A different type of "microscope": Using MD simulations to understand the conformational dynamics of bioactive peptides with non-trivial conformational ensembles

---

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy

of Freie Universität Berlin

by

MARIUS T. WENZ

Berlin

December 2022

The research of this thesis was carried out under the direction of Prof. Dr. Bettina G. Keller at the Institute of Chemistry and Biochemistry between May 2019 and December 2022.

1st reviewer **Prof. Dr. Bettina G. Keller**
Freie Universität Berlin
Department of Biology, Chemistry Pharmacy
Institute of Chemistry and Biochemistry
Physical & Theoretical Chemistry
Arnimallee 22
14195 Berlin

2nd reviewer **Prof. Dr. Roderich D. Süssmuth**
Technische Universität Berlin
Fakulty II - Mathematics and Natural Sciences
Department of Chemistry, Sekr. TC 2
Straße des 17. Juni 124
10623 Berlin

Date of defense: 23.03.2023

ACKNOWLEDGEMENT

First and foremost, I would like to express my deep gratitude to my first supervisor Prof. Dr. Bettina G. Keller for giving me the opportunity to enter the intriguing world of Computational Science. Since I joined her group in 2018 for my Bachelor's thesis and especially throughout my doctoral studies, she has been teaching me how to ask the right questions and how to distinguish good from less-suited solution strategies. Within this context, I wish to thank her for our fruitful discussions and her trust in my abilities. I am very impressed by her broad knowledge in the field, and her career inspires me.

Next, I would like to thank my second supervisor Prof. Dr. Roderich D. Süssmuth for keeping me in touch with fascinating Chemistry and for believing in the symbiosis of experiments and theory. It was a pleasure to be "intoxicated" by our joint research project. Also, I greatly admire his commitment to designing a student-friendly graduate school, and I thank him for allowing me to learn more about scientific committees during my time as student representative.

I wish to thank my Prof. Dr. Beate Paulus as my third mentor, who always strives to create a good atmosphere in our institute and is approachable for urgent matters at any time.

Next, I would like to thank my co-workers from Roderich's group, in particular Dr. Simone Kosol and Dr. Guiyang Yao, for their efforts and the exciting discussions. Their insights have been offering me a change of perspective that broadened my horizon and that was often very helpful for the thrive of my investigations.

Furthermore, I thank Prof. Dr. Christian Freund and his group, especially Dr. Miriam Bertazzon and Dr. Jana Sticht, for the perseverance they showed throughout our challenging "tandem WW project". In that regard, I do not want to leave unnoticed the time and effort my former colleague Dr. Stevan Aleksić had already spent when I entered the project.

I gratefully acknowledge the DFG for funding as well as the ZEDAT for technical support and computation time.

Also, I would like to thank our secretary Julija Djordjevic for being the permanent contact person for all beaurocratic issues.

Special thanks go to Dr. Simone Kosol, Dr. Oliver Lemke, Dr. Jan-Oliver Kapp-Joswig and Dr. Luca Donati for proofreading of this thesis.

Warm thanks go to Maria Seidel for her wonderful organisation of our graduate school and her constant efforts to achieve the best for the PhD students.

Also, I am deeply grateful to Luise Franz for sharing the responsibility of being student representative with me. It was a pleasure to organise scientific events such as conferences

and retreats with you!

Over the last years, the graduate school “GRK 2473” has been an integral part of my time as PhD student. I would like to thank all PhD students who have invested their time to make it enjoyable.

Moreover, I will keep many good memories from my time in the Takustr. 3/Arnimallee 22 thanks to the family atmosphere in the Theoretical Chemistry department and events like Christmas parties, after-work beers or movie evenings. In particular, I wish to mention my closest colleagues: Stefanie Kieninger, Jennifer Anders, Marco Manni, Simon Petry, Jan-Oliver Kapp-Joswig, Felix Witte and Oliver Lemke.

Finally, I would like to thank my friends, my family and my girlfriend Maureen who are always there for me and on whom I can rely in difficult times. You supported me in the last years and you believed in me. Without you, I would certainly not be where I am now. I am very fortunate to have you in my life.

STATUTORY DECLARATION

I, Marius T. Wenz, hereby assure that I have independently written this thesis entitled:

A different type of "microscope": Using MD simulations to understand the conformational dynamics of bioactive peptides with non-trivial conformational ensembles

I did not use any other than the declared sources, and I have explicitly marked all material which has been quoted either literally or by content from the used sources. This thesis, in a similar or identical form, has been neither submitted to an Examination Office nor published elsewhere before.

Berlin, den 22.12.2022 _____

Marius T. Wenz

Contents

LIST OF PUBLICATIONS

Part A: tandem WW domain project

A.1 "Target Recognition in Tandem WW Domains: Complex Structures for Parallel and Antiparallel Ligand Orientation in h-FBP21 Tandem WW"
M. T. Wenz†, M. Bertazzon, J. Sticht, S. Aleksić, D. Gjorgjevikj, C. Freund and B. G. Keller
J. Chem. Inf. Model **2021**
doi: 10.1021/acs.jcim.1c01426

Part B: Amanitin project

B.1 "Iodine-Mediated Tryptathionine Formation Facilitates the Synthesis of Amanitins"
G. Yao†, C. H. Knittel†, S. Kosol, M. T. Wenz, B. G. Keller, H. Gruß, A. C. Braun, C. Lutz, T. Hechler, A. Pahl and R. D. Süssmuth
J. Am. Chem. Soc. **2021**
doi: 10.1021/jacs.1c06565

B.2 "The Occurrence of Ansamers in the Synthesis of Cyclic Peptides"
G. Yao†, S. Kosol†, M. T. Wenz†, E. Irran, B. G. Keller, O. Trapp and R. D. Süssmuth
Nat.Commun. **2022**
doi: 10.1038/s41467-022-34125-8

B.3 "The influence of N-methylation on the ansamers of an amatoxin: Gly5Sar-amanullin"
M.T. Wenz†, S. Kosol, G. Yao, R. D. Süssmuth and B. G. Keller
BioRxiv preprint server **2022**
doi: 10.1101/2022.12.21.521444

First-authorship is labelled "†". A **detailed description of my contribution** to the above-mentioned publications is given together with their summaries.

LIST OF ABBREVIATIONS

CommonNN	"common-nearest-neighbour" clustering.
DHIle	(3*R*,4*R*)-4,5-dihydroxyisoleucine.
h-FBP21	*human*-formin binding protein 21.
MD	molecular dynamics.
MSM	Markov State Model.
NMR	nuclear magnetic resonance.
NOE	Nuclear Overhauser Effect.
PDB	Protein Data Bank.
PPI	protein-protein interaction.
PRM	proline-rich motif.
SAR	structure-activity relationship.
TICA	time-independent component analysis.
tWW	*tandem* WW domain.

ABSTRACT

Protein-protein interactions (PPI) are the fundamental driving forces for many biological processes ranging from intercellular communication to the functioning of complex machineries such as the ribosome. To understand and predict PPIs, the occurring complexes and resulting biological functions, a structural-dynamical perspective is often required. Molecular dynamics simulations are a powerful tool that can provide this insight with atomistic precision, and can be pivotal for the understanding of biomolecules. In this thesis, we investigate two aspects of PPI research within the framework of two projects: Firstly, we study the recognition event that mediates the PPI between h-FBP21 and the spliceosomal core Sm protein B/B'. In this recognition event, a tandem WW domain of h-FBP21 selectively binds to proline-rich sequences of Sm B/B'. Here, we will create and characterise the complex structure of h-FBP21 tWW and a selected peptide sequence of Sm B/B' based on MD simulations and experimental data. We devise a workflow, which allows the systematic investigation of binding structures from recognition domains in tandem repeats.

In the second project, we investigate strong RNA polymerase II inhibitors: the amatoxins. Amatoxins are a family of cyclic octapeptides featuring a tryptathionine bridge. In combination, the strong inhibition capacity and the exceptionally stable structure make amatoxins very attractive for the design of anti-body drug conjugates (ADCs). To this end, it is important to understand the synthesis and possible structural modifications of the amatoxins. Here, we will present a novel synthetic route to establish the tryptathionine bridge. Using MD simulations and experimental data, we show that this allows us to selectively synthesise isomers for the amatoxin scaffold. To describe the isomers unambiguously, we propose a novel nomenclature, which might also hold for other cyclic peptides. Finally, we will examine N-methylation as a possible way to modify the amatoxins. We believe that our results could encourage further SAR-studies, which are needed to exploit the potential of the amatoxins as possible ADCs.

ZUSAMMENFASSUNG

Protein-Protein-Wechselwirkungen (PPI) sind die grundlegenden Antriebskräfte für viele biologische Prozesse, von der interzellulären Kommunikation bis hin zum Funktionieren komplexer Maschinen wie dem Ribosom. Um PPI, die entstehenden Komplexe und die daraus resultierenden biologischen Funktionen zu verstehen und vorherzusagen, ist häufig eine strukturell-dynamische Perspektive erforderlich. Molekulardynamiksimulationen sind ein leistungsfähiges Instrument, das diesen Einblick mit atomarer Präzision ermöglicht und für das Verständnis von Biomolekülen von entscheidender Bedeutung sein kann. In dieser Arbeit beleuchten wir zwei Aspekte der PPI-Forschung im Rahmen von zwei Projekten: Zunächst untersuchen wir das Bindungsereignis, das die PPI zwischen h-FBP21 und dem Sm Protein B/B', einem Bestandteil des Spliceosomes, vermittelt. Bei diesem Bindungsereignis bindet eine tandem-WW-Domäne von h-FBP21 selektiv an prolinreiche Sequenzen von Sm B/B'. In diesem Projekt werden wir die Komplexstruktur von h-FBP21 tWW und einer ausgewählten Peptidsequenz von Sm B/B' auf der Grundlage von MD-Simulationen und experimentellen Daten erstellen und charakterisieren. Wir entwickeln darüber hinaus einen Arbeitsablauf, der es uns ermöglicht, die Komplexstruktur von paarweise auftretenden und funktionell der h-FBP21 tWW ähnlichen Domänen systematisch zu bestimmen. Im zweiten Projekt untersuchen wir starke RNA-Polymerase-II-Inhibitoren: die Amatoxine. Amatoxine sind eine Familie von zyklischen Oktapeptiden mit einer Tryptathionin-Brücke. Ihre Eigen- schaft, die Funktionsweise von RNA-Polymerase II stark zu hemmen, und ihre außergewöhnlich stabile Struktur machen Amatoxine für die Entwicklung von Antikörper-Wirkstoff-Konjugaten (ADKs) sehr attraktiv. Zu diesem Zweck ist es wichtig, die Synthese und mögliche strukturelle Modifikationen der Amatoxine zu verstehen. In diesem Projekt stellen wir einen neuen Syntheseweg zur Herstellung der Tryptathionin-Brücke vor. Mit Hilfe von MD-Simulationen und experimentellen Daten zeigen wir, dass auf diese Weise Isomere für das Amatoxin-Gerüst selektiv synthetisiert werden können. Zur eindeutigen Beschreibung dieser Isomere schlagen wir eine neue Nomenklatur vor, die auch für andere zyklische Peptide gelten könnte. Schließlich werden wir die N-Methylierung als eine mögliche Methode zur strukturellen Modifizierung der Amatoxine untersuchen. Wir glauben, dass unsere Ergebnisse weitere SAR-Studien anregen könnten, die notwendig sind, um das Potenzial der Amatoxine als mögliche ADKs zu nutzen.

1 INTRODUCTION

1.1 Protein-protein interactions and bioactive peptides

Protein-protein interactions (PPIs) is the generic term for highly specific physical contacts between two or more proteins in the cellular environment. They are the fundament for almost all biological processes ranging from the operating principle of molecular machines,[1–3] like the ribosome[4] or the spliceosome,[5] to the information transfer in signalling pathways.[6,7] PPIs can be classified on the basis of the interacting proteins, the intended functionality or the lifetime of the arising complex.[8,9] Note, however, that these categories are not strictly defined, and many PPIs may well combine several, but not mutually exclusive properties. In fact, PPIs are reported to be massively impacted by their physiological environment, such as concentration of the complex components or the physicochemical environment (pH, ions, etc.).[9,10] PPIs in a cellular environment form a complex network (interactome), and, for example, the human interactome alone is expected to cover more than 400,000 PPIs.[11,12] Although regularities have been found that help to classify the different PPIs, many PPIs still remain undiscovered, possibly covering explanations for unsolved biological questions.

