification (i.e., feature encoding) or regression in machine learning tasks [104].

2.1.3 Molecular Modelling

Molecular modelling is a research area concerned with 3D structural aspects of chemical or biological molecules. Various methods and algorithms have been developed to represent and manipulate three-dimensional structures to link physicochemical properties of molecules to their biological role [107].

Interactions between molecules are crucial for many biological processes. For this reason, interactions and complex formation is one specific focus of molecular modelling, and many different methods were developed to study these processes [107, 108]. For this thesis, the focus will be on the formation of protein-ligand complexes, and therefore on Molecular Docking (or docking) and Molecular Dynamics (MD) simulation. Following the definition of a ligand, molecules binding to a target will be referred to as ligand in the following sections. To understand the different approaches and theories developed for docking and MD simulation, different binding

models proposed in literature are summarised in the following. The initial model to describe a protein-ligand complex was the lock-and-key model by Fisher [109]. The lock-and-key model describes the ligand as a key, and the protein as a lock. Both are rigid and upon binding the key (ligand) recognises its lock (protein) because their shapes are complementary [109]. Certain behaviour, e. g. non-competitive inhibition or allosteric modulation, was not explained by the original binding model. Therefore, Koshland [110] proposed the idea of an induced fit model. The induced fit model is based on the idea of mutual adaption. Upon recognising each other and getting into closer contact, ligand and protein adjust their shapes until they are at the optimal fit [110]. Ma et al. [111] proposed the idea of conformational ensembles. This theory suggests that a protein can adopt many conformations and undergo conformational changes. Some conformations are preferred for binding by the ligand, and therefore stabilise a particular conformation by shifting the conformational equilibrium upon complex formation [111]. Even though the different models may seem contradictory, they describe different aspects of the recognition process and are therefore complementary. Studying the binding and building of complexes experimentally is difficult, therefore molecular modelling techniques are frequently applied to hypothesise about an interaction or complex formation [112]. In the following, docking and MD simulation are described in further detail.

2.1.3.1 Molecular Docking

Molecular docking describes a set of tools to determine the spatial conformation and binding energy of a protein-ligand complex by providing a static pose of two molecules interacting on an atomic level. The aim of docking is to be fast and accurate, and consists of two main phases: a search of different configurations or poses and a scoring or evaluation of poses [113]. To achieve this goal, several simplifications have to be made in order to calculate the result in reasonable time. Two molecules have numerous different ways of interacting with each other, and the number of conformations is huge. Therefore, it is not possible to search the entire potential energy landscape and simplifications have to be made [114].

Different approaches with different degree of flexibility were developed. In the early 1980s and 1990s, initial algorithms for molecular docking were developed, which were based on the lock-and-key model (e. g. [115, 116]). These algorithms belong to the class of rigid docking, which is the simplest method. Both, the protein and ligand are treated as rigid entities. The binding site of the protein is approximated by a series of spheres filling it, and the ligand by a set of spheres that define its volume. In the next step, the ligand is evaluated at different positions in the binding site of a protein. The ligand is placed in different orientations in the protein binding site. Neither the ligand nor the protein can change their initial three-dimensional shape and are rigid entities with six degrees of freedom (DOF) - three translational and three rotational, which are considered during docking [115, 116].

Considering prediction accuracy and computational resources, semi-flexible docking is the most popular method used in many popular docking programs (e. g. AutoDock [117], FlexX [118]). In addition to the six DOF, the ligand is modelled with conformational degrees of freedom, which provides more flexibility and aims to explore torsional degrees of freedom. Various algorithms have been developed to sample the conformational flexibility of the ligand [114], which will not be discussed here. The protein in contrast is still treated as either rigid, or semi-rigid with individual side chains being flexible [113].

With flexible docking, the number of potential conformations is huge, as the protein (i.e., all residues) and the ligand are treated fully flexible. It is more difficult to search this conformational landscape with a good balance between accuracy, speed, and cost. The speed of these

docking approaches is kept up by the improvements which were made to the individual algorithms and increase in computational power [119]. Today, flexible docking of large molecules (or protein-protein docking) is still a major challenge, especially when considering protein backbone flexibility [120]. Docking requires high-quality 3D structures of the ligand and target protein. If a 3D structures was determined experimentally, e. g. by X-ray crystallography, Nuclear magnetic resonance spectroscopy (NMR) or electron microscopy, a high-resolution and often pre-processing of the structure is necessary [121]. Experimentally determined structures can be found e. g. in the PDB [64] or Cambridge Crystallographic Database [122]. To obtain a 3D structure of a target protein if no experimentally resolved structure is available, homology modelling can be applied [123] or the structure can be predicted. A collection of protein structure predictions is e. g. the AlphaFold Protein Structure Database [124]. The AlphaFold Protein Structure Database [124] is based on AlphaFold2, which is a neural network used to predict protein structure from amino acid sequence. In 2020 at the Critical Assessment of Techniques for Protein Structure Prediction 14 competition, it achieved a stunning performance close to 90 % in the Global Distance Test (GDT). This score determines how similar a predicted structure is in comparison to an experimental structure with 100 % being completely identical, and above 90 % is considered equally to experimental techniques like X-ray crystallography. In prior years, GDT accuracy was around 60% so the development of AlphaFold2 marks a major breakthrough in the field of structural biology and will strongly influence research in life sciences and medicine in the future [125]. Although the protein structures retrieved have a high accuracy, a recent study has shown that AlphaFold models perform worse for high-throughput docking compared to experimental structures [126].

Ligand structures of small molecules or macrocycles can be found in various chemical databases (e. g. PubChem [127] or ChEMBL [105]). The structure may be available in 2D or 3D coordinates, whereby the 3D structure (or conformation) must be generated for molecular modelling approaches [121]. If a 3D structure of a ligand (i.e., small molecule or macrocycle) is not available, it can be generated using specialised software/programming languages (e. g. RDKit [128], Open Babel [129], Avogadro [130]).

After obtaining initial 3D structures of the ligand and protein, both need to be prepared and optimised for the particular docking program, as some require a particular protonation state, for example. In addition, when a structure has been determined experimentally, some residues or atoms may not be resolved and require modelling or correction [121, 131].

Once a pose has been determined, it needs to be scored and evaluated with scoring functions [132]. They approximate the binding free energy and should correspond to a global minimum of energy of bound conformation. Scoring functions have two roles in docking: 1) identify correct conformational and orientational sampling to distinguish from incorrect ones, and 2) obtain the correct ranking of poses. Unfortunately, the correct ranking is difficult, because the estimation of free energy between bound and unbound conformations depend strongly on the competition with solvent. In addition, ligand entropy changes need to be taken into account, which is not possible to calculate. Scoring functions try to take into account various factors which are important to form a protein-ligand complex, but assumptions and simplifications are necessary to ensure fast computation and to consider both - steric and chemical complementarity [113, 114].

To evaluate the quality of docking, the docked poses are compared to experimentally determined protein-ligand complex structures and are evaluated with the root-mean-square deviation (RMSD). During a virtual screening approach, that is screening a large library of molecules against a target protein, enrichment factors are calculated. The enrichment factor is a ratio

that determines the percentage of active molecules (i.e., molecules with high binding affinity) in the sample versus the percentage of active molecules in the whole library. The higher the enrichment factor, the better the docking approach worked on the studied library [114].

2.1.3.2 Molecular Dynamics Simulation

Molecular Dynamics (MD) simulation is a very powerful but time-consuming method that allows studying and understanding the temporal evolution or dynamics of a system (trajectories) on an atomistic level of detail. MD simulations are applied to study a multitude of different systems such as nucleic acids, proteins, carbohydrates, and many more [133]. To date, it is not possible with *in vivo* or *in vitro* experimental methods to study these systems in their detailed atomistic dynamics, even though there have been major breakthroughs with NMR, and recently with cryo-EM methods [66]. In contrast to the rather rigid approach of docking, all molecules in MD simulations (water, protein, ligand, others) are treated and studied as flexible, which helps to explore kinetic behaviour, such as folding rates, and assemblies of molecules [133, 134].

A difficulty during MD simulation are surface effects of a system. All atoms that are simulated by MD are contained within a simulation box. The atoms at the border of the simulation box do not have the ideal number of neighbours and therefore have a higher energy than the atoms which are in the bulk of the simulation. Those atoms will then exhibit surface effects, and thus, disturb the simulation. One possibility to avoid surface effects without having to simulate a huge box are so-called "periodic boundary conditions" (PBC), which are frequently applied [135]. This transforms the singular simulation box to an infinite 3D lattice - when a molecule moves in the central box, its periodic image will also move in the other boxes. If it exits through one side of the box, it will enter through the other side of the box, and therefore the system has no surface. The appropriate size of the simulation box is on one hand as small as possible to allow faster computation, since the larger the box the more (solvent) molecules are in there, but on the other hand large enough so that no long-range interactions are calculated with the molecule itself which can cause interaction artefacts due to the PBC [107, 135–137].

MD simulation can be used for different purposes which includes to study dynamical behaviour of a system (e. g. dynamics of protein–ligand complexes or protein function [138]), calculate properties of a system at equilibrium (e. g. stability), to sample different configurations of a system (e. g. to study the conformational ensemble of a protein or complex, or conformations of a molecule to obtain different structures), or to refine structures in conjunction with experimental data [133].

Energy and Force Calculations The calculation of forces and energies is central to MD simulations, as they determine the dynamics of the system under study. Two different methods exist - quantum mechanics and classical mechanics. The classical mechanics approach treats a molecule as a classical object, where atoms are modelled as soft balls (or spheres) and bonds are represented by elastic sticks. The laws of classical mechanics describe the dynamics of the system, and the molecular parameters are approximated to their quantum characteristics. Therefore, classical mechanics is computationally much cheaper [65].

To be able to apply classical mechanics in MD simulations, the quantum mechanical description of system needs to be reduced. This is achieved by considering two approximations: 1) nuclei are treated as point particles since they are much heavier than the electrons which are not presented explicitly and are approximated as one potential energy surface, and 2) the Born-Oppenheimer approximation [139], which states that the wave function of electrons and nuclei

can be treated separately, because the nuclei is much heavier than the electrons and therefore is approximated as fixed or stationary, while the electrons are dynamic and move so fast that they react instantaneously to the motion of their nuclei. Nevertheless, it is a challenge to find a realistic potential that accurately mimics the potential of the energy surface and results in significant computational simplifications [65, 140].

In this work only the classical mechanics approach will be considered and the following explanations and paragraphs will focus on this approach.

Potential Energy and Force Fields The molecular mechanics model is used to calculate the potential energy in the MD simulation. The energy of a configuration is related to the forces acting on the atoms, which are soft spheres or a series of charged points. The forces are calculated based on the estimate of the potential free energy function $U(\vec{R})$, for which a force field is used [141].

A force field is a set of equations and parameters used to reproduce molecular geometry and selected properties by calculating the energies on the chemical system. In addition, it includes the potential energy function with bonded and non-bonded potential terms. In the following, the simplest form of a force field will be discussed. However, the potential energy function depends on the force field [141]. Force field parameters describe the bond lengths and angles, torsions i.e., angle describing rotation around a bond, and non-bonding interaction between atoms. In the simplest form, the potential energy function is defined as the sum of the individual potential energy contributions (additive force fields) as in Equation 2.1 [141].

$$U(\vec{R}) = U_{bond} + U_{angle} + U_{dihedral} + U_{nonbond} \tag{2.1}$$

U_{bond} is the oscillation of the bond lengths or bond stretching energy, U_{angle} is the oscillation of 3 atoms around the equilibrium bond angle or bond angle energy, $U_{dihedral}$ is the torsional rotation of 4 atoms around the central bond, which is also called torsional energy, and $U_{nonbond}$ is the non-bonded energy term [141].

The calculation of the non-bonded energy term i.e., the electrostatics and dispersion, is a particular challenge in MD simulation. The calculation requires a summation over all atom pairs, which is a huge bottleneck with respect to the computational cost [142], because the MD algorithm is dominated by the calculation of long-range interactions [143]. Many methods to accurately estimate long-range interactions have been developed (e. g. [144]) and are today routinely applied [145, 146]. However, studies have shown that large cut-off values need to be applied, since neglecting long-range interactions can lead to inaccurate models [65, 147].

Several force fields have been developed for specific applications, such as proteins, nucleic acids, lipids or sugars, and have their own unique properties. The most commonly used ones cover multiple of these categories and are AMBER [148, 149], GROMOS [150], CHARMM [151], and OPLS [152]. Although a force field is parameterised on experimental data, the parameters may be inaccurate or approximate, and simplifications are applied for calculations [141].

Force fields are well suited for many systems within the respective application area. However, the design of a force field is difficult, because it needs to be accurate, yet computationally efficient and also transferable to many systems with different conditions. In addition, parameters from one force field are not transferable to other force fields [141]. For small molecules, deriving parameters is difficult and usually involves quantum mechanical treatment. To aid in parameterisation of small molecules, different web servers and tools have been developed to estimate these parameters, as e. g. ATP topology [153], CGenFF [154], SwissParam [155] or CHARMM-GUI [156] to generate the topology for small molecules.

MD Algorithm A high-level description of the MD algorithm as a whole is given in this section.

First, initial positions and velocities - and with this, a kinetic energy - are assigned to all atoms. This procedure results in the initial molecular motion [65, 157].

Then, to simulate the motion over time, Newton's equation of motion must be solved by integration to calculate the new positions of the atoms based on the known position at time t. The time is partitioned in small time steps Δt. The definition of the time step is an important issue in MD simulation. If the time step is too small, sampling can be insufficient and difficulties crossing over high energy conformational barriers can occur. If it is too large, the stability of the system may be affected. The usual time step used for simulating several hundred nanoseconds is 2 fs. The system contains N atoms, which means Newton's equations represent a set of 3N second order differential equations which cannot be solved analytically. Instead, an approximate solution is used by discretising the equation to numerically integrate it with a finite and small time step Δt, a so-called integration step [65, 157]. Several algorithms exist for numerical integration, and a popular integrator because of its simplicity and numerical stability is e. g. the velocity Verlet integrator [158]. The velocity Verlet integrator is based on the over all concept, that the position, velocity, and acceleration (which is dependent of the potential energy function and therefore the force field) of an atom i at the current time point t are used to calculate the new positions and velocities. This step is repeated for a new time step $(t + \Delta t)$, until the defined number of time steps of the simulation is reached [65, 157].

After the simulation finished, i.e., all positions, velocities, and forces have been calculated for each atom at each time step, a trajectory file is obtained containing position and velocity vectors, describing the motion of the system [65, 157].

Analysis Due to the MD simulation's ergodicity assumption that the spatial average equals the temporal average, arithmetic averages are independent of initial conditions and the whole system relaxes towards a unique equilibrium [159]. For this reason, statistical methods can be used and an ensemble, which is a collection of different states that belong to a single thermodynamic state, can be replaced with a very long trajectory. For the analysis of MD simulation trajectory data, secondary properties can be calculated from the atomic coordinates, so that arithmetic averages over time of the MD trajectories can be calculated. Commonly determined values are thermodynamic properties (i.e., entropy, Gibbs free energies, enthalpie, binding energies, specific heats and others), kinetic properties (e. g. folding rates), structural changes (e. g. conformational changes, root mean squared deviation and fluctuation, radius of gyration, solvent accessible surface area, volume, and many others) and interactions between molecules (e. g. protein-ligand complex). A key step in the analysis pipeline is the visual inspection of the trajectory and 3D conformation, which is tedious and time-consuming due to the increase in computational power [66]. Today, MD simulation trajectories can easily result in several gigabytes or terabytes in size, making them difficult to visualise and analyse [66, 160].

2.2 Computer Science

2.2.1 Feature Space Reduction

The feature space describes the set of features that characterise data. In this work, data is for example molecules, which are characterised by a vector representation (so-called feature). The feature space describes in this case the chemical space, which is in this thesis defined as a region

of particular choices of descriptors (e. g. physicochemical properties, molecular representation, or fingerprints [2]) that characterise a molecule [161]. In this view, following the definition of Bajorath [162], it is a multidimensional descriptor space or N-dimensional Cartesian space. The exploration of chemical space is usually associated with a visual representation of the multidimensional space. In general, it is assumed that related molecules (either by structure, physicochemical properties or biological activity) form groups that are located closer together in chemical space and therefore form a map [162]. This map can, for example, be visualised and analysed after applying principal component analysis (PCA) [163] or other methods (e. g. SOM maps [164], MDS [165] or t-SNE [166]) to obtain a lower dimensional representation of the chemical space [167]. For analysis, feature space reduction is often of crucial importance and aims to reduce the number of features (i.e., dimensions) while retaining features that are variable in the data set. The reduced dimensions can then be used for visualisation, machine learning or other purposes. In order to visualise the feature space, it is necessary to reduce the data to a few dimensions, usually a 2D or 3D projection that allows us to get an idea of the grouping of molecular structures. In machine learning, feature reduction can help to obtain a classifier with improved performance. It was found that a classifier performs better on a reduced set of features, as often a subclass of features is already representative of the whole set of features [168, 169]. In addition, it reduces the problem of overfitting to the data in the training set and removes correlations and noise in the data [168]. In the following, two methods for dimensionality reduction will be described in closer detail. One is PCA, which is one of the major methods used and the other one is the (α, β)-k-Feature Set-Problem, which is based on optimisation.

2.2.1.1 Principal Component Analysis

Principal component analysis (PCA) [163], an unsupervised machine learning approach, is formally a linear dimensionality reduction method. PCA is widely used to visualise or process large, high-dimensional data sets, that are more and more common and difficult to interpret, and to reduce the dimensionality of a data set while minimising data loss. For this reason, a smaller number of uncorrelated variables (i.e., principal components) are calculated by optimising for the highest variance across correlated variables in the data [163]. The following is a high-level description of the individual steps carried out during PCA. A covariance matrix is computed to calculate all possible pairs of covariance in our data set, because the covariance can only be calculated between two data points. Then, the eigenvectors and eigenvalues of the covariance matrix are calculated. The eigenvectors are orthogonal to each other and form a line along the relation between data sets. Therefore, they provide information about the pattern in the data set. The next step is to select the components or eigenvectors to keep. To decide which eigenvectors or principal components are kept, they are sorted by their eigenvalues so that the eigenvector with the highest eigenvalue (or highest variance) is first (first principal component), followed by the other eigenvectors which capture less variance in the data and are therefore less important. Usually, the first few eigenvectors capture a high percentage of the variance. To reduce dimensionality, eigenvectors with very low eigenvalues can be removed, since they are usually not significant. Of course, this step leads to some loss of information in the data, which is usually negligible. If the original data set had n dimensions and x eigenvectors are removed, $n - x$ dimensions are left. The eigenvectors remaining after dropping the uninformative ones form the feature vector. All feature vectors are summarised in a matrix consisting of the kept eigenvectors. Lastly, the new data set along our principal components is derived to express the data in the number of selected dimensions. The kept dimensions are now the principal

components, which were projected along the eigenvectors that were kept [170].

In the context of molecular structures, PCA can be applied to a variety of different data sets or representations. For example, visualising the chemical space either with physicochemical properties or molecular fingerprints [162, 167, 171], or structural motion and therefore different conformations [172–174].

2.2.1.2 (α, β)-k-Feature Set-Problem

Another approach to reduce the dimensionality of the feature set, is the application of the (α, β)-k-Feature Set-Problem [175]. It is a combinatorial problem that helps in feature selection by optimising a set of features towards a minimum cost, maximising intra-class relationships (similarities) and inter-class discrimination information (differences). The objective function aims to minimise the number of features with the constraints of 1) maximising α nodes, meaning that a pair of samples with a different target feature differ in at least α features, and 2) maximising β nodes, meaning that a pair of samples with the same target feature share at least β features. This approach ensures discrimination while managing within-class similarity [176, 177].

A possible solution of the (α, β)-k-Feature Set-Problem is a feature set S of small cardinality or length k, containing the feature indices of the original feature set. If a feasible solution exists for a data set, then at least one ($\alpha > 0$) Boolean feature will help to discriminate between any pair of samples belonging to different classes, and at least zero ($\beta \geq 0$) Boolean features have the same value between any two samples of the same class [175].

The formal mathematical definition of the (α, β)-k-Feature Set-Problem by Cotta, Sloper and Moscato is as follows [175]:

(α, β)-k-Feature Set Problem

Instance: A set of m examples $X = \{x^{(1)}, \ldots, x^{(m)}\}$,
such that for all i, $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}, t^{(i)}\} \in \{0,1\}^{n+1}$, and three integers $k > 0$, and $\alpha, \beta \geq 0$.

Question: Does there exist an (α, β)-k-Feature Set S, $S \subseteq \{1, \cdots, n\}$, with $|S| \leq k$ and such that:

- for all pairs of examples $i \neq j$, if $t^{(i)} \neq t^{(j)}$ there exists $S' = S'(i,j) \subseteq S$ such that $|S'| \geq \alpha$ and for all $l \in S'$, $x_l^{(i)} \neq x_l^{(j)}$?
- for all pairs of examples $i \neq j$, if $t^{(i)} = t^{(j)}$ there exists $S' \subseteq S$ such that $|S'| \geq \beta$ and for all $l \in S'$, $x_l^{(i)} = x_l^{(j)}$?

where the set S' is not the same for all pairs of examples so we have written $S' = S'(i,j)$.

2.2.2 Specific Aspects of Machine Learning

In the following, specific aspects of machine learning are summarised to cover the aspects relevant for this thesis. Machine learning (ML) is a powerful technique to predict a target value, (e. g. physicochemical, molecular properties, bioactivity [178–181]), from a vector representation of molecules.

There are supervised and unsupervised machine learning approaches, which differ in their training corpus. For unsupervised machine learning approaches, there are no target values

associated with the training corpus; for supervised machine learning approaches, a feature vector is associated with a target value. If a target feature is continuous, the task is called regression, if a target feature is binary, the task is called classification. Feature encoding is a crucial step in machine learning to improve performance. If a selected feature is poor, the training set is too small, the machine learning algorithm selected or hyperparameters are inappropriate, the performance of supervised machine learning is poor [182]. There are many techniques on how to encode and represent molecular data to machine learning. Briefly, a molecule or protein needs to be represented as a numerical feature to a computer, which is described in detail in Section 2.1.2. To test the performance of a machine learning approach, different metrics such as the area under the ROC curve are used [183]. To calculate these values, a data set needs to be split in a training and a test set. Different methods are available for this approach. The most common ones are a split in training and test set, k-fold cross-validation, and leave-one-out validation. For a split in training and test set, the data set is usually split in a proportion of two third to one third, respectively. For k-fold cross-validation, the data set is split into k-folds. k-1 folds are used for training data set, and 1-fold for test data set. It is then rotated until every fold was once the test set, so building an ML model is repeated for k times. In the leave-one-out approach, all data points except for one are used for training, the one remaining for testing. The data point for testing is then also exchanged until every data point has been once the test set. This is a really accurate approach for estimating performance, but computationally expensive and therefore usually applied to smaller data sets [182].
Many different machine learning approaches have been developed (e. g. Linear Discriminant Analysis [184], Supported Vector Machines [185], or artificial neural networks [100]). Since in this work we have a small data set we use for machine learning, two machine learning approaches are of greater relevance, which will be described in the following. A subclass of machine learning approaches are decision trees, which can be used for classification and regression [186]. In the decision tree, the root node is on top and marks the starting point. With the feature representation at hand, one walks down the individual branches, and at each branch (or decision point), a position in the feature vector is evaluated for a certain value. Depending on the evaluation, the walking continues until we reach the leaves of the tree representing the target values, which a feature vector is then assigned to. The advantage of decision trees is that they are easy to interpret and follow, as the rules are simple [186]. The disadvantage is, that they are prone to overfitting to training data [182]. Random Forests (RF) were developed by Breiman [187] to reduce overfitting and improve performance, and were applied here as a machine learning algorithm. RF is basically a collection of trees, a so-called ensemble. The whole ensemble is trained with random subset of samples and random subset of features. Each tree will return one prediction, and the final decision or prediction is based on majority voting of all trees [187].

The second machine learning algorithm applied is XGBoost [188], which is a fast and parallel implementation of Gradient Boosting Machines (GBM) [189]. GBM is also an ensemble of trees, which complement each other. Instead of returning a single prediction value, the individual leaves are real scores which complement each other. During training, gradient tree boosting is applied, which minimises an objective function to measure model performance. There is a term for training loss to measure the prediction error, and regularisation to describe tree complexity to avoid overfitting. Instead of having a majority vote, the trees are added sequentially without changing existing trees. When a tree is added, it is parametrised to further reduce the loss by following the gradient, then an error is calculated and the weights of individual trees are updated to minimise the prediction error. For the final prediction, each predicted value of

each tree is summed up and an own weight for each individual tree added, depending on its accuracy [188, 189].

2.3 Microcystin as Use Case

Microcystins (MC) are released during cyanobacterial blooms worldwide [17] and are potent cyanotoxins [18, 19]. They are a family of structurally similar toxins and share their overall cyclic structure. Despite the fact that MC congeners share certain structural features, 279 structural variants have been identified with different side chains and structural features to date [15]. MC are especially problematic to consumption and intoxications [18] of cattle [190–192], dogs [193, 194] and other mammals [31, 195, 196] and occur frequently. Especially for human consumption, they are problematic and can lead to intoxication or in extreme cases death [23–25] by accidental or deliberate ingestion of contaminated water or food [19–21], inhalation of aerosolised toxins [22], or parenteral exposure during dialysis [23–25, 197]. Despite these drastic short-term effects, they are also suspected to cause long-term damage in multiple organs (i.e. liver, kidney, brain) and lead to liver cancer if exposure is frequent or persistent [198]. During an acute MC release in cyanobacterial blooms, access to drinking water or bathing water needs to be closed to prevent exposure to MC. These events occur frequently worldwide, also in a local bathing water close to Böhringen (Baden-Württemberg, Germany). To determine whether a shutdown is necessary, the Umweltbundesamt and the European guideline 2006/7/EG [199] suggests to first measure the number of cyanobacterial cells, and if a critical mass is reached, to detect the amount of toxins. To determine which MC congeners are present in blooms, liquid chromatography-mass spectrometry is used [200]. For risk assessment, up-to-date guidance documents of the World Health Organization (WHO) focuses solely on MC-LR and do not consider other congeners [26]. The provisional guideline value equals 0.001 mg MC-LR (free and cell bound) per litre water. The value was calculated in 1999 and is based on a single 90-day toxicity study in mice [27]. The tolerable daily intake (TDI) was estimated to be 0.04 mg per kg of body weight, multiplied by an uncertainty factor of 1000. The uncertainty factor was determined based on extrapolation from animal to human data, to take variability in the population into account, and factors like e. g. database limitations on chronic toxicity, carcinogenicity, and others. The guidance value (see Equation 2.2) was calculated based on an allocation factor (AF = 0.8) which is based on a percentage of exposure, the body weight of a standard adult (bw = 70 kg) and the total amount of resource (C = 2 l drinking water per day), resulting in 0.001 mg MC-LR per litre drinking water [18, 26].

$$GuidanceValue = \frac{TDI \cdot bw \cdot AF}{C} \tag{2.2}$$

Even though we know today about 279 MC congeners [15], this guidance value has not been updated and risk assessment is still based solely on MC-LR. Toxicological data of other MC congeners is neglected, so the provisional approach is to express other MC congeners as concentration equivalents of MC-LR by HPLC, or toxicity equivalents of MC-LR measured by bioassays [26]. Nevertheless, it was shown that MC-LR is not the most toxic MC congener when considering not only ser/thr-protein phosphatases (PPP) 1, but also PPP2A and PPP5 [16], and that hepatocyte donor cells have extremely different sensitivity to MC congeners [28]. In addition, other MC congeners can pass the blood-brain barrier [201] and only a few have been tested and analysed [16, 33, 37]. For this reason, an improved understanding of toxicokinet-

ics and -dynamics of different MC congeners is necessary to develop a more adequate risk management strategy to account for structural diversity of MC congeners.

2.3.1 Structure

MC congeners are structurally very similar cyclic peptides and share an overall structure of seven amino acids, of which some are L- and some are D-derivatives. The consensus sequence of MC congeners is cyclo-[D-Ala1]-[L-X2]-[β-D-MeAsp3]-[L-Z4]-[Adda5]-[-D-Glu6]-[Mdha7] (see Figure 2.1) [15].

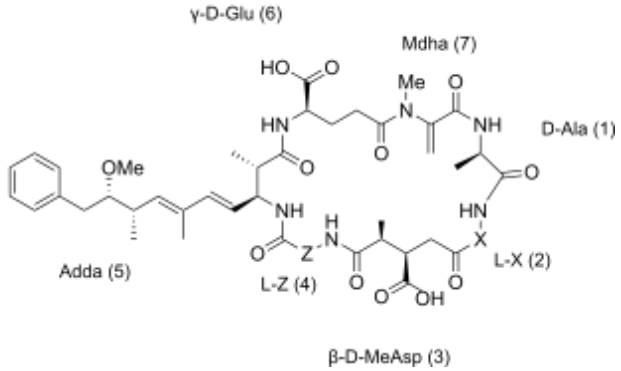


Figure 2.1: Consensus structure of MC congeners. Hyper-variable amino acids at position 2 and 4 are denoted as X and Z. The structure was created with ChemDraw JS [202].

The positions 1, 5 and 6 are highly conserved. The Adda residue (IUPAC name: `(2S,3S,8S,9S,4E,6E)-3-amino-9-methoxy-2,6,8-trimethyl-10-phenyl-4,6-decadienoicacid`) at position 5 has so far only been found in MC congeners and nodularin, and is present in nearly all MC congeners with minor modifications [15, 203]. Positions 2 and 4 are hyper-variable and are often referred to as X and Z, and are mostly occupied by standard L-amino acids. Different methylation patterns can occur at position 3 (β-D-MeAsp3, IUPAC name: β-D-methylaspartate) and position 7 (Mdha, IUPAC name: methyldehydroalanine), leading to variation in the overall structure. To describe the variable features in MC congener structures, a nomenclature was developed [204]. The one-letter code of amino acids at position 2 and 4 describe the most common source of variability and are added after MC, e. g. if leucine is at position 2 and arginine at position 4, the molecule is called MC-LR. Other modifications or alternations of the consensus structure are also included in the nomenclature and are placed in squared brackets before the previously discussed name. For example, if the molecule MC-LR had an aspartate at position 3 instead of methylaspartate, i.e., the amino acid is not methylated, the name of the modified MC congener would be [Asp3]MC-LR [203]. For a full review and overview of structural diversity of MC congeners, the reader is referred to a review by Bouaïcha and colleagues [15]. All MC congener structures relevant for this thesis are summarised in Figure 2.2.

2.3.2 Mechanisms of Toxicity

Toxicity is a complex process in the human body and involves different physiological steps in e. g. toxin uptake, distribution, and excretion (toxicokinetics) as well as effects on the cellular

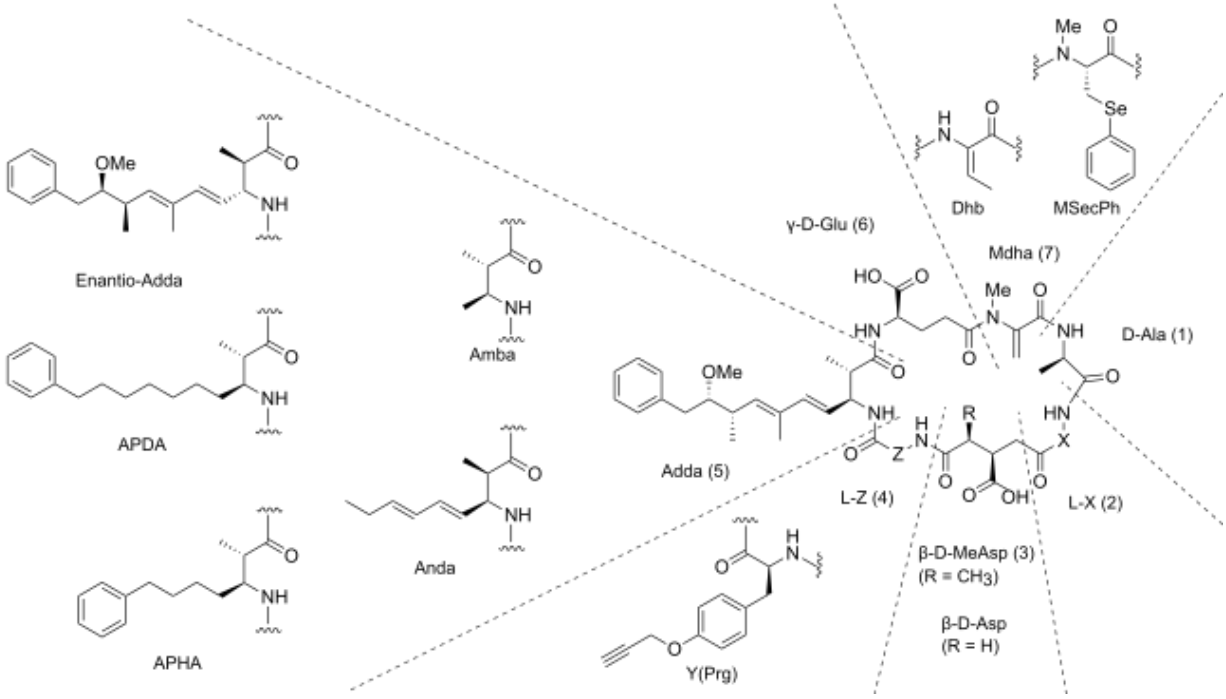


Figure 2.2: Summary of MC congener structures relevant for this thesis. Hypervariable amino acids at position 2 and 4 are denoted as X and Z. R is either a methyl group or hydrogen atom. The structure was created with ChemDraw JS [202].

level (toxicodynamics, see Figure 2.3). MC congener structure can impact toxicokinetics and -dynamics of MCs, which are defined by several protein-ligand interactions.
The uptake of MC is mostly by digestion. OATP transporters transport MCs into the blood of the portal vein, and from there into the liver cells [205]. Depending on the distribution of different types of OATP transporters on various organs, MC congeners are transported into the cells. OATP1B1 and OATP1B3 for instance are expressed in the liver and transport MC-LR [28, 29], in brain cells, OATP1A2 is required for the uptake of MC-LR [29, 206]. In addition to the different affinities of transporters to various MC congeners, the presence of other MC congeners can influence binding as well [28–30]. In contrast to the import, much less is known about the export mechanism of MCs. It was suggested that MCs are exported via MRP2, but other transporters are involved as well [36]. From the hepatocytes, MC congeners are exported to the systemic blood circulation and bile, where they can distribute in the whole body with a high residence time [36, 207] before excretion via the faeces and urine [208].

The toxicodynamics inside the cell (see Figure 2.3) involves reversible and irreversible binding to PPP1, PPP2A, PPP4, PPP5 and likely PPP6 [31–33, 209] (see Section 2.3.3), as well as detoxification via glutathione (GSH) pathway, which has the aim to detoxify substances in cells and reduces MC congener toxicity [34, 35]. Multiple different conjugates with MC congeners can be built (e. g. GSH, Cysteine (Cys), N-acetyl-Cys, and others), each having their own toxicity [210] and therefore adding a layer of complexity.

For binding of MC to PPP, the mechanism suggested is three partite: the first step is binding of the hydrophobic Adda chain to orient MC towards the binding pocket, second, double positively charged ions and charged parts of MC interact via strong electrostatics, and third, a covalent bond is formed between a thiogroup of PPP and an unsaturated carbon atom of MC [211–214]. Even though a covalent bond can be formed, inhibition of protein phosphatases

does not require one [32, 215, 216].

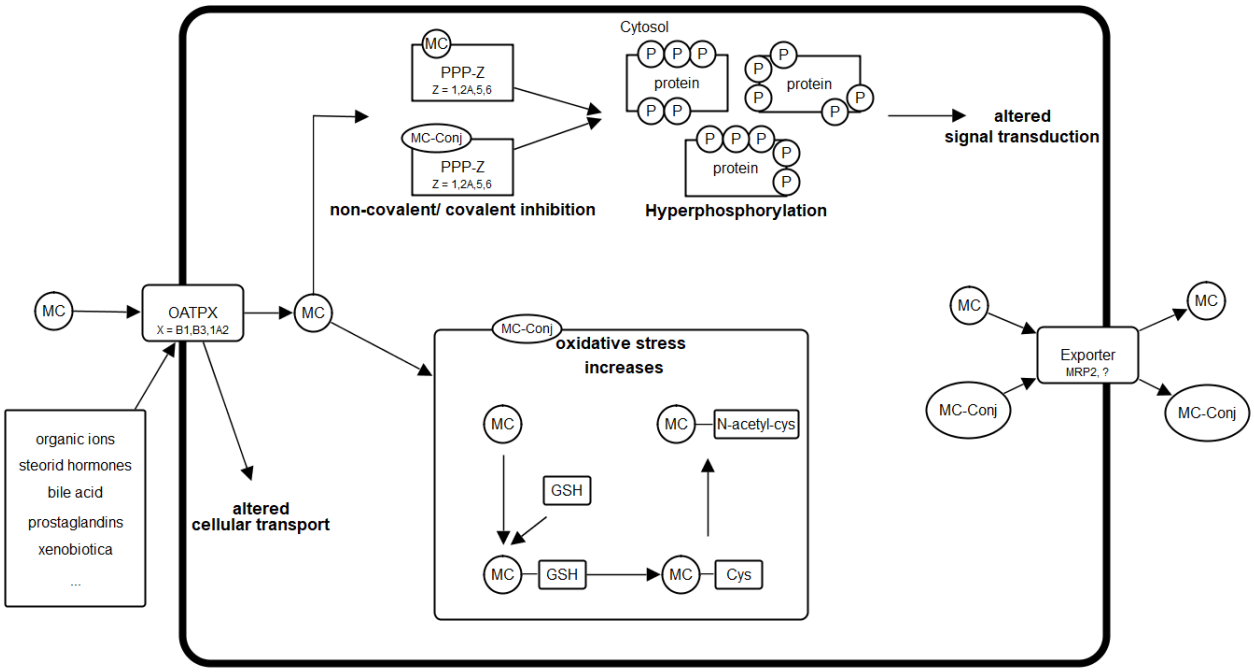


Figure 2.3: Overview of cellular mechanism triggered by Microcystins. Visualisation drawn with VANTED [217, 218].

2.3.3 Protein Phosphatases

Cell signalling is an important cellular process which involves transmitting, receiving and processing signals. Phosphorylation as part of cellular signalling acts as a molecular switch and is crucial for cellular homeostasis. Protein kinases and protein phosphatases are counterparts - kinases phosphorylate substrates while phosphatases dephosphorylate substrates. A wide range of processes in the cell are regulated through this mechanism, e. g. DNA replication, cell cycle progression, energy metabolism, cell differentiation and development [219, 220]. If the cellular homeostasis of phosphorylation is disturbed and an imbalance between kinases and phosphatases occurs, serious diseases such as cancer, diabetes, heart or neurodegenerative disease are the result [221].

In human cells, different residues are phosphorylated with different frequencies. Serines are phosphorylated most in 86.4 % of the cases, threonine in 11.2 % and tyrosines only in 1.8 % of the cases [222]. Another residue that can be potentially phosphorylated is histidine. In bacteria, fungi, and plants, histidine phosphorylation is well known. Recently, studies provided evidence that histidine phosphorylation also occurs in mammalian cells [223]. Contradictory to those findings is a study by Leijten et al. [224], who suspected that automated annotation and high-throughput workflows are overestimating the number of histidine phosphorylation and therefore the role in signalling in mammalian cells [224].

There are more than 500 genes known for human kinases [225], which phosphorylate substrates, and more than 150 genes known for serine/threonine phosphatases counteracting these kinases by dephosphorylating their substrates [226]. Serine/threonine phosphatases were found

to only have a few catalytic subunits, but many regulatory subunits, which modulate specificity [226]. Serine/threonine phosphatases can be further classified into three families based on their reaction mechanism, domain architecture and three-dimensional structures: the aspartate-based phosphatases (FCP/SCP) family, the metal-dependent protein phosphatases (PPM) family and phosphoprotein phosphatases (PPPs). Even though metal-dependent is in the family name of PPMs, the catalytic mechanism of PPPs also depends on metal ions [227].

This work focuses on the interaction between MC congeners and PPPs, and therefore on PPPs, which are the largest phosphatase family and responsible for 98.2 % of dephosphorylation events in human cells [228]. The enzymes have different functions and different substrate specificities with in the cell. PPPs form multi-subunit complexes and are made up of catalytic and multiple regulatory subunits, both of which determine substrate specificity [228]. They share a highly conserved active binding site with a binuclear metal centre. Their metal ions are coordinated between three histidines, two aspartates and one asparagines. In addition, water molecules bridge the two metal ions, which are then activated and initiate a nucleophilic attack on a phosphorus atom of a substrate to start dephosphorylation. Moreover, they have a strong preference for negatively charged substrates, which leads to strong binding [212, 227, 228].

MC congeners are potent inhibitors of the catalytic centre of PPPs i.e., PPP1, PPP2A, PPP4, PPP5 and likely PPP6 [31–33, 209], but do not inhibit PPP2B or PPP7 [209]. MC congeners bind by reversible and irreversible binding via a cysteine residue to PPPs [32, 211]. However, the irreversible bond is not essential for inhibition of PPPs by MC congeners [32, 215, 216]. Inhibition and therefore blocking of the catalytic site of PPPs leads to massively altered cellular function. Hyperphosphorylation of many phosphate-regulated enzymes leads to a dysregulated phospho-protein homeostasis and deregulation of major cellular processes [229–231] as PPPs have many important functions, which were summarised by Pereira et al. [232]:

- PPP1: Muscle relaxation, synaptic transmission, gene expression, glycogen metabolism, RNA splicing, cell-cycle progression
- PPP2: Cell-cycle regulation, cell growth control, cytoskeleton dynamics, cell mobility, metabolism, transcription, translation, RNA splicing, DNA replication, apoptosis, inflammation, differentiation
- PPP3: Response to neurotransmitters and nerve impulses, NMDA-receptor signalling pathway, T-lymphocyte activation
- PPP4: Centrosome maturation, microtubule organization, histone phosphorylation, apoptosis
- PPP5: Cell growth, ribosomal RNA transcription regulation, atrial natriuretic peptide (ANP) signalling, steroid signalling, blue-light signal transduction, ion channels regulation
- PPP6: Transcription, translation, morphogenesis, cell-cycle regulation
- PPP7: Control of phosphorylation status of G protein–coupled receptors

Chapter 3

Structural Analysis and Prediction of Microcystin Toxicodynamics

3.1 Summary

This chapter describes two different approaches to analyse and estimate the half maximal inhibitory concentration (IC_{50} value) of Microcystin (MC) on ser/thr-protein phosphatases (PPP). The largest available data set of MC IC_{50} values was determined using 18 MC congeners and three different PPPs (PPP1, PPP2A and PPP5). Depending on the PPPs/MC congener combinations that were tested, the IC_{50} values differed. One approach to investigate the MC-induced PPP inhibition computationally, is a machine learning (ML) model that we developed based on the 2D chemical structure of MC congeners and amino acid sequences of PPPs to classify MC congeners into three toxicity classes, resulting in a model with an overall 80-90 % correct prediction. The second approach to analyse this data set is a mathematical optimisation based on the (α, β)-k-Feature Set Problem, which is a dimensionality reduction technique. We identified biologically meaningful toxicity signatures by applying the (α, β)-k-Feature Set Problem to extended connectivity fingerprint (ECFP) representation of MC congeners. Boolean rules were then derived to classify toxicity, leading to insights about substructures associated with toxicity.

Both approaches use and analyse the data set published in the following paper, which is described in Section 3.2:

- Stefan Altaner*, **Sabrina Jaeger***, Regina Fotler, Ivan Zemskov, Valentin Wittman, Falk Schreiber and Daniel R. Dietrich: 'Machine learning prediction of cyanobacterial toxin (microcystin) toxicodynamics in humans' Alternatives to Animal Experimentation 37 (1): pp. 24-36, 2020. doi: 10.14573/altex.1904031.
  *: Equal contribution.

Additionally, a machine learning model was developed in this publication and is described in Section 3.3.

The second approach is described in detail in Section 3.4 and was submitted to the following journal:

- Pablo Moscato, **Sabrina Jaeger-Honz**, Mohammad N. Haque, and Falk Schreiber: 'The (α,β)-k Boolean signatures of molecular toxicity: Microcystin as a case study', bioRxiv, 2024. doi:10.1101/2024.12.29.630644.

Ideas, figures, and tables have been taken from the publications and text modules may be similar to a certain extent. For a statement of contributions, the reader is referred to Section 1.3.

3.2 Background

To predict bioactivity or molecular properties such as e. g. solubility or protein folding, machine learning is commonly employed in the fields of bio- and cheminformatics. Several methods to represent molecular data as n-dimensional feature vector, which captures molecular structure or structural information, exist. These methods include, but are not limited to, graph-based representations (e. g. [178, 181, 233, 234]), text-based representations (e. g. [89, 235]) or numerical representations (e. g. [104, 179, 236–242]) [243]. Even though machine learning methods have made significant progress over the last few years, it is still difficult to apply ML to pharmacology or toxicology, because data sets are often small and heterogeneous compared to those in other domains. Especially for small data sets, these approaches are challenging and usually result in low prediction quality [244].

A machine learning or predictive model for MC toxicodynamics (see Section 2.3.2) would be beneficial to improve current assessments of toxicity. Due to the lack of efficient techniques for synthesizing and purifying uncharacterised MC congeners in sufficient quantities required for in *vitro* or *in vivo* testing, research on MC toxicodynamics has primarily concentrated on a limited number of MC congeners (-LR, -RR, -LA, -LF) targeting PPP1 and PPP2A. Consequently, only a handful of MC congeners have undergone comprehensive testing and analysis, despite their significant variations in their ability to inhibit specific PPPs [16, 33, 37]. Given the growing number of known congeners (currently 279 [15]) and the challenges associated with their synthesis, *in-silico* approaches are imperative for further research and risk assessment to ensure the safety of human consumption [16].

Although *in-silico* methods analysing MC congener toxicity would be helpful, they have been mostly used to model various properties of water bodies, including cyanobacterial biomass or cell density [245, 246], MC occurrence [247, 248], MC or cyanotoxin concentrations [249–253], cyanotoxin mixture prediction for MC concentration [254], and cyanobacterial population dynamics combined with cyanotoxin or MC and geosmin concentrations [255, 256]. The methods used to predict and model these properties include a variety of different ML algorithms and approaches, such as multiple linear regression, cluster analysis, supported vector machines, random forests, boosted trees, Bayesian network modelling, cubist modelling, artificial neural networks or a combination of several algorithms [245–256].

Even though data on the toxicity of MC congeners is scarce, three prediction models have been developed. One prediction approach classifies 24 MC congeners using a multiple linear regression approach by representing the entire MC congener structure as average for feature representation. The aim of this model is to predict the median lethal dose (LD50) of MC congeners which was collected from various sources in the literature [257], which is often considered less reliable because experimental conditions cannot be controlled when data are collected from different sources and procedures may differ. In another prediction approach, an ANN model of the inhibition data (or curve) for two MC congeners on three different protein phosphatases (PPP1, PPP2A and mutant PPP2A) was developed to quantify the inhibition data of mixtures of two MC congeners [258]. The third prediction approach was to develop a prediction of adsorption, distribution, metabolism, elimination, and toxicity (ADMET) for four MC congeners with dedicated software to evaluate the effects of MC toxicity in different organisms such as rats and humans. There were significant differences between the models, indicating that rats

are more sensitive to MC congeners than humans [259].

All these approaches provide valuable information about MC toxicity, but the data sets are very small or heterogeneous. For this reason, our collaborators in the biology department of the University of Konstanz collected the largest available data set of MC congener inhibition values for 18 MC congeners on three different PPPs. The development of this data set and an initial toxicity model is described in Section 3.3, and a second model to investigate substructures related to toxicity is described in Section 3.4.

3.3 Machine Learning of Microcystin Toxicity

This section is based on our paper (Altaner et al. [16]) where the aim was to derive a data set of PPP inhibition capacity (toxicity) of 18 structurally diverse MC congeners and to develop an ML model to predict toxicity classes of MC congeners. Since part of this work was experimental and part computational, the experimental part, which was carried out in the Department of biology, is very briefly summarised in the introduction, as it is not the focus of this thesis, but the foundation of the computational work. For a detailed description of the experimental part, see the Materials and Methods section of the original publication [16].

3.3.1 Introduction

A comprehensive data set including inhibition of PPPs by MC congeners does not exist in the literature, as values are often distributed across a variety of publications and different methods, which makes the results difficult to compare as experimental procedures differ. For this reason, Altaner and colleagues [16] generated a comprehensive data set of PPP (PPP1, PPP2A, PPP5) inhibition (toxicodynamics) of 18 MC congeners (see Table 3.1 and Table 3.2), which is the largest data set available in literature so far.

MC congeners are potent inhibitors of the catalytic subunit of PPPs. The families of PPPs, which are described in Section 2.3.3 in detail, are structurally very similar and have a protein sequence homology of up to 65 % (Clustal Omega). However, despite this similarity, each PPP has defined substrate specificities and therefore different functions. PPPs are expressed ubiquitously, although certain organs may harbour specific PPPs. Consequently, the observed toxicity resulting from MC exposure is influenced not only by the result of MC toxicokinetics (i.e., distribution in different organs), but also by toxicodynamics (i.e., interaction with PPPs on a cellular level).

In this study, PPP inhibition was measured for PPP1, PPP2A and PPP5 in an *in vitro* colorimetric protein phosphatase inhibition assay and quantified by the IC_{50} value. The activity of the PPP has been calculated by subtracting the start value (0 h) from the end value (3 h). Then, this was compared to the 100 % activity and the IC_{50} value was calculated using GraphPad Prism 5 with a 5-PL non-linear regression (see Table 3.2). The anchorage points and constraints were set between 100 % and 0 %. The IC_{50} values were then used to fit a machine learning approach to classify MC congeners into different toxicity classes and to compare the results with the *in vitro* data.

3.3.2 Method

There are 18 MC congeners available in the data set of IC_{50} values for PPP1 and PPP5. For PPP2A, the data set is unfortunately not complete and only 11 MC congeners could be

Table 3.1: Structural information of different MCs [16].

Congener	X	Z	R_1	R_2	Adda
MC-RR	Arginine	Arginine	Methyl	C=CH_2	normal
MC-LR	Leucine	Arginine	Methyl	C=CH_2	normal
MC-YR	Tyrosine	Arginine	Methyl	C=CH_2	normal
MC-WR	Tryptophane	Arginine	Methyl	C=CH_2	normal
MC-LA	Leucine	Alanine	Methyl	C=CH_2	normal
MC-LY	Leucine	Tyrosine	Methyl	C=CH_2	normal
MC-LF	Leucine	Phenylalanine	Methyl	C=CH_2	normal
MC-LW	Leucine	Tryptophane	Methyl	C=CH_2	normal
MC-HilR	Homoisoleucine	Arginine	Methyl	C=CH_2	normal
MC-HtyR	Homotyrosine	Arginine	Methyl	C=CH_2	normal
[β-D-Asp3]-MC-RR	Arginine	Arginine	Hydrogen	C=CH_2	normal
[β-D-Asp3]-MC-LR	Leucine	Arginine	Hydrogen	C=CH_2	normal
[β-D-Asp3,Dhb7]-MC-RR	Arginine	Arginine	Hydrogen	C=C-CH_3	normal
MC-LY(Prg)	Leucine	Tyrosine with propargyl	Methyl	C=CH_2	normal
[Enantio-Adda5]-MC-LF	Leucine	Phenylalanine	Methyl	C=CH_2	enantiomeric
[Anda5]-MC-LY(Prg)	Leucine	Tyrosine with propargyl	Methyl	C=CH_2	(2S,3S,4E,6E)-3-amino-2-methylnona-4,6-dienoic acid
[MSecPh7]-MC-LY(Prg)	Leucine	Tyrosine with propargyl	Methyl	C-Se-Phenyl	normal
[Amba5]-MC-LY(Prg)	Leucine	Tyrosine with propargyl	Methyl	C=CH_2	(2S,3S)-3-amino-2-methylbutanoic acid

measured due to the discontinuation of PPP2A production by the manufacturer. Since the data set is small, the IC_{50} values have been converted to toxicity classes in order to build a model on this small set. MC congeners were classified into three classes: toxic (IC_{50} value ≤ 10 nM), less toxic (10 nM $< IC_{50}$ value $\leq 1,000$ nM) and non-toxic (IC_{50} value $> 1,000$ nM). The classification resulted in 31 toxic, 8 less toxic and 8 non-toxic MC congeners, which are 47 data points in total (see Table 3.3). The classification into different classes was based on relevance for human intoxication and guidelines by the World Health Organization WHO [18, 24, 260, 261], but were not based on any established thresholds.

For a prediction model, which in this case is a machine learning model, MC congeners and protein phosphatases have to be represented as numerical vectors to the computer. To encode molecular data as numerical vectors, several techniques are available. One method is Mol2vec [104] which is based on natural language processing (see Section 2.1.2.2). A pre-trained molecular model provided by Jaeger et al. [104] was used to generate features for MC congeners. This model is essentially a lookup table of vector representations of molecular substructures. The vectors are obtained by training a deep neural network on a database of molecular structures based on publicly available data and approximately 19.9 million compounds. The identifiers describing the substructures were retrieved during the generation of ECFPs with a radius equal to one (see Section 2.1.2.1). To obtain an order of a molecule which is necessary for the NLP representation, the identifiers are arranged according to the order in the canonical simplified molecular input line entry system (SMILES) (see Section 2.1.2.1). The two identifiers obtained for each substructure (radius 0 and radius 1) are placed next to each other to form an alternating sentence. The result is a pre-trained molecular model with vectors of 300 dimensions representing individual substructures. To represent an MC congener, substructure vectors are obtained from the lookup table and summed up to represent the whole molecule, giving a 300-dimensional representation.

Table 3.2: IC_{50} values of MC congeners with different protein phosphatases [16].

	IC_{50} [nM]		
Molecule	PPP1	PPP2A	PPP5
MC-RR	1.5	1.6	11.7
MC-LR	0.3	0.5	5.1
MC-YR	1.2	1.0	5.6
MC-WR	1.3	n.d.	5.1
MC-LA	2.0	1.4	4.7
MC-LY	0.8	n.d.	4.1
MC-LF	1.2	0.7	2.5
MC-LW	1.9	0.7	6.1
MC-HilR	0.6	n.d.	4.2
MC-HtyR	0.7	n.d.	4.7
[β-D-Asp3]-MC-RR	45.0	n.d.	167.1
[β-D-Asp3]-MC-LR	0.9	n.d.	10.2
[β-D-Asp3,Dhb7]-MC-RR	62.0	84.3	877.1
MC-LY(Prg)	1.7	0.4	1.7
[MSecPh7]-MC-LY(Prg)	1.9	0.9	18.2
[Enantio-Adda5]-MC-LF	x	x	x
[Amba5]-MC-LY(Prg)	520,817	2,135	54,063
[Anda5]-MC-LY(Prg)	1,724	n.d.	2,420

x: IC_{50} value is too high to be measured (inactive compound)

n.d.: not determined, PPP2A production discontinued by the manufacturer

Table 3.3: Classification of MC congeners in toxicity classes based on IC_{50} values. The number of data points is summed up over all PPPs tested in the original data set, as well as the number of data points after synthetic minority oversampling.

			Number of data points	
Class	IC_{50} [nM]	Description	original	after SMOTE
0	≤ 10	Toxic	31	31
1	> 10 and $\leq 1,000$	Less toxic	8	31
2	$> 1,000$	Non-toxic	8	31
		Total	47	93

Similarly, a vector representation of protein phosphatases was added to one model, as the IC_{50} values also depend on the PPP-MC congener interactions, even though some of them have high toxicity towards all PPPs. To represent them as a numerical vector of fixed size to a machine learning algorithm, ProtVec [103], which is based on the same principle as Mol2vec, was applied (see Section 2.1.2.2). In order to train a model on proteins, we had to collect a

corpus. We downloaded all protein sequences on UniProt [262, 263] (accessed: 18.04.2018), and proteins (or protein words) were encoded as 3-grams. To encode a protein as 3-gram, three amino acids were grouped to one 3-gram, and all possible 3-grams were generated by applying a sliding window over the protein sequence, resulting in three sentences per protein. The final model consists of 3-grams describing proteins in a lookup table with a 300-dimensional representation to have the same weight as the molecular vectors, as described in Jaeger et al. [104]. We then applied our generated model to encode PPPs as vectors, which were used for feature representation. The sequences of the PPPs were obtained from UniProt (UniProt ID: PPP1 (P62136), PPP2A (P67775) and PPP5 (P53041)) [262, 263].

For a schematic overview of the workflow for feature encoding and pre-processing of the data set, see Figure 3.1.

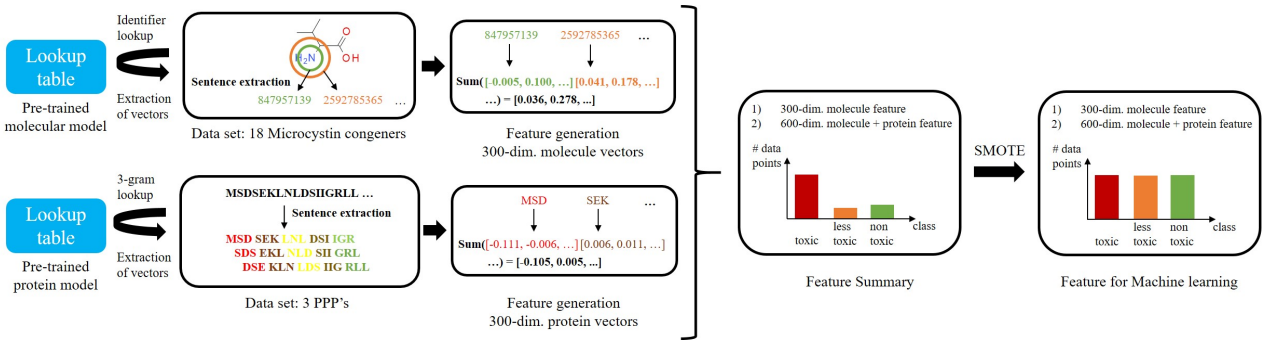


Figure 3.1: Schematic depiction of the workflow for feature encoding of MC congeners and PPPs. In addition, the pre-processing of the data set with oversampling and number of data points is visualised.

Random forests [187] (as implemented in scikit-learn v0.8.0 [264]) and extreme gradient boosting (XGBoost implementation v0.19.1 [188]) of Gradient Boosting Machines (GBM) classifier [189], were applied as machine learning approaches to the different combinations of features. The different combinations of features tested were Mol2vec, and Mol2vec+ProtVec. As the three classes of toxicity were highly imbalanced in favour of the toxic class, leading to poor performance of the ML algorithm for the other two classes, the Synthetic Minority Oversampling Technique (SMOTE [265], imbalanced learn v0.3.3 [266]) was applied with default settings and a ratio of 1.0. This resulted in 31 data points per toxicity class and a total of 93 data points (see Table 3.3). The data set was divided into training and test set using two different procedures and both were evaluated: 1) 80 % training and 20 % test set, and 2) a 5-fold cross-validation (CV). The first approach for data splitting results in one ML model, with an arbitrary split between training and test set. The second approach, results in five ML models with 1/5 of the data assigned to the test set, rotating iteratively until each data point has been in the test set once. After optimising the hyperparameters of the ML algorithms on one data split (see Table 3.4), each split was repeated 50 times to check the robustness of the prediction.

In the end, the best result was obtained with a consensus model of three different ML prediction models (Mol2vec with RF, Mol2vec with GBM and Mol2vec + ProtVec with RF), which was used for the final evaluation of the IC_{50} value prediction.

The evaluation of the different trained machine learning approaches was based on the confusion matrix and different evaluation metrics in scikit-learn (v 0.8.0 [264]). The confusion matrix compares the true label of a data point with the predicted label of a data point, in this

Table 3.4: The final set of hyperparameters for the different machine learning models.

		Tuned hyperparameters	
Feature	ML algorithm	Number of estimators	Maximum tree depth
Mol2vec + ProtVec	RF	50	5
Mol2vec	RF	50	5
Mol2vec	GBM	100	3

case it is a 3×3 matrix (3 classes). When a dark diagonal is obtained, the majority of data points are correctly predicted. This provides a good estimate of a model's performance and gives an overview of true positive/negative, as well as false positive/negative data points. True positives/negatives (TP/TN) are values that are correctly classified as positive/negative. False positives (FP) are data points that are classified as positive even though they are negative (so-called type 1 error), false negatives (FN) are data points that are classified as negative even though they are positive (so-called type 2 error). In toxicity estimation, false negative data points are a worse error than false positive values, because they underestimate toxicity. Evaluation metrics that can be used to calculate performance for a multi-class classification problem are precision (see Equation 3.1), recall (see Equation 3.2), and F1-score (see Equation 3.3), which are normalised values between 0 and 1. A high precision value is obtained, when a low number of false positives are observed, and a high recall is obtained, when a low number of false negative are observed. The F1-score is the harmonic mean of precision and recall — and thus reflects the balance between recall and precision [264].

$$Precision = \frac{TP}{TP + FP} \tag{3.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.2}$$

$$F1 - score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3.3}$$

For a schematic overview of the ML procedure and evaluation, see Figure 3.2.

3.3.3 Result and Discussion

To predict the IC_{50} value of MC congeners on PPPs, three different machine learning models (see Table 3.4) were built to predict three different toxicity classes (see Table 3.3), and a fourth one was derived by majority voting of the three trained ML models. The first one was a combination of Mol2vec and ProtVec with Random Forests, the second one Mol2vec with Random Forests and the third one Mol2vec with Extreme Gradient Boosting. A fourth one was built by combining all the trained models and forming a combined model, with majority voting deciding on the final classification. As the classes were highly imbalanced towards the toxic class and only a few data points were available for the less toxic and non-toxic class, oversampling was applied, resulting in 31 data points per class. The data was split into a training and a test set in two different approaches: 1) 80/20 split and 2) 5-fold CV. Each split and training was repeated 50 times to be able to estimate performance over a range of splits. To predict classes for each data point, a combined model was applied, which was a majority vote of all individual models built. For both data splits, high values were achieved for the

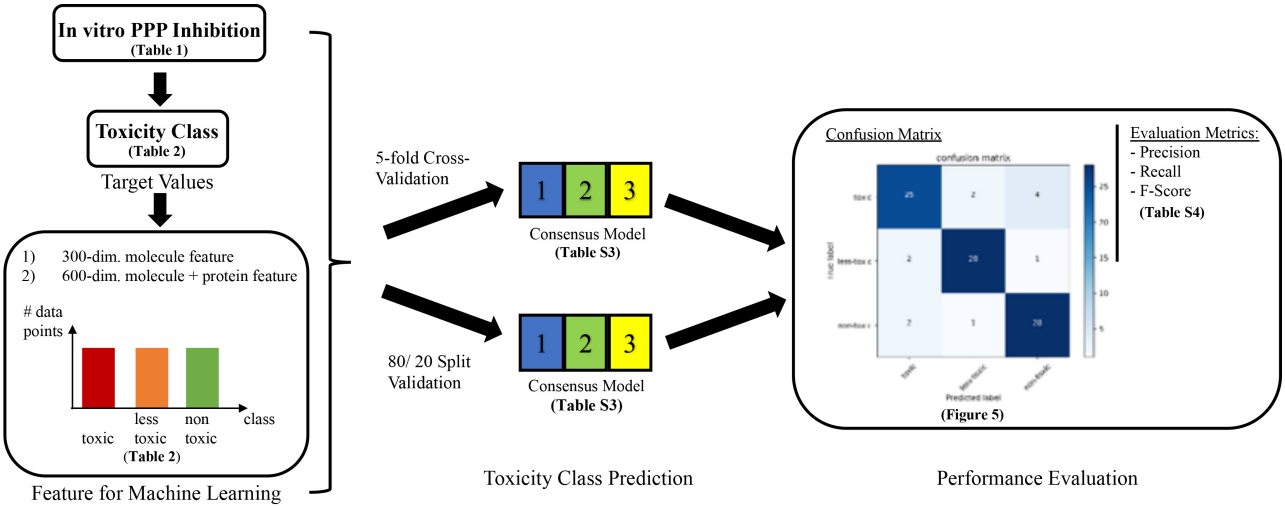


Figure 3.2: Schematic representation of Machine Learning workflow and evaluation of performance.

evaluation metrics, with precision scores above 0.80 and recall and F1-scores mostly above 0.80 (see Table 3.5 and SI Table 8.1). Cross-validation, i.e., 5-fold CV, is more appropriate for small data sets [267] and therefore the focus of this section will be on the CV results.

To assess the performance in more detail, the confusion matrix (see Figure 3.3A) was analysed. The majority of toxic data points (25 out of 31) were correctly classified. Out of the six that were incorrectly classified, two were predicted to be less toxic and four were predicted to be non-toxic. Three of the misclassified data points were the MC congener MC-LF, which was classified as non-toxic instead of toxic. In addition, the MC congener [Enantio-Adda5]-MC-LF, which is a stereoisomer of MC-LF, was classified as toxic in PPP2A, although it is non-toxic. A stereoisomer is a molecule with the same atoms, but different 3D orientation. The misclassification stems from the fact that the generation of the Mol2vec corpus (i.e., training vectors) did not take into account the chirality of the molecules. For this reason, MC-LF and its stereoisomer [Enantio-Adda5]-MC-LF are represented as the same vectors, even though they are toxic and non-toxic MC congeners, respectively. Depending on the training set, it is then placed in either the toxic or non-toxic class, leading to incorrect classification. For future work, the molecular corpus for Mol2vec training should be regenerated and a Mol2vec model retrained. Including stereoisomers would improve the representation of the individual substructures and therefore the representation of a molecule to a machine learning algorithm. This is currently a disadvantage of the method. However, the ability of Mol2vec to be combined with the protein representation of PPPs using ProtVec, presents an advantage over other methods. Furthermore, in contrast to standard representations such as ECFPs, Mol2vec is a dense vector representation, which means that instead of encoding mostly what is absent in the molecular structures, it encodes what is present. ECFPs are for example resulting in sparse vectors, as only a few substructures (usually 10 to 20) are mapped to a bit vector with usually 2048 dimension, resulting in many positions with absent (0) data. To validate this assumption, a second training run of the ML algorithm excluded [Enantio-Adda5]-MC-LF from the data set (see Figure 3.3B). After exclusion of [Enantio-Adda5]-MC-LF, MC-LF was correctly classified as well as the propargylated MC-LY variant, which is structurally similar to MC-LF.

In the toxic class, 30 out of 31 were correctly predicted. [β-D-Asp3]-MC-LR is still misclassified, but is now classified as less toxic rather than non-toxic. In addition, 28 out of 31

Table 3.5: Evaluation metrics for different machine learning models built. The 5-fold cross-validation data split and training was repeated 50 times and performance was evaluated. Mean and standard deviation are shown for precision, recall and F1-Score.

ML algorithm	Feature	Class	Precision	Recall	F1-Score
Combined model	-	non-toxic	0.81 ± 0.15	0.80 ± 0.16	0.79 ± 0.10
		less toxic	0.87 ± 0.11	0.87 ± 0.15	0.86 ± 0.10
		toxic	0.92 ± 0.10	0.87 ± 0.17	0.88 ± 0.10
RF	Mol2vec + ProtVec	non-toxic	0.80 ± 0.14	0.81 ± 0.16	0.79 ± 0.11
		less toxic	0.89 ± 0.11	0.90 ± 0.13	0.88 ± 0.09
		toxic	0.91 ± 0.11	0.83 ± 0.16	0.86 ± 0.11
RF	Mol2vec	non-toxic	0.80 ± 0.15	0.80 ± 0.16	0.78 ± 0.10
		less toxic	0.87 ± 0.11	0.85 ± 0.15	0.85 ± 0.10
		toxic	0.92 ± 0.10	0.87 ± 0.17	0.88 ± 0.11
XGB	Mol2vec	non-toxic	0.81 ± 0.14	0.78 ± 0.16	0.78 ± 0.11
		less toxic	0.86 ± 0.11	0.87 ± 0.14	0.85 ± 0.10
		toxic	0.92 ± 0.10	0.87 ± 0.15	0.88 ± 0.10

were correctly predicted for the less toxic class. [β-D-Asp3]-MC- LR (in PPP5) is not classified as non-toxic and therefore underestimated, but predicted to be toxic. Moreover, MC-RR in PPP5 is now predicted as toxic, leading to a new, incorrect classification. The non-toxic class has overall 29 out of 31 correctly predicted data points. Only (Asp3,Dhb7)MC-RR in PPP5 was assigned to the higher toxicity class less toxic, together with an artificial data point. In conclusion, removing the MC-LF stereoisomer ([Enantio-Adda5]-MC-LF) from the data set leads to a better toxicity classification performance with fewer false-negative predictions, which is important for toxicological predictions. Therefore, overestimating toxicity in comparison to true toxicity is not as critical in risk assessment, as underestimating toxicity. For this reason, the drawback of this ML model was considered acceptable, as one tries to be more cautious in assigning toxicity classes by overestimating potential risks and hazards. SMOTE provided a valid approach to sample data in such a limited data set as the one presented here. Nevertheless, SMOTE is an artificial procedure to increase the number of individual samples, which introduces uncertainties in the training data set and might also influence or cause misclassification. For this reason, it would be highly beneficial if the size of the training data set could be increased. There are several possibilities to increase the size of a data set with the aim of improving the overall model performance and prediction: collecting data from literature, conducting more experiments including rare MC variants, including other PPPs inhibitors, such as nodularins [203], structurally unrelated anabaenopeptins [268], or by collecting data from online databases such as ChEMBL. With more data points available i.e., inhibition values on PPPs, it may be possible to train deep neural networks, which were not utilised here, due to the small size of the data set, as they tend to overfit to training data and typically require thousands of data points. Nevertheless, the ML model and prediction developed here, can serve as a first estimation of toxicity for the more than 279 currently known MC congeners, since the performance of the model, despite all potential limitations and uncertainties, provided 80-90 % correct predictions of toxicity classes. In addition, the combination of our presented ML model with other methods, such as adverse outcome pathways or additional molecular modelling tech-

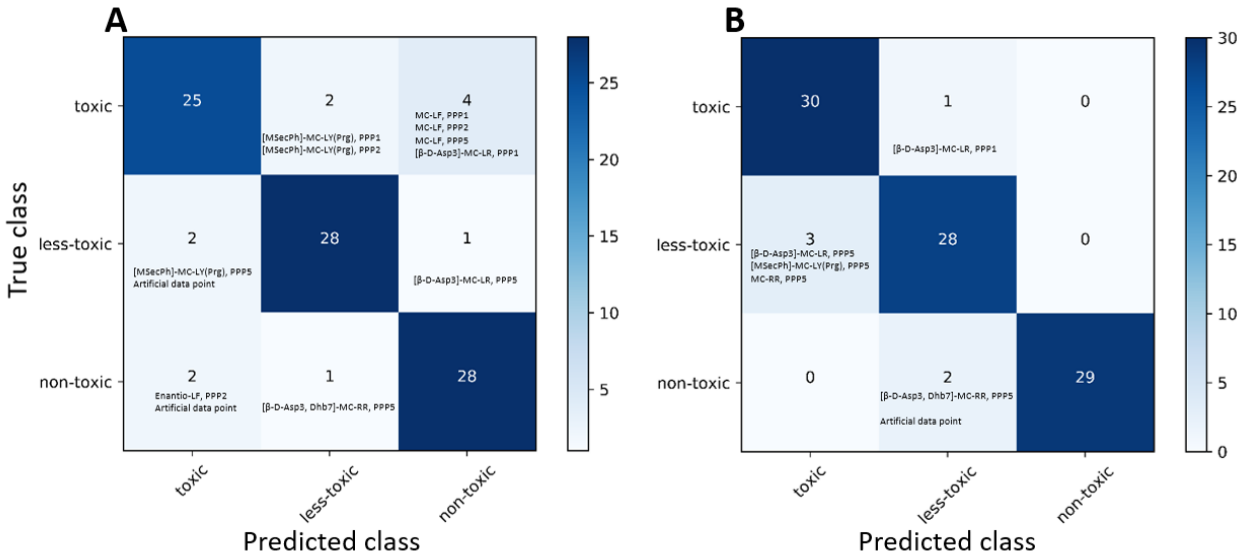


Figure 3.3: Example confusion matrix of microcystin toxicity prediction for 5-fold cross-validation. A) Whole data set used for training, B) stereoisomer [Enantio-Adda5]-MC-LF) was removed from the data set. The data points with wrong classification are labelled in the confusion matrix.

niques (e. g. Docking, Molecular Dynamics Simulation), may provide a reliable estimate for risk assessment of MC congener toxicity.

3.4 Boolean Signatures of Microcystin Toxicity

The approach presented in the previous section was an ML model which could predict toxicity reliable. Nevertheless, the model was a black box and we were not able to explain why MC congeners were classified in the different classes. This section is based on our paper (Moscato et al. [269]) which describes a feature reduction approach based on mathematical optimisation on our MC congener data set. By identifying substructures of MC congeners that are related to toxicity, Boolean rules can be derived to classify MC congeners into toxicity classes, which can be mapped back to molecular structure and therefore result in an explainable approach.

3.4.1 Introduction

Feature selection is crucial to machine learning performance or model building. However, when dealing with a large number of descriptors, interpreting the results of quantitative structure-activity relationship (QSAR) models or machine learning approaches can become challenging [270]. To address this issue, feature selection or reduction techniques can be employed to enhance model transparency and performance [271, 272]. This is a drawback of many methods, including simple ones, such as multiple linear regression, since it will calculate coefficients for all input variables if not restricted otherwise. Also, more complex methods, like artificial neural networks, suffer from this limitation, although they are routinely used in many research areas because of their impressive performance. Nevertheless, these models are often “black boxes”, hindering or even preventing the identification of the specific features that contribute

to a particular prediction. The challenge of understanding machine learning results arises from either considering all input features, many of which may be irrelevant, or using combinations of features that introduce higher-order complexities that hinder understanding. For this reason, "explainable artificial intelligence (AI)" techniques are becoming increasingly common, as they allow the user to understand the characteristics of a sample and deepen the knowledge to explain the result [273, 274]. In the research area of cheminformatics, this can lead to the discovery of interesting (sub-)structures, which can improve experimental design and structural understanding, or help to prioritise structures. One class of models, which aim at relating a set of variables to a target, are QSAR models. QSAR models are a combinatorial optimisation problem, which is rebuilt for each data set of interest, and has the goal to ensure explainability [275, 276]. To emphasise the importance of explainability in cheminformatics research, the Organisation for Economic Co-operation and Development (OECD) defined principles for the validating of QSAR models. These principles require: 1) defined endpoints, 2) an unambiguous algorithm, 3) a defined domain of application, 4) appropriate measures of goodness-of-fit, robustness and predictivity, and 5) if possible, a mechanistic interpretation [277].

In the context of cheminformatics, linking a chemical structure to a target outcome is often difficult [278], because chemical structures are often encoded in large and sparse matrices, as so-called features [240]. Here, a feature is defined as a binary characteristic of a sample, that can be either present or absent. This is a Boolean feature, which results in "True" (1) if it is present and "False" (0) if it is absent. A widely used numerical representation of molecular structures are circular fingerprints, more specifically extended connectivity fingerprints [96] (ECFP, see Section 2.1.2.1), where "1" indicates, that a substructure is present in a molecular structure, and a "0" that a substructure is absent. ECFPs are commonly used in many different application areas [279], to assess the similarity of molecules [94, 98], to cluster molecules [96], to predict molecular properties [180, 280], or to discover retrosynthetic routes [281]. The wide range of applications showed that ECFPs are a suitable representation for a variety of tasks. The ECFP is generated by encoding substructures of molecules to capture relevant information [96]. For this reason, ECFPs can also be used to identify a subset of features that explain an outcome of interest. This is a combinatorial optimisation problem known as the k-Feature Set [282].

Optimisation has been applied to classify molecular data, as e. g. the mixed integer optimisation model of the hyper-box framework [283, 284]. Multiple hyper-boxes are used and placed around data points that belong to the same class. Optimisation of variables and constraints are then applied to avoid overlapping of different classes and multi-class classification problems can be predicted. Even state-of-the-art classifiers such as logistic regression, support vector machines and neural networks could be outperformed by further developments of this framework [285]. It was used to generate a rule-based IF-THEN decision model for classification of CO_2 storage sites [286], or to derive deterministic rules for prediction of properties or potential molecular structures of fragrance molecules in an interpretable and transparent way [287]. Other optimisation approaches are mostly used for the design of molecular structures, which is reviewed in Austin et al. [288].

In this chapter, a mathematical optimisation method is presented which, for the first time, is applied directly to ECFPs of MC congeners to derive signatures or substructures relevant to toxicity. The mathematical optimisation is based on the (α, β)-k-Feature Set Problem [175], which is a generalisation of the k-Feature Set problem [282] with feasible solutions. The aim of this work was to reduce the number of features using an exact approach for dimensionality reduction, resulting in identification (α, β)-k feature sets. These are explainable positions in the ECFP or distinct substructures, which are used to predict the toxicity of MC congeners

towards PPPs. The detailed mathematical definition of this problem is given in Section 2.2.1.2. This approach provides insight into structural differences and similarities in our structure of interest. In contrast to our approach discussed previously in this chapter by Altaner et al. (see Section 3.3), the method described here yields interpretable structural characteristics that can provide insights into the chemical toxicity of a group of MC congeners. The (α, β)-k-Feature Set Problem was applied to Altaner and colleagues' data set (see Table 3.2), Boolean rules were derived to classify data points, and substructures were identified and compared to known relevant interaction sites of MC congeners in the literature (see Figure 3.4).

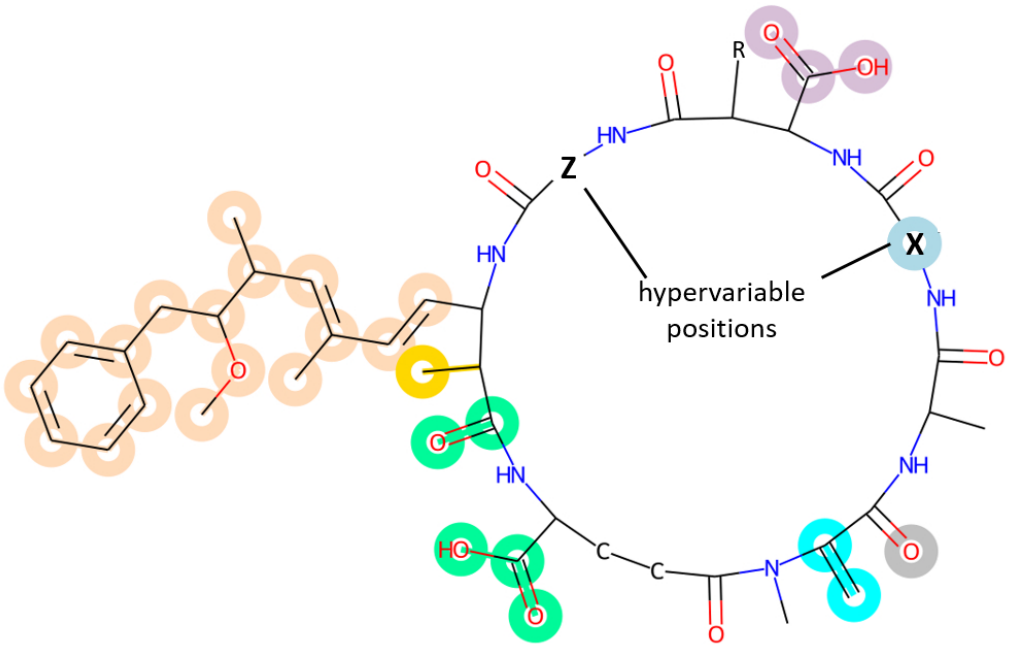


Figure 3.4: Overview of known interaction sites of MC congeners with PPP1 and PPP2A [289]. X and Z denote amino acids at hyper-variable positions 2 and 4. R is either a methyl group or hydrogen atom. Different colours indicate different types of interactions: Beige and light blue: hydrophobic interactions; gold: water molecule replacement; mint-green: indirect coordination to metals; cyan: covalent bonding; light purple: hydrogen bonds. Modified from Fontanillo and Köhn [216].

3.4.2 Method

The data set previously described in Section 3.3 and summarised in Table 3.2 was generated by Altaner et al. [16]. In this study, the same data set was used to understand which substructures in MC congeners contribute to toxicity, and to compare the mathematical optimisation approach with the original machine learning approach. The (α, β)-k-Feature Set Problem needs a binary classification label. For this reason, two classification labels have been chosen instead of the previous three — the less toxic and toxic class. All values with an IC50 above 40 nM are classified as less toxic, and those below as toxic. This results in 34 data points for the toxic class, and 13 for the less toxic class on three different protein phosphatases. Since the less toxic compounds are in the minority, [Enantio-Adda5]-MC-LF was removed from the training set for all three PPPs, and included in a validation set for evaluation of the performance. This

leaves 10 data points for the less toxic class. A validation set of unseen MC congener data was built to evaluate the performance of the Boolean rules and (α, β)-k-Feature Set-Set. The validation set was based on PPP1 inhibition, as this PPP had the most IC_{50} values available in literature. Five new values were found in the PubChem database [127], two in a publication by Fontanillo and Köhn [216] and one ([Enantio-Adda5]-MC-LF) was previously omitted from the Altaner et al. [16] set. The validation set is summarised in Table 3.6

Table 3.6: Validation data set of IC_{50} values and classification of MC congeners for PPP1.

Molecule	**IC_{50} (nM)**	**Classification**
Mol1 [216]	0.52	toxic
Mol2 [216]	253500	less toxic
44271302 [290]	0.8	toxic
44271308 [291]	0.8	toxic
44271410 [292]	0.52	toxic
44271411 [293]	0.3	toxic
44271325 [294]	3.0	toxic
[Enantio-Adda5]MC-LF [16]	x	less toxic

x: IC_{50} value is too high to be measured (inactive compound)

To encode the structure of an MC congener as a binary feature vector, extended-connectivity fingerprints were chosen. RDKit [295] was used to generate ECFPs as a bit array with a length of 2048 and a radius of 2, which results in a so-called ECFP4, where a 0 means a substructure is absent ("false") and a 1 means that a substructure is present ("true"), which is a Boolean feature. For further information on ECFPs see Section 2.1.2.1. This results in a sparse vector i.e., a sparse feature representation, as most substructures will be absent. To obtain a classifier with good performance, feature selection is crucial. It was found that a classifier performs better on a reduced set of features, as often a subclass of features is already representative of the whole set of features [271]. To reduce the dimensionality of the feature set, the (α, β)-k-Feature Set-Problem [175], a mathematical optimisation, was applied to the data set. To solve the (α, β)-k-Feature Set-Problem, an objective function is optimised towards a minimum cost set of features, to reduce the number of features. At the same time, different constraints are applied to maximise 1) the inter-class discriminative information, i.e., differences, by maximising α nodes, and 2) the intra-class relationships, i.e., similarities, by maximising β-nodes. Therefore, the approach discriminates between features of different target classes, while preserving within-class similarity [177]. If more than one solution was obtained for a problem, the one with the greatest coverage of sample pairs from different classes was preferred. The formal definition of the (α, β)-k-Feature Set-Problem is summarised in Section 2.2.1.2. In this work, an Integer Programming formulation of the (α, β)-k-Feature Set-Problem [176, 177] was used and the solution obtained by using IBM ILOG CPLEX Optimization Studio V12.8.0 (also referred to as CPLEX) [296].

To visualise and analyse the resulting feature set and Boolean rule, RDKit [295] was used to highlight bonds and atoms of the respective feature sets. Note that the radius of ECFP4 increases iteratively around a currently selected atom. Therefore, feature sets may have overlapping substructures. Inkscape [297] was used to arrange the molecular images.

3.4.3 Result and Discussion

3.4.3.1 Identification of (α, β)-k Feature Sets

Before running the solver to identify the (α, β)-k feature sets, equivalent features are grouped together, if they have the same value for all values in all samples, i.e., positions of the ECFP4 that are identical across all MC congeners. Of these values, only one representative feature is kept, together with the information which positions were combined. Thus, a representative feature is a set of features, which is defined as a set of positions in ECFPs, with identical value for all samples across all ECFP4 for all MC congeners. The features with the same value for both target values (toxic and less toxic) are removed, as they are irrelevant. After removing the duplicate feature on each set of PPPs, 25 representative features remain for PPP1, 17 for PPP2A and 25 for PPP5.

The data set is separated by PPP, resulting in three data sets, in order to develop a more precise approach targeted at each PPP. For each set of data, CPLEX is run to obtain Boolean signatures that solve the (α, β)-k-Feature Set Problem. The maximum number α is equal to two, i.e., there are two representative features for each data set (for each individual PPPs) to discriminate between at least one pair of samples with different values in each target class. Although the number is the same in each set, the individual pair may be different. These representative features are so-called feature sets, which are referred to below in capital letters, to distinguish them from individual feature positions (lower case), which group together a number of equivalent features. Once the maximum value of $\alpha = 2$ has been identified, CPLEX searches for a solution with a minimum value of $k > 0$, which means a small set of feature sets, with the constraint of $\beta \geq 0$. k, the number of identified feature sets, was determined to be $k = 7$, $k = 4$ and $k = 7$ for PPP1, PPP2A, and PPP5, respectively. Next, $\beta \geq 0$ was maximised resulting in $\beta = 1$, $\beta = 0$ and $\beta = 1$ for PPP1, PPP2A and PPP5, respectively. Thus, for PPP1 and PPP5 samples with the same target value or toxicity class have at least one feature in common, whereas for PPP2A they have zero features in common, implying lower within-class similarity. The Boolean signatures are summarised in Table 3.7 and the equivalent features, corresponding to positions in ECFP4, are summarised for each set of data in the Supplementary Information Section 8.1.1.

Table 3.7: Summary of the parameters and feature sets obtained after solving the (α, β)-k-Feature Set for each of the toxicity discrimination tasks on different PPPs.

Protein	α	β	k	Boolean signature of toxicity
PPP1	2	1	7	{F32, F130, F232, F295, F336, F695, F1346}
PPP2A	2	0	4	{F32, F130, F232, F773}
PPP5	2	1	7	{F32, F130, F232, F295, F336, F773, F1346}

The identified Boolean signatures can then be used to derive Boolean rules for classifying individual MC congeners, or as input to other ML approaches, as we now have a simpler feature representation focusing on relevant positions. The following sections discuss the individual Boolean signatures or identified substructures for each data set describing MC toxicity, the Boolean rules derived to classify toxicity, and an application of this method to a new data set.

Table 3.8: Optimal feature set on PPP2A toxicity with $\alpha = 2$, $\beta = 0$, and $k = 4$. An upper case 'F' refers to a set of equivalent features.

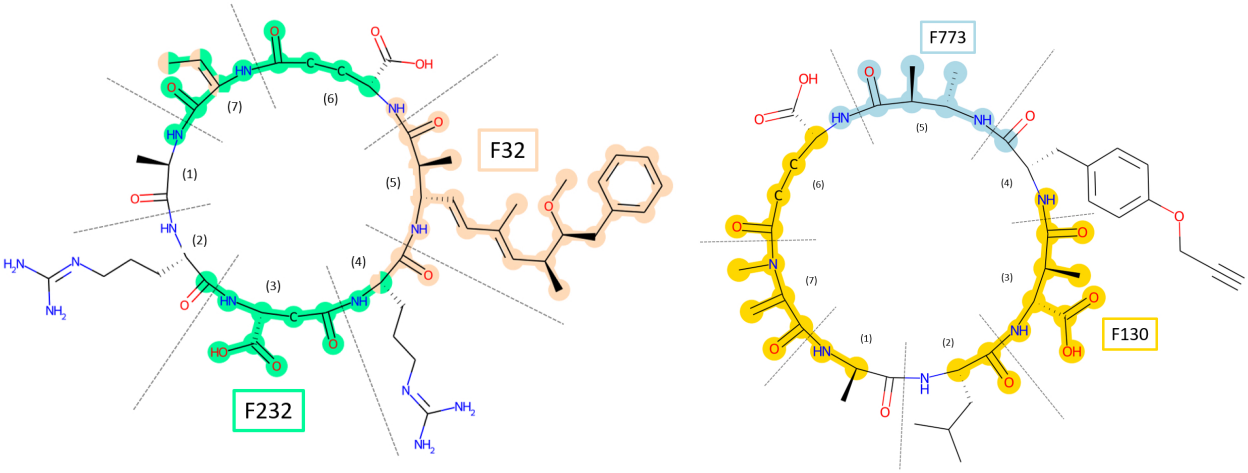
Molecule	IC_{50} [nM]	F32	F130	F232	F773
MC-RR	1.6	1	1	0	0
MC-LR	0.5	1	1	0	0
MC-YR	1	1	1	0	0
MC-LW	0.7	1	1	0	0
MC-LA	1.4	1	1	0	0
MC-LF	0.7	1	1	0	0
MC-LY(Prg)	0.4	1	1	0	0
[MSecPh7]-MC-LY(Prg)	0.9	1	1	0	0
[β-D-Asp3,Dhb7]-MC-RR	84.3	1	0	1	0
[Amba5]-MC-LY(Prg)	2135	0	1	0	1

3.4.3.2 Boolean Signature for PPP2A

The Boolean signature of toxicity obtained for PPP2A is summarised in Table 3.8. Boolean rules can be derived from the feature sets in the table to classify the toxicity of MC congeners to PPP2A. To understand which substructures are related to toxicity and influence classification, the Boolean rule is kept as simple as possible and the minimum number of features to separate the classes are chosen. To avoid a bias in the selection, the person who selected the Boolean rules was a computer scientist with no knowledge of the biological interaction of MC and PPPs. Therefore, one simple rule (out of many possible ones) for classification is as follows: *"If F32 and F130 are both 'True' (i.e., present in the structure), and neither F232 nor F773 are 'True', then the molecule is toxic for PPP2A"* and vice versa.

The individual feature sets in the toxicity signatures can be mapped back to the structure of MC congeners (see Figure 3.5) by assigning the individual position in the ECFP back to the molecular structure. The visualisation on molecular structure helps to generate knowledge about substructures related to toxicity. In order to analyse the importance of the identified feature sets, they are compared with important substructures identified in a structure-activity relationship study that is part of a review by Fontanillo and Köhn [216]. Feature sets F32 and F232 are shown on [β-D-Asp3, Dhb7]-MC-RR in Figure 3.5a and feature sets F130 and F773 on [Amba5]-MC-LY(Prg) in Figure 3.5b. Both MC congeners are classified as less toxic based on their IC_{50} value, i.e., the same class to which the Boolean rule assigns them. Comparing the feature sets with the substructures of MC congeners known to be important for interaction (Figure 3.4), the respective feature sets are less defined and more atoms are involved compared to the study by Fontanillo and Köhn. As only ten MC congeners were available in the training set, it is difficult to identify relevant substructures, so it is expected that the method might capture less meaningful signatures, simply because the training set is so small. In comparison, the feature sets of 17 MC congeners on PPP1 and PPP5 resulted in smaller feature sets, which are more defined and contain fewer atoms, except for F32. Here, F32 covers residue 5 (Adda), which was found to be important for binding to PPP1 and PPP2A. Adda is not toxic itself [298], but crucial for activity against PPP2A [299] and a methyl group reduces the activity to submicromolar activity, which is essentially a non-toxic compound [300]. Our data set is consistent with this observation, as e. g. the MC congener (see Figure 3.5b) with a methyl group (F773) instead of an Adda residue (F32) is classified as less toxic based on its IC_{50} value. Both,

F32 and F773 cover an additional methyl group attached to the backbone of MC congeners, which is important to displace water molecules in the binding site of PPPs (see Figure 3.4, gold colour). Nevertheless, it is difficult to infer activity from this methyl group, as the presence or absence of Adda seems to have a stronger effect.



(a) [β-D-Asp3, Dhb7]-MC-RR (b) [Amba5]-MC-LY(Prg)

Figure 3.5: Feature sets of PPP2A highlighted on two MC congener structures. F32 (beige) covers the Adda side chain and contributes to toxicity. F773 (light blue) has Adda replaced by a methyl group, which leads to decreased toxicity. F130 (gold) and F232 (mint-green) covering the carboxyl group at residue three is not fully understood regarding toxicity and studies showed different results. F130 (gold) covering the methyl group at residue three might block the interaction. F232 (mint-green) covers the atoms that would be responsible for the covalent bond at residue 7 which is not possible due to modification, but not essential for toxicity.

F232 and F130 are even less specific than F32 and F773. Both are distributed across the backbone of the MC congener structures and contain several important functional groups for different types of interactions, but are not really specific. These feature sets include atoms forming hydrogen bonds with PPP2A (see Figure 3.4, silver and light purple), and atoms forming a covalent bond (see Figure 3.4, cyan colour). Noteworthy, a hydrogen bond specific to the interaction with PPP2A and the carboxylic acid oxygen at residue seven of MC congeners is known (see Figure 3.4, silver) and is captured in F232 and F130. In the other data sets based on PPP1 and PPP5 this interaction is excluded and not identified. This finding highlights the biological significance of our feature sets based on molecular fingerprints. The carboxylic acid of β-D-MeAsp (residue three) was reported to interact with PPP2A and PPP1 via a hydrogen bond, but its role is not fully understood and investigated so far [33, 216]. In addition, results in literature are contradictory. For further discussion of residue three, the reader is referred to the following paragraphs, where the approach is analysed on PPP1, as there is more data available in the PPP1 set. The covalent bond between MC congener and PPPs does not affect toxicity, although it is important for irreversible binding, but MC congeners can also bind reversibly to PPPs [31, 211]. As it does not affect toxicity per se, it will not be further considered.

To summarise the important interactions of MC congeners with PPP2A, the Boolean rule defining MC congeners as toxic if F32 and F130 are present (i.e., true) and if F232 and F773 are absent (i.e., false) is consistent with the interactions described in literature [216]. The important Adda residue which results in inactivity when missing is represented by F773. F232 and F130 cover the same substructures, except for the substructure responsible for the covalent bond and

a methyl group attached to residue three. The methyl group in residue three present in F232 might be able to block a hydrogen bond formed with a nearby carbonyl group, thus influencing toxicity and binding. However, in previous studies, it was observed that a hydrophobic residue at position four affects the interaction of MC congeners with PPP2A and results in reduced toxicity [301]. This interaction or phenomenon was not found for our data set, described and discussed in the previous section. The small size of 11 MC congeners most likely makes it difficult to capture the differences with our data set.

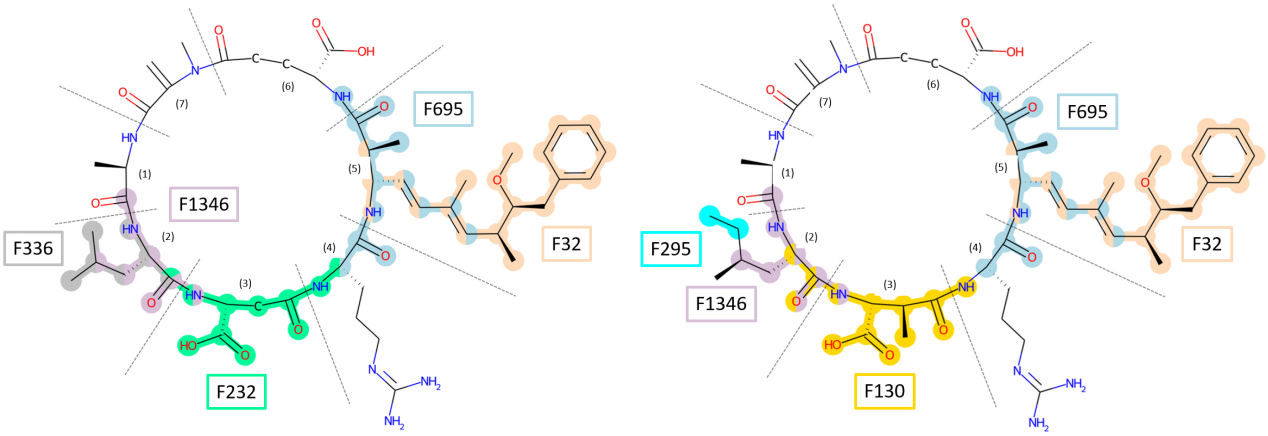
3.4.3.3 Boolean signature for PPP1

The feature sets identified on the PPP1 data set are summarised in Table 3.9. Several Boolean rules can be derived, and again a minimal set of features is chosen, even though the choice of the exact set is arbitrary. The Boolean rule derived to classify toxicity as follows: *"If F32 is True, and F130 or F336 are True (i.e., either of those or both of them), then the molecule is toxic for PPP1, otherwise it is less toxic."*

Table 3.9: Optimal feature set on PPP1 toxicity with $\alpha = 2$, $\beta = 1$, and $k = 7$. An upper case 'F' refers to a set of equivalent features.

Molecule	IC_{50} [nM]	F32	F695	F130	F1346	F336	F295	F232
MC-RR	1.5	1	1	1	0	0	0	0
MC-LR	0.3	1	1	1	1	1	0	0
MC-WR	1.3	1	1	1	0	0	0	0
MC-YR	1.2	1	1	1	0	0	0	0
MC-LW	1.9	1	1	1	1	1	0	0
MC-LY	0.8	1	1	1	1	1	0	0
MC-LA	2	1	1	1	1	1	0	0
MC-LF	1.2	1	1	1	1	1	0	0
MC-HilR	0.6	1	1	1	1	0	1	0
MC-HtyR	0.7	1	1	1	0	0	0	0
[β-D-Asp3]-MC-LR	0.9	1	1	0	1	1	0	1
MC-LY(Prg)	1.7	1	1	1	1	1	0	0
[MSecPh7]-MC-LY(Prg)	1.9	1	1	1	1	1	0	0
[β-D-Asp3]-MC-RR	45	1	1	0	0	0	0	1
[β-D-Asp3, Dhb7]-MC-RR	62	1	1	0	0	0	0	1
[Amba5]-MC-LY(Prg)	520817	0	0	1	1	1	0	0
[Anda5]-MC-LY(Prg)	1724	0	1	1	1	1	1	0

The individual feature sets and corresponding substructures for the PPP1 data set are mapped onto the structure of MC congeners in Figure 3.6. The derived feature sets and their biological relevance are also compared to the data by Fontanillo and Köhn (see Figure 3.4) [216]. Feature sets F336 and F232 are specific to [β-D-Asp3]-MC-LR (see Figure 3.6a), F130 and F295 are specific to MC-HilR (see Figure 3.6b), while F32, F695, and F1346 are common to both MC congeners. Based on their IC_{50} values in Table 3.9, both MC congeners are toxic. The feature sets derived for PPP1 are, in contrast to the feature sets of PPP2A, more defined and less distributed across the MC structures, but still less specific compared to the detailed structure activity relationship study presented by Fontanillo and Köhn [216].



(a) [β-D-Asp3]MC-LR (b) MC-HilR

Figure 3.6: Feature sets of PPP1 highlighted on two MC structures. F32 (beige) covers Adda side chain and contributes to toxicity. F130 (gold) and F232 (mint-green) covering the carboxyl group at residue three is not fully understood regarding toxicity and studies showed different results. The data set used in this study showed reduced toxicity to PPP1 when this residue is demethylated. F336 (silver) contributes to toxicity, since it covers a hydrophobic side chain at residue two. F1346 (light purple) and F295 (cyan) also partly cover this residue. F695 (light blue) covers the atoms responsible for water replacement and coordination to metal ions. Both interactions contribute to binding and toxicity.

Also for PPP1, feature set F32 covers the side chain of the Adda residue which is important for binding but not toxic itself [298], as was already seen and discussed in the previous section for PPP2A.

β-D-Me-Asp and β-D-Asp at residue three are represented by F130 (gold) and F232 (mint-green, see Figure 3.6), respectively. Feature set F130, which is the methylated residue, contributes in our Boolean rule to the toxicity that we observe in this data set. According to the Boolean rule, F130 (gold) or F336 (silver) must be true (i.e., substructure must be present) and F32 must also be true as well, to classify an MC congener as toxic. The role of the carboxyl group at residue three (F130 and F232) is not fully understood, as discussed previously. However, it is important for hydrogen bonding with PPP1 and PPP2A. Demethylation (i.e., β-D-Asp, F232) showed different results in literature. Some studies found negative effects for binding towards PPP1, others showed no effect [33, 302]. In the data set considered here (see Table 3.9), the IC_{50} values for both MC congeners are almost identical when LR is at positions 2 and 4: 0.90 nM for [β-D-Asp3]MC-LR and 0.30 nM for MC-LR. [β-D-Asp3]MC-LR belongs to the toxic class, even though it is demethylated at position 3. Therefore, F336, which covers the hydrophobic side chain of leucine at residue two, probably has a strong effect on toxicity/binding to PPP1. Even though, no effect of residue two on binding was observed in the QSAR study [216], it has been reported in literature that lipophilicity (or hydrophobicity) at residue two and a hydrophilic amino acid at residue four (which is not covered by F336) correlate with higher inhibitory potency. However, it has also been reported that residue four had a stronger influence on binding to PPPs than position 2 [303], which is not covered by our feature sets. This is probably a bias in the data set, as most of the MC congeners varying at residue four are in the toxic class. It is likely that the model did not learn to include residues at position 4, as our method optimises for similarities within the same class. To summarise, demethylation of residue three had no effect on toxicity or binding for MC-LR compared to [β-D-Asp3]MC-LR.

When comparing MC congeners with MC-RR at positions 2 and 4, demethylation results in reduced toxicity towards PPP1 and large differences are observed. MC-RR is toxic with an IC_{50} value of 1.5 nM, while [β-D-Asp3]-MC-RR is less toxic with an IC_{50} value of 45 nM. Thus, it is likely that not only demethylation at residue three influences binding to PPP1, but that modification of residue two must be present in order to see a change in the IC_{50} value. Comparing MC-LR with MC-RR, arginine (R) is a positively charged residue at residue two in contrast to Leucine (L), which is hydrophobic. Perhaps the intermolecular interactions change when residue three is demethylated and residue two is a positively charged amino acid, since MC-RR itself is almost as toxic as MC-LR is. Therefore, not only a positively charged amino acid at residue two, but also demethylation of β-D-MeAsp3 at residue three, is probably necessary to have a lower inhibitory potency and therefore a less toxic compound.

Contrary to the feature sets derived from the data set with PPP2A, the MC congener atom involved in covalent binding is not identified in our feature sets derived from the data set with PPP1. This is considered uncritical, because the covalent bond is not required for toxicity and not a key interactor for inhibition [31, 211].

3.4.3.4 Boolean signature for PPP5

For the optimisation on PPP5 data set, seven sets of features exist (see Table 3.10). The same Boolean rule we identified for PPP1, is also valid for PPP5: *"If F32 is True, and F130 or F336 are True (i.e., either of those or both of them), then the molecule is toxic for PPP1, otherwise it is less toxic."*

Table 3.10: Optimal feature sets on PPP5 toxicity with $\alpha = 2$, $k = 7$, and $\beta = 1$. An upper case 'F' refers to a set of equivalent features.

Molecule	IC_{50} [nM]	F32	F130	F1346	F336	F295	F232	F773
MC-RR	11.7	1	1	0	0	0	0	0
MC-LR	5.1	1	1	1	1	0	0	0
MC-WR	5.1	1	1	0	0	0	0	0
MC-YR	5.6	1	1	0	0	0	0	0
MC-LW	6.1	1	1	1	1	0	0	0
MC-LY	4.1	1	1	1	1	0	0	0
MC-LA	4.7	1	1	1	1	0	0	0
MC-LF	2.5	1	1	1	1	0	0	0
MC-HilR	4.2	1	1	1	0	1	0	0
MC-HtyR	4.7	1	1	0	0	0	0	0
[β-D-Asp3]-MC-LR	10.2	1	0	1	1	0	1	0
MC-LY(Prg)	1.7	1	1	1	1	0	0	0
[MSecPh7]-MC-LY(Prg)	18.2	1	1	1	1	0	0	0
[β-D-Asp3]-MC-RR	167.1	1	0	0	0	0	1	0
[β-D-Asp3, Dhb7]-MC-RR	877.1	1	0	0	0	0	1	0
[Amba5]-MC-LY(Prg)	54063	0	1	1	1	0	0	1
[Anda5]-MC-LY(Prg)	2420	0	1	1	1	1	0	0

Compared to PPP1 and PPP2A, PPP5 is less studied and there is less data available in the literature for comparison. The percent identity after alignment of the protein sequence of the catalytic subunit of PPP5 compared to PPP1 and PPP2A with Protein Blast [304], results in 44.40 % percent identity with PPP1 and 41.02 % percent identity with PPP2A. The definition

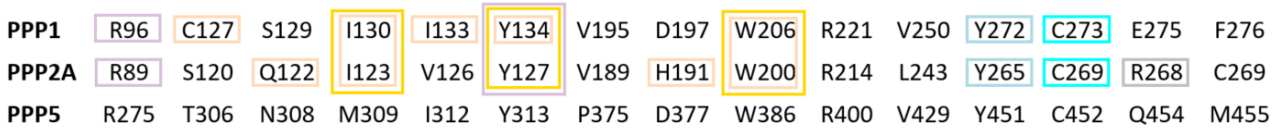


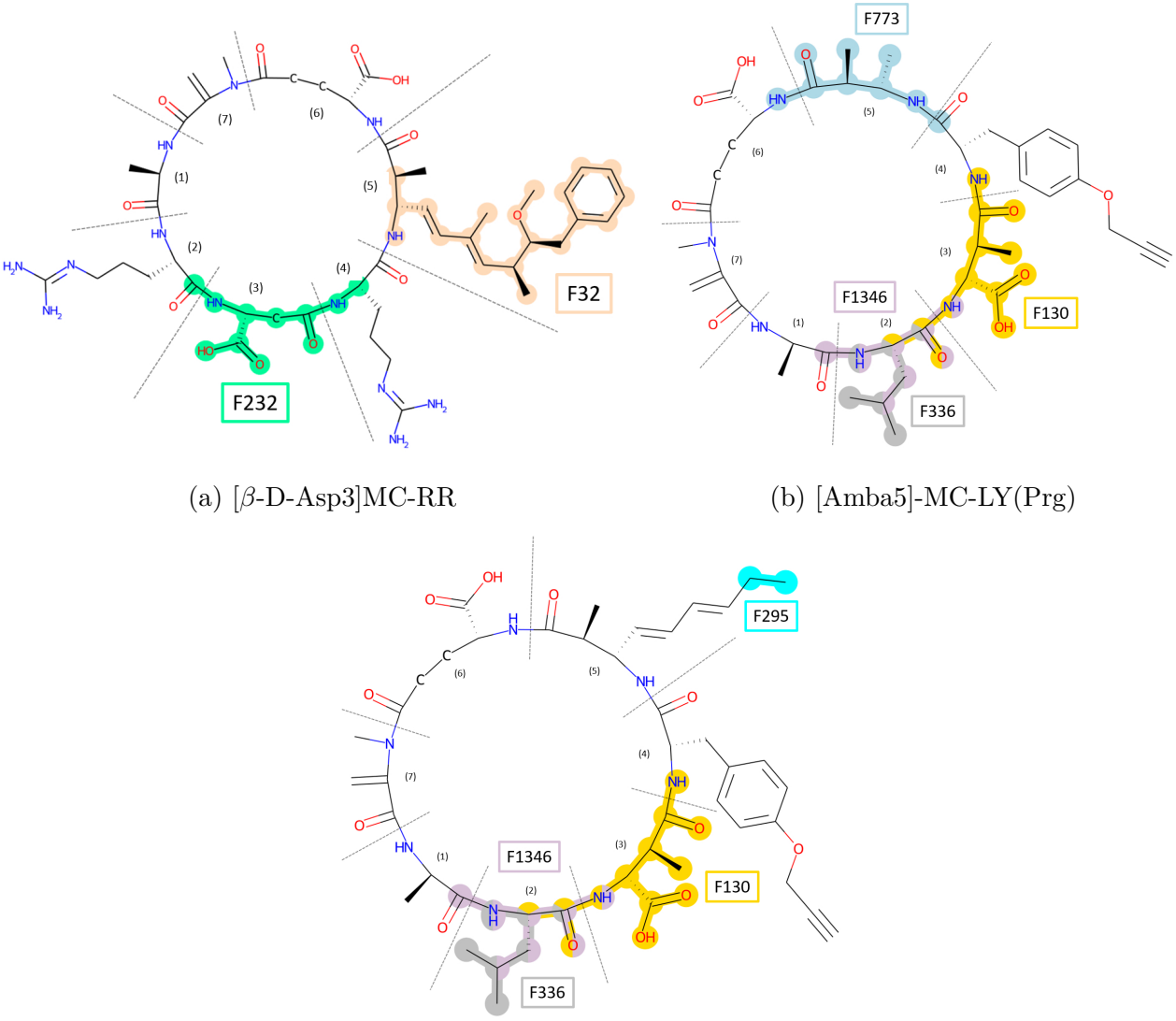
Figure 3.7: Partial sequence alignment of PPP1, PPP2A and PPP5. Amino acids interacting with MC congeners [216] are marked with a box, where different colours indicate different interactions. Beige and light blue: hydrophobic interactions; gold: water molecule replacement; mint-green: indirect coordination to metals; cyan: covalent bonding; silver (unique to PPP2A) and light purple: hydrogen bonds. The image is modified from Pereira et al [232].

of percent identity is the percentage of sequences that have identical residues at the same positions in the alignment [305]. Thus, when considering the sequence of the catalytic core, PPP5 is more closely related to PPP1 than to PPP2A, but evolutionarily distant from both, as PPP5 belongs to a different subfamily than PPP1 and PPP2A [306]. PPP1 and PPP2A share a higher percent identity with 49.27 %. RNA sequence is usually less conserved than the three-dimensional structure, which is more important for interaction. When the three-dimensional structure of the catalytic subunit of PPP5 is compared to PPP1 and PPP2A, it matches well, and a pairwise comparison of the root-mean-square deviation of α-carbon is between 1.3 and 1.5 Angström [307, 308]. From a structural point of view, this suggests an overall similar binding site (size, structural constraints), of the catalytic domains. Side chain residues could vary between the PPPs and they are usually critical for interaction with ligands. There is no structure-activity relationship study available on the interaction between PPP5 and MC congener, because, unlike for PPP1 and PPP2A, the crystal structure of PPP5 and MC has not been resolved. A partial amino acid sequence alignment of PPP1, PPP2A and PPP5 was found in literature (see Figure 3.7). With the amino acid sequence alignment, it is possible to compare individual residues that may be important for the interaction by identifying the corresponding residues of PPP1 and PPP2A. The interactions and corresponding residues of PPP1 and PPP2A are known, allowing the comparison with the amino acids of PPP5, which can be identical or have similar chemical properties compared to the amino acids of PPP1 and PPP2A. Of course, this analysis is only suggestive and a first step to support our findings, but detailed mutational studies or resolving the crystal structure of the complex are necessary to provide evidence.

Interestingly, the feature sets obtained for optimisation on the PPP5 data set, overlaps with the feature sets obtained for the other PPPs. With PPP1, where we have a higher percent identity, we share many more feature sets compared to PPP2A. The feature sets are visualised on three example structures, to cover all feature sets (see Figure 3.8). In the absence of SAR studies on interaction between PPP5 and MC congeners, we cannot conclude on the interaction. Therefore, we refer the reader to the discussion of the feature sets for PPP1 and PPP2A in the previous sections.

3.4.3.5 Considering new structures

The results obtained from solving the (α, β)-k-Feature Set Problem depend on the specific data set used. A slightly modified data set may result in different solutions, and therefore different values for α, β, and k. To assess the behaviour of our solutions (i.e., feature sets) and the performance and stability of our Boolean rules, a test data set based on IC_{50} values of MC congeners on PPP1 was collected from the literature (see Table 3.6). The new MC congeners structures were encoded as ECFP4. In comparison to the training set, the bit information at



(a) [β-D-Asp3]MC-RR

(b) [Amba5]-MC-LY(Prg)

(c) [Anda5]-MC-LY(Prg)

Figure 3.8: Feature sets of PPP5 highlighted on three example structures. F32 (beige) covers Adda side chain and contributes to toxicity. F773 (light blue) has Adda replaced by a methyl group, which leads to decreased toxicity. The role of F295 (cyan) is unclear, but toxicity is highly reduced compared to F32. F130 (gold) and F232 (mint-green) covering the carboxyl group at residue three is not fully understood regarding toxicity, and studies showed different results. F336 (silver) and F1346 (light purple) mark a hydrophobic residue which contributes to toxicity at position 2.

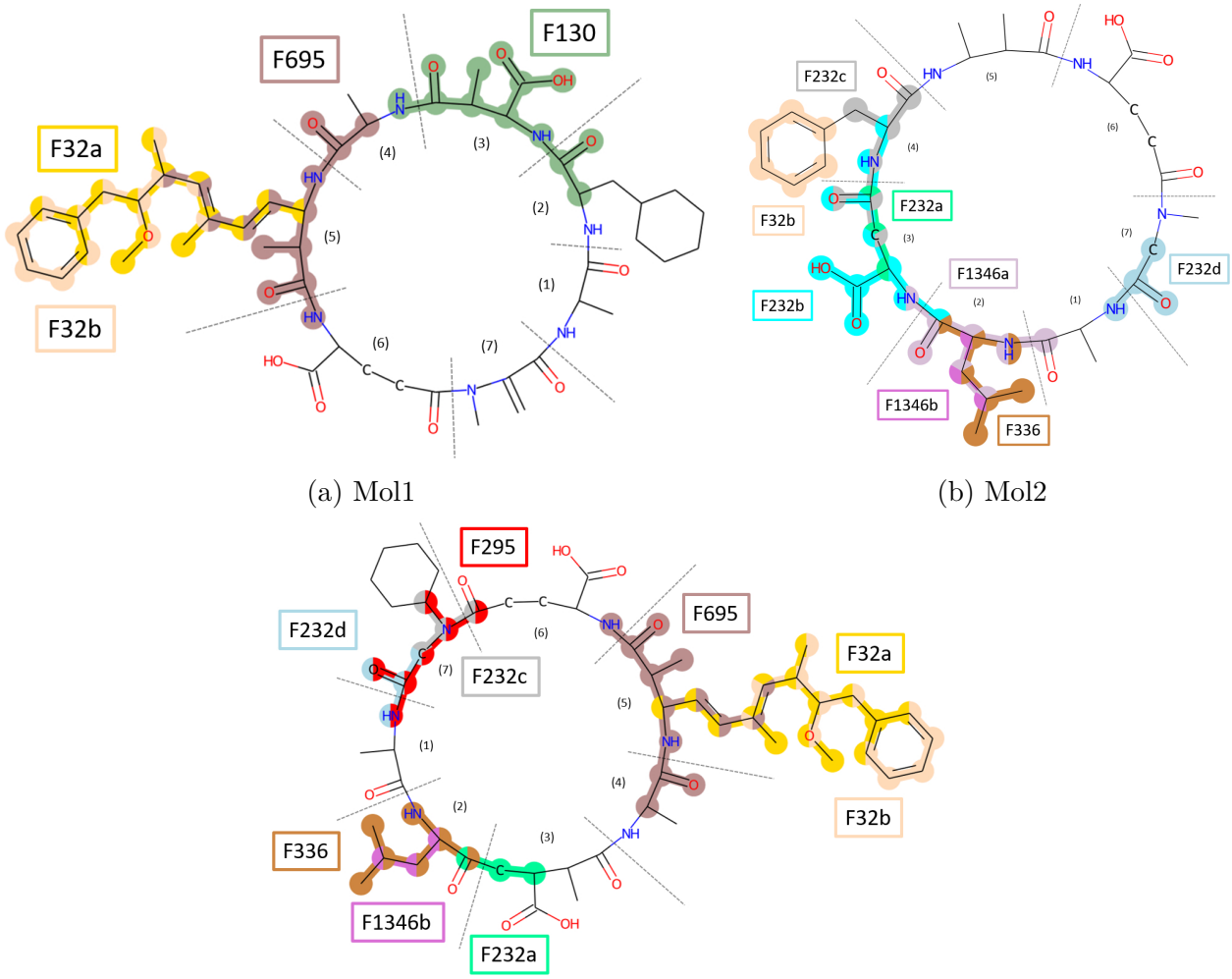
individual fingerprint positions differed for the ECFP4s derived from the test set. Previously, only positions in the fingerprint vector with identical values were summarised. So, for the new test set, the positions in the feature set sometimes result in different values for the individual positions, because the optimisation was not based on this data. For this reason, these feature sets had to be split into smaller feature sets to maintain the original condition of only combining fingerprint/feature positions with identical values. The original naming scheme has been kept for clarity, and where a feature set has been split, it is denoted by small letters in alphabetical order to retain the information about which features have been split into new ones. The new feature sets are shown in Table 3.11 and the equivalent features are listed in the Supplementary Information Section 8.1.1. The Boolean rule originally derived from the original PPP1 data set to infer MC toxicity on PPP1, is as follows: *"If F32 is True, and F130 or F336 are True (i.e., either of those or both of them), then the molecule is toxic for PPP1, otherwise it is less toxic."*. When considering our test data set, F32 splits up into two new feature sets, F32a and F32b, containing 17 and 3 fingerprint positions, respectively. In order to adhere to the original feature set as close as possible, F32a was selected as it consists of most of the original fingerprint positions. Therefore, the Boolean rule was modified as follows: *"If F32a is True, and F130 or F336 are True (i.e., either of those or both of them), then the molecule is toxic for PPP1, otherwise it is less toxic."*. The newly generated feature sets are superimposed on the MC congener structures in Figure 3.9.

By applying the modified Boolean rule to the test data set, a prediction for toxicity classification is obtained (see Table 3.11). Six out of eight MC congeners were correctly classified with five toxic and one less toxic molecule, i.e., a total of 75 %. The incorrectly classified molecules are [Enantio-Adda5]MC-LF and 44271410. [Enantio-Adda5]MC-LF is a stereoisomer of the toxic MC-LF, which belongs to a different toxicity class, the less toxic. The differences in the stereocentres are not reflected in our ECFP4, so it is misclassified because the same feature sets have already described a toxic molecule. 44271410 is incorrectly classified as less toxic, since F130 and F336 are both 0. Since the structures in the training set for PPP1 are only known to be toxic if one of these feature sets is present, a structure where both are absent is classified as less toxic.

Table 3.11: Feature sets, IC_{50} class, and predicted class for new data set on PPP1 toxicity. Each column has a capital letter representing a set of equivalent features. An upper case 'F' refers to a set of equivalent features, lower case letters indicate when a feature set has been split up into multiple feature sets.

Molecule	IC_{50} Class	Predicted Class	F32a	F32b	F130	F232a	F232b	F232c	F232d	F295	F336	F695	F1346a	F1346b
Mol1 [216]	toxic	toxic	1	1	1	0	0	0	0	0	0	1	0	0
Mol2 [216]	less toxic	less toxic	0	1	0	1	1	1	1	0	1	0	1	1
44271302 [290]	toxic	toxic	1	1	0	1	0	1	1	1	1	1	0	1
44271308 [291]	toxic	toxic	1	1	0	1	0	0	1	0	1	1	0	1
44271410 [292]	toxic	less toxic	1	1	0	1	0	0	0	0	0	1	0	0
44271411 [293]	toxic	toxic	1	1	0	1	0	0	0	0	1	1	0	1
44271325 [294]	toxic	toxic	1	1	0	1	0	0	0	0	1	1	0	1
[Enantio-Adda5] MC-LF [16]	less toxic	toxic	1	1	1	0	0	0	0	0	1	1	1	1

The detailed analysis presented here shows that Boolean rules to classify compounds into toxicity classes based on substructures derived from ECFP can reliably estimate toxicity. However, the transferability is of course limited, because the original training data set is very small with only 17 MC congeners. Thus, when the derived Boolean rules are applied on the new data set, not all MC congeners can be correctly classified, although 75 % of new MC congeners were



(a) Mol1

(b) Mol2

(c) 44271302

Figure 3.9: Newly derived feature sets of MC congeners on PPP1 superimposed on three example structures. The colour coding is as follows: FP32a is gold, FP32b is beige, FP130 dark green, FP232a is mint-green, FP232b is cyan, FP232c is silver, FP232d is light blue, FP295 is red, FP336 is brown, FP695 dark brown, FP1346a is light purple, and FP1346b is purple.

classified correctly (six out of eight). Nevertheless, more data to train on would be beneficial and a good opportunity to retrain the model on additional samples to potentially get more accurate results. Such a data set is currently not available, so we have presented here a first approach for toxicity classification of MC congeners, based on small and understandable feature sets that can be translated back to molecular substructures to uncover those related to toxicity.

Even though substructures derived from ECFP can explain and help to derive rules for classification, previous work has raised the issue that using structural alerts on molecular structures can be problematic. Stepan et al. [309] analysed approved drugs and found that for a high percentage of them one or more alerts in their structure was identified, but no significant incidents were reported. This study concluded that structural alerts may overestimate toxicity, which is problematic for drug discovery [309], but not for toxins, which is the focus of this work. This issue was further discussed by Alves et al. [310], who concluded that certain structural features, such as those used in structural alerts, cannot be considered in isolation, but rather as multiple

substructures acting together. Ultimately, a combination of various substructures define the biological activity, which questions the use of structural fragments to predict toxicity of the whole molecule [310]. While the approach presented in this work focuses on substructures, it does not treat them individually, but combines several into feature sets. Therefore, each feature set represents multiple substructures rather than individual fragments. Moreover, the Boolean rule derived from the respective data set that determines relevant structures related to toxicity, considers multiple feature sets and combines the substructures as part of a rule. Therefore, the approach does not focus on an individual substructure at one position in the molecule. For these reasons, the application of the (α, β)-k feature set problem to MC congener binding to PPPs is a valid approach to identify substructures associated with MC congener toxicity.

3.5 Conclusion

In this chapter, two different approaches to classify MC congeners based on their IC_{50} values have been proposed. We were able to show that machine learning and mathematical optimisation can be used to build reliable prediction models for MC congener toxicity.

The first approach was based on a feature representation inspired by natural language processing combined with standard ML algorithms. A vector ultimately represented the MC congener and PPP structure as a whole, by capturing chemical and sequence information, rather than individual substructures of the molecules or individual amino acid triplets of the protein. The machine learning approaches were decision trees, which are suitable for small data sets and usually tend to perform well. In addition, SMOTE has been applied to artificially increase the size of the data set. We have been able to demonstrate that we can build a reliable machine learning model using this methodology, achieving 80-90 % correct predictions of toxicity class. However, this approach is a black box approach and lacks explainability. There is no explanation as to why certain MC congeners were misclassified.

In contrast to the first approach, the second approach aimed to identify explainable and meaningful substructures related to toxicity. The (α, β)-k-Feature Set Problem is an optimisation problem that has been applied directly to ECFP4 for the first time. ECFP4, which is a circular fingerprint, is a classic representation of molecular structures, but has the disadvantage that it results in a sparse vector representation (i.e., more substructures are absent than present). The (α, β)-k method avoids this problem by reducing the dimensionality of the data sets to derive meaningful feature sets. These feature sets were then used to derive Boolean rules to classify MC congeners in toxicity classes, which were tested on a newly developed data set. We could clearly show that we were able to derive chemically meaningful signatures that could be related to toxicity by comparing the feature sets obtained with data collected in literature, and to achieve 75 % correct predictions on the newly collected data set. The identification of the Adda side chain in the different feature sets was particularly interesting, as Adda serves as an important anchor in the PPP binding site [212, 216].

Nevertheless, both approaches suffer from the fact that — even though it currently is the largest publicly available data set — our MC congener data set is very small. Both would probably give more reliable results if more data on MC congeners, or other PPP inhibitors, were available. However, it is possible to re-train the models as soon as data becomes available, so future models are likely to be more reliable. Although this is a limitation, the overall approach is useful as both approaches roughly categorise individual MC congeners into their toxicity classes, and thus provide preliminary predictions. In vitro toxicokinetic and toxicodynamic assays are not available for the majority of the approximately 279 characterised MC congeners [15],

making risk assessment very difficult. Therefore, the approaches presented are a first step towards building a reliable classification model, and gaining knowledge about the individual substructures contributing to toxicity, which is a step towards improved risk assessment. In future work, the three-dimensional conformation and interaction of MC congeners could be investigated or integrated into machine learning models, to obtain more reliable and even more explainable results.

Chapter 4

Molecular Modelling of Macrocyclic Structures

4.1 Summary

This chapter describes two approaches to investigate the conformation and binding of Microcystin (MC) congeners to ser/thr-protein phosphatases 1 (PPP1) by Molecular Dynamics (MD) simulation. As discussed in Section 2.3.1, MC congeners are structurally similar macrocycles. The prediction of inhibition and interaction with PPPs based on simple feature representation is difficult due to the small available data sets. To gain further insight, MD simulations were carried out to get a better estimation of interaction and conformational behaviour. The MD simulations are also used for the development of interaction fingerprints and their analysis and interpretation (see Chapter 5). Both approaches presented here are MD simulations of MC congeners, but differ in data set size and methodological details.

The first approach is described in detail in Section 4.3. MD simulations were carried out on four MC congeners and PPP1. This work has been published in the Journal of Chemico-Biological Interactions. Ideas, figures, and tables have been taken from the publication and text modules may be similar to a certain extent:

- **Sabrina Jaeger-Honz**, Jahn Nitschke, Stefan Altaner, Karsten Klein, Daniel R. Dietrich and Falk Schreiber: 'Investigation of Microcystin Conformation and Binding Towards PPP1 by Molecular Dynamics Simulation' Chemico-Biological Interactions 351: 15 pages, 2022. doi:10.1016/j.cbi.2021.109766.

According to the FAIR principles (Findability, Accessibility, Interoperability, and Reuse of digital assets, see Section 1.4) all data sets and scripts which are part of this manuscript have been published open source in the Zenodo repository:

1. https://doi.org/10.5281/zenodo.5017745 [48]
2. https://doi.org/10.5281/zenodo.5017839 [49]
3. https://doi.org/10.5281/zenodo.5017851 [50]

While the first approach relied on manual curation and Amber force fields, the second approach makes use of CGenFF for automation of parameterisation with CHARMM force

fields and is described in detail in Section 4.4. MD simulations were carried out on twelve MC congeners and their respective cysteine (Cys) and glutathione (GSH) conjugates with PPP1. Methodologically, the simulations are based on the CHARMM force field and the parametrisation has been automated with CGenFF. Parts of this work has been published in the Journal of Chemico-Biological Interactions. Ideas, figures, and tables have been taken from the publication and text modules may be similar to a certain extent:

- **Sabrina Jaeger-Honz**, Raymund Hackett, Regina Fotler, Daniel R. Dietrich and Falk Schreiber: 'Conformation and binding of 12 Microcystin (MC) congeners to PPP1 using molecular dynamics simulations: A potential approach in support of an improved MC risk assessment.', Chemico-Biological Interactions 407: 15 pages, 2025. doi: `10.1016/j.cbi.2025.111372`.

Parts of the MD simulation data and results of this manuscript have been published open source in the Zenodo repository:

1. https://doi.org/10.5281/zenodo.14501700 [53]

For a statement of contribution, the reader is referred to Section 1.3.

4.2 Background

In order to estimate binding dynamics and hence toxicity, it is crucial to understand the conformation of proteins and their interactions with ligands. Both parameters can be studied using computational methods such as docking or MD simulations [13]. MC congeners are macrocycles that have a scaffold function. The conformation of the scaffold or backbone is critical for interaction of MC congeners with different proteins and strongly influences their toxicokinetics and -dynamics. Toxicity is the endpoint of complex processes influencing each other and the toxicity of molecules, such as MC congeners, affects many processes in the cell and body (see Section 2.3.2). Therefore, the focus of this work was to investigate one of the experimentally most studied factors to be able to assess the reliability of our approach, namely the toxicodynamics of PPP1 inhibition (see Figure 2.3). Predicting the correct macrocyclic structure to study interaction is difficult and the available experimental data is often sparse [3]. In this case, structural data in complex with PPP1 is only available for MC-LR. Many studies have focused on docking of MC congener structures to different targets to study their interaction. Although some studies could replicate the interactions found in the crystal structure of MC-LR bound to PPP1 [210, 311, 312], the binding pose obtained by docking is unlikely to be accurate enough to provide decisive insights, because MC congeners are very large (more than 35 heavy atoms) [3] and the binding site of PPP1 is shallow [227, 313, 314]. Docking and scoring are very sensitive to changes in the 3D structure, therefore we decided to use MD simulations to study binding and conformations. Docking is applied to place the MC congener in the binding site, but not as a stand-alone technique.

To date, the use of Molecular Modelling and in particular MD simulations on MC congeners to study their conformations and binding has been limited. There are some studies focusing on different properties, as e. g. covalent binding of MC to PPP1 [311], investigating the adsorption mechanism of MC at the water-mineral interface [315], metal binding selectivity and coordination [316], or on MC-LR/MC-RR and disinfection by-products binding to PPP1 [210, 312] or PPP2A [317], or investigating the sensitivity and binding mode of MC to antibody binding [318, 319]. The first MD simulation to study MC-LR conformation was carried out in 1996

by Trogen et al. [320] and the first study of MD simulation of PPP1 MC-congener interaction to explore toxicity was carried out in 2000 by Mattila et al. [213]. Studies of MC congener conjugates on PPP1 were only done with docking so far [210, 321]. For the sake of completeness, studies investigating the conformations and interactions of MC congeners and conjugates will be summarised in the following, regardless of the methodology used (e. g. nuclear magnetic resonance (NMR), docking, MD simulation).

The backbone of MC-LR was predicted to be planar in the first attempts to understand the 3D structure of MC-LR [322, 323]. In contrast to the predictions, a subsequent experimental study by Rudolf-Böhner et al. [324] showed a different result. MC-LY and MC-LR were dissolved in dimethyl sulfoxide (DMSO) and their conformations resolved by NMR spectroscopy. MC-LY adopts to a compact boat-like structure. The Adda and tyrosine side chains of MC-LY, collapse together, resulting in restricted movement of the Adda side chain. MC-LR had three conformational families of half-chair-like arrays, which is a more compact structure than MC-LY, with the Adda and arginine side-chains facing outwards from the MC ring structure, resulting in more conformational freedom. In 1995, a NMR study by Bagu et al. [325] interpreted MC-LR as a saddle-shaped structure, with Adda also extending away from the ring structure behind the saddle, leading to conformational freedom.

In the same year, the crystal structure of PPP1 with bound MC-LR was resolved and the interactions studied [212]. The crystal structure showed a similar conformation of MC-LR to the NMR study by Bagu et al. [325]. It was shown that MC-LR interacts with the hydrophobic groove (via the Adda side chain), coordinates indirectly with metal ions and water molecules in the active binding site via carboxylate groups of the Glu residue and carbonyl oxygen of MC-LR, leucine packs closely to Tyr272, carboxylate group of Masp hydrogen bond with Arg96 and Tyr134, and Mdha forms a covalent bond with Cys273 [212].

Trogen et al. [320] generated 3D structures of MC-LR with NMR spectroscopy in water and DMSO solution. In both solvations, one conformational family with saddle-shaped form of MC-LR backbone was identified. MD simulation and other techniques were used to study conformational exchange. Also, MD simulation in solvent with a length of 372 ps showed one conformational cluster. The identified conformations were in agreement with the structures reported by Rudolf-Böhner et al. [324], Bagu et al. [325] and the conformation identified in the crystal structure [212].

In a second NMR solution study with MC-LR and MC-LL Bagu et al. [326] showed that both congeners have a similar, saddle-shaped backbone and overall structure and are similar to the bound crystal structure of MC-LR with only minor differences. Docking was used to confirm a binding pose of MC-LL into PPP1c, and a similar conformation to MC-LR bound in crystal was obtained. The authors concluded that the free solution NMR structure of MC-LR does not differ significantly in conformation upon binding and inhibition of PPP1. They assumed that the high affinity towards PPP1 is caused by the rigid saddle-shaped conformation of the backbone, as this already provides a scaffold for correct binding. Furthermore, the Adda side chain plays a crucial role by serving as an anchor, ensuring the accurate placement of MC congeners within the PPP1 groove. This positioning enables the other residues to be already aligned correctly, facilitating direct interactions. Since the conformation and binding pose of MC-LL is so similar to MC-LR, the authors proposed that other MC congeners will have a similar free and bound structure compared to MC-LR, as the second leucine of MC-LL does not lead to structural change [326]. In summary, the authors proposed that MC congeners do not influence the overall structure of PPP1 upon binding and that no conformational change in MC-LR is induced upon binding. Moreover, all MC congeners should have a similar conformation

and structure compared to MC-LR, since MC-LL behaves exactly like MC-LR does.

Two different studies of MC-LR interaction were published in 2000. One is the study by Lavigne et al. [327] where an ensemble of MC-LR NMR structures were docked into the catalytic centre of PPP1 and properties were calculated to estimate the free energy of dissociation and compared to the crystal structure of PPP1-MC-LR. Their analysis led to the conclusion that numerous docking poses exhibiting comparable dissociation energy to the crystal structure can be detected, showcasing considerable structural variability. This finding implies that MC-LR can bind to PPP1 with diverse structural arrangements, indicating the existence of conformational flexibility within the complex in solution. In 2000, the first known MD study of PPP1 in complex with MC-LR and surrounding water was carried out to define the binding characteristics [213]. In the 600 ps of simulation, no major structural changes for MC-LR or PPP1 were observed. A minor conformational change occurs at the IGlu residue, probably due to the strong attraction of IGlu carboxyl groups to the metal ion. In contrast to the resolved crystal structure of PPP1-MC-LR, four hydrogen bonds frequently formed during MD simulation instead of three. Other strong interactions observed in the crystal structure could be identified and include metal coordination with the metal ions, hydrogen bonding with Arg96, Asp220, Arg221, and Gly274, water molecule replacement and hence interaction with Asn124 and His125, ionic interaction with Glu275, and coordination of His248, His66, Asp64 and Asp92 with metal ions in the binding site. For leucine and arginine, no significant interactions were observed, and Adda was reported to ensure correct orientation, but no real interaction [213]. To summarise, MC-LR does not induce major conformational change in PPP1 binding, but stabilises the structure by causing only local perturbations in the binding site. The MD studies were very short in timescale compared to today's simulations and the timescale of toxin binding, and only a small part of the enzyme was represented explicitly.

There is even less data available on the interaction of MC congener conjugates (i.e., biothiol conjugates) with PPP1. Only two docking studies exist so far from the same authors [210, 321], which has not been further validated. Their experimental results could be validated in other experimental work. MC congener conjugates are produced by a cell to detoxify MC congeners via GSH pathway. They showed varying degrees of toxicity and the molecular mechanism is still unclear.

The first study by Zong et al. [321] investigated the inhibition capacities and the mechanism of MC-LR-GSH and MC-RR-GSH binding to PPP1. Both were found to have lower toxicity (i.e., inhibition capacity) than the respective unmodified MC conjugate. The second study by Zong et al. [210] focused on MC-LR conjugated with GSH, Cys, Cys-Gly, Hcy, AcSH and βME. MC-LR-GSH and MC-LR-Cys-Gly have high toxicity but lower than MC-LR, while MC-LR-Cys and MC-LR-Hcy had some toxicity, and MC-LR-βME and MC-LR-AcSH were toxic. In both studies, GSH conjugation resulted in reduced interaction with key interaction partners, such as the building of hydrogen bonds or hydrophobic interactions with Adda. Adda was found to be crucial for interaction with PPP1, and also plays an important role in the interaction with MC congener conjugates. In addition, conjugates reduced the number of total hydrogen bonds, and alter key interaction partners and therefore the inhibition pattern. The increased size of the MC congener by conjugation of a biothiol, leads to a larger surface area to interact with, but induces steric hindrances that regulate toxicity. Secondary toxicity of the products was diversified and non-negligible. In addition, it altered the coordination of Mn^{2+} ions with the protein residues, thus influencing the toxicity of MC conjugates, as metal ion coordination is crucial for a functional PPP1, which is a metalloenzyme [227]. Mn^{2+} interaction patterns with protein residues in the catalytic centre were altered by the conjugates, and overall weak-

ened, resulting in reduced toxicity [210]. In summary, the inhibition decreased with increase in biothiol molecular weights and polarity, therefore GSH is an effective pathway to regulate MC-LR toxicity.

The number of known MC congeners are increasing, but it is still difficult to synthesise and study them. For this reason, Molecular Modelling approaches can help to understand the interaction and conformation of MC congeners, to further investigate and assess risks of MC congeners and to guarantee safety for human consumption [16]. Due to technological processes and powerful approaches of e. g. MD simulation to study interaction of MC congeners with PPP1 and conformation of MC congeners, we considered it timely and advantageous to not only consider MC-LR but also other MC congeners.

4.3 Molecular Dynamics Simulation of 4 MC Congeners

4.3.1 Introduction

The MC congeners used in this study were chosen based on a wide range of PPP1 inhibition capacity, which was determined by a colorimetric phosphatase assay (see Table 3.2). MC-LR and MC-LF were chosen as toxic MC congeners, [β-D-Asp3]-MC-RR as less toxic congener, and [Enantio-Adda5]MC-LF as non-toxic congener. [Enantio-Adda5]MC-LF has an inverted stereochemistry of the Adda side chain in all four stereocentres compared to MC-LF [200]. Since MD simulations and especially parameterisation are expensive and time-consuming, only four MC congeners were chosen for this initial study on interaction with PPP1. For PPP1 and MC-LR, a crystal structure (PDB Code: 1fjm [212]) is available in the PDB database [328], which was chosen as starting structure. For the other MC congeners, no structure is available in complex with PPP1 and MC congeners were docked into the binding site.

Two MD simulation approaches were chosen for the data presented here: 1) a complex simulation to study interaction and conformation of the MC congener with PPP1, and 2) a (TIP3P) solvent simulation to explore the macrocycle backbone conformation and properties of MC congeners.

4.3.2 Methods

4.3.2.1 Structure Preparation

PPP1 preparation Before MD simulation, the structure of PPP1 had to be prepared. The PDB structure 1fjm [212] was downloaded from PDB [328] and subunit A chosen as receptor relevant for MC congener interaction. All water molecules were excluded from the structure, except two, which are important for coordination with the two manganese ions in the active centre. Explicit hydrogen atoms were added with UCSF Chimera [329] to compute Gasteiger partial charges. The manganese charge was fixed at +2 and Gasteiger partial charges for water ions were calculated with UCSF Chimera [329], which are necessary for docking.

MC structure preparation To prepare three-dimensional structures of MC congeners, the crystallised structure of MC-LR was deattached from the PPP1 receptor by deleting the bond to Cys273. It was noted that a carboxyl group at the D-glutamic acid in MC-LR was missing.

This was considered as an error and the structure manually corrected with UCSF Chimera [329]. Next, it was energetically minimised with Chem3D© (v16.0 by CambridgeSoft, PerkinElmer Informatics) by using a minimum root means square (RMS) Gradient of 0.01 and a termination after a maximum of 10,000 steps with a heating temperature of 0 K and a target temperature of 300 K by applying the MM2 force field. For the other MC congeners, no crystal structures were available. For this reason, the corrected MC-LR structure was used to generate presumed structures of MC-LF, [β-D-Asp3]-MC-RR, and [Enantio-Adda5]MC-LF. The structures were modified and generated with UCSF Chimera. To generate [Enantio-Adda5]MC-LF from MC-LF, the stereoisomeric moieties were rotated with Matlab R2016a [330]. For all structures, UCSF Chimera [329] was used to add hydrogen atoms to allow docking.

4.3.2.2 Docking

For the MD simulation of MC congener and PPP1, an initial pose of MC in the binding site of PPP1 is necessary. As a structure has only been crystallised for the MC-LR with PPP1, docking is required prior to the MD simulations in order to generate the initial structures and the starting position for the other three MC congeners.

Autodock 4.2 [331] was used for docking, and it was checked whether a docking of MC-LR similar to its crystal pose could be generated. Autodock 4.2 [331] was not capable to treat a macrocyclic flexible. For this reason, the MC congener backbone was treated as rigid. In addition, the number of rotatable bonds is limited to 32, so the receptor with manganese ions and water molecules was treated as rigid. The residues of MC congeners differing between congeners (residue 2 (L-Leucine), 4 (L-Arginine) and 5 (Adda)) were kept rotatable since they differ among the congeners. Each docking generated 10 poses, which were sorted according to the increasing docking score. Each docking was repeated three times, resulting in 30 poses. All poses were compared with each other, and poses closest to the crystallised MC-LR pose were selected as initial starting positions. The standard docking parameters were employed according to the user guide, except that charges were not replaced in the receptor structure, since the charges for water and manganese were set by UCSF Chimera [329]. The non-transformed coordinates were used from published structures of PPP1 [212] and the centre of the grid box selected close to the active site at (91.226, 23.163, 24.424) with dimensions of (30, 22.5, 18.75 Å) corresponding to standard length units used by Autodock 4.2 [331].

4.3.2.3 Molecular Dynamics Simulation

Preparation For MD simulation, the orientation of side chains and position of hydrogens in the PPP1 structure are important. To determine the position of hydrogens and orientation of asparagine, glutamine and histidine, Molprobity [332, 333] was used and recommended flips performed. Afterwards, terminal residues were capped with N-methyl amide group (NME) at the C-terminus and an acetyl group (ACE) at the N-terminus. The Amberff14SB force field [149] was used to parameterise the protein, and active site waters were treated as TIP3P solvent. For manganese ions, the default parameterisation of Amberff14SB force field was not appropriate, since the manganese ions and active site water dissociated in test runs of the apo structure (protein in solvent). For this reason, the parameters were exchanged with parameters from

magnesium ions, since manganese ions are only slightly larger and have similar coordination preferences to magnesium ions [334]. The exchange of parameters leads to a stable coordination in the binding site.

MC congeners are made up of standard and non-standard amino acids. Non-standard amino acids were parameterised manually. This was achieved by separate treatment of non-natural amino acids, which were cut out of the macrocycle backbone. To mimic an intact backbone, the non-natural amino acids were capped with NME and ACE, and UCSF Chimera was used to add hydrogen atoms [329]. The calculation of charge derivations and selection of force field parameters was then set up on R.E.D. Server Development tutorial IV.1 (central fragment of amino acid) to obtain charges and parameters [335–338]. GAMESS [339] was used for quantum mechanical calculations and unknown parameters manually derived from Amber ff14SB force field [149] or general amber force field (GAFF) [148]. The standard amino acids were parameterised with Amber ff14SB force field. To generate amber topology files, LEaP was applied [340] and ParmEd [341] was used to translate them to Groningen machine for chemical simulations (GROMACS) topology.

Procedure GROMACS version 2016.4 [342, 343] was used for MD simulations. For baseline comparison, an apo simulation of PPP1 in solvent was set up, as well as solvent simulations of only MC congeners. In addition, a complex simulation (PPP1-MC-congener) was set up. For all simulations, a transferable intermolecular potential with 3 points (TIP3P) water model [344] was used as solvent and the system neutralised with sodium ions. To calculate long-range electrostatic interactions, particle mesh Ewald (PME) [345] with Verlet cutoff-scheme was applied [346]. To constrain bonds to correct length, the LINCS algorithm was applied [347]. To equilibrate the system and minimize the energy, steepest descent minimisation, with a maximum of 50,000 steps or until a maximum force smaller than 10.0 kJ/mol was reached, was applied. Afterwards, the system was equilibrated in two phases. The first phase, a constant volume, constant temperature (NVT) equilibration (isothermal-isochoric) to stabilise the temperature, with a velocity-rescaling thermostat [348], was performed for 100 ps with a time step of 2 fs. The second phase, a constant pressure, constant temperature (NPT) equilibration to stabilise the pressure to 1 bar with a Parrinello-Rahman barostat [349] followed. This was running for 100 ps with a step size of 1 fs for solvent simulations, and 2 fs for complex/apo simulations. After this pre-processing, the MD run was started. For solvent simulation it was performed for 20 ns (step size of 2 fs), for complex and apo simulation for 280 ns (step size of 2 fs). For each simulation type, three replicates were set up. For the solvent simulation, the structure of MC congener from the respective complex was extracted and used as starting point. To obtain a starting structure for replicates 2 and 3, a snapshot at 10 and 15 ns during the first replicate was taken and used to restart the simulations. Replicate 1 of the apo simulation of PPP1 was started with solely the prepared structure of PPP1. Likewise, snapshots at 100 and 200 ns were done to get different conformations of the protein structure for replicates 2 and 3. Those conformations were aligned with the initial structure of PPP1 and used to generate new starting structures of MC-congener and PPP1 complex to perform replicate 2 and 3. An overview of simulation lengths and types can be found in Table 4.1.

Analysis To analyse the results of the MD simulation, mostly built-in tools from GROMACS [342, 343] were used. To judge the equilibration of the system, temperature, and pressure were calculated over the whole trajectory. For all other analysis methods, the trajectories obtained by MD simulations were cut off after the first 5 ns and 30 ns for solvent

Table 4.1: Summary of employed MD simulations and total simulation time.

Simulation type	System	Simulation time
solvent	MC-LR	20 ns
solvent	MC-LF	20 ns
solvent	[Enantio-Adda5]MC-LF	20 ns
solvent	[β-D-Asp3]-MC-RR	20 ns
apo	PPP1	280 ns
complex	PPP1-MC-LR	280 ns
complex	PPP1-MC-LF	280 ns
complex	PPP1-[Enantio-Adda5]MC-LF	280 ns
complex	PPP1-[β-D-Asp3]-MC-RR	280 ns

and complex/apo simulation, respectively, to analyse the well equilibrated system. Rotational and translational movements were eliminated with a least squares fit and periodic boundary conditions removed. The root-mean-square deviation (RMSD), volume, radius of gyration, and solvent accessible surface area were determined for PPP1 and MC congener, as well as a number of contacts and hydrogen bonds between PPP1 and the MC congener. Principal Component Analysis [350] was used to analyse the backbone structures of PPP1 and MC congeners.

Python programming language [351] was used to calculate properties and visualisation of data. NumPy (v1.18.5) [352] was used to calculate mean values, standard deviations, the median, and interquartile ranges. To compare different simulation settings with each other, linear least-squares regression with SciPy [353] function linregress was applied to calculate a 95 % confidence interval of slope and intercept for each simulation type for all replicates simultaneously to compare different simulation settings. For visualisation of data with boxplots and principal component analysis, Matplotlib (v3.2.2) [354] and Mol2vec (v0.1) helper function [104] were used, respectively. Time series data was visualised with Jscatter, (v1.2.7.2) [355] which is a python wrapper used to plot data with Grace [356]. Molecular structures and interactions were visualised with PyMOL [357] and exported as images.

4.3.3 Results and Discussion

4.3.3.1 Stability of PPP1

To determine the influence of MC congener binding on PPP1 structure, the protein stability of PPP1 was investigated. To compare to a baseline, the apo simulation is used (see Table 4.1). The protein properties volume, radius of gyration and solvent accessible surface area were investigated, since they might change upon MC-congener binding. The statistics of the data set (mean, median, interquartile ranges, statistical evaluation) are given in the Supplementary Information (see Supplementary Information (SI) Table 8.2 and SI Table 8.3). The properties' distribution over time are shown in Figure 4.1. The protein properties volume, radius of gyration and solvent accessible surface area for PPP1 have comparable distributions for all simulations. The radius of gyration has a mean value of 1.87 ± 0.01 nm and a median of 1.87 nm, and did not change for PPP1 over the different simulations. This measure is an estimate of the size and compactness of a protein, which suggests that PPP1 structure was very stable through the simulation without changes in size or compactness [358], independent of the simulation type. The same effect could be observed with the volume and solvent accessible surface area (accessibility of surface area of protein to solvent [359]). Both properties were stable and did

not change over time, with no significant differences in the mean value of range: 59.07 $\frac{nm}{S^3/N}$ - 59.31 $\frac{nm}{S^3/N}$ and 135.59 $\frac{nm}{S^2/N}$ - 136.60 $\frac{nm}{S^2/N}$, respectively. The statistical evaluation of potential differences between complex and apo simulations were done with linear least-squares regression. A 95 % confidence interval of slope and intercept was determined (see SI Table 8.3) and for all three calculated properties and simulations, the confidence intervals for slope were 0.000 and for the intercept the confidence values overlapped for all calculated properties, suggesting stable values during the simulation.

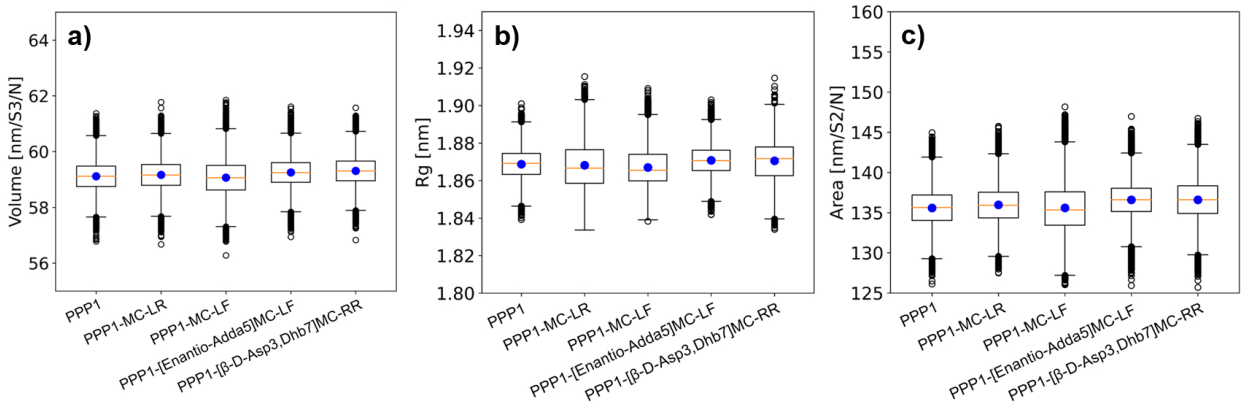


Figure 4.1: Property value distribution for PPP1 (with Mn^{2+} and active site water) during MD simulation. The orange line marks the median, the blue circle shows the mean value. The different properties shown are: a) volume, b) radius of gyration and c) solvent accessible surface area.

To highlight the deviation of a structure from a reference structure, the root mean squared deviation (RMSD) can be calculated. The smaller the RMSD, the higher is the structural stability, i.e., of PPP1. The RMSD over time is shown in the SI Figure 8.1, and calculated properties are shown in the SI Table 8.2. In comparison to apo simulation, the RMSD of the backbone of PPP1 is smaller for all PPP1-MC-congeners simulations, except for [β-D-Asp3,Dhb7]MC-RR. However, the distribution of all PPP1-MC-congener backbone (except for MC-LF) RMSD value is more distributed in comparison to apo simulation, suggesting a higher protein flexibility in PPP1-MC-congener simulations. The mean of RMSD for PPP1 backbone were for all simulations below 0.20 nm, which is considered very stable, but the intercept confidence interval showed higher RMSD values for less and non-toxic MC congeners in comparison to apo simulation. In contrast, PPP1 structure simulated with toxic MC congeners are even more stable than apo simulations. From this data, we can conclude that MC congener binding does not induce major conformational changes in the structure of PPP1.

4.3.3.2 Stability of MC Congeners

In addition to protein backbone stability of PPP1, also stability of MC congener (i.e., macrocycle) were investigated. Here, also the same properties of volume, radius of gyration, solvent accessible surface area and RMSD were calculated to investigate changes in macrocycle backbone upon binding to PPP1 (see SI Table 8.4 and SI Table 8.5).

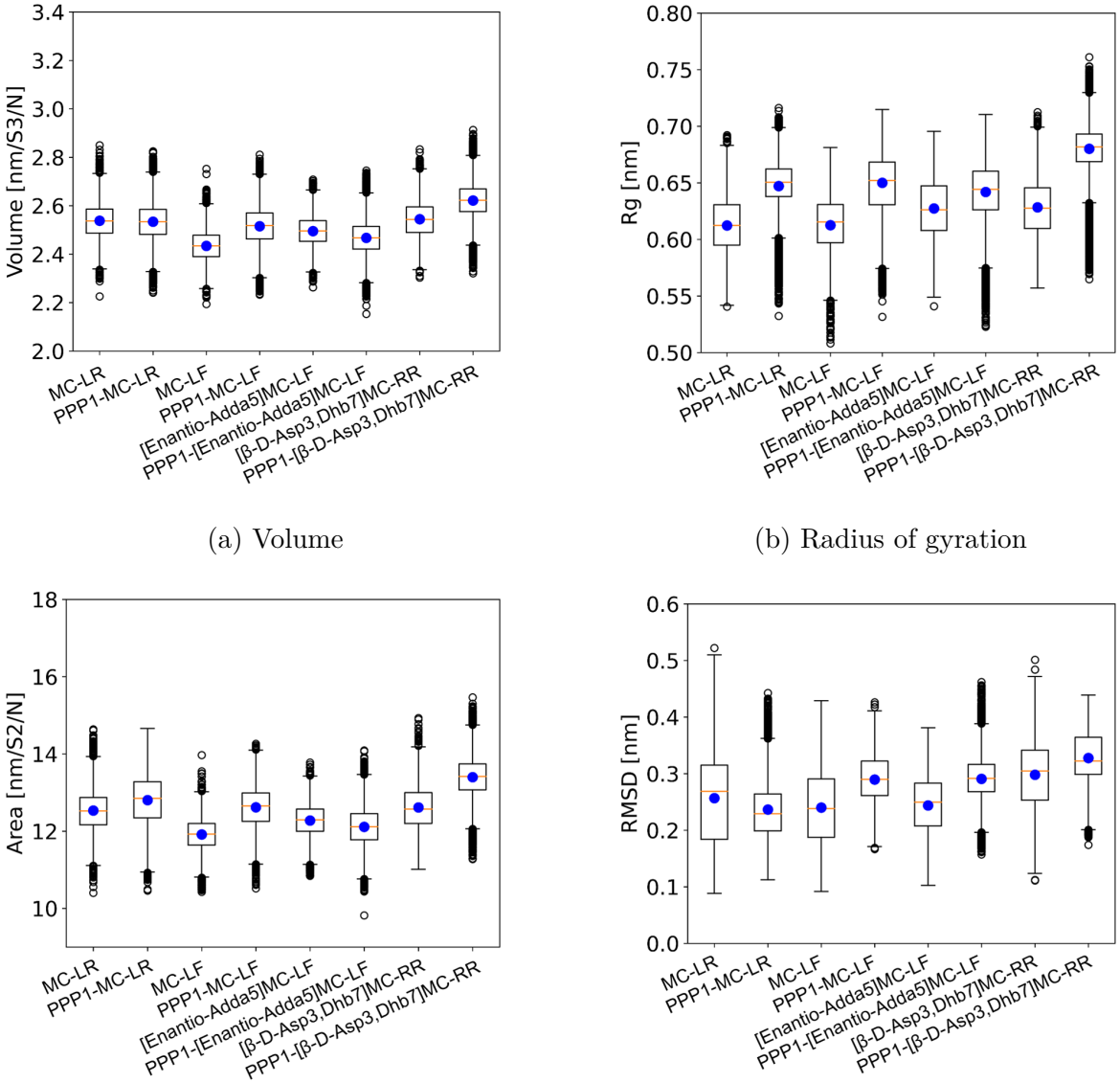
In Figure 4.2 the distribution of these calculated values are shown as boxplots for MC congener simulated in solvent, and compared to complex simulation where MC congener was

simulated in PPP1 binding site. For the volume (see Figure 4.2a), no change was observed for MC-LR. For MC-LF and [β-D-Asp3,Dhb7]MC-RR on the other hand, the volume increased while for [Enantio-Adda5]MC-LF the volume decreased. Also for the radius of gyration (see Figure 4.2b) and solvent accessible surface area (see Figure 4.2c) we observe a different trend for [Enantio-Adda5]MC-LF compared to the other MC congeners. For [Enantio-Adda5]MC-LF both properties are comparable between solvent and complex simulation, while for the other MC congeners the values increased, even though they are still within their standard deviations. Nevertheless, we can see that the distribution of values change.

When simulating MC congeners in solvent, the MC congener is surrounded by water. For this reason, it is expected that MC congeners have a more open structure when simulated in complex in comparison to simulation in solvent. Since the hydrophobic residues of MC congeners (e. g. Adda, or partly amino acids at the hyper-variable positions) try to shield themselves from the hydrophilic environment in water, the structure is more coil-like than when bound in PPP1. In fact, it was reported in literature that no major conformational change in backbone structure of MC congener occurs upon binding to PPP1 [212, 320, 326]. The most hydrophilic congener [β-D-Asp3,Dhb7]MC-RR with positively charged amino acids at position 2 and 4 adopts in contrast to the other MC congeners to a more open structure when binding to PPP1. Probably, the positively charged amino acids reduce the interaction with the hydrophobic groove of PPP1, which increases the properties calculated when simulated in complex. When compared to the other MC congeners they have similar properties for the solvent simulation, but for complex simulation we observe a higher mean for this MC congener (see Figure 4.2). These findings indicate a more open conformation of [β-D-Asp3,Dhb7]MC-RR when bound to PPP1 in comparison to other MC congeners, which is also supported by statistical analysis (see SI Table 8.5). The more hydrophobic MC congeners with phenylalanine at hyper-variable position 4 (MC-LF and [Enantio-Adda5]MC-LF) have lower values in solvent simulation compared to MC-LR and [β-D-Asp3,Dhb7]MC-RR. Nevertheless, the mean values for both congeners are within standard deviation of each other. For [Enantio-Adda5]MC-LF the values are overall lower when simulated in complex with PPP1, which is explained by the orientation of [Enantio-Adda5], which is flipped over the backbone due to flipped stereocentres, instead of extending away from the backbone towards the hydrophobic groove as we can observe for MC-LF. In Figure 4.2d the all-atom RMSD of MC congeners simulated in solvent and complex (i.e., PPP1 binding site) are shown. The distribution of RMSD values in solvent are more wide in comparison to complex simulation, which probably occurs due to conformational restriction of MC congener when simulated in PPP1 binding site. Except for MC-LR, the all-atom RMSDs for simulation of MC-congeners in complex with PPP1 result in a higher median and mean RMSD, which was confirmed by statistical analysis (see SI Table 8.5). The higher RMSD could be caused by docking poses which might not be suitable for MC congeners. MC-LR was the only one for which a crystal structure was available, so the overall conformation at the start of the simulation might have been more well suited for MC-LR than for the other MC congeners. For the other MC congeners, more rearrangement may have occurred to correct the initial position and conformation, leading to a higher deviation from the starting structure, which is what the RMSD calculation is based on.

4.3.3.3 Conformational Space Analysis

In order to assess the overall motion of the system (e. g. protein or ligand), Principal Component Analysis (PCA) was used to examine atomic fluctuations to investigate stability of PPP1 backbone. PCA can be used to reduce dimensionality of a data set (i.e., a three-dimensional



(a) Volume

(b) Radius of gyration

(c) Solvent accessible surface area

(d) All-atom RMSD

Figure 4.2: Properties of MC congeners during MD simulation, shown as distribution. The orange line marks the median, the blue circle shows the mean. The properties of MC congeners are calculated for simulation in solvent and in complex.

structure) by projecting the high-dimensional atomic fluctuations to a low-dimensional space to identify conformations. Usually, the first few eigenvectors play the most important role and are analysed to get an idea about the motion [360]. For a detailed description of PCA, the reader is referred to Section 2.2.1.1.

PCA on PPP1 Backbone To investigate stability of PPP1 backbone and investigate the movement, PCA was applied (see Figure 4.3). The principal components (or eigenvectors) of the PPP1 backbone movement of apo simulation, were projected on PPP1 backbone structure of complex simulation to allow identifying differences in dynamics, and to investigate whether interaction with MC congener leads to altered conformations. Overall, the PCA projection

of apo simulation leads to one big cluster suggesting a stable structure which does not adapt to different conformations, which is in agreement with the results of property analysis (see Section 4.3.3.1).

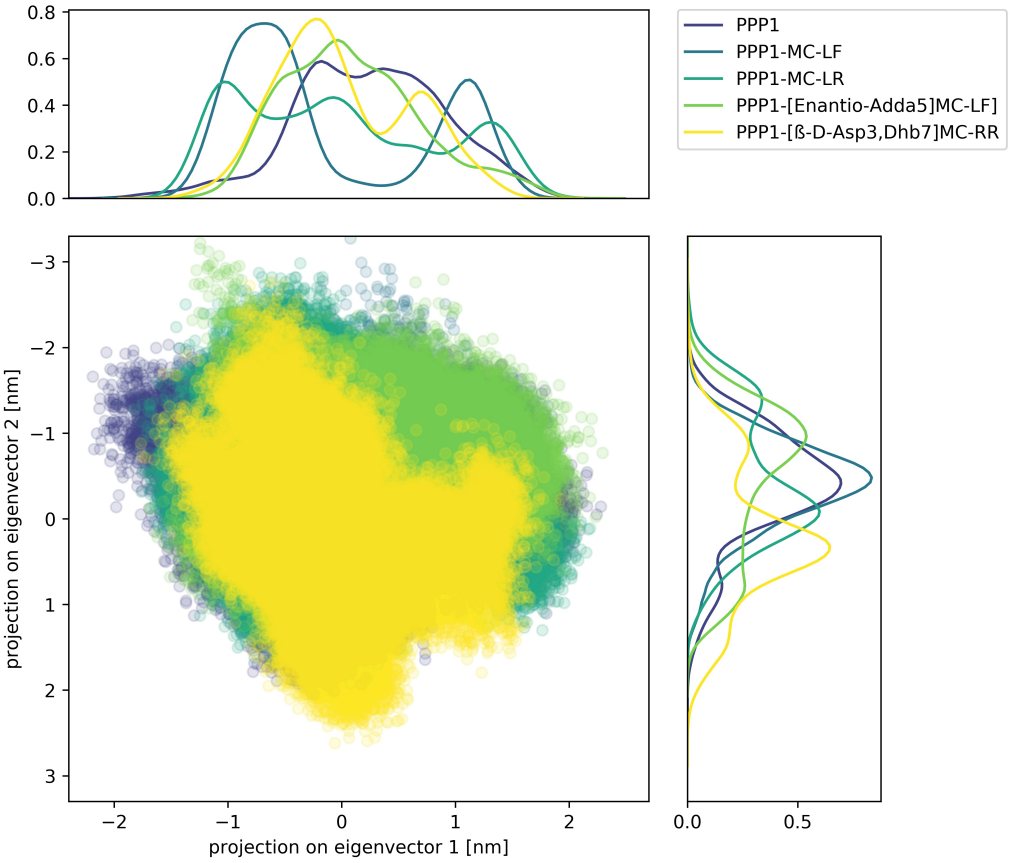


Figure 4.3: Principal components of PPP1 backbone of all complex MD simulations and apo simulation. The 2D projections are visualised by the central scatter plot, which is complemented by a distribution line at the top and right to highlight the frequency of occurrence of individual data points.

In comparison to the apo simulation (see SI Figure 8.2) the backbone structure of PPP1 of PPP1-MC-LR (see SI Figure 8.3) and PPP1-MC-LF (see SI Figure 8.4) simulation have a shifted distribution line for both principal components. Two clusters are visible in the scatter plot for PPP1-MC-LF, which suggests two major conformations. Both clusters overlap with PPP1-MC-LR and apo cluster, so it is unclear whether this really represents two clusters or just one and needs to be further investigated. The backbone of PPP1 of the non-toxic congener simulation PPP1-[Enantio-Adda5]MC-LF (see SI Figure 8.5) have a comparable distribution to apo which is slightly narrower, compared to the simulations with toxic MC congeners, suggesting that the conformation of PPP1 backbone is closer to the unbound structure when toxic MC congeners are bound. The less toxic congener simulation PPP1-[β-D-Asp3,Dhb7]MC-RR (see SI Figure 8.6) is also comparable to the apo simulation. For both simulations, PPP1-[Enantio-Adda5]MC-LF and PPP1-[β-D-Asp3,Dhb7]MC-RR principal component 1 (PC1) is similar to the apo distribution, but both are shifted along principal component 2 (PC2). In contrast, the simulations of PPP1-MC-LR and PPP1-MC-LF, which are both toxic congeners, are comparable to apo for PC2, but shifted for PC1. To summarise, one major conformational

cluster of PPP1 backbone could be identified across all simulations, which points towards a high backbone stability of PPP1, which was discussed earlier in Section 4.3.3.1 and was also reported in literature. Nevertheless, certain similarities in distribution could be addressed for the different toxicity classes.

PCA on MC Congener Backbone In addition to the PPP1 backbone, also backbone conformations of MC congeners, which basically represent different energy states [360], were analysed (see Figure 4.4). In the literature, MC-LR backbone was described as saddle-shaped conformation in water [212, 320, 324–326], which does not change significantly upon binding to PPP1 [212, 213]. In this study, MC-LR adopts to two clusters which were identified by PCA, which suggests two conformations when simulated with water (solvent simulation, see Figure 4.4a). Both conformations look similar to the saddle-shaped conformation described in literature [320, 325, 326], but the backbone is stretched differently for the bigger solvent cluster on the left side (green stick). The cluster on the right side (cyan sticks) represents a smaller cluster which is sparsely populated. It only appears towards the end of the third replicate of solvent simulation, so it needs further investigation to show whether this really is a new conformational cluster. The cluster of backbone conformations for the complex simulation is located left of the major solvent cluster. Here, MC congener backbone still has its saddle shaped structure (pink sticks), but seems to be more skewed and compact in comparison to the solvent structure. For this reason, we conclude that overall the conformation does not change much upon binding, but a distinct conformational change does occur in MC-LR backbone when simulated with PPP1. This was not found in literature so far, which could be due to the reason of much smaller simulation time.

MC-LF shows in contrast to MC-LR two clusters in solvent simulation with different backbone conformation (see Figure 4.4b). The bigger cluster on the left has a saddle-shaped backbone (green sticks) which is more bent and pushed towards a closed conformation. Overall, this one looks similar to the conformation of the MC-LR backbone structure of complex simulation. The second conformational solvent cluster is even more sparsely populated than the one for MC-LR (cyan stick) and looks less roundly shaped on the right side of the backbone. Likewise, further simulations have to prove the second conformational cluster since it is really sparsely populated and occurs only during the beginning of the first and second repetition, suggesting that it is an unstable conformation. The cluster of MC-LF backbone in complex simulation (pink sticks) is very defined and clearly separated and shifted in between the two solvent clusters. It is highly similar to the structure of the smaller solvent conformational cluster, which is probably unstable. We conclude that MC-LF in complex simulation has a similar structure to a small cluster observed in solvent, and therefore does not have a major change in conformation upon binding, but in comparison to the bigger solvent cluster has a conformational change. [Enantio-Adda5]MC-LF and [β-D-Asp3,Dhb7]MC-RR have more overlapping cluster with less distinct shape than MC-LR and MC-LF.

[Enantio-Adda5]MC-LF (see Figure 4.4c) has a widely distributed solvent cluster which is overlapping with the cluster of complex simulation, and a smaller solvent cluster which is located further apart from the other two clusters. [Enantio-Adda5]MC-LF has flipped stereocentres for the Adda side chain, making the variant non-toxic, but otherwise structurally identical to the toxic MC-LF. The MC backbone structure of the major conformational solvent cluster (green stick) is similarly shaped to the solvent structure of the small conformational cluster of MC-LF (cyan sticks, Figure 4.4b), while the backbone structure of the minor cluster (cyan sticks, Figure 4.4c) is more similar to the structure of the major solvent cluster of MC-LF (green

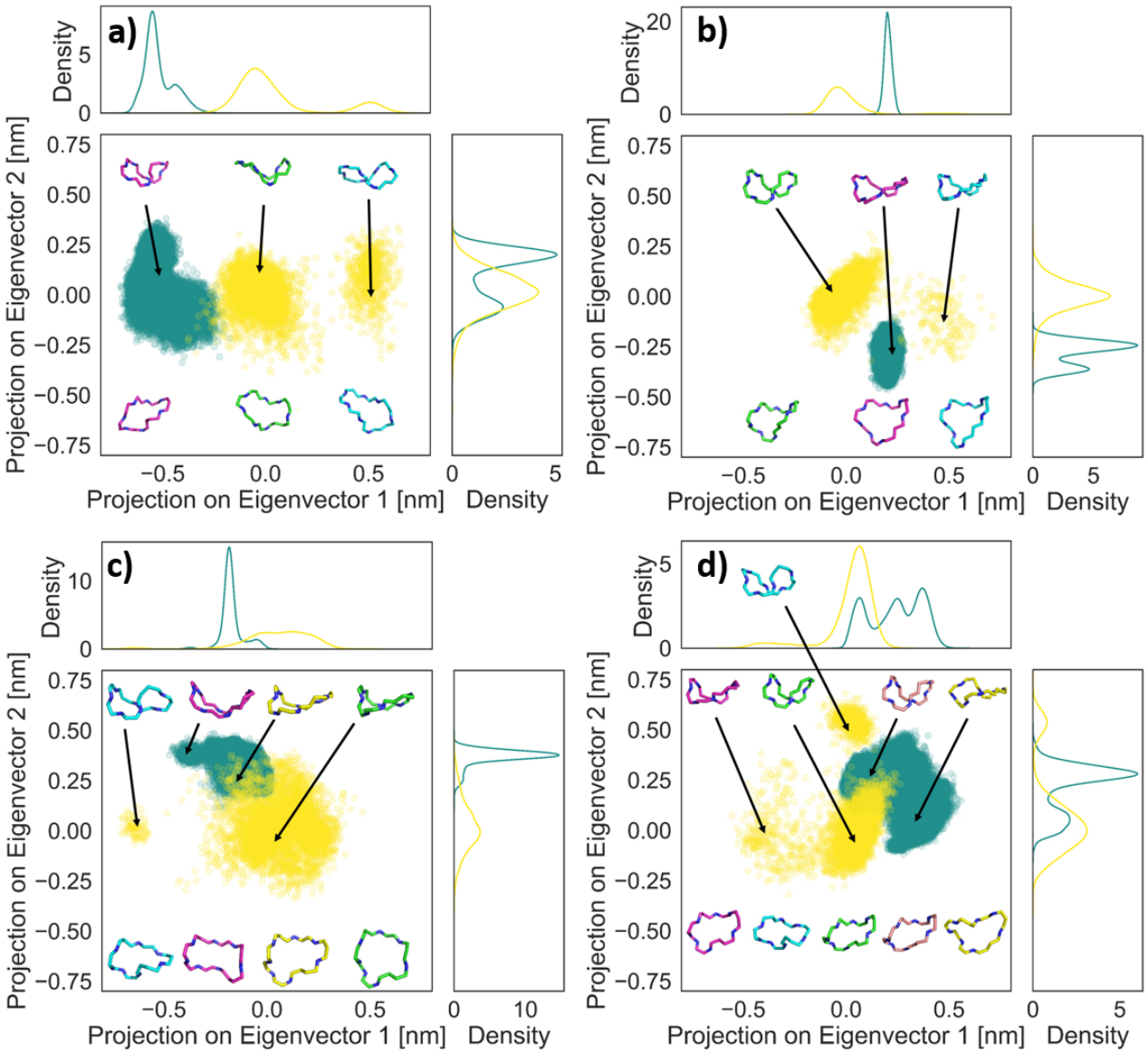


Figure 4.4: Principal components of MC congener backbone of complex (green) and solvent (yellow) MD simulations. The 2D projections are visualised by the central scatter plot, which is complemented by a distribution line at the top and right to highlight the frequency of occurrence of individual data points. The sticks represent the MC congener backbone of the different clusters, with the side view presented at the top, and the top view represented at the bottom. The MC congeners are: a) MC-LR, b) MC-LF, c) [Enantio-Adda5]MC-LF, and d) [β-D-Asp3,Dhb7]MC-RR.

sticks, Figure 4.4b). Even though we can observe similarities between the structures, they do differ in the exact conformation. For this reason, we conclude that the flipped stereocentres of Adda side chain do not only influence the side chain conformation, but also influence the backbone conformation. The cluster of complex simulation overlaps with the structure of solvent simulation. Where both of them do overlap, the conformations of complex simulation (yellow stick) looks similar to the solvent conformation (green stick). In the part of the cluster which does not overlap with the solvent simulation cluster, the conformation is a bit more bent compared to the solvent cluster, but more open and stretched in comparison to the other

conformational clusters of MC-LR and MC-LF. For this reason, we conclude that the backbone conformation of [Enantio-Adda5]MC-LF is probably not fitting as well into the binding site as the conformations of the other two toxic MC congeners.

[β-D-Asp3,Dhb7]MC-RR (see Figure 4.4d) the solvent simulation has two distinct clusters, and one sparsely populated one (pink stick) which is located close to the bigger conformational cluster (green stick) and not well separated. Most likely, they belong together. When comparing the conformations of the major (green sticks) and smaller solvent cluster (cyan sticks), the conformation is overall similar, but the smaller solvent cluster has a more bent structure and the backbone seems to be pushed together. Where the major solvent cluster (green stick) overlaps with the complex (rose stick), the structures of MC congeners backbone are very similar with no observable differences. In contrast, the representative backbone structure of the complex cluster (yellow sticks) has a differently shaped backbone than all other conformations observed before. The MC backbone is stretched flat on the right side of the ring, and overall a lot less bent than the other structures. Apparently, either the modification of position 3, where a methyl group was replaced with a hydrogen atom appears, or the arginine residue at position 2, to influence the overall backbone structure, making it more flexible compared to all other MC congeners investigated (see SI Figure 8.11 and SI Figure 8.12).

4.3.3.4 Binding of MC Congeners to PPP1

Quantitative Analysis of Binding For molecular processes (i.e., inhibition) the interaction between ligands (i.e., MC congeners) and proteins (i.e., PPP1) is crucial. For MC congeners, hyper-variable position 2 is known to provide hydrophobic interactions and the carboxyl group at position 3 builds hydrogen bonds [216, 311]. The methyl group attached to the backbone at position 5 replaces interactions with water molecules, and the Adda side chain itself was reported to be important for anchoring the MC congeners in the correct orientation in the hydrophobic groove [212, 361]. At position 6, carboxyl and carbonyl groups coordinate with the metal ions Mn^{2+} [212] and at position 7 a covalent bond is formed via the unsaturated methylene rest which is not crucial for inhibitive capacity, i.e., toxicity [212, 216, 362]. To estimate interactions and therefore binding of MC congeners to PPP1, the number of contacts closer than 0.6 nm were determined (see SI Table 8.6) and shown in Figure 4.5a. The most contacts of PPP1 to MC congeners were observed with MC-LR, with a mean value of 2583.77 $\pm$ 476.08 and a median of 2533. This is followed by MC-LF and [β-D-Asp3,Dhb7]MC-RR with a mean value of 2202.55 $\pm$ 370.54 and 2228.15 $\pm$ 335.82 and a median of 2252 and 2244, respectively. The lowest number of contacts were established between PPP1 and [Enantio-Adda5]MC-LF, with a mean value of 1863.62 $\pm$ 455.63 and a median value of 1913. Even though the mean values differ between the MC congeners, the number of contacts is not an appropriate parameter to distinguish between them, since they do not distinguish between known interaction of the binding site and general interactions somewhere in the protein. In addition, all values are within each other's standard deviations.

A more targeted approach was to calculate the number of contacts (< 0.6 nm) between the carboxylate anion at position 3 of MC congener and Mn^{2+} ions, since they are known to contribute to binding (see Figure 4.5b) and that removing the carboxylate ion at position 3 substantially reduces inhibition of PPP1 [363]. Therefore, we made the assumption that a higher number of contacts between Mn^{2+} ions and MC congener are an indicator of how well MC congener binds to PPP1. A higher number of contacts was observed for the toxic MC congeners MC-LR and MC-LF with a mean value of 13.60 $\pm$ 4.02 and 24.42 $\pm$ 8.83 and a median of 14 and 23, respectively, in comparison to the other MC congeners. The less-toxic

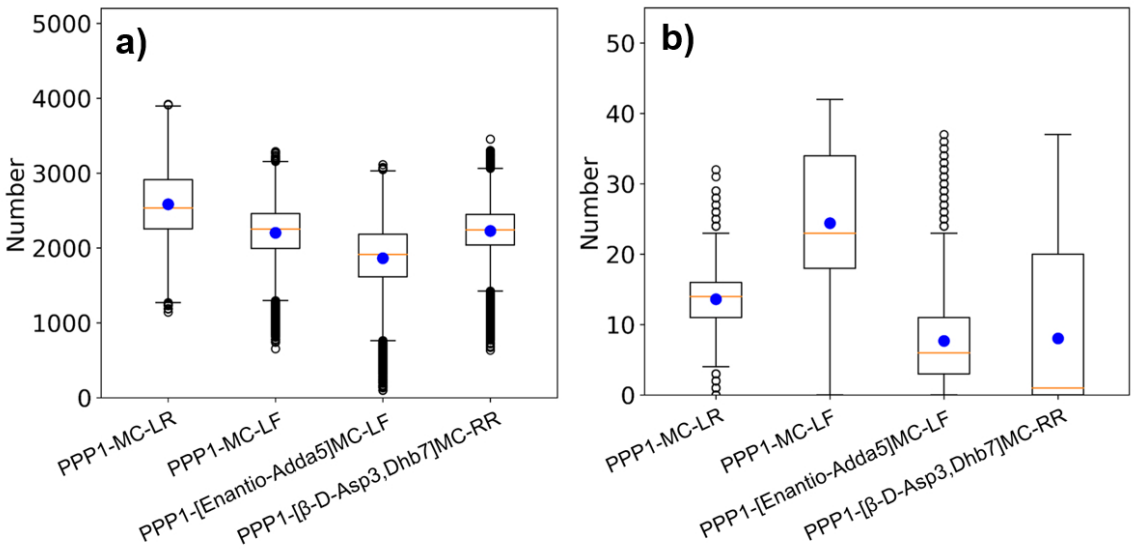


Figure 4.5: Distribution of number of contacts (< 0.6 nm) during MD simulation. The median is marked with an orange line, while the mean value is shown as a blue circle. a) number of contacts between PPP1 and MC congener and b) number of contacts between Mn^{2+} and MC congener.

congener [β-D-Asp3,Dhb7]MC-RR has a median of 1 and a mean value of 8.05 ± 9.99, and the non-toxic congener [Enantio-Adda5]MC-LF, has a median of 6 and a mean of 7.69 ± 6.15. Even though the mean and median number of contacts between Mn^{2+} and [Enantio-Adda5]MC-LF was lower than for the toxic congeners, the distribution and interquartile range of 0 to 20 is very broad, we could show that it highly differs from the number of contacts of toxic congeners, especially its enantiomer MC-LF. The less-toxic MC congener [β-D-Asp3,Dhb7]MC-RR has also a low number of contacts with Mn^{2+} ions. The number of contacts is still in the standard deviation of the toxic congener MC-LR, but the median values of 1 and 14 highly differ, which indicates less stable binding for [β-D-Asp3,Dhb7]MC-RR. For the number of contacts to Mn^{2+} ions, the same trend could be observed for the statistical analysis with linear least-squares regression (see SI Table 8.7).

Qualitative Analysis of Binding The molecular interactions were analysed to investigate the binding of MC congeners to PPP1. In the following, interactions between MC congeners and PPP1 are shown. Known interacting residues are shown as pink sticks, and residues interacting in the simulation (i.e., proximity of <4 Å) are shown as green lines. The highly toxic MC congeners MC-LR and MC-LF [16] are known to interact well with the PPP1 residues described in Fontanillo and Köhn [216] which could be also seen in our simulations (see SI Figures 8.7 to 8.9). The calculations of distances with known interacting residues were lower than for the less and non-toxic MC congener (data not shown). [Enantio-Adda5]MC-LF differs in four stereocentres of the Adda residue in comparison to the toxic MC-LF, which leads to a flipped orientation of [Enantio-Adda5] being oriented above the backbone instead of stretched away from the backbone (see SI Figure 8.10). When simulating [Enantio-Adda5]MC-LF, in the beginning of the uncut trajectory of replicate 3, [Enantio-Adda5]MC-LF is coordinated in the binding site (see Figure 4.6). It then starts moving outside the binding site by flipping over the Enantio-Adda5 residue (see Figure 4.7a), since it is not stable bound to the protein residues

(Cys127, Ile130, Ile133, Tyr134, Trp206) [216] but essential for binding to PPP1 [212].

Afterwards, it flips right back into the binding site instead of diffusing away from the protein (see Figure 4.7b and Figure 4.8a), but with Adda side chain oriented towards the wrong side of the binding site, where it remains for the rest of the simulation (see Figure 4.8b). These results suggest a less stable interaction of [Enantio-Adda5]MC-LF with PPP1, but nevertheless we could only observe this effect for the last replicate, not for all three replicates.

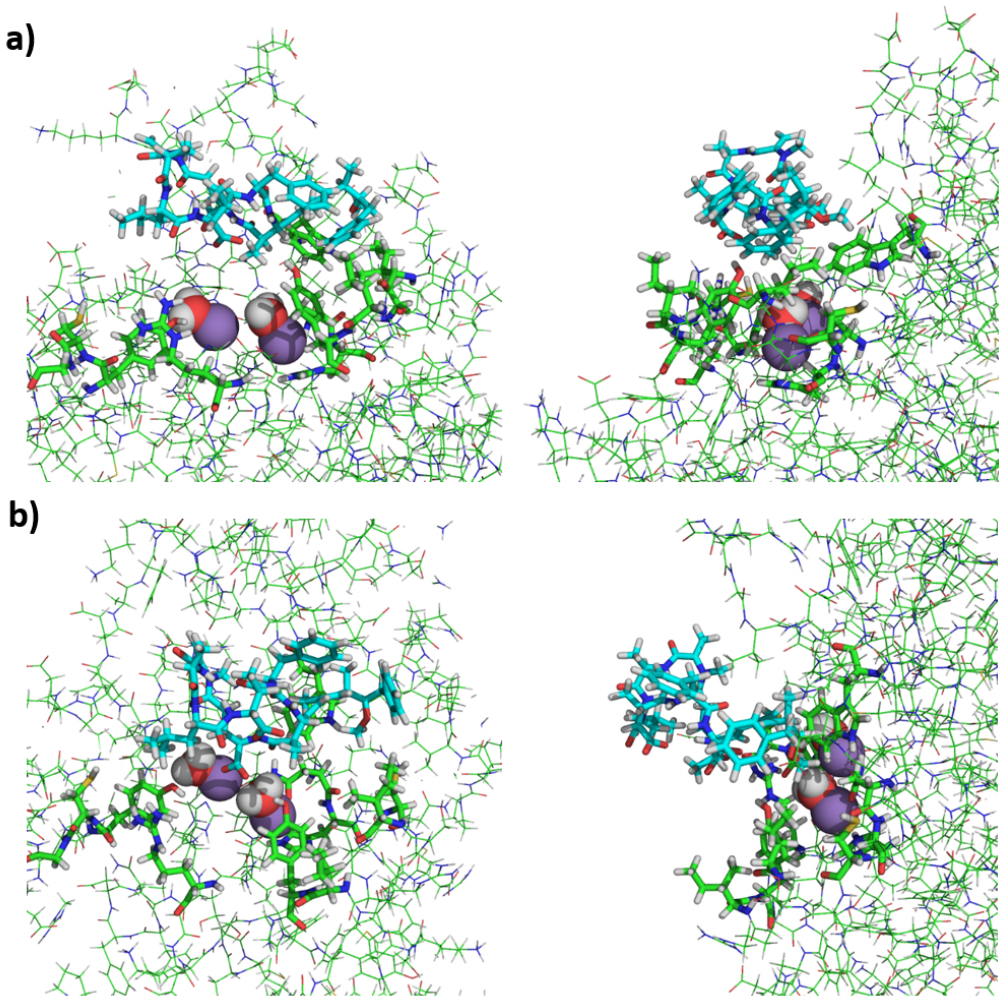


Figure 4.6: Interaction of PPP1 (green) and [Enantio-Adda5]MC-LF (cyan) uncut trajectory. Interacting residues are shown as green lines, residues interacting described in literature are displayed in pink. Mn^{2+} and the active site waters are shown as spheres. The images show the uncut MD trajectory of interaction of [Enantio-Adda5]MC-LF with PPP1 in top and side view. a) is a snapshot at the beginning and b) towards the end of the uncut MD trajectory (first 30 ns).

In comparison to the other MC congeners, [β-D-Asp3,Dhb7]MC-RR is not only modified at the hyper-variable position 2, but also demethylated at position 3, and methylated at the unsaturated carbon atom at position 7. The modifications described here lead to a more open structure in comparison to the other MC congeners, and demethylation at residue 3 seems to influence the backbone conformation (see SI Figure 8.11 and SI Figure 8.12). In contrast to the non-toxic [Enantio-Adda5]MC-LF, [β-D-Asp3,Dhb7]MC-RR is coordinated stable into the binding site and does not move out of it (see Figure 4.9).