This lack of knowledge and the versatility of the PPIs become a problem when we look at diseases such as malaria, Altzheimer's or cancer, which involve PPIs as well.[12–14] In literature also pictured "reaching for a high-hanging fruit",[15] the development of possible therapeutic agents is a multi-problem challenge: PPIs are based on interaction interfaces, generally large, flat surfaces (1500-2000 $Å^2$) with increased hydrophobicity. These interaction interfaces were considered "undruggable" for a long time.[12,15,16] Also, not all hydrophobic contacts between two or more proteins are indeed PPIs. PPIs must therefore always be defined in a biological context.[17]

Here, the identification of "hot spots", i.e. residues that are pivotal for the affinity of the binding, was a major step forwards for the design of possible modulators.[11,15,18] Yet, the high-affinity binding between the proteins makes it difficult for modulators to compete. The situation is further complicated by the different possible ways to intercept PPIs. In literature,[12] *inhibition*, i.e. the ability to block an interaction interface for possible binding partners, is differentiated from *stabilisation*, in which a protein-protein complex is captured in the bound state. Both processes can be further discriminated into *orthosteric*, i.e. direct binding to the "hot spot", or *allosteric*, binding non-interacting areas. Although this finding offers opportunities for approaching a targeted PPI differently, one of the major challenges remains the lack of reference structures for binding-competent ligands.[12,15] The availability and the increasing precision of computational methods has been substantially enriching the research on PPIs.[19,20] As we will see later (see 3.3.2), virtual screening by docking assays, for instance,

can help to generate and optimise possible modulators.[18,20]

Within the context of drug development for PPIs, especially bioactive peptides have gained a lot of interest over the last years.[21–24] Bioactive peptides represent a huge group of compounds, which can be classified according to their properties (structures, interactions), origins (extraction, synthesis) and applications (therapeutics, tissue engineering).[25] This introduction does not do justice to the versatility of bioactive peptides, but for an excellent overview please refer to Refs. [23, 25–27]. There are some aspects, though, that are especially noteworthy within the context of this thesis. Compared with small molecule inhibitors, bioactive peptides bind their targets with higher affinity and higher selectivity while covering larger binding pockets.[12,22] This is very useful to deal with the promiscuous "hot spots" of the protein interaction interfaces. From a patient's perspective, however, bioactive peptides are controversial: Although their metabolites are reported less toxic than existing drugs, most bioactive peptides have the problem that they become proteolytically decomposed by proteases in the gastrointestinal tract.[22,27] Hence, after oral uptake, most peptides either reach their targets only in small concentrations or they have to be administered in high doses, which in turn can be problematic for the patient. Second, even after intravenous or subcutaneous administration, most peptides struggle with overcoming the cell membrane, which limits their use for intracellular targets.[23,27] Both drawbacks, however, can be counteracted by functionalisation with reversible ligands or by the addition of transporters. This concept of transporting peptides into the cell is especially relevant for the design of anti-body drug conjugates using cytotoxic peptides. Altogether, the versatility in their properties and possible modifications makes bioactive peptides very promising candidates for finely tuned therapeutic agents especially for PPIs.

1.2 Overview of this thesis

In this thesis, we will deal with *transient* PPIs, i.e. PPIs with frequently changing binding situations: First, we will study the domain-based mediation of a specific PPI in the context of the spliceosome ("tWW project"; for the publication, please refer to 4.1). Second, we will take a careful look on bioactive peptides that can inhibit and down-regulate the functioning of important working apparatus during the transcription process ("amatoxin project", for the publications, please refer to 4.2, 4.3, 4.4).

The remainder of this chapter presents the systems studied in this thesis. It begins with the recognition of proline-rich sequences by specific adaptor domains (see 1.3). This also involves a short introduction on the most important models for categorising binding events (see 1.3.2). Afterwards, the bioactive peptide family "amatoxins" (see 1.4) is introduced with a focus on their structure (see 1.4.1), toxicity (see 1.4.2), residue-specific properties (see 1.4.3),

synthesis (see 1.4.4) and relatives (see 1.4.5).

The next chapter sets out the research question of this thesis and the methodology to answer it (see 2). It follows a three-part theory chapter starting with classical molecular dynamics (MD) simulations (see 3.1) with a focus on the fundamental concepts (see 3.1.1), setups and dimensionality reduction (see 3.1.2), further analyses (see 3.1.3) and issues (see 3.1.4). The second part gives an introduction to nuclear magnetic resonance (NMR) spectroscopy (see 3.2.1) and their use for MD simulations (see 3.2.2). In the end, the additional computational tools clustering (see 3.3.1) and docking (see 3.3.2) are discussed.

Chapter 4 contains the publications that were produced as part of this thesis. Each publication is preceded by a brief summary of the most important contents and methods used.
This thesis concludes with an outlook (see 5).

1.3 Proline-recognition by WW domains

Molecular machines and signalling pathways have in common that they involve frequent rearrangements of the interacting partners.[28] The involved PPIs consequently have to do the balancing between being highly specific, yet also highly dynamic.[29,30] To achieve this, highly dynamic PPIs are often mediated by protein domains, i.e. highly conserved subunits that possess a high selectivity for specific peptide sequences in the partnering protein.[30,31] One example for such PPI-mediating domains are WW domains,[29,32,33] a family of protein modules that consist of 34 - 40 amino acids. WW domains comprise a triple-stranded β-sheet and owe their name to two highly conserved tryptophan (Trp) residues: One Trp residue is located in the β-sheet and essential for the folding of the WW domain. The other Trp residue is part of the binding site of the WW domain usually together with serine (Ser) and tyrosine (Tyr).

Over the years, the structure[34–36] as well as the binding behaviour[29,37–39] of single WW domains have been studied in great detail. WW domains selectively bind to proline-rich motifs (PRMs) in the targeted sequence, which can either be independent ligands or parts of larger proteins. Principally, two different binding orientations are possible:[29] a *parallel* arrangement, in which the N-termini of both ligand and WW domain are on the same side, as well as an *antiparallel* arrangement, in which the N-termini are on opposite sides.

Based on their ligand preference, WW domains can be classified according to five consensus amino acid sequences.[29,38] These sequences have in common that they comprise at least two successive proline-residues (Pro), which will be referred to as "PP-motif".[29] For this PP-motif, the residues of the binding site of the WW domains, usually Trp and Tyr, form a hydrophobic pocket, called "XP-groove". The recognition event between PRMs and WW domains is based on the PP-motif inserting into the XP-groove.[29,38] This binding arrangement

is further stabilised by hydrogen bonds between members of the XP groove and residues flanking the PP-motif. These hydrogen bonds are possible due to the exceptional backbone conformation of the PRM. The PP-motif forms a so-called "left-handed polyproline II helix", a very restricted structural arrangement, which causes the peptide backbone carbonyl groups to be solvent exposed and therefore highly responsive for hydrogen bonding. Besides, if the PP-motif is extended by another proline-residue, the resulting PPP-motif will have pseudo-C_3 rotational symmetry.

1.3.1 The tandem problem for WW domains

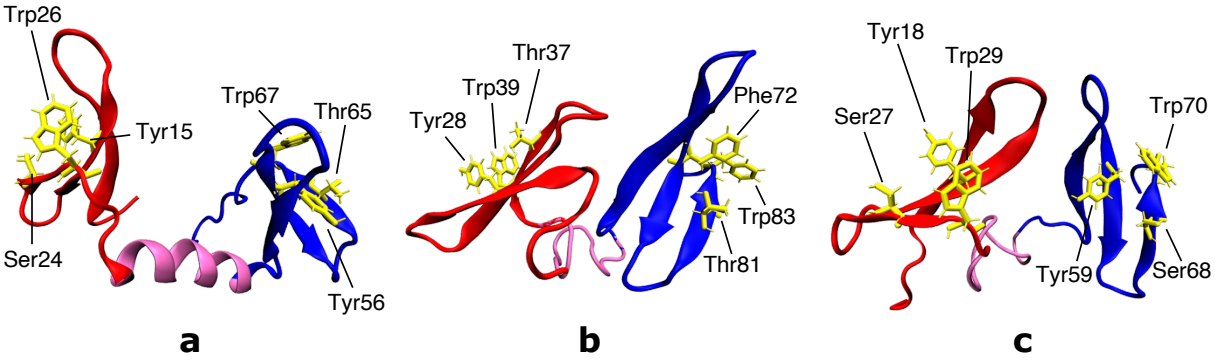


Figure 1.1: Solution structures of the *tandem* WW domains (tWWs) of (a) *yeast* splicing factor PRP40 (PDB entry: 1O6W [40]), (b) *Drosophila* suppressor of deltex Su(dx) (PDB entry: 1TK7 [41]) and (c) *human*-formin binding protein 21 (PDB entry: 2JXW [42]). The WW domains are coloured in red and blue, the interdomain region is depicted in purple. Residues that are relevant for binding PRMs are shown explicitly: they are coloured in yellow and labelled according to the residue type and index in the respective tWW structure. Only one representative structure is shown per tWW.

The versatility of WW domains is enhanced by the fact that they often occur in pairs, so-called "tandems". In these tWWs, the individual WW domains are linked by an interdomain region of varying length and flexibility. This interdomain region massively influences the structure, stability and binding behaviour of the tWW, and the few tWW structures known today raise questions rather than provide answers:[40,41,43–45] For example, the tWW of PRP40, a pre-mRNA processing protein (*yeast*),[40] contains a highly organised, α-helical interdomain region (see Figure 1.1.a). Fixed in their relative position by the rigidity of the α-helix, the two WW domains face in opposite directions. As a result, the PRP40 tWW is reported to either bind one large bridging ligand or multiple targets.[46] Other tWWs comprise flexible interdomain regions that allow for versatile arrangements of the included WW domains. Due to this flexibility, the individual WW domains cannot only bind different targets in diverse orientations, but they can also bind together to multivalent targets. It has been reported that the binding affinity of tWWs to multivalent targets can exceed the summed affinities of the individual domains.[43,44,47,48] This finding is possibly related to the ability of the individual WW domains to form a supramodule.[43] Interestingly, the formation of such a supramodule

does not necessitate direct interactions between the two WW domains.[43] In the tWWs of KIBRA, MAGI1-3 and SAV1, for instance, the WW domains form a supramodule, which is only stabilised by interactions from the interdomain region to structural elements directly following the second WW domain.[43,45]

There are reported cases, though, in which there are interactions between the WW domains: It was observed that WW domains might associate with each other and thereby block their binding sites for potential targets.[49] Hence, in terms of ligand binding, these WW domains cooperate negatively with each other. This is in contrast to what has been reported for the tWW of the suppressor of deltex (Su(dx), *Drosophila*, see Figure 1.1.b). Here, the ligand binding is expected to happen synergistically, which means that the binding of one WW domain enhances the binding ability of the second WW domain.[41,46,50] In an extreme case, the binding of one WW domain can induce the folding of the other WW domain to a binding-competent conformation ("folding-upon-binding").[50]

The tWW studied in project **A** (see Figure 1.1.c) contains two stably folded WW domains. Hence, a "folding-upon-binding" mechanism can be ruled out. Due to the stability of the WW domains and the flexibility of the linker, a synergistic binding behaviour might instead be possible.

In summary, the elucidation of possible tWW-complex structures is a challenging task for two main reasons: First, there are many factors that come together, which all influence the binding equilibrium of tWWs. Hence, unknown structures for tWW-complexes cannot possibly be predicted from known complex structures. Second, despite the characteristics we could identify for the tWWs, we still do not know what influences the orientation of the ligand. This is complicated by the fact that, in principle, every single WW domain can bind its target in two orientations (see 1.3). As consequence, eight different binding situations are conceivable without further restrictions.

1.3.2 Binding mechanisms

Since nothing was known about the complex formation of the tWW-ligand complex studied in this thesis (see 4.1), we had to mimic the binding event by using molecular docking techniques (see 3.3.2). To understand these docking methods and, more generally, to understand protein-ligand binding, we will shortly go through the existing binding models that have been developed over a long period of time. Note however, that these binding models should be considered extreme cases that only occur under very specific circumstances.

In 1894, Emil Fischer proposed the "lock-and-key" mechanism to explain the specificity of enzymes:[51]

$$\mathrm{P} + \mathrm{L} \rightleftharpoons \mathrm{P} \cdot \mathrm{L}. \tag{1.1}$$

Without any structural changes, the ligand (L) fits into the binding site of the protein (P) like a key into a lock. The specificity of the binding results from the complementary shapes of ligand and protein. As apparent from the previous sections, by considering proteins and ligands as rigid bodies, however, we cannot explain the mediation of PPIs or allosteric binding. The "lock-and-key" mechanism only holds for systems with almost no flexibility such as the tWW of PRP40 introduced before.[46]

In 1958, D.E. Koshland suggested the "induced fit" mechanism, which includes more flexibility:[52]

$$\mathrm{P}_1 + \mathrm{L} \rightleftharpoons \mathrm{P}_1 \cdot \mathrm{L} \tag{1.2}$$

$$\mathrm{P}_1 \cdot \mathrm{L} \rightleftharpoons \mathrm{P}_2 \cdot \mathrm{L}. \tag{1.3}$$

In this model, the ligand (L) interacts with a binding-competent, yet not optimal structure of the protein (P_1). This binding event then causes a conformational change in the protein to stabilise the complex, which gives the final complex ($\mathrm{P}_2 \cdot \mathrm{L}$).[53,54]

For a long time, the interpretation of binding events was limited to the "lock-and-key" and the "induced fit" mechanisms. Then, a third mechanism, called "conformational selection", entered the field.[54] This mechanism had its origin in the folding theory of proteins and the associated energy landscape.[55,56] In contrast to the "induced-fit" and the "lock-and-key" mechanisms, the "conformational selection" mechanism does not assume the protein to be in a binding-competent structure most of the time. Instead, it is believed that the protein exists in an ensemble of conformations and the ligand can "select" a certain conformation whenever formed. The binding of the ligand then causes a shift in the relative populations of the protein conformations towards the bound-one. Hence, by this model, we can also explain binding events between non-native protein conformations and possible ligands.[54,55] Using the nomenclature from eq. 1.2, the "conformational selection" could be formulated as follows:[53,54]

$$\langle \mathrm{P}_1, \ldots, \mathrm{P}_{\mathrm{n}\neq 2} \rangle \rightleftharpoons \mathrm{P}_2 \tag{1.4}$$

$$\mathrm{P}_2 + \mathrm{L} \rightleftharpoons \mathrm{P}_2 \cdot \mathrm{L}, \tag{1.5}$$

where the first equation denotes the formation of the conformation P_2 out of the ensemble of conformations $\langle \ldots \rangle$.