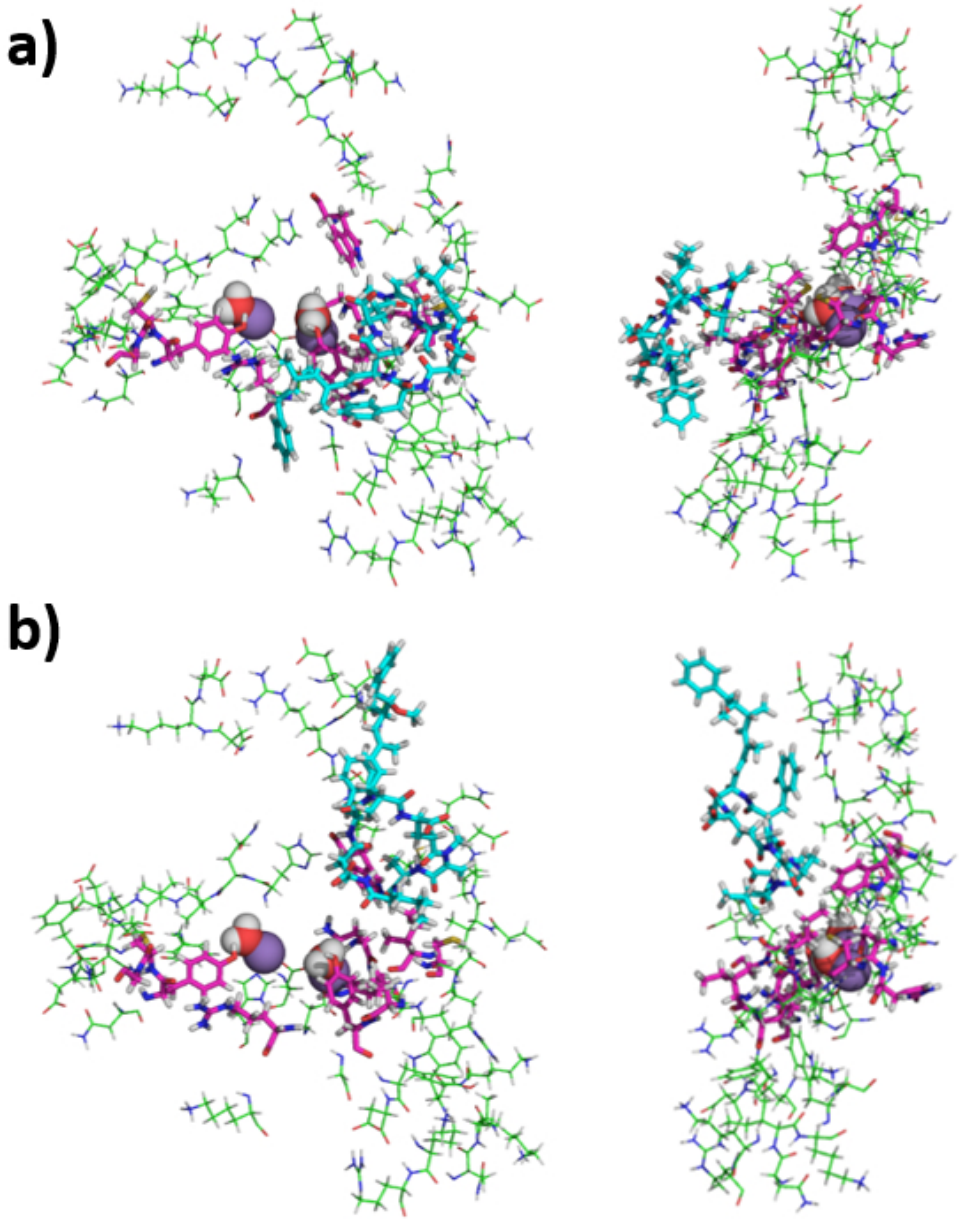


Figure 4.7: Interaction of PPP1 (green) and [Enantio-Adda5]MC-LF (cyan) flip out of the binding site. Interacting residues are shown as green lines, residues interacting described in literature are displayed in pink. Mn^{2+} and the active site waters are shown as spheres. The images show the trajectory of the system in top and side view, and a) is a snapshot at 89 ns and b) at 102 ns.

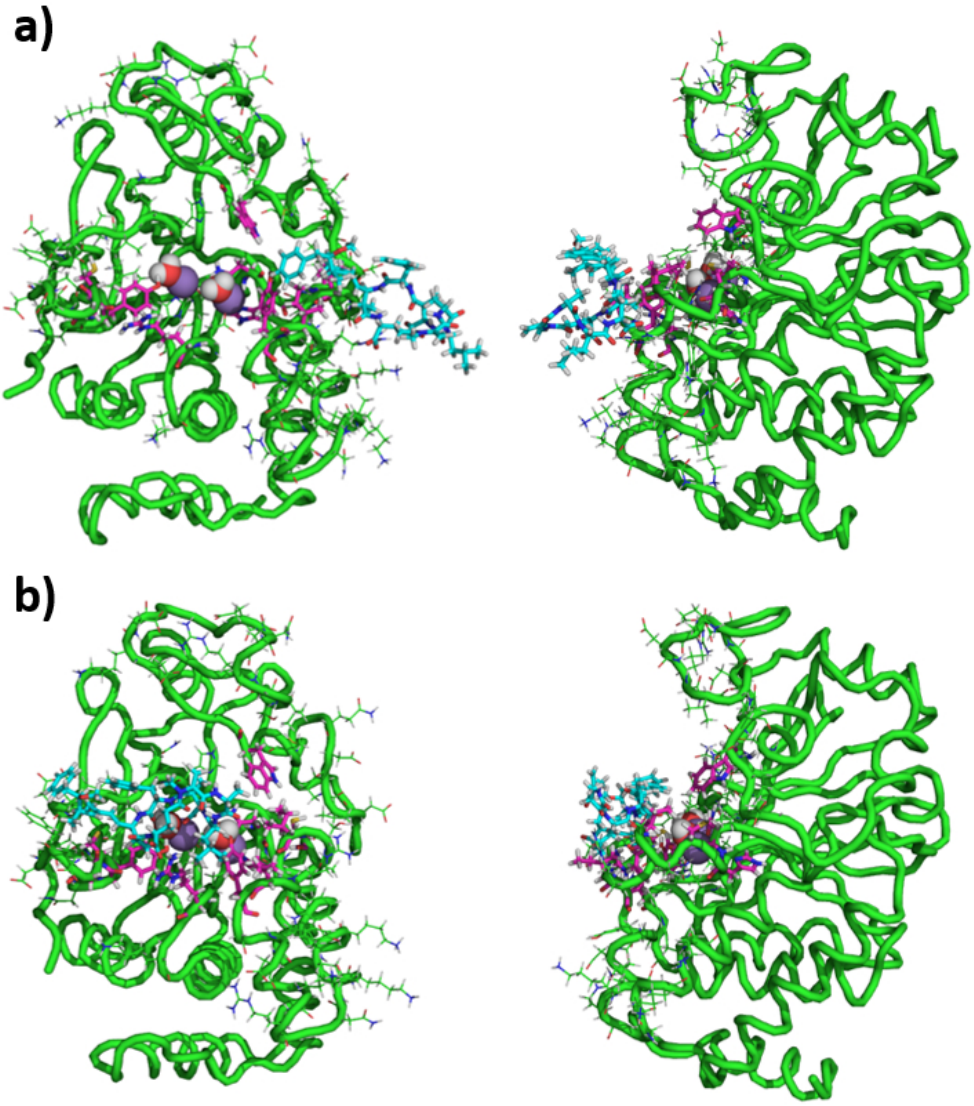


Figure 4.8: Interaction of PPP1 (green) and [Enantio-Adda5]MC-LF (cyan). Interacting residues are shown as green lines, residues interacting described in literature are displayed in pink. Mn^{2+} and the active site waters are shown as spheres. The images show the trajectory of the system in top and side view and a) are a snapshot at the beginning and b) of the third replicate.

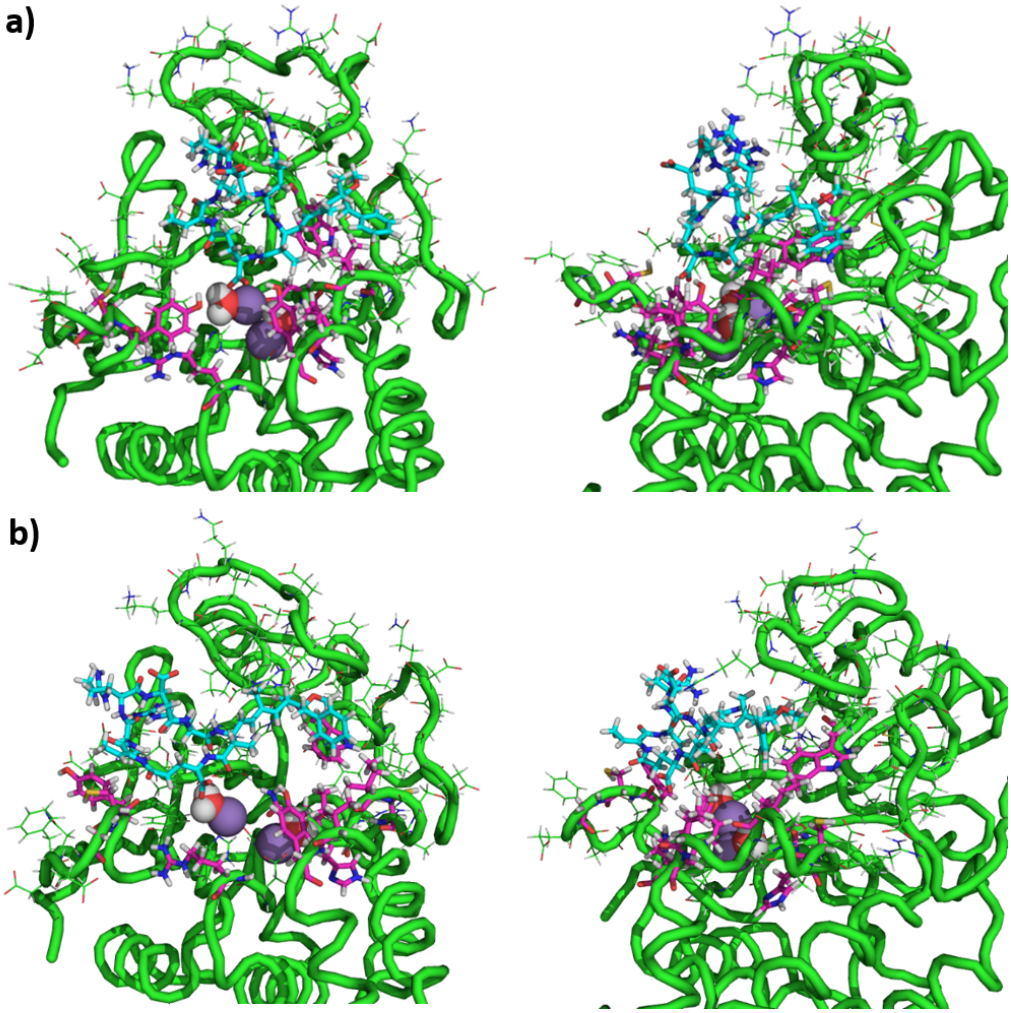


Figure 4.9: Interaction of PPP1 (green) and [β-D-Asp3,Dhb7]MC-RR (cyan). Interacting residues are shown as green lines, residues interacting described in literature are displayed in pink. Mn^{2+} and the active site waters are shown as spheres. The images show the trajectory of the third replicate in top and side view and a) is a snapshot at the beginning and b) towards the end of the third replicate.

4.4 Molecular Dynamics Simulation of 12 MC Congeners and Conjugates

4.4.1 Introduction

After simulating and studying four MC congeners, a larger data set of twelve MC congeners with conjugates was set up. The aim of the simulations was to get a better understanding of MC congener conformation and binding to PPP1 and how it relates to their toxicity values and modifications. The MC congeners selected for simulation were also based on the inhibitory concentration (IC_{50}) values determined by colorimetric phosphatase assay and cover a wide range of IC_{50} values (see Table 3.2 in Section 3.3).

The structures of MC congeners selected (see Table 4.3) are described in Table 3.1 and in Table 4.2. Two new IC_{50} values were measured for MC congeners which were used in this study, as both are non-toxic congeners which are the minority group [364]. Furthermore, not only MC congeners but also metabolised variants of Cys and GSH conjugates were simulated in both Cahn-Ingold-Prelog (CIP) configurations (R and S) to obtain an understanding of the interaction of conjugated MC congeners with PPP1. Both CIP configurations were chosen because the orientation of Cys and GSH attachment to MC congeners is unclear.

To allow simulations of such a large data set, the MC congener parameterisation method has been modified, compared to Section 4.3, to be less time-consuming. The PPP1 and MC-LR crystal structure (PDB Code: 1fjm [212]) available in the Protein Data Bank (PDB) database [328]) was chosen as starting structure. For all other MC congeners, no crystal structure is available in complex with PPP1, so docking was used to place the MC congeners in the binding site.

Two MD simulation approaches were chosen for the data presented here: 1) a simulation of the complexes of PPP1 and MC congener to study stability and conformation of PPP1, MC congener and their conjugates and 2) a (TIP3P) solvent simulation to explore the macrocycle backbone conformation and properties of MC congeners and conjugates. The aim of this study was to investigate the structural stability of PPP1 and MC congeners, as well as different conformations. The binding properties are not analysed here, as the data set is simply too large to view every simulation in detail. For an analysis of interaction and binding, the reader is referred to Section 5, where an approach to determine and evaluate interactions automatically was developed.

Table 4.2: Structural information and IC_{50} value of new MC congeners used in this MD simulation study [364].

Congener	X	Z	R_1	R_2	Adda	IC_{50} [nM]
[Apha5]-MC-LF	Leucine	Phenylalanine	Methyl	C=CH_2	(2S,3S)-3-amino-2-methyl-7-phenylheptanoic acid	138,261
[Apda5]-MC-LF	Leucine	Phenylalanine	Methyl	C=CH_2	(2S,3S)-3-amino-2-methyl-10-phenyldecanoic acid	1280

4.4.2 Methods

4.4.2.1 Structure Preparation

PPP1 preparation The structure of PPP1 was downloaded from the PDB (PDB code: 1fjm [212]). The hydrogen atoms were added for a pH of 7.4 using Avogadro and the structure model corrected by using SWISS-Model server [365–370] to add 61 atoms which were originally missing in the structure. Afterwards, hydrogen atoms were added and orientation of asparagine, glutamine and histidine analysed with Molprobity [332, 333] and recommended flips performed. The N- and C-terminus of PPP1 were capped using PyMol [371] with ACE and NME residue, respectively.

MC structure preparation MC-LR was extracted from PDB structure 1fjm by deleting the covalent bond between MC-LR and residue Cys273 of PPP1. Hydrogens were added for pH 7.4 with Avogadro [130]. To generate all other MC congener structures (see Table 4.3), the structure of MC-LR was modified with PyMol [371]. To generate metabolised MC congeners with GSH or Cys all structures, except for (D-Asp3,Dhb7)MC-RR which is likely not conjugated due to methylation of the double bond [33], were modified at the double bond of N-Me-DHa [210] in R and S (CIP) configuration, as it is unclear which metabolites are built [372, 373]. For MC congener conjugates, the nomenclature in this chapter will be adapted as follows: the name of the MC congener is followed by the conjugate (Cys or GSH), and the CIP configuration (R or S). For example, MC-LR conjugated with Cys in R configuration will be denoted as MC-LR-Cys-R. The structures were then minimised with UCSF Chimera [329] with 5000 steps, 500 conjugate steps at a step size of 0.01 to remove unfavourable clashes and search for an energy minimum of the structure. Then, Gasteiger partial charges were added. Avogadro was then used to optimise the mol2 file [130].

4.4.2.2 Docking

Preparation To generate a structure of MC congeners and conjugates with PPP1, docking was applied for MC congeners two to twelve, and for all MC conjugates, because no crystal structure was available. The structures were prepared as described in Section 4.4.2.1 and an MD simulation with MC congeners and conjugates in solvent was run to generate two more structures, so that a total of three different structures per MC congeners and conjugates were available which were used for docking. For MD simulation, the mol2 file atoms and corresponding bonds in mol2 file were sorted in ascending order with a script published by Justin Lemkul on `http://www.mdtutorials.com/gmx/complex/Files/sort_mol2_bonds.pl`.
To generate MC congener topology files, CHARMM General Force Field (CGenFF) version 2.4.0 was used [374–376] and converted to GROMACS file format.

MC congener simulations in water were performed for 30 ns in the production run. All other parameters and steps were performed identical to the procedure for solvent simulation described in Section 4.4.2.3. After the simulation, GROMACS cluster analysis was performed with a cut-off of 0.3 nm and average middle structures extracted from the first two clusters, if two or more clusters were present.

Procedure For docking with AutoDock Vina (v1.1.2) [377], bond order and connectivity were corrected and explicit polar hydrogens added in PyMOL. As previously, UCSF Chimera [329]

Table 4.3: MC congener data set and poses selected from AutoDock Vina docking for MD simulation of MC congener and respective conjugates. The background colour of the row indicates toxicity class: the toxic class is grey, less toxic is blue and non-toxic is white.

MC Congener	Congener	Cys-R	Cys-S	GSH-R	GSH-S
MC-LR	x	x-ray	cluster 2	cluster 2	cluster 2
MC-LF	x-ray	cluster 1	x-ray	cluster 2	x-ray
[Enantio-Adda5]-MC-LF	x-ray*	x-ray	cluster 2	x-ray	x-ray
[β-D-Asp3,Dhb7]-MC-RR	x-ray*	-	-	-	-
MC-RR	cluster 1	cluster 1	cluster 1	cluster 2	cluster 2
MC-LY	cluster 1	x-ray	cluster 2	x-ray	x-ray
MC-YR	x-ray	cluster 1	cluster 1	cluster 1	x-ray
MC-LY(Prg)	cluster 2	cluster 2	cluster 1	cluster 1*	cluster 2
[Anda5]-MC-LY(Prg)	cluster 2	cluster 1	cluster 2*	cluster 1	x-ray
[Amba5]-MC-LY(Prg)	cluster 1	cluster 2	cluster 1	x-ray	x-ray
[Apha5]-MC-LF	cluster 1	cluster 2	x-ray	cluster 2	x-ray
[Apda5]-MC-LF	x-ray	cluster 1	cluster 1 and cluster 2***	x-ray x-ray	x-ray and cluster 1 **

x: Not applicable, crystal structure available.
- : Not applicable, cannot be conjugated.
* Identical affinity for two poses, correct orientation preferred.
** Identical affinity for two poses with wrong orientation, x-ray orientation preferred.
*** Identical affinity of two poses with same orientation. Cluster 1 selected as it represents majority of structures.

was also used here to minimise the structure with identical settings and mol2 file format fixed with Avogadro [130].
Before docking, Autodock Tools (v1.5.7) [331] was used to generate a pdbqt file of PPP1 (preparation of structure described in 4.4.2.1) and charges for manganese ions and water were corrected manually. For all MC congener structures, torsion trees defining bond rotation and flexibility for side chain moieties of MC congeners were set on, and MC backbone was considered as rigid. The grid box was centred around the active site at 91.226, 23.163 and 24.424 with dimensions of 30 x 22.5 x 18.75 Å, in x, y and z direction, respectively. Exhaustiveness was set to 16. The complex with the highest affinity values was selected. If two complexes had identical affinity, an orientation most similar to MC-LR crystal structure was favoured, if Adda residue was unmodified. If a structure with Adda residue resulted in a different conformation from x-ray structure, the pose with the highest affinity and correct orientation was favoured, as we know that Adda highly influences the binding of MC congeners [216] and serves as an anchor in the binding site. For MC congeners with modified Adda residue or MC congener conjugates, binding poses were not corrected, since this modification might change the binding mode as the anchoring function of Adda is lost. In addition, the size of MC congener conjugates is increased a lot, which induces steric hindrances [210, 321]. The selected poses from docking are summarised in Table 4.3.

4.4.2.3 Molecular Dynamics Simulation

Preparation MD simulations were performed with GROMACS version 2020.2 with Docker images loaded from NVIDIA `nvcr.io/hpc/gromacs:2020.2` and CHARMM36 forcefield pub-

lished July 2020 [374–376, 378–381]. For PPP1 protein structure, CHARMM36 force field needed to be adjusted. It caused compatibility issues with GROMACS for NME and ACE caps (described in Section 4.4.2.1). For this reason, atom nomenclature was changed for both residues for GROMACS compatibility, as suggested in GROMACS User discussions `https://gromacs.bioexcel.eu/t/additive-forcefield-of-charmm36/489/5`.
The CHARMM36 force field has no parameters for manganese ions. The parameters for manganese ions were also adjusted here as for the simulations described in Section 4.3.2.3 and parameters from magnesium ions were used, since manganese ions are only slightly larger and have similar coordination preferences to magnesium ions [334]. The parameters for ϵ were set to 0.277000 kJ and σ to 0.01234280 Å [382]. The exchange of parameters leads to stable coordination in the binding site. MC congeners were parameterised as described in Section 4.4.2.2.

Procedure Periodic boundary conditions with a distance of 1.0 between solute and box were modelled as dodecahedron. TIP3P water model [344] was used as water solvent model. The system was neutralised with Na^+ and Cl^-. A Verlet cutoff-scheme [346] was applied with Particle-Mesh Ewald (PME) electrostatics [345]. Van der Waals interactions were cut-off at 1.2 nm. The leap-frog integrator was used and LINCS [347] applied to constrain bonds to hydrogens.

For energy minimisation, the steepest descent algorithm with 50,000 steps and a step size of ≤ 0.01 with a tolerance of $< 10.0\,\mathrm{kJ\,mol^{-1}}$ was applied. For energy minimisation and equilibration with an MC congener or conjugate, position restraints of $1.000\,\mathrm{kJ\,mol^{-1}\,nm^2}$ on all heavy atoms in all axes were applied. NVT equilibration (isothermal-isochoric) to stabilise temperature was performed for 100 ps with a time step of 2 fs and temperature coupled with a randomised velocity rescaling [348] to 300 K. Initial velocities were generated randomly from a Maxwell distribution at 300 K. NPT equilibration to stabilise pressure was performed for 100 ps with a time step of 2 fs using the Berendsen isotropic pressure coupling [383] to 1.0 bar.

For the production run, different simulation types were set up: 1) solvent simulation (MC congeners and conjugates in water), 2) apo simulation (PPP1 in water) and 3) complex simulations (MC-congener or conjugate bound to PPP1 in water). The simulation times of the production runs for each simulation type is summarised in Table 4.4.

Table 4.4: Summary of employed MD simulations and total simulation time.

Simulation Type	System	Simulation Time
Solvent	MC-congener/conjugate	35 ns
Apo	PPP1	530 ns
Complex	PPP1-MC-congener/conjugate	530 ns

The step size for all simulation was 2 fs, except for simulation of MC-LR-Cys-R conjugate, where the step size was decreased to 1 fs due to difficulties with vibrating bonds. The output was logged every 10 ps.

For the production run, Parrinello-Rahman barostat [349] and randomised velocity rescaling [348] as thermostat was used. For each simulation type, three replicates were set up. For the solvent simulation, the structure of MC congener or conjugate from the complex was extracted and used as starting point. To get starting structures for replicates 2 and 3, GROMACS cluster analysis was applied to replicate 1 of complex simulation with the gromos method (pairwise

distances) and a cut-off of 0.25 nm. Starting structures for replicate 2 and 3 for complex and solvent simulation were the average structure of the two most populated clusters.

For the apo simulation, replicate 1 was started with the prepared structure of PPP1. GROMACS cluster analysis was applied to replicate 1 with a cut-off of 0.21 nm and the average structure of the two most populated clusters were selected for replicates 2 and 3. Those structures were aligned to the crystal structure used in replicate 1 with PyMOL [371].
For the complex simulations, the crystal structure of MC-LR with PPP1 was used for replicate 1 and for all other MC congeners the docked pose. For replicate 2 and 3 the aligned structures of PPP1 were used to generate new starting structures of PPP1 MC-congener or conjugate complex to start replicate 2 and 3.

Analysis The analysis of the simulation was mostly performed with GROMACS functions. To analyse only well-equilibrated trajectories, all simulations were cut off in the beginning. For the solvent simulations 5 ns were cut off, for apo and complex simulation 30 ns of the trajectories were cut off.

After centering the protein structure, rotational and translational motions were eliminated using a least-squares fit and periodic boundary conditions were removed. The root-mean-square deviation (RMSD), volume, radius of gyration, and solvent accessible surface area were determined for PPP1 and MC congener. Principal Component Analysis (PCA) [350] was used to analyse the backbone structures of PPP1 and MC congeners. For PPP1 backbone analysis, eigenvalues, and eigenvectors (i.e., principal components) were derived from all replicates of the apo simulation and projected on all complex simulations. For MC-congener or conjugate backbone, eigenvalues and eigenvectors were derived from all replicates of the respective solvent simulation and projected on the respective complex simulation. To visualise the result of PPP1 backbone projection, Normal Mode Analysis was used as provided in a script for PyMOL [371, 384]. Normal Mode Analysis describes the flexible states accessible at an equilibrium state to a protein structure. The result are arrows which visualise the movement and dynamics along the eigenvectors 1 and 2 (derived from PCA) based on the extreme conformations (first and last) sampled during the simulation along the respective eigenvector. These images will be called porcupine plots in the following.

To visualise the results of PCA analysis for projection on MC-congener and conjugate backbone, Mol2vec (v0.1) helper function [104] was used to plot the projections along the first two eigenvectors as a scatter plot complemented by two density histograms to show the distribution of data points along both eigenvectors. A 3D density mesh grid was calculated over the data points (every 50^{th} and 5^{th} data point for complex and solvent simulations, respectively) to estimate where most of the data points are located to derive representative structures. The smoothing parameter k was set to 20 for all simulations, except the solvent simulations of MC-YR-Cys-S and [Apha5]-MC-LF-GSH-S, as the resulting 3D surface was not appropriately shaped. Therefore, k was set to 24. The structure closest to the individual maxima was determined to be a representative structure. The positions on the scatter plot of the representative structures were plotted as circle (solvent) or triangle (complex). After obtaining the representative MC congener and conjugate structure, all structures were aligned with Wizard Pair Fitting in PyMOL [371] to the MC-LR backbone structure obtained from the crystal structure (PDB code: 1fjm [212]) and exported as image. Python programming language [351] was used to calculate properties and visualise data. NumPy (v1.21.2) [352] was used to calculate mean values and standard deviations over all replicates. Jscatter, (v1.2.7.2) [355] is a python

wrapper to plot data with Grace [356] and was used to read xvg files obtained by GROMACS analysis function. Matplotlib (v3.5.1) [354] was used to visualise time series data and the three replicates were merged into one trajectory for better comparison. Each 25^{th} or 2^{nd} data point was plotted for the complex and solvent simulations, respectively, to reduce visual clutter in the time series visualisation.

4.4.3 Results and Discussion

The analysis methods used on this data set are as close as possible to the ones used for the small data set (see Section 4.3.3). There are two main reasons for the differences that could be observed between the simulation sets: 1) the force field used for this simulation differs from that of the small data set, and 2) each replicate of each simulation was simulated twice as long as before.

The MD simulation data set analysed here is much larger than before. For this reason, some analysis methods have been adopted compared to the previous section in order to analyse the data in a compact way and make it understandable and readable for the reader. For the MC congener conjugates, we can only compare the MD simulations with those of the MC congeners, since there is no reported work on MD simulation of MC congener conjugates, only analysis of interactions obtained by docking are reported in the literature [210, 321].

4.4.3.1 Stability of PPP1

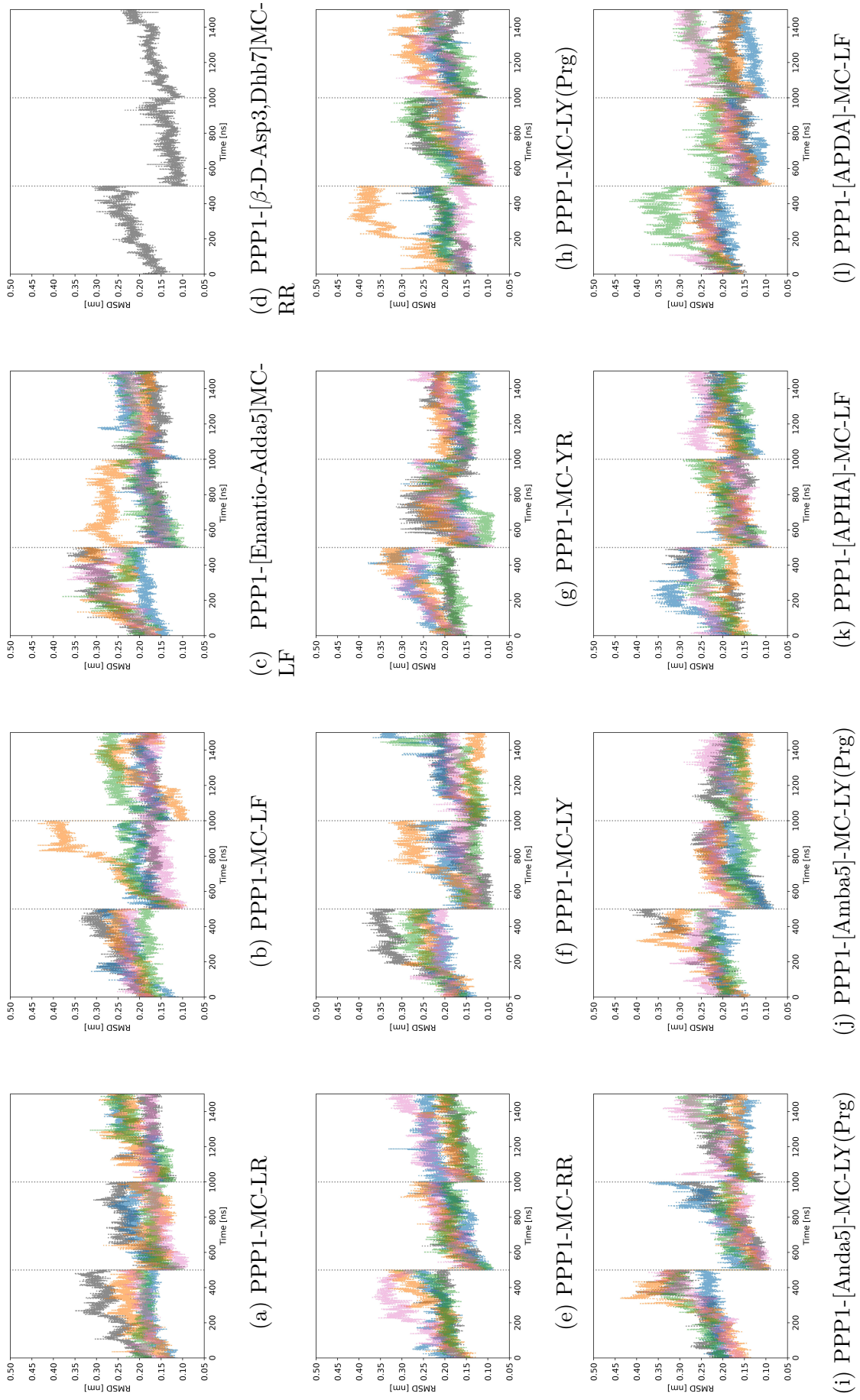
The properties calculated to estimate the structural stability of PPP1 and described below are volume, solvent accessible surface area, radius of gyration and root mean squared deviation. These properties give an estimate of how compact and folded the protein is, how much it is accessible to solvent and how much the protein structure does change upon MC congener binding (see SI Table 8.8). Independent of MC congener binding, PPP1 has similar properties.

The volume of PPP1 has a mean value of 60.83 to $61.11 \frac{nm}{S^3/N} \pm 0.57$ to $0.72 \frac{nm}{S^3/N}$ and the time series data look similar over the whole trajectory, i.e., no major changes are visible (data not shown, see SI Table 8.8). For MC congener conjugates, the volume of PPP1 has mean values from 60.34 to 61.18 ± 0.54 to $0.74 \frac{nm}{S^3/N}$, which is similar to the volume when PPP1 is simulated with MC congeners and apo simulation with a mean value of $60.83 \pm 0.61 \frac{nm}{S^3/N}$. The same can be observed for the solvent accessible surface area of PPP1, no differences in time series are visible and mean values range from 142.8 to 146.06 ± 2.71 to $4.36 \frac{nm}{S^2/N}$ for MC congeners and 140.43 to 145.89 ± 2.58 to $4.55 \frac{nm}{S^2/N}$ for MC congener conjugates. PPP1 apo simulation has a comparable value with a mean of $143.99 \pm 3.09 \frac{nm}{S^2/N}$. The radius of gyration has also stable mean values of 1.88 to 1.91 ± 0.01 to 0.02 nm for all simulations. Nevertheless, in the time series visualisation of the radius of gyration (see SI Figure 8.14) differences can be observed for the individual replicates. For replicate 1 the radius of gyration seems to be lower on average across the MC congeners and conjugates (except for 5 simulations, namely PPP1-MC-LY(Prg)-Cys-R,PPP1-[Anda5]-MC-LY(Prg) and PPP1-[Anda5]-MC-LY(Prg)-Cys-R, PPP1-[Apha5]-MC-LF-Cys-S and PPP1-[Apda5]-MC-LF-GSH-S). For the apo simulation, the trend is exactly opposite, with replicate 3 having lower values on average.

The RSMD values for the backbone of PPP1 indicates the structural stability of PPP1 compared to the beginning of the simulation (reference structure). Here, the backbone RMSD values of the apo simulations results in a mean value of 0.21 ± 0.05 nm. For all other complex simulations with MC congener, the mean value ranges between 0.17 and 0.22 nm, with standard

deviations between 0.02 to 0.06 nm. The backbone structure of PPP1 is therefore considered to be highly stable. However, the time series visualisation shows an increase for most MC congeners complex and the apo simulation during the first replicate. In particular, after 250 ns, there is an increase in the RMSD (see SI Figure 4.10). For the conjugate simulations, we also observe a stable structure of the backbone PPP1 and see the same trend with an increase in the RMSD during the first replicate. For some conjugates we have a peak in the RMSD during one replicate, so there may be some structural rearrangement for some Cys-R conjugates (orange line, PPP1-MC-LF, PPP1-[Anda5]-MC-LY(Prg), PPP1-MC-LY(Prg), PPP1-[Amba5]-MC-LY(Prg)), one Cys-S conjugate (blue line, PPP1-[Apha5]-MC-LF) and one GSH-S conjugate (green line, PPP1-[Apda5]-MC-LF). The mean and standard deviations range from 0.16 to 0.25 and 0.02 to 0.07, respectively.

Overall, all mean values and standard deviations of the four properties give stable values for PPP1. Therefore, the structure of PPP1 is considered stable throughout the simulations, which was previously observed in the simulation set on four MC congeners (see Section 4.3.3.1), but not shown so far for MC congener conjugates. On average, the values are slightly higher in this simulation data set, which can be explained by the length of simulation and different force field. Here, the simulations are almost three times as long as before, so these values are considered similar to the ones before and confirm the result of the previous simulations. If we compare the RMSD values with the simulations on the smaller data set (see Section 4.3.3.1), we see that the means are lower for the smaller set (below 0.20 nm), and we could not observe an increase in the RMSD for replicate 1 of previous simulations as we do here. Again, this observation could be due to longer simulation times. In addition, the RMSD is a calculation against a reference structure, which is the start of the simulation, and the crystal structure of PPP1. Maybe some further equilibration in the starting structure was necessary, which of course has a higher increase in RMSD over time. For replicate 2 and replicate 3 we cannot see the same trend, so the structure was probably better equilibrated by then. The same trend is observed for the apo simulation. The RMSD increases during the first replicate, while it is lower for replicates 2 and 3. In contrast, the radius of gyration gets lower for the third replicate, indicating a more compact folded structure (see SI Figure 8.13). Deletion of the MC-LR from the binding site before starting the simulation may lead to this result. Some minor rearrangements are likely to occur as the ligand is no longer bound to the binding site. MD simulations have not yet been performed for MC congener conjugates, so there is no data in the literature or from our own work to compare with. However, on average the PPP1 structure has a similar stability to the MC congener simulation, although we see some structural rearrangements for some MC congener conjugates, but not for all of them. A closer analysis of PPP1 backbone structure is discussed in Section 4.4.3.3 and analyses the conformational space in more detail. The results here indicate an overall stable backbone structure of PPP1 independent of the simulation.



(a) PPP1-MC-LR (b) PPP1-MC-LF (c) PPP1-[Enantio-Adda5]MC-LF (d) PPP1-[β-D-Asp3,Dhb7]MC-RR

(e) PPP1-MC-RR (f) PPP1-MC-LY (g) PPP1-MC-YR (h) PPP1-MC-LY(Prg)

(i) PPP1-[Anda5]-MC-LY(Prg) (j) PPP1-[Amba5]-MC-LY(Prg) (k) PPP1-[APHA]-MC-LF (l) PPP1-[APDA]-MC-LF

Figure 4.10: Time series of PPP1 backbone RMSD of different MC congeners and their respective conjugates. The three replicates are shown as one timeline and are separated by a grey dashed line at 500 and 1000 ns. The MC congeners and conjugates are shown in different colours: unconjugated in grey, Cys-S in blue, Cys-R in orange, GSH-S in green and GSH-R in pink.

4.4.3.2 Stability of MC Congeners

The stability of the MC congener was investigated using the same properties as for PPP1, so volume, radius of gyration, solvent accessible surface area and RMSD were calculated. Therefore, changes in MC congener structure upon binding to PPP1 were investigated by comparing the properties to the ones derived from solvent simulation (see SI Table 8.9). In summary, for the conjugated MC congeners there is a large effect on the volume, radius of gyration and solvent accessible surface area. Since Cys and GSH increase the number of atoms and therefore the size of the MC congener, the GSH conjugates have the largest properties, followed by Cys conjugates and MC congeners. Time series visualisation of volume and solvent accessible surface area (data not shown), shows only minor changes for solvent and complex simulation. Comparing the means and standard deviations of the four MC congeners simulated previously, they are overall similar and only slightly higher in the simulations presented here (see SI Table 8.4 for previous and SI Table 8.9 for current simulation).

The three properties evaluated are to some extent related and similar trends are observed overall. Comparing the means and standard deviations of all properties between solvent and complex simulation, we see an increase for all MC congeners except for [Amba5]-MC-LY(Prg), [Anda5]-MC-LY(Prg), MC-YR, MC-LY, MC-LF and MC-LR, were we have similar or slightly higher mean values. For most MC congener conjugates, the volume also increases for complex simulation, and mean values are mostly comparable for R and S CIP configurations. For some conjugates, congener and CIP configuration dependent effects are observable when comparing the mean values obtained in the solvent simulation with the complex simulation. These effects vary dependent on the respective congener. For example, GSH-R has a higher means compared to GSH-S for MC-LF, [Apha5]-MC-LF, MC-LY(Prg), [Anda5]-MC-LY(Prg). Cys-R has a higher mean compared to Cys-S for MC-RR, but vice versa for [Amba5]-MC-LY(Prg), MC-YR, and [Apha5]-MC-LF. Interestingly, MC-RR-GSH-S values are higher compared to MC-RR-GSH-R, but decrease in the complex simulation while GSH-R slightly increases, leading to similar values for complex simulation. For some conjugates, no differences or only small differences between solvent and complex simulation are observable, as e. g. MC-LF R conjugates, [Amba5]-MC-LY(Prg)-Cys-R, and Cys conjugates of MC-RR and [Apha5]-MC-LF, as well as GSH conjugates of MC-YR, MC-LY(Prg), [Amba5]-MC-LY(Prg) and [Apda5]-MC-LF and [Anda5]-MC-LY(Prg)-GSH-S. For other MC congeners conjugates, the properties (i.e., volume, radius of gyration, solvent accessible surface area) decrease, e. g. for MC-LY(Prg)-Cys-R, or Cys-S (e. g. for MC-YR, [Anda5]-MC-LY(Prg), and [Amba5]-MC-LY(Prg)), or GSH conjugates of [Enantio-Adda5]MC-LF, while it increases for MC-LY(Prg)-Cys-S.

Nevertheless, the changes are still within the standard deviations of the individual values. This is not entirely unexpected, since the structure of MC-LR was proposed not to change when in solvent compared to the binding site [212, 213, 320]. In our previous work we were able to show that there are several conformational clusters in the solvent, the number of which depends on the MC congener, so there are indeed some differences between the structures or conformations. Nevertheless, a majority of the conformations are still close to one structure observed in both simulations. Therefore, it is difficult to quantify differences with mean and standard deviation, and we could not observe any difference in the values for MC-LR compared to less or non-toxic MC congeners. Our previous simulations suggested that MC-LF and [β-D-Asp3,Dhb7]MC-RR increase their properties upon binding to PPP1, which was observed here for [β-D-Asp3,Dhb7]MC-RR. For MC-LF, however, the properties are now more constant, with only a slight increase in radius of gyration and solvent accessible surface area. Previously, it was assumed that MC-LF forms a more coil-like structure when simulated in solvent compared

to complex simulation in order to shield its more hydrophobic phenylalanine residue from the surrounding water compared to other MC congeners. It is a little more compact in solvent, as the radius of gyration indicates, but not much. In addition, [Enantio-Adda5]MC-LF also shows a different trend here compared to our previous simulation. The volume does not decrease but increases this time, while the radius of gyration still has a constant mean value, which is more similar to the other MC congeners than in previous simulations. When comparing the more hydrophilic MC congeners ([β-D-Asp3,Dhb7]MC-RR, MC-RR) with the more hydrophobic MC congeners without modified Adda side chain (e. g. MC-LF, MC-LY) we observe overall higher mean values for all three properties, which is in agreement with the results from previous simulations. [β-D-Asp3,Dhb7]MC-RR and MC-RR seem to adapt to a more open structure in solvent, as well as in complex simulation. In solvent, the positively charged arginine residues interact better with the surrounding water than hydrophobic amino acids do, therefore the structure is probably less coil-like in complex. Perhaps the charged residue reduces interaction with the hydrophobic groove of PPP1, leading to a more open conformation. Interestingly, the MC congeners [Anda5]-MC-LY(Prg), [Amba5]-MC-LY(Prg), [Apha5]-MC-LF and [Apda5]-MC-LF, which have a modified Adda side chain, have overall lower values for all properties compared to the corresponding MC congener without modified Adda side chain (MC-LY(Prg), MC-LF). Since [Anda5]-MC-LY(Prg), [Amba5]-MC-LY(Prg) and [Apha5]-MC-LF are smaller compared to Adda, the results are in line with the expectations. For [Apda5]-MC-LF, the results are surprising because APDA is similar in size to Adda, but seems to cause a more compact folding. The differences observed in the size, compactness, and accessibility of the MC congeners could be due to their structural conformations. Therefore, the structures and conformations will be investigated in more detail in Section 4.4.3.3 to explain some differences we can observe for the properties here.

For most of the MC congeners and conjugates, a similar trend to the previous study is observed for the all-atom RMSD (see SI Figure 8.15 and SI Figure 8.16). The all-atom RMSD for the complex simulations have a higher mean compared to the solvent simulation, although some values are still within their standard deviations (see SI Table 8.9). For MC-LR, MC-LF and MC-YR the RMSD decreases from solvent to complex simulation, indicating a stable starting structure in complex simulation. For MC-LR the crystal structure was used as the initial conformation, which already fits well into the binding site and therefore results in lower values. For MC-LF, MC-YR, and MC-LY and [Amba5]-MC-LY(Prg), where we have a similar RMSD, the docking pose may have fit well as the overall starting structure, resulting in less rearrangement during the simulation. For the conjugates, there is also an overall increase from solvent to complex simulation. Here, the RMSD values are on average higher than for the MC congeners. The structures are much larger than the MC congeners. Therefore, more rearrangement is possible and the docking poses are likely to be less reliable, as the larger molecules are more difficult to dock, leading to rearrangement in the MC congener conjugate in the simulation due to less favourable binding. For some conjugates we observe a different trend, where the RMSD is either constant or decreasing, leading in the case of MC-LR-GSH-R to much higher values than GSH-S when simulated in complex (also for other properties), or to lower values than the respective other CIP configuration for MC-RR-GSH-S, MC-LY-Cys-R, Anda-Cys-S and [Anda5]-MC-LY(Prg)-GSH-R). For MC-LY(Prg)-GSH-R, [Amba5]-MC-LY(Prg)-Cys-S and [Apda5]-MC-LF-Cys-S it results in an adjustment of the values similar to the other CIP configuration. For this reason, we conclude that for some MC congener conjugates there is an effect of the CIP configuration on the conformation and therefore on the properties of the respective conjugate, which will also be further investigated in Section 4.4.3.3.

4.4.3.3 Conformational Space Analysis

PCA on PPP1 Backbone The PCA projection on 2D for the PPP1 backbone is similar between MC congener and conjugates. In contrast to apo, the movements are smaller and therefore the data points are less distributed than apo but are located in a similar conformational space (see e. g. SI Figure 8.17, rest of the data not shown). The normal mode vector analysis of the extreme movement of the backbone results in porcupine plots (see Figure 4.11a and Figure 4.11b). The porcupine plots visualise the movement of PPP1 backbone along the eigenvectors. Along eigenvector 1(see Figure 4.11a), mainly the lower part of PPP1 moves outwards, from the A' alpha-helix of the N-subdomain to the 1' beta sheet of the N-subdomain (see Figure 4.11c). Similarly, the loop at the top of the structure, which extends from the beta sheet 8 of the C-subdomain to beta sheet 9 of the C-subdomain, has some twisting movement, directed towards the binding site. Instead of vertical, the loop becomes horizontally oriented. The other loop in the H-alpha helix of the C subdomain at the top moves slightly backwards. All other simulations with MC congeners and conjugates have similar movement to apo along eigenvector 1, but less extensive, resulting in less movement and shorter arrows. Otherwise, little difference is noticeable or visible.

The movement along eigenvector 2 (see Figure 4.11b) is mostly located in the outer parts of PPP1 backbone and different loops. There is some significant movement at the bottom and top of the protein, similar to eigenvector 1, but in a different direction, with the loops folding more inwards. In addition, there is a lot of rearrangement in the middle and on the left-hand side of the protein, where the loop structures are also in motion. In the centre, the eigenvector 2 stretches forward in between beta sheet 2 and alpha-helix B of the N-subdomain. On the left, there is an outward movement between beta sheets 12 and 13 of the C subdomain.

Comparing the PPP1 projection of the complex simulation, the main movement in PPP1 along eigenvector 1 is similar for apo and all complex simulations, although it is slightly more restricted in the complex simulation. In contrast, movement along eigenvector 2 which considers more the smaller loops in between alpha helices and beta sheets of PPP1, there are some differences between MC congeners and conjugates in the extent of the movement, but still highly similar (data not shown). These results are consistent with the structural properties discussed in Section 4.4.3.1 and indicate a very stable structure of PPP1, with small changes occurring upon binding.

(a) Eigenvector 1 (b) Eigenvector 2 (c) PPP1 structural elements

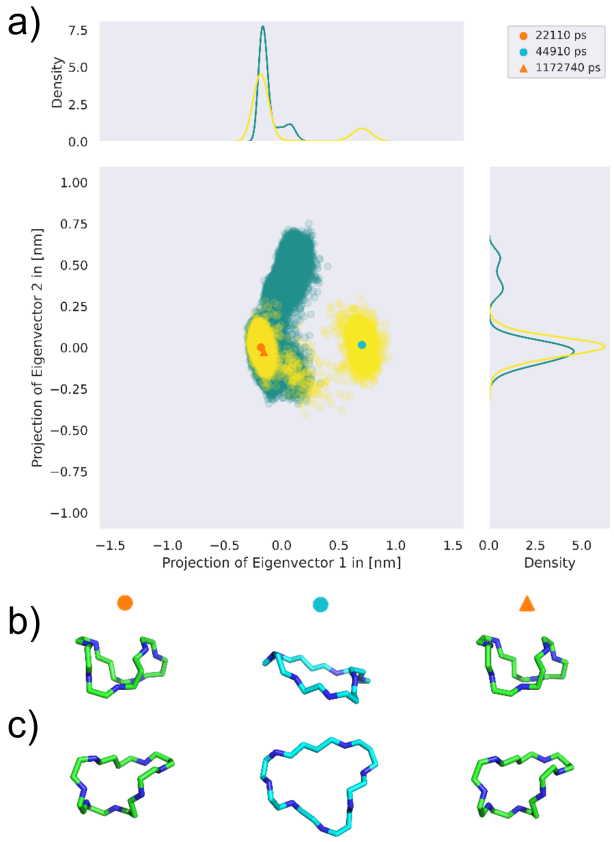
Figure 4.11: Porcupine plot of backbone movement of PPP1 compared to structural elements of PPP1. a) and b) The movement of PPP1 backbone of the apo simulation was projected by principal component analysis and the first two eigenvectors are shown. The arrows indicate the movement above 4 Å. In c) the structural elements where the movement occurs are shown. The A' alpha-helix of the N-subdomain is shown in pink, the beta sheet 1' of the N-subdomain is shown in cyan. The H-alpha-helix of the C-subdomain is shown in green, and the beta sheets 8 and 9 of sheet 3 in the C-subdomain are shown in dark blue and orange, respectively. The beta sheet 12 in sheet 2 of the C-subdomain is shown in yellow, beta sheet 13 of the C-subdomain in pale yellow. Beta sheet 2 of the N-Subdomain of sheet one is shown in marine, and the alpha-helix B in the N-subdomain in light blue. The image was generated with PyMol [371] and the notation of the structural elements is taken from Goldberg et al. [212].

PCA on MC Congener Backbone A PCA analysis was also performed on the MC congener backbone to analyse the different conformations that the MC backbone adopts during the simulation, which represent different energy states [360].

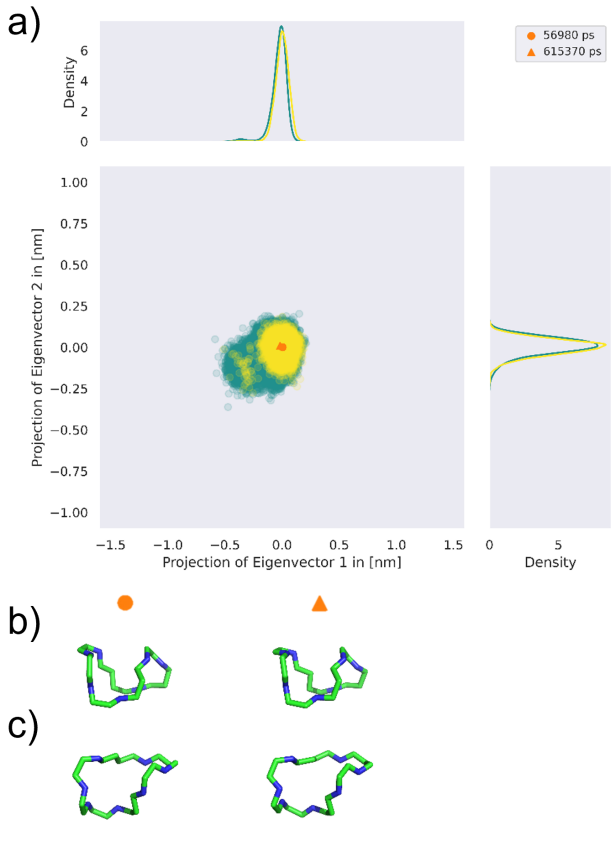
MC-LR was previously described as saddle-shaped conformation in water [212, 320, 324–326], and when binding to PPP1 [212, 213]. The saddle-shaped conformation was also found here for MC-LR in solvent and in complex simulation. Similar to our previous simulations (see Section 4.3.3.3) a second cluster for MC-LR conformation was also observed. In contrast to our previous simulations, the structures here overlap more between solvent and complex simulation. This suggests that at least the overlapping structures are shared between the different types of simulations, which is also evident when looking at the stick representation of the two overlapping conformations (orange circle and triangle, see Figure 4.12). In contrast, the second conformational cluster of the MC-LR backbone on the right (cyan circle) is a more planar structure compared to what we found previously. Interestingly, in the 1990s it was predicted that MC-LR would have a planar backbone [322, 323]. A study by Lavigne et al. [327] proposed that different structures of MC-LR may exist and bind to PPP1, and that the complex is flexible in solution. In our previous work (see Section 4.3.3.3) we were able to show that at least two conformations of MC-LR exist in solution, and these are also observed in these MD simulations.

For MC-LF, the overlap between the two clusters is even higher than for MC-LR (orange circle and triangle, see Figure 4.12). Only one conformation could be derived for each simulation, which is in both cases also saddle-shaped and almost identical to MC-LR backbone structure. Compared to previous simulations, the second conformational cluster disappeared

and merged into one. We initially assumed that the second conformational cluster was unstable as it was only very sparsely populated, which is supported by our data here.



(a) MC-LR

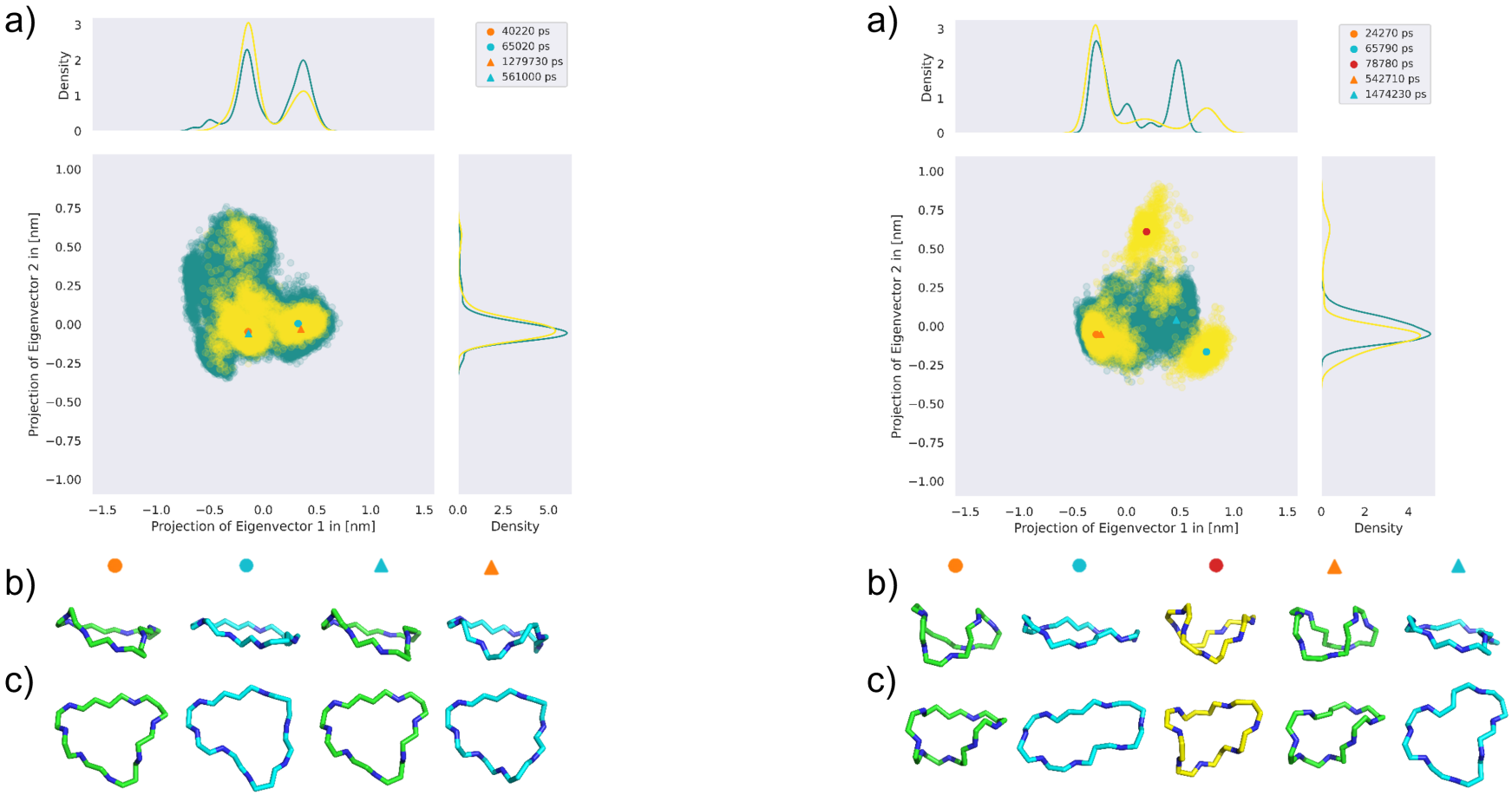


(b) MC-LF

Figure 4.12: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

[Enantio-Adda5]MC-LF (see Figure 4.13) has a wider spread distribution of the scatter plot compared to the previous MC congeners. Several clusters for solvent are visible here, and the complex structures are assigned to a similar position and therefore to the overall structure. In contrast to the previous simulations, the conformations of [Enantio-Adda5]MC-LF are planar. The ring is more or less open and quite stretched towards the outside. The structures overlap between the two simulation types and look different compared to the narrow and defined structure of MC-LF. MC-LF differs from [Enantio-Adda5]MC-LF only in the stereocentres of the Adda side chain, resulting in a non-toxic rather than a toxic congener. Thus, we conclude that side-chain stereocentres influence the whole backbone structure and not just the side-chain orientation, which is consistent with our previous results. In contrast to our previous results, the two projections of the different simulations overlap more, probably due to more sampling caused by the longer simulation time.

[β-D-Asp3,Dhb7]MC-RR has three scattered clusters for solvent simulation and one large cluster for complex simulation which is in the centre of the solvent clusters (see Figure 4.13). The saddle-shaped structure found in solvent and complex (orange circle and triangle) was also found for this MC congener. In addition, a more open, almost planar structure could be identified for both simulation types (cyan circle and triangle), but the shape is slightly different for both. The cyan triangle occupies a conformational space, which seems to be only accessible to the MC congener in PPP1 and not in the solvent simulation. Therefore, MC congener seems to undergo some conformational change upon binding to PPP1. The last conformation found in solvent looks like a more skewed saddle-shaped cluster and is differently shaped from the previous observed structures in the smaller simulation set.



[Enantio-Adda5]MC-LF

[β-D-Asp3,Dhb7]MC-RR

Figure 4.13: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

The conformations reported for the remaining MC congeners and conjugates have not yet been studied and only one MC congener has a reported structure derived from experiments in the literature. Therefore, the discussion is mostly limited to the other reported MC congener conformations.

Compared to [β-D-Asp3,Dhb7]MC-RR, MC-RR (see Figure 4.14), which is the same MC congener but without the modifications at position 3 and 7, the clusters are less scattered and more defined. Two clusters are visible in the solvent simulation, where the major cluster has a V-shaped backbone, and the minor cluster has a planar structure which is surrounded by the complex cluster. The major structure of the MC-RR backbone in complex simulation looks different to the solvent simulation and is shaped like an inverted U. MC-RR is also a toxic MC congener, but has a different backbone structure in the complex simulation, as previously observed for MC-LR and MC-LF. Compared to [β-D-Asp3,Dhb7]MC-RR the backbone looks completely different. Therefore, we conclude that the modification of side chains attached to the backbone (i.e., [β-D-Asp3,Dhb7]) seems to have a large influence on the backbone structure, even though only small groups attached to the backbone were modified.

MC-LY also has a well-defined and narrow distribution (see Figure 4.14), similar to MC-LF (see Figure 4.12), and is also a toxic MC congener. Solvent and complex clusters overlap almost completely and have one major structural representative that looks very similar for both of them. The structure looks mostly planar, but has a skewed part at both ends of the ring making the backbone look like a boat conformation, which is in agreement with an experimental study in 1994 which determined the structure of MC-LY in DMSO using NMR spectroscopy [324]. Interestingly, the backbone structure between MC-LF and MC-LY looks completely different (MC-LF is saddle-shaped), although only one amino acid (F for Y) has been replaced both of which are hydrophobic and have a very similar overall structure.

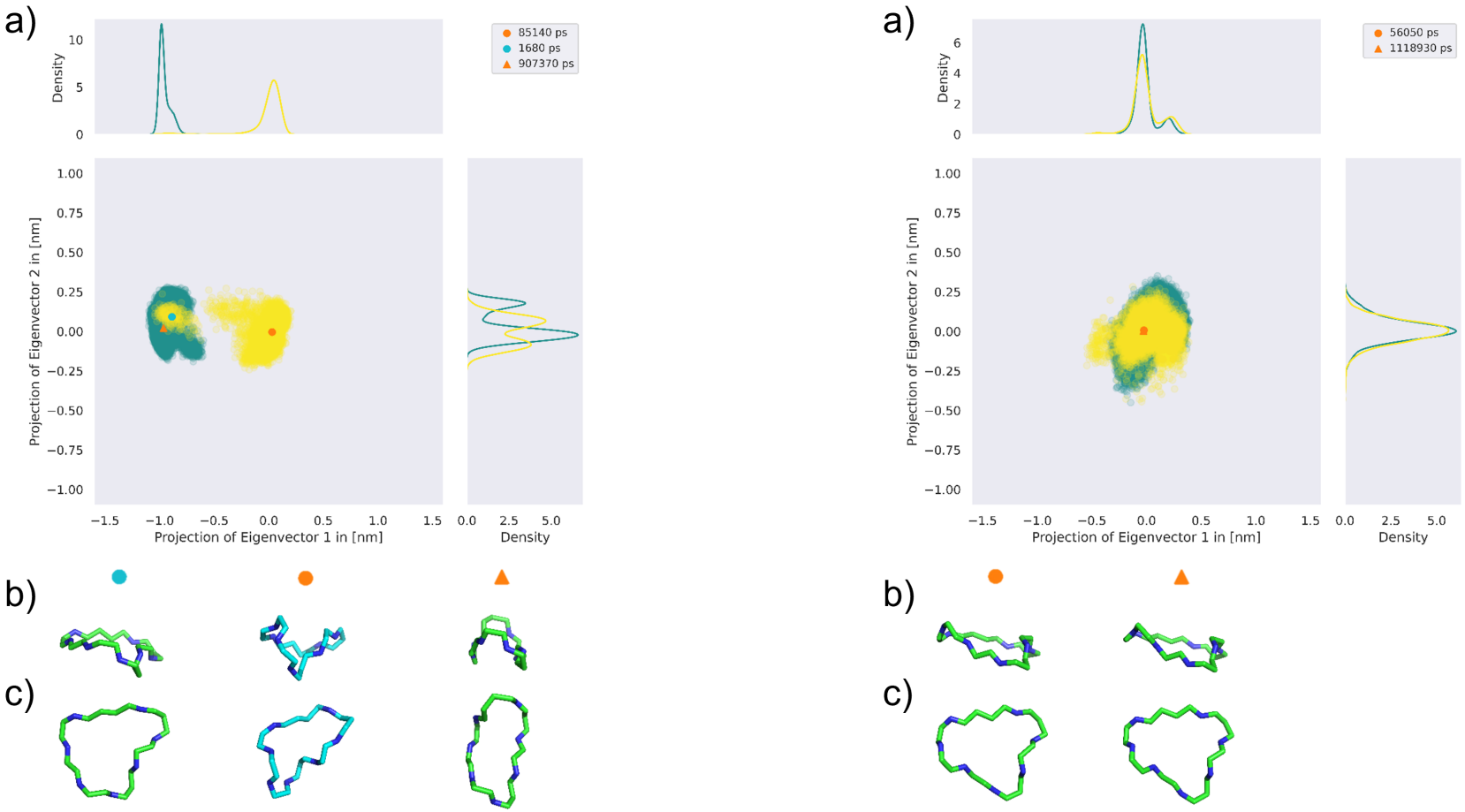


Figure 4.14: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

MC-YR (see Figure 4.15) is a toxic MC congener with two solvent clusters and a large, widely distributed complex cluster. A structure similar to MC-LY, which looks like a boat, is observed in the large solvent cluster (orange circle). The minority solvent cluster (cyan circle) has a V-shape, which was also observed for the majority solvent cluster of MC-RR (see orange circle in Figure 4.14). In contrast, the majority of structures for complex simulations are saddle-shaped, as for MC-LR. Therefore, it appears that a conformational shift occurs in the backbone of MC-YR upon binding to PPP1.

MC-LY(Prg) is a toxic MC congener and has one large solvent cluster which is visible with the complex cluster organised around the solvent cluster (see Figure 4.15). The solvent structure is bent to the right of the MC congener backbone and has an inverted U-shape (orange circle). The complex simulation at this position (cyan triangle) also shows the same structure. The cluster of complex structures on the right side has an upward bent U-like structure (orange triangle). The small modification compared to MC-LY (see Figure 4.14) results in different conformations to which the backbone adapts.

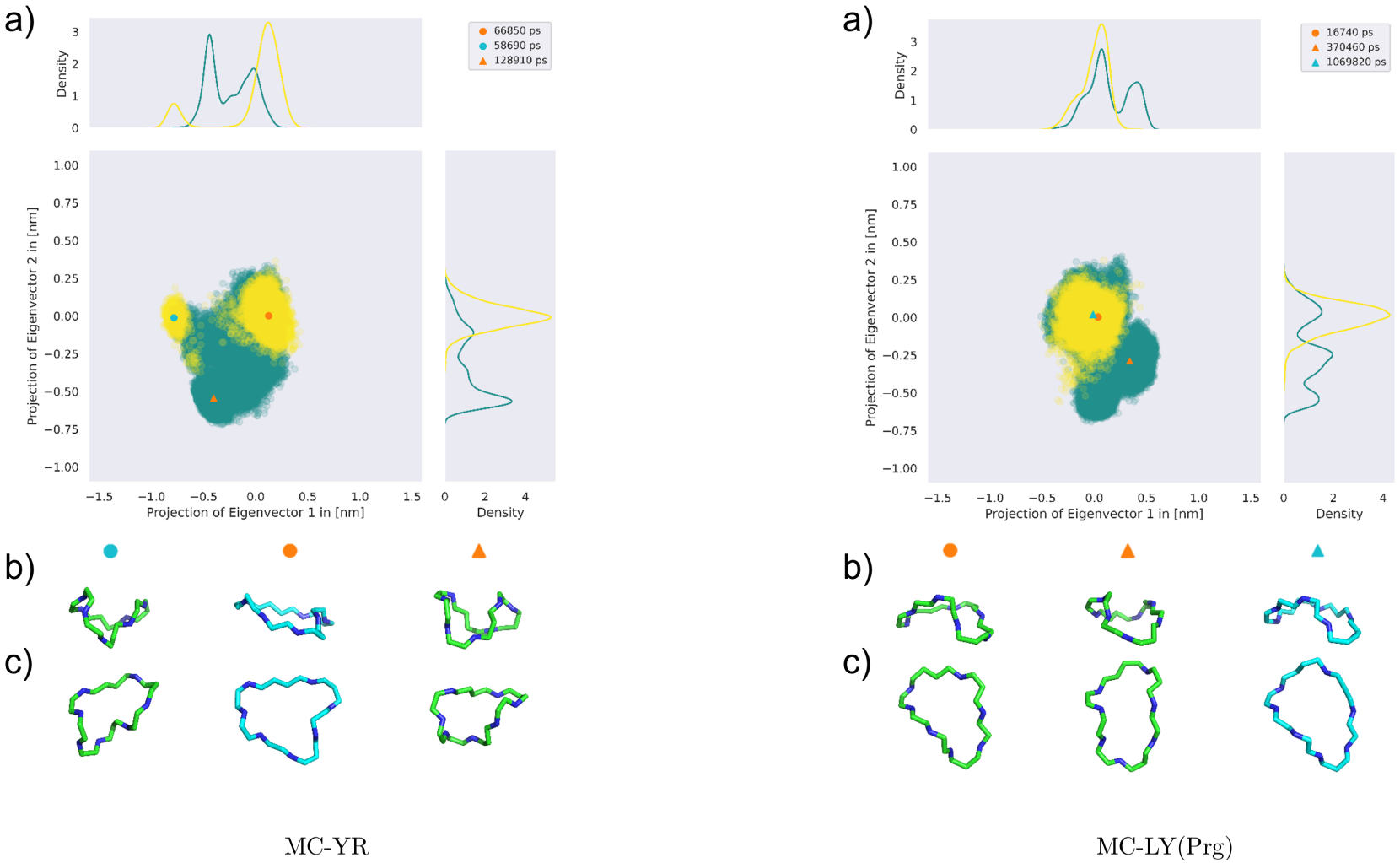
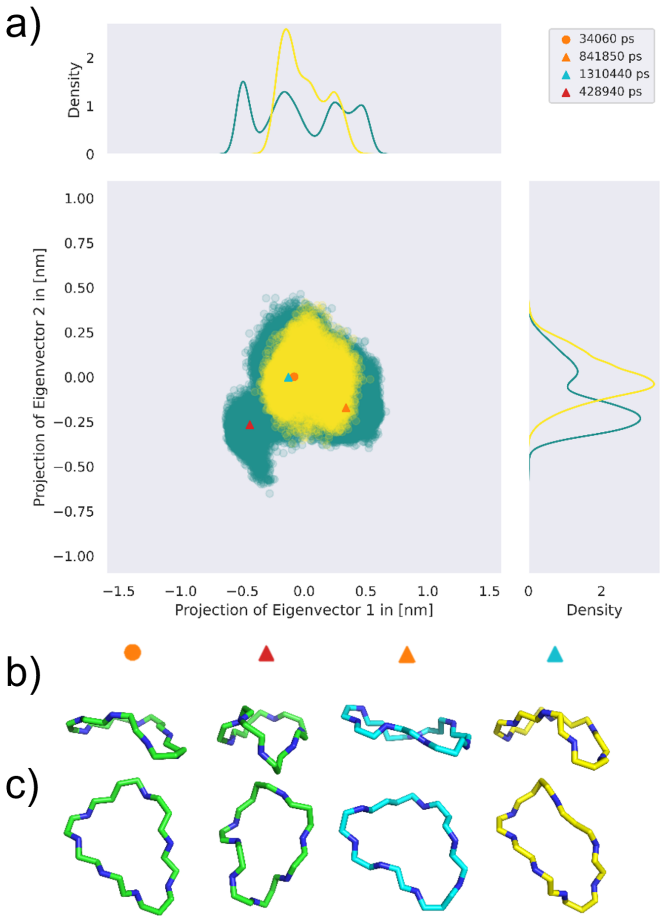


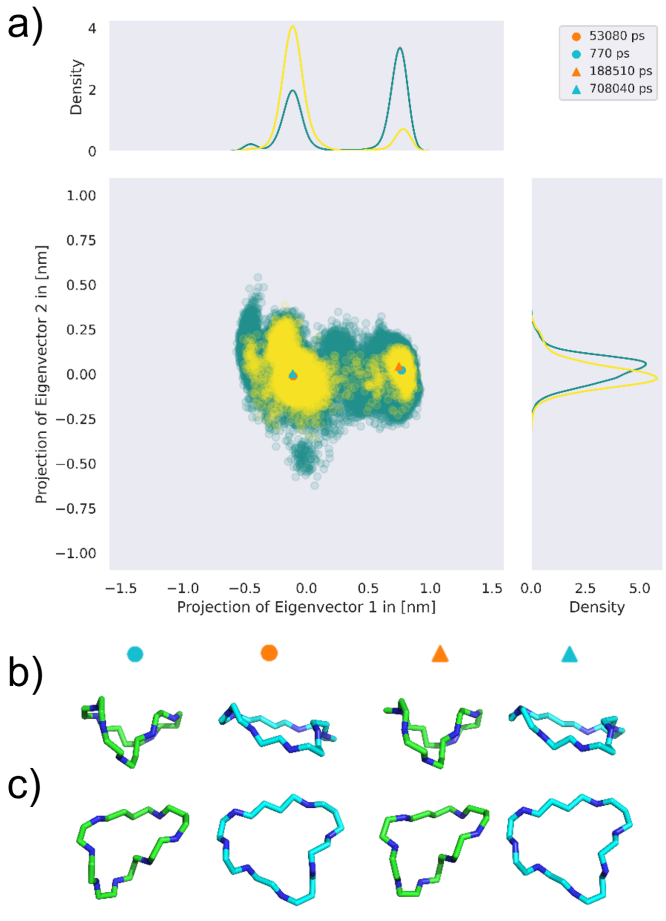
Figure 4.15: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

[Anda5]-MC-LY(Prg) (see Figure 4.16) is a non-toxic MC congener and modified at the Adda side chain compared to MC-LY(Prg). The distribution and scatter plot of [Anda5]-MC-LY(Prg) is similar to that of MC-LY(Prg). A large solvent cluster is also visible here with a mostly planar, slightly bent downwards shape (orange circle), which is similar to the structure identified in the complex simulation (cyan triangle). Both of these structures are similar to the solvent and complex structure that has been identified for MC-LY(Prg) (orange circle and cyan triangle, see Figure 4.15). In addition, two more conformations could be identified which were not found for MC-LY(Prg). The cluster of complex structures on the right side has a mostly planar structure (orange triangle), while the structure on the left side is highly skewed (red triangle). Therefore, we conclude that the modification of the Adda side chain probably affects the backbone conformation, as we have identified some conformations that are unique to either [Anda5]-MC-LY(Prg) or MC-LY(Prg), but they also have some conformations in common that occur during the simulation. Nevertheless, [Anda5]-MC-LY(Prg) is a non-toxic MC congener compared to the toxic MC-LY(Prg) so the influence of the Adda side chain seems to be more important than the exact backbone conformation, as [Anda5]-MC-LY(Prg) is closer to the saddle-shape of the toxic MC-LR.

[Amba5]-MC-LY(Prg) is a non-toxic MC congener which is also modified at the Adda residue (see Figure 4.16). Two major conformational clusters are visible for the solvent and complex simulations, which overlap considerably, and the representatives lie directly on top of each other. The part of the scatter plot which is more populated in the solvent is less populated in the complex simulation and vice versa. The more populated structure is an almost planar ring (orange circle and cyan triangle), which is only slightly curved at the left and right end of the ring structure. This structure is similar to that observed in the solvent simulation of MC-YR (orange circle, major cluster, see Figure 4.15) and MC-LR (blue circle, minor cluster, see Figure 4.12), as well as solvent and complex simulation of MC-LY (orange circle and triangle, see Figure 4.14). The second structure observed (cyan circle and orange triangle), which is more populated in complex simulation is V-shaped and similar to the structure observed for MC-YR solvent simulation (cyan circle, minor cluster, see Figure 4.15), and MC-RR solvent simulation (orange circle, major cluster, see Figure 4.14).



[Anda5]-MC-LY(Prg)

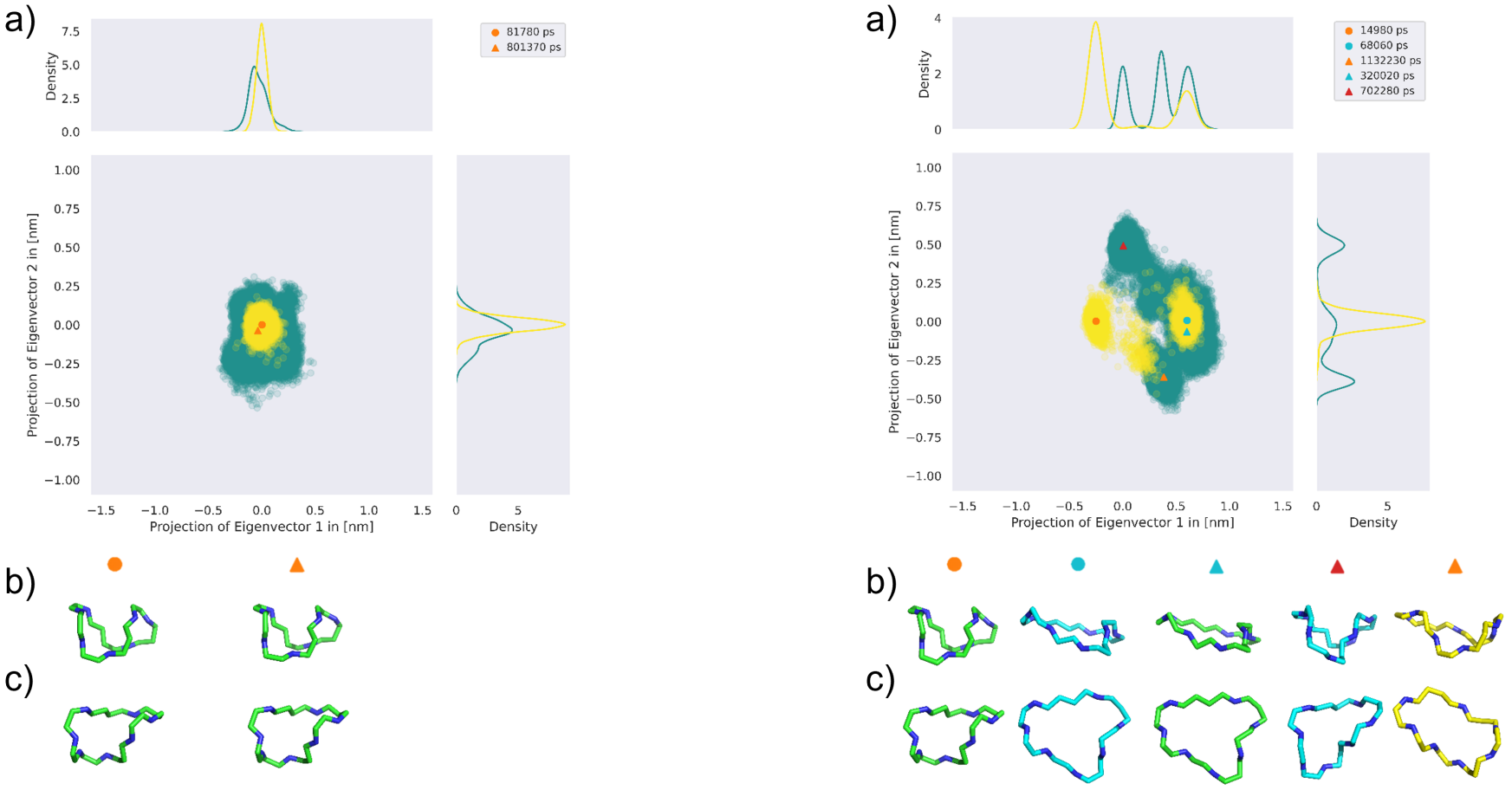


[Amba5]-MC-LY(Prg)

Figure 4.16: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

[Apha5]-MC-LF is a non-toxic MC congener and the scatter plot and structure (see Figure 4.17) are very similar to the toxic MC congener MC-LF (see Figure 4.12) and also has a saddle-shaped structure for solvent and complex simulation. For the solvent simulation, only a very narrow cluster is observed, around which the cluster of the complex simulation is located. The observed backbone structure for [Apha5]-MC-LF is similar to that of the toxic congeners (i.e., MC-LF), suggesting that the influence of the Adda side chain is more important than the exact backbone conformation.

[Apda5]-MC-LF is a non-toxic MC congener with two major conformations for solvent simulation and three for complex simulation (see Figure 4.17). The major solvent conformation is saddle-shaped (orange circle) as for other MC congeners and does not overlap with structures of the complex projection. The second cluster of solvent projections overlaps with the complex projection, and both have a similar planar structure that is slightly skewed at the left and right end of the structure (cyan circles and triangles). This structure is similar to that observed in solvent simulations of MC-YR (orange circle, major cluster, see Figure 4.15) and MC-LR (blue circle, minor cluster, see Figure 4.12) as well as solvent and complex simulations of MC-LY (orange circle and triangle, see Figure 4.14) and [Amba5]-MC-LY(Prg) (orange circle and cyan triangle, see Figure 4.16).



[Apha5]-MC-LF

[Apda5]-MC-LF

Figure 4.17: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

In conclusion, the saddle-shaped structure of MC congeners can be observed for several MC congeners, but not for all of them. Modification of the Adda side chain can lead to a change in the backbone structure of MC congeners, i.e., [Enantio-Adda5]MC-LF, [Anda5]-MC-LY(Prg) and [Amba5]-MC-LY(Prg) that are non-toxic congeners. However, it does not necessarily change the conformation, as e. g. non-toxic congeners [Apha5]-MC-LF, and [Apda5]-MC-LF can adapt to conformations which are also observed for toxic MC congeners. Therefore, in these cases, it is not the conformation of the backbone that is changed, but probably the interaction with the side chains to such an extent that it does not bind as well. Overall, the conformational differences between all MC congeners are limited and partly overlap between solvent and complex simulation, even when the conformations are not saddle-shaped. Therefore, we propose that MC congeners do not have to undergo a major conformational change upon binding, but can do so, and that there are conformations other than the saddle-shaped structure reported so far. An NMR solution study by Bagu et al. [326] found that MC-LR and MC-LL have a similar saddle-shaped backbone and structure with only minor differences. Subsequent docking of MC-LL showed that the conformation of MC-LL at the binding site is similar to that of MC-LR. Based on these findings, Bagu et al. [326] assumed that since the conformation and binding pose of MC-LL is so similar to MC-LR, that all other MC-congeners will have a similar structure in solution and conformation bound to PPP1. We were able to show that some MC congeners adapt to a saddle-shaped structure, but also that different MC congeners actually adapt to different backbone structures, depending on their modification and the amino acids at the hypervariable positions 2 and 4.

The structures of MC congener conjugates have not been reported in literature and have not been simulated. We found that they adapt to a wider range of conformations compared to unconjugated MC congeners (see SI Figure 8.18 to SI Figure 8.39). Many conjugates also adopt a saddle conformation, such as MC-LF-GSH-S, MC-LF-GSH-R, MC-LR-Cys-R or MC-LR-Cys-S. We also see an effect of the CIP configuration on the cluster distribution. For some conjugates, Cys-S and Cys-R or GSH-S and GSH-R have very different distributions which are widely scattered or narrow, and some have many representative structures, i.e., maxima in the free energy distribution, (e. g. [Apha5]-MC-LF-Cys-S or [Apda5]-MC-LF-GSH-R) while others have only two representative conformations (e. g. [Apha5]-MC-LF-Cys-R or [Apda5]-MC-LF-GSH-S). Planar structures can also be identified (e. g. [Amba5]-MC-LY(Prg)-Cys-R), which may be partially tilted at the edges (e. g. MC-YR-GSH-R), boat, chair or half-chair conformations, e. g. MC-RR-GSH-R, MC-YR-GSH-R or [Anda5]-MC-LY(Prg)-Cys-S, respectively. In addition, there are also V-shaped conformations (e. g. MC-RR-GSH-R) or inverted U-form conformations which can be more open (e. g. MC-LY-GSH-R) or closed (e. g. MC-YR-Cys-R) in the centre of the ring structure. In summary, conjugates can adopt an even greater variety of backbone conformations than MC congeners. Some of these conformations have already been observed for the MC congeners, others are specific to conjugates. As there are no data in literature to compare with, more data and investigations are needed.

4.4.3.4 Binding of MC Congeners to PPP1 - Qualitative Analysis

Due to the large amount of data generated in this data set, a proper visual analysis of the trajectory to evaluate binding of MC congeners and conjugates was not feasible. Therefore, the trajectory was quickly inspected by highly summarising the trajectory to 150 frames. Sometimes, water molecules that have been coordinated in the binding site move out of the binding site. This is considered uncritical, as the MC congeners should replace the water molecules. In addition, active site water can be pulled out when a MC congener diffuses out of the binding

pocket.

In contrast to our previous simulation set, toxic congeners (e. g. MC-LR, MC-LY(Prg)) also move out of the binding pocket for replicates 2 and 3 (see Figure 4.18). This effect could be caused by some interactions that developed unfavourably for certain MC congeners, or by inappropriate starting structures, as this is mostly the case for replicates 2 and 3, although this was not observed for some non-toxic MC congeners and should be investigated in more detail. Of course, this has an effect on the conformations and properties calculated in this section to analyse the data set. Therefore, the results should be viewed with caution and more simulations with better starting structures for replicates are necessary to evaluate this issue in more detail. Also, for MC congener conjugates (see SI Figure 8.40 and SI Figure 8.41), many move out of the binding sites. For some of them this can already be observed for replicate 1, for most of them in replicate 2 or 3. This could also be caused by inappropriate starting structures which were received by docking. The conjugated MC congener is even larger, making it more difficult to dock, because macrocycles have complex 3D structures [14, 38, 39] and scoring functions are very sensitive to changes in 3D conformation [3]. In addition, it is unknown how MC congener conjugates interact with PPP1, because there are, to the best of our knowledge, no experimental studies on MC congener conjugate interaction with PPP1. Therefore, as for MC congeners, the results should be viewed critically and more simulations with better starting structures for replicates are necessary to evaluate this issue in more detail.

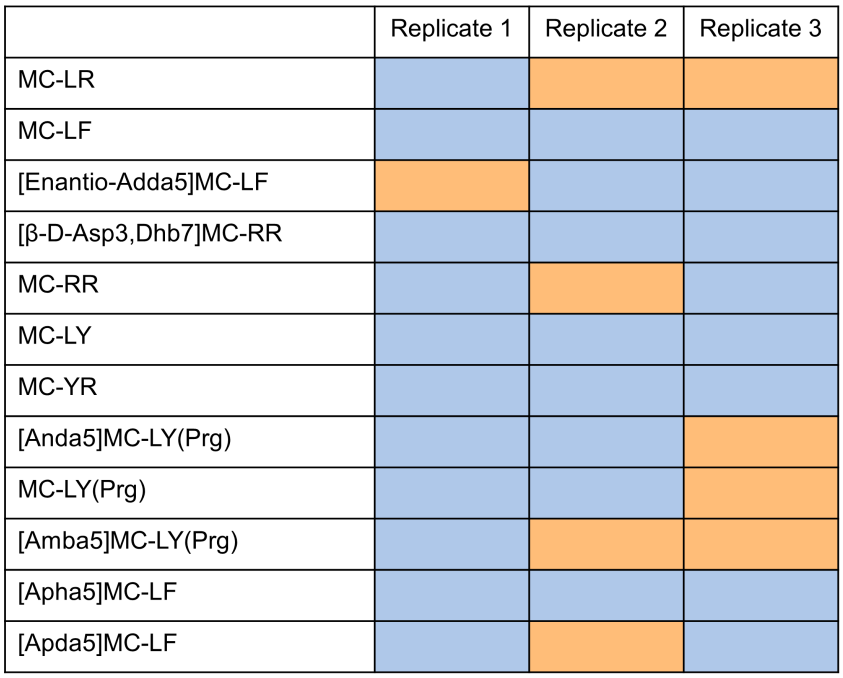


Figure 4.18: Summary of qualitative evaluation of binding stability of MC congeners in PPP1. For each MC congener, the individual replicates are shown. Blue colour indicates stable binding, orange colour indicates diffusion out of the binding site.

4.5 Conclusion

In conclusion, we presented for the first time MD simulations not only for MC-LR, but also for 3 other MC congeners, which were later extended in a second simulation data set of a total of

12 MC congeners with selected conjugates. We were able to show that MD simulation studies help to understand MC-congener specific backbone conformation and dynamics, and the simulations presented here are a step towards a congener-specific investigation of structures that have been neglected so far.

For the conformational analysis of the MC congener backbones, we were able to demonstrate that different conformational clusters and conformations exist for the individual MC congeners, which might partly explain the tight binding to PPP1. For MC-LR, we showed two conformational clusters in solvent, which were confirmed in our second simulation data set. In contrast to the first data set, we were able to identify a planar structure in solvent for MC-LR, which was originally predicted in the 1990s but never experimentally confirmed [322, 323].

In addition, we have reported backbone conformations for 11 MC congeners in complex and solvent. Many of these can also adapt to the saddle-shaped structure identified for MC-LR, but also to other conformations that can be described as planar, inverted U-shape or V-shape. In both data sets, we showed that the less and non-toxic MC congeners ([β-D-Asp3,Dhb7]-MC-RR and [Enantio-Adda5]-MC-LF) have a higher overlap and distribution of clusters in contrast to the toxic MC congeners, indicating a less stable backbone structure. For our second data set, we were able to show that some non-toxic MC congeners share backbone conformations with toxic congeners, supporting the assumption that the Adda side chain is indeed an important factor in binding, and does only partly influence the backbone structure. Therefore, we conclude that the backbone conformation is unlikely to be the only reason for the tight binding. However, for [Enantio-Adda5]-MC-LF, a completely different backbone was observed to its stereoisomer MC-LF, suggesting that modification of Adda can, but does not necessarily have to, alter the backbone conformation. The MC congener conjugates adapt to an even greater variety of structures, and the orientation of the conjugation in R or S for some conjugates strongly influences the conformations and the conformational landscape.

The partly observed differences between the data sets probably result from different methodology and overall longer simulation time. Nevertheless, the results of the second study should be interpreted with caution, as for many replicates the MC congeners or conjugates move out of the binding pocket, which may be caused by inappropriate starting structures for replicates 2 and 3 and therefore, the effects of this need to be investigated in more detail.

The next step is to analyse the binding and interaction of MC congeners and conjugates, which will be evaluated and studied in the next chapter, as it is not feasible to analyse such a large data set manually.

Chapter 5

Interaction Fingerprints for Molecular Dynamics Simulation

5.1 Summary

This chapter describes the development of an approach to automatically derive so-called interaction fingerprints (IFP) from Molecular Dynamics (MD) simulations and how to systematically aggregate, analyse and visualise them. IFPs encode interactions between molecular structures, which are important for molecular recognition and therefore for properties such as binding or toxicity. For this reason, the analysis of interactions is often a crucial step in MD simulation analysis.

The data sets of MD simulations collected in Chapter 4 on Microcystin (MC) congeners and conjugates interacting with PPP1 are used here for the development of IFP analysis, interpretation and visualisation. The analysis of IFPs will give an overview of how MC congeners interact with PPP1 and therefore aid in understanding why some MC congeners are strong inhibitors of ser/thr protein phosphatase (PPP1) and others do not inhibit PPP1. Even though the methods developed here are studied in the context of MC toxicity, they are applicable to a wide range of molecular systems, because its input is an MD simulation of two molecules interacting with each other.

This work has been published in the Journal of Cheminformatics. Ideas, figures and tables have been taken from the publication and text modules may be similar to a certain extent:

- **Sabrina Jaeger-Honz**, Karsten Klein and Falk Schreiber: 'Systematic analysis, aggregation and visualisation of interaction fingerprints for molecular dynamics simulation data', Journal of Cheminformatics, 16 (28): 15 pages, 2024. doi:10.1186/s13321-024-00822-3.

According to the FAIR principles (Findability, Accessibility, Interoperability, and Reuse of digital assets, see Section 1.4) all data sets and scripts which are part of this manuscript have been published open source in the Zenodo repository:

1. https://doi.org/10.5281/zenodo.10423389 [51]

2. https://doi.org/10.5281/zenodo.10424417 [52]

For a statement of contribution, the reader is referred to Section 1.3.

5.2 Background

Fingerprints are easy and fast to compute, analyse and compare. Therefore, they are a popular concept and widely used in cheminformatics [41, 98, 385–387]. The most commonly used fingerprints are molecular fingerprints, which often encode 2D chemical structures as a 1D bit vector [41, 385] (see Section 2.1.2.1). There are many different similarity metrics, the most common being the Tanimoto Coefficient (Tc), or distance measures that help to analyse and compare molecular structures quickly [41, 385]. The concept of molecular similarity is based on the structural similarity principle, which states that structurally similar molecules tend to have the same physical properties and biological activities, and that small changes have only a small effect on the biological activity [41, 98, 385–387]. Although this principle is true for many molecules in different data sets, there are also molecules, known as activity cliffs, for which this principle is not true. For these molecules, small changes in structure lead to large changes in biological activity [98, 388]. Therefore, the threshold at which a molecule is considered similar is dependent on the molecular representation and data set studied, because the concept of similarity of molecular structure is not well-defined [41, 98, 385, 386, 388]. Ligand-centric methods were developed early to predict or model bioactivity. They are based on the structure similarity principle, which has been successfully applied to many applications, although it is difficult to link molecular similarity to biological activity, because molecular similarity often involves chemical intuition and subjective criteria such as filter values. In addition, they ignore interactions that are crucial for biological activity [386]. To overcome this limitation of 2D similarity concepts, 3D similarity methods have been proposed for target screening, but they have limited predictive power, and they are often ligand-centric as well, since they lack target information (e. g. key residues), or are more difficult to compute and analyse [41, 389].

Modelling the 3D conformation and interactions of molecular structures is crucial for understanding the molecular recognition process, which is important for understanding biological activity and processes (see Section 2.1.3) [390, 391]. Different methods and tools have been developed to study interactions computationally, such as molecular docking and MD simulations (see Section 2.1.3). Interaction fingerprints (IFP) have been developed to analyse and visualise the interactions obtained by those methods. IFPs are a combination of the fields of molecular modelling and cheminformatics, where fingerprints are frequently applied.

IFPs are based on the idea that interactions derived from 3D coordinates from molecular modelling techniques or experimental studies and are transformed into a 1D bit vector, which is a typical representation in the field of cheminformatics. The interactions can either be sampled by methods from molecular modelling, such as docking, MD simulation, or obtained from experiments [108, 392]. Most commonly, they describe interactions between proteins and ligands, but also protein-protein interaction, or interactions with nucleic acid can be encoded [391]. Many different methods have been developed to derive IFPs of protein-ligand interactions, and most have been used as post-processing methods for virtual screening approaches (i.e., large-scale docking approaches) and conformational space analysis [389, 392–401], but have also been used for machine learning approaches [390, 402–404], and, more recently, to study interactions derived from MD simulations [391, 405].

An early example, which was developed to rescore docking poses in virtual screening approaches to identify active molecules, is SIFt (structural interaction fingerprint). SIFt [393] was one of the first approaches and encoded 3D interactions between a ligand and a residue in seven different types of interactions. This approach results in an explicit 7 bit representation of interactions as a 1D bit vector for each residue, which can be reduced if computation needs to

be accelerated. The derived SIFt vectors can be used for visualisation (e. g. hierarchical clustering), analysis and organisation and can be applied to different types of systems [393]. Many extensions of this method have been proposed, as e. g. weighted SIFTs [406], profile SIFTs [407], as well as other methods to derive IFPs [394, 396]. Mpamhanga et al. [395] proposed an interaction fingerprint that also takes into account knowledge-based scores, which are derived from comparing the similarity of a bit string of a docking solution to a reference binding mode. With this approach, the method aims to identify compounds that have a similar mode of binding. Using the bit string representation of the binding mode or pairwise interactions, the fingerprints can be scored and clustered to explore relationships between binding modes. To derive a score for reranking docking poses, the Tc and Euclidean distance, as well as simple counts, were combined with the scores obtained from docking to generate a modified score [395].

Marcou and Rognan [396] introduced a molecular interaction fingerprint which was also a modification of SIFt, that became an important tool for rescoring of docking poses and to identifying new potential fragments in virtual screening approaches. It considers more interaction types than SIFt and also increased quality of prediction in comparison to former methods [408]. The scoring of IFPs was compared to a reference, which seemed to work better than conventional scoring. Encoding the 3D information as IFPs allows quick and efficient comparison by calculating the Tc, which works better than conventional RMSD investigation. In addition, decoys (inactive molecules with similar properties to active molecules) in the data set were efficiently identified, and truly active scaffolds were reranked to the top of the list. The authors determined a threshold of $Tc = 0.6$ to distinguish good from bad docking solutions on the studied benchmark data set. It was also possible to cluster IFPs extracted from known PDB structures according to their binding patterns by generating a phylogenetic tree using neighbour-joining clustering [396].

Other IFP approaches encode interactions implicitly, such as the Structural Protein-Ligand Interaction Fingerprints (SPLIF). SPLIF encodes fragments of the ligand and protein which are interacting, instead of the interaction itself. They evaluate the similarity between different SPLIFs by calculating a normalised quantitative score, which was then compared to the SPLIF of a reference protein-ligand complex [399].

Several other approaches have been proposed, as for example PLIP [400], Arpeggio [401], which are web services to determine interactions from structures [400, 401], PyPLIF Hippos, which is a Python-based programming tool [408] and many others. To visualise the interaction with 3D structure, many use PyMOL sessions [400, 401] or Tc to estimate similarity of IFPs [408]. The Tc is often visualised as a matrix, where colour indicates the similarity value between IFPs.

Recently, the concept of IFPs has also been applied to MD simulations [391, 405] (see Section 2.1.3.2), which is more challenging compared to virtual screening approaches, because of the additional temporal dimension. MD simulations are powerful tools to study temporal motion and therefore dynamics (i.e., conformation, interaction, etc.) of a system over time (so-called trajectories) [65]. MD simulations result in long trajectories of individual atoms and their motion at specific time points, leading to a massive amount of time-dependent data. The analysis of these trajectories is difficult and time-consuming, as mostly well-established measures are calculated and analysed, such as root-mean-square deviation, root-mean-square fluctuation (RMSF), radius of gyration and energy-based approaches [66, 133]. These properties are usually studied to identify interesting points in the trajectory where e. g. changes occur, and these frames or time points are then investigated directly by visual inspection of the

3D conformations. Data analysis is currently the bottleneck, as it is difficult to analyse the massive amount of data, which are increasing in size and length as computing power increases, making a frame-by-frame analysis and observation impossible [160, 409, 410]. Especially if multiple simulations are compared and differences between simulations are highlighted, as i.e., for interactions, analysis is difficult [411].

To aid in this process, methods, and tools have been developed to e. g. investigate the interaction between a protein and a ligand. However, most of these methods are based either on visual inspection (e. g. VMD [412]) and visualisation of contact maps [413] (e. g. as a contact frequency map (MDContactCom [414]), or as dynamic matrix (CONAN [415])), or a list of interaction partners or distances (e. g. GROMACS [342] or MDAnalysis [416]). These methods all come with their disadvantages, as it is difficult to see and perceive differences in multiple matrix visualisations. A list is complicated to analyse and lacks 3D representation, and a trajectory usually consists of many frames or time steps. For these reasons, there is a strong need to develop methods to improve the analysis of these dynamic interactions.

One of the first approaches of IFPs for MD simulation considering the interaction between a ligand and a protein was a workflow described by Kokh et al. [405]. The underlying idea of this approach was to provide the user with a workflow to calculate and analyse MD-IFPs for large systems in a reasonable time for several hundred compounds. MD-IFP was developed to study the unbinding trajectory (i.e., dissociation routes), conformation and residence time for a series of compounds. To read and iterate over the MD simulation frames, MDAnalysis [416] was used and combined with RDKit [128] to identify and compute interactions which were mapped to a bit vector. The resulting MD-IFPs were mapped based on the ligand centre of mass on a 3D grid in a physical or IFP space, and clustered with k-means or Gaussian methods. This resulted in 8 clusters, where transitions between the clusters were visualised and used to study intermediate states (metastable structures) relevant to dissociation. In addition, the interactions of the MD-IFPs of the eight clusters within a certain threshold or with Euclidean distance were visualised as a matrix [405].

Bouysset and Fiorucci [391] created a Python library called ProLIF (Protein-Ligand Interaction Fingerprint) to systematically calculate IFPs from MD simulation data, experimental data or docking poses for a variety of different molecules. Many different interaction types are supported for analysis, and additional interactions can be added by the user or existing ones modified. While most programs are not compatible with MD simulation data, ProLIF overcomes this challenge by integrating MDAnalysis. MDAnalysis can infer bond order and charges which are often missing in MD topology and are necessary to convert molecular structures to RDKit, if hydrogens are represented explicitly. The interactions can then be analysed either per residue, or atomic level and after analysis the results are provided as data frame for further processing. The individual interactions can be visualised as a timeline. The library also provides functions to visualise and calculate a so-called aggregated frame. For the aggregated frame, all interactions in the IFP are summed up over time and interactions that occur more than 30 % of the time are part of the aggregated frame (see Figure 5.1). The aggregated frame, as well as a specific frame at a specific time point, can be interactively visualised at atom-level for the ligand, and residue level for the protein. Only one atom group of a ligand is highlighted as interacting, which is the one that most frequently interacts with the protein. ProLIF also provides functions to visualise a residue interaction network when interactions within a protein are explored, and a Tanimoto similarity matrix for each frame of the MD simulation for a protein-ligand MD trajectory to estimate how binding behaviour changes over time [391].

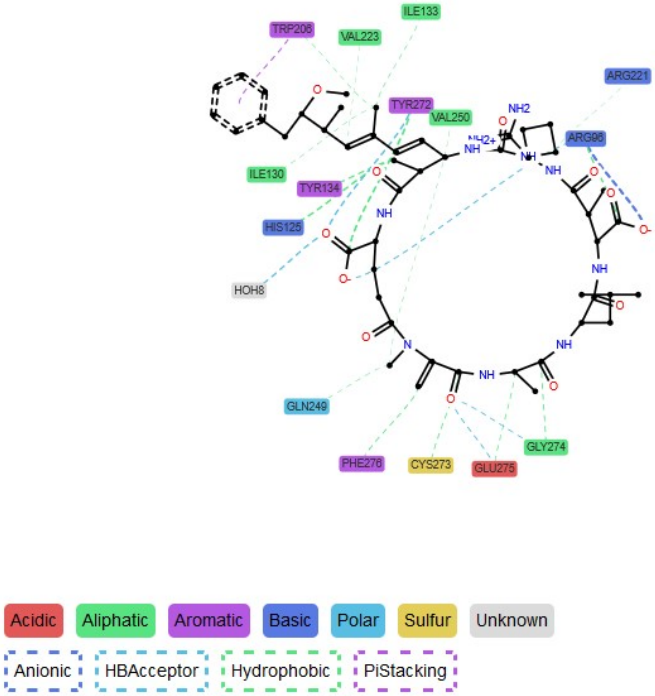


Figure 5.1: Visualisation of aggregated frame with interactions occurring more than 30 % of the time proposed in ProLIF [391] (own data, PPP1-MC-LR complex, see Section 4.3).

IFPs derived from MD simulation provide a valuable tool for systematically exploring and studying the interactions that occur in large simulations. After computation of the IFPs they are easy to handle and analyse since they are 1D bit vectors. The previously proposed methods have the disadvantage of either massively aggregating data by considering e. g. only interactions that occur more than 30 % of the time, which results in a representative IFP that does not exist in reality, or in loosing information by aggregating to clusters. Therefore, the aim of this chapter is to propose new methods for analysis and visualisation of IFPs derived from MD simulation data in order to systematically aggregate relevant information and thereby reduce the number of time frames. In addition, the methods developed here should facilitate the comparison of multiple simulations, which is still difficult so far.

5.3 Development of Automatic Calculations of Interaction Fingerprints

5.3.1 Introduction

The aim of the MD simulations derived from Chapter 4 was to analyse and understand differences in binding behaviour of MC congeners and conjugates to PPP1. Visual inspection of the trajectories was difficult on the first data set (see Section 4.3) and not possible for the second data set (see Section 4.4) because of the high numbers of simulations. Therefore, an automated method was needed to derive interactions automatically. In this section, a method for auto-

matically deriving interaction fingerprints was developed, which was supposed to be extended to a library. In the middle of the development, ProLIF [391] was published and released, which had the same idea based on the same Python libraries. For this reason, the development was stopped, as ProLIF was already more mature, is actively maintained and available open-source to avoid parallel development of similar libraries. Nevertheless, the method is described in this section because it was developed to the point that an evaluation of the data presented in the last chapter was possible.

5.3.2 System and Methods

The overall approach developed here to automatically derive interactions between ligands and proteins is summarised in Figure 5.2. The approach consists of two steps: the assignment, where the system is analysed for its molecules and atom types, and the interaction processing, where interactions are detected for the whole trajectory based on specific criteria. Criteria for interaction were described by De Freitas and Schapira who analysed structures deposited in the PDB and evaluated interactions in X-ray structures to determine thresholds [417]. According to their work, interactions can be defined as listed in Table 8.10.

The method presented here has been implemented in Python. Python was chosen as programming language because it is widely used in the community. In addition, specialised libraries exist which were used to parse and process MD simulation data (MDAnalysis [416]), and to process molecular structure (RDKit [128]). For fast and efficient calculations, libraries such as math [351] and NumPy [352] are used to optimise processing of this large amount of data.

In the following, the method developed is described as a two-step approach in closer detail (for an overview, see Figure 5.2).

Assignment Step In the first part of the method, the trajectories and structure files derived from MD simulation are initially processed and assigned to specific atom types and interactions. The assignment is the first step, to later calculate the interactions efficiently, and is shown as orange boxes in Figure 5.2. To keep the description more concise and easier to understand, MC congener will be referred to as ligand, and interacting molecules (i.e., PPP1 with H_2O and Mn^{2+} ions) as protein. The individual steps in the first part are as follows:

1. Load trajectory and structure file of MD simulation data (can be in variable data formats as implemented in MDAnalysis)

2. Assign bond types, order and charges to MD simulation data using MDAnalysis to obtain correct chemical information.

3. Select the ligand and protein atoms within a user-defined cut-off radius r (here: 5Å) to process only nearby atoms. As MD simulation potentially contain hundreds of thousands of atoms, this pre-selection is necessary to avoid long computation times. MDAnalysis offers a dynamic selection, so that all atoms within the cut-off value r at any time point in the trajectory are selected.

4. Assign aromaticity to ligand atoms. Where possible, aromatic atoms are assigned to ring structures with RDKit to obtain the correct atom types for interaction analysis. This step is not necessary for protein atoms, as this information is already assigned to the atom types of the amino acids and well-known.

5. Assign all atoms (ligand and protein) to predefined atom types which are involved in specific interactions (see Supplementary Information (SI) Table 8.10 [417]). The assignment is based on the MDAnalysis selection syntax, which has an interface to RDKit to process SMARTS correctly. SMARTS are used to assign correct atom types. The selection syntax is summarised in SI Table 8.11.

6. Assign predefined atom types to particular interactions that may occur. Based on the atom type, certain interactions can be established, which are summarised in SI Table 8.10 [417]. The atom types are assigned to an interaction as pairs (i.e., atom pairs), where one atom type is from the ligand and the other is from the protein. To evaluate all interactions, all possible atom pairs are generated.

Interaction Processing Step To process the interactions, a second step, the so-called interaction processing step, is necessary (see red boxes in Figure 5.2). Within the second step, three for-each loops are run to process all data. First, all frames obtained by the simulation are iteratively processed, then all interactions are processed, and last, each atom pair within the interactions are processed. This results in the following steps:

1. for each $frame \in frames$,

2. for each $interaction \in interactions$,

3. for each $atompair \in atompairs$,

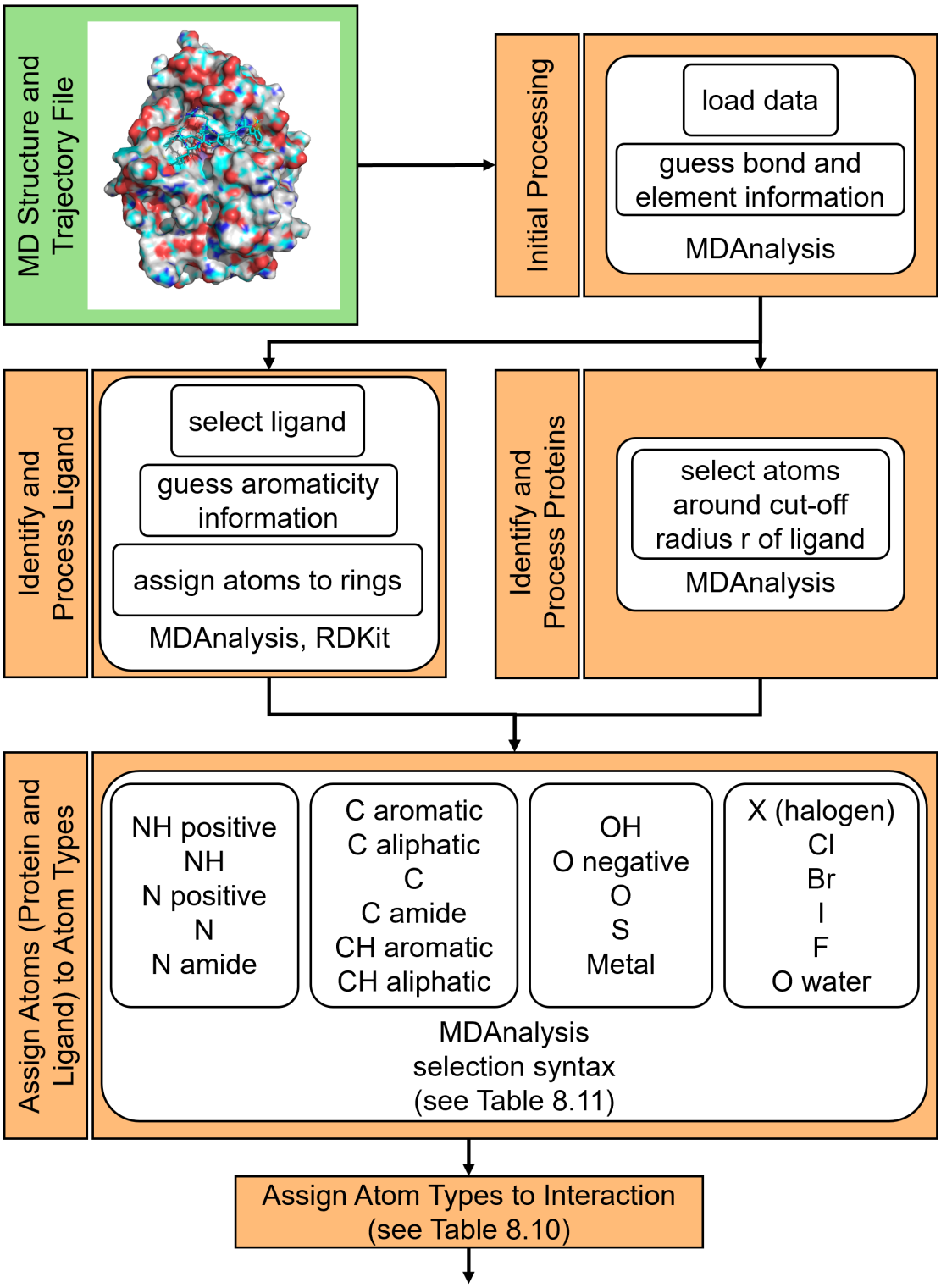


Figure 5.2: Flow chart of method to derive IFPs from MD simulation data. The green boxes describe the input, orange boxes the assignment steps, red boxes the interaction processing steps of the pipeline, and blue boxes the output. Continued on next page.

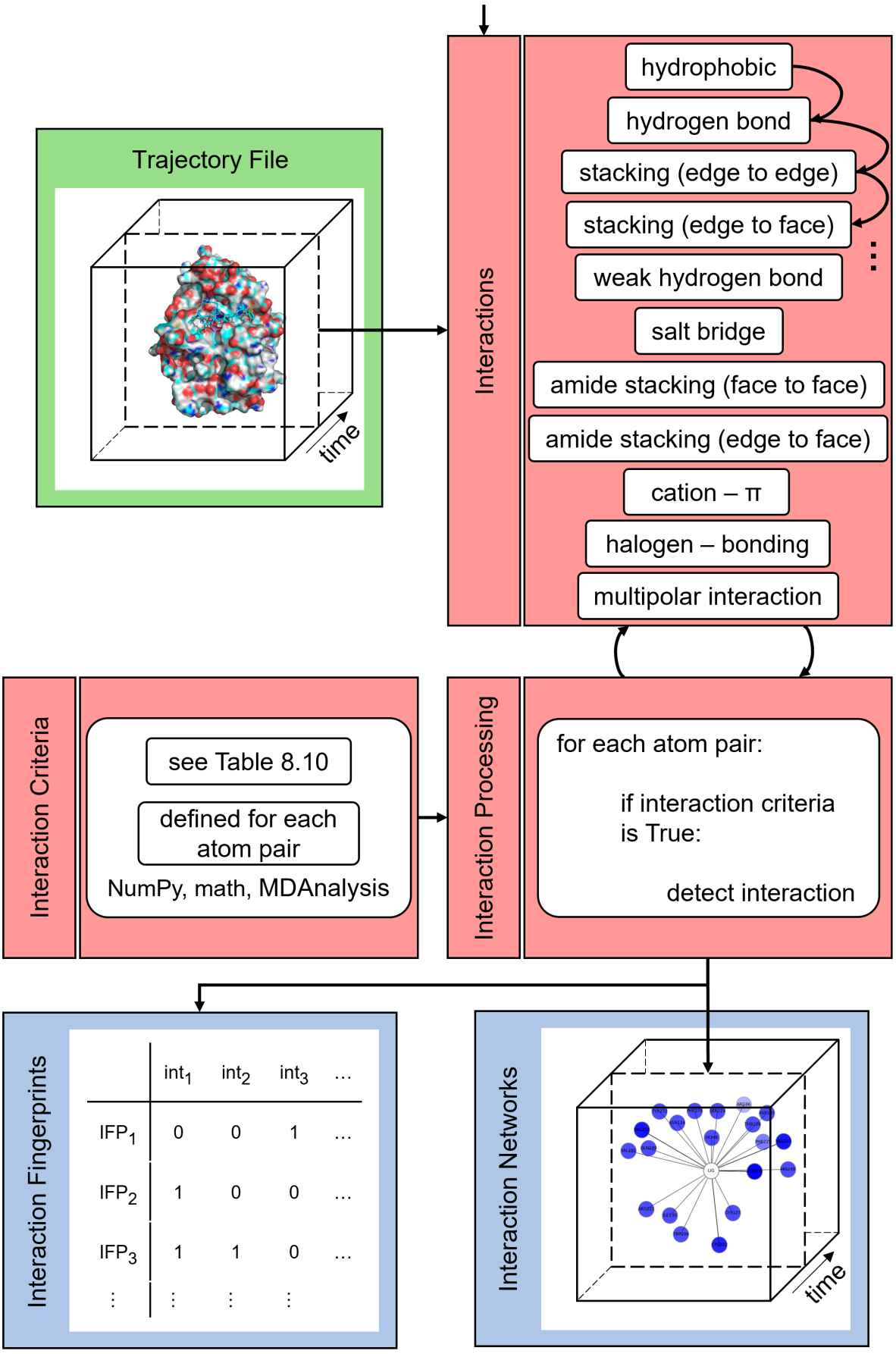


Figure 5.2: Continued from previous page.

4. If the interaction criteria are *True* for the respective atom pair and interaction (see SI Table 8.10 [417]), which can be computed efficiently with NumPy and math, the interaction is detected.

After processing and detecting all interactions, two different output formats are generated (see blue boxes, Figure 5.2):

- IFP matrix of bit vectors, where the rows correspond to individual IFPs (or frames), and the columns to interactions. The column names are assigned to the protein residue name and number, and type of interaction detected, as the ligand is treated as a single entity. If an interaction is present, the bit is turned on (1), if an interaction is absent, the bit is turned off (0).
- Interaction network, which is generated for each IFP (or frame) with NetworkX. The interaction networks are star graphs, with the ligand as the central node, and the interacting residues arranged around the ligand.

5.3.3 Conclusion

An automatic method to determine IFPs from MD simulation data in Python was described here. The derived IFPs can be used to investigate and visualise interactions of large data sets to efficiently study them. As ProLIF [391], which is based on the same ideas and libraries as the method presented here, was published and released, the development of a library was stopped to avoid parallel development of software. ProLIF is a library which is available open-source, actively maintained and ready to use for different biomolecular systems. Therefore, the focus of this chapter shifted to the analysis, visualisation, and comparison of IFPs, which is described in Section 5.4. These approaches are based on the ProLIF Library to provide a comprehensive set of tools to develop.

5.4 Development of Methods for Aggregation and Comparison of Interaction Fingerprints

5.4.1 Introduction

In this section, a new method to systematically aggregate IFPs from MD simulations is described. In addition, the aggregated IFPs are systematically visualised and processed. This allows the analyst to compare IFPs within and in between MD simulations to study differences and similarities in interactions or, in the use case of MC congeners studied here, to understand binding to PPP1.

5.4.2 Methods for Interaction Fingerprint Calculation

The generation of the data set analysed and preliminary results are described in Section 4.3 and Section 4.4. Both data sets are based on different types of simulation: 1) solvent simulation, where MC congener is solvated in water, 2) apo simulation, where PPP1 is simulated in water, and 3) MC congeners simulated in the binding site of PPP1. For the analysis and study of interactions, only the latter one is relevant. In contrast to Chapter 4, the simulations will be referred to by MC congener name instead of PPP1-MC congener name for the ease of reading.

To calculate interaction fingerprints from MD simulation, MDAnalysis(v2.4.0) [416, 418] was used to read trajectories and select atoms. Manganese ions (Mn^{2+}) do not have a van der Waals radius assigned in MDAnalysis. Therefore, it has been assigned manually to the topology table by using the van der Waals radius of magnesium ions (Mg^{2+}). Mg^{2+} were selected because the parameters of Mn^{2+} in MD simulations (see Chapter 4) have also been replaced to Magnesium parameters to result in more stable binding in the first data set, and to obtain parameters in the second data set. Mn^{2+} parameters can be exchanged with Mg^{2+} parameters, as they are similar in size and have similar coordination preferences to Mg^{2+} [334]. The parameters were also exchanged for the MD simulation itself, to have a stable coordination of Mn^{2+} in the binding site of PPP1. To calculate the IFPs, ProLIF (v1.0.0) [391] was used with RDKit (v2021.03.5) [128]. Many different interactions are available for calculation in ProLIF: Anionic, CationPi, Cationic, EdgeToFace, FaceToFace, HBAcceptor, HBDonor, Hydrophobic, Interaction, MetalAcceptor, MetalDonor, PiCation, PiStacking, XBAcceptor, XBDonor and VdWContact. X denotes halogen atoms. VdWContact was removed from the IFP calculation between MC congeners and PPP1, water and Mn^{2+}, because test runs showed that all interactions were changed to van der Waals contact rather than more specific ones. Mn^{2+} were treated separately to include the VdWContact with metal ions, as this interaction might be more unspecific. The IFPs were calculated for all frames in our MD simulation data, and replicates were treated as a single entity.

5.4.3 System and Methods for Aggregation, Visualisation and Comparison of IFPs

In the course of this thesis, an implementation to calculate interactions between ligands and proteins from MD simulation data was developed.

The system and methods developed here to aggregate, visualise and compare IFPs of MD simulation data are summarised in Figure 5.3 and Figure 5.4. This approach consists of two steps:

- the pre-processing, where IFPs are modified to summarise relevant interactions, and the aggregation step, where IFPs are summarised based on structural or temporal component (see orange boxes, Figure 5.3),

- and the visualisation and comparison of IFPs where similarity within IFPs of the same simulation (see orange box, Figure 5.4) and in between simulations (see red box, Figure 5.4) are evaluated and visually assessed.

The pre-processing and visualisation method presented in the following, has been also implemented as a Python library, because of previously mentioned advantages, and is called NetworkMD. The following specialised libraries were used to pre-process, calculate and visualise IFPs: MDAnalysis (v2.4.0) [416, 418], RDKit (v2021.03.5) [128], ProLIF (v1.0.0 [391]), NumPy (v1.21.2) [352], tqdm (v4.62.3) [419], pandas (v1.3.3) [420], scikit-learn (v1.0) [264], Matplotlib (v3.4.3) [354], imageio (v2.28.0) [421], Networkx (v2.6.3) [422] and DyNetx (v0.3.1) [423].

In the following, the individual steps of pre-processing and visualisation, as well as comparison of IFPs are shortly summarised and explained.

5.4.3.1 Pre-Processing of Interaction Fingerprints

An overview of the newly developed method to pre-process IFPs of MD simulation data is shown in Figure 5.3. The individual steps of the pre-processing step are (see orange boxes, Figure 5.3):

1. Load data frame of IFPs which were pre-calculated with ProLIF (see Section 5.4.2 and green box in Figure 5.3).

2. Restructure the data frame to resolve the multi-index, map the Boolean values (True/False) to a bit vector (1/0) for presence or absence of interactions

3. Sliding window processing. For each column (i.e., interaction) a sliding window calculation is performed with pandas. x_1 is a percentage value and used to calculate the size of the sliding window based on the number of all IFPs (i.e., trajectory length). The currently calculated data point is centred in the window, so that interactions close in time (before and after) are considered symmetrically. The new value of the current data point is the mean value across the window, which gets assigned to the position of the current data point.

4. Filtering of the calculated mean. The calculated mean values are then filtered based on x_2, which is the percentage of occurrences of an interaction within the window. If the mean value is greater than x_2, the interaction is set to 1 and therefore present; if it is below the threshold, the interaction is set to 0 and therefore absent.

5. Aggregation of pre-processed IFPs. The pre-processed and filtered IFPs are then aggregated with two different approaches: 1) structure-based, which means that any identical IFP in the data frame, regardless of where they are located in time, and 2) time-based, which means that IFPs that are identical and occur immediately after each other are summarised. When frames are summarised, an additional column is added to the data frame to store the number of occurrence of summarised frames (see green box, Figure 5.3).

The proposed method of x_1 and x_2 filtering smooths the data of the retrieved IFPs. The x_1 filter evaluates the occurrence of interactions within a time window. Filtering with x_2 considers only interactions that occurred more often than a certain threshold within the window defined by x_1. This procedure is applied because numerical simulation data only has a certain accuracy and errors in computation can occur. Also, some interactions may occur really rarely, even within the small time window, and can therefore probably be neglected because they are either artefacts or not relevant. Both filters can be adjusted by the user, as the x_1 and x_2 filter are dependent on the MD simulation and probably also on the data set studied. Different thresholds were studied here, to investigate the effect of filtering on the data set. For x_1, the different percentages evaluated were: 0.5 %, 1 %, 1.5 %, 2 %, 2.5 %, 5 %, 7.5 % and 10 %. For x_2 the filtering of the calculated mean values was evaluated for: 0.00, 0.01, 0.02, 0.025, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40. The mean values are between 0 and 1, so the values correspond to the percentage of occurrence of an interaction within the window. The x_2 filter will be referred to as percentage values for the remainder of this chapter, to improve readability. The values chosen for x_1 and x_2 have been chosen to cover a wide range of different values, to evaluate appropriate values for this data set because the filter values probably depend on the data set and simulation settings. Both filter values have been limited at the upper end, as the smoothing of the data is already high at these percentage values and is unlikely to be meaningful. The steps

between the filter values are smaller at the beginning and larger towards the end, as changes to smaller values are considered more sensitive than to larger ones. Structure-based and time-based aggregation have different advantages and disadvantages. Filtering by structure results in unique IFPs collected throughout the simulation and the lowest possible number without further aggregation, but the information about the temporal evolution of IFPs is lost. Filtering by time preserves the temporal component but may result in duplicates in the data set of IFPs, when they are revisited at a later stage and result in a higher number of IFPs. As the individual filters and aggregation methods depend on the data set, the MD simulation and how the IFPs were derived, the scripts provided for the workflow and NetworkMD are modular so that the user can easily change applied thresholds and aggregation type.

5.4.3.2 Similarity Calculation

To compare the aggregated IFPs with each other, the number of differences was calculated to estimate similarity of IFPs within the simulation. The number of differences is defined as shown in Equation 5.1.

$$N_{Diff} = \sum_{i=0}^{n-1} |IFP1_i - IFP2_i| \tag{5.1}$$

To compare the IFPs between MD simulations with each other, the Rogers-Tanimoto dissimilarity metric was computed. The distances were computed with scikit-learn (v1.0) [264] and SciPy distance functions (v1.7.1) [424], as these are optimised for efficient calculation on large amounts of data. The Rogers-Tanimoto dissimilarity is defined in Equation 5.2. c_{ij} is the number of occurrences in two 1-D vectors at position i and j, c_{TT} is the number of bits set on (1, interaction present) in both vectors, c_{FF} is the number of bits set off (0, interaction absent) in both vectors, and c_{TF} and c_{FT} is the number of bits set on in the first vector and off in the second vector and vice versa. In our case, the dissimilarity is between 0 and 1. Dissimilar IFP have a value close to 1, while similar IFP have values close to 0. The similarity of two IFPs is defined by Equation 5.3. To evaluate similarity for molecular fingerprints or also IFP fingerprints so far, most researchers consider the Tanimoto coefficient. A study by Racz at al. [425] showed that there are other coefficients that achieve consistent results on very diverse benchmark data sets and are viable alternatives to the Tanimoto coefficient. Rogers-Tanimoto is one of them and therefore selected in this work due to the possibility of fast computation. The assignment of a function to calculate similarity or dissimilarity is also modular, so the user can exchange the distance metrics to any other offered by the Python libraries scikit-learn [264] or SciPy [424] for pairwise distance calculation.

$$Dissimilarity_{Rogers-Tanimoto} = \frac{2 \times (c_{TF} + c_{FT})}{c_{TT} + c_{FF} + 2 \times (c_{TF} + c_{FT})} \tag{5.2}$$

$$Similarity = 1 - Dissimilarity_{Rogers-Tanimoto} \tag{5.3}$$

5.4.3.3 Visualisation and Comparison of Interaction Fingerprints

To support analysis and comparison of IFPs, a number of different visualisations were proposed. An individual visualisation can not cover all aspects of IFPs, which are relevant for analysis, therefore multiple visualisations are proposed and are partly provided as summarised

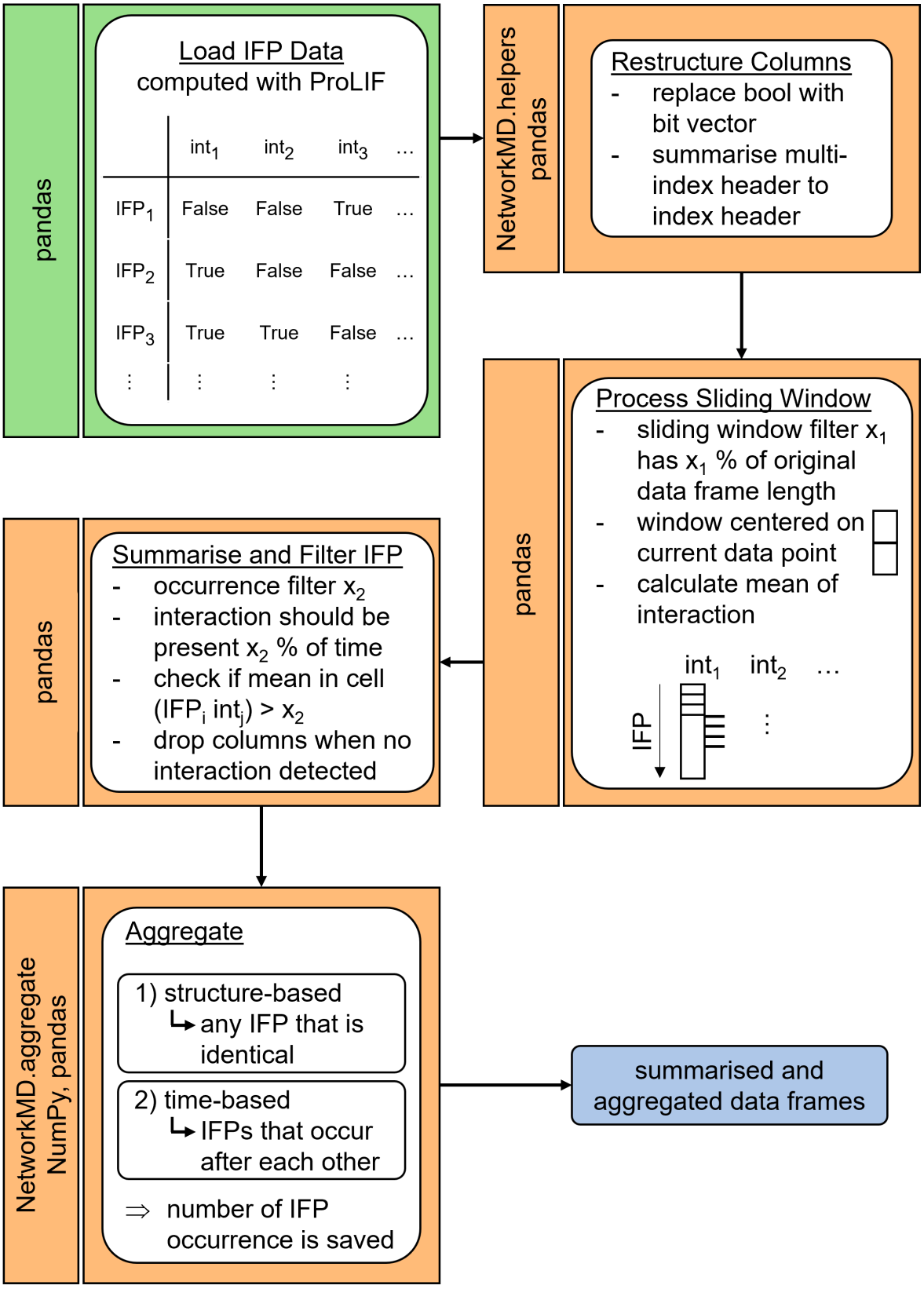


Figure 5.3: Flow chart of methods developed to pre-process and aggregate IFP data frame derived from MD simulation.

visualisations. The proposed combination of visualisations shows different aspect about the IFPs, which together help to analyse and understand the IFPs derived from MD simulation. For an overview of the different visualisations developed, see Figure 5.4. The visualisations are described individually, but are partly provided as combined images to the user. The visualisation process is a two-step approach. First, IFPs within the same simulation are compared, then, IFPs between two different simulations are visualised and compared. For both approaches, the summarised and aggregated IFPs are used as input (see green box, Figure 5.4). To compare interactions and IFPs within the same simulation, circular charts, line plots, histograms, a similarity matrix and a network visualisation were developed (see orange box, Figure 5.4).

The circular chart shows the interaction for each residue individually, but summarises the different types of interactions. This visualisation gives an impression of the length of the interaction, as the full circle would be the whole simulation. In addition, it is easier to compare, which interaction appear and disappear over time, or are constantly present/absent over longer periods of time.

The histograms show the number of differences within all IFPs present in an MD simulation. To give the user an impression of how the number of differences is distributed, and how dissimilar IFPs are on average from each other.

The matrix visualisation also shows the number of differences to all other IFPs, which is colour coded with the viridis colour map, which is perceptually uniform and robust to colour-blindness. Here, we can visualise how the differences develop over the course of the MD simulation and therefore keep the temporal or individual frame component.

The line plot shows the number of occurrence of an individual aggregated frame, i.e., how many IFPs have been aggregated structure-based or time-based to the respective IFP. The second line plot visualisation shows how many IFPs are identical within an MD simulation, by symbolising the frames as two horizontal lines, which are connected by vertical lines if the frames are identical, i.e., their number of differences is zero.

The network visualisation shows the star graphs of the interactions. The ligand is shown in the centre, and the residues are arranged around them. If a structure is provided by the user, the position of the residue nodes is oriented at approximately the same position provided in the structure file, to keep 3D information of the structure. In addition, a line plot of the number of occurrences of each frame is provided at the top, so that the user can directly estimate how often an individual network occurs in the data set. The different types of interactions are shown with different visualisation (see Table 5.1). Not all interactions are relevant in this work, but as a library for general use was developed, all interactions provided by ProLIF were encoded.

To compare interactions and IFPs between two different simulations, the aggregated data frames need to be further processed, as some interactions may be unique to either of the two simulations. In this case, the two IFP data sets are merged by adding all columns which are present in each simulation. If an interaction was not previously present in an IFP data set, the columns are filled with 0 (interaction absent), as it could potentially interact with the residue but was not detected (see red box, Figure 5.4).

To quantify and compare differences between IFPs of two simulations, similarity or dissimilarity of IFPs has to be evaluated. Since comparison between different simulations lead to higher differences between the IFPs, counting the number of differences was not considered appropriate any more. Therefore, the Rogers-Tanimoto dissimilarity is calculated and converted into similarity as described in Paragraph 5.4.3.2. To assess similarity, thresholds are chosen to

Table 5.1: Explanation of glyphs and colours used to visualise the interaction as a network.

Interaction	Glyph	Colour
Hydrophobic	circle	blue
HBAcceptor	square	blue
HBDonor	square	red
Anionic	arrow down	blue
Cationic	arrow down	red
CationPi	arrow left	red
PiCation	arrow left	blue
PiStacking	arrow up	blue
EdgeToFace	arrow right	red
FaceToFace	arrow right	blue
MetalAcceptor	tick up	red
MetalDonor	tick up	blue
XBAcceptor	tick down	red
XBDonor	tick down	blue
VdWContact	circle	red

categorise IFPs in three different classes: identical, similar and dissimilar. Similarity is a fuzzy concept, and different thresholds for similarity measures at which a molecule is considered similar have been proposed. Some authors suggest, e. g. that molecules with a Tanimoto coefficient (TC) of $Tc > 0.85$ are considered as structurally similar [386, 426], others consider an $Tc > 0.5$ as similar, and $Tc \leq 0.5$ as dissimilar [387]. The exact threshold is also data set dependent, and thresholds for IFPs have not been evaluated. Therefore, we decided to consider a similarity ≥ 0.975 as identical, and between 0.85 and 0.975 as similar, and have been evaluated by visual inspection of the simulation data. The thresholds for the different similarity classes can be adjusted by the user.

To visualise and compare the IFPs between different data sets, a line plot was constructed to evaluate similarity between and within simulations. Each simulation is represented as three lines: one simulation as dark blue lines (a, b, c) and the other simulation as bright blue lines (d, e, f). Between a and b, and e and f the identical IFPs within the same MD simulation are shown, between b and c, and d and e the similar IFPs within the same MD simulation are shown. To encode identical or similar IFPs between simulations (i.e., between c and d), a colour coding has been developed, where identical IFPs are shown in cyan, and similar IFPs in red.

All visualisations can be exported as image files, and the network visualisation can be additionally export as GIF to have a movie of IFP development over time.

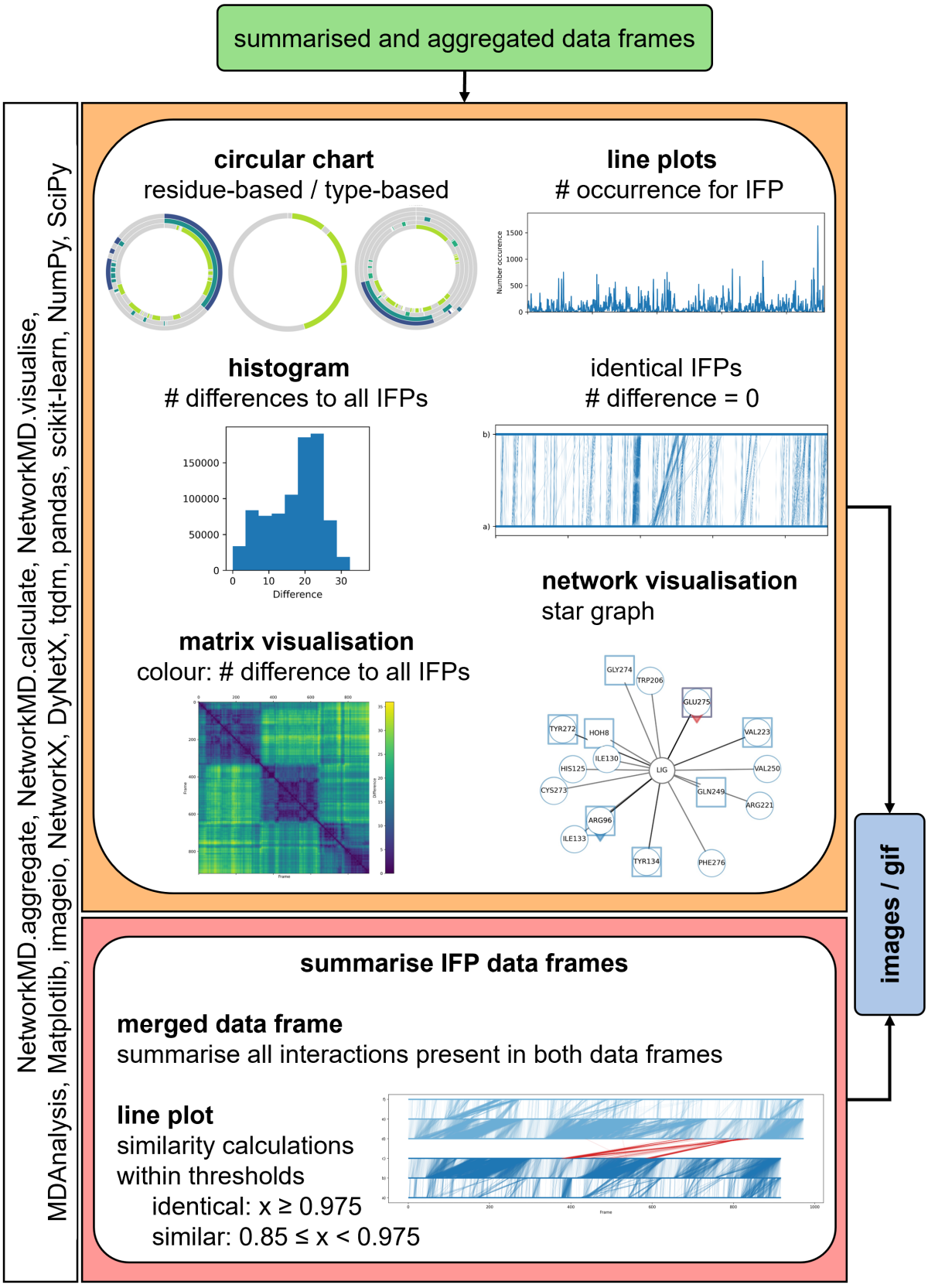


Figure 5.4: Flow chart of visualisations developed to compare IFP data within and in between MD simulation.

5.4.4 Results and Discussion

In the following, the analyses of influence of filtering and aggregation on IFPs will be discussed. To improve readability of the results and discussion section, a simplified naming scheme is introduced: As only complex simulation can be considered for interaction analysis, the simulations are referred to by MC congener name instead of PPP1-MC congener, as in the previous chapter. The aggregated frame derived based on the ProLIF paper (with occurrence more than 30 %) is referred to as aggregated$_{occ30}$ IFP to distinguish from temporal (time) and structural (structure) aggregation.

5.4.4.1 Influence of Aggregation and Filtering of Interaction Fingerprints

To investigate the influence of pre-processing (see Section 5.4.3.1), two experiments were performed and the number of interactions and IFPs were calculated for the small data set of four MC congeners (see Section 4.3). First, the influence of aggregation on IFPs based on structure, time or to one aggregated$_{occ30}$ IFP was investigated, and second, the influence of different filters on the IFPs were investigated. To filter the IFPs, two different filtering options are available:

- x_1 filter, which determines the size of a sliding window running over the individual interactions and calculates the mean of interactions, and
- x_2 filter, which only considers interactions if their mean is higher than a threshold.

Both values are percentage values, i.e., they are calculated based on the length of the data set, and the time an interaction is occurring, respectively. The IFP aggregation is then based either on structure, where all identical IFPs are aggregated, regardless of when they occur in the data set, or based on time, where all IFPs that are identical and occur immediately after each other are aggregated.

For the first experiment, we were able to show that, regardless of aggregation, the number of interactions differ between the MC congeners (see Table 5.2). The number of interactions is identical for both aggregation methods, as the IFPs are aggregated (i.e., the rows of the data frame) and not filtered. The lowest number of interactions detected are 55 for MC-LF and the highest number of interactions are 108 for [β-D-Asp3,Dhb7]MC-RR. These results indicate, that the interactions of MC congeners with PPP1 are different, which is what we also found in Section 4.3.3, by calculating secondary properties from the MD simulation. In comparison, if the aggregated$_{occ30}$ frame is calculated the number of detected interactions drops from 86 to 14 for MC-LR, 55 to 20 for MC-LF, 96 to 8 for [Enantio-Adda5]MC-LF, and 108 to 11 for [β-D-Asp3,Dhb7]MC-RR. This finding suggests, that there are only a few interactions that occur frequently, and that the aggregated$_{occ30}$ frame gives a good overview over the main interactions, but also that a lot of information on interaction is lost when aggregating to one frame. The number of IFPs vary after aggregation, dependent on MC congener and type of aggregation. The original data set had 75,000 IFPs. The structural aggregation results in overall less IFPs, which is expected because it ignores the temporal component and therefore aggregates the IFPs more coarsely. This results in 12,634 to 23,725 individual IFPs per MC congener, which is still a too high number to analyse visually. In percentage, 17 % to 32 % of the original number of IFPs we had in the original data set. When aggregated over time, the number of IFPs is even higher, ranging from 64,140 to 71,764 which is 86 % to 96 % of the size of the original data frame.

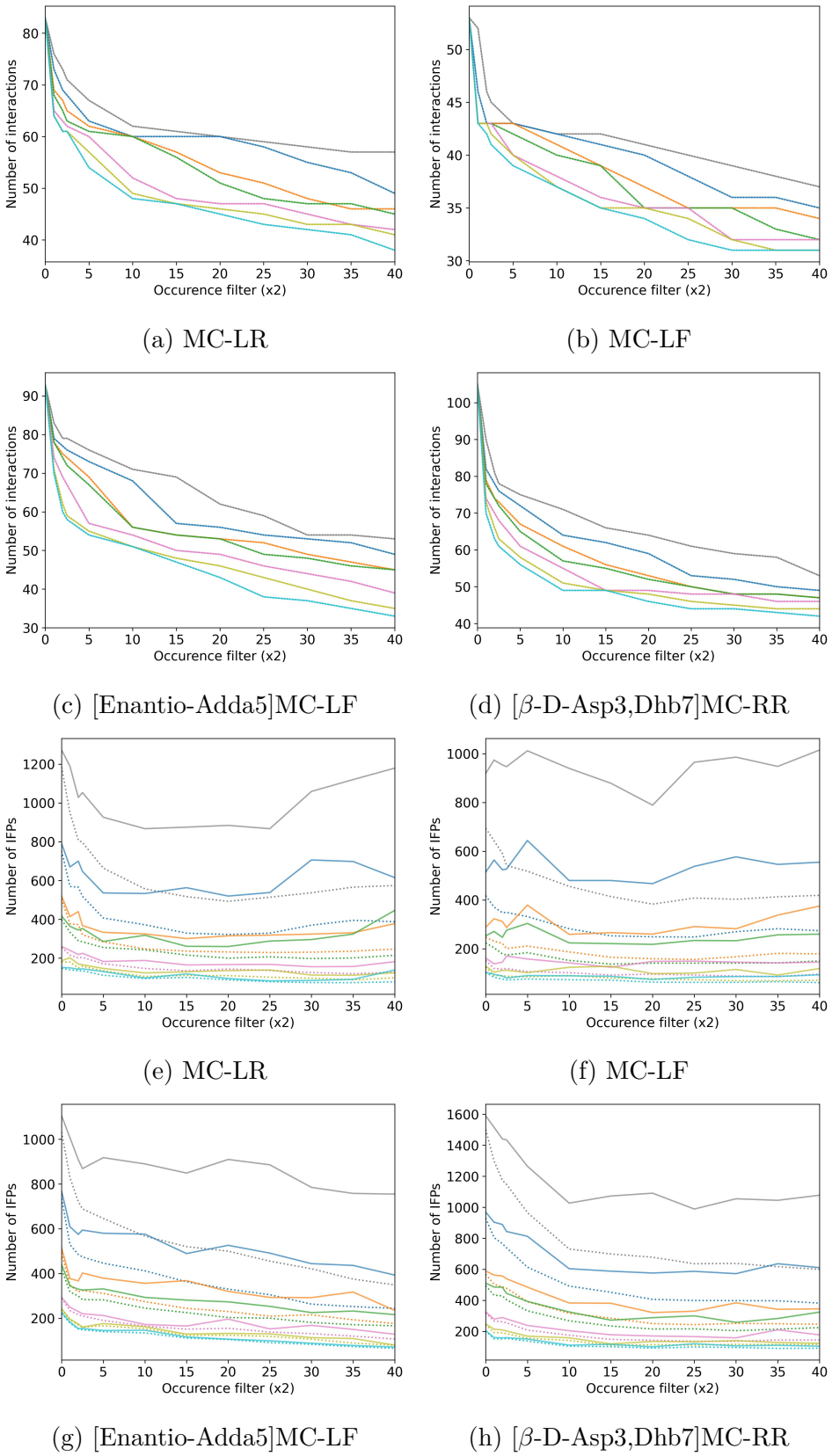
Table 5.2: Number of interactions and number of IFPs without filtering. The IFPs have been aggregated by structure or time. In addition, the number of interactions derived from $\text{aggregated}_{\text{occ30}}$ frame is compared to the original number of interactions.

	MC-LR		MC-LF		[Enantio-Adda5]MC-LF		[β-D-Asp3,Dhb7]MC-RR	
	structure	time	structure	time	structure	time	structure	time
number of interactions for $\text{aggregated}_{\text{occ30}}$ frame	14		20		8		11	
number of interactions	86		55		96		108	
number of IFPs	27529	71764	23009	71613	12634	64140	23725	68694

The second experiment to investigate the influence of x_1 (size of sliding window) and x_2 (occurrence within window) filtering showed that the choice of the filters has a strong effect on the number of interactions detected and the number of IFPs retrieved. Our main findings are that the smaller the sliding window is chosen, the more interactions are retrieved (see Figure 5.5a to Figure 5.5d), and that even the largest window size is chosen, more interactions could be retrieved than when simply summarising all IFPs to the $\text{aggregated}_{\text{occ30}}$ frame, as the number of interactions does not drop below 30 for any MC congener.

The number of interactions decreases depending on the filter value chosen for the occurrence within a window. In the beginning, when small values are filtered (0 to 2.5 %) there is a large decrease in the number of interactions for all MC congeners. This suggests that there are many interactions that occur very rarely, even within a smaller window. Therefore, we conclude that they are probably not relevant to the overall interactions, as they only exist for a really short time span in the simulation, and might either be not relevant or Artefacts. Beyond an occurrence filter of around 10 to 20 % for most window sizes, increasing the filter does not lead to much further reduction of interactions. For MC-LF and [Enantio-Adda5]MC-LF the number of interactions still drops slightly more compared to MC-LR and [β-D-Asp3,Dhb7]MC-RR, but overall not as fast as for windows smaller than 0.05. We therefore decided, that the x_2 filter must be at least 5 %. Also, for the number of IFPs there is no large difference between 10 % and 20 % for most MC congeners. For this reason, we decided on a x_2 filter value of 20 % for this data set, as this corresponds to 1.5 ns in our simulation. This is roughly the timescale on which side-chain rotation or fluctuation occurs (10^{-9} s) [427] and therefore seems to be biologically appropriate, since shorter time scales are less relevant to the interaction.

The number of IFPs (see Figure 5.5e to Figure 5.5h) is strongly dependent on the aggregation type. As hypothesised, the number of IFPs is lower for structural aggregation than for temporal aggregation because the temporal aggregation can revisit states of the simulation that are summarised in the structural aggregated IFPs. Overall, filtering the data based on a sliding window allows a massive aggregation of the data set. Irrespective of the window size chosen, less than 2.1 % of the original data set remains, resulting in a maximum of 1737 detected IFPs. For smaller window sizes (0.5 % and 1 %) we see a higher difference between structural and temporal aggregation, which disappears more and more for larger windows. Because the effect is so large for a window size of 0.5 %, we decided to exclude this window size here, as it probably also still contains a lot of noise in the temporal data. In contrast, for larger window sizes, almost no differences in number of IFPs between structural and temporal aggregation were present. For this reason, the large window sizes of 5 %, 7.5 % and 10 % were found to be unsuitable for the temporal aggregation of the data set evaluated here.



(a) MC-LR (b) MC-LF

(c) [Enantio-Adda5]MC-LF (d) [β-D-Asp3,Dhb7]MC-RR

(e) MC-LR (f) MC-LF

(g) [Enantio-Adda5]MC-LF (h) [β-D-Asp3,Dhb7]MC-RR

Figure 5.5: Number of interactions and interaction fingerprints after x_1 and x_2 filters are applied. a) to d) show the number of interactions, e) to h) the number of IFPs. The different x_1 filters are coloured by their value: 0.5 % is grey, 1 % is blue, 2 % is orange, 2.5 % is green, 5 % is pink, 7.5 % is yellow and 10 % is cyan. The x-axis shows the different x_2 filters. The solid and dashed lines represent the aggregation types, time and structure, respectively.

The filtering based on occurrence of interactions within a window can reduce the number of IFPs retrieved. Our main finding was that up to a certain threshold, the number of IFPs does indeed decrease, but after a threshold is reached, the number of IFPs can increase, resulting in a U-shaped curve. The observed effect is dependent on: 1) the type of aggregation, as structural aggregation is less prone to this observation, 2) the sliding window size, as the interactions are overall more stable for larger window sizes, and 3) the MC congener, as the effect is higher for MC-LR and MC-LF. At first glance, this seems counterintuitive. However, the filtering of occurrences does not necessarily reflect the number of IFPs left. As some interactions have their mean close to the thresholds evaluated, some of them are set to 0 while others are set to 1. The result may be the generation of previously unseen IFPs. For example, if the threshold is set to 20 % and some values are very close, e. g. 19.99 % and 20.001 %, the first one will be set to 0 (absent) and the second one to 1 (present). Therefore, at least one different IFP was generated, which was not present when the filter value was smaller than 20 %. For most MC congeners, the number of IFPs decrease until 10 % of occurrence within a window and increase again after 20 % of occurrence within a window.

For further comparison and visualisation of IFPs we have chosen the following aggregation and filtering settings based on our data set:

- Temporal aggregation: Since the temporal information of interactions and whether certain IFPs are revisited is of importance, the structural aggregation was dropped. The structural aggregation was initially considered since it was assumed it reduces the number of IFPs a lot, as it did on the unfiltered data set. As the number of IFPs between the two aggregation methods do not vary much, the temporal development is considered more important than summarising a few more IFPs.
- Sliding window filter of 1 %: Below 1 % the number of interactions decreases quickly, probably due to noise in the data. Large window sizes of 5 % and above indicate a high level of data smoothing, as structural and temporal aggregation do almost not differ in the number of interactions. For window sizes of 2 and 2.5 %, there were hardly any differences. These were therefore also excluded. The 1 % sliding window on this data set corresponds to approximately 7.5 ns.
- Occurrence filter of 20 %: the number of IFPs and interactions are relatively stable with an occurrence filter of 10 % to 20 %. We therefore decide on a filter of 20 % as this refers to a timescale which is biologically appropriate for our study.

5.4.4.2 Analysis of Interaction Fingerprints of Molecular Dynamics Simulation of 4 MC Congeners

In this section, two approaches to compare IFPs are discussed: 1) within the same simulation and 2) between different simulations. All IFPs presented were filtered based on a window size of 1% and an occurrence of interaction of 20 % within the window and IFPs aggregated over time.

IFP Comparison within MC Congeners Simulation To compare IFPs within the same simulation, the visualisation shown in Figure 5.6a has been proposed. Different visualisations and representations were combined to show different aspects of the data. At the top, the number of occurrences of each individual IFP is shown, complemented by a line plot showing

identical IFPs, which are connected by vertical lines between the horizontals a) and b). On the top right, the distribution of the number of differences to all IFPs is shown as a histogram, and at the bottom, as a matrix visualisation. The number of differences is mapped as colour on the matrix visualisation, and low numbers of differences are shown in blue and high numbers of differences in yellow. The matrix visualisation was also used in previous work to indicate development of trajectory similarity over time, as e. g. shown in ProLIF [391]. Here, the matrix visualisation is improved since it is complemented by additional visualisations which show different aspects. The results were then compared to the $\text{aggregated}_{\text{occ30}}$ frame proposed by Bouysset and Fiorucci [391], which was derived from the filtered data set i.e., before aggregation. The $\text{aggregated}_{\text{occ30}}$ frames for all MC congeners are summarised in SI Table 8.12.

In the following, the main findings will be summarised, and then known interaction sites reported in literature will be briefly described. The results and visualisations for each individual MC congener are then presented, by first addressing the overall development of IFPs and then describing individual representative IFPs.

To summarise, the main findings of this section are:

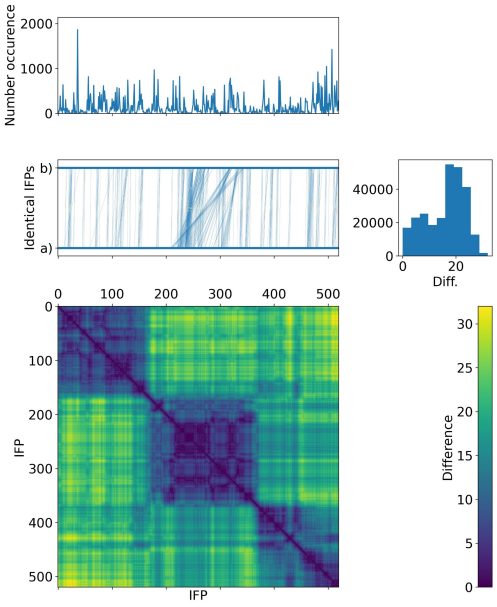
- IFPs close in time are highly similar with low number of differences. Over the course of the trajectory, they start accumulating small changes which result in overall larger changes in interaction patterns.

- Interactions that were described in literature could be identified in our IFP network visualisation, but they never occur all at once.

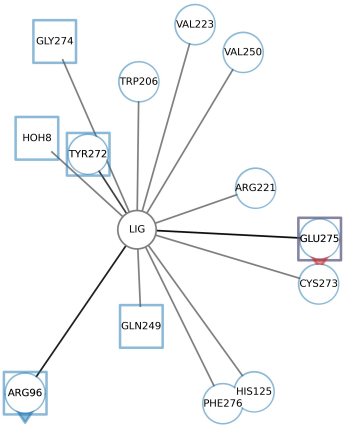
- The $\text{aggregated}_{\text{occ30}}$ frame identifies important interactions, but could not identify all interactions that occur in highly frequent IFPs.

The known interactions of MC congeners with PPP1 have been thoroughly summarised in a review by Fontanillo and Köhn [216] and discussed in Section 3.4.3. They are briefly summarised here to give the reader a quick overview. Hydrogen bonds can be formed with Arg96, Tyr134, and water (indirect coordination to manganese ions). MC congeners replace interaction with water by interacting with residues Asn124, His125, Ile130, Tyr134 and Trp206. Hydrophobic interactions are built with Cys127, Ile130, Ile133, Trp206, Tyr272 and Gly274. A covalent bond could potentially form with Cys273, which is not observed here, as no bonds can be broken or formed with classical MD simulations. Several residues can have multiple different interactions with MC congeners, so an amino acid is not restricted to one type of interaction.

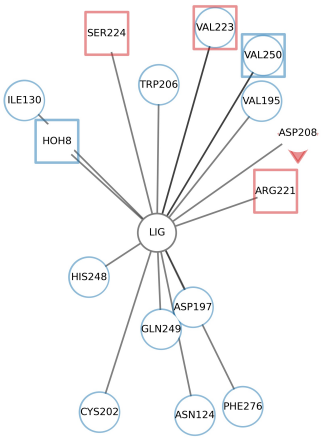
For **MC-LR**, the visualisation in Figure 5.6a shows three areas of higher similarity, which are indicated by large blue squares. Within those areas, there are well-defined areas of higher similarity and therefore a low number of differences. Over time, the IFPs get more diverse as changes accumulate, which is indicated by more yellow areas. The line plot above the matrix visualisation shows a vertical connection between identical IFPs and confirms that the majority of the identical IFPs are located in proximity to each other in time. The number of distances to all other IFPs in the histogram shows two peaks: a small peak around 10 differences, and a second peak around 20 differences. This finding also suggests that there are some IFPs that are closely related to each other and more distant from others. The line plot of occurrence of each IFP shows that there are several peaks which identify representative IFPs.



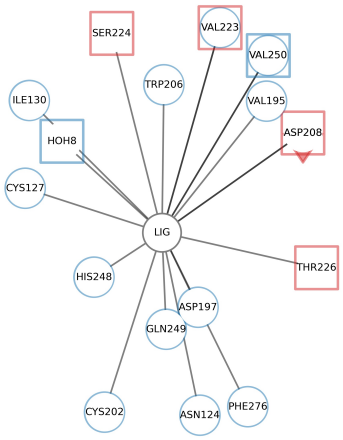
(a) Number of differences and occurrence



(b) IFP 36: 1864 occurrences



(c) IFP 497: 1042 occurrences



(d) IFP 507: 1425 occurrences

Figure 5.6: Comparison of IFP similarity within MC-LR. a) shows the occurrence and identical IFPs (connected by vertical lines) as line plot. The number of differences to all IFPs are shown as histogram (top right), and as colour on the matrix visualisation. In b), c) and d) the most frequent IFPs are shown sorted by time.

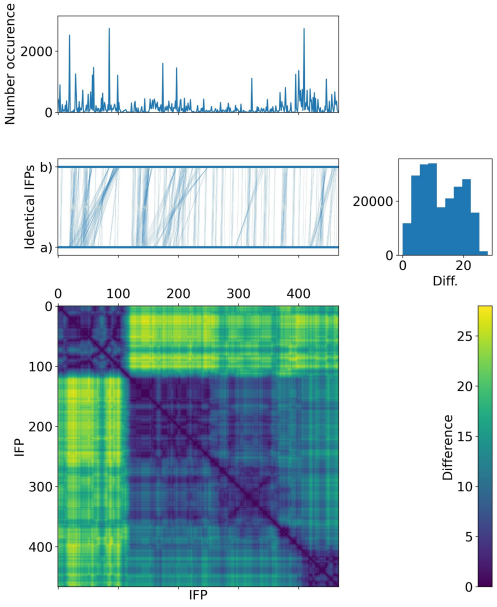
For a more detailed analysis of the IFP patterns, the three most frequently occurring IFPs are analysed and compared in the following. IFP 36 (Figure 5.6b) is located at the beginning of the simulation, while IFP 497 (Figure 5.6c) and IFP 507 (Figure 5.6d) occur towards the end of the simulation. Interestingly, some known interactions are found for all IFPs (Trp206, HOH), others only for IFP 36 (His125, Tyr272, Cys273, Gly274) or for IFP 497 and IFP 507 (Asn124,Ile130). Cys127 can only be identified for IFP 507 and three known interactions cannot be found for any of the three IFPs (Ile133, His125, Tyr134). The interactions are not necessarily of the same type as described in literature, but we could show for MC-LR that the IFP could recover interactions which were described in literature and are known to be

important for interaction, although they do not all occur at the same time. Phe276, Val223 and Gln249 occurred in the three IFPs, but were not described in literature so far and could be interesting for further studies and evaluation. Phe276 and Val223 could also be identified with the aggregated$_{occ30}$ frame (see SI Table 8.12), but Gln249 was not detected.

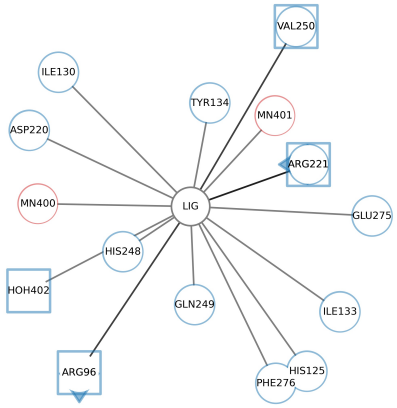
IFP 497 and IFP 507 are closer to each other in time and are also more similar to each other than to IFP 36. IFP 497 and IFP 507 have the same overall interaction pattern and differ in only four residues. IFP 497 has an additional interaction as hydrogen bond donor with Arg221 (red square), while IFP 507 has no interaction with this residue. In contrast, IFP 507 has hydrophobic interactions (blue circle) with Cys127 and cationic interactions (red square) with Thr226, which are both not present in IFP 497. Both interact with Asp208, but IFP 497 has a cationic interaction (red arrow down) and IFP 507 is a hydrogen bond donor (red square). IFP 36 differs quite a lot from the other two IFPs with 6 different interaction partners (i.e., residues) and different interaction types.

For **MC-LF**, the visualisation in Figure 5.7a shows the separation into two major squares, where the first one is a well-defined, and the second one is a larger lighter blue square, in which smaller nested squares exist, that indicate also regions with high similarity. The distribution of differences to all other IFPs shows a similar trend and is divided into two peaks — a smaller one below 10 differences, and a larger one around 20 differences. In contrast to MC-LR, there are more identical IFPs visible as vertical lines, which tend to be close together in time. Again, we see two areas that are somewhat separated from each other. The further the IFPs are away from each other, the less similar they are. In the occurrence line plot we can observe that at the beginning of the simulation and towards the end of the simulation there are three IFPs that occur very frequently and are probably more relevant than the remaining IFPs, as they occur less frequently.

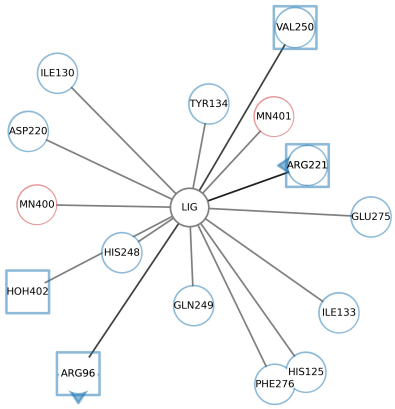
For the most frequent IFP patterns (see Figure 5.7b to d) of MC-LF, we can observe similar trend to those of MC-LR. The first two IFPs (IFP 19 and 85, see Figure 5.7b and c) are not similar, but identical and very frequent (5260 occurrences) and therefore probably highly relevant. IFP 409, which is further towards the end, has very different interactions and overall fewer interactions, but is also frequently occurring in the simulation with 2733 occurrences. Both unique IFPs share interactions known from literature (Arg96, Ile130 and Tyr134) and some are unique to IFP 19 and 85 (His125, Ile133, water, manganese) and some to IFP 409 (Cys127, Trp206, Tyr272). The only two interactions known in literature that do not occur for any of the IFPs are Asn124 and Cys273. With Cys273 MC congeners can form a covalent bond, which cannot be observed in the simulations, but proximity or non-covalent interaction to this residue would point towards correct orientation. In comparison to the aggregated$_{occ30}$ IFP, two interactions which were reported in literature could be found in IFP 409 here, but not in the IFP (Trp206 and Tyr272). In addition, four other interactions of IFP 19 and IFP 85 are missing, which are Glu275, Asp220, Gln249 and Phe276. The latter two were also found in the IFP of MC-LR, where Gln249 was also not found in the aggregated$_{occ30}$ IFP.



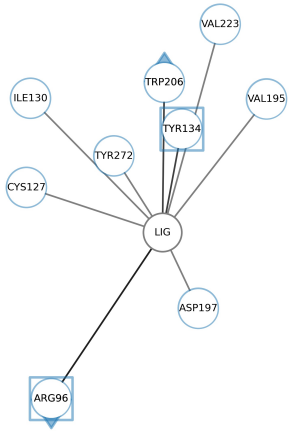
(a) Number of differences and occurrence



(b) IFP 19: 2520 occurrences



(c) IFP 85: 2740 occurrences

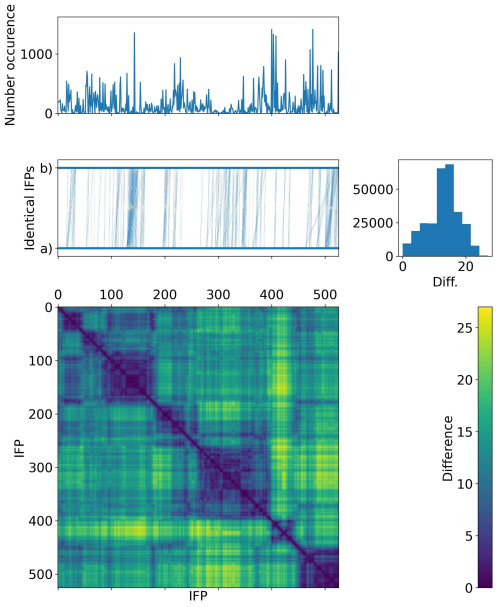


(d) IFP 409: 2733 occurrences

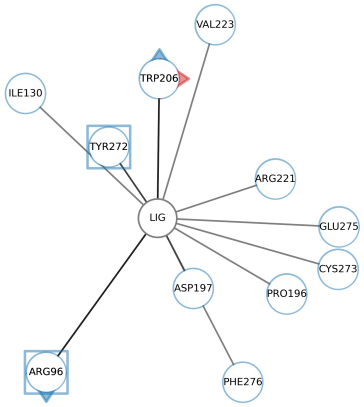
Figure 5.7: Comparison of IFP similarity within MC-LF. a) shows the occurrence and identical IFPs (connected by vertical lines) as line plot. The number of differences to all IFPs are shown as histogram (top right), and as colour on the matrix visualisation. In b), c) and d) the most frequent IFPs are shown sorted by time.

For **[Enantio-Adda5]MC-LF** the visualisation in Figure 5.8a shows, in contrast to the first two MC congeners, that the simulation is broken down into smaller squares with less high order arrangements. The squares are more defined at the beginning and end of the simulation, and become larger towards the middle. A yellow area with the appearance of a line is clearly visible and is an indication of an interaction pattern, that is different from the rest of the simulation. Towards the end, there is another well-defined dark blue square, indicating similar interactions. In contrast to the other MC congeners, the number of differences between all IFPs is normally distributed and the peak is located in the range of 15 to 20. Compared to the previous two IFP sets, the line plot of identical IFPs shows that they are even more concentrated around

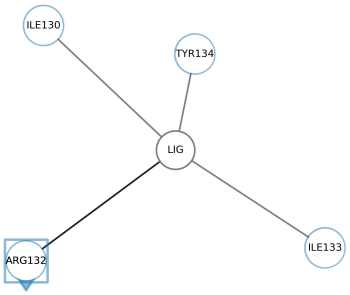
the IFPs close in time. This is probably the result of the similarity being broken down into smaller blocks compared to the other MC congeners. In addition, the vertical lines appear less dense than for MC-LF, indicating less similar IFPs within [Enantio-Adda5]MC-LF.



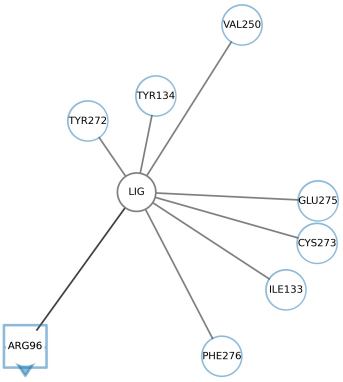
(a) Number of differences and occurrence



(b) IFP 143: 1355 occurrences



(c) IFP 400: 1412 occurrences



(d) IFP 477: 1412 occurrences

Figure 5.8: Comparison of IFP similarity within [Enantio-Adda5]MC-LF. a) shows the occurrence and identical IFPs (connected by vertical lines) as line plot. The number of differences to all IFPs are shown as histogram (top right), and as colour on the matrix visualisation. In b), c) and d) the most frequent IFPs are shown sorted by time.

The most frequent IFP patterns (see Figure 5.8b to 5.8d) of [Enantio-Adda5]MC-LF, show that towards the end of the simulation, the two frequent IFPs 400 and 477 built much fewer interactions than the other IFPs analysed so far. IFP 400 has only six interactions, three of them are with Arg132. The other three interactions are formed with known important residues (Ile130, Ile133, Tyr134). The same observations as for IFP 400 are true for IFP 477. It also interacts with Glu275 and Phe276 (also for IFP 143), which were also observed for MC-LF and

MC-LR. IFP 143 forms many hydrophobic interactions, some of which have been described in literature (Ile130, Trp206, Tyr272). The interaction with Pro196 was not observed so far for any IFP. The aggregated$_{occ30}$ IFP is also sparse in comparison to other MC congeners. From the simulations presented in Section 4.3, we know that [Enantio-Adda5]MC-LF moves out of the binding pocket in the third replicate and flips back in the binding site with the wrong orientation. This is probably the reason for the low number of interactions observed here for the major IFPs. Nevertheless, it is surprising that an IFP with so few interactions appears so often and therefore seems to be a major representation.

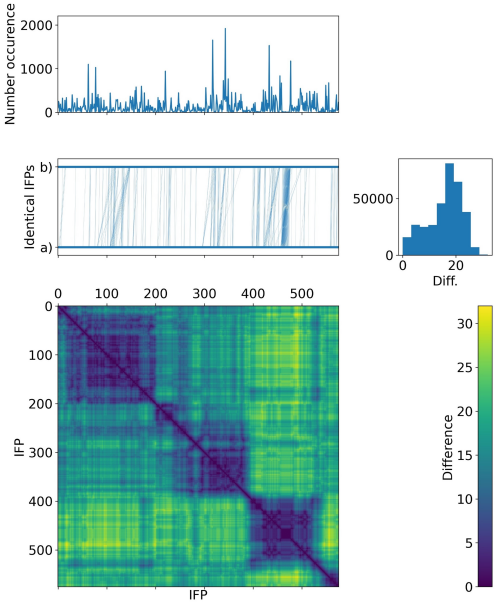
For [β**-D-Asp3,Dhb7]MC-RR** the visualisation in Figure 5.9a shows three larger and two smaller dark blue squares, indicating 5 IFP patterns in the simulations. The blocks are better defined as for [Enantio-Adda5]MC-LF, but less well-defined than for MC-LR and MC-LF. The number of differences is mostly shifted towards 20 to a large peak. There are fewer identical IFPs for [β-D-Asp3,Dhb7]MC-RR, as the vertical lines in the line plot are less dense than for the other MC congeners. The number of occurrences indicates few structures that occur frequently.

The most frequent IFP patterns (see Figure 5.9b to 5.9d) show that the interaction pattern for [β-D-Asp3,Dhb7]MC-RR is different compared to the other MC congeners, at least from a perspective of the most common IFPs.

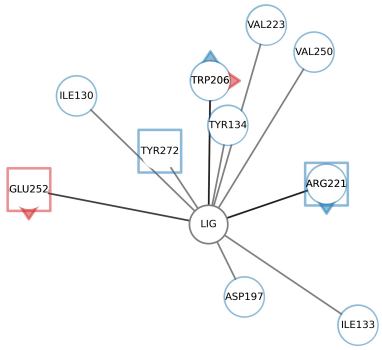
IFP 317 and IFP 343 are very similar to each other and have only two unique residues, Asp197 and Arg96, respectively. Surprisingly, IFP 433 interacts with three aspartates, and differs from the other representative IFPs. The interaction with aspartates has not been reported in literature and were also not identified in the aggregated$_{occ30}$ IFP. Only one of the aspartates has been found in the other IFPs, which was for MC-LR. In contrast to the other IFPs we could not identify the residues Glu275 and Phe276, which were found for all other MC congeners.

To **summarise**, our results suggest that changes in interactions accumulate slowly over the course of the trajectory. We can also observe that there are some IFPs that occur more frequently than others and are probably good representatives of the respective blocks. This is also reflected in the most frequently occurring IFPs within MC congeners, which are more similar to each other when they are also close in time. In addition, interactions described in literature can be identified. However, they often occur in different IFPs and not at the same time, i.e., within the same IFP. Furthermore, a comparison with the aggregated$_{occ30}$ IFP showed, that considering individual IFPs indeed reveals differences in interaction patterns, especially for less frequent interactions. The less frequent interactions are still relevant, because they occur in IFPs which have a high occurrence and therefore exist over a period of time. In addition, we could identify interactions which were not reported in literature so far, and were present for 3 out of 4 MC congeners (Phe276 and Glu275). Overall, further investigation of less frequent IFP patterns is needed to get a better picture of the binding processes and interactions. The results described here are an interesting starting point and show that a closer look at individual IFPs derived from the aggregated IFPs can yield interesting information.

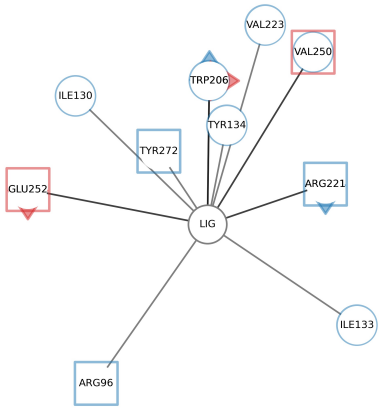
To visualise the temporal change of interactions per amino acid residue, a circular chart visualisation has been developed (see Figure 5.10). The different types of interactions that can occur per residue are visualised on the same circular chart using different colours. These charts give a summary and impression of which interactions are frequent or persistent and which disappear from time to time. However, this visualisation has been developed to help the user to navigate through the simulations and serve as a guidance but not to stand alone for data exploration, as this results in too many images to be included in this thesis.



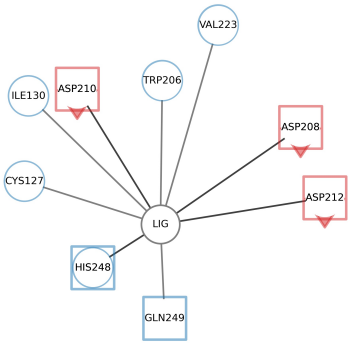
(a) Number of differences and occurrence



(b) IFP 317: 1655 occurrences



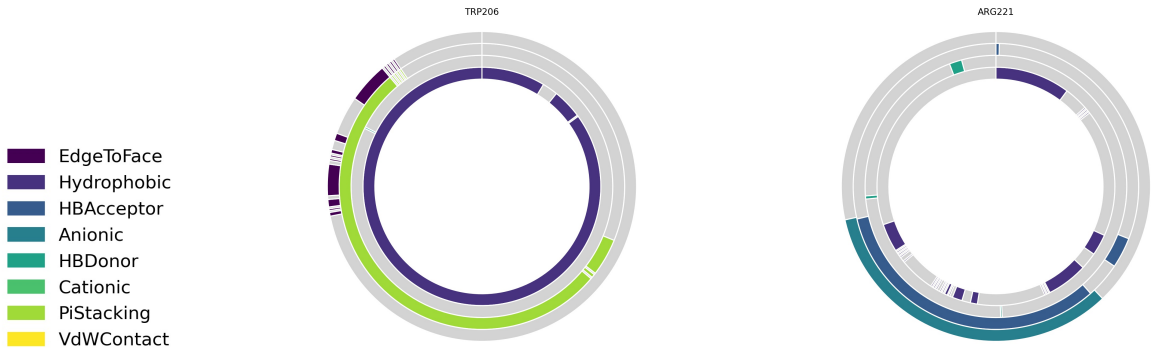
(c) IFP 343: 1921 occurrences



(d) IFP 433: 1531 occurrences

Figure 5.9: Comparison of IFP similarity within [β-D-Asp3,Dhb7]MC-RR. a) shows the occurrence and identical IFPs (connected by vertical lines) as line plot. The number of differences to all IFPs are shown as histogram (top right), and as colour on the matrix visualisation. In b), c) and d) the most frequent IFPs are shown sorted by time.

IFP Comparison between MC Congeners Simulation This section analyses the similarity of IFPs between different simulations and MC congeners. To compare two different sets of IFPs, the sets were merged based on their interactions, so that all interactions present in one set of IFP were also present in the other, because similarity calculation requires identical vector lengths, i.e., number of interactions. If an interaction is not present in an IFP set, it is set to 0 and encoded as absent, because both were simulated with the same protein and could have had an interaction with that residue. As IFPs from different simulations are now compared, the concept of similarity has been adapted, as counting the number of differences to separate identical from similar values is no longer appropriate. Therefore, the similarity



(a) Legend (b) TRP206 (c) ARG221

Figure 5.10: Examples of interacting residues of IFPs derived from MC-LR simulation. Some interacting residues are selected as an example, to show an example of circular chart visualisation.

is assessed using the inverse Rogers-Tanimoto dissimilarity and three classes were chosen to categorise IFPs. Different values for similarity have been proposed in literature, classifying a molecular structure as similar if the Tc is ≥ 0.5 [387] or ≥ 0.85 [386, 426]. The different values have been proposed because similarity is a fuzzy concept and also dependent on the data set [41, 98, 385, 386, 388]. Therefore, we considered two IFPs as identical, if the achieved similarity was ≥ 0.975, as similar if it was between 0.85 and 0.975, and as dissimilar if it was below 0.5. The values were assessed by visual inspection of the IFP data, and thresholds can be adjusted by the user.

The main results are summarised below, followed by a brief description of the assessment of similarity between IFP sets, before considering similarity for each possible pair of MC congeners.

To summarise, the main findings of this section are:

- Even with the new definition of identical and similar, identical IFPs from the same simulation tend to be close in time, while similar values may be further apart.
- When comparing IFPs between different simulations, they are overall less similar than within the same simulation. Therefore, we conclude that although MC congeners share some binding patterns, they also have different binding patterns.
- IFPs identified as similar (or identical) between simulations share their interactions overall and have similar patterns. The patterns overlap with interactions reported in literature, but not all of them occur together in the same network. The same is true when comparing the individual IFPs with the aggregated$_{\text{occ30}}$ IFP.
- The toxic MC-LR and MC-LF have the highest overlap of similar IFPs, and at the same time the highest proportion of dissimilar IFPs, suggesting common interaction patterns, but also distinct interactions.
- The toxic and non-toxic stereoisomers MC-LF and [Enantio-Adda5]MC-LF share almost no interaction patterns and have very different interaction patterns

- [Enantio-Adda5]MC-LF has a surprisingly high proportion of similar IFPs to MC-LR

To quantify the similarity between the IFP sets, the proportions of identical, similar and dissimilar values were calculated by occurrence and percentage of IFPs and are summarised in SI Table 8.13. To visually compare similarity between IFP sets, line plots were generated as described in Section 5.4.3.3. When comparing the similarity for two MC congeners in one plot, the similarity within the MC congener IFPs is also visualised, as the similarity criteria have been redefined and a comparison is necessary to interpret the results appropriately. In all images (see Figures 5.12a to Figures 5.17a) one MC congener is shown in dark blue colour at the bottom, the other one in light blue at the top. Between *a*-*b* and *e*-*f* identical IFPs within the same simulation are shown, between *b*-*c* and *d*-*e* similar IFPs within the same simulation are shown. Between *c*-*d* identical and similar IFPs are shown by comparing the IFP sets between the simulations in cyan and red, respectively. The identical values are not visible for this data set, as only one position was found for a pair of MC congeners. To compare the IFPs in more detail, similar or identical IFPs, that map from one MC congener to the IFPs of the other are selected. For this approach, the most frequently occurring IFP was selected together with the corresponding IFP it maps to in the other set of IFPs and is shown as an addition to the line plots.

In SI Table 8.13 the number and percentages of IFPs in a similarity class are summarised. The numbers do not add up to 100 %, since one IFP can fall into multiple classes depending on which reference IFP is used for similarity calculation, which means that all values reflect the total number in comparison to the data set, i.e., a percentage value can be a maximum of 100 %. The values are summarised visually in Figure 5.11 to provide a quick overview. The individual values percentage values (see SI Table 8.13) will be discussed together with the visualisations of line plots in the following.

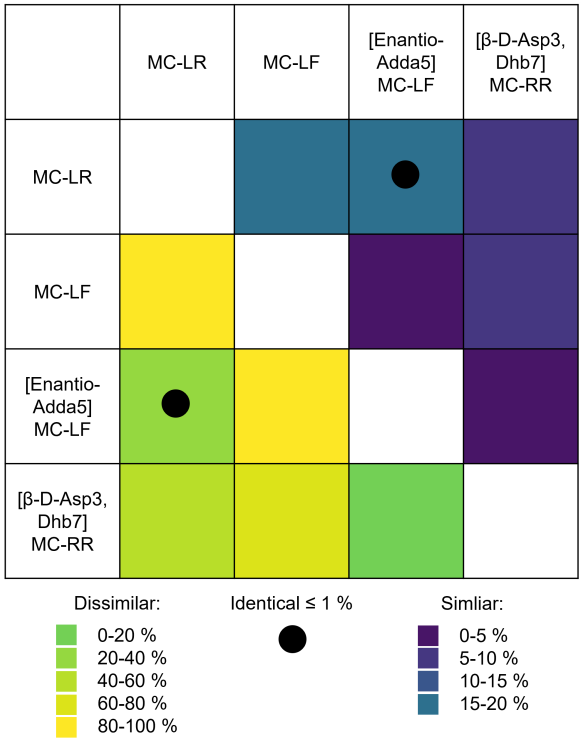
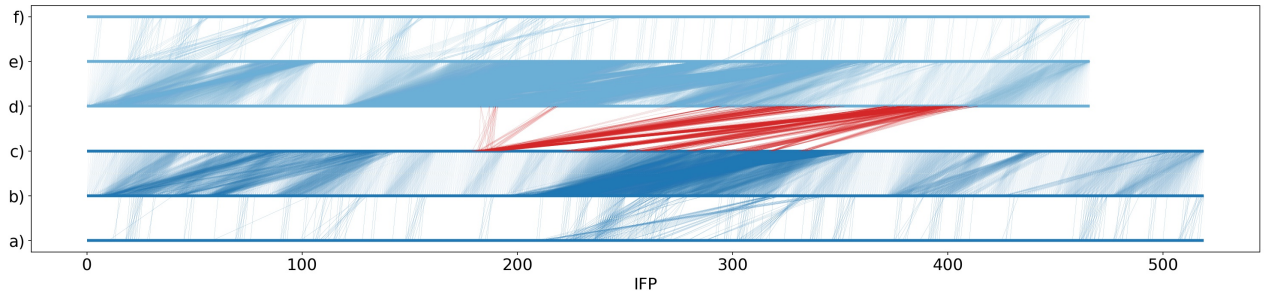


Figure 5.11: Summary of similarity achieved for IFPs of different simulations for the small MC congener data set. The percentage of IFPs that are similar or dissimilar are colour coded according to the legend on the plot. Identical values are indicated by a glyph, if present. The upper right of the diagonal represents similar IFPs, the lower left diagonal represents dissimilar IFPs.

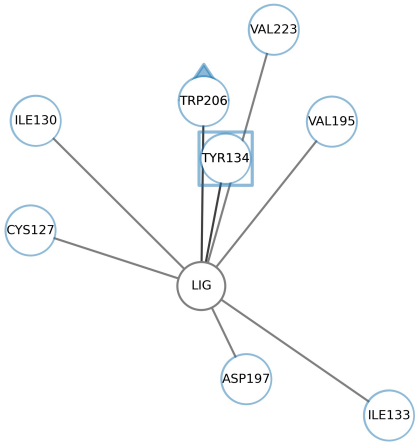
The IFP sets of **MC-LR and MC-LF** are summarised in Figure 5.11, Figure 5.12, and SI Table 8.13. We could show that the toxic MC congeners share certain binding patterns which confirms our initial hypothesis, but also that they have distinct IFP patterns, which are MC congener specific. No identical IFPs could be identified between the two sets, even though both of them are toxic. On the other hand, MC-LR and MC-LF have the highest proportion of similar IFPs (about 18 %) identified over all comparisons with different MC congeners.

Several regions of MC-LR map back to approximately four larger areas in MC-LF. Consequently, there seems to be a larger connected area that is similar for both MC congeners, but otherwise not similar. MC-LR has the highest overlap with MC-LF in comparison to all MC congeners, which indicates common interaction patterns. The overlap is nevertheless centred around a certain area, so they do also differ partly in their interaction behaviour. In Figure 5.12b and 5.12c, a set of two similar IFPs of both MC congeners that map to each other were selected based on the majority of occurrences. The two reference IFPs for comparison occur 755 times in MC-LR (IFP 185) and 2733 in MC-LF (IFP 409). Both share an overall interaction pattern, and all interactions present in MC-LR network are also present in MC-LF, except for one interaction with Ile133, which is known to be important for interaction. MC-LF builds two interactions that are known to be important, which is with Arg96 and Tyr272, but those could be also identified for MC-LR, which was described in the previous section. The additional interactions described in literature for Tyr134, Ile130, Ile133, Trp206 and Cys127 could be identified and retrieved. The aggregated$_{occ30}$ IFP is quite similar to the IFPs presented here,

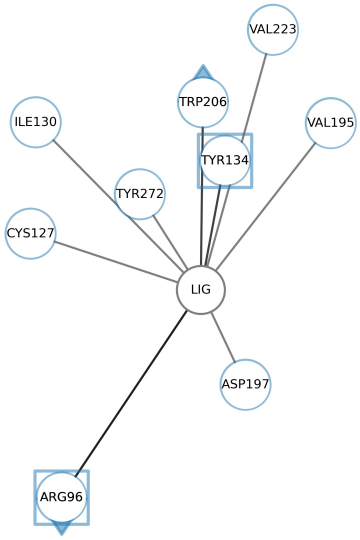
but has a total of two missed interactions in comparison to the respective IFP. Even though MC-LR and MC-LF share many of their interactions in the selected representative IFPs and also by percentage compared to the other MC congeners, they also share a high percentage of dissimilar IFPs (about 86 %).



(a) MC-LR & MC-LF



(b) MC-LR IFP 185: 755 occ.



(c) MC-LF IFP 409: 2733 occ.

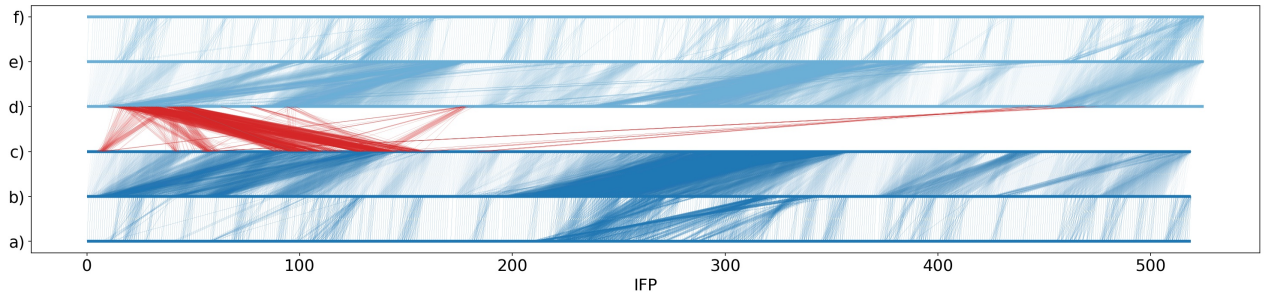
Figure 5.12: Comparison of IFPs between MC-LR and MC-LF simulations. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* similar IFPs between simulations are shown as red lines. In b) and c) one representative of similar IFPs is shown per MC congener.

Surprisingly, the comparison of **MC-LR with [Enantio-Adda5]MC-LF** (see Figure 5.13a) showed that they share an identical interaction (0.16 %), a number of similar IFPs (16.9 %), and low amount of dissimilar IFPs (29.11 %, see Figure 5.11 and SI Table 8.13). This result is not what we hypothesised, as both congeners belong to different toxicity classes, with MC-LR being highly toxic and [Enantio-Adda5]being non-toxic.
Interestingly, the overlap occurs mostly at the beginning of the simulation and only a little towards the end. Therefore, the IFP sets seem to become more diverse over time. Nevertheless, the number of similar interactions detected is close to that of MC-LR with MC-LF: 16.9 %

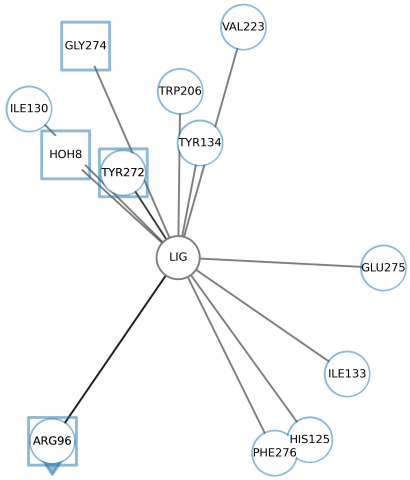
compared to 18.1 %, respectively.

We observed one overlap with identical IFPs, which is not visible in the visualisation because it is a very thin line. When studying identical representative IFPs (see Figure 5.13b and 5.13c), we observe the same interactions except for one, which is Gly274 which only occurs in MC-LR. Also, here we can detect almost all (except for Cys273, Asn124, Cys127) interactions that are known from literature (such as HOH8, Arg94, Tyr134, Ile133, His125, Trp206, Tyr272, Gly274). At the beginning of the simulation, the [Enantio-Adda5]MC-LF was probably closer to its docked pose, which might explain the higher similarity to MC-LR, which was used as a reference for a correct docking pose. In addition, the two frames occur only 102 and 129 times, which is in comparison to other representative networks much less and therefore probably not as important as it is not really conserved.