As apparent from the eq. 1.2 and eq. 1.4, the "induced fit" and "conformational selection" mechanisms are competing binding models. They can be put in a kinetic and thermodynamic context, which will not be further elucidated here. For excellent reviews on the ongoing controversy between the two mechanisms, please refer to Refs. [54, 56]. In any case, the

close rivaly between the two binding mechanisms makes it very difficult to chose, and more recent reports even suggest that the truth might lie in combining them.[57]

1.4 Amatoxins

For more than 70 years, the scientific community is intrigued by the fascinating chemistry and properties of the amatoxins.[58,59] The name "amatoxins" stands for a family of small bicyclic octapeptides many of which are lethal in very small doses. Their most prominent representative is α-amanitin (see Figure 1.2), which is found in several species of the *Amanita* fungus family such as the "death cap" mushroom (*Amanita phalloides*).[60–62] The deadly character of the amatoxins results from their ability to inhibit the RNA polymerase II in the nucleotide addition cycle.[59,63–65] Driven by the idea to use α-amanitin or analogues as antibody-drug conjugates,[66,67] two major objectives have been pursued over the years: First, the exploration of the total synthesis of α-amanitin and analogues to allow for syntheses on larger scale and with access to more derivatives.[68–71] Second, the investigation of the structure-activity relationship (SAR) of the involved residues in order to understand their contribution to the inhibition of RNA polymerase II.[72–76]

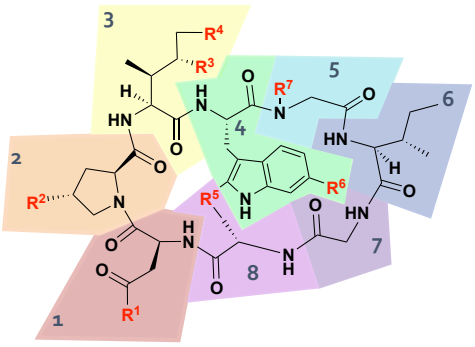


Figure 1.2: Structural formula of amatoxins coloured according to the residue index. Functional groups are highlighted in red. Their definitions are given in Table 1.1.

Table 1.1: Functional groups for the amatoxin scaffold (see Figure 1.2). Only selected amatoxins are shown. Please note, "Gly5Sar" denotes the structure that has been studied in subproject **B3**. According to Ref. [77], the correct name would be "*S*-deoxo[Ile3]aminamide".

Name	R^1	R^2	R^3	R^4	R^5	R^6	R^7
α-amanitin	NH_2	OH	OH	OH	SO	OH	H
β-amanitin	OH	OH	OH	OH	SO	OH	H
amanullin	NH_2	OH	H	H	SO	OH	H
Gly5Sar	NH_2	OH	OH	OH	S	OH	H
aminamide	NH_2	OH	OH	OH	SO	H	H

1.4.1 Structural aspects

Amatoxins comprise an astonishing degree of complexity despite their small sizes. Eight amino acids are linked to one cyclic peptide chain ("macrolactam") with an additional linkage over the side chains of tryptophan and cysteine ("tryptathionine bridge", Figure 1.2). The resulting compact structure makes amatoxins very resistant against temperature changes, enzymes or acid degradation.[59,78] In addition, amatoxins are greatly soluble in water, which

facilitates the oral uptake. The robustness and solubility alone, however, only make it troublesome to get rid of amatoxins after ingestion. So, what exactly makes amatoxins, especially α-amanitin so life-threatening?

1.4.2 Poisoning

α-Amanitin poisoning has been studied in great detail.[59,78,79] It is a slowly acting toxin, which means that shortly after ingestion (1.5 - 2 h), α-amanitin can be detected in the urine, but the affected body does not show any symptoms.[78,80,81] Then, for the first of three stages of the poisoning process, symptoms such as dehydration and abdominal pain are reported. It follows a pseudo-recovery stage, which illustrates the treachery of the amatoxin poisoning: although the symptoms subside,[59] the liver is seriously damaged so that after 24 - 48 h, clinical signs such as coagulopathy appear.[79] Finally, in the last stage, the affected body starts suffering from hepato-renal syndrome, which results in coma and death (within 4 - 7 d).[78,79]

Two questions arise from this description of the poisoning process: Why do α-amanitin and analogues show a preference for hepatocytes? What happens on a molecular level?
Hepatocytes express two transporters, the organic anion-transporting polypeptide (OATP) and Na^+/taurocholate cotransporter polypeptide (NTCP), which both have been identified as carriers for α-amanitin.[78,82,83] Once α-amanitin is inside the cell, a cascade of effects begins. In brief, α-amanitin inhibits the RNA polymerase II, or more precisely: the RNA polymerase elongation complex (EC), by trapping it in an intermediate state during the translocation process.[63,84] In this trapped state, α-amanitin impedes both the entry of new nucleotides to the active site of the RNA polymerase II and the translocation of DNA and RNA. As consequence, the protein synthesis of e.g. coagulation factors is slowed down significantly.[78,85] The cell death caused by α-amanitin occurs via apoptosis triggered either indirectly by the stress of the RNA polymerase inhibition or directly by the interaction of α-amanitin with the mitochondrial membrane.[78,86]

Please note, in literature the terms "(cyto)toxicity" and "inhibition capacity" have been used interchangeably in the context of α-amanitin, but there is a fine difference between them: cytotoxicity is a measure of how much the substrate harms the cell, and inhibition capacity is a measure of how strong an inhibitor binds to its target. While the strong inhibition of RNA polymerase II by α-amanitin has a strong cytotoxic effect, it was reported that without access to the RNA polymerase II, the cytotoxicity of α-amanitin is much weaker.[87]

1.4.3 Structure-activity relationship

Due to the excellent work of Florian Brueckner, Patrick Cramer, David A. Bushnell and Roger D. Kornberg, we know that the reason for the strong inhibition is based on the interactions between the subunits of the RNA polymerase and the residues in the α-amanitin.[63,84,88]

Their structural insight was an important contribution to the SARs that scientists have been investigating for many years. Nearly whenever a new α-amanitin derivative was synthesised, its inhibition capacity was tested to draw conclusions on the impact of varying side chain lengths,[72,74] (de)functionalisations[72] or mutations.[71] However, these studies do not always relate to the RNA polymerase II of the same species. Here, we will therefore discuss some consensus on the SAR in α-amanitin and analogues. Based on the co-crystal structures of RNA polymerase II with α-amanitin,[63,89] it is clear that each residue in the amatoxin scaffold interacts with the target via hydrogen bonds or hydrophobic interactions. Thereby, the amatoxin scaffold is the incontrovertible requirement for inhibition: deviations from the bicyclic structure with the tryptathionine bridge cause a complete loss of inhibition capacity.[71,90,91] We will now go through the SARs of the different residual positions. The amino acid present in α-amanitin is indicated in brackets, respectively:

- Residue 1 (Asn1) should either be asparagine or aspartic acid in order to maintain the inhibition capacity.[92]
- Residue 2 (Hyp2) comprises one of the pivotal features for effective inhibition: the hydroxyl group at the C_γ-atom, which is directly involved in hydrogen bonding to RNA polymerase II.[72,74,84]
- For residue 3 (DHIle3), it was reported that the stereochemistry of the C_γ-atom (α-amanitin: *(S)*) is very important for the inhibition,[74,90] and the hydroxyl-group attached to it influences the toxicity of the amatoxin.[73,90]
- Due to the relevance of the tryptathionine bridge, there is not much variation reported for residue 4 (Trp4). Solely the 6'-hydroxyl-group in the indole side chain can be missing without any impact on the inhibition capacity.[93]
- The fifth residue (Gly5) is suggested to be a direct supporter for the positioning of residue 6 on the RNA polymerase II. If Gly5 is replaced by bulkier residues, the inhibition capacity is massively reduced.[71,72]
- For residue 6 (Ile6), a lipophilic side chain is favourable, as mutations by alanine nearly abolish the inhibition capacity.[74,94]
- Residue 7 (Gly7) appears to be more tolerant to structural modifications than residue 5.[71,72] By altering the side chains of residue 6 and residue 7, the group of David M. Perrin has recently synthesised an α-amanitin derivative, which they claim to be even more toxic than α-amanitin.[95]

- In the SAR of residue 8 (Cys8), there is some disagreement, which relates to the stereochemistry of the sulphur atom if it is oxidised.[71,72,74,95] However, if the sulphur is part of a thioether, the inhibition capacity remains unchanged.[59]

1.4.4 Synthesis

Taken together, the elucidation of the SAR is a complex and non-trivial task, which was hampered for a long time due to the lack of a library of amatoxins. A major issue in building such a library was the only partial understanding of the total synthesis of the amatoxins. In nature, amatoxins are ribosomally synthesised,[61,62] and in principle they could be extracted from their natural source. However, for modifying the amatoxin scaffold, an in-depth understanding of the intermediate steps is necessary.

In 2018, the group of David M. Perrin published the first total synthetic strategy for α-amanitin. Then, two further synthetic protocols followed: one from our colleagues, the group of Roderich D. Süssmuth from Technische Universität Berlin (2019), and one from Heidelberg Pharma Research GmbH (2020). In 2021, our colleagues published another synthetic route together with us (publication **B1**, see 4.2).

In all these strategies, there are two pivotal steps: the macrolactamisation and the formation of the tryptathionine bridge. The different approaches show that it makes a difference in which order these two major steps are performed.
In the synthetic route from Perrin *et al.*, a heptapeptide is prepared, which already includes tryptophan with an oxidised side chain. This intermediate allows for a Savige-Fontana reaction to form the tryptathionine bridge.[91,96] In the end, the macrolactam is formed by inserting (3*R*,4*R*)-4,5-dihydroxyisoleucine (DHIle) via a coupling reaction.[68]

The scientists from Heidelberg Pharma Research GmbH followed a very similar approach. Using solid-phase peptide synthesis, they create a linear octapeptide in which the side chain of tryptophan is also oxidised. Then, they establish the tryptophan-cysteine linkage also via Savige-Fontana reaction before they close the macrolactam.[70]

The 2019 synthetic route from Süssmuth *et al.* is quite different: Here, they pursue a "[5+1+2]"-strategy, in which the tryptophan-cysteine linkage is formed prior to the formation of any peptide chain. Once formed, the cyclic subunit Trp4-Gly5-Ile6-Gly7-Cys8 is generated yielding the 5-membered fragment. It follows the addition of the DHIle3 (1-membered fragment) to the N-terminus and afterwards the addition of the 2-membered fragment (Asn1-Hyp2) to the C-terminus, each via coupling reaction. In the last step the macrolactam is closed between DHIle3 and Hyp2.[69]

The most recent synthetic strategy by Süssmuth *et al.* is presented in subproject **B1**. In short, linear octapeptides are formed via solid-phase peptide synthesis. This time, how-

ever, the tryptophan side chain is not oxidised. Instead of a Savige-Fontana reaction, an iodine-mediated cyclisation forms the tryptathionine bridge. Afterwards the macrolactamisation takes place between the C- and N-termini. In comparison to the synthetic strategies discussed before, this alternative route has two major advantages: (1) strong oxidants are not necessary, since the side chain of tryptophan does not need to be oxidised. (2) The iodine-mediated tryptathionine bridge formation can be applied to various linear precursors allowing to synthesise a variety of α-amanitin derivatives.[71]

These four total syntheses were a big achievement for the research community dealing with amatoxins. After more than 70 years of research, it is now finally possible to build up an amatoxin library to further explore the amatoxin family. In subproject **B2**, for instance, we identified a novel type of isomerism, the discovery of which can be attributed to the unravelled total synthesis of amatoxins. Also, the investigation of the conformations of the Gly5Sar-amanullin (see Figure 1.2) is an exciting benefit from being able to synthesise a variety of α-amanitin derivatives (see publications **B2** and **B3**)

1.4.5 Other fungal toxins related to the amatoxins

Dealing with amatoxins, it should be noted that there are other (related) peptide families, which share their unique feature, the tryptathionine bridge. The closest relatives to the amatoxins are the "phallotoxins", which can also be extracted from the death cap mushroom.[58,92] In fact, phalloidin, the most-prominent member of that peptide family, was discovered prior to α-amanitin.[97] Unlike amatoxins, phallotoxins only comprise seven amino acids, and they are not orally-toxic. If inserted to the blood-stream, however, phallotoxins also target the liver. The toxicity of the phallotoxins, though, comes from their ability to bind to filamentous F-actin in hepatocytes. Bound to the phallotoxin, F-actin is prevented from depolymerisation, which, however, is necessary for the cell to operate.[91,98]

Within the context of phallotoxins, the "virotoxins" are noteworthy, since they also comprise seven amino acids and show a similar toxic behaviour like the phallotoxins. However, virotoxins are monocyclic peptides that lack the tryptophan-cysteine linkage.[99]

Finally, there are "cortinarins", cyclic decapeptides that were isolated in 1984 from a poisonous species of the *Cortinarius* fungus genus.[100,101] Three cortinarins were extracted (cortinarin A, B, C), of which two exhibit the tryptathionine bridge (A, B) and one does not (C). Interestingly, the cortinarins with the tryptathionine bridge showed renal toxicity, while cortinarin C appears to be non-toxic.[100,101] The mechanism of the toxicity, however, remains to be discovered in future studies.

2 RESEARCH QUESTION

In this thesis, we will tackle two challenging aspects of PPI-related research with classical MD simulations:

1. We will take a closer look at the domain-based interaction interface of a particular PPI within the spliceosomal context.

2. We will study the design and synthesis of a specific bioactive peptide, which is known as strong inhibitor in the context of the PPIs in the transcription process.

MD simulations are a tool, which - given the continuous improvement and increasing availability of computational resources - has earned a well-recognised role in the PPI research.[102,103] As described in the introduction (see 1.1), PPIs represent a vast and versatile research field, which often lacks the atomistic view on the underlying processes. It is true though, that there are experimental methods to gain access to the structure and dynamics of the molecules involved. Examples are (protein) X-ray crystallography and/or NMR experiments. However, these methods have some major drawbacks in comparison to MD simulations:

Despite its sometimes cumbersome implementation, (protein) X-ray crystallography can provide static structures in a specific environment. This can be very helpful for structurally restricted peptides as we will see in one of the projects (see 4.4). Yet, many biological functions are defined by motion of the interacting molecules. This is not reflected in the crystal structure.[104]

NMR experiments on the other hand can offer dynamical insights into the systems, but they are unsuited to monitor structural changes below their spatial and temporal resolution. NMR experiments rather provide ensemble averages over the measurement time.[104,105] Moreover, with increasing size of the system of interest, the resulting NMR spectra become more difficult and complex. This can make the assignment of atom-specific contributions to the calculated averages almost impossible. Note, these drawbacks do not devalue the important insights we can gain by (protein) X-ray crystallography or NMR experiments on the investigated system. On the contrary, their data can often be used as reference or guidance for MD simulations as shown in all projects.