For the similar IFPs (see Figure 5.13d and 5.13e) more differences are observable compared to identical representative IFPs. Nevertheless, the IFPs for MC-LR only have one difference to each other, as IFP 129 has one additional interaction for Glu275. The two IFPs 42 and 63 of [Enantio-Adda5]MC-LF IFP have in total six differences from each other, as three interactions disappear (Phe276 and Glu275, HOH8) and three appear (Cys202, Val250, HOH402), although the exchange of the water molecule by the other water molecule is probably not that relevant. When comparing MC-LR IFP 129 with [Enantio-Adda5]MC-LF IFP 63, both of them have much higher occurrences than the identical IFP before. In addition, they have six differences, which is so far a relatively high number. Both of them have interactions reported in literature, and some of them are shared across the IFPs.

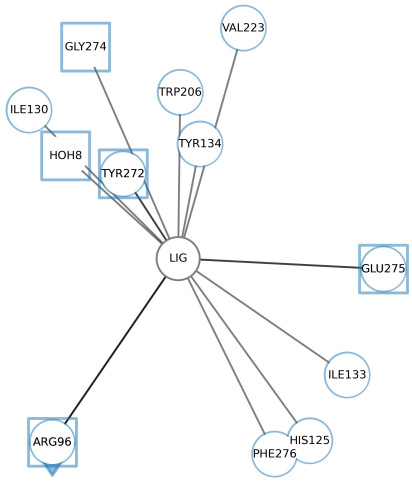


(a) MC-LR & [Enantio-Adda5]MC-LF

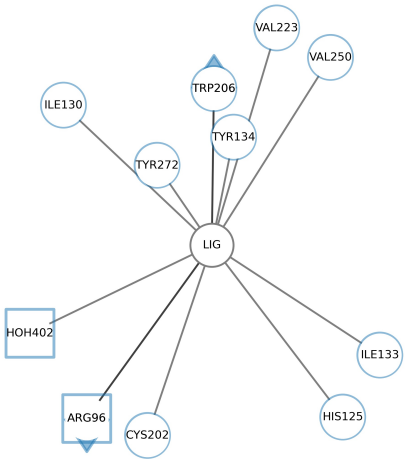


(b) MC-LR IFP 98: 102 occ.

(c) [Enantio-Adda5]MC-LF IFP 42: 129 occ.



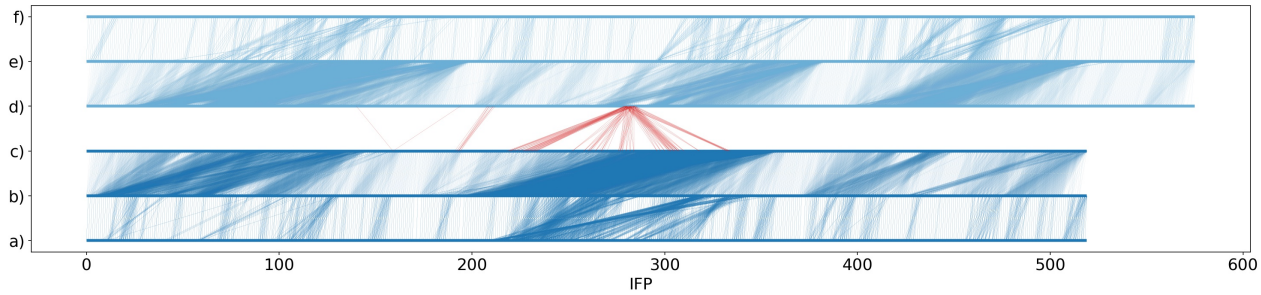
(d) MC-LR IFP 129: 741 occ.



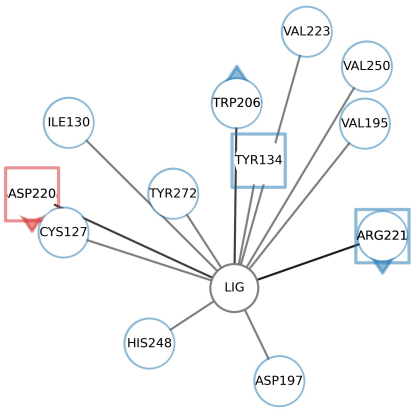
(e) [Enantio-Adda5]MC-LF IFP 63: 661 occ.

Figure 5.13: Comparison of IFPs between MC-LR and [Enantio-Adda5]MC-LF simulations. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* identical and similar IFPs between simulations are shown as cyan and red lines, respectively. In b) and c) one representative of identical and in d) and e) of similar IFPs are shown per MC congener.

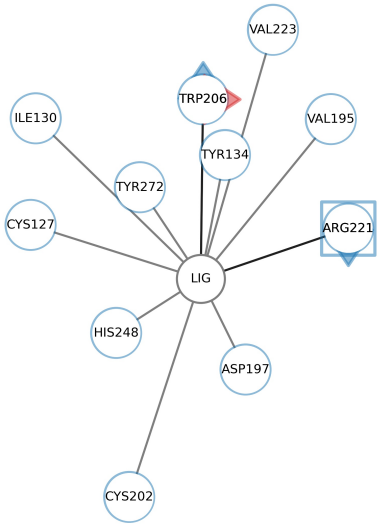
MC-LR and [β-D-Asp3,Dhb7]MC-RR have a low similarity for both sets (see Figure 5.11 and SI Table 8.13). The percentage of similar values is only 5.31 %, which is much lower than the values previously observed for the other MC congeners. The dissimilar IFPs are close to 57 % which is lower than compared to other MC congeners. Therefore, a large proportion of compounds must exist that is neither similar nor dissimilar. Several positions of MC-LR in the middle of the IFPs map to one region in IFPs of [β-D-Asp3,Dhb7]MC-RR, but the overlap is small (see Figure 5.14a). This might be caused by the modification of a hydrophobic leucine residue to a charged arginine residue, which could influence interaction a lot. In addition, the modified backbone of [β-D-Asp3,Dhb7]MC-RR leads to different backbone conformation of this MC congener, which might also highly influence binding (see Section 4.3.3.3). In total, they differ at three positions (Asp220, Val250, Cys202) in the representative IFPs (see Figure 5.14b and 5.14c).



(a) MC-LR & [β-D-Asp3,Dhb7]MC-RR



(b) MC-LR IFP 218: 611 occ.

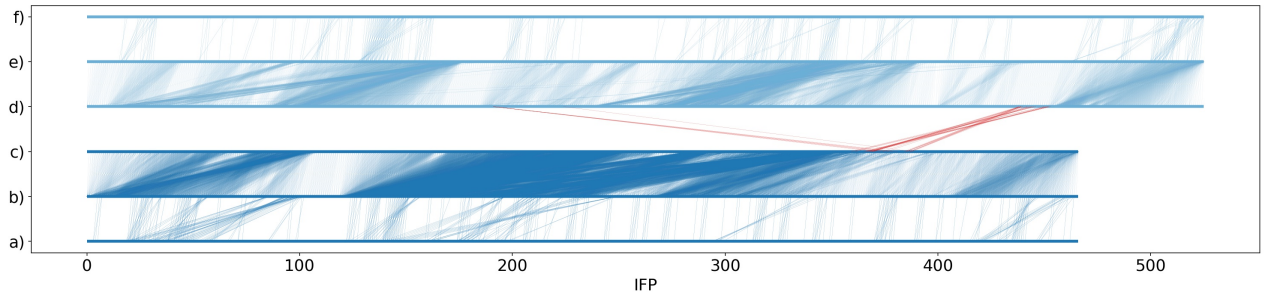


(c) [β-D-Asp3,Dhb7]MC-RR IFP 281: 116 occ.

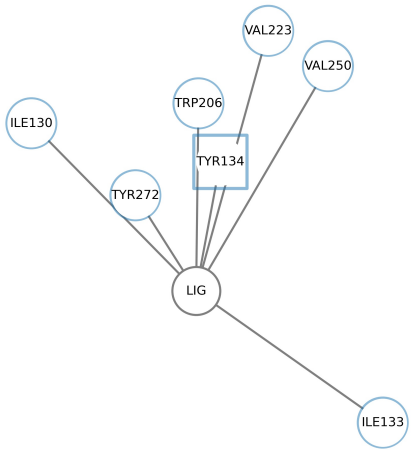
Figure 5.14: Comparison of IFPs between MC-LR and [β-D-Asp3,Dhb7]MC-RR simulations. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* similar IFPs between simulations are shown as red lines. In b) and c) one representative of similar IFPs is shown per MC congener.

MC-LF and [Enantio-Adda5]MC-LF do only share a few similar IFPs in the merged

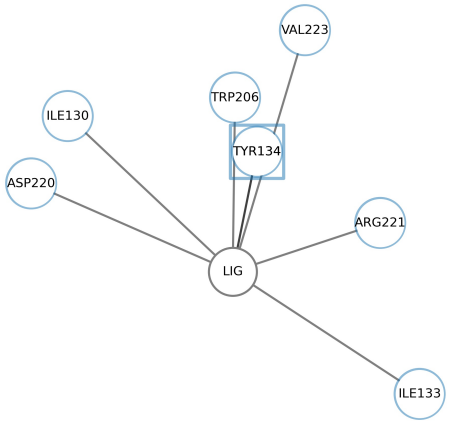
interaction set with 2.76 % of the IFPs, which is the lowest number detected (see Figure 5.11 and SI Table 8.13). We could clearly show that the IFP patterns do not overlap, so despite their small structural difference, they differ a lot in their IFPs which might help to analyse their toxicity (see Figure 5.15a). In addition, they also have a high number of dissimilar IFPs detected, with 84.73 %. Both compounds belong to different toxicity classes, i.e., the toxic and non-toxic class, but do only differ in the stereocenters of the Adda side chain. Therefore, the compounds are very similar to each other, and difficult to distinguish for many computational methods.



(a) MC-LF & [Enantio-Adda5]MC-LF



(b) MC-LF IFP 369: 648 occ.



(c) [Enantio-Adda5]MC-LF IFP 444: 289 occ.

Figure 5.15: Comparison of IFPs between MC-LF and [Enantio-Adda5]MC-LF simulations. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* similar IFPs between simulations are shown as red lines. In b) and c) one representative of similar IFPs is shown per MC congener.

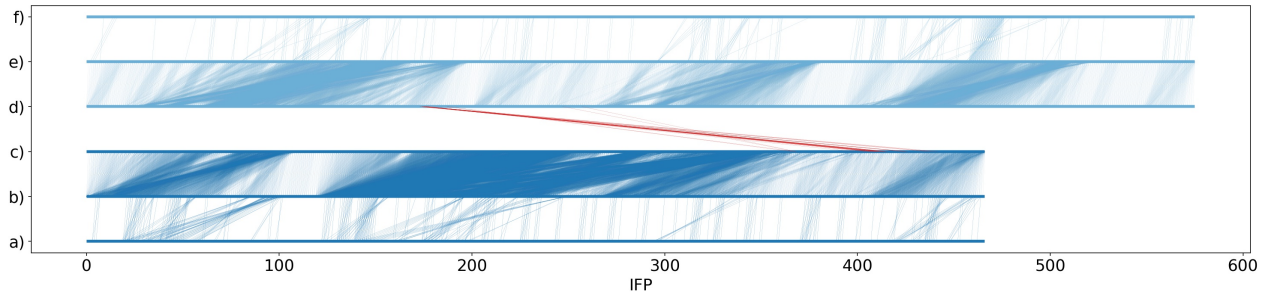
The representative IFPs in Figures 5.15b and 5.15c, are both IFPs with a low number of interactions (seven for MC-LF and eight for [Enantio-Adda5]MC-LF) compared to the other representative IFPs analysed so far. However, most of the interactions were in agreement with literature. Probably this is a stable core that almost always forms when MC interact with

PPPs, and because both IFPs are rather sparse with not many interactions, they could match here, even though they both occur not that long in comparison to other IFPs in their respective sets. For MC-LF the IFP even occurs quite often with 648 occurrences compared to other IFPs, but with the representative IFPs studied in the previous section with interactions occurring more than 2700, it is not that frequent. Also, for [Enantio-Adda5]MC-LF, the occurrence of 289 is considered low. Both representative IFPs only differ in two interacting residues, namely Val250 and Tyr272 which is missing for [Enantio-Adda5]MC-LF, and Asp220 and Arg221 which is missing for MC-LF.

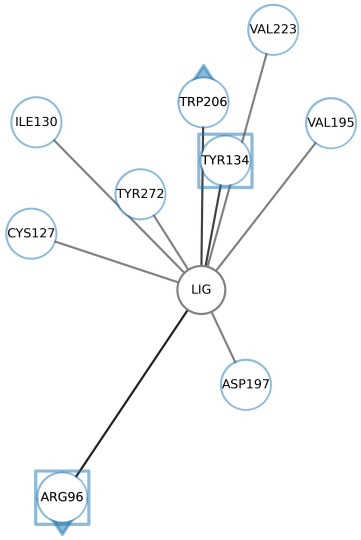
MC-LF and [β-D-Asp3,Dhb7]MC-RR share some interactions for IFPs. A small area of MC-LF towards the end of the simulation maps to a small area of [β-D-Asp3,Dhb7]MC-RR in the beginning of the simulation (see Figure 5.16a). This results in 5.85 % of IFPs that are classified as similar to each other (see Figure 5.11 and SI Table 8.13). On the other hand, a high percentage (almost 78 %) of IFPs are considered dissimilar, which is quite a lot. The representative IFP of MC-LF is nevertheless interesting, as it is one that occurs very frequently in the simulations, with an occurrence of 2733 (see Figure 5.16b). In contrast, the IFP of [β-D-Asp3,Dhb7]MC-RR appears less frequently (446 occurrences). Both IFPs have only 4 differences to each other, with Asp197 being unique to MC-LF and Mn400, Glu275 and Val250 being unique to [β-D-Asp3,Dhb7]MC-RR.

For comparison of **[Enantio-Adda5]-MC-LF and [β-D-Asp3,Dhb7]MC-RR** interaction pattern, we could show that they do not share similar binding patterns, but they are also not dissimilar either, which would be in accordance with the toxicity classes of non and less toxic. When calculating the proportion of similar IFPs, [Enantio-Adda5]and [β-D-Asp3,Dhb7]MC-RR have only 3.81 % of frames which are considered similar, which is the second-lowest number for all MC congeners (see Figure 5.11 and SI Table 8.13). On the other hand, the proportion of dissimilar IFPs was only identified to be 14.14 % which is the lowest number found so far.

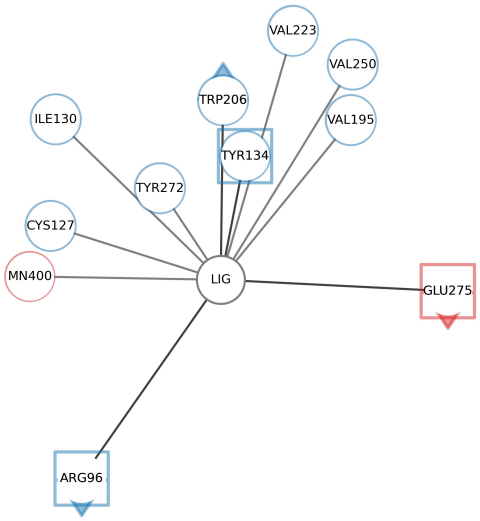
The interaction patterns of [Enantio-Adda5]-MC-LF and [β-D-Asp3,Dhb7]MC-RR do overlap in some small areas (see Figure 5.17a). Nevertheless, the overlap is mostly located at one area located towards the end of the simulation for [Enantio-Adda5]-MC-LF, and maps to the first third of the simulation of [β-D-Asp3,Dhb7]MC-RR. Since the position is so narrow on [Enantio-Adda5]-MC-LF and highly scattered across [β-D-Asp3,Dhb7]MC-RR, we assume that the similarity between IFPs is rather low. When analysing the representative IFPs, it is also obvious that here the IFPs are very sparse with eight and eleven interactions (see Figure 5.17b and 5.17c) and that the most frequent occurring IFPs are also not occurring often with only 289 and 447 occurrences, respectively.



(a) MC-LF & [β-D-Asp3,Dhb7]MC-RR

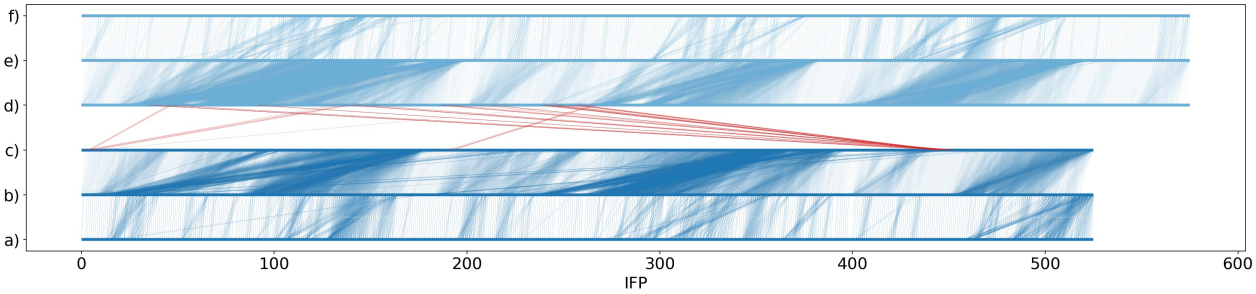


(b) MC-LF IFP 409: 2733 occ.

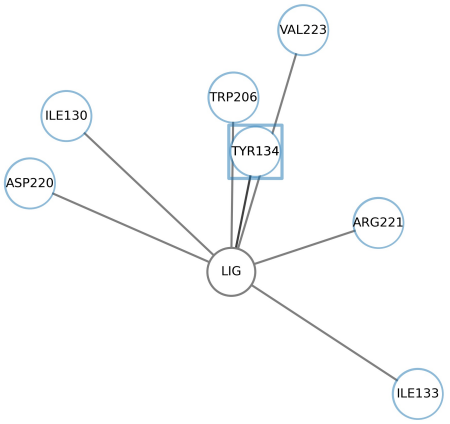


(c) [β-D-Asp3,Dhb7]MC-RR IFP 177: 446 occ.

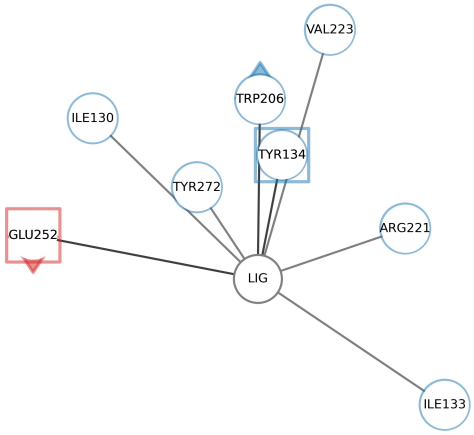
Figure 5.16: Comparison of IFPs between MC-LF and [β-D-Asp3,Dhb7]MC-RR simulations. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* similar IFPs between simulations are shown as red lines. In b) and c) one representative of similar IFPs is shown per MC congener.



(a) [Enantio-Adda5]MC-LF & [β-D-Asp3,Dhb7]MC-RR



(b) [Enantio-Adda5]MC-LF IFP 444: 289 occ.



(c) [β-D-Asp3,Dhb7]MC-RR IFP 247: 447 occ.

Figure 5.17: Comparison of IFPs between [Enantio-Adda5]MC-LF and [β-D-Asp3,Dhb7]MC-RR simulations. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* similar IFPs between simulations are shown as red lines. In b) and c) one representative of similar IFPs is shown per MC congener.

5.4.4.3 Analysis of Interaction Fingerprints of Molecular Dynamics Simulation of 12 MC Congeners

In this section, the IFPs generated from the second MD data set which were described in Section 4.4 are analysed in more detail. To the best of our knowledge, no data is available on the interaction of MC congener GSH/Cys conjugates with PPP1. Therefore, we restrict our analysis to the MC congeners rather than the full set because of the lack of data in the literature. In addition, many MC congeners move out of the binding pocket during the simulation (see Section 4.4.3.4). Therefore, the results must be viewed with caution, although they provide a good test case for the library developed here to evaluate how IFPs perform when very diverse interactions exist within a simulation.

The second data set was analysed based on the filtering evaluated on the previous data set. Therefore, the x_1 filter (sliding window) was calculated to be 0.5 %, because this data set is twice as long as the previous data set and we want to keep the amount of time (7.5 ns) which is considered within a window, to be able to compare the results with the results from the previous section. The x_2 filter (occurrence filter) is then applied as before with 20 %. Due to the size of the data set, the results presented and the discussion will be limited to the overall trends for the IFP development over time and comparison.

IFP Comparison within MC Congeners Simulation To analyse the IFPs within MC congeners MD simulation, different types of visualisation were generated to visualise the number of differences, the size of aggregation, which IFPs are identical and also the distribution of number of differences across all IFPs. Some representative visualisations have been selected and are shown in Figure 5.18, while the other visualisations can be found in SI Figure 8.42.

The new visualisation as well as their combination with existing visualisations (i.e., the similarity matrix) aid in understanding a data set and the temporal evolution of IFPs better, especially in cases where we have as many simulations as we have here. Even though the visualisations help to understand the data better, it is still difficult to describe and evaluate such a large data set. Therefore, the following paragraphs will focus on the general trends that we can observe in the data.

The visualisation of the occurrence of individual IFPs help to estimate how many IFPs have been grouped into one representative IFP and indicates which IFPs occur frequently, and which do not. This information can aid in identifying representative IFPs within a certain area. It also provides a quick overview of the distribution along the IFPs over time. For the data presented here, we can see that for some MC congeners many frequently occurring IFPs are located within the same area, as e. g. MC-LR in the middle, MC-RR, [Anda5]MC-LY(Prg), [Amba5]MC-LY(Prg), [Apha5]MC-LF and [Apda5]MC-LF towards the end. For all other MC congeners (e. g. [Enantio-Adda5]MC-LF) the occurrence is rather distributed over the trajectory and therefore many representative structures in different areas exist.

With the matrix visualisation, we can identify regions of high similarity (dark blue) and regions of lower similarity (yellow). IFPs accumulate changes slowly over time, so there are often locally defined regions of high similarity along the diagonal. Interestingly, for MC-LR we can see a region that is highly similar within a square, but highly dissimilar to all other IFPs in the simulation, which is visible by green-yellow stripes (see Figure 5.18a). In addition, we can easily identify if a state is revisited, as it is the case for e. g. [Enantio-Adda5]MC-LF where a small region, is revisited, which is indicated by a visible dark blue stripe located further towards the end of the simulation (see Figure 5.18c). This finding is also clearly visible in the line plot

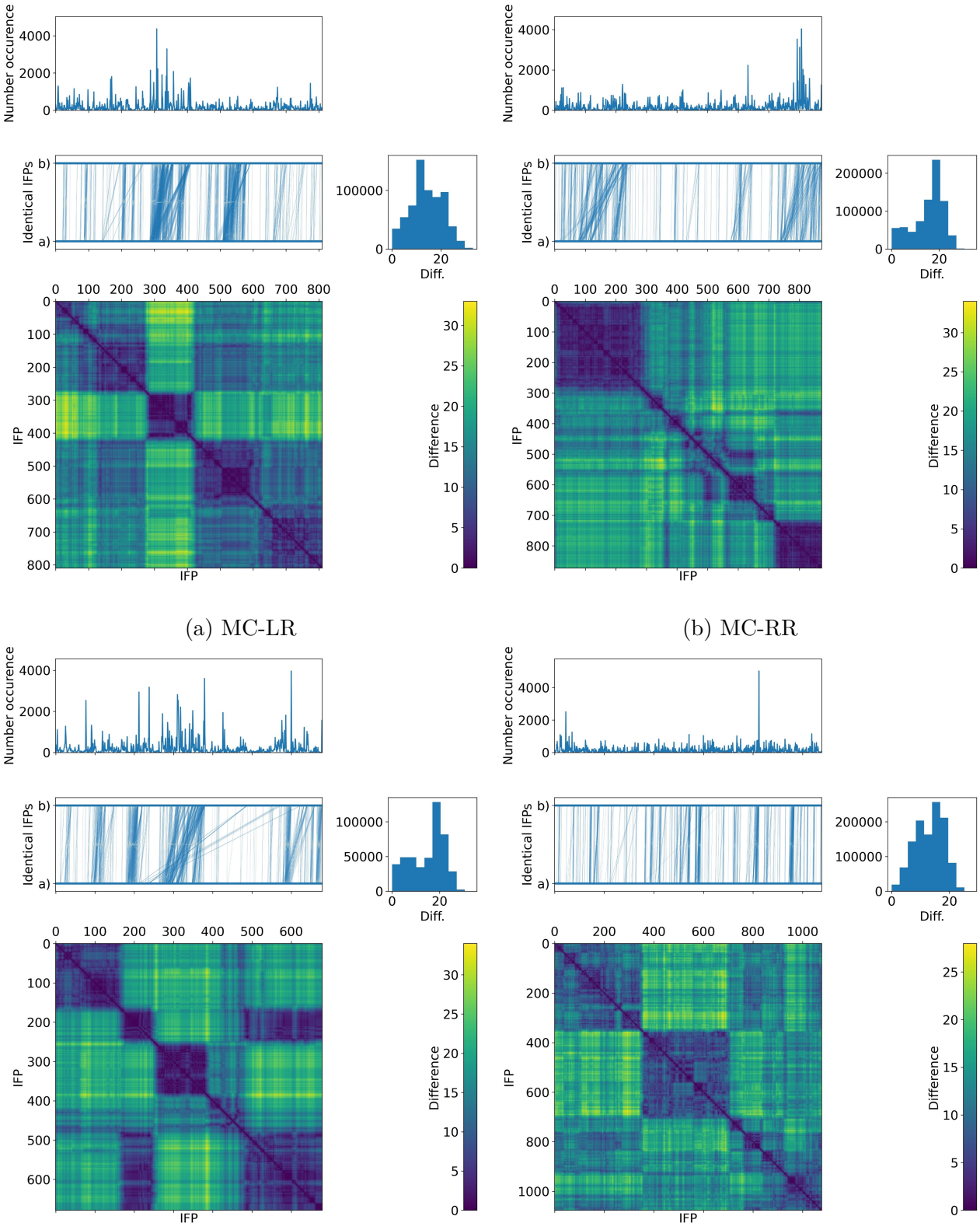
representation, which shows vertical connections for identical frames. Here, we see a diagonal line connecting the two squares, which are highly similar. Many other MC congeners have these well-defined larger squares of similarity, while for others there are superior squares indicating regions of partly lower similarity, divided into smaller squares that look a little blurred at first glance, as e. g. for MC-RR (see Figure 5.18b or SI Figure 8.42g and h). The area of blurred and diffuse squares is often the area that is less connected in the line plot with identical frames and therefore shows less structural similarity.

In summary, the visualisations and data processing presented here helps the user to understand the temporal evolution and occurrence of IFPs within a simulation for a large data set. Even though we have some MC congeners which moved out of the binding site (e. g. MC-LR, MC-RR, [Anda5]MC-LY(Prg), see Section 4.4.3.4), we were able to group the IFPs of the individual simulations into regions of higher similarity. We assume that the small window size and a higher occurrence filter helped to smooth out the really rare interactions during major rearrangements. However, the user is able to change the filtering values and perhaps with smaller values these differences would be easier to detect.

IFP Comparison between MC Congeners Simulation In this section, the IFPs of the different MC congener simulations are compared with each other based on the same procedure as the first data set, which was based on the inverse Rogers-Tanimoto dissimilarity with adjusted thresholds of 0.975 where two IFPs are considered identical, and in between 0.85 and 0.975 they are considered as similar. Again, the two different IFP sets were merged based on the interactions present to generate an IFP with the same length of vectors to be able to calculate similarity.

The proportions of identical, similar and dissimilar values were calculated by occurrence and percentage of IFPs and are summarised visually in Figure 5.19 to provide a quick overview. The individual values are summarised in SI Table 8.14 and will be discussed together with the visualisations of line plots in the following. The line plots to compare similar and identical IFPs between two MC congeners are shown in Figure 5.20 and SI Figure 8.43.

Overall, no dissimilar IFPs could be identified for any comparison of all simulations (see Figure 5.19 and SI Table 8.14), all of them share some similar IFPs and a few share identical IFPs. These findings indicate, that most IFPs result in similarity values between 0.5 and 0.85, which is a different range compared to the previous simulations. The force field and parameterisation of MC congeners for the simulations presented here were different from the first data set. This could lead to slightly different MD simulation trajectories, so the similarity thresholds chosen here are probably not as appropriate and should be adjusted if only this one data set was analysed. In order to be able to compare the data set with the previous simulation set, the thresholds are retained here.



(a) MC-LR

(b) MC-RR

(c) [Enantio-Adda5]MC-LF

(d) [Apha5]MC-LF

Figure 5.18: Comparison of IFP similarity within MC congener. The occurrence and identical IFPs (connected by vertical lines) as line plot. The number of differences to all IFPs are shown as histogram (top right), and as colour on the matrix visualisation.

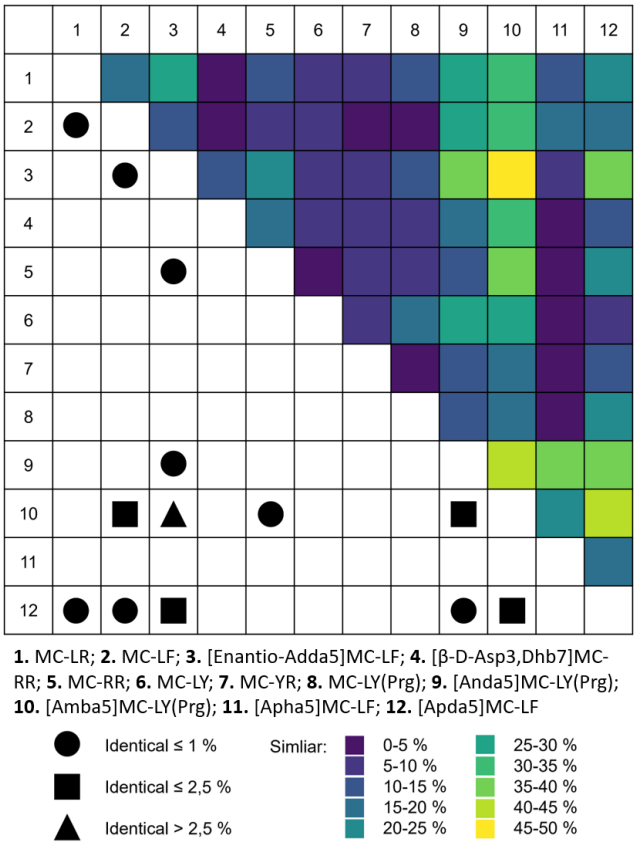


Figure 5.19: Summary of similarity achieved for IFPs of different simulations for the larger MC congener data set. The percentage of similar IFPs is colour coded according to the legend on the plot. Identical values are indicated by a glyph, when present. The upper right of the diagonal represents similar IFPs, the lower left diagonal identical IFPs. No dissimilar IFPs were identified for this data set.

For some simulations, we could observe a higher proportion of identical IFPs compared to previous results of the small simulation set. For the MC congeners where we could find identical IFPs, the proportion is mostly below 1 %, meaning that only a few IFPs are identical to each other, except for [Enantio-Adda5]MC-LF, which has 1.31 % of identical values with [Apda5]-MC-LF and even 9.02 % identical values with [Amba5]-MC-LY(Prg). Also, [Amba5]-MC-LY(Prg) shares 2.5 % and 1.54 % identical IFPs with [Apda5]-MC-LF and [Anda5]-MC-LY(Prg), respectively.
Interestingly, these are all non-toxic MC congeners that seem to share a lot of IFPs i.e., they have identical binding patterns, which seems to be quite unique within this class. In addition, all of those MC congeners have been modified at the Adda side chain, which seems to lead to certain binding patterns. The only MC congener with modified Adda side chain that has different IFPs compared to the ones compared here, is [Apha5]-MC-LF, where we could not identify any MC congener with identical IFPs. Therefore, this MC congener seems to have rather unique binding patterns, which are shared neither across the non-toxic nor the toxic class.

For each of the simulations compared, we can find similar IFPs (see Figure 5.19 and SI Table 8.14). The proportions of similar IFPs identified can vary widely, ranging from 0.32 % to 48.57 %. These findings indicate that some MC congeners have similar binding patterns, while

others are more distinct from each other.

In line with our previous findings, we could observe a higher proportion of IFPs being similar for MC-LR and [Enantio-Adda5]MC-LF, than for MC-LR and MC-LF with 26.8 and 17.46 %, respectively. In addition, the proportion of similar values for [β-D-Asp3,Dhb7]MC-RR, is low for MC-LR (2.59 %), MC-LF (3.44 %). In contrast to previous findings, [Enantio-Adda5]MC-LF and [β-D-Asp3,Dhb7]MC-RR share 15.14 % of similar IFPs, while MC-LR and MC-LF share 17.46 %, which is much higher than for the previous simulation. Since the simulations were generated differently, and some MC congeners were not as stable in the binding site as before (e. g. MC-LR), we probably sampled new interaction patterns that were not present before. In addition, the simulations were twice as long — so we may have been able to visit states of interactions that were simply not visible before due to the shorter simulation time.

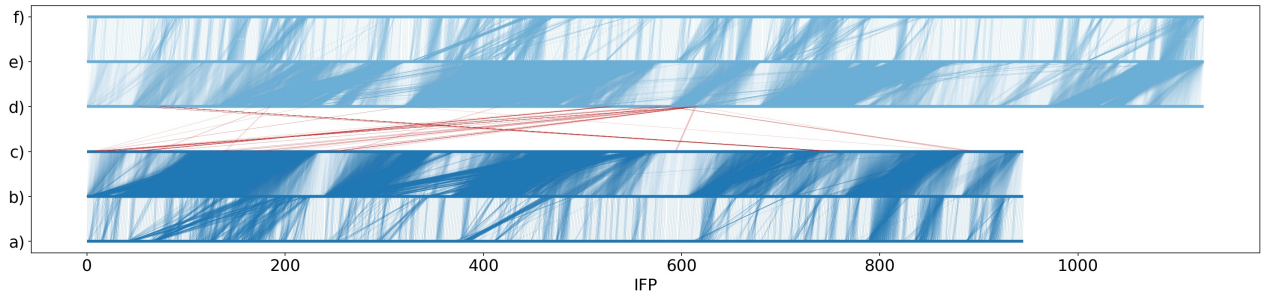
Also, closely related MC congeners with a low number of modifications were found to have a high amount of similar IFPs. For example, MC-LY and MC-LY(Prg) have 16.74 % of IFPs which are similar, and both are in the toxic class. Therefore, it seems that the binding pattern is conserved. Nevertheless, MC-LY shares higher similarity with [Anda5]-MC-LY(Prg) (27.55 %), [Amba5]-MC-LY(Prg) (25.71 %), which are all non-toxic MC congeners. This might partly be due to the identical hyper-variable positions, but maybe also caused by more unspecific patterns, which might be introduced due to diffusion out of the binding site and modification of the Adda residue. [β-D-Asp3,Dhb7]MC-RR and MC-RR which only differ based on two modifications of the backbone, share 15.14 % of similar IFPs but do not show any identical IFPs.

[Amba5]-MC-LY(Prg) shares a lot of similarity across all MC congeners, ranging from 16.04 % to 48.57 %, even though it is highly modified at the Adda residue which is replaced by a methyl group. Therefore, Adda cannot serve as anchor in the binding site and indeed we see [Amba5]-MC-LY(Prg) moving out of the binding site and wandering along the protein during the simulation. Therefore, the high similarity of IFPs might be an artefact of the method. Due to merging the two IFP sets, probably a lot of absent interactions are introduced for [Amba5]-MC-LY(Prg). This may artificially increase the score if this interaction is also absent in the MC congener as the increasing denominator will lead to values closer to 0, and, therefore, inverting the dissimilarity will lead to overall higher similarity. For this reason, the similarity should be reassessed using a different metric to investigate this finding in closer detail. Then, the IFPs should be systematically compared, and probably also the thresholds for the definition of similarity adjusted to improve understanding of the interactions.

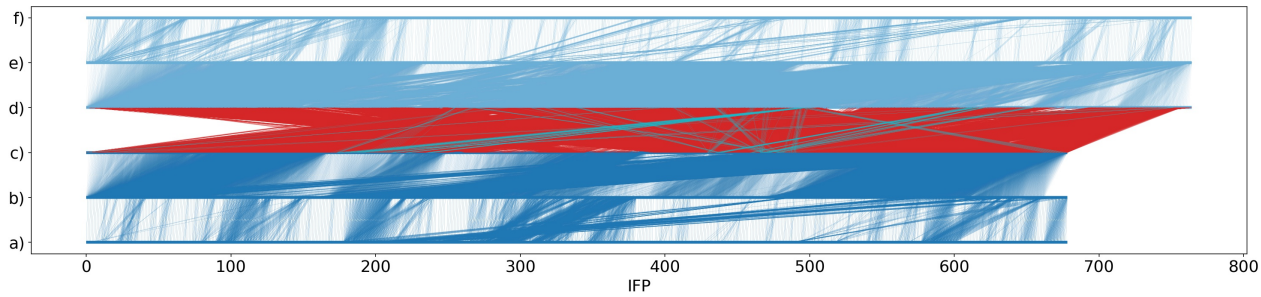
The line plots help to easily identify which regions of the IFPs map to which other regions of the IFP and to get a fast overview on the similarity of interaction patterns. For some interaction we see at first glance that they map to many positions and are rather similar, e. g. for [Amba5]-MC-LY(Prg) and [Enantio-Adda5]MC-LF, for others we can identify MC congeners that share almost no interactions, as e. g. MC-LF and MC-YR. We can also identify MC congeners where small regions in one IFP set map to larger regions located in the other IFP set, as e. g. for [Enantio-Adda5]MC-LF and [Apha5]-MC-LF, and therefore assess differences in the amount of areas with similar IFPs. The mentioned examples are shown in Figure 5.20.

When comparing the IFPs to the previous simulation, most line plots show a similar trend to before, but have more similar areas overall, i.e., more red vertical lines connecting two areas than before. The trends remain the same, suggesting that the method and visualisation are efficient at capturing the overall interaction patterns in the simulation, regardless of the exact procedure (see SI Figure 8.43, not all data is provided). Only for [Enantio-Adda5]MC-LF and [β-D-Asp3,Dhb7]MC-RR the mapping looks very different to the previous simulation, which is

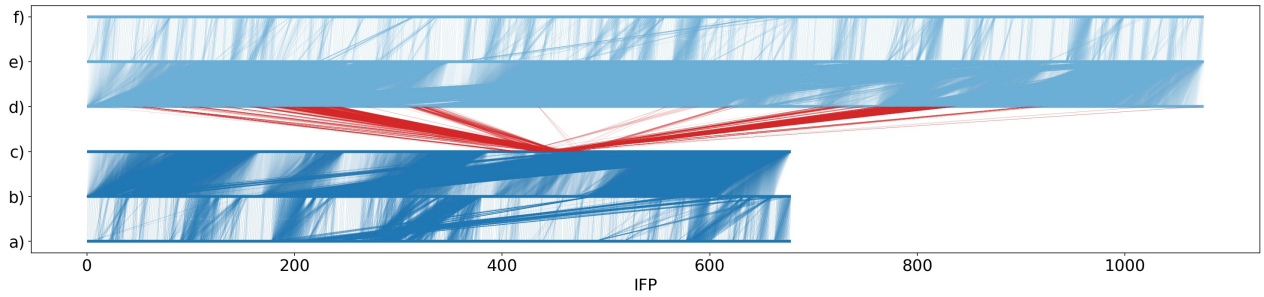
probably because of the higher amount of interactions.



(a) MC-LF & MC-YR



(b) [Enantio-Adda5]& [Amba5]-MC-LY(Prg)



(c) [Enantio-Adda5]MC-LF & [Apha5]-MC-LF

Figure 5.20: Similarity calculation between IFPs of different simulation and networks of major IFPs. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* identical and similar IFPs between simulations are shown as cyan and red lines, respectively.

5.4.5 Conclusion

To conclude, we have clearly shown that the approach developed to aggregate IFPs helps to massively reduce the temporal data set of simulations analysed in a biologically meaningful way. In addition, the visualisations provided give a good overview of the temporal development of IFPs, how often they occur and therefore how important they are likely to be. Therefore, we were able to represent the simulations with a few individual IFPs by deriving representative IFPs on an example data set. Our aggregation method offers the advantage that we can analyse more realistic networks and look at time specific differences, since non-covalent interactions can

form and break.

We were able to derive biological insights and could show that our method is able to reproduce known binding partners from literature, which interestingly never do all occur together in our representative IFPs. In addition, we were able to show that summarising the whole simulation into an $\text{aggregated}_{\text{occ30}}$ IFP is a valid approach for a first overview and quick comparison of the dataset, but suffers from missing interactions that occur frequently in individual IFPs and are therefore likely to be important. The second data set we analysed showed that the thresholds to group IFPs as similar need to be adjusted dependent on the data set and that the method to calculate similarity or dissimilarity is also dependent on the data set and should be evaluated based on the data set.

For our use case, we were able to show that the toxic MC-LR and MC-LF have less similar binding patterns than expected. In addition, we demonstrated that the stereoisomers MC-LF (toxic) and [Enantio-Adda5]MC-LF (non-toxic) do not share many similar IFPs, and that [β-D-Asp3,Dhb7]MC-RR has different interactions compared to the other MC congeners, which is probably caused by changes in the backbone conformation. A more thorough analysis of all the IFPs identified is needed to confirm this finding.

We believe that our approach developed and presented here can be easily adjusted to other systems and will be useful to derive representative IFPs for different use cases, such as understanding binding and interaction, or for machine learning.

In future work, further developments are needed to evaluate the IFPs in greater detail, and additional aggregation methods should be developed to get a better understanding of the temporal development of interactions. In addition, the approach should be extended to compare more than two simulations at the same time.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

This thesis presented new methods and workflows to analyse small molecular data sets, which are a particular challenge in the research fields of bio-and cheminformatics. Due to the lack of data, many methods result in limited performance [7, 8]. The focus of this work has therefore been to combine both areas of research to develop new methods that can be applied with sufficient performance to small data sets.
The new methods and workflows were evaluated by studying the toxicity of Microcystin (MC) congeners as a use case. MC congeners are particularly interesting and difficult to study because they represent a class of closely related macrocyclic structures [7], with different bioactivities, and limited available data [16, 33, 37], making computational modelling and prediction difficult. Most MC congeners are highly toxic to humans, but the current risk assessment is only based on MC-LR, which has been shown not to be the most toxic MC congener [16]. Therefore, this thesis investigated the different aspects related to MC congener toxicity, such as binding and interaction with ser/thr protein phosphatases (PPP), to evaluate MC congener specific differences. The aim was to understand why and how MC congeners have different bioactivities despite being structurally closely related, and to develop methods to aid in understanding and analysis in order to provide methods for an MC congener-specific risk assessment.

The major research questions (RQ) of this thesis were therefore:
[RQ1] What new methods can help to overcome the shortcomings of small molecular data sets?
[RQ2] What new methods can help to understand macrocyclic influence and dynamics and hence bioactivity?
[RQ3] How can we aggregate and compare interaction fingerprints based on Molecular Dynamics simulation?

To answer these RQ, this thesis proposed different new methods and workflows. To investigate RQ1, prediction models based on two-dimensional structures were built and targeted towards small data sets. The first approach combined synthetic minority oversampling techniques with feature representation from natural language processing, to train decisions trees to predict toxicity classes of MC congeners. We could demonstrate that majority voting of three different models resulted in 80-90 % correct prediction of toxicity classes, which is considered a good performance on such a small data set. However, this approach is a black box and lacks explainability of classification decisions, which is important for risk assessment. Consequently,

there has been a need for the development of methods to ensure the explainability of the predictions. For this reason, an optimisation method based on the (α, β)-k-Feature Set Problem has been applied for the first time to molecular data related to toxicity. The optimisation has been directly applied to the extended connectivity fingerprints of MC congeners. We clearly demonstrated that this method was able to derive biologically meaningful substructures related to toxicity and that Boolean rules could subsequently be applied for classification. Of course, with more data available, both approaches would probably lead to improved performance. As no *in-vitro* toxicokinetic or toxicodynamic assays are available for the majority of the 279 MC congeners, this work provided a first step towards MC congener specific toxicity classification and we showed that prediction models can be built on this small molecular data set.

As bioactivity depends on three-dimensional interactions between two molecules, i.e., MC congeners and PPP1, the aim of RQ2 was to develop and evaluate methods that take into account three-dimensional interaction and conformation. Macrocycles have specific structural properties, that allow them to adapt to specific conformations that influence their bioactivity [40, 78, 79]. Therefore, the three-dimensional conformation of MC congeners and their interactions with PPP1 were modelled by Molecular Dynamics (MD) simulation. The aim of the MD simulations was to understand how their conformation might affect their bioactivity and how they interact with PPP1. A new method has been developed to automatically derive interactions between MC congeners and PPP1 from the trajectories of MD simulations. For this approach, two data sets of MD simulations were generated which, for the first time, included not only MC-LR but also other MC congeners and their conjugates. We were able to show that there are MC congener specific backbone conformations, but also identified conformations that are shared by several MC congeners, even in different toxicity classes.

The method developed to automatically derive interactions from MD simulations based on specified interaction criteria resulted in two-dimensional representations of three-dimensional interactions, so-called interaction fingerprints (IFP). This resulted in a huge number of IFPs, leading to RQ3 and the aim to systematically aggregate and compare IFP. We developed a method of aggregation based on filters and sliding windows, and we were able to massively reduce the number of IFPs, while retaining important interactions, which were reported in the literature, and preserving temporal information. We were also able to show that combining and developing visualisations helps to understand interaction patterns, and also to compare them across MC congeners and therefore different simulations. This led to the identification of interaction patterns that were common to all MC congeners, but also to the identification of interaction patterns that were specific to MC congeners.

In summary, this thesis has presented several new approaches and methods to overcome the shortcomings of small molecular data sets by: 1) establishing prediction models tailored to small data sets, 2) by analysing and processing MD simulation data to understand conformation and interactions and 3) by developing a new method to aggregate, visualise and compare IFPs from MD simulations. Although the methods were developed using the data set of MC congener toxicity, they can in principle be applied to other molecular data sets, as the approaches and methods presented are generally applicable to molecular data.

6.2 Future Perspectives

Although this work has introduced several new approaches and methods for studying small molecule data sets and has contributed to the understanding of MC congener-specific conformation and interaction, there are several directions for future research.

The predictive models developed in this thesis could be expanded as more data becomes available. One approach to collect data or make the prediction model available to others would be a web server where researchers could provide data to improve the model, or upload their MC congener structures to get a classification of toxicity. The classification could be accompanied by the Boolean rules derived in this thesis to highlight which substructures might be relevant for which toxicity class, especially if several are present in a molecular structure, to provide guidance to the user.

Another direction for future research would be to investigate further conformations and interactions with MD simulations. So far, the simulations have been limited to PPP1, but MC congeners also inhibit other PPPs, such as PPP2A or PPP5, which were included in our prediction model and for which bioactivity data is available. This is of particular interest because some MC congeners belong to a different toxicity class, depending on the PPP with which they interact [16]. To date, only molecular docking approaches with PPP2A have been described in the literature, which do not take into account conformational flexibility of MC congeners or the dynamics in the interactions [301, 428].

Finally, there are several future research directions for the aggregation and visualisation of IFPs. Although we have massively aggregated the number of IFPs, it is still not feasible to compare all IFPs and their interactions with each other. We started by selecting representative IFPs with high occurrence for comparison within the simulation. To select IFPs for comparison between different simulations, we analysed them for similarity and also selected pairs of IFPs with high occurrence. The selection process for deriving representative IFPs could be improved, since it is currently mostly based on occurrence of individual IFPs. An automated method of identifying higher regions of similarity would be useful to identify areas of conformational difference from which a representative IFP could be identified.

In addition, an approach similar to the graph transitions proposed by Bouguera et al. [429] to combine IFPs into one visualisation would be advantageous to analyse temporal development of IFPs, but also to compare IFPs from different simulations. This would be a good addition to the visualisation approach we developed, which is currently limited to two simulations. Further development of new visualisation methods for improved analysis is needed. In addition, it would be interesting to develop visualisations for distantly related IFPs to highlight differences in interactions.

Furthermore, the here presented IFPs could be also used for machine learning, similar to the approaches of e. g. Szulc et al. [404], Sato et al. [430] or Rodríguez-Pérez et al. [431], to combine the prediction models with molecular dynamics simulation, as information about the three-dimensional interactions might lead to improved performance.

Chapter 7

References

7.1 List of Publications

- **Sabrina Jaeger-Honz**, Martin Kern, Hanna Borlinghaus, Mei-Ling Han, Yan Zhu, Xukai Jiang, Jian Li, Falk Schreiber and Björn Sommer: 'Molecular dynamics simulations of eight *Pseudomonas aeruginosa* membranes aid in understanding resistance mechanism to Polymyxin B1', in preparation.

- **Sabrina Jaeger-Honz**, Raymund Hackett, Regina Fotler, Daniel R. Dietrich and Falk Schreiber: 'Conformation and binding of 12 Microcystin (MC) congeners to PPP1 using molecular dynamics simulations: a potential approach for improved MC risk assessment.' Chemico-Biological interactions 407: 15 pages, 2025. doi:10.1016/j.cbi.2025.111372.

- Pablo Moscato, **Sabrina Jaeger-Honz**, Mohammad N. Haque, and Falk Schreiber: 'The (α,β)-k Boolean signatures of molecular toxicity: microcystin as a case study', bioRxiv, 2024. doi:10.1101/2024.12.29.630644.

- **Sabrina Jaeger-Honz**, Karsten Klein, and Falk Schreiber: 'Systematic analysis, aggregation and visualisation of interaction fingerprints for molecular dynamics simulation data', Journal of Cheminformatics, 16 (28): 15 pages, 2024. doi: 10.1186/s13321-024-00822-3.

- Denis Bienroth, Natalie Charitakis, **Sabrina Jaeger-Honz**, Dimitar Garkov, David A. Elliott, Enzo R. Porrello, Karsten Klein, Hieu T. Nim, Falk Schreiber, Mirana Ramialison: 'Spatially resolved transcriptomics mining in 3D and virtual reality environments with VR-Omics', bioRxiv, 2023. doi: 10.1101/2023.03.31.535025.

- Martin Kern, **Sabrina Jaeger-Honz**, Falk Schreiber and Björn Sommer: 'APL@voro — interactive visualization and analysis of cell membrane simulations', Bioinformatics 39 (2): 3 pages, 2023. doi: 10.1093/bioinformatics/btad083.

- Denis Bienroth, Hieu T. Nim, Dimitar Garkov, Karsten Klein, **Sabrina Jaeger-Honz**, Mirana Ramialison and Falk Schreiber: 'Spatially resolved transcriptomics in immersive environments', Visual Computing for Industry, Biomedicine, and Art 5(2): 13 pages, 2022. doi: 10.1186/s42492-021-00098-6.

- Lucas Joos, **Sabrina Jaeger-Honz**, Falk Schreiber, Daniel A Keim and Karsten Klein: 'Visual Comparison of networks in VR', IEEE Transactions on Visualization and Computer Graphics 28 (11): pp. 3651-3661, 2022. doi: 10.1109/TVCG.2022.3203001.

- **Sabrina Jaeger-Honz**, Jahn Nitschke, Stefan Altaner, Karsten Klein, Daniel R. Dietrich and Falk Schreiber: 'Investigation of microcystin conformation and binding towards PPP1 by molecular dynamics simulation' Chemico-Biological Interactions 351: 15 pages, 2022. doi: 10.1016/j.cbi.2021.109766.

- Karsten Klein, **Sabrina Jaeger**, Jörg Melzheimer, Bettina Wachter, Heribert Hofer, Artur Baltabayev and Falk Schreiber: 'Visual analytics of sensor movement data for cheetah behaviour analysis', Journal of Visualization 24(4): 807–825, 2021. doi: 10.1007/s12650-021-00742-6.

- Stefan Altaner*, **Sabrina Jaeger***, Regina Fotler, Ivan Zemskov, Valentin Wittman, Falk Schreiber and Daniel R. Dietrich: 'Machine learning prediction of cyanobacterial toxin (microcystin) toxicodynamics in humans' Alternatives to Animal Experimentation 37 (1): pp. 24-36, 2020. doi: 10.14573/altex.1904031.

- Karsten Klein, **Sabrina Jaeger**, Jörg Melzheimer, Bettina Wachter, Heribert Hofer, Artur Baltabayev, and Falk Schreiber: 'Visual analytics for cheetah behaviour analysis' in Proceedings of the 12th International Symposium on Visual Information Communication and Interaction (VINCI '19). Association for Computing Machinery, New York, NY, USA, Article 16, 1–8, 2019. doi: 10.1145/3356422.3356435.

- **Sabrina Jaeger**, Karsten Klein, Lucas Joos, Johannes Zagermann, Michael de Ridder, Jinman Kim, Jean Yang, Ulrike Pfeil, Harald Reiterer and Falk Schreiber: 'Challenges for brain data analysis in VR environments' IEEE Pacific Visualization Symposium (PacificVis), Bangkok, Thailand, 2019, pp. 42-46, doi: 10.1109/PacificVis.2019.00013.

7.2 Bibliography

1. Engel, T. Basic overview of chemoinformatics. *J. Chem. Inf. Model.* **46,** 2267–2277. doi:10.1021/ci600234z (2006).
2. Leach, A. & Gillet, V. *An introduction to chemoinformatics* 1st ed. doi:10.1007/978-1-4020-6291-9 (Springer Dordrecht, 2007).
3. Chen, Y. & Kirchmair, J. Cheminformatics in natural product-based drug discovery. *Mol. Inform.* **39,** 2000171. doi:10.1002/minf.202000171 (2020).
4. Wishart, D. S. Introduction to cheminformatics. *Curr. Protoc. Bioinformatics* **18.** doi:10.1002/0471250953.bi1401s18 (2007).
5. Sukumar, N., Krein, M. & Breneman, C. M. Bioinformatics and cheminformatics: where do the twain meet? *Curr. opin. drug discov. dev.* **11,** 311–319. ISSN: 1367-6733 (2008).
6. *Structural bioinformatics* 1st ed. (eds Bourne, P. E. & Weissig, H.) ISBN: 978-0-471-20200-4. doi:10.1002/0471721204 (Wiley, 2003).
7. Driggers, E. M., Hale, S. P., Lee, J. & Terrett, N. K. The exploration of macrocycles for drug discovery — an underexploited structural class. *Nat. Rev. Drug Discov.* **7,** 608–624. doi:10.1038/nrd2590 (2008).

8. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3,** 283–293. doi:10.1021/acscentsci.6b00367 (2017).

9. Zin, P., Williams, G. & Fourches, D. Cheminformatics-based enumeration and analysis of large libraries of macrolide scaffolds. *J. Cheminform.* **10,** 53. doi:10.1186/s13321-018-0307-6 (2018).

10. Yudin, A. K. Macrocycles: lessons from the distant past, recent developments, and future directions. *Chem. Sci.* **6,** 30–49. doi:10.1039/C4SC03089C (2015).

11. Marsault, E. & Peterson, M. Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery. *J. Med. Chem.* **54,** 1961–2004. doi:10.1021/jm1012374 (2011).

12. Ermert, P. Design, properties and recent application of macrocycles in medicinal chemistry. *Chimia* **71,** 678. doi:10.2533/chimia.2017.678 (2017).

13. Hospital, A., Goñi, R., Orozco, M. & Gelpí, J. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.* **10,** 37. doi:10.2147/AABC.S70333 (2015).

14. Medina-Franco, J. L. & Saldívar-González, F. I. Cheminformatics to characterize pharmacologically active natural products. *Biomolecules* **10,** 1566. doi:10.3390/biom10111566 (2020).

15. Bouaïcha, N. *et al.* Structural diversity, characterization and toxicology of microcystins. *Toxins* **11,** 714. doi:10.3390/toxins11120714 (2019).

16. Altaner, S. *et al.* Machine learning prediction of cyanobacterial toxin (microcystin) toxicodynamics in humans. *ALTEX - Alternatives to Animal Experimentation.* doi:10.14573/altex.1904031 (2020).

17. De Figueiredo, D., Azeiteiro, U., Esteves, S., Gonçalves, F. & Pereira, M. Microcystin-producing blooms-a serious global public health issue. *Ecotoxicol. Environ. Saf.* **59,** 151–63. doi:10.1016/j.ecoenv.2004.04.006 (2004).

18. Dietrich, D. & Hoeger, S. Guidance values for microcystins in water and cyanobacterial supplement products (blue-green algal supplements): a reasonable or misguided approach? *Toxicol. Appl. Pharmacol.* **203,** 273–289. doi:10.1016/j.taap.2004.09.005 (2005).

19. Codd, G. *et al.* Cyanobacterial toxins, exposure routes and human health. *Eur. J. Phycol.* **34,** 405–415. doi:10.1080/09670269910001736462 (1999).

20. Li, Y. *et al.* A cross-sectional investigation of chronic exposure to microcystin in relationship to childhood liver damage in the three Gorges Reservoir Region, China. *Environ. Health Perspect.* **119,** 1483–1488. doi:10.1289/ehp.1002412 (2011).

21. Poste, A. E., Hecky, R. E. & Guildford, S. J. Evaluating microcystin exposure risk through fish consumption. *Environ. Sci. Technol.* **45,** 5806–5811. doi:10.1021/es200285c (2011).

22. Wood, S. A. & Dietrich, D. R. Quantitative assessment of aerosolized cyanobacterial toxins at two New Zealand lakes. *J. environ. monit.* **13,** 1617. doi:10.1039/c1em10102a (2011).

23. Pouria, S. *et al.* Fatal microcystin intoxication in haemodialysis unit in Caruaru, Brazil. *Lancet* **352,** 21–26. doi:10.1016/S0140-6736(97)12285-1 (1998).

24. Azevedo, S. M. F. O. *et al.* Human intoxication by microcystins during renal dialysis treatment in Caruaru—Brazil. *Toxicology* **181-182,** 441–446. doi:10.1016/S0300-483X(02)00491-2 (2002).

25. Yuan, M., Carmichael, W. W. & Hilborn, E. D. Microcystin analysis in human sera and liver from human fatalities in Caruaru, Brazil 1996. *Toxicon* **48,** 627–640. doi:10.1016/j.toxicon.2006.07.031 (2006).

26. World Health Organization. *Guidelines for drinking-water quality: fourth edition incorporating first and second addenda* 4th ed + 2st add. ISBN: 978-92-4-004506-4 (World Health Organization, 2022).

27. Fawell, J. K., Mitchell, R. E., Everett, D. J. & Hill, R. E. The toxicity of cyanobacterial toxins in the mouse: I microcystin-LR. *Hum. Exp. Toxicol.* **18,** 162–167. doi:10.1177/096032719901800305 (1999).

28. Fischer, A. *et al.* The role of organic anion transporting polypeptides (OATPs/SLCOs) in the toxicity of different microcystin congeners in vitro: a comparison of primary human hepatocytes and OATP-transfected HEK293 cells. *Toxicol. Appl. Pharmacol.* **245,** 9–20. doi:10.1016/j.taap.2010.02.006 (2010).

29. Fischer, W. *et al.* Organic anion transporting polypeptides expressed in liver and brain mediate uptake of microcystin. *Toxicol. Appl. Pharmacol.* **203,** 257–263. doi:10.1016/j.taap.2004.08.012 (2005).

30. Lu, H. *et al.* Characterization of organic anion transporting polypeptide 1b2-null mice: essential role in hepatic uptake/toxicity of phalloidin and microcystin-LR. *Toxicol. Sci.* **103,** 35–45. doi:10.1093/toxsci/kfn038 (2008).

31. MacKintosh, C., Beattie, K. A., Klumpp, S., Cohen, P. & Codd, G. A. Cyanobacterial microcystin-LR is a potent and specific inhibitor of protein phosphatases 1 and 2A from both mammals and higher plants. *FEBS Lett.* **264,** 187–192. doi:10.1016/0014-5793(90)80245-E (1990).

32. Hastie, C. J., Borthwick, E. B., Morrison, L. F., Codd, G. A. & Cohen, P. T. W. Inhibition of several protein phosphatases by a non-covalently interacting microcystin and a novel cyanobacterial peptide, nostocyclin. *Biochim. Biophys. Acta Gen. Subj.* **1726,** 187–193. doi:10.1016/j.bbagen.2005.06.005 (2005).

33. Hoeger, S. J., Schmid, D., Blom, J. F., Ernst, B. & Dietrich, D. R. Analytical and functional characterization of microcystins [Asp3]MC-RR and [Asp3,Dhb7]MC-RR: consequences for risk assessment? *Environ. Sci. Technol.* **41,** 2609–2616. doi:10.1021/es062681p (2007).

34. Kondo, F. *et al.* Formation, characterization, and toxicity of the glutathione and cysteine conjugates of toxic heptapeptide microcystins. *Chem. Res. Toxicol.* **5,** 591–596. doi:10.1021/tx00029a002 (1992).

35. Zong, W.-S. *et al.* Evaluation of the direct and indirect regulation pathways of glutathione target to the hepatotoxicity of microcystin-LR. *BioMed Res. Int.* **2018,** 5672637. doi:10.1155/2018/5672637 (2018).

36. Kaur, G., Fahrner, R., Wittmann, V., Stieger, B. & Dietrich, D. R. Human MRP2 exports MC-LR but not the glutathione conjugate. *Chem. Biol. Interact.* **311,** 108761. doi:10.1016/j.cbi.2019.108761 (2019).

37. Garibo, D. *et al.* Inhibition equivalency factors for microcystin variants in recombinant and wild-type protein phosphatase 1 and 2A assays. *Environ. Sci. Pollut. Res.* **21,** 10652–10660. doi:10.1007/s11356-014-3065-7 (2014).

38. Ertl, P. & Schuffenhauer, A. *Cheminformatics analysis of natural products: lessons from nature inspiring the design of new drugs* (eds Petersen, F. & Amstutz, R.) 217–235. doi:10.1007/978-3-7643-8595-8_4 (Birkhäuser Basel, Basel, 2008).

39. Lucas, X., Grüning, B. A., Bleher, S. & Günther, S. The purchasable chemical space: a detailed picture. *J. Chem. Inf. Model.* **55,** 915–924. doi:10.1021/acs.jcim.5b00116 (2015).

40. Cummings, M. D. & Sekharan, S. Structure-based macrocycle design in small-molecule drug discovery and simple metrics to identify opportunities for macrocyclization of small-molecule ligands. *J. Med. Chem.* **62,** 6843–6853. doi:10.1021/acs.jmedchem.8b01985 (2019).

41. Kumar, A. & Zhang, K. Y. J. Advances in the development of shape similarity methods and their application in drug discovery. *Front. Chem.* **6,** 315. doi:10.3389/fchem.2018.00315 (2018).

42. Ramachandran, R., Bugbee, K. & Murphy, K. From open data to open science. *Earth Space Sci.* **8.** doi:10.1029/2020EA001562 (2021).

43. Wilson, S. L. *et al.* Sharing biological data: why, when, and how. *FEBS Lett.* **595,** 847–863. doi:10.1002/1873-3468.14067 (2021).

44. Gewin, V. Data sharing: an open mind on open data. *Nature* **529,** 117–119. doi:10.1038/nj7584-117a (2016).

45. Burgelman, J.-C. *et al.* Open science, open data, and open scholarship: european policies to make science fit for the twenty-first century. *Front. Big Data* **2,** 43. doi:10.3389/fdata.2019.00043 (2019).

46. European Organization For Nuclear Research & OpenAIRE. *Zenodo* 2013. doi:10.25495/7GXK-RD71. https://www.zenodo.org/.

47. Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3,** 160018. doi:10.1038/sdata.2016.18 (2016).

48. Jaeger-Honz, S. *et al. Molecular dynamics simulation of MC-congeners in complex with PPP1 - replicate 1* version 1.0 (Zenodo, 2021). doi:10.5281/zenodo.5017745.

49. Jaeger-Honz, S. *et al. Molecular dynamics simulation of MC-congeners in complex with PPP1 - replicate 2* version 1.0 (Zenodo, 2021). doi:10.5281/zenodo.5017839.

50. Jaeger-Honz, S. *et al. Molecular dynamics simulation of MC-congeners in complex with PPP1 - replicate 3* version 1.0 (Zenodo, 2021). doi:10.5281/zenodo.5017851.

51. Jaeger-Honz, S., Klein, K. & Schreiber, F. *Interaction fingerprints for molecular dynamics simulation of MC-LR and MC-LF with PPP1 - Data, scripts and libraries* 2023. doi:10.5281/zenodo.10423389.

52. Jaeger-Honz, S., Klein, K. & Schreiber, F. *Interaction fingerprints for molecular dynamics simulation of MC-LR and MC-LF with PPP1 - Libraries and scripts* 2023. doi:10.5281/zenodo.10424417.

53. Jaeger-Honz, S., Hackett, R., Fotler, R., Dietrich, D. R. & Schreiber, F. *Molecular dynamics simulations of 12 Microcystin congeners in solvent and complex with PPP1* version 0.1 (Zenodo, 2024). doi:10.5281/zenodo.14501700.

54. Hann, M. & Green, R. Chemoinformatics — a new name for an old problem? *Curr. Opin. Chem. Biol.* **3,** 379–383. doi:10.1016/S1367-5931(99)80057-X (1999).

55. Brown, F. K. *Chemoinformatics: what is it and how does it impact drug discovery.* (ed Bristol, J. A.) 375–384. doi:10.1016/S0065-7743(08)61100-8 (Academic Press, 1998).

56. *Chemoinformatics: a textbook* (eds Gasteiger, J. & Engel, T.) 649 pp. ISBN: 978-3-527-30681-7. doi:10.1002/3527601643 (Wiley-VCH, Weinheim, 2003).

57. *Chemoinformatics in drug discovery* 1 ed., reprint (ed Oprea, T. I.) *Methods and principles in medicinal chemistry* **23.** 493 pp. ISBN: 978-3-527-30753-1. doi:10.1002/3527603743 (Wiley-VCH, Weinheim, 2006).

58. Wermuth, C. G., Aldous, D. J., Raboisson, P. & Rognan, D. *The practice of medicinal chemistry* 4th edition. ISBN: 978-0-12-417205-0. doi:10.1016/C2012-0-03066-9 (Elsevier Academic Press, London, 2015).

59. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature* **181,** 662–666. doi:10.1038/181662a0 (1958).

60. Meyer, E. F. Interactive computer display for the three-dimensional study of macromolecular structures. *Nature* **232,** 255–257. doi:10.1038/232255a0 (1971).

61. Berman, H. M. The Protein Data Bank: a historical perspective. *Acta Crystallogr. A Found. Crystallogr.* **64,** 88–95. doi:10.1107/S0108767307035623 (2008).

62. Meyer, E. E. The first years of the Protein Data Bank. *Prot. Sci.* **6,** 1591–1597. ISSN: 0961-8368, 1469-896X. doi:10.1002/pro.5560060724 (July 1997).

63. Berman, H. M., Kleywegt, G. J., Nakamura, H. & Markley, J. L. The Protein Data Bank at 40: reflecting on the past to prepare for the future. *Structure* **20,** 391–396. doi:10.1016/j.str.2012.01.010 (2012).

64. Burley, S. K. *et al.* RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Res.* **51,** D488–D508. doi:10.1093/nar/gkac1077 (2023).

65. Adcock, S. A. & McCammon, J. A. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* **106,** 1589–1615. doi:10.1021/cr040426m (2006).

66. Shukla, R. & Tripathi, T. *Molecular dynamics simulation of protein and protein–ligand complexes* (ed Singh, D. B.) 133–161. ISBN: 9789811568145. doi:10.1007/978-981-15-6815-2_7 (Springer Singapore, Singapore, 2020).

67. Alder, B. J. & Wainwright, T. E. Phase transition for a hard sphere system. *J. Chem. Phys.* **27,** 1208–1209. doi:10.1063/1.1743957 (1957).

68. Rahman, A. Correlations in the motion of atoms in liquid argon. *Phys. Rev.* **136,** A405–A411. doi:10.1103/PhysRev.136.A405 (1964).

69. Rahman, A. & Stillinger, F. H. Molecular dynamics study of liquid water. *J. Chem. Phys.* **55,** 3336–3359. doi:10.1063/1.1676585 (1971).

70. McCammon, J. A., Gelin, B. R. & Karplus, M. Dynamics of folded proteins. *Nature* **267,** 585–590. doi:10.1038/267585a0 (1977).

71. McCammon, J. A., Karim, O. A., Lybrand, T. P. & Wong, C. F. Ionic association in water: from atoms to enzymes. *Ann. N. Y. Acad. Sci.* **482,** 210–221. doi:10.1111/j.1749-6632.1986.tb20952.x (1986).

72. Duan, Y. & Kollman, P. A. Pathways to a protein folding intermediate observed in a 1-μs simulation in aqueous solution. *Science* **282,** 740–744. doi:10.1126/science.282.5389.740 (1998).

73. Freddolino, P. L., Arkhipov, A. S., Larson, S. B., McPherson, A. & Schulten, K. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **14,** 437–449. doi:10.1016/j.str.2005.11.014 (2006).

74. The Royal Swedish Academy of Sciences. *The Nobel Prize in Chemistry 2013* [*press release*] https://www.nobelprize.org/prizes/chemistry/2013/press-release/, accessed 2023-07-14. 2013.

75. Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief. Bioinformatics* **20,** 1981–1996. doi:10.1093/bib/bby063 (2019).

76. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **23,** 3–25. doi:10.1016/S0169-409X(96)00423-1 (1997).

77. Mallinson, J. & Collins, I. Macrocycles in new drug discovery. *Future Med. Chem.* **4,** 1409–1438. doi:10.4155/fmc.12.93 (2012).

78. Doak, B. C., Zheng, J., Dobritzsch, D. & Kihlberg, J. How beyond rule of 5 drugs and clinical candidates bind to their targets. *J. Med. Chem.* **59,** 2312–2327. doi:10.1021/acs.jmedchem.5b01286 (2016).

79. Poongavanam, V. *et al.* Conformational sampling of macrocyclic drugs in different environments: can we find the relevant conformations? *ACS Omega* **3,** 11742–11757. doi:10.1021/acsomega.8b01379 (2018).

80. Raboisson, P. *Macrocycles* (eds Wermuth, C. G., Aldous, D., Raboisson, P. & Rognan, D.) 267–275. ISBN: 978-0-12-417205-0. doi:10.1016/B978-0-12-417205-0.00010-9 (Academic Press, San Diego, 2015).

81. Villar, E. A. *et al.* How proteins bind macrocycles. *Nat. Chem. Biol.* **10,** 723–731. doi:10.1038/nchembio.1584 (2014).

82. Heinis, C. Tools and rules for macrocycles. *Nat. Chem. Biol.* **10,** 696–698. doi:10.1038/nchembio.1605 (2014).

83. Giordanetto, F. & Kihlberg, J. Macrocyclic drugs and clinical candidates: what can medicinal chemists learn from their properties? *J. Med. Chem.* **57,** 278–295. doi:10.1021/jm400887j (2014).

84. Bonnet, P., Agrafiotis, D. K., Zhu, F. & Martin, E. Conformational analysis of macrocycles: finding what common search methods miss. *J. Chem. Inf. Model.* **49,** 2242–2259. doi:10.1021/ci900238a (2009).

85. Gasteiger, J. & Funatsu, K. Chemoinformatics – An important scientific discipline. *J. Comput. Chem., Jpn.* **5,** 53–58. doi:10.2477/jccj.5.53 (2006).

86. David, L., Thakkar, A., Mercado, R. & Engkvist, O. Molecular representations in AI-driven drug discovery: a review and practical guide. *J. Cheminformatics* **12,** 56. doi:10.1186/s13321-020-00460-5 (2020).

87. Mauri, A., Consonni, V. & Todeschini, R. *Molecular descriptors* (eds Leszczynski, J. *et al.*) 2065–2093. doi:10.1007/978-3-319-27282-5_51 (Springer International Publishing, Cham, 2017).

88. Wigh, D. S., Goodman, J. M. & Lapkin, A. A. A review of molecular representation in the age of machine learning. *WIREs Comput. Mol. Sci.* **12.** doi:10.1002/wcms.1603 (2022).

89. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28,** 31–36. doi:10.1021/ci00057a005 (1988).

90. Morgan, H. L. The generation of a unique machine description for chemical structures – a technique developed at chemical abstracts service. *J. chem. doc.* **5,** 107–113. doi:10.1021/c160017a018 (1965).

91. O'Boyle, N. M. Towards a universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *J. Cheminformatics* **4,** 22. doi:10.1186/1758-2946-4-22 (2012).

92. Schneider, N., Sayle, R. A. & Landrum, G. A. Get your atoms in order—An open-source implementation of a novel and robust molecular canonicalization algorithm. *J. Chem. Inf. Model.* **55,** 2111–2120. doi:10.1021/acs.jcim.5b00543 (2015).

93. Daylight Theory. *SMARTS - A language for describing molecular patterns* https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html, accessed 24-07-2023.

94. Riniker, S. & Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *J. Cheminformatics* **5,** 26. doi:10.1186/1758-2946-5-26 (2013).

95. Gao, K. *et al.* Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys.* **22,** 8373–8390. doi:10.1039/D0CP00305K (2020).

96. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50,** 742–754. doi:10.1021/ci100050t (2010).

97. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput.* **29,** 97–101. ISSN: 0095-2338, 1520-5142. doi:10.1021/ci00062a008 (May 1989).

98. O'Boyle, N. M. & Sayle, R. A. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminformatics* **8,** 36. doi:10.1186/s13321-016-0148-0 (2016).

99. Shin, W.-H., Zhu, X., Bures, M. & Kihara, D. Three-dimensional compound comparison methods and their application in drug discovery. *Molecules* **20,** 12841–12862. doi:10.3390/molecules200712841 (2015).

100. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536. doi:10.1038/323533a0 (1986).

101. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. *arXiv.* doi:10.48550/arXiv.1301.3781 (2013).

102. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. *arXiv.* doi:10.48550/arXiv.1310.4546 (2013).

103. Asgari, E. & Mofrad, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE* **10,** e0141287. doi:10.1371/journal.pone.0141287 (2015).

104. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58,** 27–35. doi:10.1021/acs.jcim.7b00616 (2018).

105. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42,** D1083–D1090. doi:10.1093/nar/gkt1031 (2014).

106. Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S. & Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **52,** 1757–1768. doi:10.1021/ci3001277 (2012).

107. Leach, A. R. *Molecular modelling: principles and applications* 2nd ed. 744 pp. ISBN: 978-0-582-38210-7 (Prentice Hall, Harlow, England ; New York, 2001).

108. Medina-Franco, J. L., Méndez-Lucio, O. & Martinez-Mayorga, K. *The interplay between molecular modeling and chemoinformatics to characterize protein–ligand and protein–protein interactions landscapes for drug discovery* (ed Karabencheva-Christova, T.) 1–37. doi:10.1016/bs.apcsb.2014.06.001 (Elsevier, 2014).

109. Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft* **27,** 2985–2993 (1894).

110. Koshland, D. E. Correlation of structure and function in enzyme action: theoretical and experimental tools are leading to correlations between enzyme structure and function. *Science* **142,** 1533–1541. doi:10.1126/science.142.3599.1533 (1963).

111. Ma, B., Kumar, S., Tsai, C.-J. & Nussinov, R. Folding funnels and binding mechanisms. *Protein Eng. Des. Sel.* **12,** 713–720. doi:10.1093/protein/12.9.713 (1999).

112. Csermely, P., Palotai, R. & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35,** 539–546. doi:10.1016/j.tibs.2010.04.009 (2010).

113. Meng, X.-Y., Zhang, H.-X., Mezei, M. & Cui, M. Molecular docking: a powerful approach for structure-based drug discovery. *Curr. Comput. Aided Drug Des.* **7,** 146–157. doi:10.2174/157340911795677602 (2011).

114. Brooijmans, N. & Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **32,** 335–373. doi:10.1146/annurev.biophys.32.110601.142532 (2003).

115. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161,** 269–288. doi:10.1016/0022-2836(82)90153-X (1982).

116. Miller, M. D., Kearsley, S. K., Underwood, D. J. & Sheridan, R. P. FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8,** 153–174. doi:10.1007/BF00119865 (1994).

117. Morris, G. M. *et al.* Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **19,** 1639–1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B (1998).

118. Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **261,** 470–489. doi:10.1006/jmbi.1996.0477 (1996).

119. Agnihotry, S., Pathak, R. K., Srivastav, A., Shukla, P. K. & Gautam, B. *Molecular docking and structure-based drug design* (ed Singh, D. B.) 115–131. doi:10.1007/978-981-15-6815-2_6 (Springer Singapore, Singapore, 2020).

120. Ozdemir, E. S., Nussinov, R., Gursoy, A. & Keskin, O. Developments in integrative modeling with dynamical interfaces. *Curr. Opin. Struct. Biol.* **56,** 11–17. doi:10.1016/j.sbi.2018.10.007 (2019).

121. Bender, B. J. *et al.* A practical guide to large-scale docking. *Nat. Protoc.* **16,** 4799–4832. doi:10.1038/s41596-021-00597-z (2021).

122. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta. Crystallogr. B. Struct. Sci. Cryst. Eng. Mater.* **72,** 171–179. doi:10.1107/S2052520616003954 (2016).

123. Spyrakis, F. & Cavasotto, C. N. Open challenges in structure-based virtual screening: receptor modeling, target flexibility consideration and active site water molecules description. *Arch. Biochem. Biophys.* **583,** 105–119. doi:10.1016/j.abb.2015.08.002 (2015).

124. Varadi, M. *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50,** D439–D444. doi:10.1093/nar/gkab1061 (2022).

125. Callaway, E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature* **588,** 203–204. doi:10.1038/d41586-020-03348-4 (2020).

126. Scardino, V., Di Filippo, J. I. & Cavasotto, C. N. How good are AlphaFold models for docking-based virtual screening? *iScience* **26,** 105920. doi:10.1016/j.isci.2022.105920 (2023).

127. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49,** D1388–D1395. doi:10.1093/nar/gkaa971 (D1 2021).

128. Landrum, G. *RDKit: open-source cheminformatics software* https://github.com/rdkit/rdkit/, Release 2021.03.5. 2021.

129. O'Boyle, N. M. *et al.* Open Babel: an open chemical toolbox. *J. Cheminformatics* **3,** 33. doi:10.1186/1758-2946-3-33 (2011).

130. Hanwell, M. D. *et al.* Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics* **4,** 17. doi:10.1186/1758-2946-4-17 (2012).

131. Petukh, M., Stefl, S. & Alexov, E. The role of protonation states in ligand-receptor recognition and binding. *Curr. Pharm. Des.* **19,** 4182–4190. doi:10.2174/1381612811319230004 (2013).

132. Li, J., Fu, A. & Zhang, L. An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdiscip. Sci. Comput. Life Sci.* **11,** 320–328. doi:10.1007/s12539-019-00327-w (2019).

133. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9,** 646–652. doi:10.1038/nsb0902-646 (2002).

134. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. Protein folding kinetics and thermodynamics from atomistic simulation. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 17845–17850. doi:10.1073/pnas.1201811109 (2012).

135. Bekker, H., Dijkstra, E. J., Renardus, M. K. R. & Berendsen, H. J. C. An efficient, box shape independent non-bonded force and virial algorithm for molecular dynamics. *Mol. Simul.* **14,** 137–151. doi:10.1080/08927029508022012 (1995).

136. Allen, M. P. & Tildesley, D. J. *Introduction* doi:10.1093/oso/9780198803195.003.0001 (Oxford University Press, 2017).

137. Braun, E. *et al.* Best practices for foundations in molecular simulations. *LiveCoMS* **1.** doi:10.33011/livecoms.1.1.5957 (2019).

138. Karplus, M. & Kuriyan, J. Molecular dynamics and protein function. *Proc. Natl. Acad. Sci. U.S.A* **102,** 6679–6685. doi:10.1073/pnas.0408930102 (2005).

139. Born, M. & Oppenheimer, R. Zur Quantentheorie der Molekülen. *Ann. Phys.* **389,** 457–484. doi:10.1002/andp.19273892002 (1927).

140. Lubich, C. *From quantum to classical molecular dynamics: reduced models and numerical analysis* 1st ed. ISBN: 978-3-03719-067-8. doi:10.4171/067 (EMS Press, Sept. 2008).

141. González, M. Force fields and molecular dynamics simulations. *École thématique de la Société Française de la Neutronique* **12,** 169–200. doi:10.1051/sfn/201112009 (2011).

142. Norberg, J. & Nilsson, L. On the truncation of long-range electrostatic interactions in DNA. *Biophys. J.* **79,** 1537–1553. doi:10.1016/S0006-3495(00)76405-8 (2000).

143. Gilson, M. K. Theory of electrostatic interactions in macromolecules. *Curr. Opin. Struct. Biol.* **5,** 216–223. doi:10.1016/0959-440X(95)80079-4 (1995).

144. Steinbach, P. J. & Brooks, B. R. New spherical-cutoff methods for long-range forces in macromolecular simulation. *J. Comput. Chem.* **15,** 667–683. doi:10.1002/jcc.540150702 (1994).

145. Sagui, C. & Darden, T. A. Molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **28,** 155–179. doi:10.1146/annurev.biophys.28.1.155 (1999).

146. Yeh, I.-C. & Berkowitz, M. L. Ewald summation for systems with slab geometry. *J. Chem. Phys.* **111,** 3155–3162. doi:10.1063/1.479595 (1999).

147. Allen, M. P. & Tildesley, D. J. *Long-range forces* ISBN: 9780198803195. doi:10.1093/oso/9780198803195.003.0006 (Oxford University Press, 2017).

148. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **25,** 1157–1174. doi:10.1002/jcc.20035 (2004).

149. Maier, J. A. *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J. Chem. Theory Comput.* **11,** 3696–3713. doi:10.1021/acs.jctc.5b00255 (2015).

150. Scott, W. R. P. *et al.* The GROMOS biomolecular simulation program package. *J. Phys. Chem. A* **103,** 3596–3607. doi:10.1021/jp984217f (1999).

151. Brooks, B. R. *et al.* CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **30,** 1545–1614. doi:10.1002/jcc.21287 (2009).

152. Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* **118,** 11225–11236. doi:10.1021/ja9621760 (1996).

153. Malde, A. K. *et al.* An Automated Force Field Topology Builder (ATB) and repository: version 1.0. *J. Chem. Theory Comput.* **7,** 4026–4037. doi:10.1021/ct200196m (2011).

154. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.,* 671–690. doi:10.1002/jcc.21367 (2009).

155. Zoete, V., Cuendet, M. A., Grosdidier, A. & Michielin, O. SwissParam: A fast force field generation tool for small organic molecules. *J. Comput. Chem.* **32,** 2359–2368. doi:10.1002/jcc.21816 (2011).

156. Jo, S., Kim, T., Iyer, V. G. & Im, W. CHARMM-GUI: A web-based graphical user interface for CHARMM. *J. Comput. Chem.* **29,** 1859–1865. doi:10.1002/jcc.20945 (2008).

157. Allen, M. P. & Tildesley, D. J. *Molecular dynamics* ISBN: 9780198803195. doi:10.1093/oso/9780198803195.003.0003 (Oxford University Press, 2017).

158. Verlet, L. Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159,** 98–103. doi:10.1103/PhysRev.159.98 (1967).

159. Skeel, R. D. What makes molecular dynamics work? *SIAM J. Sci. Comput.* **31,** 1363–1378. doi:10.1137/070683660 (2009).

160. Yu, I., Feig, M. & Sugita, Y. High-performance data analysis on the big trajectory data of cellular scale all-atom molecular dynamics simulations. *J. Phys.: Conf. Ser.* **1036,** 012009. doi:10.1088/1742-6596/1036/1/012009 (2018).

161. Dobson, C. M. Chemical space and biology. *Nature* **432,** 824–828. doi:10.1038/nature03192 (2004).

162. Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **41,** 233–245. doi:10.1021/ci0001482 (2001).

163. Wold, S., Esbensen, K. & Geladi, P. Principal component analysis. *Chemometr. Intell. Lab. Syst.* **2,** 37–52. doi:10.1016/0169-7439(87)80084-9 (1987).

164. Kohonen, T. The self-organizing map. *Proc. IEEE* **78,** 1464–1480. doi:10.1109/5.58325 (1990).

165. Kruskal, J. B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29,** 115–129. doi:10.1007/BF02289694 (1964).

166. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9.** http://jmlr.org/papers/v9/vandermaaten08a.html (2008).

167. Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Comm.* **1,** 30. doi:10.1039/c0md00020e (2010).

168. Guyon, I. M. & Elisseeff, A. An introduction to Variable and feature selection. *J. Mach. Learn. Res.* **3,** 1157–1182. ISSN: 1532-4435 (2003).

169. Wang, S., Tang, J. & Liu, H. in *Encyclopedia of machine learning and data mining* (eds Sammut, C. & Webb, G. I.) 503–511 (Springer US, Boston, MA, 2017). doi:10.1007/978-1-4899-7687-1_101.

170. Smith, L. I. *A tutorial on principal components analysis* in (2002). https://api.semanticscholar.org/CorpusID:60161425.

171. Cihan Sorkun, M., Mullaj, D., Koelman, J. M. V. A. & Er, S. ChemPlot, a python library for chemical space visualization. *Chem. Methods* **2.** doi:10.1002/cmtd.202200005 (2022).

172. García, A. E. Large-amplitude nonlinear motions in proteins. *Phys. Rev. Lett.* **68,** 2696–2699. doi:10.1103/PhysRevLett.68.2696 (1992).

173. Amadei, A., Linssen, A. B. M. & Berendsen, H. J. C. Essential dynamics of proteins. *Proteins* **17,** 412–425. doi:10.1002/prot.340170408 (1993).

174. Palma, J. & Pierdominici-Sottile, G. On the uses of PCA to characterise molecular dynamics simulations of biological macromolecules: basics and tips for an effective use. *ChemPhysChem* **24.** doi:10.1002/cphc.202200491 (2023).

175. Cotta, C., Sloper, C. & Moscato, P. *Evolutionary search of thresholds for robust feature set selection: application to the analysis of microarray data* in *Applications of Evolutionary Computing, EvoWorkshops* (eds Raidl, G. R. *et al.*) **3005** (Springer, 2004), 21–30. doi:10.1007/978-3-540-24653-4_3.

176. Moscato, P., Berretta, R., Hourani, M., Mendes, A. & Cotta, C. *Genes related with alzheimer's disease: a comparison of evolutionary search, statistical and integer programming approaches* in *Applications of evolutionary computing* (eds Rothlauf, F. *et al.*) (Springer Berlin Heidelberg, 2005), 84–94. doi:10.1007/978-3-540-32003-6_9.

177. Berretta, R., Mendes, A. & Moscato, P. Selection of discriminative genes in microarray experiments using mathematical programming. *J. Res. Pract. Inf. Technol.* **39,** 287–299. doi:10.3316/ielapa.938056748782756 (2007).

178. Duvenaud, D. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *arXiv.* doi:10.48550/arXiv.1509.09292 (2015).

179. Lusci, A., Pollastri, G. & Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* **53,** 1563–1575. doi:10.1021/ci400187y (2013).

180. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* **3,** 80. doi:10.3389/fenvs.2015.00080 (2016).

181. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *J. Chem. Inf. Model.* **57,** 1757–1772. doi:10.1021/acs.jcim.6b00601 (2017).

182. Kotsiantis, S. B. *Supervised machine learning: a review of classification techniques* in *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies* (IOS Press, NLD, 2007), 3–24. ISBN: 9781586037802.

183. Marzban, C. The ROC Curve and the Area under It as performance measures. *Weather Forecasting* **19,** 1106–1114. doi:10.1175/825.1 (2004).

184. Friedman, J. H. Regularized discriminant analysis. *J. Am. Stat. Assoc.* **84,** 165–175. doi:10.1080/01621459.1989.10478752 (1989).

185. Vapnik, V. N. *The Nature of Statistical Learning Theory* ISBN: 978-1-4757-2442-4. doi:10.1007/978-1-4757-2440-0 (Springer New York, New York, NY, 1995).

186. Murthy, S. K. Automatic construction of decision trees from data: a multi-disciplinary survey. *Data Min. Knowl. Discov.* **2,** 345–389. doi:10.1023/A:1009744630224 (1998).

187. Breiman, L. Random forests. *Mach. Learn.* **45,** 5–32. doi:10.1023/A:1010933404324 (2001).

188. Chen, T. & Guestrin, C. *XGBoost: a scalable tree boosting system* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA, 2016), 785–794. doi:10.1145/2939672.2939785.

189. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **29,** 1189–1232. doi:10.1214/aos/1013203451 (2001).

190. Mez, K. *et al.* Identification of a microcystin in benthic cyanobacteria linked to cattle deaths on alpine pastures in Switzerland. *Eur. J. Phycol.* **32,** 111–117. doi:10.1080/09670269710001737029 (1997).

191. Frazier, K., Colvin, B., Styer, E., Hullinger, G. & Garcia, R. Microcystin toxicosis in cattle due to overgrowth of blue-green algae. *Vet. hum. toxicol.* **40,** 23–24 (1998).

192. Puschner, B. *et al.* Blue-green algae toxicosis in cattle. *J. Am. Vet. Med. Assoc.* **213,** 1605–1607 (1998).

193. Wood, S. A. *et al.* Identification of a benthic microcystin-producing filamentous cyanobacterium (Oscillatoriales) associated with a dog poisoning in New Zealand. *Toxicon* **55,** 897–903. doi:10.1016/j.toxicon.2009.12.019 (2010).

194. Van der Merwe, D. *et al.* Investigation of a *Microcystis aeruginosa* cyanobacterial freshwater harmful algal bloom associated with acute microcystin toxicosis in a dog. *J. Vet. Diagn. Invest.* **24,** 679–687. doi:10.1177/1040638712445768 (2012).

195. Hooser, S. B., Beasley, V. R., Lovell, R. A., Carmichael, W. W. & Haschek, W. M. Toxicity of microcystin LR, a cyclic heptapeptide hepatotoxin from Microcystis aeruginosa, to rats and mice. *Vet. Pathol.* **26,** 246–252. doi:10.1177/030098588902600309 (1989).

196. Falconer, I. R., Burch, M. D., Steffensen, D. A., Choice, M. & Coverdale, O. R. Toxicity of the blue-green alga (cyanobacterium) Microcystis aeruginosa in drinking water to growing pigs, as an animal model for human injury and risk assessment. *Environ. Toxicol. Water Qual.* **9,** 131–139. doi:10.1002/tox.2530090209 (1994).

197. Da Gloria Lima Cruz Teixeira, M. *et al.* Gastroenteritis epidemic in the area of the Itaparica Dam, Bahia, Brazil. *Bull. Pan. Am. Health Organ.* **27,** 244–253. ISSN: 0085-4638 (1993).

198. Grosse, Y. *et al.* Carcinogenicity of nitrate, nitrite, and cyanobacterial peptide toxins. *Lancet Oncol.* **7,** 628–629. doi:10.1016/S1470-2045(06)70789-6 (2006).

199. Amtliche Mitteilungen. Empfehlung zum Schutz von Badenden vor Cyanobakterien-Toxinen. *Bundesgesundheitsblatt - Gesundheitsforsch. - Gesundheitsschutz* **58,** 908–920. doi:10.1007/s00103-015-2192-8 (2015).

200. Zemskov, I., Altaner, S., Dietrich, D. R. & Wittmann, V. Total synthesis of microcystin-LF and derivatives thereof. *J. Org. Chem.* **82,** 3680–3691. doi:10.1021/acs.joc.7b00175 (2017).

201. Feurstein, D., Stemmer, K., Kleinteich, J., Speicher, T. & Dietrich, D. R. Microcystin congener– and concentration-dependent induction of murine neuron apoptosis and neurite degeneration. *Toxicol. Sci.* **124,** 424–431. doi:10.1093/toxsci/kfr243 (2011).

202. PerkinElmer. *ChemDraw JS* https://chemdrawdirect.perkinelmer.cloud/js/sample/index.html. Version 19.0.0. 2023.

203. Rinehart, K. L. *et al.* Nodularin, microcystin, and the configuration of Adda. *J. Am. Chem. Soc.* **110,** 8557–8558. doi:10.1021/ja00233a049 (1988).

204. Carmichael, W. *et al.* Naming of cyclic heptapeptide toxins of cyanobacteria (blue-green algae). *Toxicon* **26,** 971–973. doi:10.1016/0041-0101(88)90195-X (1988).

205. Runnegar, M., Berndt, N. & Kaplowitz, N. Microcystin uptake and inhibition of protein phosphatases: effects of chemoprotectants and self-inhibition in relation to known hepatic transporters. *Toxicol. Appl. Pharmacol.* **134,** 264–272. doi:10.1006/taap.1995.1192 (1995).

206. Feurstein, D., Holst, K., Fischer, A. & Dietrich, D. OATP-associated uptake and toxicity of microcystins in primary murine whole brain cells. *Toxicol. Appl. Pharmacol.* **234,** 247–255. doi:10.1016/j.taap.2008.10.011 (2009).

207. Greer, B., Meneely, J. P. & Elliott, C. T. Uptake and accumulation of Microcystin-LR based on exposure through drinking water: an animal model assessing the human health risk. *Sci. Rep.* **8,** 4913. doi:10.1038/s41598-018-23312-7 (2018).

208. Robinson, N. A., Pace, J. G., Matson, C. F., Miura, G. A. & Lawrence, W. B. Tissue distribution, excretion and hepatic biotransformation of microcystin-LR in mice. *J. Pharmacol. Exp. Ther.* **256,** 176–182. ISSN: 0022-3565 (1991).

209. Swingle, M., Ni, L. & Honkanen, R. E. *Small-molecule inhibitors of ser/thr protein phosphatases: specificity, use and common forms of abuse* 23–38. doi:10.1385/1-59745-267-X:23 (Humana Press, New Jersey, 2007).

210. Zong, W., Wang, Q., Zhang, S., Teng, Y. & Du, Y. Regulation on the toxicity of microcystin-LR target to protein phosphatase 1 by biotransformation pathway: effectiveness and mechanism. *Environ. Sci. Pollut. Res.* **25,** 26020–26029. doi:10.1007/s11356-018-2676-9 (2018).

211. MacKintosh, R. W. *et al.* The cyanobacterial toxin microcystin binds covalently to cysteine-273 on protein phosphatase 1. *FEBS Lett.* **371,** 236–240. doi:10.1016/0014-5793(95)00888-G (1995).

212. Goldberg, J. *et al.* Three-dimensional structure of the catalytic subunit of protein serine/threonine phosphatase-1. *Nature* **376,** 745–753. doi:10.1038/376745a0 (1995).

213. Mattila, K., Annila, A. & Rantala, T. T. *Metals ions mediate the binding of cyanobacterial toxins to human protein phosphatase 1 - a computational study Acta Universitatis Ouluensis A* **351** (University of Oulu, 2000).

214. Gulledge, B., Aggen, J., Eng, H., Sweimeh, K. & Chamberlin, A. Microcystin analogues comprised only of Adda and a single additional amino acid retain moderate activity as PP1/PP2A inhibitors. *Bioorganic Med. Chem. Lett.* **13,** 2907–11. doi:10.1016/S0960-894X(03)00588-2 (2003).

215. Craig, M. *et al.* Molecular mechanisms underlying the interaction of motuporin and microcystins with type-1 and type-2A protein phosphatases. *Biochem. Cell Biol.* **74,** 569–578. doi:10.1139/o96-061 (1996).

216. Fontanillo, M. & Köhn, M. Microcystins: synthesis and structure–activity relationship studies toward PP1 and PP2A. *Bioorg. Med. Chem.* **26,** 1118–1126. doi:10.1016/j.bmc.2017.08.040 (2018).

217. Junker, B. H., Klukas, C. & Schreiber, F. VANTED: a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinform.* **7,** 109. doi:10.1186/1471-2105-7-109 (2006).

218. Rohn, H. *et al.* VANTED v2: a framework for systems biology applications. *BMC Syst. Biol.* **6,** 139. doi:10.1186/1752-0509-6-139 (2012).

219. Hunter, T. Protein kinases and phosphatases: the Yin and Yang of protein phosphorylation and signaling. *Cell* **80,** 225–236. doi:10.1016/0092-8674(95)90405-0 (1995).

220. Day, E. K., Sosale, N. G. & Lazzara, M. J. Cell signaling regulation by protein phosphorylation: a multivariate, heterogeneous, and context-dependent process. *Curr. Opin. Biotechnol.* **40,** 185–192. doi:10.1016/j.copbio.2016.06.005 (2016).

221. De Munter, S., Köhn, M. & Bollen, M. Challenges and opportunities in the development of protein phosphatase-directed therapeutics. *ACS Chem. Biol.* **8,** 36–45. doi:10.1021/cb300597g (2013).

222. Olsen, J. V. *et al.* Global, *in vivo*, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127,** 635–648. doi:10.1016/j.cell.2006.09.026 (2006).

223. Hardman, G. *et al.* Strong anion exchange-mediated phosphoproteomics reveals extensive human non-canonical phosphorylation. *EMBO J.* **38,** e100847. doi:10.15252/embj.2018100847 (2019).

224. Leijten, N. M., Heck, A. J. R. & Lemeer, S. Histidine phosphorylation in human cells; a needle or phantom in the haystack? *Nat. Methods* **19,** 827–828. doi:10.1038/s41592-022-01524-0 (2022).

225. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934. doi:10.1126/science.1075762 (2002).

226. Swingle, M. R. & Honkanen, R. E. Inhibitors of serine/threonine protein phosphatases: biochemical and structural studies provide insight for further development. *Curr. Med. Chem.* **26,** 2634–2660. doi:10.2174/0929867325666180508095242 (2019).

227. Shi, Y. Serine/threonine phosphatases: mechanism through structure. *Cell* **139,** 468–484. doi:10.1016/j.cell.2009.10.006 (2009).

228. Zhang, M., Yogesha, S. D., Mayfield, J. E., Gill, G. N. & Zhang, Y. Viewing serine/threonine protein phosphatases through the eyes of drug designers. *FEBS J.* **280,** 4739–4760. doi:10.1111/febs.12481 (2013).

229. Falconer, I. R. & Yeung, D. S. Cytoskeletal changes in hepatocytes induced by Microcystis toxins and their relation to hyperphosphorylation of cell proteins. *Chem. Biol. Interact.* **81,** 181–196. doi:10.1016/0009-2797(92)90033-h (1992).

230. Ohta, T. *et al.* Hyperphosphorylation of cytokeratins 8 and 18 by microcystin-LR, a new liver tumor promoter, in primary cultured rat hepatocytes. *Carcinogenesis* **13,** 2443–2447. doi:10.1093/carcin/13.12.2443 (1992).

231. Wickstrom, M. L. *et al.* Alterations in microtubules, intermediate filaments, and microfilaments induced by microcystin-LR in cultured cells. *Toxicol. Pathol.* **23,** 326–337. doi:10.1177/019262339502300309 (1995).

232. Pereira, S., Vasconcelos, V. & Antunes, A. The phosphoprotein phosphatase family of ser/thr phosphatases as principal targets of naturally occuring toxins. *Crit. Rev. Toxicol.* **41,** 83–110. doi:10.3109/10408444.2010.515564 (2011).

233. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30,** 595–608. doi:10.1007/s10822-016-9938-8 (2016).

234. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv.* doi:10.48550/arXiv.1802.04364 (2018).

235. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminformatics* **5,** 7. doi:10.1186/1758-2946-5-7 (2013).

236. Kier, L. B. A shape index from molecular graphs. *Quant. Struct.-Act. Relat.* **4,** 109–116. doi:10.1002/qsar.19850040303 (1985).

237. Joback, K. G. & Reid, R. C. Estimation of pure-component properties from group-contributions. *Chem. Eng. Commun.* **57,** 233–243. doi:10.1080/00986448708960487 (1987).

238. Marrero, J. & Gani, R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilib.* **183-184,** 183–208. doi:10.1016/S0378-3812(01)00431-9 (2001).

239. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep neural nets as a method for quantitative structure–activity relationships. *J. Chem. Inf. Model.* **55,** 263–274. doi:10.1021/ci500747n (2015).

240. Chakravarti, S. K. Distributed representation of chemical fragments. *ACS Omega* **3,** 2825–2836. doi:10.1021/acsomega.7b02045 (2018).

241. Feinberg, E. N., Joshi, E., Pande, V. S. & Cheng, A. C. Improvement in ADMET prediction with multitask deep featurization. *J. Med. Chem.* **63,** 8835–8848. doi:10.1021/acs.jmedchem.9b02187 (2020).

242. Shao, J. *et al.* S2DV: converting SMILES to a drug vector for predicting the activity of anti-HBV small molecules. *Brief. Bioinformatics.* doi:10.1093/bib/bbab593 (2022).

243. Alshehri, A. S., Gani, R. & You, F. Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: State-of-the-art and future directions. *Comput. Chem. Eng.* **141,** 107005. doi:10.1016/j.compchemeng.2020.107005 (2020).

244. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *arXiv.* doi:10.48550/arXiv.1703.00564 (2018).

245. Chang, N.-B., Vannah, B. & Jeffrey Yang, Y. Comparative sensor fusion between hyperspectral and multispectral satellite sensors for monitoring microcystin distribution in Lake Erie. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7,** 2426–2442. doi:10.1109/JSTARS.2014.2329913 (2014).

246. Mishra, S. *et al.* Evaluation of a satellite-based cyanobacteria bloom detection algorithm using field-measured microcystin data. *Sci. Total Environ.* **774,** 145462. doi:10.1016/j.scitotenv.2021.145462 (2021).

247. Jiang, P., Liu, X., Zhang, J. & Yuan, X. A framework based on hidden Markov model with adaptive weighting for microcystin forecasting and early-warning. *Decis. Support Syst.* **84,** 89–103. doi:10.1016/j.dss.2016.02.003 (2016).

248. Taranu, Z. E., Gregory-Eaves, I., Steele, R. J., Beaulieu, M. & Legendre, P. Predicting microcystin concentrations in lakes and reservoirs at a continental scale: a new framework for modelling an important health risk factor. *Glob. Ecol. Biogeogr.* **26,** 625–637. doi:10.1111/geb.12569 (2017).

249. Francy, D. S. *et al.* Predicting microcystin concentration action-level exceedances resulting from cyanobacterial blooms in selected lake sites in Ohio. *Environ. Monit. Assess.* **192,** 513. doi:10.1007/s10661-020-08407-x (2020).

250. Shan, K. *et al.* Application of Bayesian network including Microcystis morphospecies for microcystin risk assessment in three cyanobacterial bloom-plagued lakes, China. *Harmful Algae* **83,** 14–24. doi:10.1016/j.hal.2019.01.005 (2019).

251. Yuan, L. L. & Pollard, A. I. Combining national and state data improves predictions of microcystin concentration. *Harmful Algae* **84,** 75–83. doi:10.1016/j.hal.2019.02.009 (2019).

252. Shan, K. *et al.* Use statistical machine learning to detect nutrient thresholds in Microcystis blooms and microcystin management. *Harmful Algae* **94,** 101807. doi:10.1016/j.hal.2020.101807 (2020).

253. He, X., Wang, H., Zhuang, W., Liang, D. & Ao, Y. Risk prediction of microcystins based on water quality surrogates: a case study in a eutrophicated urban river network. *Environ. Pollut.* **275,** 116651. doi:10.1016/j.envpol.2021.116651 (2021).

254. Christensen, V. G. *et al.* Cyanotoxin mixture models: relating environmental variables and toxin co-occurrence to human exposure risk. *J. Hazard. Mater.* **415,** 125560. doi:10.1016/j.jhazmat.2021.125560 (2021).

255. Recknagel, F., Orr, P. T., Bartkow, M., Swanepoel, A. & Cao, H. Early warning of limit-exceeding concentrations of cyanobacteria and cyanotoxins in drinking water reservoirs by inferential modelling. *Harmful Algae* **69,** 18–27. doi:10.1016/j.hal.2017.09.003 (2017).

256. Harris, T. D. & Graham, J. L. Predicting cyanobacterial abundance, microcystin, and geosmin in a eutrophic drinking-water reservoir using a 14-year dataset. *Lake Reserv. Manag.* **33,** 32–48. doi:10.1080/10402381.2016.1263694 (2017).

257. Tardaguila, A., Sy, J. & Punzalan, E. *QSAR models for predicting toxicities of microcystins in cyanobacteria using getaway descriptors* Research Congress 2013, De La Salle University Manila. 2013.

258. Covaci, O. I. *et al.* Highly sensitive detection and discrimination of LR and YR microcystins based on protein phosphatases and an artificial neural network. *Anal. Bioanal. Chem.* **404,** 711–720. doi:10.1007/s00216-012-6092-6 (2012).

259. Da Silva, C. G., Duque, M. D., Freire Nordi, C. S. & Viana-Niero, C. New insights into toxicity of microcystins produced by cyanobacteria using in silico ADMET prediction. *Toxicon* **204,** 64–71. ISSN: 00410101. doi:10.1016/j.toxicon.2021.11.002 (2021).

260. *Toxic cyanobacteria in Water: a guide to their public health consequences, monitoring and management* (eds Bartram, J. & Chorus, I.) 0th ed. (CRC Press, 1999). ISBN: 978-0-429-17845-0. doi:10.1201/9781482295061.

261. Berry, M. A. *et al.* Cyanobacterial harmful algal blooms are a biological disturbance to Western Lake Erie bacterial communities: bacterial community ecology of CHABs. *Environ. Microbiol.* **19,** 1149–1162. doi:10.1111/1462-2920.13640 (2017).

262. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46,** 2699–2699. doi:10.1093/nar/gky092 (2018).

263. The UniProt Consortium *et al.* UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51,** D523–D531. doi:10.1093/nar/gkac1052 (2023).

264. Pedregosa, F. *et al.* Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12,** 2825–2830 (2011).

265. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16,** 321–357. doi:10.1613/jair.953 (2002).

266. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18,** 1–5. doi:10.48550/arXiv.1609.06570 (2017).

267. Beleites, C., Neugebauer, U., Bocklitz, T., Krafft, C. & Popp, J. Sample size planning for classification models. *Anal. Chim. Acta* **760,** 25–33. doi:10.1016/j.aca.2012.11.007 (2013).

268. Spoof, L., Błaszczyk, A., Meriluoto, J., Cegłowska, M. & Mazur-Marzec, H. Structures and activity of new anabaenopeptins produced by baltic sea cyanobacteria. *Mar. Drugs* **14,** 8. doi:10.3390/md14010008 (2015).

269. Moscato, P., Jaeger-Honz, S., Haque, M. N. & Schreiber, F. The (α,β)-k Boolean signatures of molecular toxicity: microcystin as a case study. *bioRxiv.* doi:10.1101/2024.12.29.630644 (2024).

270. Ponzoni, I. *et al.* QSAR classification models for predicting the activity of inhibitors of β-secretase (BACE1) associated with alzheimer's disease. *Sci. Rep.* **9,** 9102. doi:10.1038/s41598-019-45522-3 (2019).

271. Eklund, M., Norinder, U., Boyer, S. & Carlsson, L. Choosing feature selection and learning algorithms in QSAR. *J. Chem. Inf. Model.* **54,** 837–843. doi:10.1021/ci400573c (2014).

272. Grisoni, F., Consonni, V. & Todeschini, R. *Impact of molecular descriptors on computational models* (ed Brown, J.) 171–209. doi:10.1007/978-1-4939-8639-2_5 (Springer New York, 2018).

273. Holzinger, A. *From machine learning to explainable AI* in *World Symposium on DISA* (2018), 55–66. doi:10.1109/DISA.2018.8490530.

274. Goebel, R. *et al. Explainable AI: the new 42?* in *Machine Learning and Knowledge Extraction* (eds Holzinger, A., Kieseberg, P., Tjoa, A. M. & Weippl, E.) (2018), 295–303. doi:10.1007/978-3-319-99740-7_21.

275. Hansch, C. & Fujita, T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **86,** 1616–1626. doi:10.1021/ja01062a035 (1964).

276. Martínez, M. J., Razuc, M. & Ponzoni, I. MoDeSuS: a machine learning tool for selection of molecular descriptors in QSAR studies applied to molecular informatics. *BioMed Res. Int.* **2019,** 2905203. doi:10.1155/2019/2905203 (2019).

277. Organization for Economic Cooperation and Development. *OECD principles for the validation, for regulatory purposes, of (quantitative) structure-activity relationship models* https://www.oecd.org/chemicalsafety/risk-assessment/37849783.pdf (accessed 2020-11-27). 2004.

278. Chakravarti, S. K. & Alla, S. R. M. Descriptor free QSAR modeling using deep learning with long short-term memory neural networks. *Front. Artif. Intell.* **2,** 17. doi:10.3389/frai.2019.00017 (2019).

279. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* **12,** 43. doi:10.1186/s13321-020-00445-4 (2020).

280. Xia, X., Maliski, E. G., Gallant, P. & Rogers, D. Classification of kinase inhibitors using a bayesian model. *J. Med. Chem.* **47,** 4463–4470. doi:10.1021/jm0303195 (2004).

281. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555,** 604–610. doi:10.1038/nature25978 (2018).

282. Davies, S. & Russell, S. *NP-Completeness of searches for smallest possible feature sets* in *Proceedings of the 1994 AAAI Fall Symposium on Relevance* (AAAI Press, 1994), 37–39.

283. Üney, F. & Türkay, M. A mixed-integer programming approach to multi-class data classification problem. *Eur. J. Oper. Res.* **173,** 910–920. doi:10.1016/j.ejor.2005.04.049 (2006).

284. Xu, G. & Papageorgiou, L. G. A mixed integer optimisation model for data classification. *Comput. Ind. Eng.* **56,** 1205–1215. doi:10.1016/j.cie.2008.07.012 (2009).

285. Yang, L., Liu, S., Tsoka, S. & Papageorgiou, L. G. Sample re-weighting hyper box classifier for multi-class data classification. *Comput. Ind. Eng.* **85,** 44–56. doi:10.1016/j.cie.2015.02.022 (2015).

286. Tan, R. R., Aviso, K. B., Janairo, J. I. B. & Promentilla, M. A. B. A hyperbox classifier model for identifying secure carbon dioxide reservoirs. *J. Clean. Prod.* **272,** 122181. doi:10.1016/j.jclepro.2020.122181 (2020).

287. Ooi, Y. J. *et al.* Design of fragrance molecules using computer-aided molecular design with machine learning. *Comput. Chem. Eng.* **157,** 107585. doi:10.1016/j.compchemeng.2021.107585 (2022).

288. Austin, N. D., Sahinidis, N. V. & Trahan, D. W. Computer-aided molecular design: an introduction and review of tools, applications, and solution techniques. *Chem. Eng. Res. Des.* **116,** 2–26. doi:10.1016/j.cherd.2016.10.014 (2016).

289. Jaeger-Honz, S. *et al.* Investigation of microcystin conformation and binding towards PPP1 by molecular dynamics simulation. *Chem. Biol. Interact.* **351,** 109766. doi:10.1016/j.cbi.2021.109766 (2022).

290. NCBI. *PubChem compound summary for CID* 44271302 https://pubchem.ncbi.nlm.nih.gov/compound/44271302. 2021.

291. NCBI. *PubChem compound summary for CID* 44271308 https://pubchem.ncbi.nlm.nih.gov/compound/44271308. 2021.

292. NCBI. *PubChem compound summary for CID* 44271410 https://pubchem.ncbi.nlm.nih.gov/compound/44271410. 2021.

293. NCBI. *PubChem compound summary for CID* 44271411 https://pubchem.ncbi.nlm.nih.gov/compound/44271411. 2021.

294. NCBI. *PubChem compound summary for CID* 44271325 https://pubchem.ncbi.nlm.nih.gov/compound/44271325. 2021.

295. Landrum, G. *RDKit: open-source cheminformatics software* https://github.com/rdkit/rdkit/, Release 2020-03 (accessed 2021-04-21). 2020.

296. IBM. *CPLEX optimization studio V12.8* https://www.ibm.com/support/pages/cplex-optimization-studio-v128. 2021.

297. Inkscape Project. *Inkscape* https://inkscape.org, Version Number: 0.92.3. 2017.

298. Namikoshi, M., Rinehart, K. L., Dahlem, A. M., Beasley, V. R. & Carmichael, W. W. Total synthesis of Adda, the unique C20 amino acid of cyanobacterial hepatotoxins. *Tetrahedron Lett.* **30,** 4349–4352. ISSN: 00404039. doi:10.1016/S0040-4039(00)99357-2 (1989).

299. Taylor, C., Quinn, R. J., Suganuma, M. & Fujiki, H. Inhibition of protein phosphatase 2A by cyclic peptides modelled on the microcystin ring. *Bioorg. Med. Chem. Lett.* **6,** 2113–2116. doi:10.1016/0960-894X(96)00378-2 (1996).

300. Fontanillo, M. *et al.* Synthesis of highly selective submicromolar microcystin-based inhibitors of protein phosphatase (PP)2A over PP1. *Angew. Chem. Int. Ed.* **55,** 13985–13989. doi:10.1002/anie.201606449 (2016).

301. Xu, Y., Cui, J., Yu, H. & Zong, W. Insight into the molecular mechanism for the discrepant inhibition of microcystins (MCLR, LA, LF, LW, LY) on protein phosphatase 2A. *Toxins* **14,** 390. doi:10.3390/toxins14060390 (2022).

302. Sivonen, K. *et al.* Isolation and characterization of hepatotoxic microcystin homologs from the filamentous freshwater cyanobacterium Nostoc sp. strain 152. *Appl. Environ. Microbiol.* **56,** 2650–2657. ISSN: 0099-2240, 1098-5336. doi:10.1128/aem.56.9.2650-2657.1990. (2024) (1990).

303. Rinehart, K. L., Namikoshi, M. & Choi, B. W. Structure and biosynthesis of toxins from blue-green algae (cyanobacteria). *J. Appl. Phycol.* **6,** 159–176. doi:10.1007/BF02186070 (1994).

304. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402. doi:10.1093/nar/25.17.3389 (1997).

305. Fassler, J. & Cooper, P. *BLAST Glossary* https://www.ncbi.nlm.nih.gov/books/NBK62051/, accessed 2023-06-07.

306. Andreeva, A. V. & Kutuzov, M. A. PPP family of protein ser/thr phosphatases: two distinct branches? *Mol. Biol. Evol.* **18,** 448–452. doi:10.1093/oxfordjournals.molbev.a003823 (2001).

307. Swingle, M. R., Honkanen, R. E. & Ciszak, E. M. Structural basis for the catalytic activity of human serine/threonine protein phosphatase-5. *J. Biol. Chem.* **279,** 33992–33999. doi:10.1074/jbc.m402855200 (2004).

308. Swingle, M. R. *et al.* Structure-activity relationship studies of fostriecin, cytostatin, and key analogs, with PP1, PP2A, PP5, and (β12–β13)-chimeras (PP1/PP2A and PP5/PP2A), provide further insight into the inhibitory actions of fostriecin family inhibitors. *J. Pharmacol. Exp. Ther.* **331,** 45–53. doi:10.1124/jpet.109.155630 (2009).

309. Stepan, A. F. *et al.* Structural alert/reactive metabolite concept as applied in medicinal chemistry to mitigate the risk of idiosyncratic drug toxicity: a perspective based on the critical examination of trends in the top 200 drugs marketed in the United States. *Chem. Res. Toxicol.* **24,** 1345–1410. doi:10.1021/tx200168d (2011).

310. Alves, V. M. *et al.* Alarms about structural alerts. *Green Chem.* **18,** 4348–4360. doi:10.1039/C6GC01492E (2016).

311. Pereira, S. R., Vasconcelos, V. M. & Antunes, A. Computational study of the covalent bonding of microcystins to cysteine residues - a reaction involved in the inhibition of the PPP family of protein phosphatases. *FEBS J.* **280,** 674–680. doi:10.1111/j.1742-4658.2011.08454.x (2013).

312. Hu, Z., Wang, X., Zhang, S. & Zong, W. Research on the discrepant inhibition mechanism of microcystin-LR disinfectant by-products target to protein phosphatase 1. *Environ. Sci. Pollut. Res.* doi:10.1007/s11356-021-12472-1 (2021).

313. Virshup, D. M. & Shenolikar, S. From promiscuity to precision: protein phosphatases get a makeover. *Mol. Cell* **33,** 537–545. doi:https://doi.org/10.1016/j.molcel.2009.02.015 (2009).

314. Peti, W., Nairn, A. C. & Page, R. Structural basis for protein phosphatase 1 regulation and specificity. *FEBS J.* **280,** 596–611. doi:https://doi.org/10.1111/j.1742-4658.2012.08509.x (2013).

315. Pochodylo, A. L., Aoki, T. G. & Aristilde, L. Adsorption mechanisms of microcystin variant conformations at water–mineral interfaces: a molecular modeling investigation. *J. Colloid Interface Sci.* **480,** 166–174. doi:10.1016/j.jcis.2016.07.016 (2016).

316. Pochodylo, A. L., Klein, A. R. & Aristilde, L. Metal-binding selectivity and coordination dynamics for cyanobacterial microcystins with Zn, Cu, Fe, Mg, and Ca. *Environ. Chem. Lett.* **15,** 695–701. doi:10.1007/s10311-017-0639-x (2017).

317. Yu, H., Xu, Y., Cui, J. & Zong, W. Mechanism for the potential inhibition effect of microcystin-LR disinfectant by-products on protein phosphatase 2A. *Toxins* **14,** 878. doi:10.3390/toxins14120878 (2022).

318. McPartlin, D. A. *et al.* Understanding microcystin-LR antibody binding interactions using in silico docking and in vitro mutagenesis. *Protein Eng. Des. Sel.* **32,** 533–542. doi:10.1093/protein/gzaa016 (2019).

319. Liu, Y. *et al.* Epitopes prediction for microcystin-LR by molecular docking. *Ecotoxicol. Environ. Saf.* **227,** 112925. doi:10.1016/j.ecoenv.2021.112925 (2021).

320. Trogen, G.-B. *et al.* Conformational studies of microcystin-LR using NMR spectroscopy and molecular dynamics calculations. *Biochemistry* **35,** 3197–3205. doi:10.1021/bi952368s (1996).

321. Zong, W. *et al.* Molecular mechanism for the regulation of microcystin toxicity to protein phosphatase 1 by glutathione conjugation pathway. *BioMed Res. Int.* **2017,** 9676504. doi:10.1155/2017/9676504 (2017).

322. Lanaras, T., Cook, C., Eriksson, J., Meriluoto, J. & Hotokka, M. Computer modelling of the 3-dimensional structures of the cyanobacterial hepatotoxins microcystin-LR and nodularin. *Toxicon* **29,** 901–906. doi:10.1016/0041-0101(91)90228-J (1991).

323. Taylor, C., Quinn, R. J., McCullock, R., Nishiwaki-Matsushima, R. & Fujiki, H. An alternative computer model of the 3-dimensional structural of microcystin-LR and nodularin rationalising their interactions with protein phosphatases 1 and 2A. *Bioorganic Med. Chem. Lett.* **2,** 299–302. doi:10.1016/S0960-894X(01)80204-3 (1992).

324. Rudolph-Böhner, S., Mierke, D. F. & Moroder, L. Molecular structure of the cyanobacterial tumor-promoting microcystins. *FEBS Lett.* **349,** 319–323. doi:10.1016/0014-5793(94)00680-6 (1994).

325. Bagu, J. R. *et al.* Comparison of the solution structures of microcystin-LR and motuporin. *Nat. Struct. Mol. Biol.* **2,** 114–116. doi:10.1038/nsb0295-114 (1995).

326. Bagu, J. R., Sykes, B. D., Craig, M. M. & Holmes, C. F. A molecular basis for different interactions of marine toxins with protein phosphatase-1 molecular models for bound motuporin, microcystins, okadaic acid, and calyculin A. *J. Biol. Chem.* **272,** 5087–5097. doi:10.1074/jbc.272.8.5087 (1997).

327. Lavigne, P. *et al.* Structure-based thermodynamic analysis of the dissociation of protein phosphatase-1 catalytic subunit and microcystin-LR docked complexes. *Protein Sci.* **9,** 252–264. doi:10.1110/ps.9.2.252 (2000).

328. Bernstein, F. C. *et al.* The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112,** 535–542. doi:10.1016/S0022-2836(77)80200-3 (1977).

329. Pettersen, E. F. *et al.* UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25,** 1605–1612. doi:10.1002/jcc.20084 (2004).

330. MATLAB. *version 7.10.0 (R2016a)* (The MathWorks Inc., Natick, Massachusetts, 2016).

331. Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* **30,** 2785–2791. doi:10.1002/jcc.21256 (2009).

332. Chen, V. B. *et al. MolProbity*: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21. doi:10.1107/S0907444909042073 (2010).

333. Williams, C. J. *et al.* MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27,** 293–315. doi:10.1002/pro.3330 (2018).

334. Bock, C. W., Katz, A. K., Markham, G. D. & Glusker, J. P. Manganese as a replacement for magnesium and zinc: functional comparison of the divalent ions. *J. Am. Chem. Soc.* **121,** 7360–7372. doi:10.1021/ja9906960 (1999).

335. Bayly, C. I., Cieplak, P., Cornell, W. & Kollman, P. A. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **97,** 10269–10280. doi:10.1021/j100142a004 (1993).

336. Dupradeau, F.-Y. *et al.* The R.E.D. tools: advances in RESP and ESP charge derivation and force field library building. *Phys. Chem. Chem. Phys.* **12,** 7821–7839. doi:10.1039/C0CP00111B (2010).

337. Vanquelef, E. *et al.* R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. *Nucleic Acids Res.* **39,** W511–W517. doi:10.1093/nar/gkr288 (2011).

338. Wang, F., Becker, J.-P., Cieplak, P. & Dupradeau, F.-Y. *R.E.D. Python: object oriented programming for Amber force fields* 2013.

339. Schmidt, M. *et al.* General atomic and molecular electronic structure system. *J. Comput. Chem.* **14,** 1347–1363. doi:10.1002/jcc.540141112 (1993).

340. Case, D. *et al. Amber 16, University of California, San Francisco.* 2016. doi:10.13140/RG.2.2.27958.70729.

341. Shirts, M. R. *et al.* Lessons learned from comparing molecular dynamics engines on the SAMPL5 dataset. *J. Comput. Aided Mol. Des.* **31,** 147–161. doi:10.1007/s10822-016-9977-1 (2017).

342. Bekker, H. *et al.* GROMACS - a parallel computer for molecular-dynamics simulations. *Physics Computing.* 4th International Conference on Computational Physics (PC 92), 252–256 (1993).

343. Abraham, M. J. *et al.* GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *Softwarex* **1-2,** 19–25. doi:10.1016/j.softx.2015.06.001 (2015).

344. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79,** 926–935. doi:10.1063/1.445869 (1983).

345. Darden, T., York, D. & Pedersen, L. Particle mesh Ewald: an $N \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **98,** 10089–10092. doi:10.1063/1.464397 (1993).

346. Páll, S. & Hess, B. A flexible algorithm for calculating pair interactions on SIMD architectures. *Comput. Phys. Commun.* **184,** 2641–2650. doi:10.1016/j.cpc.2013.06.003 (2013).

347. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18,** 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H (1997).

348. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126,** 014101. doi:10.1063/1.2408420 (2007).

349. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52,** 7182–7190. doi:10.1063/1.328693 (1981).

350. Abraham, M., van der Spoel, D., Lindahl, E., Hess, B. & the GROMACS development team. *GROMACS user manual version 2016.4, www.gromacs.org* (2017).

351. Van Rossum, G. & Drake Jr, F. L. *Python reference manual* (Centrum voor Wiskunde en Informatica Amsterdam, 1995).

352. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585,** 357–362. doi:10.1038/s41586-020-2649-2 (2020).

353. Jones, E., Oliphant, T., Peterson, P., *et al. SciPy: open source scientific tools for python* http://www.scipy.org/, accessed 2021-06-10. 2001.

354. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Chem. Eng.* **9,** 90–95. doi:10.1109/MCSE.2007.55 (2007).

355. Biehl, R. Jscatter, a program for evaluation and analysis of experimental data. *PLoS ONE* **14,** 1–18. doi:10.1371/journal.pone.0218789 (2019).

356. *Grace - 2D plotting tool* https://plasma-gate.weizmann.ac.il/Grace/, accessed 10-06-2021. 2008.

357. Schrödinger, LLC. *PyMOL - The PyMOL Molecular Graphics System, Version 0.99, Schrödinger, LLC.* 2015.

358. Lobanov, M. Y., Bogatyreva, N. S. & Galzitskaya, O. V. Radius of gyration as an indicator of protein structure compactness. *Mol. Biol.* **42,** 623–628. doi:10.1134/S0026893308040195 (2008).

359. Lee, B. & Richards, F. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55,** 379–IN4. doi:10.1016/0022-2836(71)90324-X (1971).

360. Maisuradze, G. G., Liwo, A. & Scheraga, H. A. Principal component analysis for protein folding dynamics. *J. Mol. Biol.* **385,** 312–329. doi:10.1016/j.jmb.2008.10.018 (2009).

361. Harada, K. *et al.* Structural determination of geometrical isomers of Microcystins LR and RR from cyanobacteria by two-dimensional NMR spectroscopic techniques. *Chem. Res. Toxicol.* **3,** 473–481. doi:10.1021/tx00017a014 (1990).

362. Moorhead, G., MacKintosh, R. W., Morrice, N., Gallagher, T. & MacKintosh, C. Purification of type 1 protein (serine/threonine) phosphatases by microcystin-sepharose affinity chromatography. *FEBS Lett.* **356,** 46–50. doi:10.1016/0014-5793(94)01232-6 (1994).

363. Stotts, R. R. *et al.* Structural modifications imparting reduced toxicity in microcystins from Microcystis spp. *Toxicon* **31,** 783–789. doi:10.1016/0041-0101(93)90384-U (1993).

364. Fotler, R. *Toxicological profile of microcystins: determination of the toxicokinetic and toxicodynamic mechanisms of microcystins in human cells* Available at http://nbn-resolving.de/urn:nbn:de:bsz:352-2-28s9cke9kuei6. PhD thesis (University of Konstanz, 2024).

365. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46,** W296–W303. doi:10.1093/nar/gky427 (2018).

366. Benkert, P., Biasini, M. & Schwede, T. Toward the estimation of the absolute quality of individual protein structure models. *Bioinformatics* **27,** 343–350. doi:10.1093/bioinformatics/btq662 (2011).

367. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci. Rep.* **7,** 10480. doi:10.1038/s41598-017-09654-8 (2017).

368. Bienert, S. *et al.* The SWISS-MODEL repository-new features and functionality. *Nucleic Acids Res.* **45,** D313–D319. doi:10.1093/nar/gkw1132 (2017).

369. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis* **30,** S162–S173. doi:10.1002/elps.200900140 (2009).

370. Studer, G. *et al.* QMEAND is co-distance constraints applied on model quality estimation. *Bioinformatics* **36,** 1765–1771. doi:10.1093/bioinformatics/btz828 (2020).

371. *Open source PyMOL: the PyMOL Molecular Graphics System* https://github.com/bieniekmateusz/pymol-mdanalysis/tree/fellows_mp_2018, Version 2.4.0.

372. Miles, C. O. *et al.* Conjugation of microcystins with thiols is reversible: base-catalyzed deconjugation for chemical analysis. *Chem. Res. Toxicol.* **29,** 860–870. doi:10.1021/acs.chemrestox.6b00028 (2016).

373. Zemskov, I., Kropp, H. M. & Wittmann, V. Regioselective cleavage of thioether linkages in microcystin conjugates. *Chem. Eur. J.* **22,** 10990–10997. doi:10.1002/chem.201601660 (2016).

374. Vanommeslaeghe, K. *et al.* CHARMM general force field: a force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.* **31,** 671–690. doi:10.1002/jcc.21367 (2010).

375. Vanommeslaeghe, K. & Mackerell, A. D. Automation of the CHARMM General Force Field (CGenFF) I: bond perception and atom typing. *J. Chem. Inf. Model.* **52,** 3144–3154. doi:10.1021/ci300363c (2012).

376. Yu, W., He, X., Vanommeslaeghe, K. & MacKerell, A. D. Extension of the CHARMM General Force Field to sulfonyl-containing compounds and its utility in biomolecular simulations. *J. Comput. Chem.* **33,** 2451–2468. doi:10.1002/jcc.23067 (2012).

377. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31,** 455–461. doi:10.1002/jcc.21334 (2010).

378. Vanommeslaeghe, K., Raman, E. P. & Mackerell, A. D. Automation of the CHARMM General Force Field (CGenFF) II: assignment of bonded parameters and partial atomic charges. *J. Chem. Inf. Model.* **52,** 3155–3168. doi:10.1021/ci3003649 (2012).

379. Soteras Gutiérrez, I. *et al.* Parametrization of halogen bonds in the CHARMM general force field: improved treatment of ligand-protein interactions. *Bioorg. Med. Chem.* **24,** 4812–4825. doi:10.1016/j.bmc.2016.06.034 (2016).

380. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14,** 71–73. doi:10.1038/nmeth.4067 (2017).

381. Best, R. B. *et al.* Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi(1)$ and $\chi(2)$ dihedral angles. *J. Chem. Theory Comput.* **8,** 3257–3273. doi:10.1021/ct300400x (2012).

382. Allnér, O., Nilsson, L. & Villa, A. Magnesium ion-water coordination and exchange in biomolecular simulations. *J. Chem. Theory Comput.* **8,** 1493–1502. doi:10.1021/ct3000734 (2012).

383. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81,** 3684–3690. doi:10.1063/1.448118 (1984).

384. Law, S. M. *Modevectors* pymolwiki.org/index.php/Modevectors, (accessed 2023-03-01).

385. Bero, S. A., Muda, A. K., Choo, Y. H., Muda, N. A. & Pratama, S. F. Similarity measure for molecular structure: a brief review. *J. Phys. Conf. Ser.* **892,** 012015. doi:10.1088/1742-6596/892/1/012015 (2017).

386. Stumpfe, D. & Bajorath, J. Similarity searching. *WIREs Comput. Mol. Sci.* **1,** 260–282. doi:10.1002/wcms.23 (2011).

387. Fassio, A. V. *et al.* Prioritizing virtual screening with interpretable interaction fingerprints. *J. Chem. Inf. Model.* **62,** 4300–4318. doi:10.1021/acs.jcim.2c00695 (2022).

388. Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71,** 58–63. doi:10.1016/j.ymeth.2014.08.005 (2015).

389. Li, G.-B. *et al.* IFPTarget: a customized virtual target identification method based on protein–ligand interaction fingerprinting analyses. *J. Chem. Inf. Model.* **57,** 1640–1651. doi:10.1021/acs.jcim.7b00225 (2017).

390. Kumar, S. & Kim, M.-h. SMPLIP-score: predicting ligand binding affinity from simple and interpretable on-the-fly interaction fingerprint pattern descriptors. *J. Cheminform.* **13,** 28. doi:10.1186/s13321-021-00507-1 (2021).

391. Bouysset, C. & Fiorucci, S. ProLIF: a library to encode molecular interactions as fingerprints. *J. Cheminformatics* **13,** 72. doi:10.1186/s13321-021-00548-6 (2021).

392. Desaphy, J., Raimbaud, E., Ducrot, P. & Rognan, D. Encoding protein–ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.* **53,** 623–637. doi:10.1021/ci300566n (2013).

393. Deng, Z., Chuaqui, C. & Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein−ligand binding interactions. *J. Med. Chem.* **47,** 337–344. doi:10.1021/jm030331x (2004).

394. Jasper, J. B., Humbeck, L., Brinkjost, T. & Koch, O. A novel interaction fingerprint derived from per atom score contributions: exhaustive evaluation of interaction fingerprint performance in docking based virtual screening. *J. Cheminform.* **10,** 15. doi:10.1186/s13321-018-0264-0 (2018).

395. Mpamhanga, C. P., Chen, B., McLay, I. M. & Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *J. Chem. Inf. Model.* **46,** 686–698. doi:10.1021/ci050420d (2006).

396. Marcou, G. & Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.* **47,** 195–207. doi:10.1021/ci600342e (2007).

397. Tan, L., Lounkine, E. & Bajorath, J. Similarity searching using fingerprints of molecular fragments involved in protein−ligand interactions. *J. Chem. Inf. Model.* **48,** 2308–2312. doi:10.1021/ci800322y (2008).

398. Crisman, T. J., Sisay, M. T. & Bajorath, J. Ligand-target interaction-based weighting of substructures for virtual screening. *J. Chem. Inf. Model.* **48,** 1955–1964. doi:10.1021/ci800229q (2008).

399. Da, C. & Kireev, D. Structural protein–ligand interaction fingerprints (SPLIF) for structure-based virtual screening: method and benchmark study. *J. Chem. Inf. Model.* **54,** 2555–2561. doi:10.1021/ci500319f (2014).

400. Salentin, S., Schreiber, S., Haupt, V. J., Adasme, M. F. & Schroeder, M. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Res.* **43,** W443–W447. doi:10.1093/nar/gkv315 (W1 2015).

401. Jubb, H. C. *et al.* Arpeggio: a web server for calculating and visualising interatomic interactions in protein structures. *J. Mol. Biol.* **429,** 365–371. doi:10.1016/j.jmb.2016.12.004 (2017).

402. Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M. & Siedlecki, P. Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35,** 1334–1341. doi:10.1093/bioinformatics/bty757 (2019).

403. Thangapandian, S. *et al. Quantitative target-specific toxicity prediction modeling (QTTPM): coupling machine learning with dynamic protein–ligand interaction descriptors (DyPLIDs) to predict androgen receptor-mediated toxicity* (ed Hong, H.) 263–295. doi:10.1007/978-3-031-20730-3_11 (Springer International Publishing, Cham, 2023).

404. Szulc, N. A., Mackiewicz, Z., Bujnicki, J. M. & Stefaniak, F. Structural interaction fingerprints and machine learning for predicting and explaining binding of small molecule ligands to RNA. *Brief. Bioinformatics* **24,** bbad187. doi:10.1093/bib/bbad187 (2023).

405. Kokh, D. B. *et al.* A workflow for exploring ligand dissociation from a macromolecule: efficient random acceleration molecular dynamics simulation and interaction fingerprint analysis of ligand trajectories. *J. Chem. Phys.* **153,** 125102. doi:10.1063/5.0019088 (2020).

406. Nandigam, R. K., Kim, S., Singh, J. & Chuaqui, C. Position specific interaction dependent scoring technique for virtual screening based on weighted protein−ligand interaction fingerprint profiles. *J. Chem. Inf. Model.* **49,** 1185–1192. doi:10.1021/ci800466n (2009).

407. Chuaqui, C., Deng, Z. & Singh, J. Interaction profiles of protein kinase−inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **48,** 121–133. doi:10.1021/jm049312t (2005).

408. Istyastono, E. P., Radifar, M., Yuniarti, N., Prasasty, V. D. & Mungkasi, S. PyPLIF HIPPOS: a molecular interaction fingerprinting tool for docking results of AutoDock Vina and PLANTS. *J. Chem. Inf. Model.* **60,** 3697–3702. doi:10.1021/acs.jcim.0c00305 (2020).

409. Gelpi, J., Hospital, A., Goñi, R. & Orozco, M. Molecular dynamics simulations: advances and applications. *Adv. Appl. Bioinform. Chem.,* 37. doi:10.2147/AABC.S70333 (2015).

410. Schlick, T. & Portillo-Ledesma, S. Biomolecular modeling thrives in the age of technology. *Nat. Comput. Sci.* **1,** 321–331. doi:10.1038/s43588-021-00060-9 (2021).

411. Bedart, C. *et al.* SINAPs: a software tool for analysis and visualization of interaction networks of molecular dynamics simulations. *J. Chem. Inf. Model.,* acs.jcim.1c00854. doi:10.1021/acs.jcim.1c00854 (2022).

412. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. of Mol. Graph.* **14,** 33–38. doi:10.1016/0263-7855(96)00018-5 (1996).

413. Badaczewska-Dawid, A. E., Nithin, C., Wroblewski, K., Kurcinski, M. & Kmiecik, S. MAPIYA contact map server for identification and visualization of molecular interactions in proteins and biological complexes. *Nucleic Acids Res.* **50,** W474–W482. doi:10.1093/nar/gkac307 (2022).

414. Motono, C., Yanagida, S., Sato, M. & Hirokawa, T. MDContactCom: a tool to identify differences of protein molecular dynamics from two MD simulation trajectories in terms of interresidue contacts. *Bioinformatics* **38,** 273–274. doi:10.1093/bioinformatics/btab538 (2021).

415. Mercadante, D., Gräter, F. & Daday, C. CONAN: a tool to decode dynamical information from molecular interaction maps. *Biophys. J.* **114,** 1267–1273. doi:10.1016/j.bpj.2018.01.033 (2018).

416. Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J. Comput. Chem.* **32,** 2319–2327. doi:10.1002/jcc.21787 (2011).

417. Ferreira de Freitas, R. & Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. *Med. Chem. Comm.* **8,** 1970–1981. doi:10.1039/C7MD00381A (2017).

418. Gowers, R. J. *et al. MDAnalysis: a python package for the rapid analysis of molecular dynamics simulations* in *Proceedings of the 15th Python in Science Conference* (eds Benthall, S. & Rostrup, S.) (Austin, TX, 2016), 98–105. doi:10.25080/Majora-629e541a-00e.

419. Da Costa-Luis, C. *et al. tqdm: a fast, extensible progress bar for python and CLI* version v4.62.3. 2021. doi:10.5281/zenodo.5517697.

420. The pandas development team. *pandas-dev/pandas: Pandas* https://github.com/pandas-dev/pandas, Version 1.3.3. 2021.

421. Klein, A. *et al. imageio/imageio* version v2.28.0. 2023. doi:10.5281/zenodo.7857504.

422. Hagberg, A., Swart, P. & S Chult, D. *Exploring network structure, dynamics, and function using NetworkX* tech. rep. https://www.osti.gov/biblio/960616 (Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008).

423. Rossetti, G., pyup.io bot, Norman, U., Dormán, H. & Dorner, M. *GiulioRossetti/dynetx:* version v0.3.1. 2021. doi:10.5281/zenodo.5599265.

424. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in python. *Nat. Methods* **17,** 261–272. doi:10.1038/s41592-019-0686-2 (2020).

425. Rácz, A., Bajusz, D. & Héberger, K. Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J. Cheminformatics* **10,** 48. doi:10.1186/s13321-018-0302-y (2018).

426. Martin, Y. C., Kofron, J. L. & Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **45,** 4350–4358. doi:10.1021/jm020155c (2002).

427. Fisette, O., Lagüe, P., Gagné, S. & Morin, S. Synergistic applications of MD and NMR for the study of biological systems. *Biomed Res. Int.* **2012,** 254208. doi:10.1155/2012/254208 (2012).

428. Yu, H., Cui, J., Xu, Y., Feng, L.-J. & Zong, W. Regulation effectiveness and mechanism of biotransformation pathway on the toxicity of microcystin-LR target to Protein Phosphatase 2A. *Int. J. Environ. Res. Public Health* **20,** 964. doi:10.3390/ijerph20020964 (2023).

429. Bougueroua, S. *et al.* Graph theory for automatic structural recognition in molecular dynamics simulations. *J. Chem. Phys.* **149,** 184102. doi:10.1063/1.5045818 (2018).

430. Sato, T., Honma, T. & Yokoyama, S. Combining machine learning and pharmacophore-based interaction fingerprint for in silico screening. *J. Chem. Inf. Model.* **50,** 170–185. doi:10.1021/ci900382e (2010).

431. Rodríguez-Pérez, R., Miljković, F. & Bajorath, J. Assessing the information content of structural and protein–ligand interaction representations for the classification of kinase inhibitor binding modes via machine learning and active learning. *J. Cheminformatics* **12,** 36. doi:10.1186/s13321-020-00434-7 (2020).

Chapter 8

Supplementary Information

8.1 Structural Analysis and Prediction of Microcystin Toxicodynamics

This table has been published by Altaner et al. in Alternatives to Animal Experimentation [16].

Table 8.1: Evaluation metrics for different machine learning models built with 80-20 data split (80 % of data used for training set, 20 % of data used as test set). The data split and training was repeated for 50 times and performance evaluated. Mean and standard deviation are shown for precision, recall and F1-Score.

ML algorithm	Feature	Class	Precision	Recall	F1-Score
Combined model	-	non-toxic	0.80 ± 0.15	0.79 ± 0.17	0.77 ± 0.11
		less toxic	0.82 ± 0.16	0.86 ± 0.14	0.83 ± 0.11
		toxic	0.91 ± 0.11	0.88 ± 0.15	0.88 ± 0.11
RF	Mol2vec + ProtVec	non-toxic	0.80 ± 0.17	0.80 ± 0.16	0.78 ± 0.12
		less toxic	0.84 ± 0.16	0.90 ± 0.15	0.85 ± 0.12
		toxic	0.89 ± 0.12	0.83 ± 0.16	0.85 ± 0.12
RF	Mol2vec	non-toxic	0.79 ± 0.16	0.79 ± 0.18	0.77 ± 0.11
		less toxic	0.83 ± 0.16	0.85 ± 0.14	0.83 ± 0.10
		toxic	0.91 ± 0.11	0.88 ± 0.15	0.88 ± 0.11
XGB	Mol2vec	non-toxic	0.81 ± 0.16	0.78 ± 0.17	0.78 ± 0.11
		less toxic	0.82 ± 0.16	0.87 ± 0.13	0.83 ± 0.10
		toxic	0.89 ± 0.12	0.88 ± 0.14	0.88 ± 0.11

8.1.1 Equivalent Features of (α, β)-k-Feature Set Solution

In the following, all feature sets for the respective PPP data set are summarised. The feature set is denoted with a capital F and named after the first feature in the set. The positions listed are equivalent positions with the same Boolean feature. The results of this section have been published by Moscato et al. [269].

8.1.1.1 PPP1

For PPP1 the Boolean signature of toxicity is: {F32, F130, F232, F295, F336, F695, F1346}.

- $F32 = \{f32, f80, f144, f614, f746, f842, f858, f890, f898, f976, f1018, f1067, f1089, f1179, f1200, f1258, f1352, f1520, f1563, f1743\}$
- $F130 = \{f130, f1319, f1780, f1838, f1842\}$
- $F232 = \{f232, f546, f685, f1412, f1434, f1576\}$
- $F295 = \{f295\}$ (i. e. no other equivalent feature)
- $F336 = \{f336, f1848, f1877\}$
- $F695 = \{f695, f716, f734, f1125, f1229, f2026\}$
- $F1346 = \{f1346, f1719\}$

8.1.1.2 PPP2A

For PPP2A the Boolean signature of toxicity is: {F32, F130, F232, F773}.

- $F32 = \{f32, f80, f144, f614, f695, f716, f734, f746, f842, f858, f890, f898, f976, f1018, f1067, f1089, f1125, f1179, f1200, f1229, f1258, f1352, f1520, f1563, f1743, f2026\}$
- $F130 = \{f130, f799, f936, f1001, f1146, f1319, f1780, f1838, f1842, f2009\}$
- $F232 = \{f232, f546, f685, f747, f845, f906, f969, f1061, f1070, f1175, f1256, f1412, f1434, f1576, f1822, f1998\}$
- $F773 = \{f773, f1301, f1554, f1604\}$

8.1.1.3 PPP5

For PPP5 the Boolean signature of toxicity is: {F32, F130, F232, F295, F336, F773, F1346}.

- $F32 = \{f32, f80, f144, f614, f746, f842, f858, f890, f898, f976, f1018, f1067, f1089, f1179, f1200, f1258, f1352, f1520, f1563, f1743\}$
- $F130 = \{f130, f1319, f1780, f1838, f1842\}$
- $F232 = \{f232, f546, f685, f1412, f1434, f1576\}$
- $F295 = \{f295\}$ (i. e. no other equivalent feature)

- $F336 = \{f336, f1848, f1877\}$
- $F773 = \{f773, f1301, f1554, f1604\}$
- $F1346 = \{f1346, f1719\}$

8.1.1.4 PPP1 and data set 2

For the test data set, the feature set of the respective PPP will split up in new feature sets, as new information about the molecular structure can be present. After re-analysing PPP1 Boolean signature of toxicity with the new data set, our feature sets change as follows:

- $F32a = \{f32, f80, f144, f614, f746, f842, f858, f890, f898, f976, f1018, f1179, f1258, f1352, f1520, f1563, f1743\}$
- $F32b = \{f1067, f1089, f1200\}$
- $F130 = \{f130, f1319, f1780, f1838, f1842\}$
- $F232a = \{f232\}$ (i. e. no other equivalent feature)
- $F232b = \{f546, f1412, f1434\}$
- $F232c = \{f685\}$ (i. e. no other equivalent feature)
- $F232d = \{f1576\}$ (i. e. no other equivalent feature)
- $F295 = \{f295\}$ (i. e. no other equivalent feature)
- $F336 = \{f336, f1848, f1877\}$
- $F695 = \{f695, f716, f734, f1125, f1229, f2026\}$
- $F1346a = \{f1346\}$ (i. e. no other equivalent feature)
- $F1346b = \{f1719\}$ (i. e. no other equivalent feature)

8.2 Molecular Dynamics Simulation of Macrocyclic Structures

8.2.1 Molecular Dynamics Simulation of 4 MC congeners

All images and tables in this section have been published by Jaeger-Honz et al. in Chemico-Biological Interactions [289].

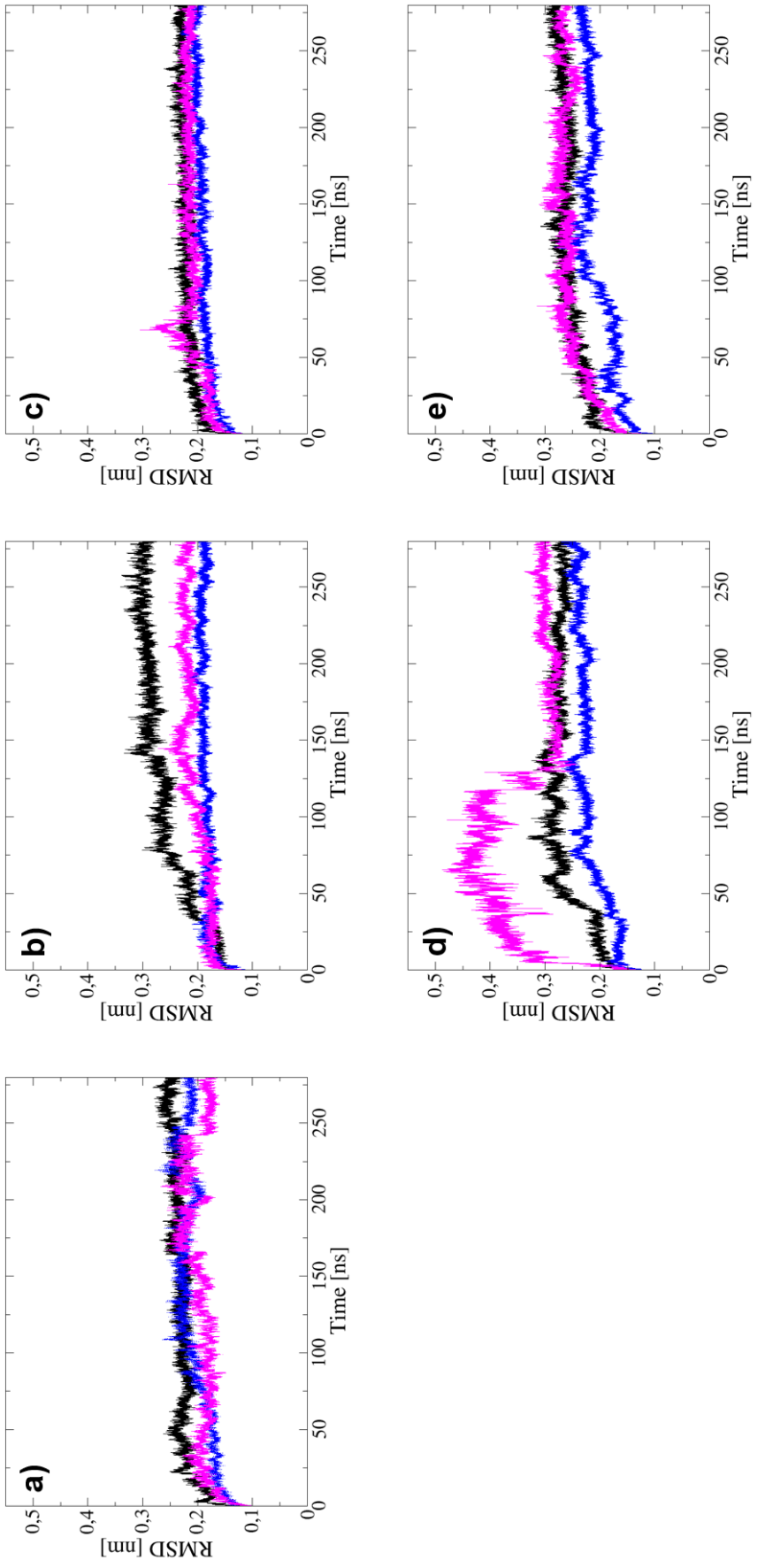


Figure 8.1: Root mean squared deviation of PPP1 (apo) and PPP1-MC-congener (complex) MD simulation shown as time series. Replicate 1 is shown in black, replicate 2 in blue and replicate 3 in purple. a) PPP1, b) PPP1-MC-LR, c) PPP1-MC-LF, d) PPP1-[*Enantio*-Adda5]MC-LF, and **e** PPP1-[β-D-Asp3, Dhb7]-MC-RR.

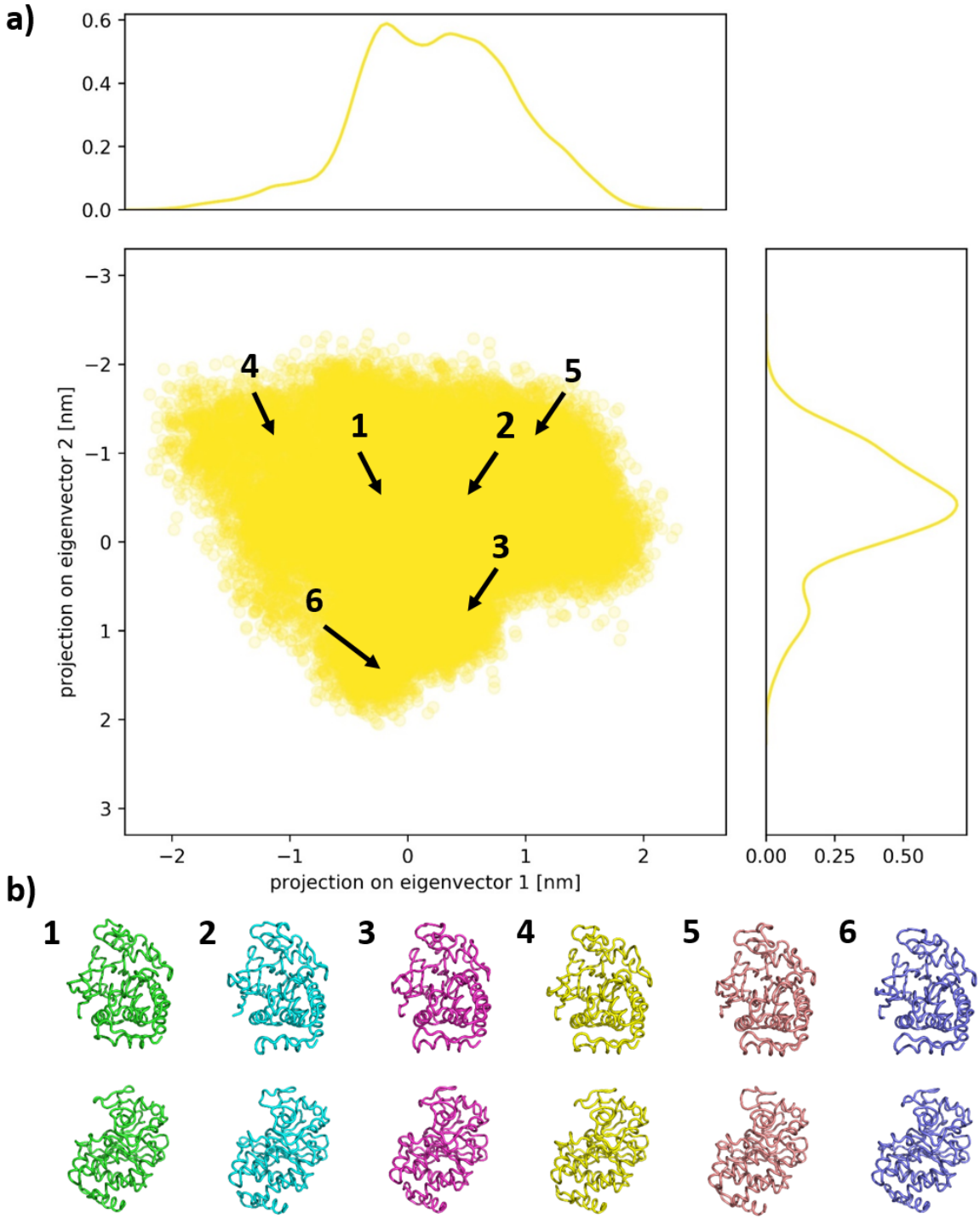


Figure 8.2: Principal component analysis of PPP1 MD Simulation. a) The 2D projection of the PPP1 backbone structure is shown in the central scatter plot. A distribution line showing the frequency of occurrence is shown at the top and left. The numbers and arrows in the scatter plot point to the position of b) respective 3D structures of PPP1.

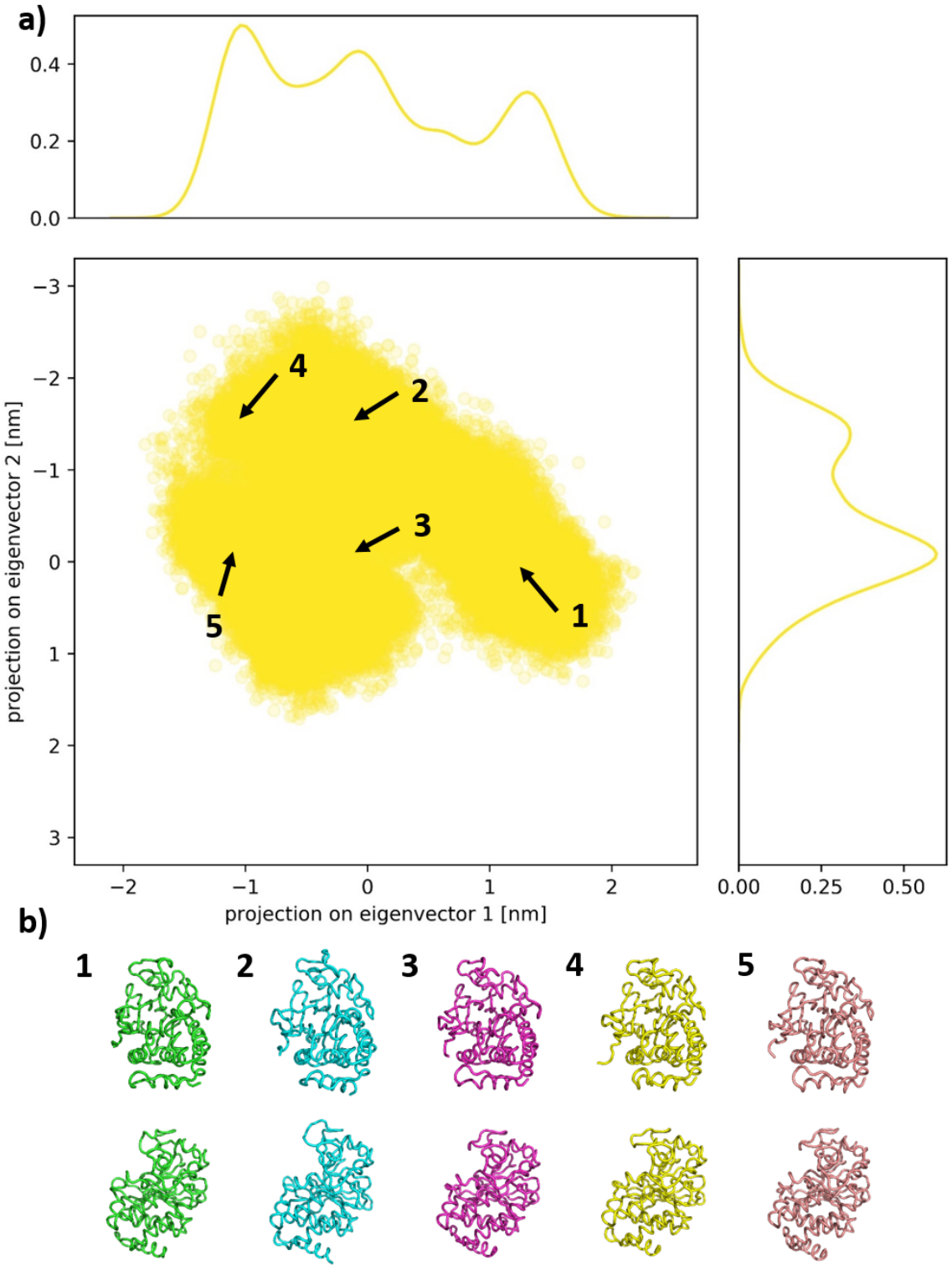


Figure 8.3: Principal component analysis of PPP1-MC-LR MD Simulation. a) The 2D projection of the PPP1 backbone structure is shown in the central scatter plot. A distribution line showing the frequency of occurrence is shown at the top and left. The numbers and arrows in the scatter plot point to the position of b) respective 3D structures of PPP1.

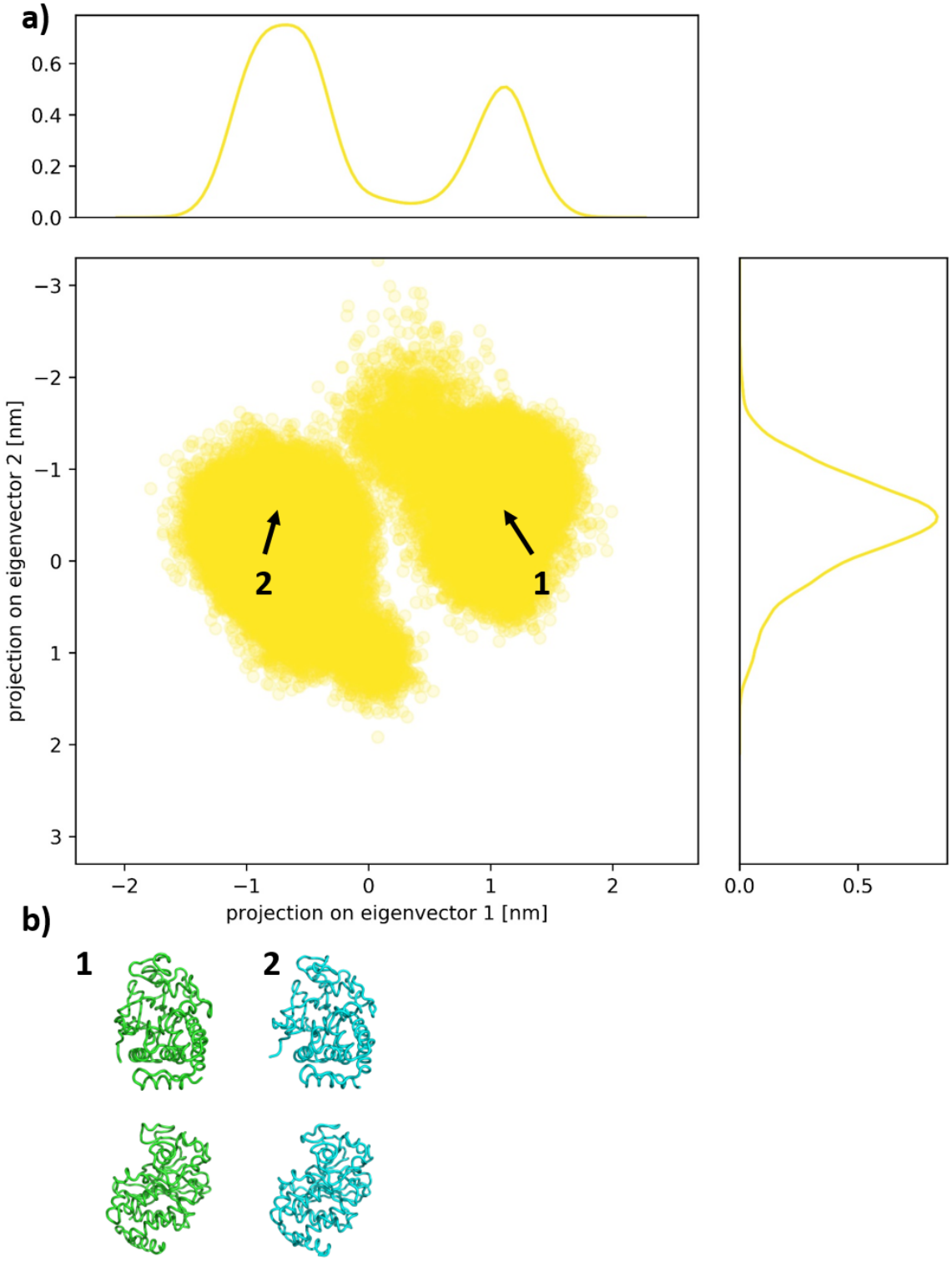


Figure 8.4: Principal component analysis of PPP1-MC-LF MD Simulation. a) The 2D projection of the PPP1 backbone structure is shown in the central scatter plot. A distribution line showing the frequency of occurrence is shown at the top and left. The numbers and arrows in the scatter plot point to the position of b) respective 3D structures of PPP1.

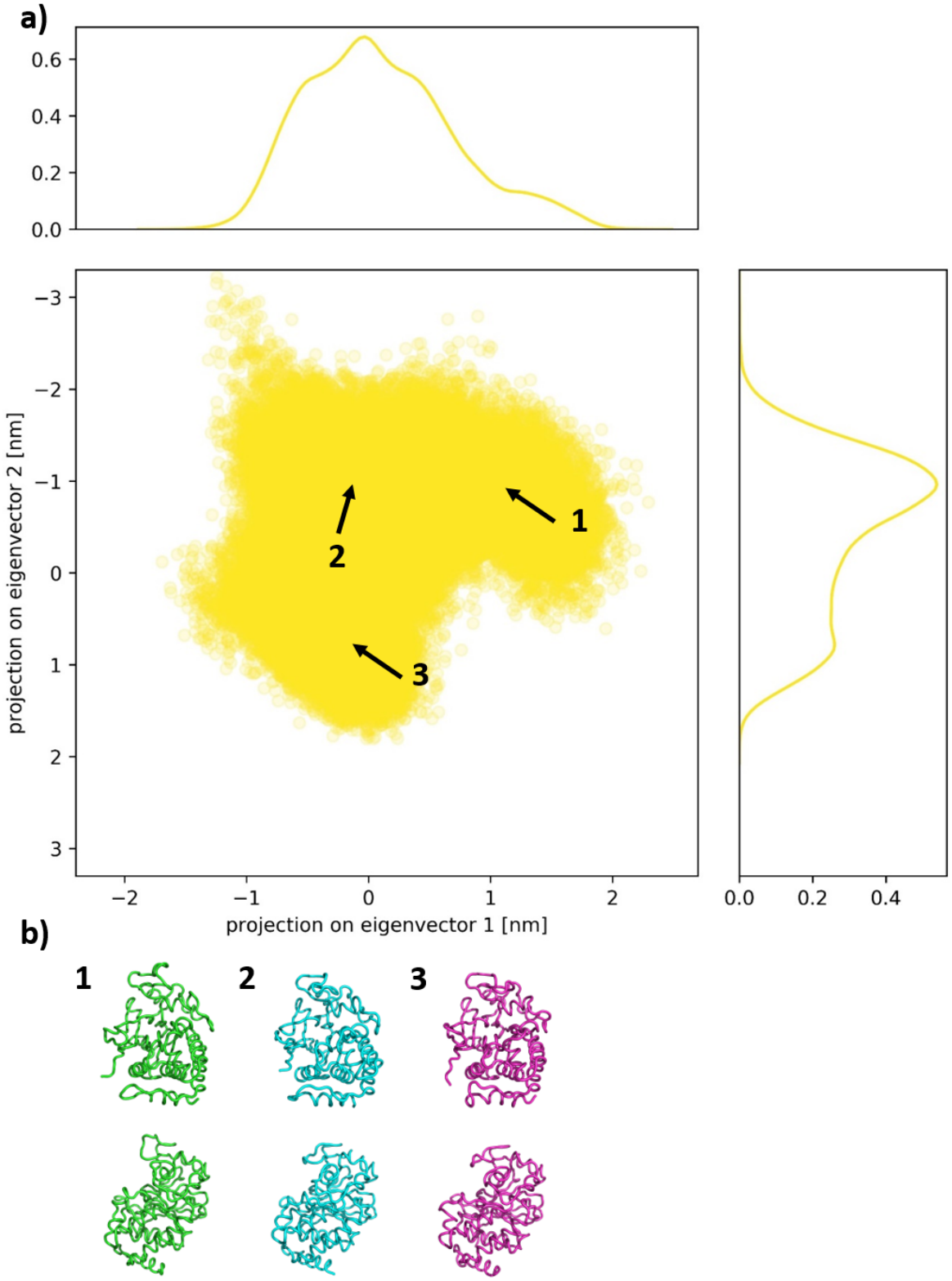


Figure 8.5: Principal component analysis of PPP1-[*Enantio*-Adda5]MC-LF MD Simulation. a) The 2D projection of the PPP1 backbone structure is shown in the central scatter plot. A distribution line showing the frequency of occurrence is shown at the top and left. The numbers and arrows in the scatter plot point to the position of b) respective 3D structures of PPP1.

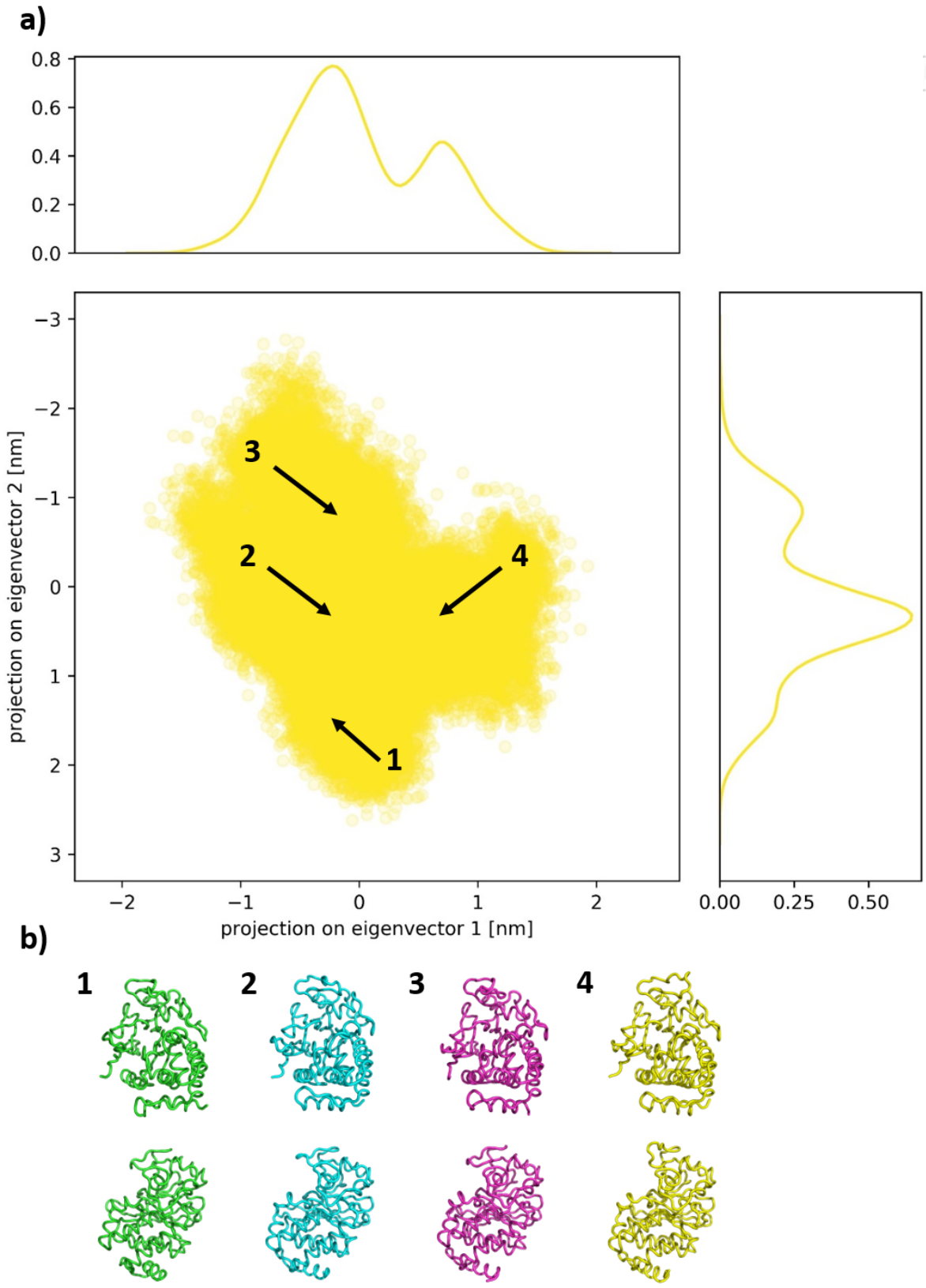


Figure 8.6: Principal component analysis of PPP1-[β-D-Asp3, Dhb7]-MC-RR MD Simulation. a) The 2D projection of the PPP1 backbone structure is shown in the central scatter plot. A distribution line showing the frequency of occurrence is shown at the top and left. The numbers and arrows in the scatter plot point to the position of b) respective 3D structures of PPP1.

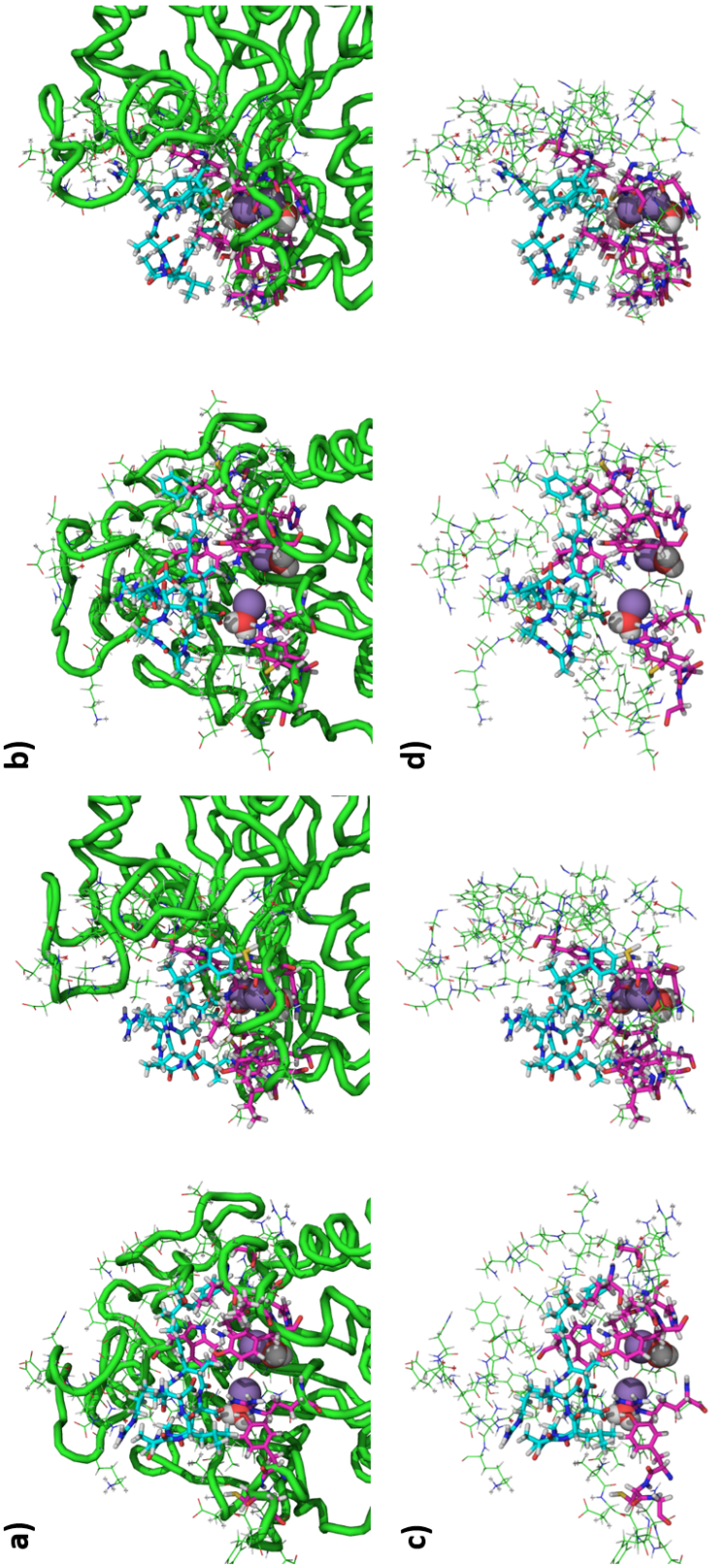


Figure 8.7: Interaction of MC-LR (cyan) with PPP1 (green). Known interacting residues are displayed in pink, interacting residues as lines in green. Mn^{2+} and the active site water is shown as spheres. Snapshots of a) PPP1 (cartoon representation) and c) interacting residues at the beginning, and b) PPP1 (cartoon representation) and d) interacting residues at the end of the third MD simulation.

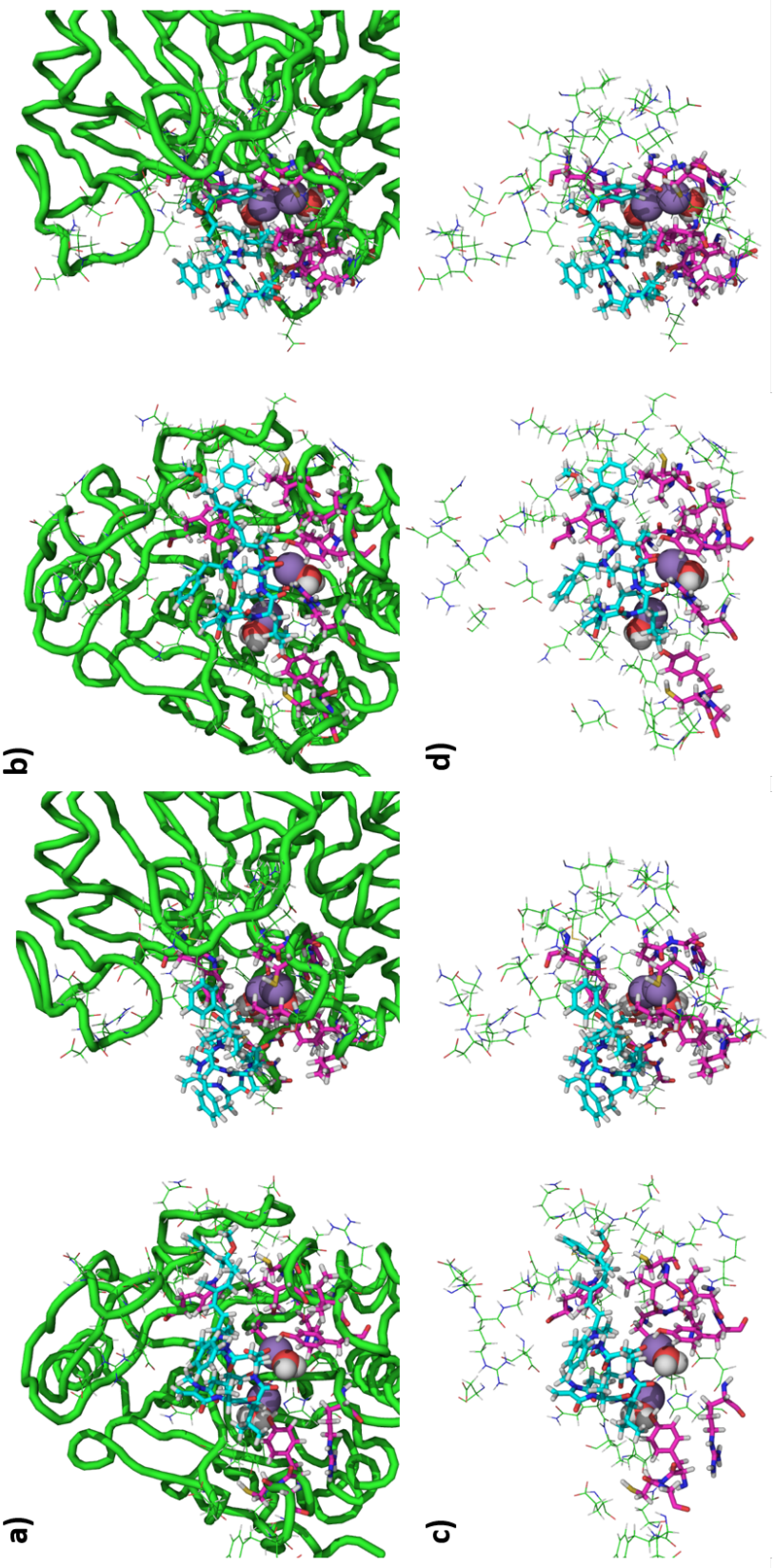


Figure 8.8: Interaction of MC-LF (cyan) with PPP1 (green). Known interacting residues are displayed in pink, interacting residues as lines in green. Mn^{2+} and the active site water is shown as spheres. Snapshots of a) PPP1 (cartoon representation) and c) interacting residues at the beginning, and b) PPP1 (cartoon representation) and d) interacting residues at the end of the third MD simulation.

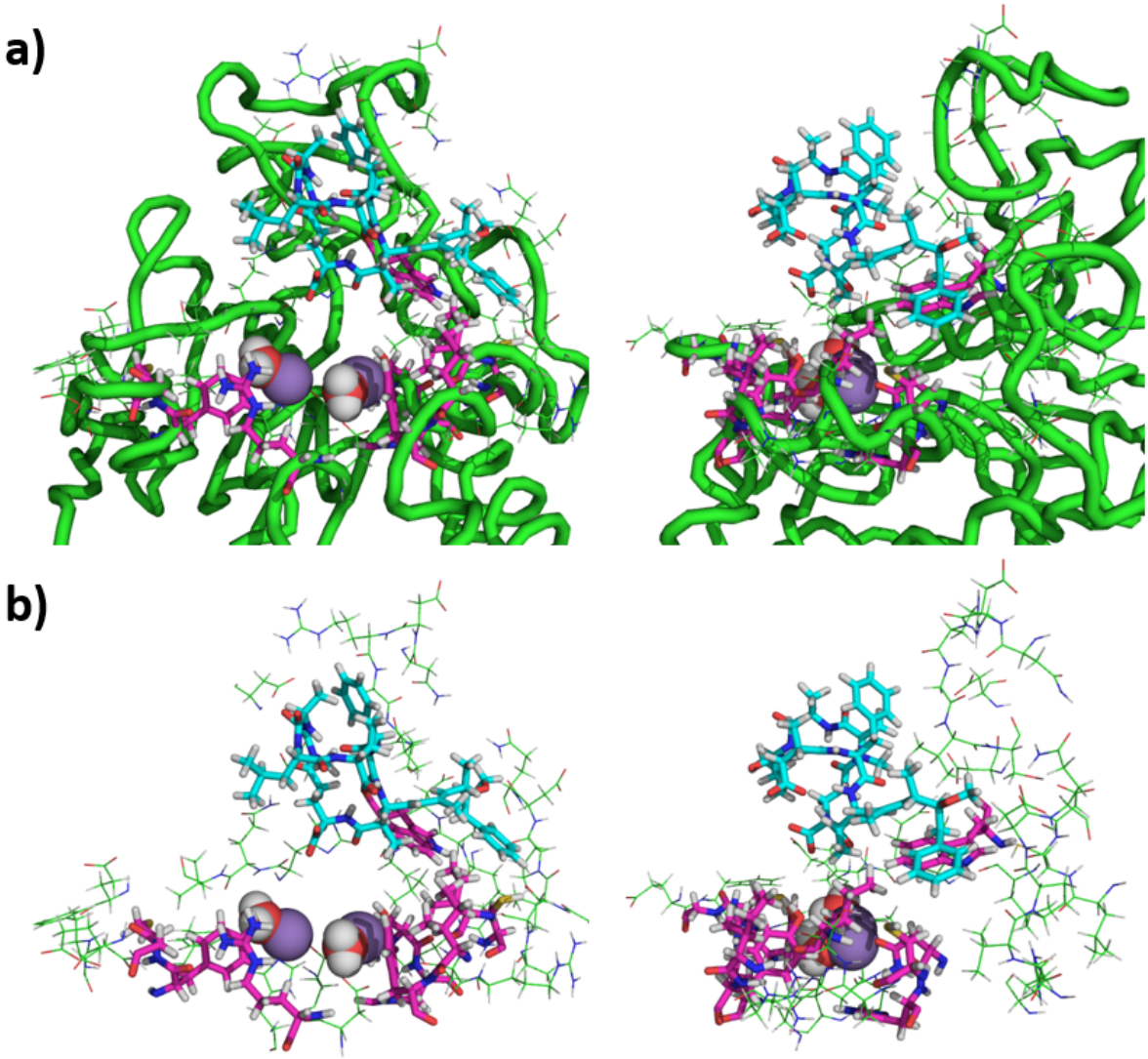


Figure 8.9: Interaction of MC-LF (cyan) with PPP1 (green) during MD simulation. Known interacting residues are displayed in pink, interacting residues as lines in green. Mn^{2+} and the active site water is shown as spheres. Even though MC-LF is toxic, the ligand is not as well coordinated into the binding site as MC-LR and moves up and down. a) PPP1 (cartoon representation) and b) interacting residues are shown.

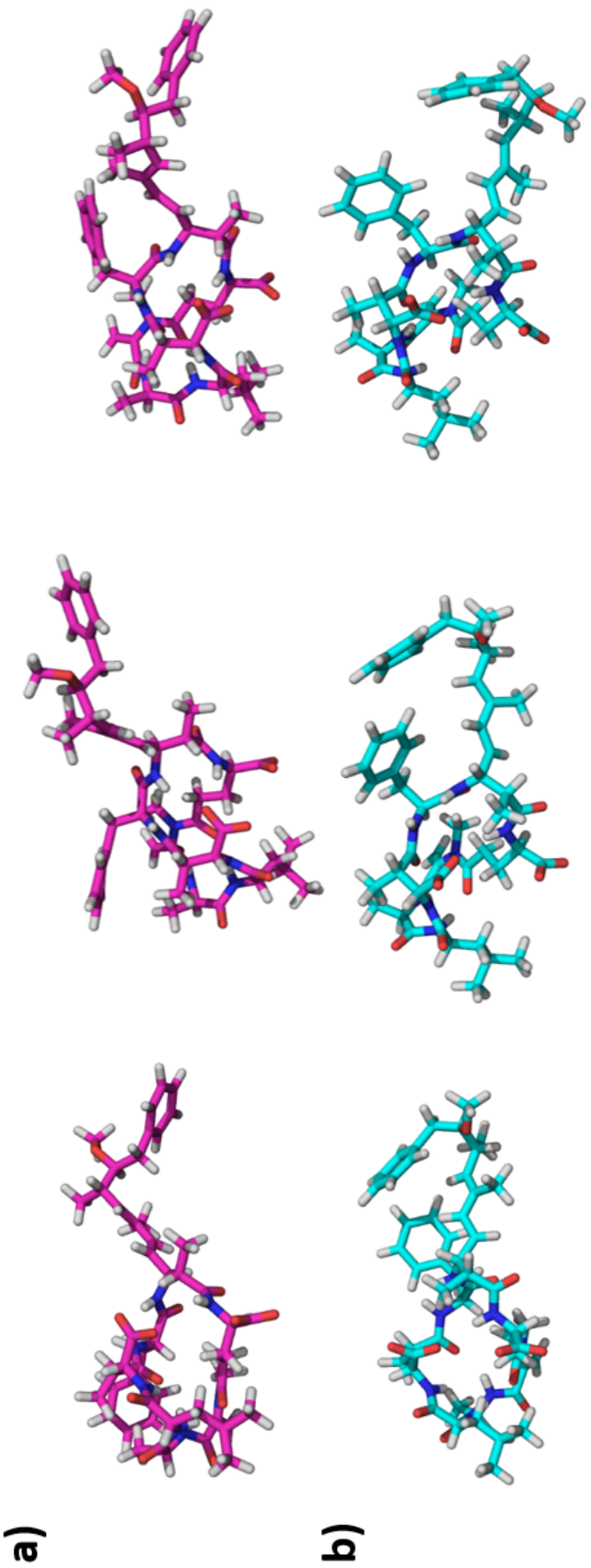


Figure 8.10: MC-congener of MC-LF and [Enantio-Adda5]MC-LF. The difference between both structures is that the stereocenters at the Adda residue are flipped. This leads to a different orientation of the Adda residue of a) MC-LF (pink) and b) [Enantio-Adda5]MC-LF (cyan).