By performing classical MD simulations, we can calculate the physical motion of the molecule(s) using a much higher time resolution than NMR experiments. This allows us (1) to monitor (atom-specific) individual contributions to the overall dynamics of the system and (2) to resolve processes faster than the temporal resolution of spectroscopical methods.[19] Therefore, I will denote MD simulations as “computational microscope”, a term that has been

introduced and shaped by Klaus Schulten from Havard Medical School.[104,106–108]

In this thesis, we will discuss the design and implementation of workflows that combine MD simulations with data analysis strategies. Considering experimental evidence as reference and quality measure, we will elaborate an in-depth perspective on the structural dynamics of the individual systems. This thesis is a tribute to MD simulations and their reliability to provide a structural and dynamical perspective that can either complete observations from experiments, or stand on its own.

The first project (**A**) aimed at unravelling the complex structure between an adaptor domain and proline-rich motifs (PRMs). In particular, we dealt with the flexible tWW of *human*-formin binding protein 21 (h-FBP21) and a proline-rich peptide sequence of the spliceosomal Sm protein B/B' (SmB2 ligand). From experiments by our colleagues from Freie Universität Berlin, we knew that (1) the h-FBP21 tWW forms a stable complex with the SmB2 ligand, and (2) the SmB2 ligand can adopt two orientations relative to h-FBP21 tWW. However, there remained some pivotal questions: How is the h-FBP21 tWW-SmB2 ligand formed? What are the stabilising forces keeping it together? How is the complex structurally organised?

In recent years, complex structures for other tWWs have been published (see 1.3.1). It turned out that the individuality of the WW domains in terms of their structure and binding behaviours translates to the characteristics of them in tandem repeats. This situation is further complicated by the interdomain region, which can vary substantially among the tWWs. By its length and flexibility, the interdomain region has a massive impact on the overall properties of the respective tWW. In regard of all these factors, which can determine the binding properties of the tWW, it becomes clear that unknown tWW-ligand complexes cannot be derived by known-ones. Lacking a complex structure for h-FBP21 tWW, predictions on the h-FBP21 tWW-SmB2 ligand complex are therefore maximal a lucky guess.[47]

In this project, we designed a workflow to direct our computational microscope towards the structural secrets of the h-FBP21 tWW-SmB2 ligand complex. We will see how structural insights into the dynamics of the isolated h-FBP21 tWW can help to generate useful starting points for ensuing docking experiments. We will cover how experimental evidence from our co-workers and our computational microscope can be effectively combined to evaluate the docking poses and finally make an educated guess on the h-FBP21 tWW-SmB2 ligand complex.

The second project (**B**) was dedicated to the investigation of amatoxins, a family of fungal toxins. Amatoxins are characterised by their exceptional structure: Although they only consist of eight amino acids, amatoxins are highly ordered in a bicyclic scaffold (see 1.4.1). We will see that this unusual structure has a serious influence on the synthesis of amatoxins and arising structures. Without a detailed structural analysis, the cause for the peculiar chemistry

of the amatoxins, however, remains unclear. Together with our colleagues from Technische Universität Berlin, we focused on three distinct aspects:

B1. In the synthesis of amatoxins, two of the most important steps are the creation of the cyclic peptide chain and the tryptophan-cysteine linkage (tryptathionine bridge). Our colleagues had experimental evidence that the selectivity for the tryptathionine bridge formation can differ substantially among different precursors. The reason for that, however, was not clear. We assumed that the precursors with higher selectivity might pre-organise in favour of the reaction, and to test this hypothesis, we took a closer look with our computational microscope.

B2. As our colleagues were further analysing the synthesis of amatoxins, they were able to isolate two isomers for the amatoxin scaffold whose existence and nomenclature is highly-debated in literature.[74,75,109] These two isomers differ in the arrangement of the tryptathionine bridge relative to cyclic peptide chain. It was unclear how the two isomers behave in solution, whether they can convert into each other and which structural conditions could favour their formation. In this subproject, we designed a model to unambiguously describe the position of the tryptathionine bridge relative to the cyclic peptide chain. We will see how this model can be used as "lens" for our computational microscope to monitor the dynamics of the tryptathionine bridge in both isomers. In addition, we will discuss the structure-determining interactions unveiled by the MD simulations.

B3. To further explore the SARs of the amatoxins, our colleagues also focused on N-methylation. N-methylation is a well-known strategy to enforce structural changes in monocyclic peptides by suppressing hydrogen bonding to a targeted amide.[110,111] Our colleagues synthesised Gly5Sar-amanullin and gained spectroscopical evidence for a conformational exchange between two long-living conformations. They found evidence that the synthesised Gly5Sar-amanullin is in the M_{ansa}-form, but the close relationship between the two conformations made it impossible to extract clear evidence for both of them. In this subproject, we will learn about the difficulties we have to encounter with our computational microscope if there is only precarious experimental evidence to validate the results. However, we will demonstrate that the level of detail provided by our microscope can unravel structural conditions, which would have stayed hidden from the sights of the experimental methods. Although our colleagues only synthesised the M-ansamer of Gly5Sar-amanullin, we simulated both ansamers extensively. Combining MD simulations with spectroscopical evidence, we thereby provide a structural-dynamical perspective on the influence of N-methylation for the amatoxin scaffold.

Throughout the described projects, we would like to answer the following question: Can we claim the MD simulations to be the computational microscope needed as urgently as the experiments to do profound research on the introduced aspects of PPI research?

3 THEORY

This chapter is divided into three parts: At first, we will discuss MD simulations covering the following parts: (1) the fundamentals of MD simulations (see 3.1.1), (2) Setup and (post)processing of MD simulations (see 3.1.2), (3) Analysis of MD simulation, in particular ensemble averages, kinetic networks (see 3.1.3) and (4) Issues and limitations of MD simulations (see 3.1.4). This is followed by a brief introduction to NMR spectroscopy and its usability for MD simulations (see 3.2) is given. In the end, we will discuss further computational tools such as clustering (see 3.3.1) and docking (see 3.3.2). Please note, we will only discuss *classical* MD simulations in this thesis if not explicitly stated otherwise.

3.1 Molecular dynamics simulations

3.1.1 The fundamentals

Newton's law of motion. In classical molecular dynamics (MD) simulations, the dynamics of a molecular system are described by classical mechanics.[112] In Cartesian space, this means that the system's dynamics are approximated by solving the differential equations from Newton's second law of motion ($F = m \cdot a$):[113,114]

$$\frac{\mathrm{d}^2\mathbf{r}_i(t)}{\mathrm{d}t^2} = \frac{\mathbf{F}_i(t)}{\mathrm{m}_i}, \tag{3.1}$$

with $\mathbf{r}_i = (x_i, y_i, z_i)^{\mathrm{T}}$ as the position of particle i with mass m_i. $\mathbf{F}_i = (f_{x,i}, f_{y,i}, f_{z,i})^{\mathrm{T}}$ denotes the forces that act on the particle i. Solving eq. 3.1 for all N particles at different times t yields a *trajectory*, a time series of states $\mathbf{x}(t) = \{\mathbf{r}_1(t), ..., \mathbf{r}_N(t)\} \in \Omega \subset \mathbb{R}^{3N}$ with Ω denoting the state space of the system. On a molecular level, $\mathbf{x}(t)$ can be understood as the configuration of the system at time t. Please note, throughout this chapter, the configuration $\mathbf{x}(t)$ will be described in *position space*, only. For a full description of the system, also the momenta $\mathbf{p}_i$ of the individual particles have to be taken into account.

At this point, some very important assumptions are already made:

(a) Electrons and their movement are neglected as the motion of the nuclei is assumed to be determining for the dynamics of the system (Born-Oppenheimer-Approximation).[113–115]

(b) The particles, either single atoms or rigid bodies, are treated as point charges with position $\mathbf{r}_i$, mass m_i and electric charge q_i.[112]

(c) The total force $\mathbf{F}_i$ acting on a particle i is defined by the negative gradient of the potential energy function $V(\mathbf{r}_1, ..., \mathbf{r}_N)$:[114]

$$\mathbf{F}_i(t) = -\frac{\partial V(\mathbf{r}_1(t), ..., \mathbf{r}_N(t))}{\partial \mathbf{r}_i(t)} = -\frac{\partial V(\mathbf{x}(t))}{\partial \mathbf{r}_i(t)}, \tag{3.2}$$

where $V(\mathbf{x}(t))$ represents the (potential) energy surface of the system on which the atoms move, and $\mathbf{x}(t)$ denotes the trajectory defined earlier.[116] Assumption (c) is justified by the Ehrenfest theorem, which states that the classical mechanics coincide with quantum mechanical expectations in macroscopical processes.[117]

The force field. Within the context of classical MD simulations, $V(\mathbf{x})$ comprises several energy terms parametrized against quantum mechanical calculations and experimental data:[113,116,118]

$$\begin{aligned} V\left(\mathbf{x}(t)\right) = & \sum_{i}^{N_{\text{bonds}}} \frac{k_{l,i}}{2}\left(l_i(t) - l_{i,0}\right)^2 + \sum_{i}^{N_{\text{angles}}} \frac{k_{\theta,i}}{2}\left(\theta_i(t) - \theta_{i,0}\right)^2 \\ & + \sum_{i}^{N_{\text{torsions}}} \frac{V_n}{2}\left[1 + \cos\left(n\omega(t) - \gamma\right)\right] \\ & + \sum_{i}^{N_{\text{particles}}} \sum_{j>i}^{N_{\text{particles}}} \left(4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}(t)}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}(t)}\right)^{6}\right] + \frac{q_i q_j}{4\pi\epsilon_0\epsilon_1 r_{ij}(t)}\right) \end{aligned} \tag{3.3}$$

$V(\mathbf{x}(t))$ and all the parameters used for its definition are together called 'force field'. Eq. 3.3 shows the collection of energy contributions according to the AMBER force field.[119] Note, there are several other definitions of force fields.[116,118]

Bonded interactions. The first three terms in eq. 3.3 represent the *bonded* interactions of the system, of which the first term incorporates a harmonic potential for the deviation of a covalent bond i with length l_i from its reference value $l_{i,0}$ at time t. $k_{l,i}$ denotes the force constant according to Hooke's law.[113]

The second term includes the deviation of an angle θ_i between three covalently linked atoms A-B-C from the reference angle $\theta_{i,0}$ at time t. The second term is also defined as a harmonic potential with force constant $k_{\theta,i}$.

The third term considers torsion angles ω between four covalently linked atoms A-B-C-D. Eq. 3.3 includes the energy contribution of the *proper* dihedral angles, which describe the position of atom A relative to atom D in terms of rotations around the bond B-C (see Figure 3.1) As the dihedral angles are intrinsic periodic, their energy contribution is described by a cosine function with amplitude $V_n/2$.

Although V_n is often named 'barrier height' of this potential, it is rather a qualitative measure

for the comparison of different dihedral angles that informs about how easy the system can be rotated around a certain dihedral angle. For instance, due to the double-bond character of the amide bond, the rotation around the dihedral angle $\angle(\mathrm{OCNH})$ is much more restricted than the rotation around the covalent bond linking two sp^3-hybridised carbon atoms in butane ($\angle(\mathrm{HCCH})$). Therefore, V_n^{amide} is larger than V_n^{butane}.[103,113] The parameter n is called multiplicity. It gives the number of minima in the dihedral potential whose positions can be influenced by the phase factor γ.

Besides the proper dihedral angles, many force fields also include an energy term for *improper* dihedrals or 'out-of-plane' motions. If included, these angles are most often defined by harmonic potentials with force constant $k_{\zeta,i}$:

$$v^{\mathrm{imp.}}(\zeta(t)) = \sum_i^{N_{\mathrm{imp.}}} \frac{k_{\zeta,i}}{2} \left(\zeta_i(t) - \zeta_{i,0}\right)^2 \tag{3.4}$$

describing the deviation of angle ζ_i at time t from the reference value $\zeta_{i,0}$. Some force fields also include cross terms combining two or more motions, but these are not discussed here.

Non-bonded interactions. The last term in eq. 3.3 describes the *non-bonded* interactions of the system for all pairs of particles that are not covered in the bonded interactions. The non-bonded interactions include van-der-Waals interactions modelled by the Lennard-Jones 12-6 potential (eq. 3.5), and electrostatic interactions defined in the Coulomb potential (eq. 3.6):

$$v^{\mathrm{LJ}}(r_{ij}(t)) = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}(t)} \right)^{12} - \left(\frac{\sigma}{r_{ij}(t)} \right)^{6} \right], \tag{3.5}$$

$$v^{\mathrm{C}}(r_{ij}(t)) = \frac{1}{4\pi\epsilon_0\epsilon_1} \frac{q_i q_j}{r_{ij}(t)}, \tag{3.6}$$

with $r_{ij}(t) = |\mathbf{r}_j(t) - \mathbf{r}_i(t)|$ as the spatial distance between two point particles i and j as common feature. The Lennard-Jones 12-6 potential (eq. 3.5) consists of a repulsive part ($\propto r^{-12}$) that can be understood in terms of the Pauli exclusion principle, and an attractive part ($\propto r^{-6}$) accounting for dispersive interactions (see Figure 3.1).

Due to these decays, the Lennard-Jones potential is a short-range potential. ϵ denotes the depth of the minimum in the Lennard-Jones 12-6 potential with respect to $v^{\mathrm{LJ}} = 0$. σ, the interatomic distance at which $v^{\mathrm{LJ}} = 0$ can be used to determine the equilibrium distance $r_{ij,0} = 2^{1/6}\sigma$ between the two particles. The Coulomb potential varies as r^{-1}, and it is therefore a long-range potential. ϵ_0 is the vacuum permittivity, ϵ_1 is the relative permittivity and q_i denotes the charge of particle i.

As apparent from eq. 3.5 and eq. 3.6, all non-bonded interactions are treated as pair interactions. Although not all non-bonded interactions are well-described by pair interactions,

many multi-body effects, e.g. polarisation effects, can be expressed in pairwise models with appropriate parametrisation. These approximations for the multi-body effects called "effective" pair potentials, are then incorporated in the force fields.

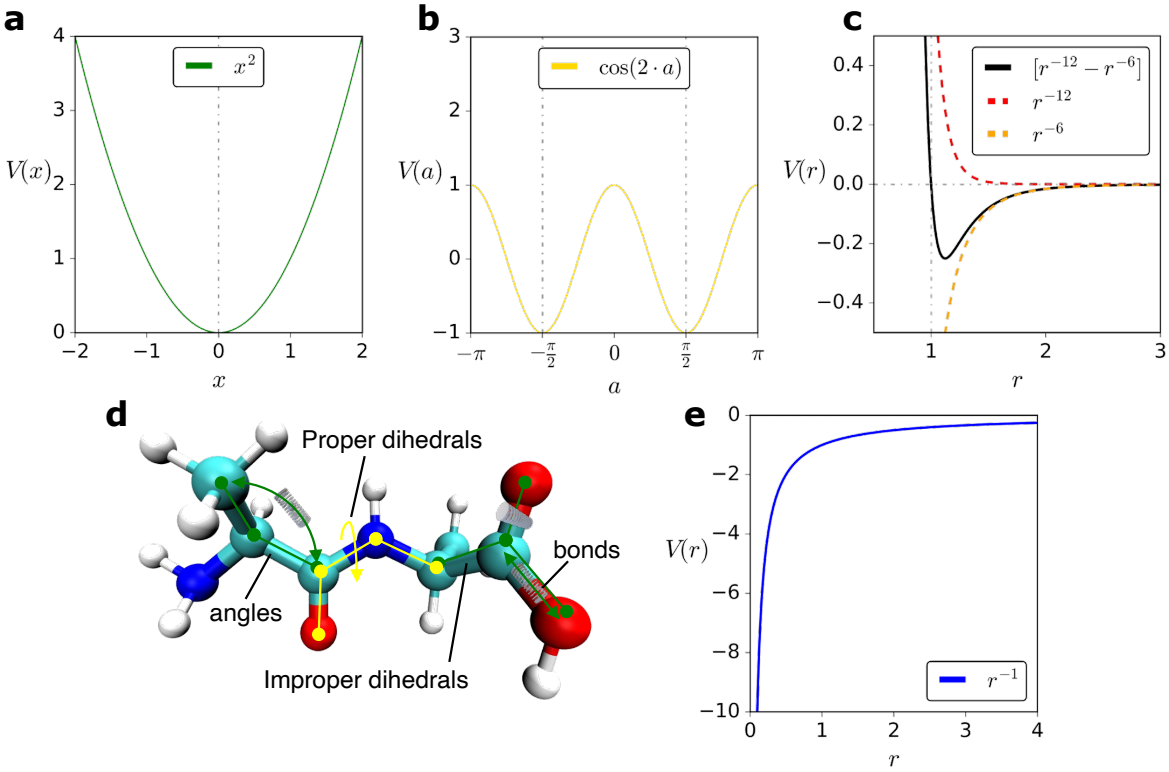


Figure 3.1: Basic functional forms used as contributing terms in the force field eq. 3.3. (a) harmonic potential, which is considered for bonds, angles and improper dihedral angles (d, green). (b) cosine based angle potential used for dihedral angles (d, yellow). (c) Lennard-Jones potential ($r^{-12} - r^{-6}$, black line), repulsion term (r^{-12}, dotted red line) and attractive term (r^{-6}, dotted orange line). The repulsive term describes the Pauli repulsion at short distances, the attractive term accounts for attraction due to London dispersion forces.

Integrators. Although the usage of force fields facilitates the handling of the equations of motion (eq. 3.1), $V(\mathbf{x}(t))$ is still a continuous function, which cannot be solved analytically for many particles.[103,113]

In classical MD simulations, the equations of motion are solved numerically by using integration algorithms that rely on a finite difference method. This means that the integration of the equations of motion is split up into many time steps of constant size Δt. In doing so, the integration schemes make two assumptions:

(1) The forces $\mathbf{F}_i = m_i \mathbf{a}_i = \partial V(\mathbf{x}(t))/\partial \mathbf{r}_i$ are constant during a time step Δt.

(2) The positions and the velocities of the particles of the system at time $t + \Delta t$ can be approximated by Taylor series expansions:

$$\mathbf{r}_i(t+\Delta t) = \sum_{n=0}^{\infty} \frac{\mathbf{r}_i^{(n)}(t)}{n!} (\underbrace{t+\Delta t-t}_{=\Delta t})^n \approx \mathbf{r}_i(t) + \Delta t \mathbf{v}_i(t) + \frac{(\Delta t)^2}{2}\mathbf{a}_i(t) + \mathcal{O}(\Delta t^4) \tag{3.7}$$

$$\mathbf{r}_i(t-\Delta t) = \sum_{n=0}^{\infty} \frac{\mathbf{r}_i^{(n)}(t)}{n!} (\underbrace{t-\Delta t-t}_{=-\Delta t})^n \approx \mathbf{r}_i(t) - \Delta t \mathbf{v}_i(t) + \frac{(\Delta t)^2}{2}\mathbf{a}_i(t) + \mathcal{O}(\Delta t^4) \tag{3.8}$$

Note, that eq. 3.7 and eq. 3.8 already include the definitions of the derivatives of the position $\mathbf{r}_i$ at time t being the velocity $\mathbf{r}_i'(t) = \mathbf{v}_i(t)$ and acceleration $\mathbf{r}_i''(t) = \mathbf{a}_i(t)$. For both steps $t \pm \Delta t$, the Taylor series is truncated after the third term leaving out terms in the order of $\mathcal{O}(\Delta t^4)$. Likewise, the velocities can also be approximated by Taylor series, however they are usually truncated after the second term ($\mathbf{v}_i' = \mathbf{a}_i$) leaving out terms in the order of $\mathcal{O}(\Delta t^3)$.

The different integration algorithms available today, differ in their choice of what is calculated at which time step. By adding eq. 3.7 and eq. 3.8, one obtains:

$$\mathbf{r}_i(t+\Delta t) = 2\mathbf{r}_i(t) - \mathbf{r}_i(t-\Delta t) + (\Delta t)^2 \underbrace{\frac{\mathbf{F}_i(t)}{m_i}}_{=\mathbf{a}_i(t)} + \mathcal{O}(\Delta t^4). \tag{3.9}$$

In other words, the next position of particle i at time $t+\Delta t$ can be obtained by the position and the force at the current time step t and the position from the previous step $t-\Delta t$. The algorithm using this integration scheme is called *Verlet* algorithm.[120]

This algorithm has two drawbacks:
(1) A term of order $\mathcal{O}(\Delta t^4)$ is added to terms of order $\mathcal{O}(\Delta t^2)$, which can cause numerical problems.
(2) Velocities are not directly computed by the integration algorithm, which is problematic as they are needed to calculate specific properties such as the system's kinetic energy.

The *leap-frog* algorithm[121] used within this thesis, improves on both drawbacks. Here, the position is updated at a half-time step ($\Delta t/2$). Subtracting the Taylor series of $\mathbf{r}_i(t-\frac{\Delta t}{2})$ from $\mathbf{r}_i(t+\frac{\Delta t}{2})$, yields (after a shift $t \to t+\frac{\Delta t}{2}$):

$$\mathbf{r}_i(t+\Delta t) = \mathbf{r}_i(t+\Delta t) + \Delta t \mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) + \mathcal{O}(\Delta t^3) \tag{3.10}$$

with

$$\mathbf{v}_i\left(t+\frac{\Delta t}{2}\right) = \mathbf{v}_i\left(t-\frac{\Delta t}{2}\right) + \Delta t \mathbf{a}(t) + \mathcal{O}(\Delta t^3). \tag{3.11}$$

The name of the algorithm refers to the successive updating interval ('leaping') of the positions and forces at time t and the velocities at time $t+\Delta t$.

Although the precision-problem of the Verlet algorithm is resolved, the leap-frog algorithm

has the disadvantage that the velocities and the positions are not synchronised. As consequence, the kinetic energy:

$$E_{\mathrm{kin},i}(t) = \frac{1}{2} m_i \mathbf{v}_i(t)^2 \tag{3.12}$$

can only be calculated with an average at time $t \pm \frac{\Delta t}{2}$:

$$\mathbf{v}_i(t) = \frac{1}{2}\left(\mathbf{v}_i\left(t - \frac{\Delta t}{2}\right) + \mathbf{v}_i\left(t + \frac{\Delta t}{2}\right)\right), \tag{3.13}$$

which means that the contribution of the kinetic energy to the total energy (eq. 3.14) cannot be determined at the same time as the potential energy which is position-dependent:

$$E_{\mathrm{tot}} = E_{\mathrm{kin}}(\mathbf{v}) + E_{\mathrm{pot}}(\mathbf{r}) \tag{3.14}$$

The time step. Both described integration schemes are deterministic and time-reversible, which means that if one integrates n steps forward and afterwards n steps backward, one will end up at the starting position. The question remains, however, which time step to choose for the integration algorithm? In this regard, the integrators do the splits between (a) ensuring the correct physical properties such as energy and total (angular) momentum and (b) staying as close as possible to the analytical trajectory of the system, i.e. the 'real solution' according to eq. 3.1.[103,113] Criterion (a) demands that the MD simulations have to estimate the correct probability distribution of the properties of the system's ensemble in order to give reliable averages for a given observable a. According to the 'ergodic hypothesis', this can be achieved by either considering many systems at the same time t or long simulation times for a single system:

$$\langle a \rangle = \lim_{\zeta \to \infty} \frac{1}{\zeta} \sum_{i=1}^{\zeta} a_i = \lim_{\tau \to \infty} \frac{1}{\tau} \sum_{t=0}^{\tau} a(\mathbf{x}(t)) \tag{3.15}$$

with a as the targeted observable, ζ as the number of systems and τ as the number of points in time. If possible, criterion (a) would therefore favour large time steps.

Criterion (b), however, demands a small time step, as the forces, which rely on the atomic positions, are kept constant during the time step. But why is that an issue?
Assume two particles moving towards each other. At time t, their distance is large enough so that the calculated forces do not include a significant repulsive contribution, which would directly influence the movement of the atoms. If the time step Δt is chosen too large, it might happen that the two atoms at time $t + \Delta t$ are too close to each other resulting in enormous repulsive forces between them. These large forces then cause vast changes in the velocities and thereby positions/forces, which might cause the simulation to crash ('blowing up'). To minimise the error in the calculation of the forces (introduced by neglecting atom movements), it would therefore be best to recalculate the forces within a time step Δt that is smaller than the fastest positional change in the system. In biological systems, this is usually

the vibration of the C-H bond, which has an oscillation frequency of roughly 10 fs. So, 1 fs or less would be a well-suited time step.[113]

Bond constraints. To reconcile both criteria, algorithms like "LINCS"[122] were developed. LINCS introduces distance constraints for specific distances such as the C-H bond and thereby allows for larger time steps. Basically, these constraints can be understood as harmonic potentials that can be added to the potential energy function $V(\mathbf{x}(t))$ (eq. 3.3):

$$v^{\mathrm{bc}} = \sum_{k=1}^{N_{\mathrm{bc}}} \lambda_k \left((\mathbf{r}_i(t) - \mathrm{r}_j(t))^2 - d_{ij}^2 \right) \tag{3.16}$$

with N_{bc} as the number of considered constraints for the atom pairs (i, j) and d_{ij} as the reference distance. A single bond constraint is zero, if the respective bond between atom i and j equals the (user-defined) reference distance d_{ij}. Please note, λ_k is a Langrange multiplier that is necessary to solve all bond constraints simultaneously, as the bonds are not independent of each other.[113] Using LINCS algorithm on bonds including hydrogen atoms, 2 fs can be used as time step.

Ensemble types. As mentioned in criterion (a), there are different ensemble types with different properties: Without further additions, Newtonian dynamics that are calculated by the integrators presented before, would yield the *microcanonical ensemble*, a closed system with constant number of particles, constant volume and constant energy (NVE ensemble). For most applications, however, other ensembles are better suited as they better mimic the experimental conditions: The *canonical* ensemble includes a constant temperature allowing for energy fluctuations, but keeping the volume and the amount of particles constant (NVT ensemble). The constant temperature is accomplished by the usage of thermostats, i.e. algorithms that modify the Newtonian dynamics so that the temperature T^{MD} defined by the time average of the kinetic energy $\langle E_{\mathrm{kin}} \rangle$ matches a (user-defined) reference value T^{ref}:

$$T^{\mathrm{MD}} = \frac{2}{N_f k_{\mathrm{B}}} \langle E_{\mathrm{kin}} \rangle = \frac{2}{N_f k_{\mathrm{B}}} \left\langle \sum_i^{N_f} \frac{1}{2} m_i v_i^2 \right\rangle, \tag{3.17}$$

with k_{B} as the Boltzmann constant and N_f as the total number of degrees of freedom, which is defined by the number of particles in the system N and the number of constraints $N_{\mathrm{constraints}}$:

$$N_f = 3N - N_{\mathrm{constraints}}. \tag{3.18}$$

$N_{\mathrm{constraints}}$ could be, for instance, N_{bc} defined in eq. 3.16. Please note, in eq. 3.17, we are making use of the *equipartition theorem*, which states that, on average, the kinetic energy is equally distributed over all degrees of freedom, yielding an energy contribution of $1/2 k_{\mathrm{B}} T$ per degree of freedom.[112,113,123] The equipartition theorem also allows us to calculate the

temperature for a single time step of the MD simulation, called *instantaneous temperature*:[113]

$$T^{\mathrm{MD}}(t) = \frac{2}{k_{\mathrm{B}}} E_{\mathrm{kin}}(t). \tag{3.19}$$

Thermostats now add factors to eq. 3.19 in order to influence the velocities incorporated in the kinetic energy. The most straight forward approach would be to correct the velocities by a factor so that the instantaneous temperature exactly matches the targeted reference temperature T^{ref}. This approach is called 'velocity-rescaling', and without further modifications, simple velocity-rescaling does not correctly reproduce the canonical ensemble.[112,113] To improve on this, G. Bussi, D. Donadio and M. Parrinello developed a modified velocity-rescaling algorithm,[123] which rescales the velocities to a kinetic energy that is drawn from the canonical kinetic energy distribution at equilibrium. As the thermostat by G. Bussi and co-workers gives the correct canonical ensemble,[112] it was used in this thesis.