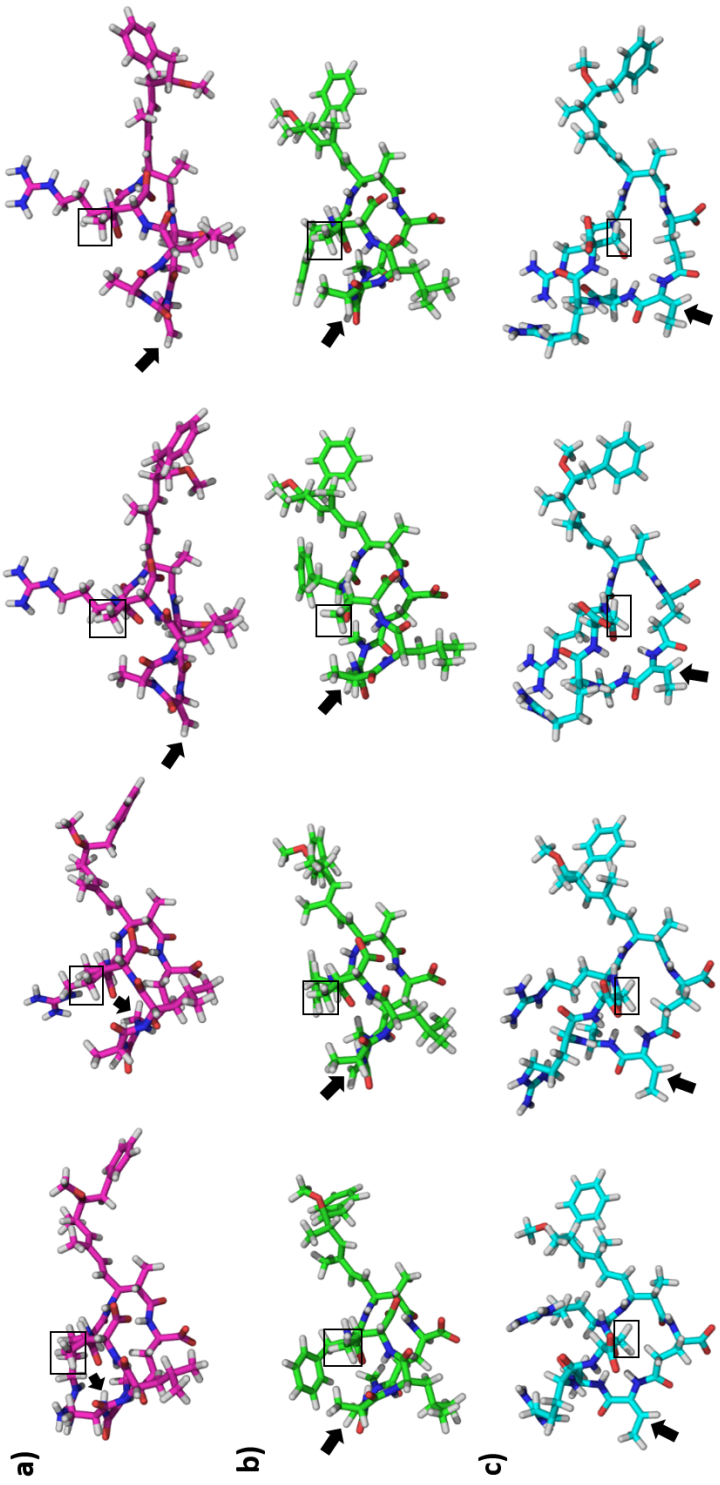


Figure 8.11: Snapshots of MD simulation in position 1 of MC-congener MC-LR, MC-LF and [β-D-Asp3,Dhb7]MC-RR. a) MC-LR (pink), b) MC-LF (green) and c) [β-D-Asp3,Dhb7]MC-RR (cyan) is in contrast to MC-LR and MC-LF modified at two positions: 1) at residue 3 a hydrogen atom instead of a methyl group is attached as rest (marked with black square), and 2) at residue 7 a methyl group is attached to the unsaturated carbon atom (marked with black arrow). This leads to a more open backbone conformation compared to MC-LR and MC-LF.

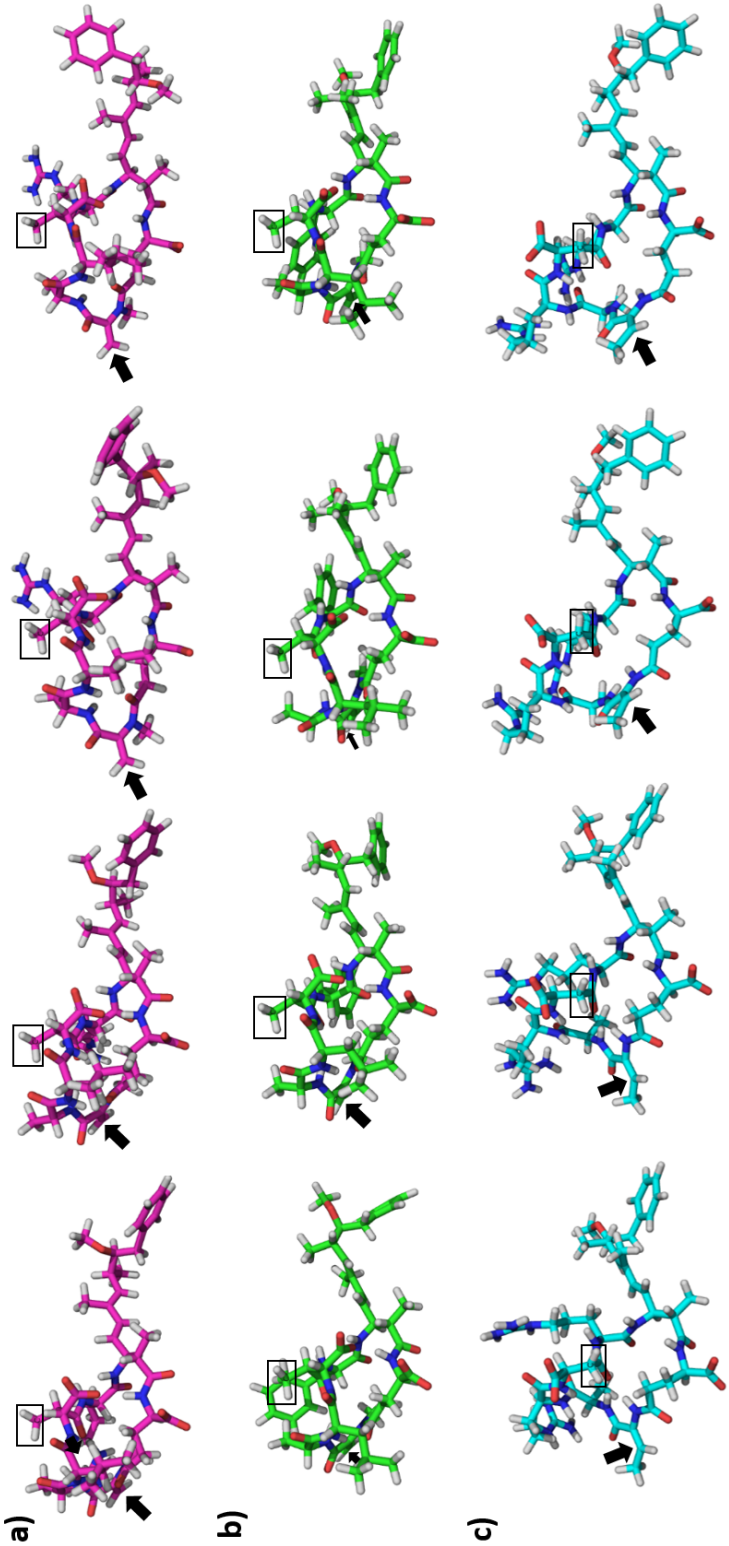


Figure 8.12: Snapshots of MD simulation in position 2 of MC-congener MC-LR, MC-LF and [β-D-Asp3,Dhb7]MC-RR. a) MC-LR (pink), b) MC-LF (green) and c) [β-D-Asp3,Dhb7]MC-RR (cyan) is in contrast to MC-LR and MC-LF modified at two positions: 1) at residue 3 a hydrogen atom instead of a methyl group is attached as rest (marked with black square), and 2) at residue 7 a methyl group is attached to the unsaturated carbon atom (marked with black arrow). This leads to a more open backbone conformation compared to MC-LR and MC-LF.

Table 8.2: Summary of volume ($\frac{nm}{S^3/N}$), radius of gyration (nm), solvent accessible surface area (SASA, $\frac{nm}{S^2/N}$) and root mean squared deviation (RMSD, nm) of ligand-complex and apo-simulations.

Simulation	Property	Mean ± Std	Median	IQR[1]
PPP1	Volume	59.12 ± 0.54	59.12	58.75 - 59.48
PPP1-MC-LR	(Protein)	59.17 ± 0.55	59.16	58.79 - 59.53
PPP1-MC-LF		59.07 ± 0.64	59.06	58.63 - 59.50
PPP1-[Enantio-Adda5]MC-LF		59.25 ± 0.52	59.25	58.90 - 59.60
PPP1-[β-D-Asp3,Dhb7]MC-RR		59.31 ± 0.53	59.31	58.95 - 59.66
PPP1	Radius of	1.87 ± 0.01	1.87	1.86 - 1.87
PPP1-MC-LR	Gyration	1.87 ± 0.01	1.87	1.86 - 1.88
PPP1-MC-LF	(Protein)	1.87 ± 0.01	1.87	1.86 - 1.87
PPP1-[Enantio-Adda5]MC-LF		1.87 ± 0.01	1.87	1.87 - 1.88
PPP1-[β-D-Asp3,Dhb7]MC-RR		1.87 ± 0.01	1.87	1.86 - 1.88
PPP1	SASA	135.59 ± 2.37	135.63	134.02 - 137.18
PPP1-MC-LR	(Protein)	135.97 ± 2.30	135.91	134.35 - 137.54
PPP1-MC-LF		135.59 ± 3.01	135.34	133.43 - 137.58
PPP1-[Enantio-Adda5]MC-LF		136.59 ± 2.20	136.59	135.13 - 138.05
PPP1-[β-D-Asp3,Dhb7]MC-RR		136.60 ± 2.57	136.61	134.89 - 138.33
PPP1	RMSD	0.15 ± 0.03	0.16	0.13 - 0.17
PPP1-MC-LR	(Protein	0.15 ± 0.05	0.13	0.11 - 0.20
PPP1-MC-LF	backbone)	0.14 ± 0.02	0.13	0.12 - 0.16
PPP1-[Enantio-Adda5]MC-LF		0.16 ± 0.04	0.14	0.12 - 0.19
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.17 ± 0.03	0.16	0.14 - 0.20
PPP1-MC-LR	RMSD	0.22 ± 0.04	0.21	0.19 - 0.26
PPP1-MC-LF	(Protein,ions,	0.21 ± 0.02	0.21	0.20 - 0.22
PPP1-[Enantio-Adda5]MC-LF	active site	0.28 ± 0.06	0.27	0.24 - 0.30
PPP1-[β-D-Asp3,Dhb7]MC-RR	water, ligand)	0.24 ± 0.03	0.25	0.22 - 0.27

[1] IQR = Interquartile Range;

Table 8.3: Summary of statistical results for volume, radius of gyration, solvent accessible surface area (SASA) and root mean squared deviation (RMSD) of PPP1 for apo and complex simulation. Linear regression was applied and slope and interception were calculated within a 95 % confidence interval. The lower and upper bound is referred to as low and up, respectively.

Simulation	Property	slope	slope low	slope up	intercept	intercept low	intercept up
PPP1	Volume	0.000	0.000	0.000	59.139	59.132	59.147
PPP1-MC-LR	(Protein)	0.000	0.000	0.000	59.052	59.044	59.060
PPP1-MC-LF		0.000	0.000	0.000	59.066	59.056	59.075
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	59.323	59.316	59.331
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	59.245	59.237	59.252
PPP1	Radius of	0.000	0.000	0.000	1.865	1.864	1.865
PPP1-MC-LR	Gyration	0.000	0.000	0.000	1.862	1.862	1.862
PPP1-MC-LF	(Protein)	0.000	0.000	0.000	1.864	1.864	1.864
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	1.872	1.871	1.872
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	1.870	1.870	1.870
PPP1	SASA	0.000	0.000	0.000	135.472	135.439	135.506
PPP1-MC-LR	(Protein)	0.000	0.000	0.000	135.120	135.088	135.152
PPP1-MC-LF		0.000	0.000	0.000	135.574	135.531	135.618
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	136.428	136.396	136.459
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	135.764	135.728	135.800
PPP1	RMSD	0.000	0.000	0.000	0.134	0.133	0.134
PPP1-MC-LR	(Protein	0.000	0.000	0.000	0.124	0.123	0.124
PPP1-MC-LF	backbone)	0.000	0.000	0.000	0.127	0.127	0.128
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	0.146	0.145	0.147
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	0.153	0.153	0.154
PPP1-MC-LR	RMSD	0.000	0.000	0.000	0.196	0.196	0.197
PPP1-MC-LF	(Protein,ions,	0.000	0.000	0.000	0.200	0.200	0.200
PPP1-[Enantio-Adda5]MC-LF	active site	0.000	0.000	0.000	0.295	0.294	0.296
PPP1-[β-D-Asp3,Dhb7]MC-RR	water, ligand)	0.000	0.000	0.000	0.227	0.226	0.227

Table 8.4: Summary of volume ($\frac{nm}{S^3/N}$), radius of gyration (nm), solvent accessible surface area (SASA, $\frac{nm}{S^2/N}$) and root mean squared deviation (RMSD, nm) of MC backbone of solvent simulation in comparison to complex simulation.

Simulation	Property	Mean ± Std	Median	IQR[1]
MC-LR	Volume	2.54 ± 0.08	2.54	2.49 - 2.59
MC-LF		2.43 ± 0.07	2.43	2.39 - 2.48
[Enantio-Adda5]MC-LF		2.50 ± 0.06	2.50	2.45 - 2.54
[β-D-Asp3,Dhb7]MC-RR		2.54 ± 0.08	2.54	2.49 - 2.60
PPP1-MC-LR	Volume	2.53 ± 0.07	2.53	2.48 - 2.58
PPP1-MC-LF		2.52 ± 0.08	2.52	2.46 - 2.57
PPP1-[Enantio-Adda5]MC-LF		2.47 ± 0.07	2.47	2.42 - 2.51
PPP1-[β-D-Asp3,Dhb7]MC-RR		2.62 ± 0.07	2.62	2.58 - 2.67
MC-LR	Radius of	0.61 ± 0.03	0.61	0.59 - 0.63
MC-LF	Gyration	0.61 ± 0.03	0.62	0.60 - 0.63
[Enantio-Adda5]MC-LF		0.63 ± 0.03	0.63	0.61 - 0.65
[β-D-Asp3,Dhb7]MC-RR		0.63 ± 0.03	0.63	0.61 - 0.65
PPP1-MC-LR	Radius of	0.65 ± 0.02	0.65	0.64 - 0.66
PPP1-MC-LF	Gyration	0.65 ± 0.02	0.65	0.63 - 0.67
PPP1-[Enantio-Adda5]MC-LF		0.64 ± 0.03	0.64	0.63 - 0.66
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.68 ± 0.02	0.68	0.67 - 0.69
MC-LR	SASA	12.53 ± 0.57	12.52	12.17 - 12.87
MC-LF		11.91 ± 0.43	11.92	11.64 - 12.20
[Enantio-Adda5]MC-LF		12.28 ± 0.43	12.29	12.00 - 12.57
[β-D-Asp3,Dhb7]MC-RR		12.61 ± 0.57	12.57	12.20 - 13.00
PPP1-MC-LR	SASA	12.80 ± 0.61	12.85	12.34 - 13.28
PPP1-MC-LF		12.62 ± 0.50	12.65	12.25 - 12.99
PPP1-[Enantio-Adda5]MC-LF		12.11 ± 0.47	12.12	11.78 - 12.45
PPP1-[β-D-Asp3,Dhb7]MC-RR		13.40 ± 0.49	13.42	13.07 - 13.47
MC-LR	RMSD	0.10 ± 0.06	0.08	0.05 - 0.16
MC-LF	Backbone	0.06 ± 0.03	0.04	0.03 - 0.09
[Enantio-Adda5]MC-LF		0.08 ± 0.03	0.07	0.05 - 0.11
[β-D-Asp3,Dhb7]MC-RR		0.08 ± 0.04	0.06	0.04 - 0.11
MC-LR	RMSD	0.26 ± 0.08	0.27	0.18 - 0.32
MC-LF		0.24 ± 0.06	0.24	0.19 - 0.29
[Enantio-Adda5]MC-LF		0.24 ± 0.05	0.25	0.21 - 0.28
[β-D-Asp3,Dhb7]MC-RR		0.30 ± 0.06	0.30	0.25 - 0.34
PPP1-MC-LR	RMSD	0.24 ± 0.05	0.23	0.20 - 0.26
PPP1-MC-LF		0.29 ± 0.04	0.29	0.26 - 0.32
PPP1-[Enantio-Adda5]MC-LF		0.29 ± 0.04	0.29	0.27 - 0.32
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.33 ± 0.04	0.32	0.30 - 0.36

[1] IQR = Interquartile Range;

Table 8.5: Summary of statistical results for volume, radius of gyration, solvent accessible surface area (SASA) and root mean squared deviation (RMSD) for MC congener of solvent and complex simulation. Linear regression was applied and slope and interception were calculated within a 95 % confidence interval. The lower and upper bound is referred to as low and up, respectively.

Simulation	Property	slope	slope low	slope up	intercept	intercept low	intercept up
MC-LR	Volume	0.000	0.000	0.000	2.531	2.526	2.535
MC-LF		0.000	0.000	0.000	2.442	2.439	2.446
[Enantio-Adda5]MC-LF		0.000	0.000	0.000	2.492	2.489	2.496
[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	2.545	2.541	2.550
PPP1-MC-LR	Volume	0.000	0.000	0.000	2.530	2.529	2.531
PPP1-MC-LF		0.000	0.000	0.000	2.511	2.510	2.512
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	2.456	2.455	2.457
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	2.631	2.630	2.632
MC-LR	Radius of	0.000	0.000	0.000	0.611	0.610	0.613
MC-LF	Gyration	0.000	0.000	0.000	0.616	0.614	0.617
[Enantio-Adda5]MC-LF		0.000	0.000	0.000	0.626	0.624	0.628
[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	0.626	0.624	0.627
PPP1-MC-LR	Radius of	0.000	0.000	0.000	0.650	0.649	0.650
PPP1-MC-LF	Gyration	0.000	0.000	0.000	0.656	0.656	0.657
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	0.631	0.630	0.631
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	0.673	0.672	0.673
MC-LR	SASA	0.000	0.000	0.000	12.488	12.455	12.522
MC-LF		0.000	0.000	0.000	11.964	11.939	11.989
[Enantio-Adda5]MC-LF		0.000	0.000	0.000	12.246	12.221	12.271
[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	12.643	12.610	12.676
PPP1-MC-LR	SASA	0.000	0.000	0.000	12.761	12.752	12.769
PPP1-MC-LF		0.000	0.000	0.000	12.603	12.596	12.610
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	11.985	11.979	11.992
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	13.452	13.445	13.459
MC-LR	RMSD	0.007	0.007	0.008	0.201	0.197	0.205
MC-LF		0.000	0.000	0.001	0.239	0.235	0.242
[Enantio-Adda5]MC-LF		0.003	0.002	0.003	0.223	0.220	0.226
[β-D-Asp3,Dhb7]MC-RR		0.004	0.004	0.005	0.266	0.262	0.269
PPP1-MC-LR	RMSD	0.000	0.000	0.000	0.209	0.209	0.210
PPP1-MC-LF		0.000	0.000	0.000	0.283	0.282	0.283
PPP1-[Enantio-Adda5]MC-LF		0.000	0.000	0.000	0.288	0.287	0.288
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	0.328	0.327	0.328

Table 8.6: Summary of additional calculated properties for number of contacts of ligand-complex simulations.

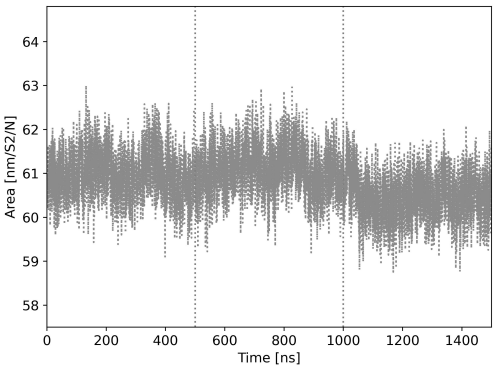
Simulation	Property	Mean ± Std	Median	IQR[1]
PPP1-MC-LR	Number of	2583.77 ± 476.08	2533	2256 - 2912
PPP1-MC-LF	Contacts	2202.55 ± 370.54	2252	1994 - 2459
PPP1-[Enantio-Adda5]MC-LF	(Protein,	1863.62 ± 455.63	1913	1615 - 2183
PPP1-[β-D-Asp3,Dhb7]MC-RR	ligand)	2228.15 ± 335.82	2244	2039 - 2449
PPP1-MC-LR	Number contacts	291.12 ± 155.32	302	181 - 420
PPP1-MC-LF	(Protein,	221.52 ± 152.59	248	77 - 335
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 2	329.50 ± 154.47	366	249 - 434
PPP1-[β-D-Asp3,Dhb7]MC-RR	sidechain))	296.87 ± 153.73	289	196 - 423
PPP1-MC-LR	Number contacts	1100.44 ± 395.12	1231	708 - 1394
PPP1-MC-LF	(Protein,	892.68 ± 231.20	963	687 - 1082
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 5	837.09 ± 339.30	864	622 - 1070
PPP1-[β-D-Asp3,Dhb7]MC-RR	sidechain))	961.18 ± 204.91	983	869 - 1100
PPP1-MC-LR	Number contacts	13.60 ± 4.02	14	11 - 16
PPP1-MC-LF	(Ions,	24.42 ± 8.83	23	18 - 34
PPP1-[Enantio-Adda5]MC-LF	ligand)	7.69 ± 6.15	6	3 - 11
PPP1-[β-D-Asp3,Dhb7]MC-RR		8.05 ± 9.99	1	0 - 20
PPP1-MC-LR	Number contacts	148.06 ± 36.66	146	129 - 173
PPP1-MC-LF	(Protein,	209.48 ± 32.06	214	192 - 232
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 5	128.40 ± 60.53	134	75 - 171
PPP1-[β-D-Asp3,Dhb7]MC-RR	methyl group)	149.67 ± 63.96	161	110 - 193
PPP1-MC-LR	Number contacts	107.69 ± 22.72	111	94 - 125
PPP1-MC-LF	(Protein,	107.83 ± 28.46	114	101 - 128
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 6	56.83 ± 29.86	63	35 - 80
PPP1-[β-D-Asp3,Dhb7]MC-RR	carboxylic acid))	68.10 ± 38.85	66	35 - 106
PPP1-MC-LR	Number contacts	55.56 ± 13.48	59	45 - 66
PPP1-MC-LF	(Protein,	53.13 ± 11.72	55	48 - 61
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 6	32.20 ± 15.15	34	23 - 43
PPP1-[β-D-Asp3,Dhb7]MC-RR	carbonyl group)	38.63 ± 18.49	40	22 - 53
PPP1-MC-LR	Number contacts	25.33 ± 14.16	27	14 - 34
PPP1-MC-LF	(Protein,	41.38 ± 13.72	41	34 - 49
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 3	40.19 ± 21.85	29	27 - 53
PPP1-[β-D-Asp3,Dhb7]MC-RR	carboxylic acid))	27.50 ± 23.36	26	1 - 52
PPP1-MC-LR	Number contacts	49.03 ± 49.18	31	11 - 74
PPP1-MC-LF	(Protein,	76.13 ± 46.11	71	43 - 124
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 7	6.54 ± 18.82	0	0 - 4
PPP1-[β-D-Asp3,Dhb7]MC-RR	methylene group))	62.88 ± 76.97	6	0 - 133

[1] IQR = Interquartile Range;

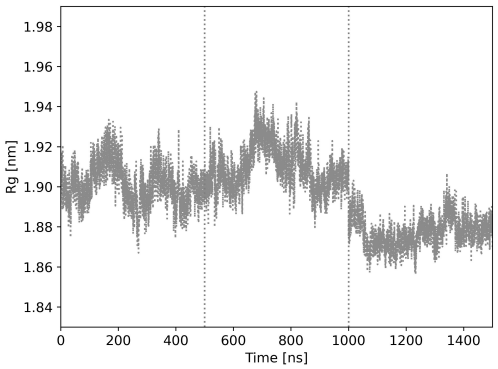
Table 8.7: Summary of statistical results for number of contacts for complex simulation. Linear regression was applied and slope and interception were calculated within a 95 % confidence interval. The lower and upper bound is referred to as low and up, respectively.

Simulation	Property	slope	slope low	slope up	intercept	intercept low	intercept up
PPP1-MC-LR	Number of	0.000	0.000	0.000	2571.148	2564.334	2577.961
PPP1-MC-LF	Contacts	0.001	0.001	0.001	2059.570	2054.400	2064.740
PPP1-[Enantio-Adda5]MC-LF	(Protein,	0.003	0.003	0.003	1519.285	1513.417	1525.153
PPP1-[β-D-Asp3,Dhb7]MC-RR	ligand)	0.000	0.000	0.000	2184.664	2179.871	2189.457
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	327.270	325.068	329.473
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	161.413	159.286	163.540
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 2	0.001	0.001	0.001	192.135	190.238	194.033
PPP1-[β-D-Asp3,Dhb7]MC-RR	sidechain))	0.000	0.000	0.000	298.319	296.119	300.519
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	1112.556	1106.901	1118.210
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	920.782	917.480	924.083
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 5	0.001	0.001	0.001	696.541	691.826	701.257
PPP1-[β-D-Asp3,Dhb7]MC-RR	sidechain))	0.000	0.000	0.000	975.073	972.142	978.004
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	13.733	13.675	13.790
PPP1-MC-LF	(Ions,	0.000	0.000	0.000	23.000	22.874	23.126
PPP1-[Enantio-Adda5]MC-LF	ligand)	0.000	0.000	0.000	3.864	3.782	3.946
PPP1-[β-D-Asp3,Dhb7]MC-RR		0.000	0.000	0.000	7.752	7.609	7.895
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	145.663	145.139	146.187
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	188.388	187.964	188.813
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 5	0.000	0.000	0.000	133.112	132.247	133.978
PPP1-[β-D-Asp3,Dhb7]MC-RR	methyl group)	0.000	0.000	0.000	165.042	164.135	165.948
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	110.080	109.755	110.404
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	94.737	94.344	95.129
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 6	0.000	0.000	0.000	50.349	49.925	50.773
PPP1-[β-D-Asp3,Dhb7]MC-RR	carboxylic acid))	0.000	0.000	0.000	59.918	59.366	60.470
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	54.550	54.357	54.743
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	48.044	47.882	48.207
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 6	0.000	0.000	0.000	29.942	29.726	30.158
PPP1-[β-D-Asp3,Dhb7]MC-RR	carbonyl group)	0.000	0.000	0.000	38.444	38.180	38.709
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	31.659	31.463	31.855
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	34.631	34.443	34.819
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 3	0.000	0.000	0.000	38.832	38.519	39.144
PPP1-[β-D-Asp3,Dhb7]MC-RR	carboxylic acid))	0.000	0.000	0.000	24.209	23.875	24.542
PPP1-MC-LR	Number contacts	0.000	0.000	0.000	67.274	66.587	67.962
PPP1-MC-LF	(Protein,	0.000	0.000	0.000	68.845	68.188	69.502
PPP1-[Enantio-Adda5]MC-LF	ligand (res. 7	0.000	0.000	0.000	5.751	5.482	6.020
PPP1-[β-D-Asp3,Dhb7]MC-RR	methylene group))	0.000	0.000	0.000	48.904	47.808	49.999

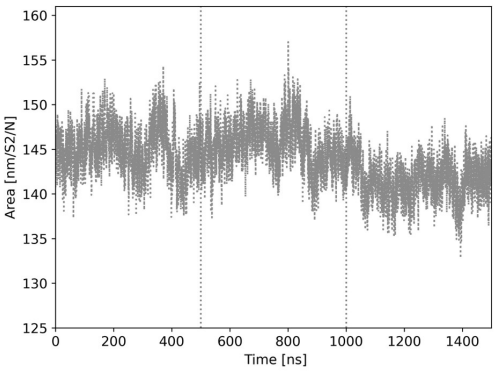
8.2.2 Molecular Dynamics Simulation of 12 MC congeners



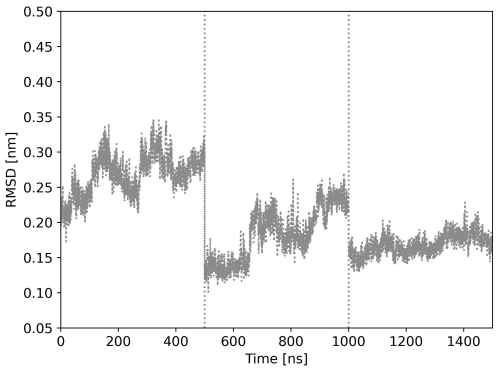
(a) Volume



(b) Radius of gyration



(c) Solvent accessible surface area



(d) Backbone RMSD

Figure 8.13: Time series data of PPP1 in apo simulation. The three replicates are shown as one time line and are separated by a grey dashed line at 500 and 1000 ns.

Table 8.8: PPP1 properties of apo and complex (MC congener and conjugates) simulations. Mean and standard deviations of PPP1 volume, radius of gyration (rgyr), solvent accessible surface area (sasa) and PPP1 backbone root mean squared deviation (RMSD) are shown.

Simulation	Volume $[\frac{nm}{S^3/N}]$	Rgyr $[nm]$	Sasa $[\frac{nm}{S^3/N}]$	RMSD $[nm]$
PPP1	60.83 ± 0.61	1.9 ± 0.02	143.99 ± 3.09	0.21 ± 0.05
PPP1-MC-LR	61.11 ± 0.69	1.9 ± 0.01	144.63 ± 3.73	0.22 ± 0.05
PPP1-MC-LR-Cys-S	60.64 ± 0.63	1.89 ± 0.02	142.58 ± 3.13	0.19 ± 0.03
PPP1-MC-LR-Cys-R	61.03 ± 0.66	1.90 ± 0.02	144.7 ± 3.30	0.19 ± 0.04
PPP1-MC-LR-GSH-S	60.68 ± 0.62	1.89 ± 0.01	142.23 ± 3.45	0.18 ± 0.03
PPP1-MC-LR-GSH-R	60.57 ± 0.61	1.88 ± 0.01	141.61 ± 3.41	0.17 ± 0.03
PPP1-MC-LF	60.83 ± 0.58	1.90 ± 0.01	142.8 ± 2.86	0.20 ± 0.04
PPP1-MC-LF-Cys-S	60.84 ± 0.65	1.89 ± 0.01	143.27 ± 3.22	0.20 ± 0.03
PPP1-MC-LF-Cys-R	61.08 ± 0.72	1.90 ± 0.02	144.97 ± 4.31	0.23 ± 0.07
PPP1-MC-LF-GSH-S	60.88 ± 0.65	1.89 ± 0.01	144.14 ± 3.91	0.21 ± 0.04
PPP1-MC-LF-GSH-R	61.15 ± 0.55	1.90 ± 0.01	145.00 ± 2.66	0.18 ± 0.03
PPP1-[Enantio-Adda5]MC-LF	60.98 ± 0.61	1.90 ± 0.01	144.3 ± 2.95	0.20 ± 0.06
PPP1-[Enantio-Adda5]MC-LF-Cys-S	60.34 ± 0.73	1.88 ± 0.02	140.43 ± 3.51	0.19 ± 0.03
PPP1-[Enantio-Adda5]MC-LF-Cys-R	60.78 ± 0.61	1.90 ± 0.02	143.13 ± 3.30	0.23 ± 0.05
PPP1-[Enantio-Adda5]MC-LF-GSH-S	60.84 ± 0.64	1.90 ± 0.01	143.46 ± 3.17	0.21 ± 0.04
PPP1-[Enantio-Adda5]MC-LF-GSH-R	61.18 ± 0.62	1.91 ± 0.01	145.53 ± 3.32	0.21 ± 0.05
PPP1-[β-D-Asp3,Dhb7]MC-RR	60.95 ± 0.59	1.90 ± 0.01	143.82 ± 3.26	0.17 ± 0.04
PPP1-MC-RR	60.89 ± 0.57	1.89 ± 0.02	144.17 ± 2.71	0.18 ± 0.02
PPP1-MC-RR-Cys-S	60.88 ± 0.62	1.90 ± 0.02	143.88 ± 3.87	0.20 ± 0.04
PPP1-MC-RR-Cys-R	60.66 ± 0.58	1.89 ± 0.01	142.58 ± 2.64	0.20 ± 0.04
PPP1-MC-RR-GSH-S	60.91 ± 0.58	1.90 ± 0.02	143.50 ± 3.11	0.18 ± 0.03
PPP1-MC-RR-GSH-R	60.78 ± 0.61	1.88 ± 0.01	143.27 ± 2.94	0.24 ± 0.06
PPP1-MC-LY	61.02 ± 0.59	1.91 ± 0.01	144.43 ± 2.85	0.20 ± 0.08
PPP1-MC-LY-Cys-S	60.76 ± 0.70	1.90 ± 0.02	143.35 ± 4.17	0.20 ± 0.04
PPP1-MC-LY-Cys-R	60.91 ± 0.64	1.90 ± 0.02	143.53 ± 3.85	0.19 ± 0.06
PPP1-MC-LY-GSH-S	60.86 ± 0.59	1.90 ± 0.01	143.90 ± 2.90	0.18 ± 0.06
PPP1-MC-LY-GSH-R	60.8 ± 0.74	1.89 ± 0.02	143.70 ± 4.55	0.18 ± 0.03
PPP1-MC-YR	60.94 ± 0.61	1.90 ± 0.02	144.29 ± 3.44	0.19 ± 0.03
PPP1-MC-YR-Cys-S	61.05 ± 0.58	1.91 ± 0.01	145.30 ± 3.26	0.20 ± 0.05
PPP1-MC-YR-Cys-R	60.93 ± 0.65	1.90 ± 0.01	144.09 ± 3.63	0.22 ± 0.04
PPP1-MC-YR-GSH-S	60.54 ± 0.72	1.88 ± 0.02	141.33 ± 4.45	0.17 ± 0.03
PPP1-MC-YR-GSH-R	60.85 ± 0.59	1.89 ± 0.01	143.53 ± 2.85	0.21 ± 0.05
PPP1-MC-LY(Prg)	60.8 ± 0.72	1.89 ± 0.02	143.25 ± 3.96	0.19 ± 0.04
PPP1-MC-LY(Prg)-Cys-S	60.96 ± 0.67	1.9 ± 0.02	144.91 ± 3.29	0.20 ± 0.03
PPP1-MC-LY(Prg)-Cys-R	61.07 ± 0.59	1.91 ± 0.02	145.44 ± 3.10	0.25 ± 0.07
PPP1-MC-LY(Prg)-GSH-S	60.81 ± 0.71	1.90 ± 0.02	143.58 ± 3.89	0.20 ± 0.04
PPP1-MC-LY(Prg)-GSH-R	60.71 ± 0.74	1.89 ± 0.02	142.33 ± 3.61	0.18 ± 0.04
PPP1-[Anda5]-MC-LY(Prg)	61.11 ± 0.62	1.91 ± 0.02	146.06 ± 3.39	0.21 ± 0.06
PPP1-[Anda5]-MC-LY(Prg)-Cys-S	60.99 ± 0.60	1.89 ± 0.01	144.08 ± 3.17	0.20 ± 0.04
PPP1-[Anda5]-MC-LY(Prg)-Cys-R	60.80 ± 0.57	1.90 ± 0.02	143.89 ± 2.84	0.19 ± 0.07
PPP1-[Anda5]-MC-LY(Prg)-GSH-S	60.85 ± 0.75	1.89 ± 0.01	143.68 ± 3.85	0.21 ± 0.05
PPP1-[Anda5]-MC-LY(Prg)-GSH-R	61.18 ± 0.65	1.90 ± 0.01	146.15 ± 3.44	0.20 ± 0.06

Table 8.8: Continued from previous.

Simulation	Volume $[\frac{nm}{S^3/N}]$	Rgyr $[nm]$	Sasa $[\frac{nm}{S^3/N}]$	RMSD $[nm]$
PPP1-[Amba5]-MC-LY(Prg)	60.84 ± 0.57	1.90 ± 0.01	143.15 ± 2.93	0.21 ± 0.05
PPP1-[Amba5]-MC-LY(Prg)-Cys-S	60.71 ± 0.61	1.89 ± 0.02	142.59 ± 3.13	0.18 ± 0.03
PPP1-[Amba5]-MC-LY(Prg)-Cys-R	60.81 ± 0.59	1.89 ± 0.01	143.60 ± 3.09	0.20 ± 0.06
PPP1-[Amba5]-MC-LY(Prg)-GSH-S	60.87 ± 0.58	1.90 ± 0.01	143.68 ± 2.90	0.18 ± 0.03
PPP1-[Amba5]-MC-LY(Prg)-GSH-R	61.16 ± 0.58	1.90 ± 0.01	145.72 ± 3.31	0.21 ± 0.03
PPP1-[Apha5]-MC-LF	60.85 ± 0.66	1.90 ± 0.02	143.47 ± 3.74	0.18 ± 0.04
PPP1-[Apha5]-MC-LF-Cys-S	61.09 ± 0.58	1.91 ± 0.01	145.30 ± 3.29	0.21 ± 0.05
PPP1-[Apha5]-MC-LF-Cys-R	60.81 ± 0.65	1.89 ± 0.01	143.00 ± 3.45	0.19 ± 0.02
PPP1-[Apha5]-MC-LF-GSH-S	60.85 ± 0.61	1.90 ± 0.02	144.24 ± 3.46	0.19 ± 0.04
PPP1-[Apha5]-MC-LF-GSH-R	61.18 ± 0.54	1.91 ± 0.01	145.87 ± 2.58	0.21 ± 0.05
PPP1-[Apda5]-MC-LF	60.94 ± 0.66	1.90 ± 0.02	144.27 ± 4.36	0.19 ± 0.03
PPP1-[Apda5]-MC-LF-Cys-S	60.97 ± 0.59	1.89 ± 0.01	144.66 ± 2.97	0.16 ± 0.04
PPP1-[Apda5]-MC-LF-Cys-R	60.91 ± 0.65	1.90 ± 0.02	143.98 ± 3.63	0.19 ± 0.04
PPP1-[Apda5]-MC-LF-GSH-S	61.21 ± 0.63	1.91 ± 0.02	145.89 ± 3.11	0.25 ± 0.05
PPP1-[Apda5]-MC-LF-GSH-R	61.12 ± 0.57	1.91 ± 0.01	145.19 ± 2.82	0.22 ± 0.04

Table 8.9: MC congener and conjugates properties of solvent and complex simulations. Mean and standard deviations of MC congener and conjugate volume, radius of gyration (rgyr), solvent accessible surface area (sasa), and root mean squared deviation (RMSD) are shown.

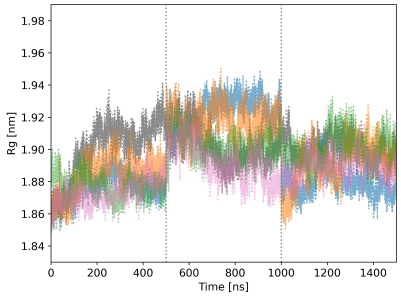
Simulation	Volume $[\frac{nm}{S^3/N}]$	Rgyr $[nm]$	Sasa $(\frac{nm}{S^3/N})$	RMSD $[nm]$
MC-LR	2.54 ± 0.07	0.62 ± 0.02	12.51 ± 0.48	0.30 ± 0.08
PPP1-MC-LR	2.55 ± 0.07	0.64 ± 0.03	12.69 ± 0.47	0.28 ± 0.08
MC-LR-Cys-S	2.80 ± 0.08	0.66 ± 0.03	13.73 ± 0.67	0.27 ± 0.06
PPP1-MC-LR-Cys-S	2.89 ± 0.07	0.71 ± 0.02	14.56 ± 0.49	0.38 ± 0.09
MC-LR-Cys-R	2.83 ± 0.07	0.67 ± 0.03	14.05 ± 0.54	0.30 ± 0.11
PPP1-MC-LR-Cys-R	2.88 ± 0.07	0.68 ± 0.03	14.39 ± 0.49	0.38 ± 0.07
MC-LR-GSH-S	3.28 ± 0.10	0.73 ± 0.05	15.97 ± 0.83	0.33 ± 0.09
PPP1-MC-LR-GSH-S	3.30 ± 0.11	0.73 ± 0.05	16.27 ± 0.98	0.36 ± 0.08
MC-LR-GSH-R	3.36 ± 0.09	0.78 ± 0.05	16.68 ± 0.72	0.52 ± 0.08
PPP1-MC-LR-GSH-R	3.37 ± 0.09	0.75 ± 0.04	16.73 ± 0.72	0.49 ± 0.12
MC-LF	2.49 ± 0.06	0.62 ± 0.03	12.24 ± 0.37	0.24 ± 0.06
PPP1-MC-LF	2.49 ± 0.06	0.63 ± 0.03	12.28 ± 0.41	0.23 ± 0.04
MC-LF-Cys-S	2.79 ± 0.08	0.67 ± 0.03	13.68 ± 0.60	0.28 ± 0.06
PPP1-MC-LF-Cys-S	2.83 ± 0.08	0.69 ± 0.03	14.04 ± 0.60	0.33 ± 0.10
MC-LF-Cys-R	2.79 ± 0.07	0.66 ± 0.02	13.57 ± 0.53	0.28 ± 0.07
PPP1-MC-LF-Cys-R	2.79 ± 0.07	0.64 ± 0.02	13.63 ± 0.51	0.39 ± 0.07
MC-LF-GSH-S	3.18 ± 0.08	0.71 ± 0.03	15.43 ± 0.63	0.33 ± 0.08
PPP1-MC-LF-GSH-S	3.21 ± 0.08	0.72 ± 0.04	15.71 ± 0.59	0.43 ± 0.16
MC-LF-GSH-R	3.29 ± 0.08	0.75 ± 0.03	16.54 ± 0.64	0.40 ± 0.06
PPP1-MC-LF-GSH-R	3.29 ± 0.09	0.75 ± 0.03	16.57 ± 0.75	0.43 ± 0.05
[Enantio-Adda5]MC-LF	2.51 ± 0.06	0.62 ± 0.03	12.20 ± 0.39	0.26 ± 0.07
PPP1-[Enantio-Adda5]MC-LF	2.54 ± 0.07	0.62 ± 0.03	12.49 ± 0.56	0.38 ± 0.08
[Enantio-Adda5]MC-LF-Cys-S	2.78 ± 0.07	0.63 ± 0.03	13.49 ± 0.57	0.40 ± 0.08
PPP1-[Enantio-Adda5]MC-LF-Cys-S	2.81 ± 0.08	0.66 ± 0.03	13.80 ± 0.61	0.43 ± 0.08
[Enantio-Adda5]MC-LF-Cys-R	2.82 ± 0.07	0.66 ± 0.03	13.89 ± 0.50	0.31 ± 0.12
PPP1-[Enantio-Adda5]MC-LF-Cys-R	2.84 ± 0.07	0.67 ± 0.03	14.06 ± 0.58	0.41 ± 0.06
[Enantio-Adda5]MC-LF-GSH-S	3.27 ± 0.08	0.74 ± 0.04	16.23 ± 0.65	0.43 ± 0.08
PPP1-[Enantio-Adda5]MC-LF-GSH-S	3.22 ± 0.08	0.71 ± 0.04	15.77 ± 0.68	0.49 ± 0.05
[Enantio-Adda5]MC-LF-GSH-R	3.27 ± 0.09	0.75 ± 0.04	16.12 ± 0.83	0.41 ± 0.09
PPP1-[Enantio-Adda5]MC-LF-GSH-R	3.26 ± 0.09	0.74 ± 0.04	16.03 ± 0.75	0.51 ± 0.07
[β-D-Asp3,Dhb7]MC-RR	2.62 ± 0.08	0.67 ± 0.04	13.14 ± 0.59	0.26 ± 0.10
PPP1-[β-D-Asp3,Dhb7]MC-RR	2.65 ± 0.10	0.65 ± 0.04	13.51 ± 0.80	0.34 ± 0.05
MC-RR	2.66 ± 0.07	0.66 ± 0.03	13.25 ± 0.50	0.30 ± 0.10
PPP1-MC-RR	2.70 ± 0.07	0.66 ± 0.04	13.62 ± 0.49	0.33 ± 0.05
MC-RR-Cys-S	2.92 ± 0.08	0.66 ± 0.02	14.26 ± 0.57	0.33 ± 0.08
PPP1-MC-RR-Cys-S	2.91 ± 0.08	0.65 ± 0.02	14.31 ± 0.59	0.35 ± 0.04
MC-RR-Cys-R	3.01 ± 0.07	0.76 ± 0.03	15.19 ± 0.51	0.29 ± 0.07
PPP1-MC-RR-Cys-R	3.01 ± 0.07	0.76 ± 0.03	15.36 ± 0.50	0.34 ± 0.04
MC-RR-GSH-S	3.37 ± 0.10	0.72 ± 0.04	16.52 ± 0.86	0.37 ± 0.07
PPP1-MC-RR-GSH-S	3.32 ± 0.10	0.68 ± 0.04	16.02 ± 0.80	0.36 ± 0.09
MC-RR-GSH-R	3.27 ± 0.12	0.68 ± 0.05	15.51 ± 1.18	0.43 ± 0.13
PPP1-MC-RR-GSH-R	3.29 ± 0.11	0.69 ± 0.04	15.77 ± 0.90	0.46 ± 0.11

Table 8.9: Continued from previous page.

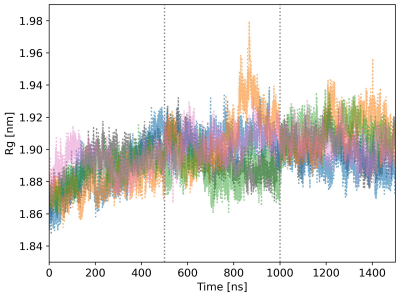
Simulation	Volume $[\frac{nm}{S^3/N}]$	Rgyr $[nm]$	Sasa $(\frac{nm}{S^3/N})$	RMSD $[nm]$
MC-LY	2.54 ± 0.07	0.62 ± 0.03	12.37 ± 0.58	0.21 ± 0.06
PPP1-MC-LY	2.54 ± 0.06	0.61 ± 0.02	12.31 ± 0.44	0.21 ± 0.04
MC-LY-Cys-S	2.76 ± 0.08	0.63 ± 0.02	13.37 ± 0.58	0.28 ± 0.09
PPP1-MC-LY-Cys-S	2.84 ± 0.09	0.67 ± 0.04	13.99 ± 0.68	0.36 ± 0.04
MC-LY-Cys-R	2.78 ± 0.07	0.61 ± 0.02	13.40 ± 0.53	0.37 ± 0.11
PPP1-MC-LY-Cys-R	2.84 ± 0.07	0.66 ± 0.03	13.75 ± 0.53	0.24 ± 0.05
MC-LY-GSH-S	3.25 ± 0.08	0.69 ± 0.03	15.92 ± 0.72	0.35 ± 0.11
PPP1-MC-LY-GSH-S	3.28 ± 0.09	0.75 ± 0.05	16.15 ± 0.80	0.43 ± 0.07
MC-LY-GSH-R	3.24 ± 0.10	0.69 ± 0.05	15.57 ± 0.83	0.44 ± 0.09
PPP1-MC-LY-GSH-R	3.29 ± 0.09	0.74 ± 0.06	16.28 ± 0.80	0.42 ± 0.08
MC-YR	2.65 ± 0.07	0.63 ± 0.03	12.94 ± 0.53	0.28 ± 0.11
PPP1-MC-YR	2.66 ± 0.08	0.66 ± 0.03	13.32 ± 0.56	0.27 ± 0.05
MC-YR-Cys-S	2.96 ± 0.07	0.73 ± 0.03	14.91 ± 0.50	0.30 ± 0.05
PPP1-MC-YR-Cys-S	2.93 ± 0.09	0.69 ± 0.04	14.56 ± 0.80	0.43 ± 0.13
MC-YR-Cys-R	2.87 ± 0.08	0.63 ± 0.02	13.79 ± 0.67	0.32 ± 0.08
PPP1-MC-YR-Cys-R	2.92 ± 0.08	0.66 ± 0.02	14.31 ± 0.66	0.43 ± 0.07
MC-YR-GSH-S	3.35 ± 0.08	0.76 ± 0.03	16.43 ± 0.67	0.40 ± 0.08
PPP1-MC-YR-GSH-S	3.36 ± 0.09	0.75 ± 0.05	16.55 ± 0.80	0.49 ± 0.08
MC-YR-GSH-R	3.37 ± 0.08	0.76 ± 0.04	16.49 ± 0.70	0.39 ± 0.08
PPP1-MC-YR-GSH-R	3.37 ± 0.09	0.75 ± 0.05	16.68 ± 0.76	0.50 ± 0.09
MC-LY(Prg)	2.63 ± 0.07	0.60 ± 0.02	12.73 ± 0.57	0.34 ± 0.12
PPP1-MC-LY(Prg)	2.67 ± 0.07	0.64 ± 0.03	13.13 ± 0.57	0.41 ± 0.11
MC-LY(Prg)-Cys-S	2.94 ± 0.09	0.67 ± 0.03	14.48 ± 0.71	0.33 ± 0.09
PPP1-MC-LY(Prg)-Cys-S	2.98 ± 0.08	0.67 ± 0.03	14.82 ± 0.64	0.31 ± 0.08
MC-LY(Prg)-Cys-R	2.95 ± 0.08	0.67 ± 0.03	14.52 ± 0.58	0.35 ± 0.12
PPP1-MC-LY(Prg)-Cys-R	2.94 ± 0.08	0.65 ± 0.02	14.27 ± 0.60	0.48 ± 0.07
MC-LY(Prg)-GSH-S	3.35 ± 0.11	0.70 ± 0.04	16.12 ± 0.97	0.34 ± 0.06
PPP1-MC-LY(Prg)-GSH-S	3.37 ± 0.11	0.71 ± 0.05	16.45 ± 0.99	0.39 ± 0.09
MC-LY(Prg)-GSH-R	3.40 ± 0.09	0.75 ± 0.05	16.83 ± 0.73	0.43 ± 0.10
PPP1-MC-LY(Prg)-GSH-R	3.47 ± 0.08	0.86 ± 0.03	17.65 ± 0.51	0.39 ± 0.04
[Anda5]-MC-LY(Prg)	2.29 ± 0.07	0.57 ± 0.02	11.47 ± 0.47	0.28 ± 0.06
PPP1-[Anda5]-MC-LY(Prg)	2.28 ± 0.07	0.57 ± 0.03	11.29 ± 0.55	0.36 ± 0.11
[Anda5]-MC-LY(Prg)-Cys-S	2.60 ± 0.07	0.61 ± 0.02	13.10 ± 0.55	0.33 ± 0.10
PPP1-[Anda5]-MC-LY(Prg)-Cys-S	2.57 ± 0.08	0.61 ± 0.02	12.82 ± 0.54	0.41 ± 0.07
[Anda5]-MC-LY(Prg)-Cys-R	2.57 ± 0.07	0.63 ± 0.02	12.79 ± 0.50	0.21 ± 0.06
PPP1-[Anda5]-MC-LY(Prg)-Cys-R	2.61 ± 0.06	0.63 ± 0.02	13.28 ± 0.47	0.35 ± 0.08
[Anda5]-MC-LY(Prg)-GSH-S	3.07 ± 0.07	0.78 ± 0.03	15.71 ± 0.54	0.37 ± 0.08
PPP1-[Anda5]-MC-LY(Prg)-GSH-S	3.08 ± 0.08	0.79 ± 0.04	15.80 ± 0.59	0.40 ± 0.06
[Anda5]-MC-LY(Prg)-GSH-R	3.03 ± 0.08	0.73 ± 0.03	15.24 ± 0.57	0.37 ± 0.08
PPP1-[Anda5]-MC-LY(Prg)-GSH-R	3.05 ± 0.08	0.74 ± 0.03	15.60 ± 0.64	0.37 ± 0.05

Table 8.9: Continued from previous page.

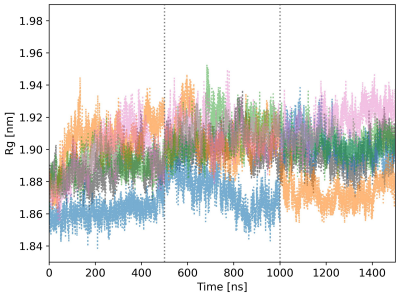
Simulation	Volume $[\frac{nm}{S^3/N}]$	Rgyr $[nm]$	Sasa $(\frac{nm}{S^3/N})$	RMSD $[nm]$
[Amba5]-MC-LY(Prg)	2.05 ± 0.05	0.50 ± 0.01	10.06 ± 0.35	0.26 ± 0.13
PPP1-[Amba5]-MC-LY(Prg)	2.06 ± 0.06	0.52 ± 0.02	10.25 ± 0.44	0.26 ± 0.07
[Amba5]-MC-LY(Prg)-Cys-S	2.41 ± 0.06	0.62 ± 0.03	12.31 ± 0.45	0.30 ± 0.08
PPP1-[Amba5]-MC-LY(Prg)-Cys-S	2.39 ± 0.06	0.61 ± 0.03	12.24 ± 0.43	0.30 ± 0.04
[Amba5]-MC-LY(Prg)-Cys-R	2.36 ± 0.06	0.56 ± 0.01	11.73 ± 0.38	0.24 ± 0.07
PPP1-[Amba5]-MC-LY(Prg)-Cys-R	2.36 ± 0.06	0.57 ± 0.02	11.78 ± 0.50	0.31 ± 0.06
[Amba5]-MC-LY(Prg)-GSH-S	2.79 ± 0.07	0.66 ± 0.03	13.99 ± 0.61	0.39 ± 0.06
PPP1-[Amba5]-MC-LY(Prg)-GSH-S	2.81 ± 0.07	0.66 ± 0.02	14.25 ± 0.53	0.43 ± 0.05
[Amba5]-MC-LY(Prg)-GSH-R	2.79 ± 0.08	0.66 ± 0.03	13.95 ± 0.68	0.40 ± 0.14
PPP1-[Amba5]-MC-LY(Prg)-GSH-R	2.80 ± 0.06	0.67 ± 0.03	14.03 ± 0.51	0.51 ± 0.08
[Apha5]-MC-LF	2.18 ± 0.06	0.53 ± 0.02	10.48 ± 0.41	0.22 ± 0.07
PPP1-[Apha5]-MC-LF	2.23 ± 0.07	0.54 ± 0.03	10.85 ± 0.52	0.37 ± 0.11
[APHA]-MC-LF-Cys-S	2.50 ± 0.07	0.59 ± 0.03	12.20 ± 0.53	0.33 ± 0.14
PPP1-[Apha5]-MC-LF-Cys-S	2.51 ± 0.07	0.60 ± 0.02	12.31 ± 0.48	0.38 ± 0.12
[APHA]-MC-LF-Cys-R	2.46 ± 0.06	0.57 ± 0.02	11.89 ± 0.45	0.23 ± 0.07
PPP1-[Apha5]-MC-LF-Cys-R	2.48 ± 0.06	0.58 ± 0.02	12.08 ± 0.45	0.34 ± 0.10
[APHA]-MC-LF-GSH-S	2.88 ± 0.09	0.63 ± 0.04	13.74 ± 0.80	0.38 ± 0.07
PPP1-[Apha5]-MC-LF-GSH-S	2.92 ± 0.09	0.65 ± 0.04	14.19 ± 0.87	0.43 ± 0.08
[APHA]-MC-LF-GSH-R	2.94 ± 0.09	0.66 ± 0.04	14.27 ± 0.80	0.42 ± 0.10
PPP1-[Apha5]-MC-LF-GSH-R	2.99 ± 0.09	0.69 ± 0.05	14.81 ± 0.84	0.47 ± 0.08
[Apda5]-MC-LF	2.35 ± 0.07	0.58 ± 0.04	11.45 ± 0.53	0.30 ± 0.11
PPP1-[Apda5]-MC-LF	2.37 ± 0.07	0.59 ± 0.04	11.59 ± 0.62	0.35 ± 0.13
[APDA]-MC-LF-Cys-S	2.63 ± 0.08	0.60 ± 0.04	12.55 ± 0.74	0.36 ± 0.10
PPP1-[Apda5]-MC-LF-Cys-S	2.70 ± 0.08	0.66 ± 0.04	13.41 ± 0.70	0.34 ± 0.12
[APDA]-MC-LF-Cys-R	2.64 ± 0.07	0.60 ± 0.02	12.79 ± 0.57	0.28 ± 0.08
PPP1-[Apda5]-MC-LF-Cys-R	2.66 ± 0.07	0.61 ± 0.02	13.13 ± 0.53	0.31 ± 0.05
[APDA]-MC-LF-GSH-S	3.05 ± 0.08	0.66 ± 0.03	14.68 ± 0.75	0.40 ± 0.16
PPP1-[Apda5]-MC-LF-GSH-S	3.05 ± 0.10	0.66 ± 0.03	14.73 ± 0.90	0.52 ± 0.07
[APDA]-MC-LF-GSH-R	3.07 ± 0.10	0.67 ± 0.03	14.79 ± 0.94	0.44 ± 0.13
PPP1-[Apda5]-MC-LF-GSH-R	3.04 ± 0.10	0.67 ± 0.04	14.48 ± 0.90	0.56 ± 0.14



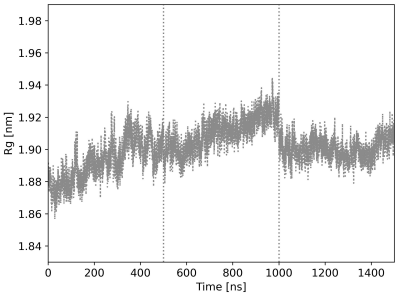
(a) PPP1-MC-LR



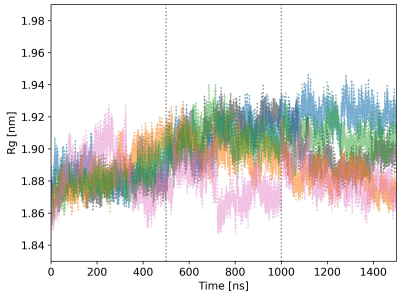
(b) PPP1-MC-LF



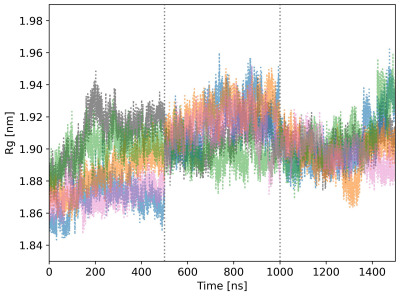
(c) PPP1-[Enantio-Adda5]MC-LF



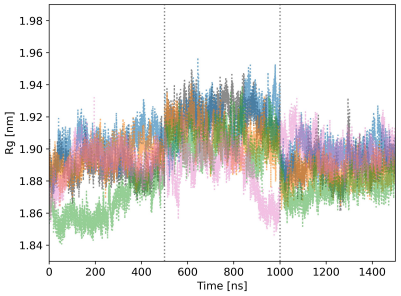
(d) PPP1-[β-D-Asp3,Dhb7]MC-RR



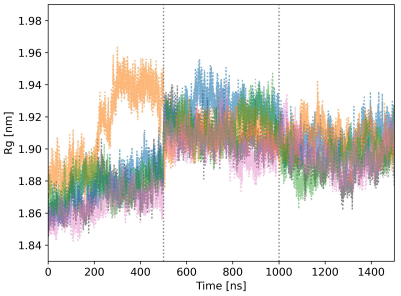
(e) PPP1-MC-RR



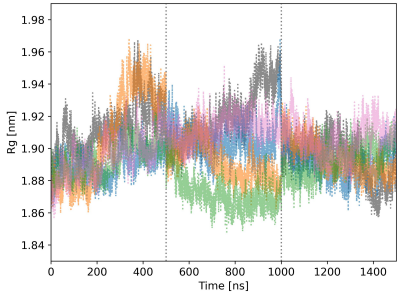
(f) PPP1-MC-LY



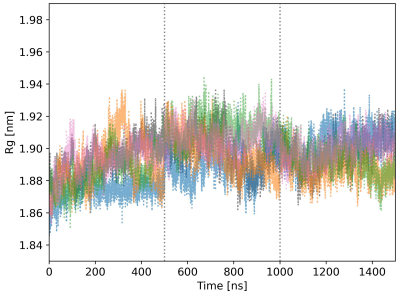
(g) PPP1-MC-YR



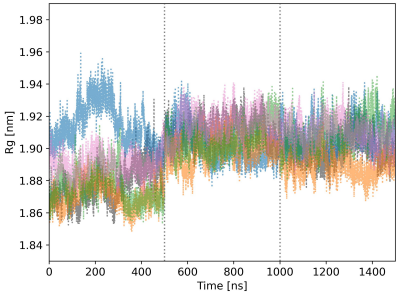
(h) PPP1-MC-LY(Prg)



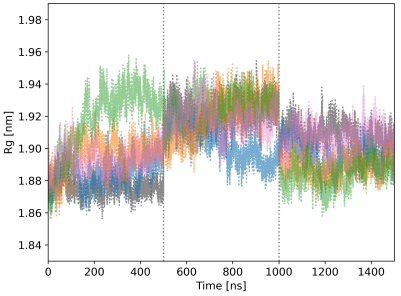
(i) PPP1-[Anda5]-MC-LY(Prg)



(j) PPP1-[Amba5]-MC-LY(Prg)

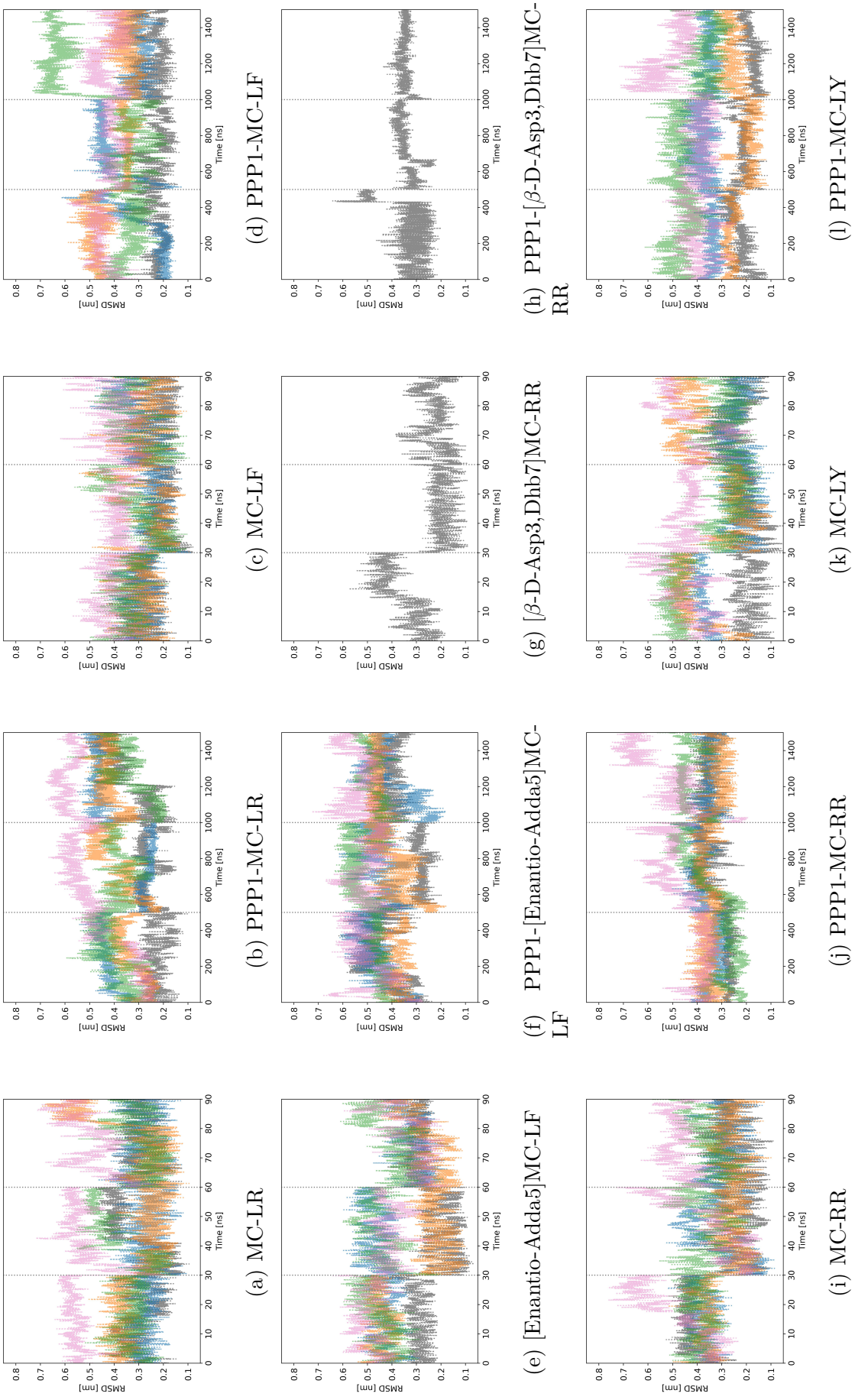


(k) PPP1-[APHA]-MC-LF



(l) PPP1-[APDA]-MC-LF

Figure 8.14: Time series of PPP1 radius of gyration of different MC congeners and their respective conjugate. The three replicates are shown as one timeline and are separated by a grey dashed line at 500 and 1000 ns. The MC congeners and conjugates are shown in different colours: unconjugated in grey, Cys-S in blue, Cys-R in orange, GSH-S in green and GSH-R in pink.



(a) MC-LR (b) PPP1-MC-LR (c) MC-LF (d) PPP1-MC-LF

(e) [Enantio-Adda5]MC-LF (f) PPP1-[Enantio-Adda5]MC-LF (g) [β-D-Asp3,Dhb7]MC-RR (h) PPP1-[β-D-Asp3,Dhb7]MC-RR

(i) MC-RR (j) PPP1-MC-RR (k) MC-LY (l) PPP1-MC-LY

Figure 8.15: Time series of MC congeners all-atom RMSD. The three replicates are shown as one line. The solvent and complex simulations are separated by a grey dashed line at 30 and 500, and 60 and 1000 ns, respectively. The MC congeners and conjugates are shown in different colours: unconjugated in grey, Cys-S in blue, Cys-R in orange, GSH-S in green and GSH-R in pink.

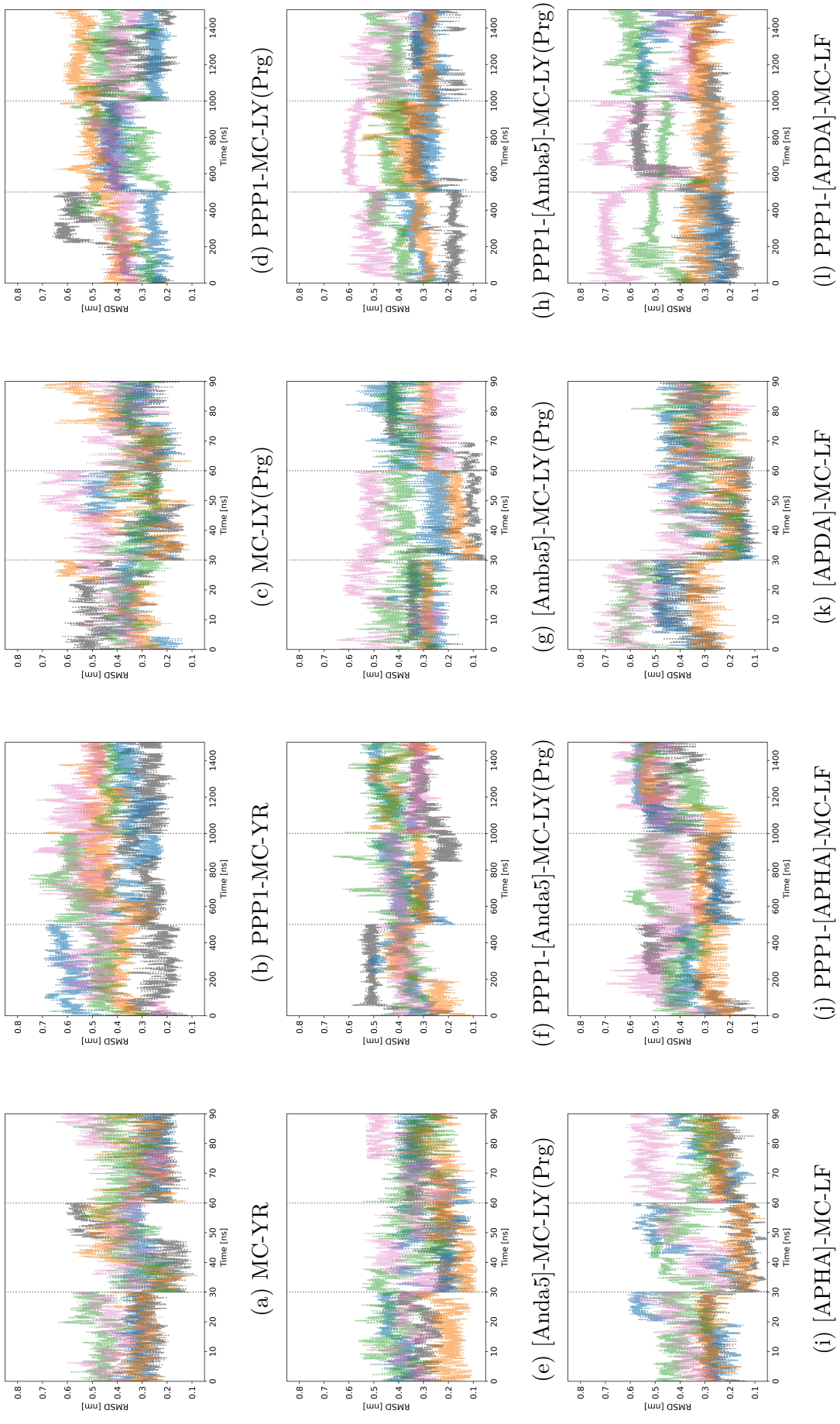


Figure 8.16: Time series of MC congeners all-atom RMSD. The three replicates are shown as one line. The solvent and complex simulations are separated by a grey dashed line at 30 and 500, and 60 and 1000 ns, respectively. The MC congeners and conjugates are shown in different colours: unconjugated in grey, Cys-S in blue, Cys-R in orange, GSH-S in green and GSH-R in pink.

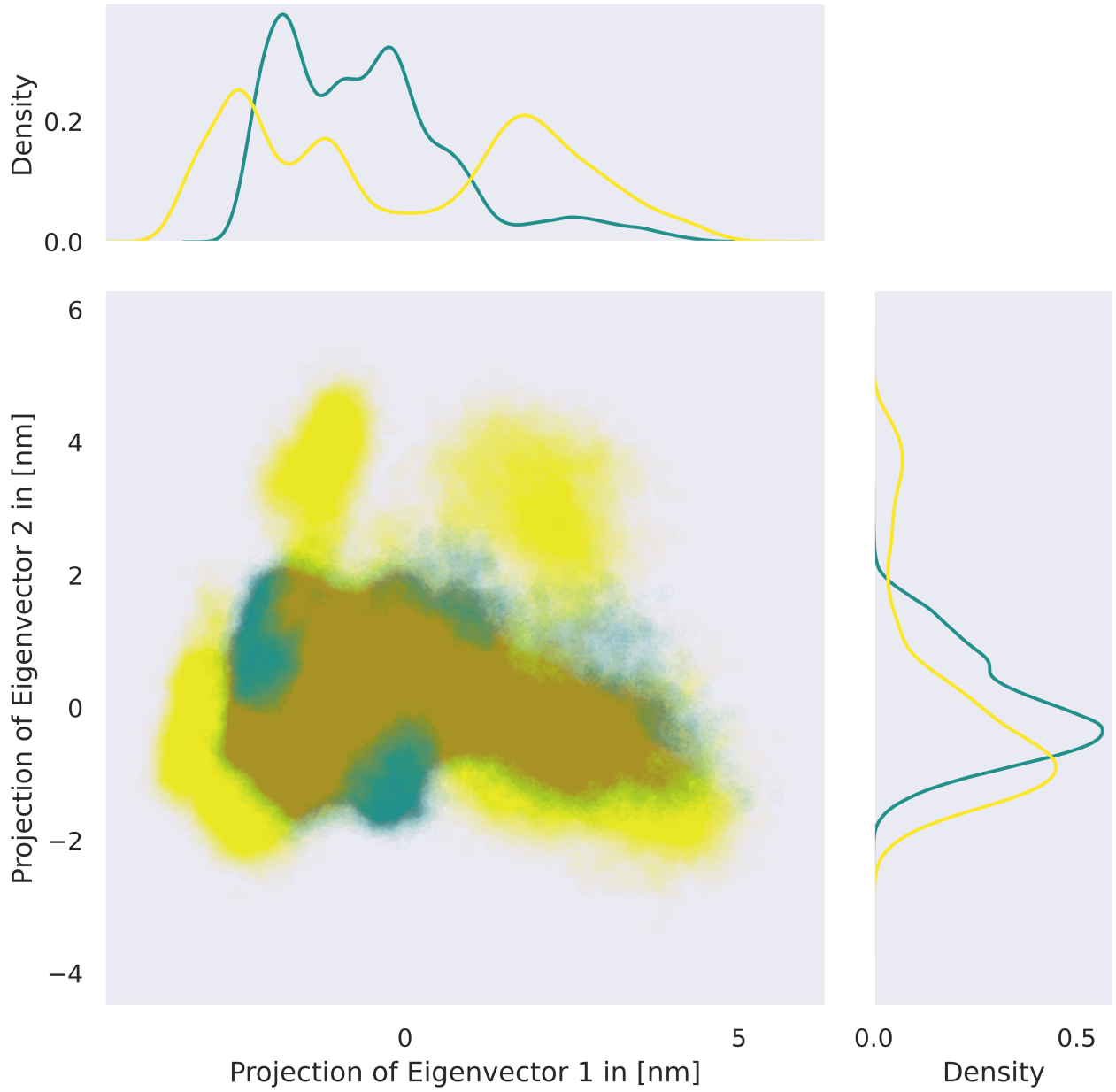


Figure 8.17: Principal components of PPP1 backbone of MC-LR complex (green) and apo (yellow) MD simulations. The 2D projections are visualised by the central scatter plot with a distribution line at the top and right to highlight the frequency of occurrence of individual data points.

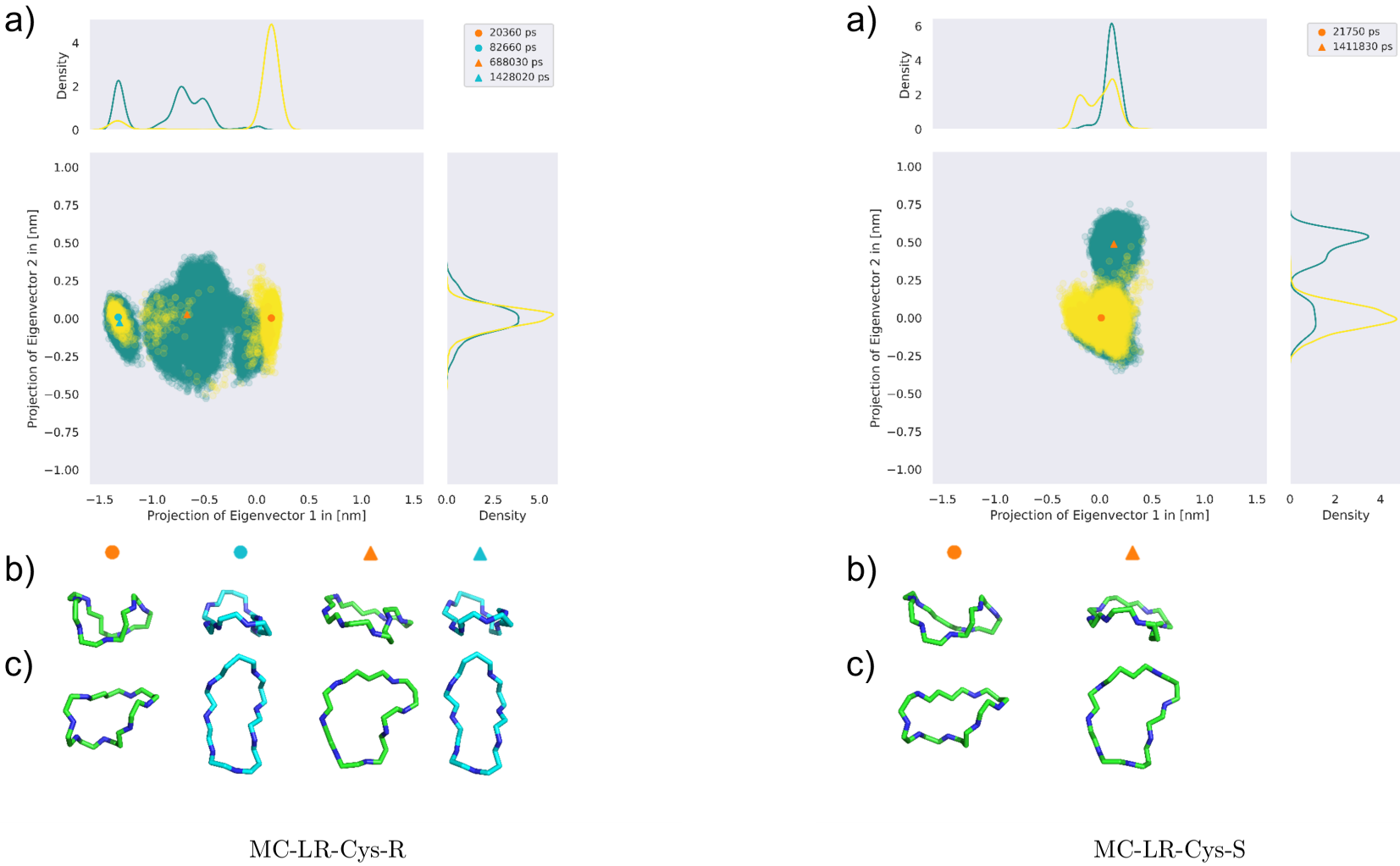


Figure 8.18: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

Figure 8.19: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

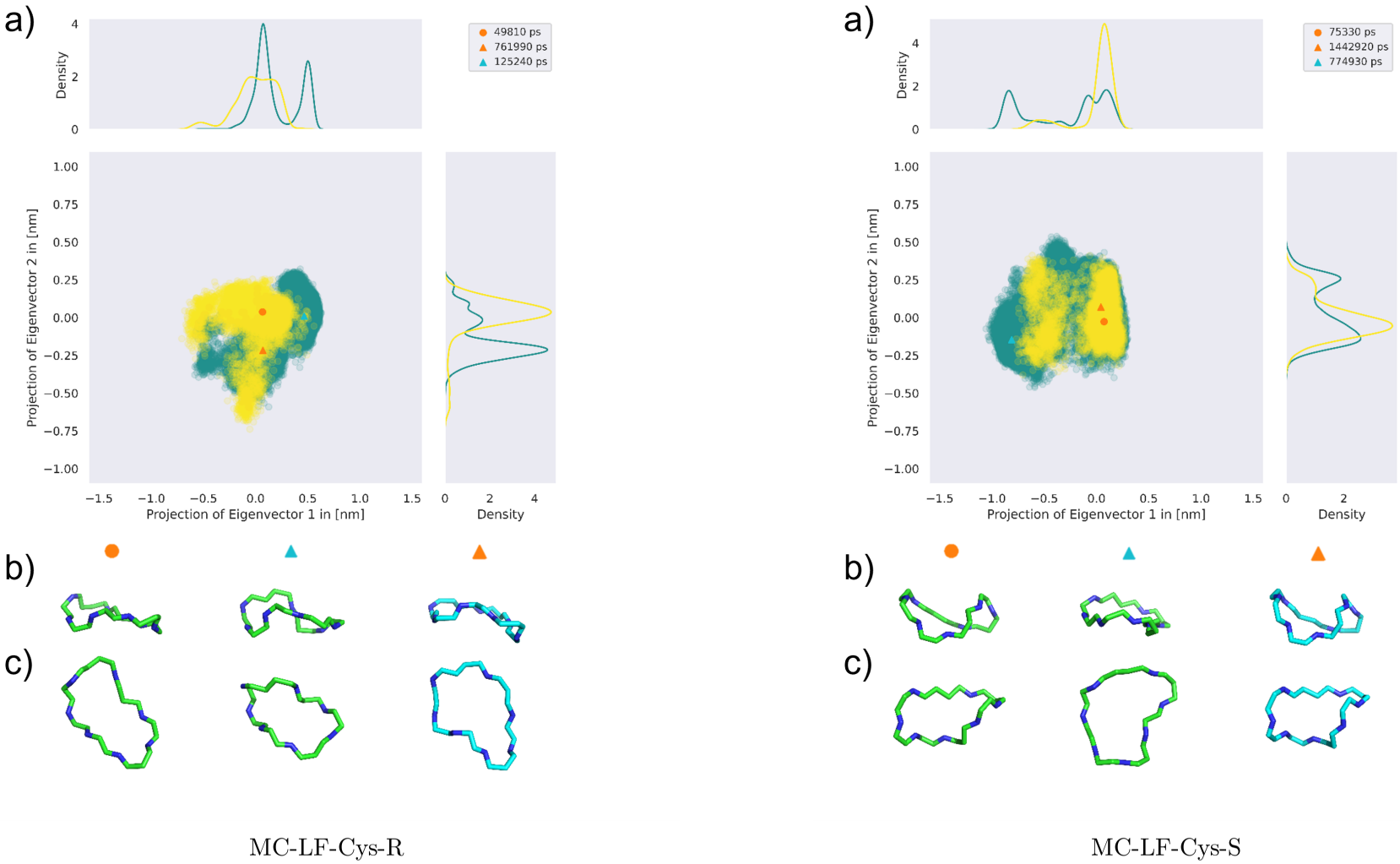
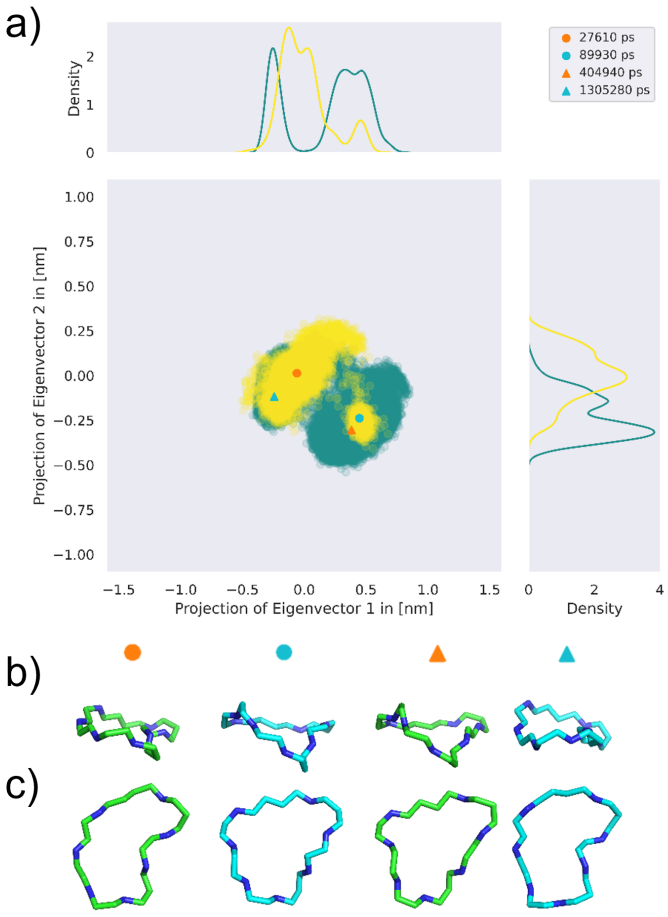
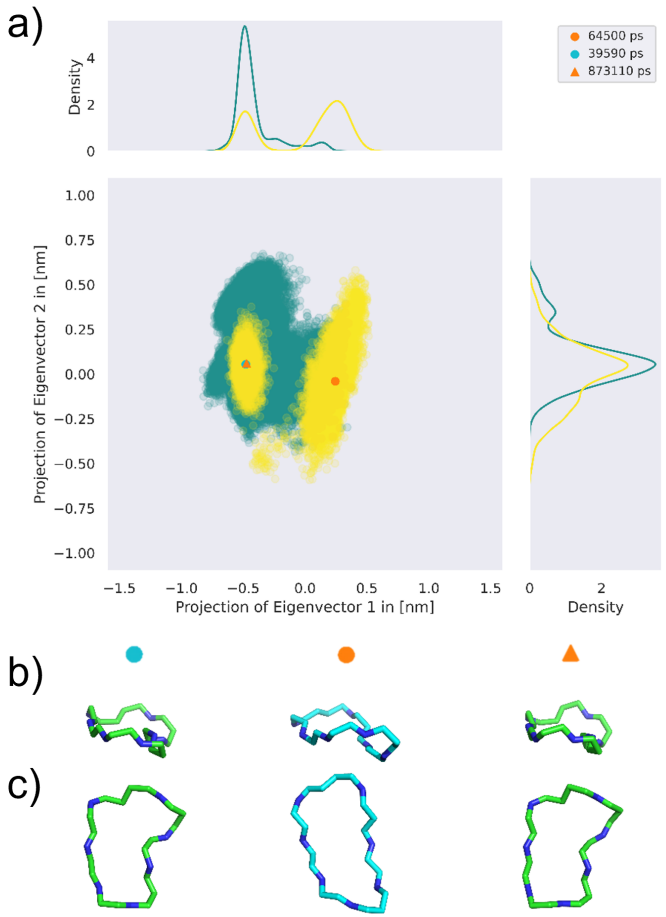


Figure 8.20: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

Figure 8.21: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

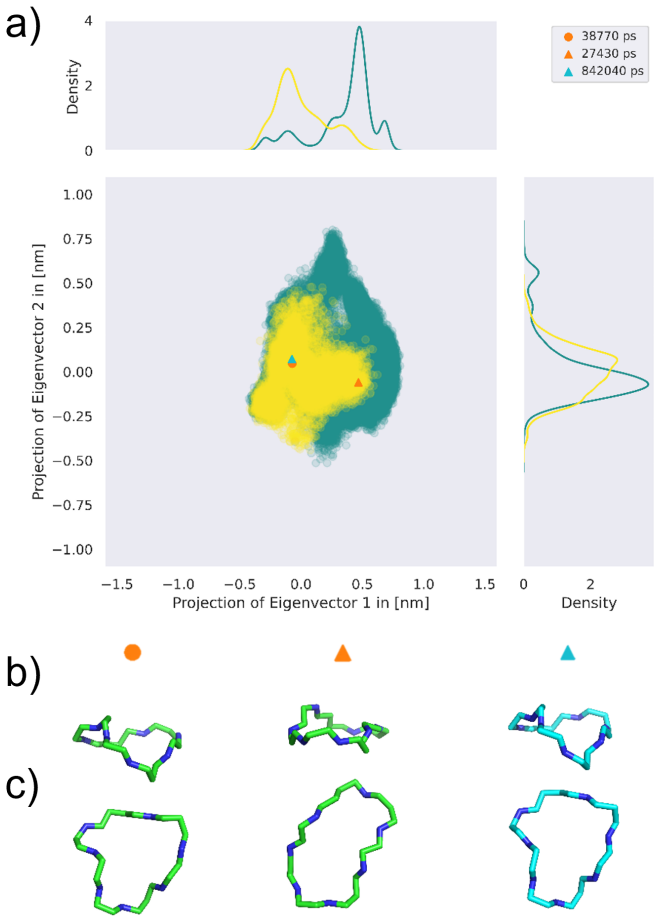


[Enantio-Adda5]MC-LF-Cys-R

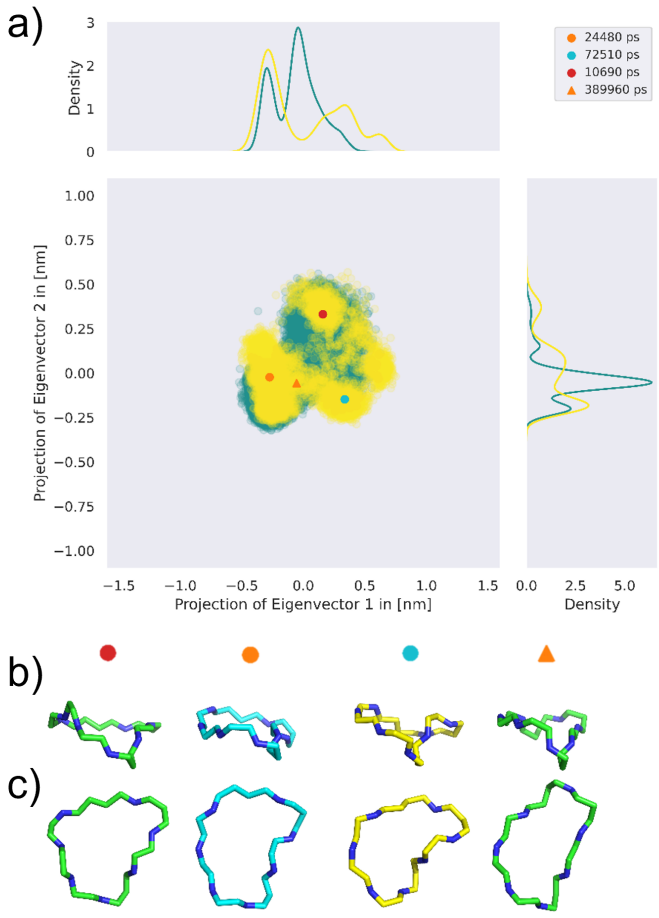


[Enantio-Adda5]MC-LF-Cys-S

Figure 8.22: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.