Thermostats can be combined with barostats yielding the *isothermal-isobaric ensemble*, in which both temperature and pressure are kept constant, while fluctuations in energy and volume are allowed (NpT ensemble). Analogously to thermostats, barostats are algorithms targeting the instantaneous pressure of the system. Instead of scaling the velocities, however, barostats are rescaling the volume. In this work, mostly the 'Parrinello-Rahman' barostat was used,[124] as it gives the correct NpT ensemble if combined with the G. Bussi thermostat.

Instead of the amount of particles, the chemical potential can be set constant yielding the *grand-canonical* (μVT) or *generalized* ensemble (μpT). Both ensemble types require more-sophisticated algorithms.[113] Since they are not used in this thesis, they will not be discussed here.

Periodic boundaries. So far, we have left out the relevance of the system's size on performing classical MD simulations. We shortly touched the distance dependency of the Coulomb- and the Lennard-Jones 12-6 potential, but we did not discuss any implications for the computational demands. For ideal comparison to experiments, one would precisely reconstruct the experimental setting including all solvent molecules, the solute and the boundaries. This would imply, however, calculating the force field contributions (eq. 3.3) for several *mol* of particles, which is computationally not feasible.

Another option would be to simulate only a tiny fraction of the system with boundaries to vacuum, with which we yet cannot model the bulk properties correctly.[125] To mimic the bulk properties, one instead makes use of *periodic boundaries* together with cut-offs, neighbourlists and specific summation techniques.

By applying periodic boundaries to the system of interest, it will be surrounded by multiple (infinite) copies of itself in all directions, who together form a lattice (see Figure 3.2). Under the condition that the amount of particles has to remain constant in the box, particles that leave

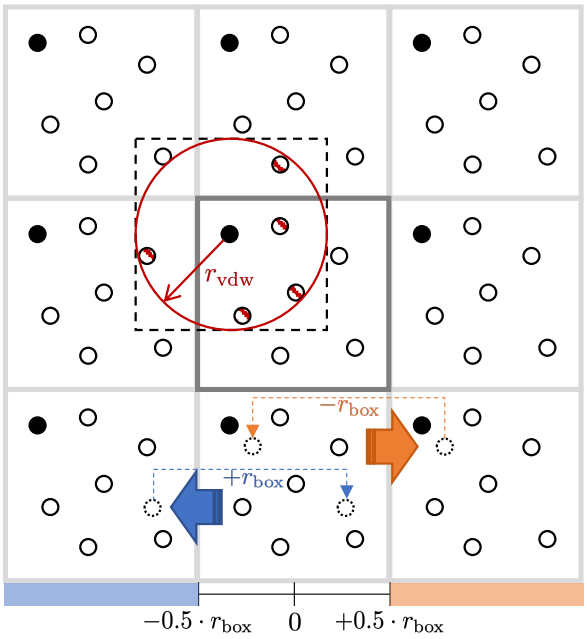


Figure 3.2: *Periodic boundaries*: The system of interest is depicted as square box (center) with box length r_{box}. It is surrounded by eight copies (shaded). For two particles, the scenario "leaving the box" is shown (orange, blue arrows). *Minimum-image convention*: The black particle is surrounded by a sphere defined by the cut-off radius $r_{\mathrm{vdw}} < L_x/2$ for the non-bonded interactions.

the box on one side are replaced by image particles entering the box from the opposing side. At any time, the coordinates of the particles can be described by multiples of the box lengths (see Figure 3.2) Two questions arise from this approach: First, how to choose a proper cell size for the original system? Second, how to treat the non-bonded interactions?

Cut-off schemes. As stated earlier, the non-bonded interactions can be differentiated into long-range and short-range interactions. Referring back to the short-range Lennard-Jones 12-6 potential, we can see that for distances $\geq 2.5\sigma$ the energy contribution is $< 1\%$ of the energy value at best distance σ. Hence, a cut-off scheme seems reasonable, in which interactions beyond a cut-off are set to zero. This cut-off, in simulation software often referred to as 'Van-der-Waals' cut-off (r_{vdw}), defines a sphere around each particle of the system, in which all short-range interactions will be calculated. Simulation software like GROMACS uses the *minimum-image convention*, which means that every atom only "sees" one image of all the other atoms of the system.[126] As consequence, the box size for the system has to be chosen such that the sphere defined r_{vdw} fits into the box.

For instance, in a cubic system, which has the same box lengths in all three dimensions (in Cartesian coordinates: $L_x = L_y = L_z$), the minimum-image convention requires that r_{vdw} is smaller than $1/2L$ (see Figure 3.2). Assigning the particles to the sphere with radius

r_{vdw} around a particle i, necessitates the calculation of all pair distances within the periodic image. Yet, with time steps in the fs-regime, we can assume that in fluids, the particles in vicinity r_{vdw} of i do not significantly change over a certain amount of time steps. Therefore, simulation software such as GROMACS calculates all distances in regular time intervals only, and creates so-called 'neighbourlists', which it uses in between.[126] These neighbourlists save the indices of all particles in the sphere with radius r_{vdw} around i as well as the indices of all atoms, which are a bit further away. The latter is done to avoid that particles entering that sphere were not included in the neighbourlist before its update.[113,126]Between the updates of the neighbourlist, the simulation software directly uses the saved indices of the particles in the vicinity to particle i to calculate the short-range interactions.

Ewald summation. While the usage of cut-offs seems justifiable for the short-range interactions, the handling of the long-range interactions is trickier. As discussed previously, the Coulomb potential decays as r^{-1}, which is problematic for two reasons:
First, by applying periodic boundary conditions, we have to account for all charges even of the periodic images.

Second, the calculation of all pairwise distances is computationally very expensive as it scales with the square of the number of particles. Fortunately, we can deal with both problems using Ewald summation.[112,127,128] This technique is based on the relation between the charge density $\rho(\mathbf{r}_i)$ and the electrostatic field $\phi(\mathbf{r}_i)$, formulated by the Poisson equation:[112]

$$\nabla^2\phi(\mathbf{r}) = -\frac{1}{\epsilon}\rho(\mathbf{r}), \tag{3.20}$$

with ϵ as the permittivity of the medium. As we use point charges in MD simulations, the charge density $\rho(\mathbf{r})$ is a set of delta functions, which complicates solving the partial differential equation eq. 3.20. Using Ewald summation, we make a detour by splitting up the charge density into a short-range and a long-range part, which we then solve separately. The short-range part $\rho^{\mathrm{SR}}(\mathbf{r})$ contains the original point charges with charge q_i as well as Gaussian-shaped charge distributions of magnitude $-q_i$:

$$\rho_i^{\mathrm{SR}}(\mathbf{r}) = q_i\delta(\mathbf{r}-\mathbf{r}_i) - q_i G_\alpha(\mathbf{r}-\mathbf{r}_i), \tag{3.21}$$

with G_α as a Gaussian distribution defined as:[113]

$$G_\alpha(\mathbf{r}) = \frac{\alpha^3}{\pi^{3/2}}\exp\left(-\alpha^2\mathbf{r}^2\right). \tag{3.22}$$

In other words, we limit the area, in which the point charge q_i at position $\mathbf{r}_i$ can be felt by others. The long-range part $\rho^{\mathrm{SR}}(\mathbf{r})$ is a compensation for the Gaussians added in the short-range term:

$$\rho_i^{\mathrm{LR}}(\mathbf{r}) = q_i G_\alpha(\mathbf{r}-\mathbf{r}_i). \tag{3.23}$$

Hence, adding $\rho_i^{\mathrm{SR}}(\mathbf{r})$ to $\rho_i^{\mathrm{LR}}(\mathbf{r})$ still gives the original charge distribution of eq. 3.20.

However, the usage of Gaussian-shaped functions has computational advantages:[112]
For the short-range part, we can now use cut-offs as the charges are shielded and therefore their range is limited. For the long-range part, we can now consider Fourier transformations as $q_i G_\alpha$ by its own is periodic and smooth. The electrostatic contribution to the potential energy finally is:

$$E^{\mathrm{C}} = E^{\mathrm{SR}} + E^{\mathrm{LR}} - E^{\mathrm{self}}, \tag{3.24}$$

where E^{self} is the interaction between the particle i with q_i and its own compensating charge cloud $q_i G_\alpha$.

The usage of Ewald sums can be even more tuned-up by discretising the positions of the charges into mesh positions, which allows for the usage of Fast Fourier Transformations. This approach is then called 'Particle-Mesh-Ewald' method.[129,130] It is well established in MD programs and also used in this thesis, but its discussion is beyond the scope of this chapter.

For background information on the mathematical context, Refs. [103, 112, 113, 125, 131, 132] are strongly recommended.

3.1.2 Setup and (Post)Processing

Working with MD simulations basically comprises three different steps: (1) Setup and data production, (2) post-processing and dimensionality reduction (if needed) and (3) analysis.

MD setup

There is excellent literature on the successful setup of MD simulations[112,133] and for the detailed simulation protocols used in this thesis, please refer to the individual papers (see 4.1, 4.2, 4.3, 4.4). However, let us call to mind some pivotal decisions during the setup before discussing the points (2) and (3). As apparent from the previous section (see 3.1.1), the force field represents the centrepiece of MD simulations. This includes both the formulation of the potential energy functions and the parameters set for the individual atoms. Hence, the force field has to be chosen in accordance to the system. For peptides and proteins comprising standard amino acids, the force field parameters of the "AMBER" family are well-suited.[116,118] For the amatoxins investigated in the subprojects **B1**-**B3**, however, these force fields are of no use since the unusual tryptathionine linkage is not covered by the AMBER force field parameters. Here, programs like "ACPYPE"[134] can help out as they create molecule-specific topologies in close relation to the AMBER force field parameters.[135,136]

Next, the origin of the starting structure for an MD simulations can have massive impact on the conformational ensemble. Structures from X-ray crystallography are usually good choices for starting structures, but they can sometimes be misleading, especially if the behaviour of the

peptide or protein alters substantially going from solid (crystal) state to solution.[104] For that reason, NMR solution structures are better starting structures. In subproject **A**, for instance, the MD simulations were started from the solution structure of tWW of h-FBP21. Both X-ray crystal and NMR solution structures are available in the Protein Data Bank (PDB).[137–139]

Lacking experimental evidence for the structure of the targeted system, the choice of the starting structure is difficult. Small molecules like the amatoxins investigated in subprojects **B1** - **B3**, can be constructed with molecular editors comparable to the molecule construction kit from school. For larger molecules like proteins, however, a construction by hand is neither feasible nor justifiable without further leads due to the many construction possibilities. In general, this problem is known as "protein folding problem",[140] in which context one important paradox has been introduced by Cyrus Levinthal in 1969.[141] The amino acid chain in a protein has many degrees of freedom for self-assembling. Imagine an amino acid sequence of roughly 100 residues that can adopt 10^{70} configurations based on the combinations of (torsion) angles. According to a "random-search" problem as suggested by C. Levinthal, all these configurations except for the correct configuration (native state) are equally likely.[140,142] Let us say it takes one picosecond (10^{-12} s) to find one of these configurations, then the time to find the native state would be $10^{70} \cdot 10^{-12} = 10^{58}$ s (roughly 10^{50} years), an enormous amount of time. This expenditure of time to find the correct protein folding has been puzzling scientists who work on protein structure prediction until the game-changing program "AlphaFold" became publicly available.[143] The "AlphaFold" program is an artificial intelligence-based program, which has convincingly proven to predict protein structures in atomic precision.[144] Hence, using protein folding prediction software such as "AlphaFold" also represents a possible source for starting structures.

After selecting a suitable starting structure, it usually follows an energy minimisation, the solvation in (explicit) solvent and equilibration simulations under the desired conditions. This will not be further discussed here. Though, in terms of the energy minimisation one should keep in mind, however, that the algorithms such as "steepest-descent" or "conjugate-gradient" always find the nearest minimum, which is not necessarily the global minimum on the potential energy surface.[113]

Dimensionality reduction

Once the MD simulations are done, we analyse the MD data to draw conclusions about the properties of the investigated system. For that, we first need to interpret the typical outcome of MD simulations, the *trajectory*. The trajectory is the output of the integration schemes described earlier (see eq. 3.10, eq. 3.11) in a specific time interval, usually multiples of the integration time steps. The trajectory therefore contains the time series of the Cartesian coordinates $\left(\mathbf{r}_i(t) = (x_i, y_i, z_i)^{\mathrm{T}}\right)$ of each atom i selected in the output control. Trajectories can be seen as a collection of snapshots (frames) provided by our computational microscope,

where the entries are the atom coordinates in the Euclidean space. Note that the atom selection (N) and the frequency, in which the positions of these atoms is written to file, can be set by the user.

The trajectories themselves do not contain any readable information about the system. Only in combination with a *topology* file, in which each entry is assigned to a specific atom of our system, we can directly extract properties, which can be computed with linear algebra, e.g. distances, angles. Usually, these properties have in common that we analyse them as ensemble averages or distributions (see 3.1.3) and to get a first impression of the system, these kind of properties are certainly useful.

In most cases though, we would like to learn more about the different conformations and possible transitions between them. For that purpose, we have to subdivide the MD data into meaningful subsets. This can be achieved by clustering techniques (see 3.3.1). Working with raw MD data, we deal with very high dimensionality since each frame of the trajectory in principle contains $3N$ dimensions for an output of N atoms. For clustering and kinetic analyses, raw MD is unsuited since its high dimensionality is not computationally feasible and not all atoms are equally important for the dynamics of the system.[145] As consequence, a sub-dimensional representation of the MD data is needed, which can be obtained by dimensionality reduction techniques.

The basic idea of proper dimensionality reduction is to put as much important information in as little dimensions as possible. Here, we will discuss time-independent component analysis (TICA), a technique whose application to MD data was introduced by Refs. [146, 147]. Please note, before using TICA, we should have an idea of meaningful input features. For peptides and proteins, especially the backbone torsion angles (ϕ, ψ, ω) are interesting, as they capture best the backbone dynamics. Sometimes also the end-to-end distance is useful, for instance if we consider folding events like in the hexapeptide NFGAIL.[148]

The following description of the TICA method is based on Refs. [146, 147]:

Let us assume we have translated the trajectory of atomic positions $\mathbf{x}(t) = \{\mathbf{r}_1(t), ..., \mathbf{r}_N(t)\} \in \Omega \subset \mathbb{R}^{3N}$ into a trajectory of N' features with d-dimensions $\mathbf{x}'(t) = \{\mathbf{f}_1(t), \cdots \mathbf{f}_{N'}(t)\} \in \Omega' \subset \mathbb{R}^d$. For example, these features could contain the backbone torsion angles for each residue of our system: $\mathbf{f}_i = (\phi_i, \psi_i, \omega_i)^{\mathrm{T}}$. Ω' is called *feature space*, a sub-dimensional representation of the conformational space Ω. The first step in TICA is that our input data are transformed into mean-free data by:

$$\mathbf{z}_i(t) = \mathbf{f}_i(t) - \langle \mathbf{f}_i(t) \rangle_t \,, \tag{3.25}$$

where $\langle \mathbf{f}_i(t) \rangle_t$ denotes the time average. Then, "time-lagged" covariance matrices $\mathbf{C}(\tau)$ are computed for $\tau = 0$ and another positvely-valued lag time τ, usually multiples of our MD time step Δt. For finite data such as our MD data, the elements of these matrices are defined

as:

$$C_{ij}(\tau) = \frac{1}{N_{\text{steps}} - \tau - 1} \sum_{t=1}^{N_{\text{steps}}-\tau} \mathbf{z}_i(t)\mathbf{z}_j(t+\tau), \tag{3.26}$$

where N_{steps} denotes the length of our MD trajectories in time steps. Before going on, $\mathbf{C}(\tau)$ is symmetrised:

$$\mathbf{C}(\tau) = \frac{1}{2}\left(\mathbf{C}_b(\tau) + \mathbf{C}_b^{\mathrm{T}}(\tau)\right), \tag{3.27}$$

where $\mathbf{C}_b(\tau)$ denotes the covariance matrix **b**efore enforcing the symmetry.

For the symmetrised covariance matrix, the generalised eigenvalue problem is solved:

$$\mathbf{C}(\tau)\mathbf{W} = \mathbf{C}(0)\mathbf{W}\mathbf{\Lambda}, \tag{3.28}$$

with $\mathbf{W}$ as the eigenvector matrix containing the *time-independent components* ("TICs"), $\mathbf{\Lambda}$ as the diagonal eigenvalue matrix and $\mathbf{C}(0)$ as the covariance matrix ($\tau = 0$).

Our mean-free feature vectors are finally projected onto the TICA space by:

$$\zeta^{\mathrm{T}}(t) = \mathbf{z}^{\mathrm{T}}(t)\mathbf{W}. \tag{3.29}$$

In eq. 3.28, we try to manifest two desired properties of the TICs in order to minimise the projection error, i.e. the loss of information due to the dimensionality reduction: To cover as much of the information from the input data, (1) the TICs shall be uncorrelated and (2) their autocovariances (diagonal elements of $\mathbf{\Lambda}$) maximal for a given lag time τ. Thus, $\mathbf{W}$ is defined to diagonalise $\mathbf{C}(\tau)$ while maximising the autocovariances λ_{ii}. To achieve this, i.e. to solve eq. 3.28, Refs. [146, 147] suggest a detour involving the AMUSE algorithm.[149] In the implementation of TICA that has been used in this thesis,[150] this detour is also used.

In brief, the detour adds a coordinate transformation prior to the computation of the "time-lagged" covariance matrix $\mathbf{C}(\tau)$. By using principal-component analysis (PCA),[151,152] the mean-free input data $\mathbf{z}_i(t)$ is transformed into principal components $\mathbf{l}_i(t)$:

$$\mathbf{l}_i(t) = \mathbf{O}\mathbf{z}_i(t), \tag{3.30}$$

where $\mathbf{O}$, the matrix containing the principal components can be obtained by solving the PCA eigenvalue problem:

$$\mathbf{C}(0)\mathbf{O} = \mathbf{O}\mathbf{Y}, \tag{3.31}$$

with $\mathbf{C}(0)$ as the covariance matrix. The (diagonal) eigenvalue matrix $\mathbf{Y}$ is then used to normalise $\mathbf{l}_i(t)$ by $\mathbf{l}'_i(t) = \mathbf{Y}^{-1}\mathbf{l}_i(t)$. Then, eq. 3.26 is solved for $\mathbf{l}'_i(t)$ yielding $\mathbf{C}'(\tau)$. As described for the time-lagged covariance matrix on the mean-free feature data, we enforce the symmetry of $\mathbf{C}'(\tau)$ (eq. 3.27).

As final step, an eigenvalue problem of $\mathbf{C}'(\tau)$ is solved:

$$\mathbf{C}'(\tau)\mathbf{V} = \mathbf{\Lambda}'\mathbf{V}, \tag{3.32}$$

where $\mathbf{V}$ denotes the eigenvector matrix of $\mathbf{C}'(\tau)$. Then, $\mathbf{l}_i(t)$ is projected onto the most dominant eigenvectors to obtain the desired TICA projection. Referring to eq. 3.29, the detour can be formulated as:

$$\zeta_i^{\mathrm{T}}(t) = \mathbf{z}_i^{\mathrm{T}}(t)\mathbf{O}\mathbf{Y}^{-1}\mathbf{V}. \tag{3.33}$$

By design, the TICs in $\mathbf{W}$ are ordered according to the magnitude of their autocovariances and the TICs with the largest autocovariances together yield a linear subspace of Ω', in which most of the slow processes are contained. Hence, for ensuing analyses, e.g. clustering (see 3.3.1), it is sufficient to focus on the dominant TICs as this should not increase the projection error. To select the most dominant TICs from negligible ones, a look into the eigenvalue spectrum ($|\lambda_{ii}(\tau)|$ against the TICs) is quite useful. Here, we can look out for a gap, i.e. a significant difference in the magnitude of λ_{ii} between two consecutive TICs.

Taken together, time-independent component analysis (TICA) is a valuable method to reduce the dimensionality of the MD data. Note, however, that we have to consider at least two dimensionality reductions: first on the feature space and second on the (most dominant) TICs. As consequence, it is difficult to give the TICs an exact physical meaning. Whether the dimensionality reduction went successful can then only be proven by the ensuing steps such as clustering (see 3.3.1) and Markov-state-models (see 3.1.3). In the next section (see 3.1.3), we will review some aspects on the analysis of MD simulations either with or without prior dimensionality reduction.

3.1.3 Analysis of MD simulations

Once one has obtained the trajectories of the system of interest, these trajectories are usually fitted and centred on the key-molecule before any analysis is performed. Also, to speed up the analyses the time resolution is often decreased by focussing on every Xth frame instead of on every single frame.

Properties directly derived from the MD trajectory

Assuming proper sampling, the trajectories resulting from the MD simulations fulfil the ergodic hypothesis (eq. 3.15). In other words: We assume that after an infinite time, each state will be visited an infinite number of times proportional to the stationary distribution. Since we are considering a system with constant amount of atoms N, we can use this assumption to calculate the ensemble average of a given observable A as time average

over our trajectories:[113,153]

$$\langle A \rangle = \frac{1}{N_{\mathrm{rep}}} \sum_{i=1}^{N_{\mathrm{rep}}} \langle A\left(\mathbf{x}_i(t)\right) \rangle = \frac{1}{N_{\mathrm{rep}}} \sum_{i=1}^{N_{\mathrm{rep}}} \left[\frac{1}{N_{\tau,i}} \sum_{\tau=0}^{N_{\tau,i}-1} A(\mathbf{x}_i(\tau)) \right] \tag{3.34}$$

with N_{rep} as the number of considered trajectories $\mathbf{x}_i$ with $N_{\tau,i}$ time steps τ. As apparent from eq. 3.34, the validity of the ensemble averages derived from our MD simulations massively depends on the sampling. Aside, assuming proper sampling and an NVT- or NpT-ensemble, the possible configurations $\mathbf{x}(t)$ are distributed according to the Boltzmann distribution:[113]

$$\pi(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{V(\mathbf{x})}{k_{\mathrm{B}} T} \right). \tag{3.35}$$

$\pi(\mathbf{x})$ is the probability for the molecule to adapt configuration $\mathbf{x}$ in equilibrium. Z is called "partition function" ("*Zustandssumme*"), the weighted collection of all possible configurations. $k_{\mathbf{B}}$ is the Boltzmann constant, T the temperature and $V(\mathbf{x})$ the potential energy function, which is expressed by the force field (see eq. 3.2, eq. 3.3).

For biological systems, particularly structure-related features such as interatomic distances, (backbone) dihedral angles or hydrogen bonds are of special interest. Based on eq. 3.34, these properties can be directly computed from the MD data - as average values, populations or distributions.

Kinetic networks

Up to this point, we have discussed properties, which are directly extracted from a (sub)data set of the MD ensemble. Yet, the time information of the MD simulations is only included qualitatively. It is more decisive where the system is most of the time than how it changes in between. In the introduction, however, we have seen that within the context of biological systems, we are especially interested in understanding the conformational dynamics of the systems. Here, not all conformational transitions are of the same interest. Usually, we are interested in the conformational changes on the longest timescales, since they represent the processes, which are more interesting from a biochemical point of view. Hence, we would like to describe the transitions between the most important states, where each state can contain multiple conformations in fast conformational exchange.

For that purpose, Markov State Models (MSMs) have proven very effective. In the following, we will shortly review some features of these models using Refs. [154, 155]. For the complete picture, please also refer to Refs. [156–158]. MSMs are based on the idea that the dynamics of a system can be analysed as Markov jump process if the process $\mathbf{x}(t)$ is (1) ergodic (see

above 3.1.3) and (2) reversible. If $\mathbf{x}(t)$ is reversible, the condition of *detailed balance* is fulfilled:

$$\pi(\mathbf{x})p(\mathbf{x},\mathbf{y};\tau) = \pi(\mathbf{y})p(\mathbf{y},\mathbf{x};\tau), \tag{3.36}$$

which means that within a lag time $\tau \in \mathbb{R}_{0+}$, the number of transitions from configuration $\mathbf{x}$ to $\mathbf{y}$ is equal to the number of transitions in the opposite direction, if the system is in thermal equilibrium.

In MSMs, we consider probability distributions of the configurations $\mathbf{x}$ at different time steps, and since we assume the process $\mathbf{x}(t)$ to be *markovian*, we can formulate:[154]

$$p(\mathbf{x},\mathbf{y};\tau) = \mathbb{P}\left[\mathbf{y} = \mathbf{x}\left(t+\tau\right) \in \Omega \mid \mathbf{x} = \mathbf{x}(t).\right] \tag{3.37}$$

Eq. 3.37 states that the probability of being in state $\mathbf{y}$ at time $t+\tau$ only depends on the state $\mathbf{x}$ at time t. Ω denotes the full state space of our system. Hence, the dynamics of the MD ensemble can be expressed as propagation of probability densities in time satisfying the Chapman-Kolmogorov condition:

$$p(\mathbf{x};t+k\tau) = \left[\mathcal{P}(\tau)\right]^k\ p(\mathbf{x};t) = \mathcal{P}(k\tau)\,p(\mathbf{x};t) \quad k \in \mathbb{N}, \tag{3.38}$$

where $\mathcal{P}$ is the operator called "propagator" and $p(\mathbf{x};t)$ is the probability of being in configuration $\mathbf{x}$ at time t. For mathematical reasons,[154,159] MSMs consider the closely-related "transfer operator" instead of the propagator:

$$u(\mathbf{x};t+k\tau) = \left[\mathcal{T}(\tau)\right]^k\ u(\mathbf{x};t) = \mathcal{T}(k\tau)\,u(\mathbf{x};t) \quad k \in \mathbb{N}, \tag{3.39}$$

where $u(\mathbf{x};t)$ is the probability of being in configuration $\mathbf{x}$ at time t weighted by the Boltzmann density: $u = \frac{p}{\pi}$ (eq. 3.35). Note, that eq. 3.38 and eq. 3.39 hold for *continuous* dynamics for the full state space Ω.