[Enantio-Adda5]MC-LF-GSH-R



[Enantio-Adda5]MC-LF-GSH-S

Figure 8.23: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

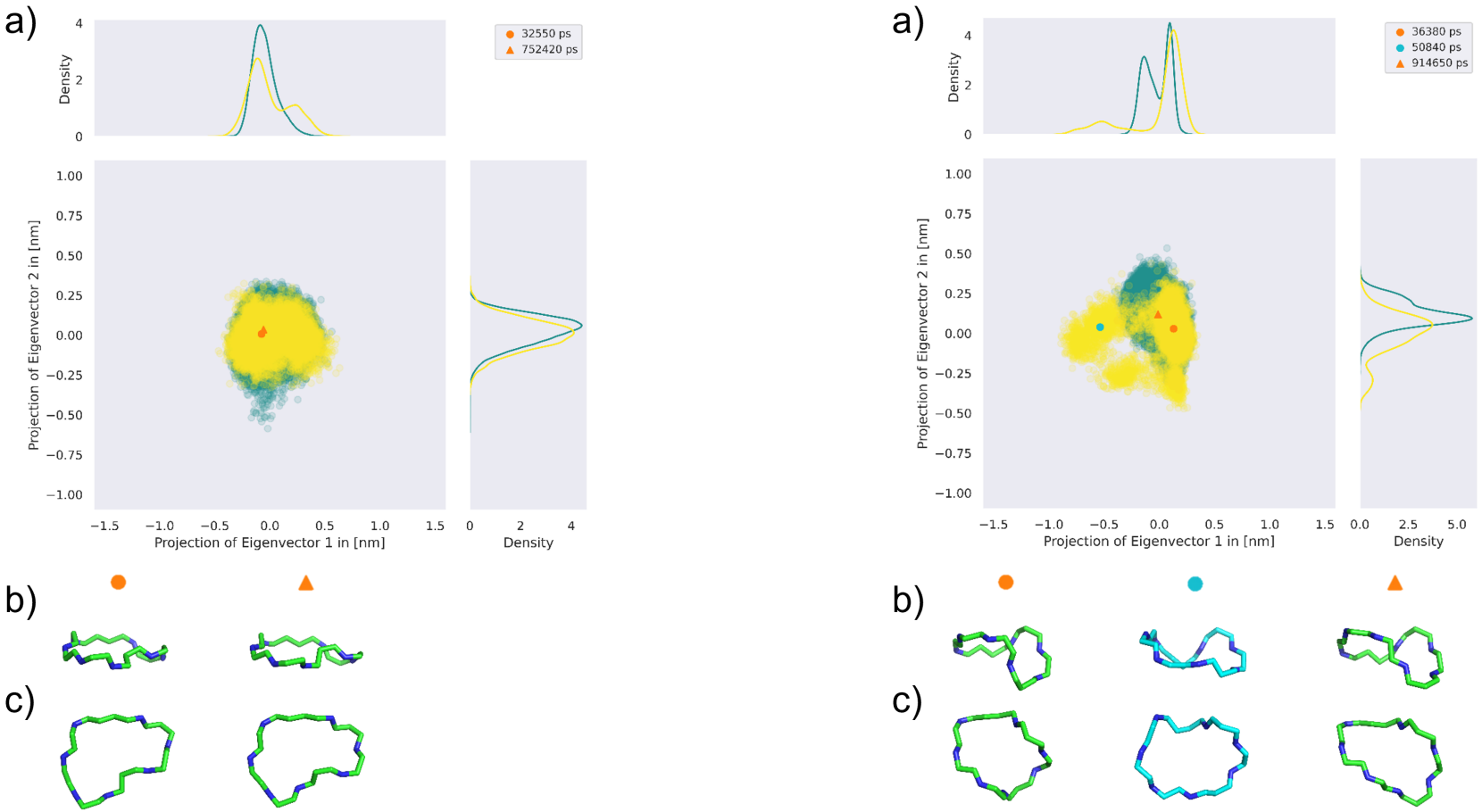


Figure 8.24: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

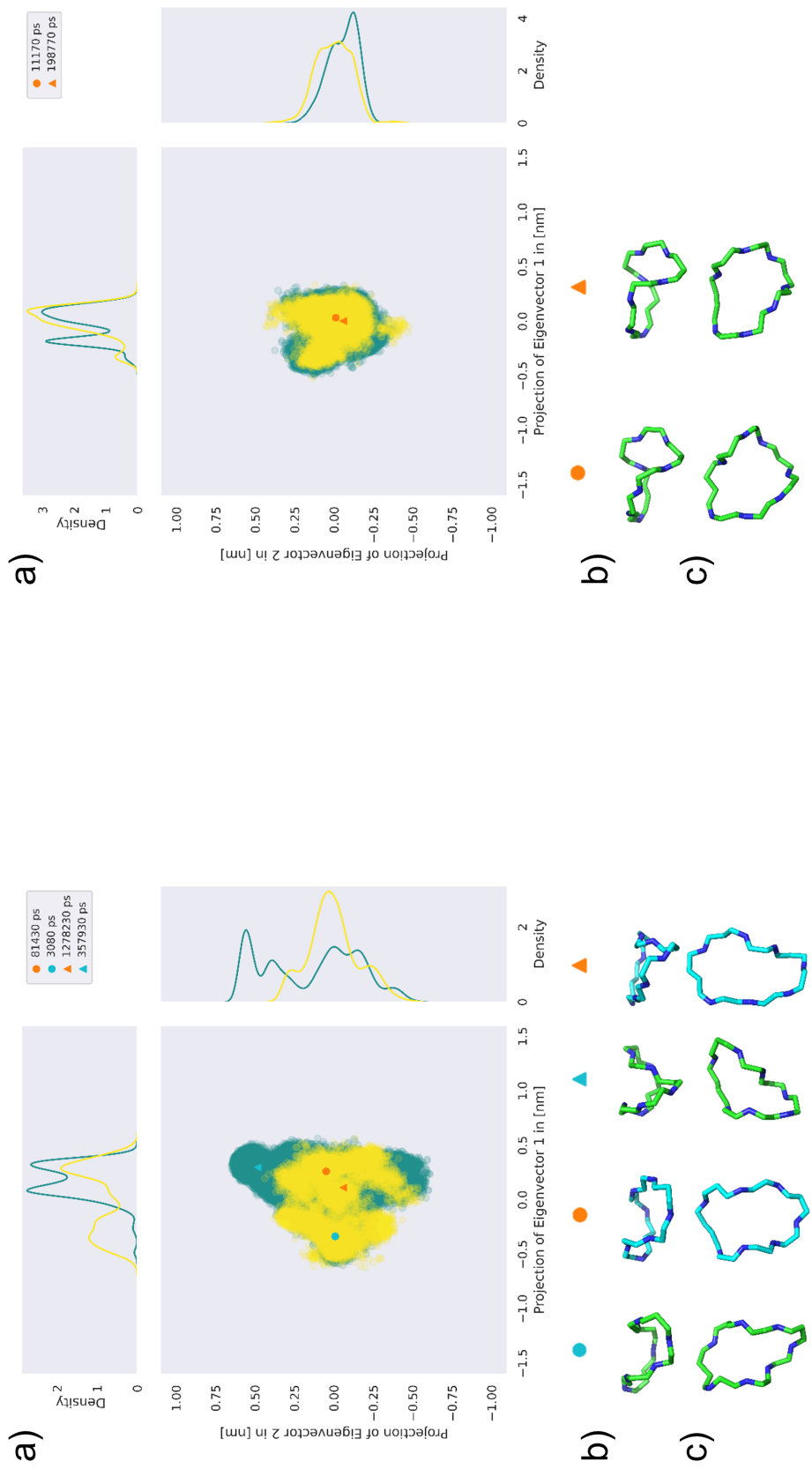


Figure 8.25: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

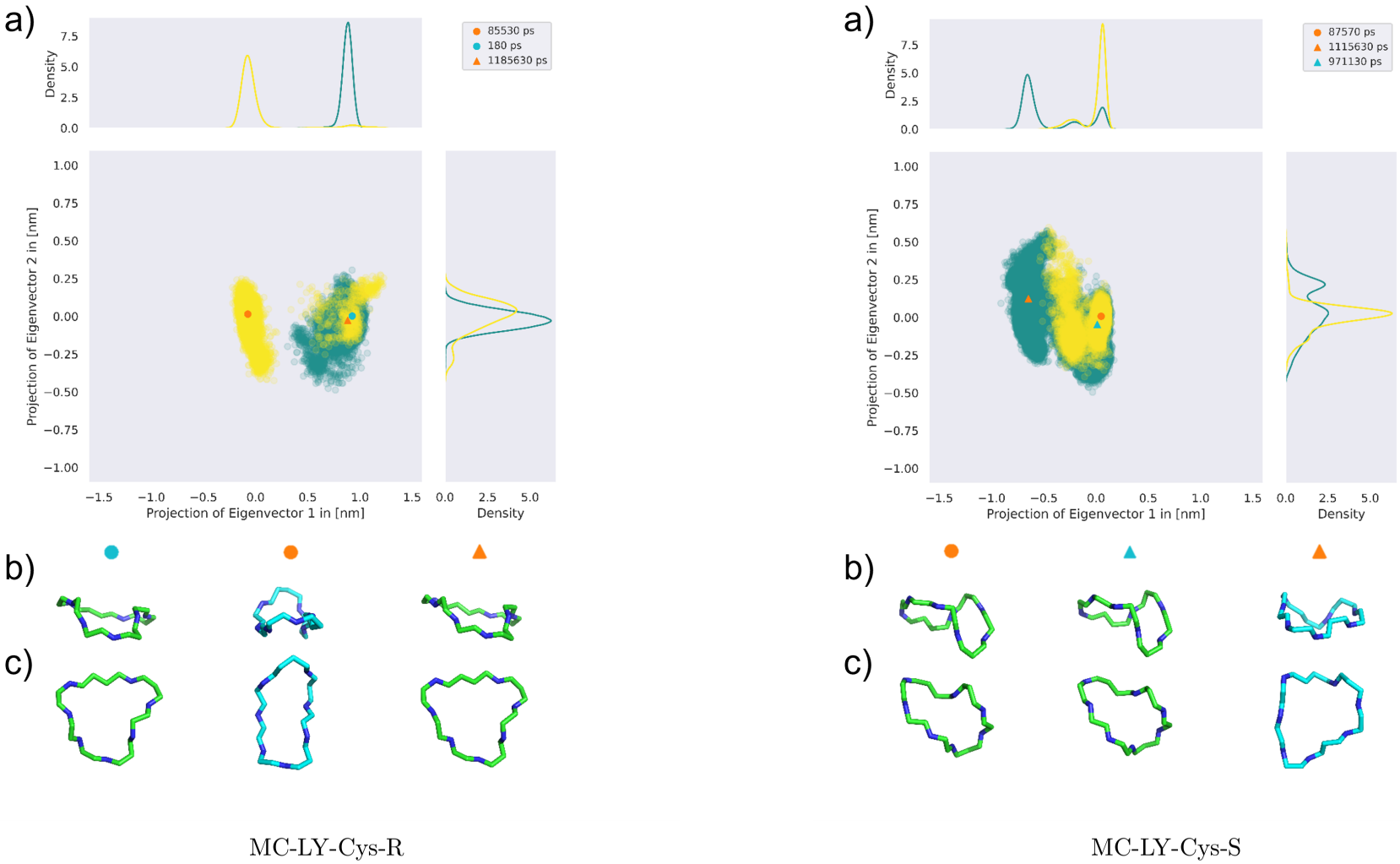


Figure 8.26: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

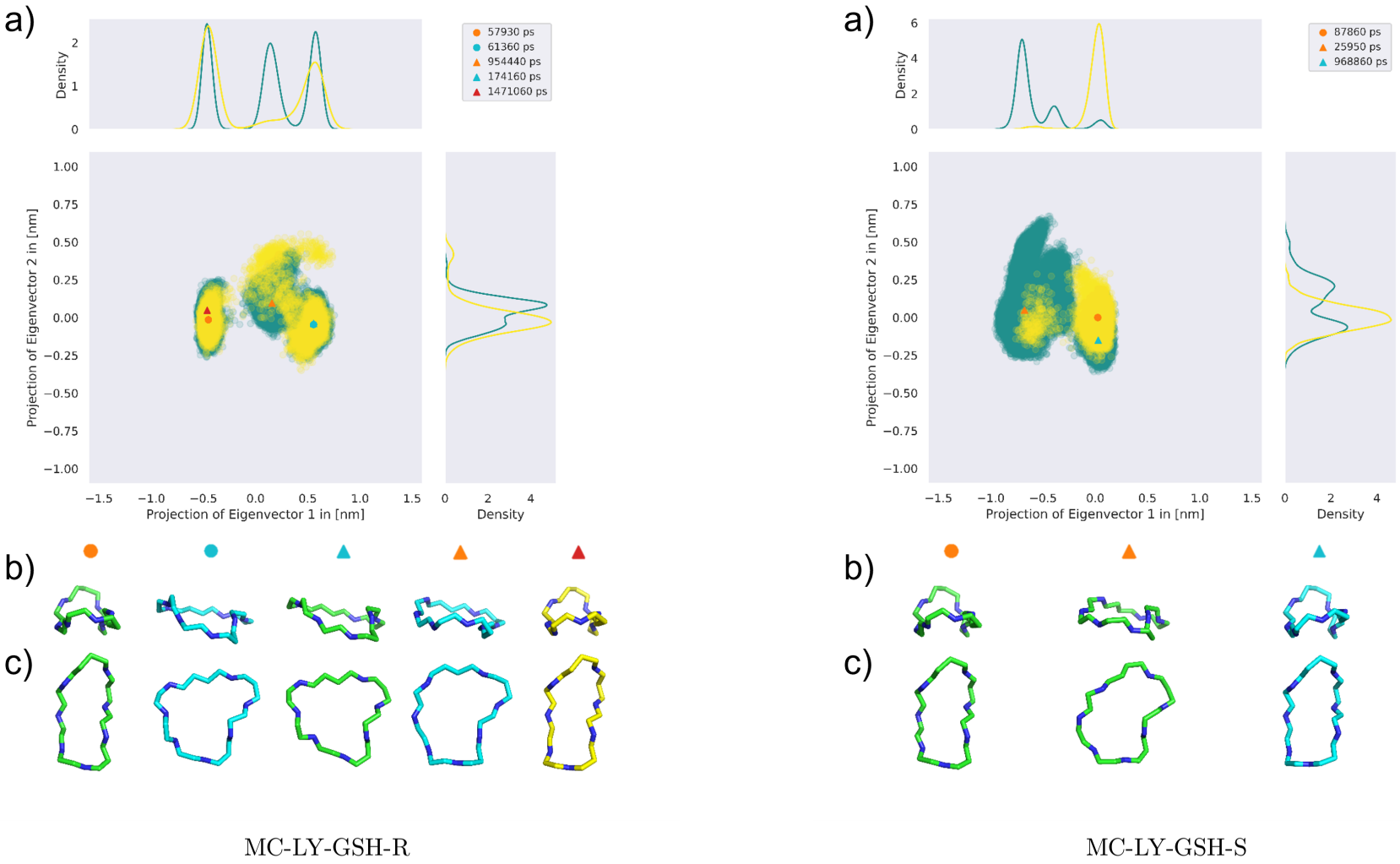


Figure 8.27: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

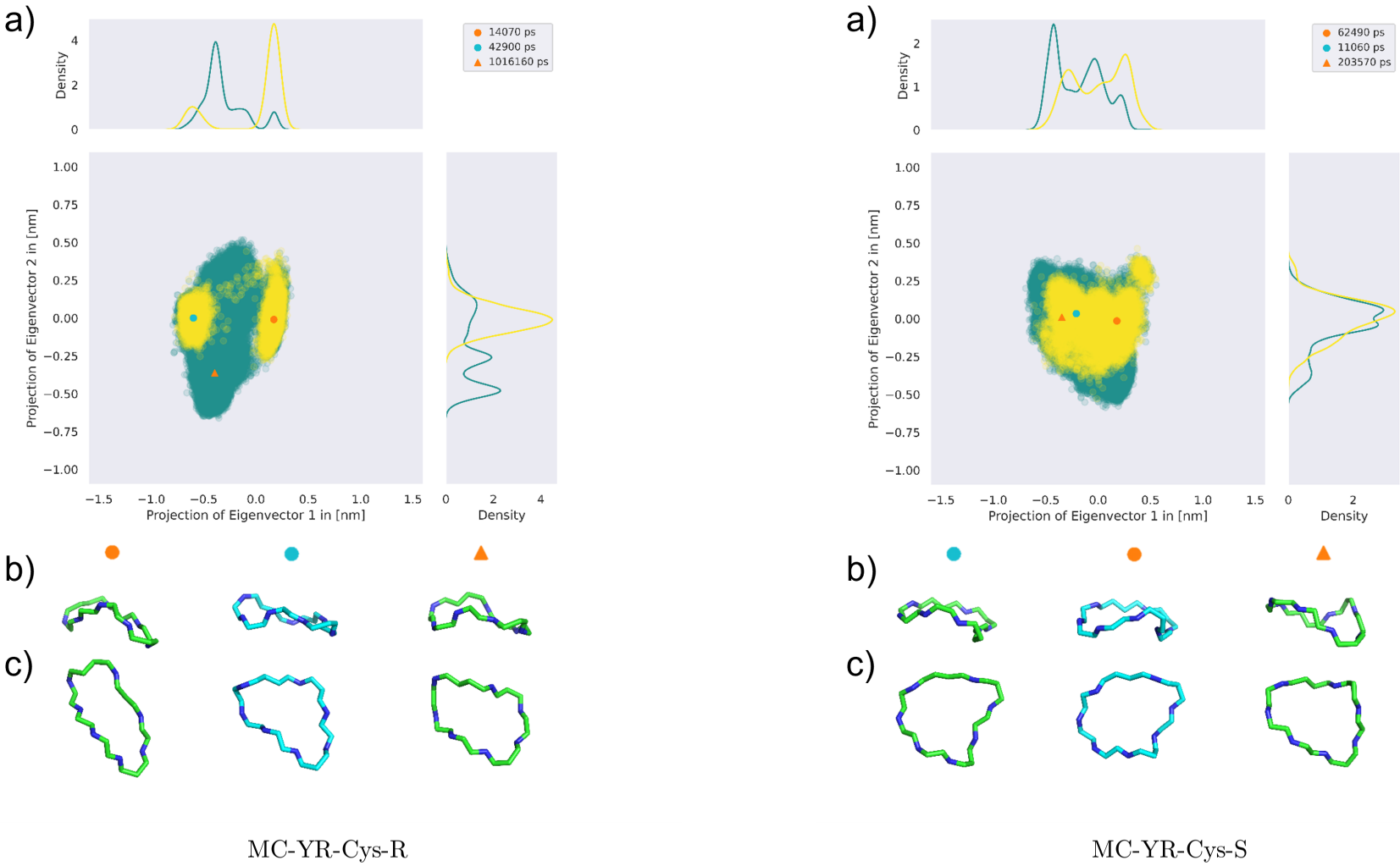
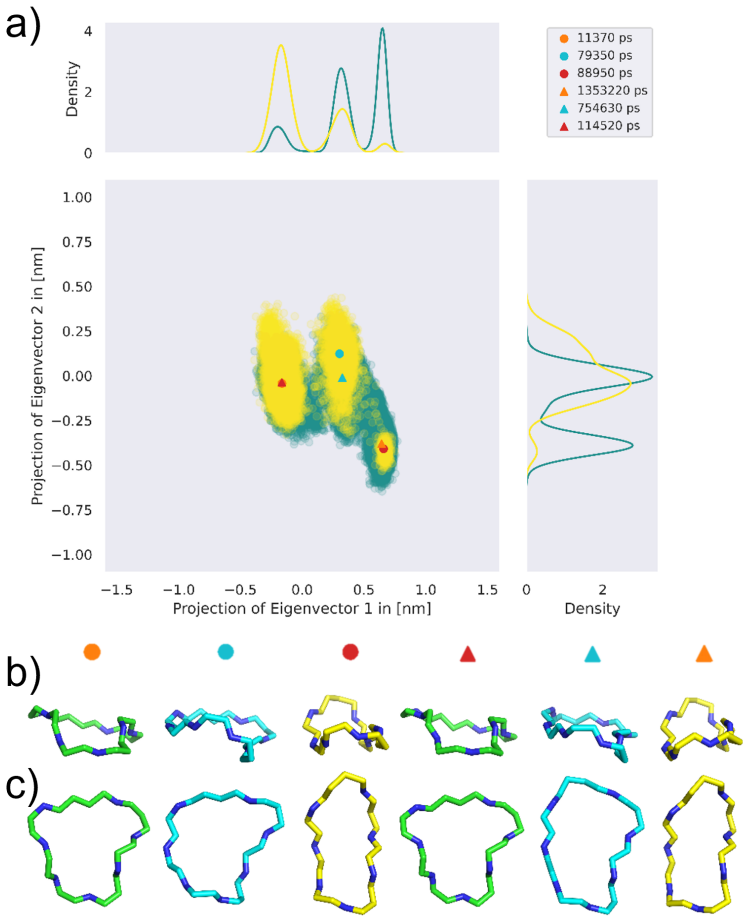
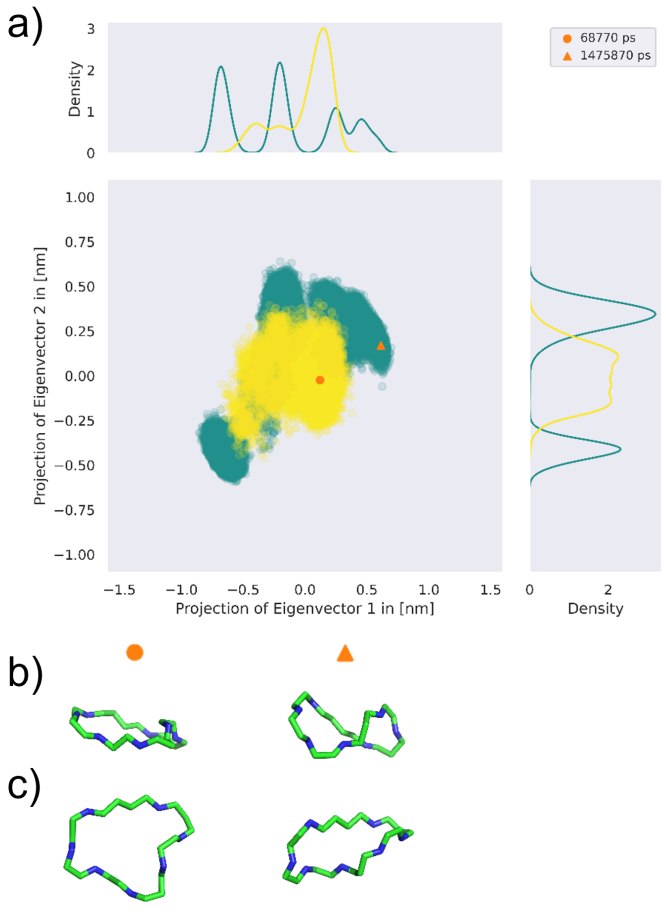


Figure 8.28: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

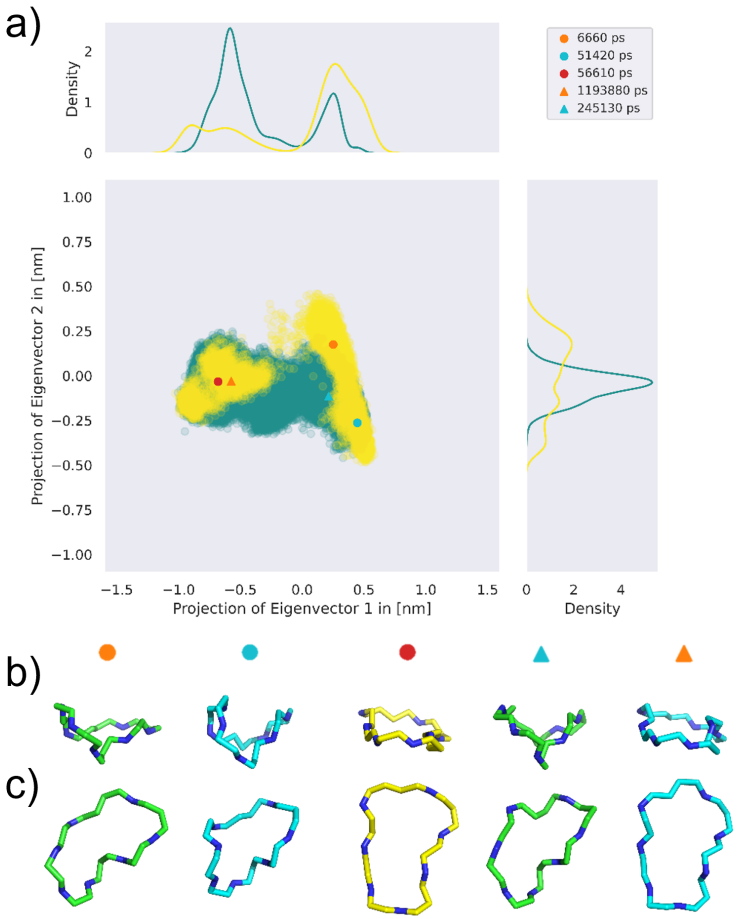


MC-YR-GSH-R

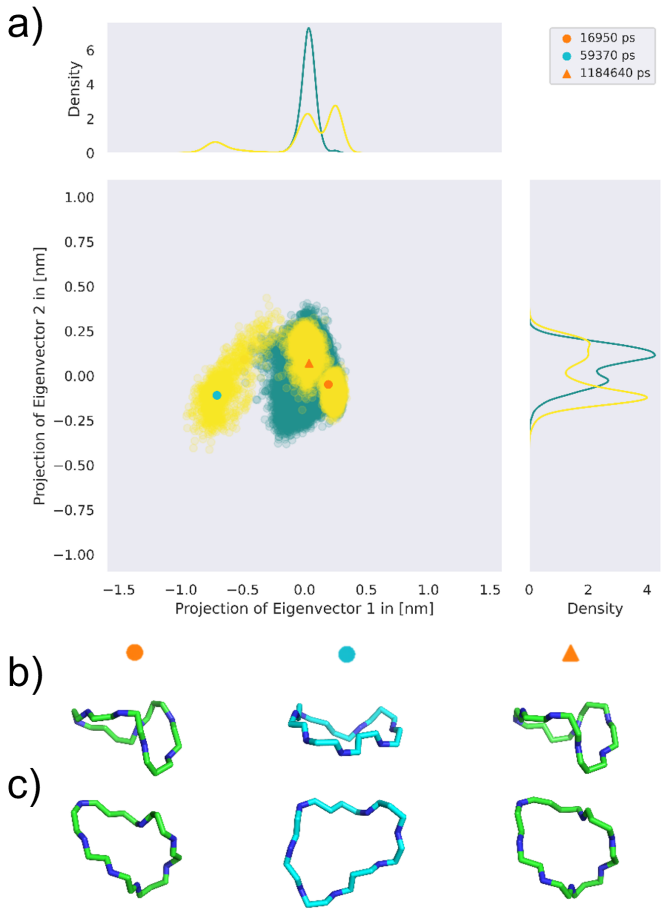


MC-YR-GSH-S

Figure 8.29: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

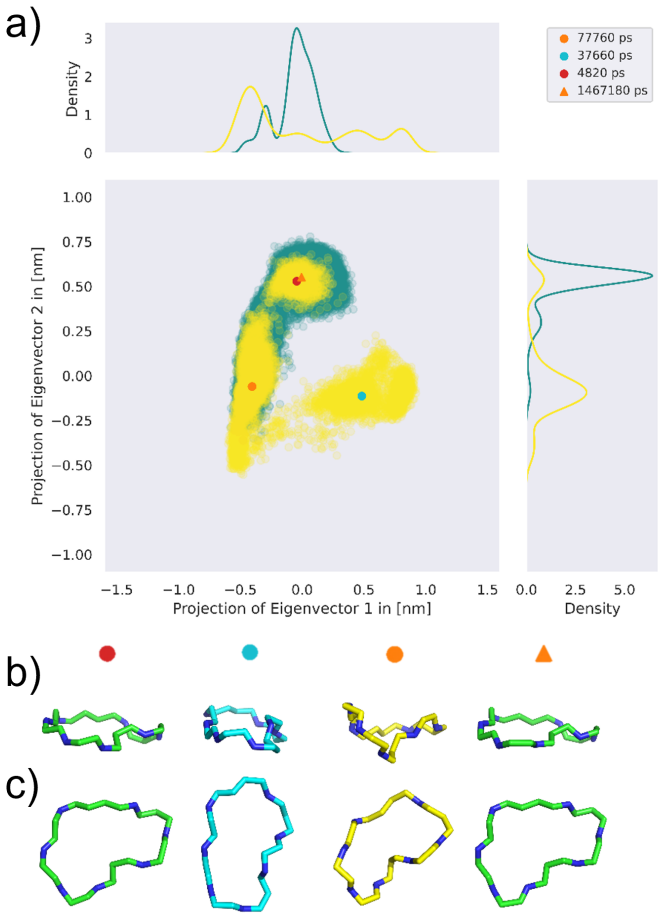


MC-LY(Prg)-Cys-R

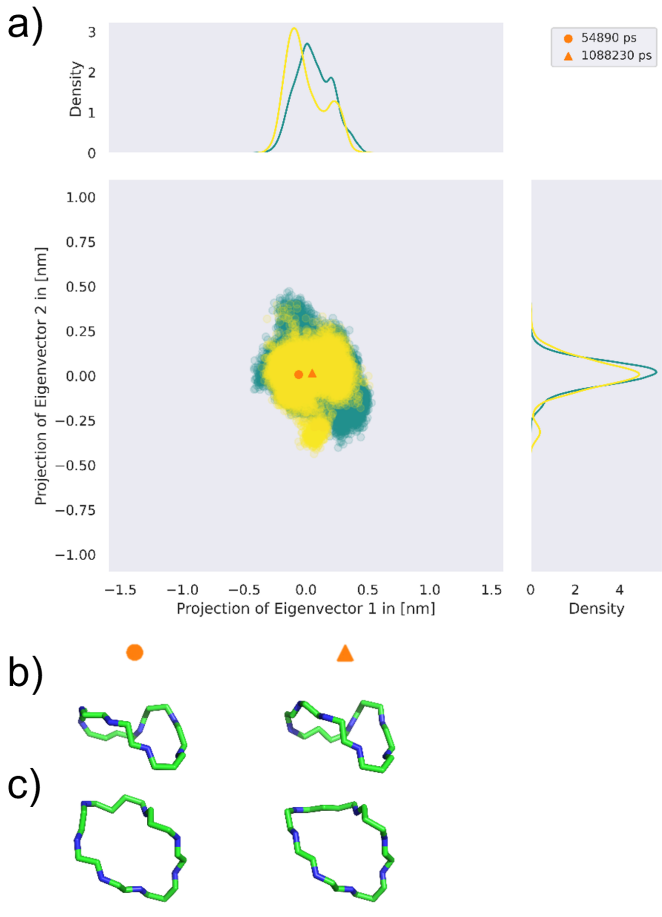


MC-LY(Prg)-Cys-S

Figure 8.30: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

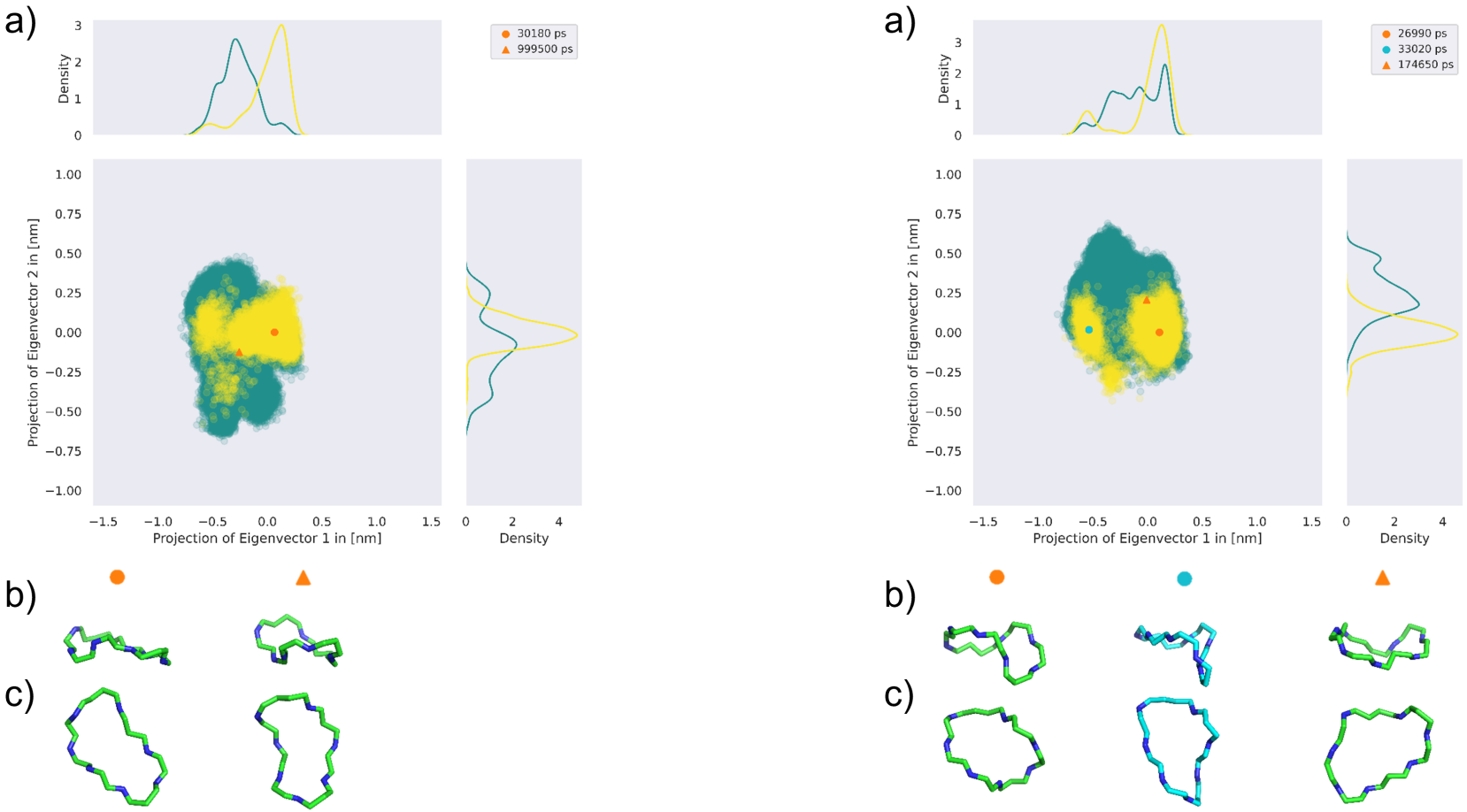


MC-LY(Prg)-GSH-R



MC-LY(Prg)-GSH-S

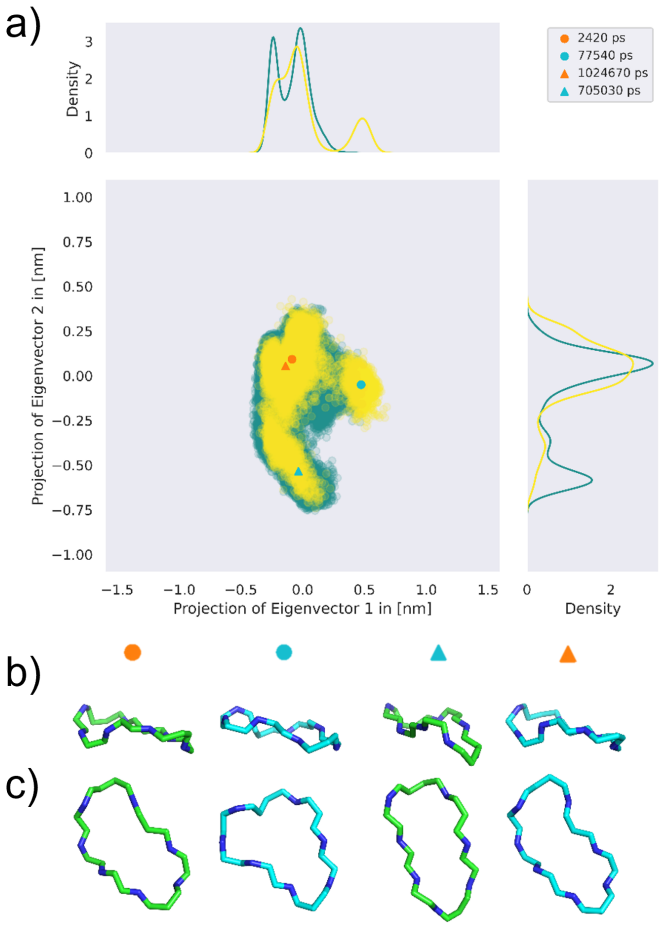
Figure 8.31: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.



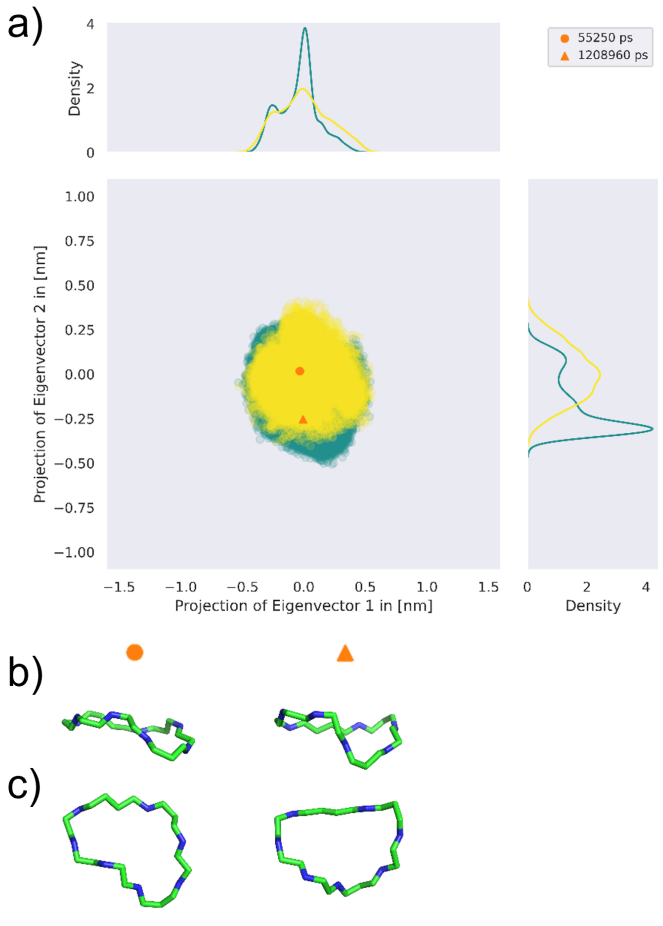
[Anda5]-MC-LY(Prg)-Cys-R

[Anda5]-MC-LY(Prg)-Cys-S

Figure 8.32: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.



[Anda5]-MC-LY(Prg)-GSH-R



[Anda5]-MC-LY(Prg)-GSH-S

Figure 8.33: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

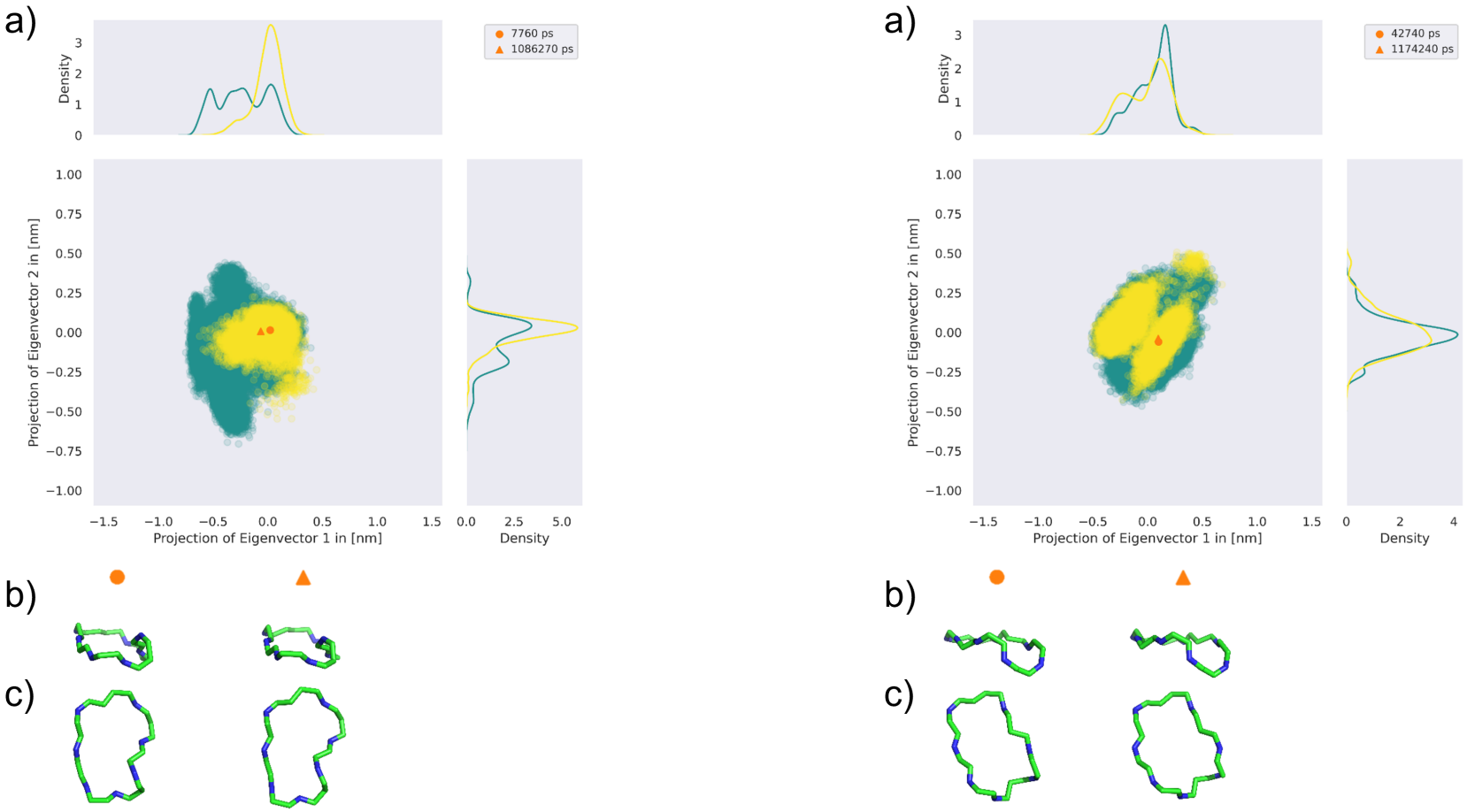
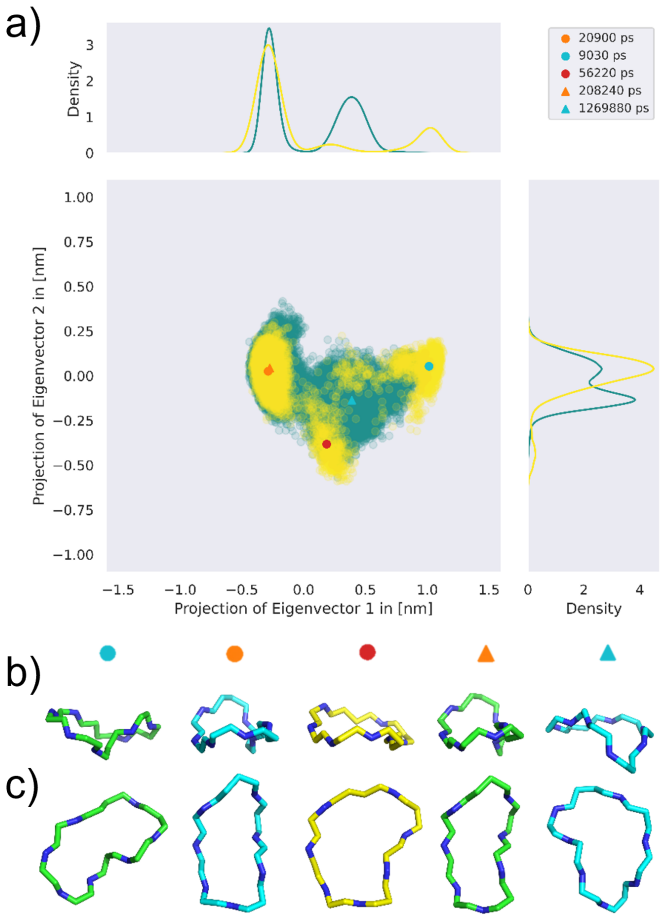
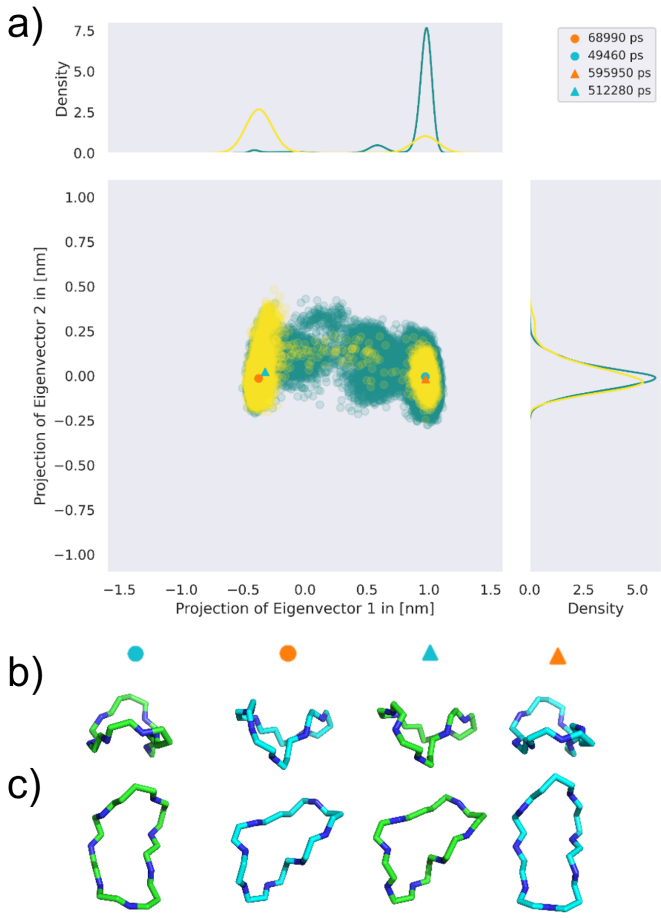


Figure 8.34: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

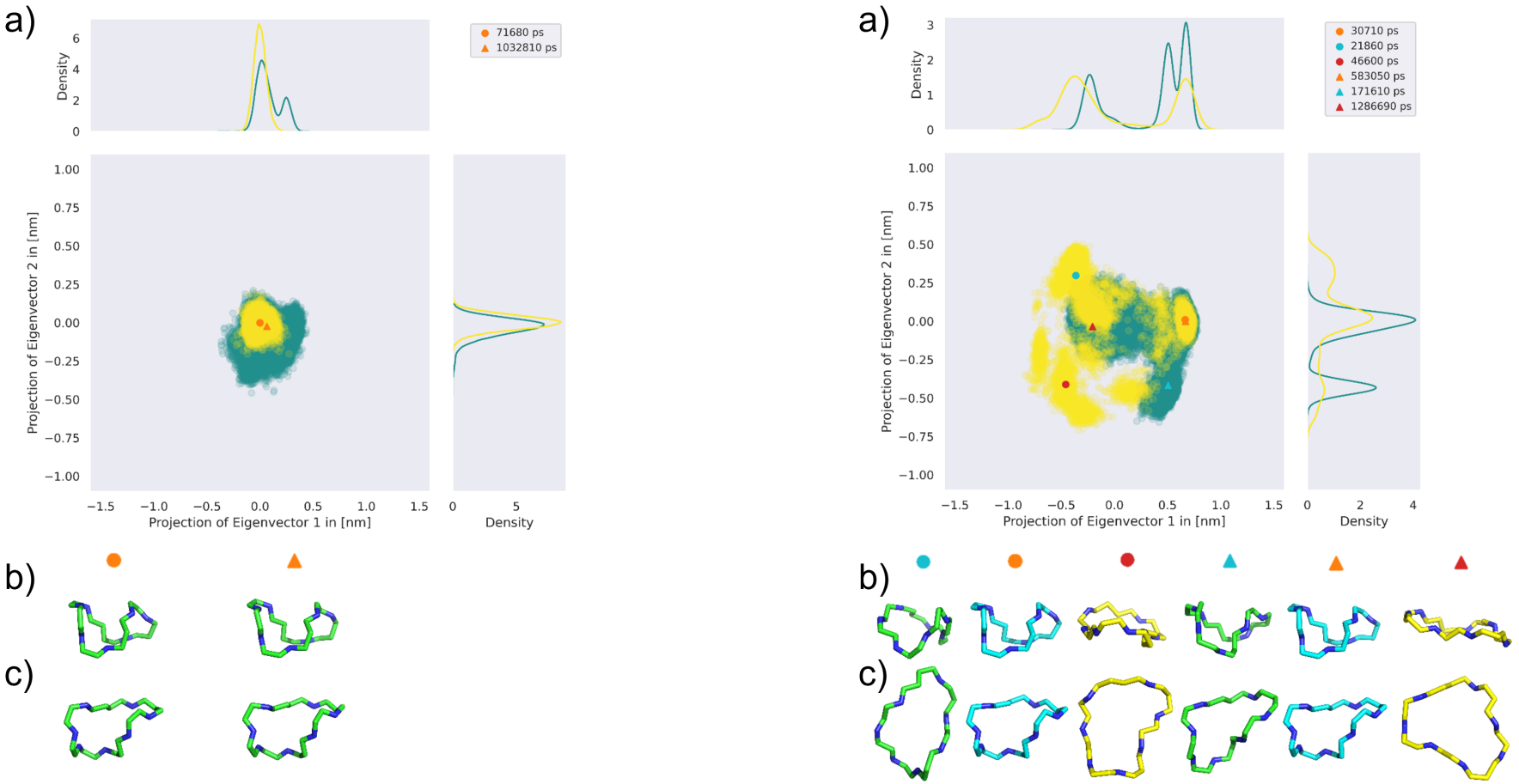


[Amba5]-MC-LY(Prg)-GSH-R



[Amba5]-MC-LY(Prg)-GSH-S

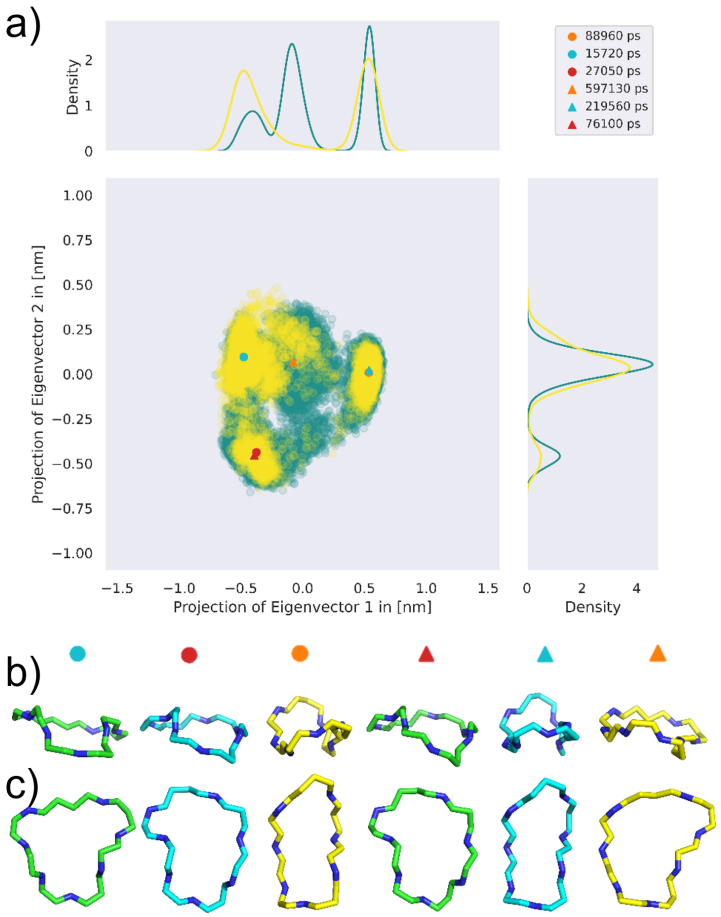
Figure 8.35: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.



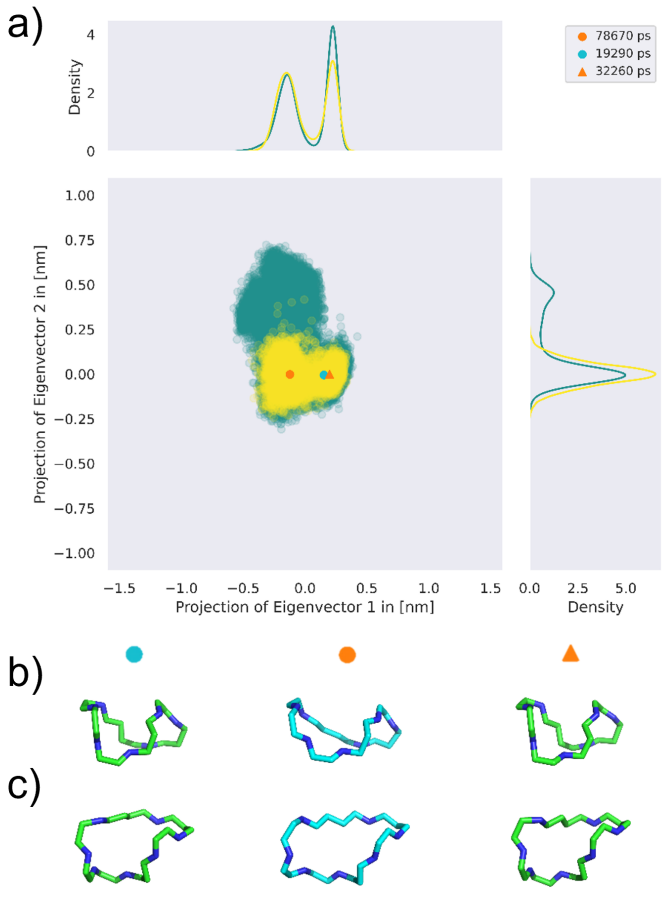
[Apha5]-MC-LF-Cys-R

[Apha5]-MC-LF-Cys-S

Figure 8.36: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

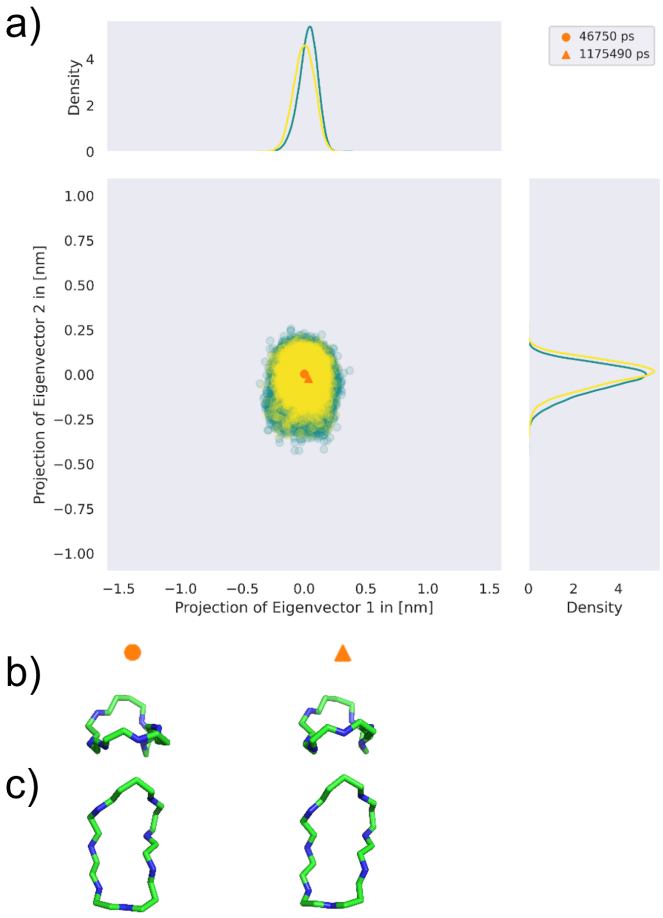


[Apha5]-MC-LF-GSH-R

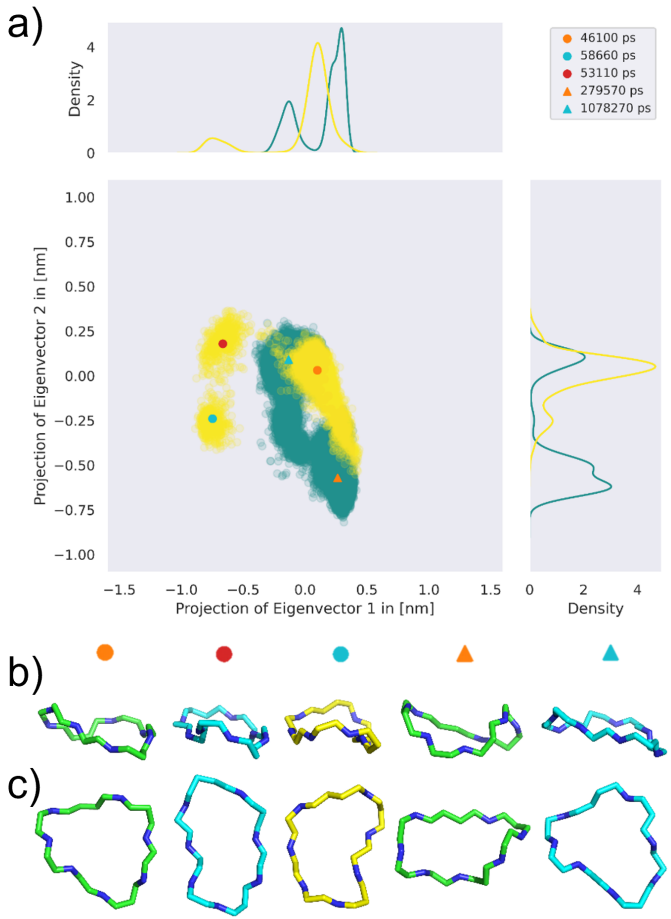


[Apha5]-MC-LF-GSH-S

Figure 8.37: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

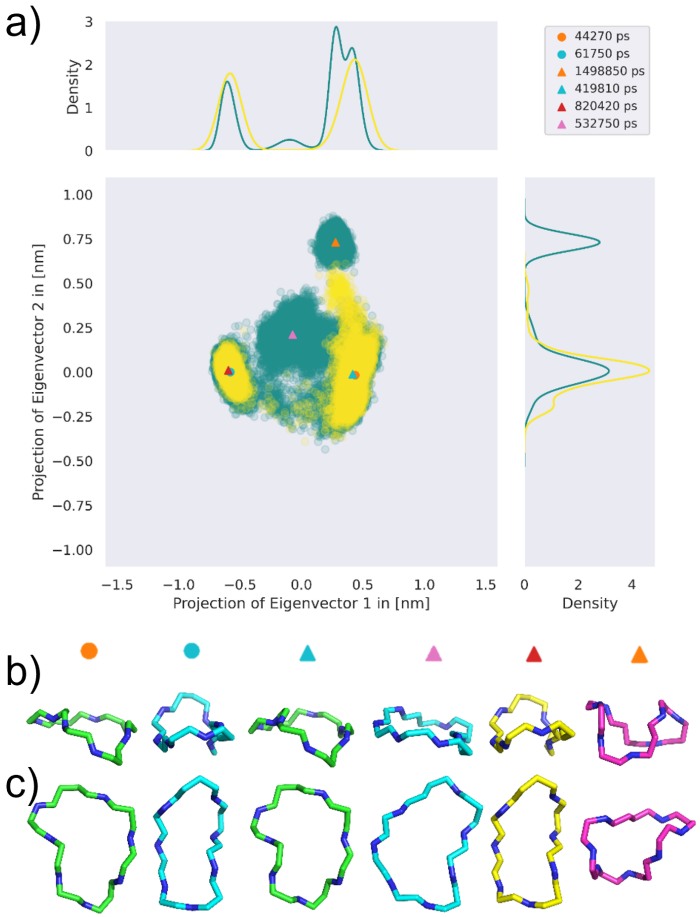


[Apda5]-MC-LF-Cys-R

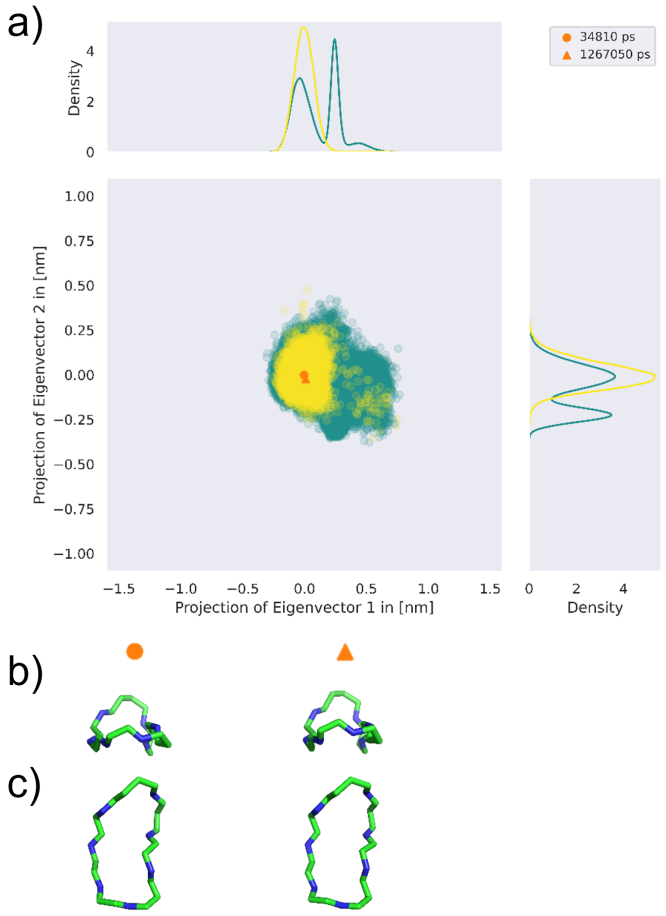


[Apda5]-MC-LF-Cys-S

Figure 8.38: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.



[Apda5]-MC-LF-GSH-R



[Apda5]-MC-LF-GSH-S

Figure 8.39: Principal components of MC congener backbone of complex (green) and solvent (yellow) simulations. a) The 2D projections are visualised as a central scatter plot, with distribution lines at the top/right to highlight the frequency of individual conformations. The circle and triangle are representative structures in solvent and complex simulation, respectively. The sticks represent the MC congener backbone of different clusters, as indicated by the symbols in b) side and c) top view.

	Replicate 1	Replicate 2	Replicate 3
MC-LR-Cys-R			
MC-LR-Cys-S			
MC-LR-GSH-R			
MC-LR-GSH-S			
MC-LF-Cys-R			
MC-LF-Cys-S			
MC-LF-GSH-R			
MC-LF-GSH-S			
[Enantio-Adda5]MC-LF-Cys-R			
[Enantio-Adda5]MC-LF-Cys-S			
[Enantio-Adda5]MC-LF-GSH-R			
[Enantio-Adda5]MC-LF-GSH-S			
MC-RR-Cys-R			
MC-RR-Cys-S			
MC-RR-GSH-R			
MC-RR-GSH-S			
MC-LY-Cys-R			
MC-LY-Cys-S			
MC-LY-GSH-R			
MC-LY-GSH-S			

Figure 8.40: Summary of qualitative evaluation of binding stability of MC congeners conjugates (part I) in PPP1. For each MC congener, the individual replicates are shown. Blue colour indicates stable binding, orange colour indicates diffusion out of the binding site.

	Replicate 1	Replicate 2	Replicate 3
MC-YR-Cys-R			
MC-YR-Cys-S			
MC-YR-GSH-R			
MC-YR-GSH-S			
MC-LY(Prg)-Cys-R			
MC-LY(Prg)-Cys-S			
MC-LY(Prg)-GSH-R			
MC-LY(Prg)-GSH-S			
[Anda5]MC-LY(Prg)-Cys-R			
[Anda5]MC-LY(Prg)-Cys-S			
[Anda5]MC-LY(Prg)-GSH-R			
[Anda5]MC-LY(Prg)-GSH-S			
[Amba5]MC-LY(Prg)-Cys-R			
[Amba5]MC-LY(Prg)-Cys-S			
[Amba5]MC-LY(Prg)-GSH-R			
[Amba5]MC-LY(Prg)-GSH-S			
[Apha5]MC-LF-Cys-R			
[Apha5]MC-LF-Cys-S			
[Apha5]MC-LF-GSH-R			
[Apha5]MC-LF-GSH-S			
[Apda5]MC-LF-Cys-R			
[Apda5]MC-LF-Cys-S			
[Apda5]MC-LF-GSH-R			
[Apda5]MC-LF-GSH-S			

Figure 8.41: Summary of qualitative evaluation of binding stability of MC congeners conjugates (part II) in PPP1. For each MC congener, the individual replicates are shown. Blue colour indicates stable binding, orange colour indicates diffusion out of the binding site.

8.3 Interaction Fingerprints for Molecular Dynamics Simulation

Table 8.10: Summary of criteria for protein-ligand interaction. Table adjusted after de Freitas and Schapira [417]. The chemical elements are abbreviated by their chemical symbol. θ is the planar angle between two rings, or one ring and an amide bond.

Interaction	Protein atom	Ligand atom	Distance (Å)	Angle (°)
Hydrophobic	C ali	C aro	≤ 4.0	-
	C aro	C ali		
	C ali	C ali		
	C	X		
	S	C aro		
Hydrogen bonding	O neg	OH	≤ 3.9	≥ 90
	OH	O neg		
	O	OH		
	OH	O		
	NH	O		
	O	NH		
	NH	O neg		
	O neg	NH		
	NH pos	O		
	O	NH pos		
	NH	N		
	O wat	O		
	O wat	N		
Stacking (edge to edge)	C aro	C aro	≤ 4.0	$60 \leq \theta \leq 120$
Stacking (edge to face)	C aro	C aro	≤ 4.0	≤ 30 or ≥ 150
Weak hydrogen bond	O	CH aro	≤ 3.6	≥ 130
	CH aro	O		
	CH ali	O		
	O	CH ali		
Salt bridge	N pos	O pos	≤ 4.0	-
	O neg	N pos		
Amide Stacking (face to face)	C amide	C aro	≤ 4.0	$30 \leq$ or ≥ 150
Amide Stacking (edge to face)	C amide	C aro	≤ 4.0	$60 \leq \theta \leq 120$
$Cation - \pi$	N pos	C aro	≤ 4.0	-
	C aro	N pos		
Halogen Bonding	O	Cl	≤ 3.47	$130 \leq \alpha 1 \leq 180$ $90 \leq \alpha 2 \leq 150$
	N	Cl	≤ 3.5	
	S	Cl	≤ 3.75	
	O	Br	≤ 3.57	
	N	Br	≤ 3.6	
	S	Br	≤ 3.85	
	O	I	≤ 3.7	
	N	I	≤ 3.73	
	S	I	≤ 3.98	
Multipolar halogen interaction	C amide	F	≤ 3.37	$\Theta 1 \geq 140$ $70 \leq \Theta 2 \leq 110$
	N amide	F	≤ 3.22	
	C amide	Cl	≤ 3.65	
	N amide	Cl	≤ 3.50	

Table 8.11: Summary of atom selection syntax in MDAnalysis for specific atom types. The chemical elements are abbreviated by their chemical symbol.

Atom	selection syntax
NH pos	(type H) and (bonded smarts [N+])
NH	(type H) and (bonded type N) and (not bonded smarts [N+])
N pos	smarts [N+]
N	type N and (not smarts [N+])
N amide	(type N) and (smarts [NX3][CX3](=[OX1])[#6])
C aromatic	smarts c
C aliphatic	smarts C
C	type C
C amide	(type C) and (smarts [NX3][CX3](=[OX1])[#6]) and (bonded type O)
CH aromatic	(type H) and (bonded smarts c)
CH aliphatic	(type H) and (bonded smarts C)
OH	(type H) and (bonded type O)
O negative	smarts [O-]
O	(type O) and (not smarts [O-])
Anion	smarts [-]
S	(type S)
Metal	(type CA) or (type CD) or (type CO) or (type CU) or (type FE) or (type MG) or (type MN) or (type NI) or (type ZN)
X	(type CL) or (type BR) or (type I) or (type F)
Cl	type CL
Br	type BR
I	type I
F	type F
O water	(resname HOH) and (type O)

Table 8.12: Number of interactions and number of IFPs without filtering. The IFPs have been aggregated by structure or time. In addition, the number of interactions derived from $aggregated_{occ30}$ frame is compared to the original number of interactions.

Interaction	MC-LR	MC-LF	[Enantio-Adda5] MC-LF	[β-D-Asp3,Dhb7] MC-RR
HOH8 HBAcceptor	0.62	-	-	-
ARG96 HBAcceptor	0.35	0.55	0.84	0.56
ARG96 Anionic	0.38	0.59	0.85	0.45
HIS125 Hydrophobic	-	0.33	-	-
CYS127 Hydrophobic	0.35	0.38	-	-
ILE130 Hydrophobic	0.86	0.99	0.70	0.92
ILE133 Hydrophobic	-	0.38	0.35	0.42
TYR134 HBAcceptor	-	0.59	-	0.36
TYR134 Hydrophobic	0.33	0.56	-	0.50
VAL195 Hydrophobic	0.47	0.50	-	-
ASP197 Hydrophobic	0.43	0.43	-	-
TRP206 Hydrophobic	0.91	0.73	0.63	0.95
TRP206 PiStacking	-	0.39	-	-
ARG221 Hydrophobic	-	0.33	-	-
VAL223 Hydrophobic	0.74	0.58	0.43	0.77
HIS248 Hydrophobic	-	0.40	-	-
VAL250 Hydrophobic	0.59	0.71	-	0.58
VAL250 HBAcceptor	-	0.31	-	-
GLU252 HBDonor	-	-	-	0.33
GLU252 Cationic	-	-	-	0.35
TYR272 Hydrophobic	0.60	-	0.72	0.59
TYR272 HBAcceptor	0.47	-	-	-
CYS273 Hydrophobic	-	-	0.30	-
PHE276 Hydrophobic	0.36	-	-	
MN400 VdWContact	-	0.34	-	-
MN401 VdWContact	-	0.67	-	-
HOH402 HBAcceptor	-	0.35	-	-

Table 8.13: Comparison of IFPs between different MC congener simulations. The similarity (x) was calculated by inverting the Rogers-Tanimoto dissimilarity. An IFP is considered identical if $x \geq 0.975$, similar if $0.85 \leq x < 0.975$, and dissimilar if $x \leq 0.5$. Please note that the percentages do not add up to 100 %, because IFP can belong to several classes. For example, if one IFP is close to another one nearby, it can be still dissimilar from another IFP at the other end of the simulations. The numbers still do reflect the proportions.

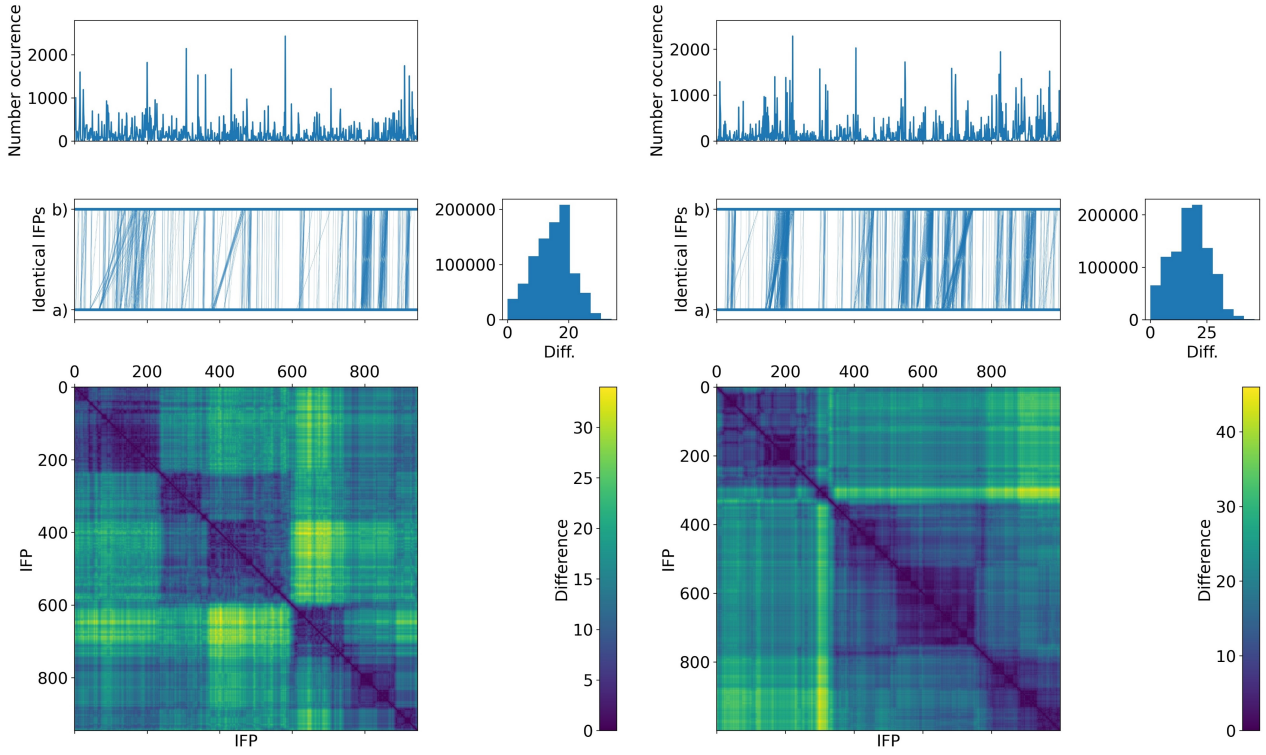
MC congener	Identical		Similar		Dissimilar	
	Number	Percent	Number	Percent	Number	Percent
MC-LR & MC-LF	0	0.00	26913	18.12	127550	85.89
MC-LR & [Enantio-Adda5]MC-LF	231	0.16	25055	16.87	43226	29.11
MC-LR & [β-D-Asp3,Dhb7]MC-RR	0	0.00	7892	5.31	84159	56.67
MC-LF & [Enantio-Adda5]MC-LF	0	0.0	4093	2.76	124361	83.74
MC-LF & [β-D-Asp3,Dhb7]MC-RR	0	0.00	8689	5.85	115771	77.96
[Enantio-Adda5]MC-LF & [β-D-Asp3,Dhb7]MC-RR	0	0.00	5659	3.81	21003	14.14

Table 8.14: Comparison of IFPs between different MC congener simulations. The similarity (x) was calculated by inverting the Rogers-Tanimoto Dissimilarity. An IFP is considered identical if $x \geq 0.975$, similar if $0.85 \leq x < 0.975$, and dissimilar if $x \leq 0.5$. Please note that the percentages do not add up to 100 %, because IFP can belong to several classes. For example, if one IFP is close to another one nearby, it can be still dissimilar from another IFP at the other end of the simulations. The numbers still do reflect the proportions.

MC congener	Identical		Similar		Dissimilar	
	Number	Percent	Number	Percent	Number	Percent
MC-LR-MC-LF	1230	0.21	104224	17.46	0	0.0
MC-LR-[Enantio-Adda5]MC-LF	0	0.0	159993	26.8	0	0.0
MC-LR-[β-D-Asp3,Dhb7]MC-RR	0	0.0	15489	2.59	0	0.0
MC-LR-MC-RR	0	0.0	69453	11.63	0	0.0
MC-LR-MC-LY	0	0.0	49493	8.29	0	0.0
MC-LR-MC-YR	0	0.0	43740	7.33	0	0.0
MC-LR-MC-LY(Prg)	0	0.0	76666	12.84	0	0.0
MC-LR-[Anda5]-MC-LY(Prg)	0	0.0	164903	27.62	0	0.0
MC-LR-[Amba5]-MC-LY(Prg)	0	0.0	200213	33.54	0	0.0
MC-LR-[Apha5]-MC-LF	0	0.0	71138	11.92	0	0.0
MC-LR-[Apda5]-MC-LF	645	0.11	122777	20.57	0	0.0
MC-LF-[Enantio-Adda5]MC-LF	2213	0.37	62583	10.48	0	0.0
MC-LF-[β-D-Asp3,Dhb7]MC-RR	0	0.0	20540	3.44	0	0.0
MC-LF-MC-RR	0	0.0	34413	5.76	0	0.0
MC-LF-MC-LY	0	0.0	39079	6.55	0	0.0
MC-LF-MC-YR	0	0.0	11148	1.87	0	0.0
MC-LF-MC-LY(Prg)	0	0.0	20345	3.41	0	0.0
MC-LF-[Anda5]-MC-LY(Prg)	0	0.0	158789	26.6	0	0.0
MC-LF-[Amba5]-MC-LY(Prg)	7551	1.26	183428	30.72	0	0.0
MC-LF-[Apha5]-MC-LF	0	0.0	100954	16.91	0	0.0
MC-LF-[Apda5]-MC-LF	1372	0.23	117866	19.74	0	0.0
[Enantio-Adda5]MC-LF-[β-D-Asp3,Dhb7]MC-RR	0	0.0	86151	14.43	0	0.0
[Enantio-Adda5]MC-LF-MC-RR	1876	0.31	131523	22.03	0	0.0
[Enantio-Adda5]MC-LF-MC-LY	0	0.0	52946	8.87	0	0.0
[Enantio-Adda5]MC-LF-MC-YR	0	0.0	47972	8.04	0	0.0
[Enantio-Adda5]MC-LF-MC-LY(Prg)	0	0.0	67624	11.33	0	0.0
[Enantio-Adda5]MC-LF-[Anda5]-MC-LY(Prg)	3509	0.59	228764	38.32	0	0.0
[Enantio-Adda5]MC-LF-[Amba5]-MC-LY(Prg)	53839	9.02	289943	48.57	0	0.0
[Enantio-Adda5]MC-LF-[Apha5]-MC-LF	0	0.0	47494	7.96	0	0.0
[Enantio-Adda5]MC-LF-[Apda5]-MC-LF	7803	1.31	232902	39.01	0	0.0
[β-D-Asp3,Dhb7]MC-RR-MC-RR	0	0.0	90398	15.14	0	0.0
[β-D-Asp3,Dhb7]MC-RR-MC-LY	0	0.0	32676	5.47	0	0.0
[β-D-Asp3,Dhb7]MC-RR-MC-YR	0	0.0	55896	9.36	0	0.0
[β-D-Asp3,Dhb7]MC-RR-MC-LY(Prg)	0	0.0	50079	8.39	0	0.0
[β-D-Asp3,Dhb7]MC-RR-[Anda5]-MC-LY(Prg)	0	0.0	117158	19.62	0	0.0
[β-D-Asp3,Dhb7]MC-RR-[Amba5]-MC-LY(Prg)	0	0.0	180008	30.15	0	0.0
[β-D-Asp3,Dhb7]MC-RR-[Apha5]-MC-LF	0	0.0	2801	0.47	0	0.0
[β-D-Asp3,Dhb7]MC-RR-[Apda5]-MC-LF	0	0.0	61285	10.27	0	0.0
MC-RR-MC-LY	0	0.0	6792	1.14	0	0.0
MC-RR-MC-YR	0	0.0	37135	6.22	0	0.0
MC-RR-MC-LY(Prg)	0	0.0	31385	5.26	0	0.0
MC-RR-[Anda5]-MC-LY(Prg)	0	0.0	83098	13.92	0	0.0
MC-RR-[Amba5]-MC-LY(Prg)	742	0.12	213163	35.7	0	0.0
MC-RR-[Apha5]-MC-LF	0	0.0	2107	0.35	0	0.0
MC-RR-[Apda5]-MC-LF	0	0.0	124802	20.9	0	0.0

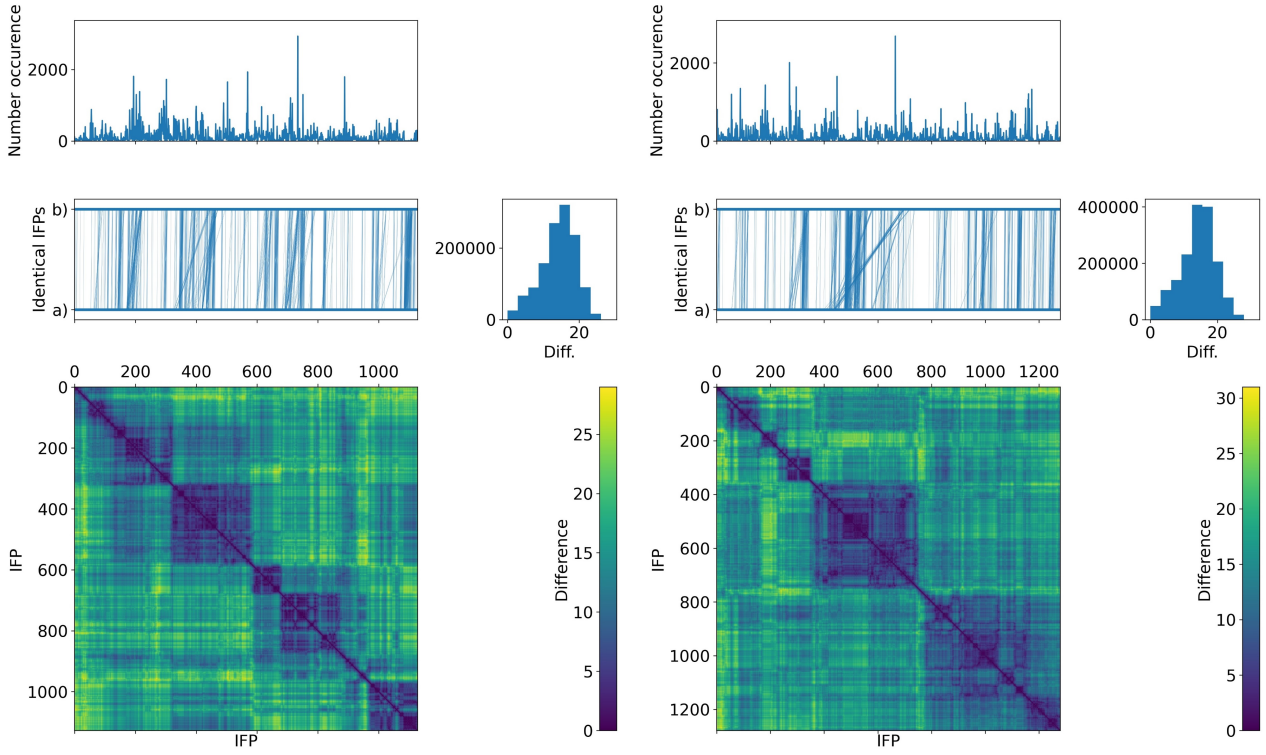
Table 8.14: Continued from previous.

MC congener	Identical		Similar		Dissimilar	
	Number	Percent	Number	Percent	Number	Percent
MC-LY-MC-YR	0	0.0	30037	5.03	0	0.0
MC-LY-MC-LY(Prg)	0	0.0	99922	16.74	0	0.0
MC-LY-[Anda5]-MC-LY(Prg)	0	0.0	164502	27.55	0	0.0
MC-LY-[Amba5]-MC-LY(Prg)	0	0.0	153493	25.71	0	0.0
MC-LY-[Apha5]-MC-LF	0	0.0	27589	4.62	0	0.0
MC-LY-[Apda5]-MC-LF	0	0.0	55270	9.26	0	0.0
MC-YR-MC-LY(Prg)	0	0.0	1932	0.32	0	0.0
MC-YR-[Anda5]-MC-LY(Prg)	0	0.0	65317	10.94	0	0.0
MC-YR-[Amba5]-MC-LY(Prg)	0	0.0	108845	18.23	0	0.0
MC-YR-[Apha5]-MC-LF	0	0.0	6047	1.01	0	0.0
MC-YR-[Apda5]-MC-LF	0	0.0	68942	11.55	0	0.0
MC-LY(Prg)-[Anda5]-MC-LY(Prg)	0	0.0	83641	14.01	0	0.0
MC-LY(Prg)-[Amba5]-MC-LY(Prg)	0	0.0	95790	16.04	0	0.0
MC-LY(Prg)-[Apha5]-MC-LF	0	0.0	27244	4.56	0	0.0
MC-LY(Prg)-[Apda5]-MC-LF	0	0.0	133419	22.35	0	0.0
[Anda5]-MC-LY(Prg)-[Amba5]-MC-LY(Prg)	9217	1.54	267947	44.88	0	0.0
[Anda5]-MC-LY(Prg)-[Apha5]-MC-LF	0	0.0	220400	36.92	0	0.0
[Anda5]-MC-LY(Prg)-[Apda5]-MC-LF	3812	0.64	220138	36.87	0	0.0
[Amba5]-MC-LY(Prg)-[Apha5]-MC-LF	0	0.0	144151	24.15	0	0.0
[Amba5]-MC-LY(Prg)-[Apda5]-MC-LF	14943	2.5	256233	42.92	0	0.0
[Apha5]-MC-LF-[Apda5]-MC-LF	0	0.0	99179	16.61	0	0.0



(a) MC-LF

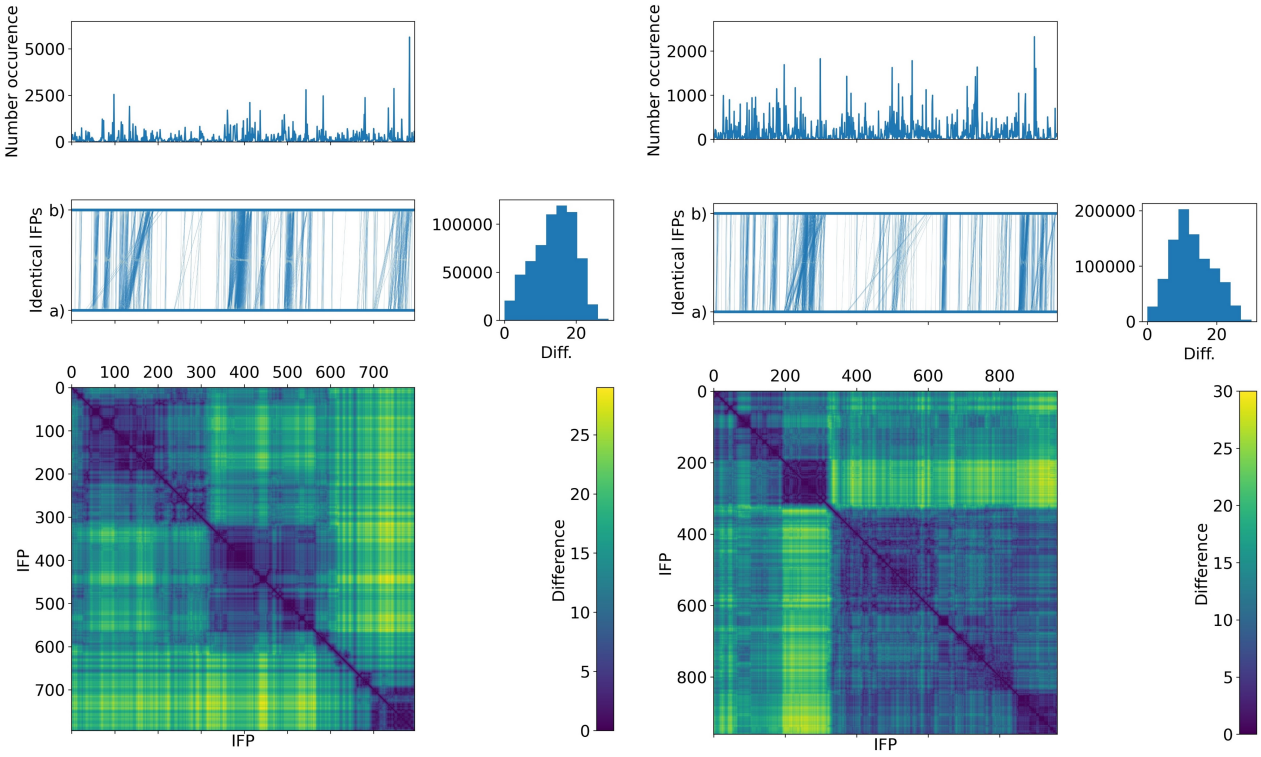
(b) [β-D-Asp3,Dhb7]MC-RR



(c) MC-YR

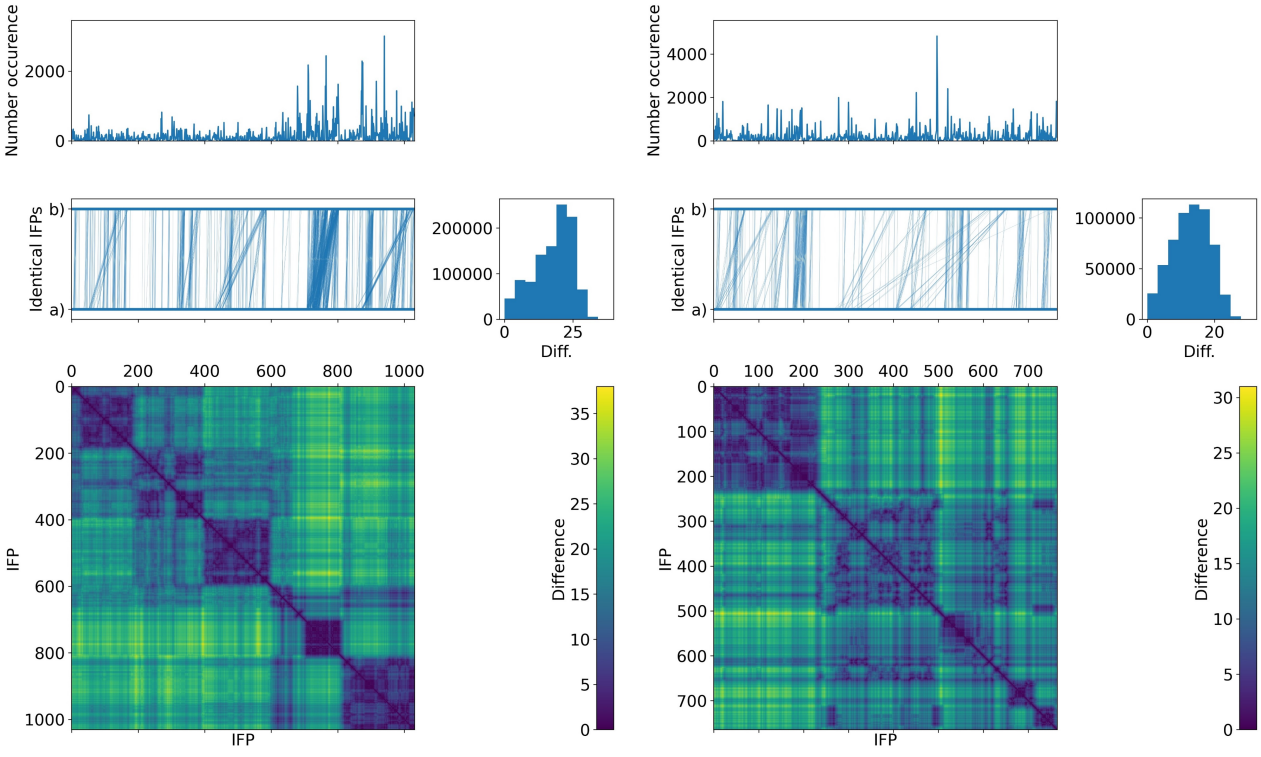
(d) MC-LY(Prg)

Figure 8.42: Comparison of IFP similarity within MC congener. The occurrence and identical IFPs (connected by vertical lines) as line plot. The number of differences to all IFPs are shown as histogram (top right), and as colour on the matrix visualisation.



(e) [Anda5]-MC-LY(Prg)

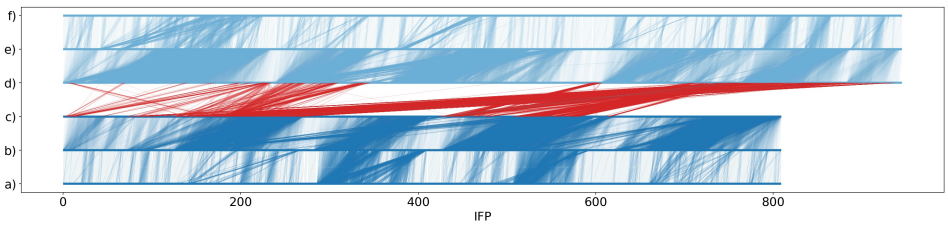
(f) MC-LY



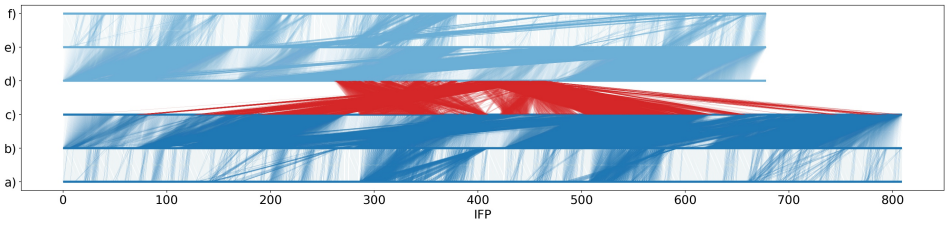
(g) [Apda5]MC-LF

(h) [Amba5]MC-LY(Prg)

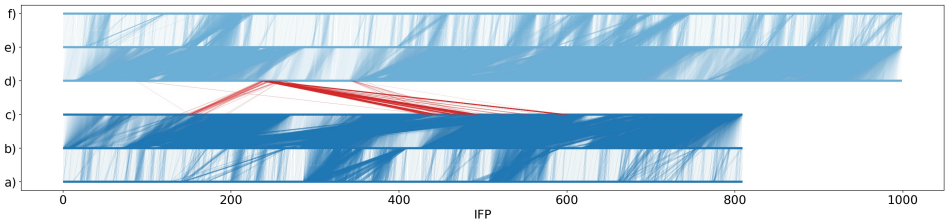
Figure 8.42: Continued from previous figure.



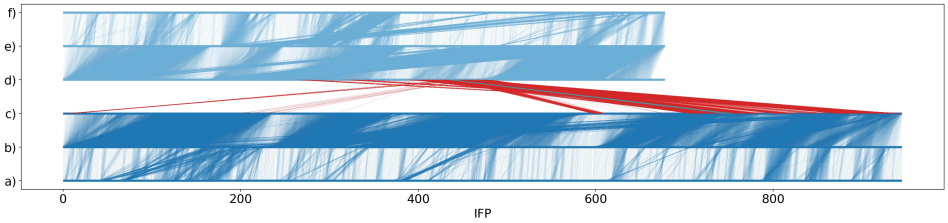
(a) MC-LR & MC-LF



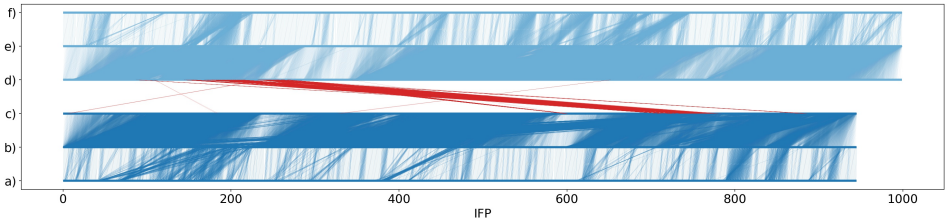
(b) MC-LR & [Enantio-Adda5]MC-LF



(c) MC-LR & [β-D-Asp3,Dhb7]MC-RR

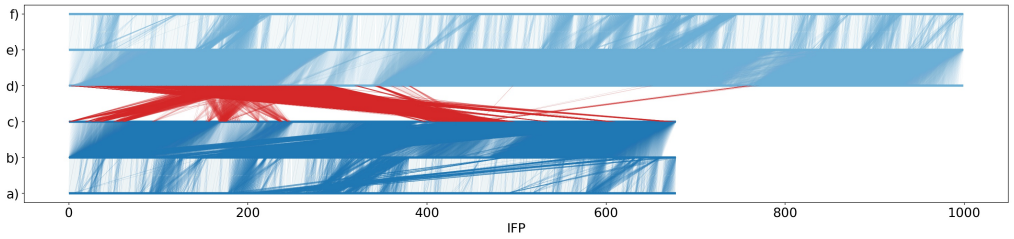


(d) MC-LF & [Enantio-Adda5]MC-LF



(e) MC-LF & [β-D-Asp3,Dhb7]MC-RR

Figure 8.43: Similarity calculation between IFPs of different simulation and networks of major IFPs. a) The first MC congener is shown in dark blue, the second in light blue. Between *a*-*b* and *e*-*f*, identical IFPs within the same simulation are shown. Between *b*-*c* and *d*-*e*, similar IFPs within the same simulation are shown. Between *c*-*d* identical and similar IFPs between simulations are shown as cyan and red lines, respectively.



(f) [Enantio-Adda5]MC-LF & [β-D-Asp3,Dhb7]MC-RR

Figure 8.43: Continued from previous.