In MD simulations, we are bound to the integration time step $\tau = k \cdot \Delta t$ (see eq. 3.10, eq. 3.11), which means that we have a discrete time. Also, the MD ensemble is too complex to derive the propagator or the transfer operator, which means that we also have to consider a dimensionality reduction and discretisation for the state space. As dimensionality reduction technique, methods such as TICA (see 3.1.2) come into play as they reduce the system to select dimensions while preserving the kinetic information. For the discretisation of the sub-dimensional state space Ω', clustering methods are usually applied (see 3.3.1). Working on a sub-dimensional, discretised representation of Ω and with discrete time allows us to approximate the operators in eq. 3.38 and eq. 3.39 by matrices:

$$p^{\mathrm{T}}(t+k\tau) = p^{\mathrm{T}}(t)\,\mathbf{T}^k(\tau) \quad k \in \mathbb{N}, \tag{3.40}$$

where $\mathbf{T}$ is called "transition probability matrix" and p^{T} is a vector containing the probabilities of the system to be in a certain subset $S_i \subset \Omega'$ at time t. Note, that our feature space Ω' is discretised in disjoint subsets $S_i \cap S_j = \emptyset \quad \forall i, j$. The matrix elements of $\mathbf{T}$ are defined as:

$$T_{ij}(\tau) = \mathbb{P}\left[\mathbf{x}(t+\tau) \in S_j \,|\, \mathbf{x}(t) \in S_i\right]. \tag{3.41}$$

In words, each matrix element is the conditional probability that the system will be in subset S_j at time $t + \tau$ given that it has been in subset S_i at time t.

At this point, please note three things:
First, the subsets S_i are directly the clusters obtained by clustering Ω'.[154]

Second, in terms of definition of the subsets and their relation to Ω', we can either consider a "full-state" representation or a "core-set" approach. In the "full-state" representation, the subsets S_i fully cover Ω', while in the "core-set" approach, we reduce the system's dynamics on select sets, which are only a subset of Ω'.

Third, by discretising the state space and by reducing the dimensionality, we introduce a *discretisation error*, which reflects the deviation in the dynamics between the discretised and the original system. The reason for the deviation is that the discretised dynamics are not truly Markovian for all lag times τ. Here, a "implied timescale test" helps to choose a lag time, at which the processes in the discretised space are approximatly Markovian.

In practical, we can determine the transition matrix elements by considering correlation matrices as shown for TICA:

$$T_{ij}(\tau) = \frac{C_{ij}(\tau)}{\sum_j C_{ij}(\tau)} \tag{3.42}$$

with

$$C_{ij}(\tau) = \frac{1}{N_{\mathrm{frames}} - \tau} \sum_{t=0}^{N_{\mathrm{frames}}-\tau} \chi_i(t)\chi_j(t+\tau), \tag{3.43}$$

where N_{frames} denotes the number of frames in our discretised, sub-dimensional MD trajectories. χ_i is a vector that indicates whether the system is in S_i at time t:

$$\chi_i(t) = \begin{cases} 1 & \text{if} \quad \mathbf{x}(t) \in S_i \\ 0 & \text{else} \end{cases} \tag{3.44}$$

T_{ij} contains the counts of the transitions $S_i \to S_j$ normalised by the total amount of arrival subsets S_j. Note, $\mathbf{C}$ is calculated for each MD trajectory individually. For the entire ensemble, the correlation matrix can be obtained by summing over the different correlation matrices.[159]
Second, to enforce detailed balance, the correlation matrix can be artificially symmetrised

as shown for TICA in eq. 3.27.

For the actual analysis of the (core-set) MSM, we consider the eigenvectors and eigenvalues of $\mathbf{T}$:

$$\mathbf{T}(\tau) = \sum_{s}^{N_{\text{sets}}} r_s \, \lambda_s(\tau) \, l_s^{\mathrm{T}}, \tag{3.45}$$

where r and l are the right and the left eigenvectors of $\mathbf{T}$ and N_{sets} is the number of subsets in the system. Due to the assumed reversibility, $\mathbf{T}(\tau)$ is self-adjoint, and if $\mathbf{T}(\tau)$ is self-adjoint, the eigenvectors are orthonormal:[154]

$$l_i^{\mathrm{T}} r_j = \delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{3.46}$$

This allows to write any solution $p^{\mathrm{T}}(t)$ (see eq. 3.40) as a linear combination of eigenvectors:

$$p^{\mathrm{T}}(t + k\tau) = \sum_{s}^{N_{\text{sets}}} c_s \, \lambda_s^k(\tau) \, l_s^{\mathrm{T}}, \tag{3.47}$$

where the coefficients c_s come from projecting $p^{\mathrm{T}}(t)$ on the left eigenvectors of $\mathbf{T}$: $c_s = \langle p(t), l_s \rangle$. Given the properties of $\mathbf{T}$, the first eigenvalue $\lambda_1 = 1$ and the first (left) eigenvector only contains positive entries. All remaining eigenvalues hold $0 \geq \lambda_{s>1} < 1$. In terms of the interpretation, the eigenvectors contain subset-specific weights for the different processes. The first (left) eigenvector can be used to obtain the equilibrium distribution. The other eigenvectors $s > 1$ can be used to project the dynamics onto the different processes. Typically, the eigenvalue spectrum shows a gap, where one group of eigenvalues is close to the first eigenvalue and the other group clearly falls off. For the analysis of the dynamics of the system, we only focus on the eigenvalues close to 1 as well as the associated eigenvectors as they represent the slowest processes in the system. But what does that mean for the physical interpretation of the dynamics?

As you can see from eq. 3.45 and eq. 3.47, the time information does only concern the eigenvalues λ_k. Hence, the eigenvalues can be given a physical meaning by calculating the "implied timescales":

$$t_s = -\frac{\tau}{\ln(\lambda_s(\tau))}. \tag{3.48}$$

Since $\lambda_{s>1} < 1$, they exponentially decay as s goes to infinity. Therefore, the implied timescales represent the decay time of all the processes $s > 1$. As $\lambda_s = 1$, after a long time, only the eigenvector corresponding to the equilibrium distribution remains.

If the dynamics are Markovian, the Chapman-Kolmogorov condition holds (eq. 3.38,eq. 3.39), which means that the implied timescale is independent of the chosen lag time τ. If we then plot the implied time scales t_s against the lag time τ, we will observe plateaus ("implied times-

cale test"). Note however, that the convergence of t_s in τ alone is not a proof for Markovianity, since for Markovianity we must also check the convergence of the eigenvectors.[159]

Implied timescales, which hold $t_s > \tau$, can be considered to evaluate the MSM. Due to the fact that $\lambda_{s=1} = 1$ and $\ln(1) = 0$, the first defined implied timescale belongs to the second (left) eigenvector of $\mathbf{T}$. To visualise the corresponding processes, we can plot (1) the values of the eigenvectors against the subset indices or (2) the projection of the dynamics on the eigenvectors. The eigenvectors $s > 1$ represent a kinetic exchange between regions with positive and negative values. The transition at the timescale t_s is then the transition between the regions of opposite sign. By extracting the structures belonging to the respective subsets (clusters), we can thus easily associate the kinetic information to structural changes, which makes MSMs very powerful.

3.1.4 Issues and limitations of MD simulations

In the preceding sections, we have discussed important aspects for the theoretical fundament (see 3.1.1), the setup (see 3.1.2) and the analysis (see 3.1.3) of classical MD simulations. In this section, we will address two main issues of classical MD simulations: (1) the "sampling problem" and (2) problems that arise from our model representation.

Since MD simulations are a computational method, we are restricted to the available resources (CPUs,GPUs). However, it has been a major achievement that parallelisation schemes are available that allow us to use the resources more efficiently.[160,161] But even with more computational resources, there are systems which are still very hard to sample with classical MD. For instance, if the system becomes very large, and/or if we are trying to model transitions on very large timescales (multiples of seconds). Very large systems have the disadvantage that the potential energy surface becomes very complex due to the large number of degrees of freedoms.[125] Transitions in the regime of seconds imply very high energy barriers, which the system rarely overcomes at room temperature. Both aspects are very problematic if we are interested in ensemble properties, since we cannot justifiably assume that we have sampled the entire conformational space. To tackle these problems, enhanced sampling methods such as metadynamics or replica-exchange might be advised.[103,162] Depending on the size or complexity of the system though, high-temperature simulations might also be sufficient (see 4.4).

The most decisive feature of MD simulations is the force field, and the approximations made therein have great influence on the results.[125,163] It is difficult to compare simulation that use different force fields, as they can differ in the definition of the potential energy function $V(\mathbf{x})$ and the atom-specific parameters. The difference in the atom-specific parameters results from the different experimental data based on which the force fields have been parametrised.[116,163] Hence, we have to consider the force field that has been parametrised

for molecules comparable to our targeted system.
Next, the structure and the atom-specific parameters are not changeable during an MD simulation: all atoms are hard spheres with a definite radius, partial charge and specific binding partners. As consequence, bond-breaking or -forming, e.g. deprotonation events, as well as excited states cannot be modelled with MD. Also, polarisation effects are only treated implicitly by partial atomic charges, which however overestimate dipoles.[163] To get access to these properties, one should consider hydride methods that combine classical mechanics with quantum-mechanical calculations (QM/MM).[125,164]

Especially non-bonded interactions are troublesome in MD: They are defined as pairwise potentials, which involves increased computational costs if the system becomes very large, and we still would like to study it in all-atom precision. Depending on the targeted property, "coarse-grained" models can however help to still model these systems efficiently.[125] Also, the non-bonded interaction between directly linked atoms (up to two three bonds) are usually neglected in favour of the desired geometries and bond angles. The force fields differ in the way they incorporate the non-bonded interactions in torsion angles.[112,163] This again emphasises that MD simulations using different force fields are difficult to compare.

Last, we use periodic-boundary conditions, which is advantageous for the computation of long-range interactions. However, this implicates a very high effective concentration due to the multiple copies of the cell, and we cannot calculate transport properties correctly.[112,125] In summary, the atomic precision of MD simulations can give a false sense of correctness in terms of the results. The fact that we obtain trajectories from our MD algorithm does not guarantee that our choice of the model, force field, system or protocol were indeed correct to be able to analyse the property of interest.[112,125] Thus, like for other experimental methods, we should be careful to handle the results correctly.

3.2 Nuclear magnetic resonance spectroscopy

NMR spectroscopy is an extremely powerful tool for the elucidation of biological systems. Using different techniques, NMR experiments can provide valuable information about: structural distances and geometries, flexibility of structural elements or entire regions, conformations in solution and possible transitions between them, solvent accessibility of amides, structural changes due to binding processes or other intermolecular interactions.[165] However, like other spectroscopical techniques, NMR experiments always provide average values over the time of the measurement. Nevertheless, their deep structural insight makes NMR-derived data very important for screening or verifying MD simulations.

In this chapter, we will recall some physical aspects of NMR spectroscopy, before we will discuss features that are usable for MD simulations.

3.2.1 Physical aspects in NMR spectroscopy

This section is based on Refs. [166–169]. Please refer to these references for further details.

In NMR spectroscopy, we depart from the idea that atoms are hard balls with a specific partial charge. Instead, each atom consists of elementary particles (electrons, protons, neutrons), which are either in the atomic nucleus (protons, neutrons) or surround it (electrons). In **nuclear** magnetic resonance spectroscopy, we make use of the principle that a nucleus, as it contains charged particles, has a magnetic moment (nuclear spin). According to quantum mechanics, the resulting angular momentum (nuclear spin angular momentum) is a vector quantity characterised by quantum numbers I:[166]

$$|\mathbf{p}| = \frac{\mathrm{h}}{2\pi}\sqrt{I \cdot (I+1)}, \tag{3.49}$$

where $|\mathbf{p}|$ is the magnitude of the vector and h is Planck's constant. The nuclear spin quantum number I depends on the number of protons and neutrons in the nucleus. Here, we can differentiate three classes of nuclei:

1. Nuclei containing odd numbers of protons and neutrons: $I = 0$ (example: ^{12}C)
2. Nuclei containing (a) an odd number of protons and an even number of neutrons or (b) an even number of protons and an odd number of neutrons: $I = n \cdot \frac{1}{2}$ with $n \in \mathbb{N}$ (example: ^{1}H, ^{13}C)
3. Nuclei containing even numbers of protons and neutrons: $I = 1$ (example: ^{2}H)

For the angular momentum, there are $2I + 1$ possible orientations in space:

$$p_z = \frac{h}{2\pi}m, \tag{3.50}$$

where m denotes the magnetic quantum numbers in the range $m = (-I, -I+1, \ldots, I-1, I)$ and p_z is the magnitude of the angular momentum in the z-direction. Without external influences, the $2I+1$ different orientations are energetically degenerated.

The magnetic moment $\boldsymbol{\mu}$ (also a vector) and the angular momentum $\mathbf{p}$ of the nucleus are linked by the gyromagnetic ratio γ, which differs for each isotope:

$$\gamma = \frac{\boldsymbol{\mu}}{\mathbf{p}}, \tag{3.51}$$

and hence:

$$|\boldsymbol{\mu}| = \gamma\,|\mathbf{p}| = \gamma\,\frac{\mathrm{h}}{2\pi}\sqrt{I(I+1)}, \tag{3.52}$$

$$\mu_z = \gamma\,p_z = \gamma\,\frac{\mathrm{h}}{2\pi}m. \tag{3.53}$$

In NMR spectroscopy, we place the nuclei in a strong magnetic field, which forces the magnetic moment to align along the magnetic field axis. For our convenience, let us assume that the external magnetic field is also oriented in positive z-direction. The presence of an external magnetic field has two consequences: First, the nuclear spin rotates around the magnetic field axis ("precession") with a specific frequency, called "Larmor frequency":[168]

$$\omega = -\gamma B_0, \tag{3.54}$$

where B_0 denotes the magnetic field strength. Here, confusion could arise, since in the context of NMR measurements, γB_0 is also called "magnetic field strength" and given in Hertz ([Hz]).[166]

Second, the $2I+1$ orientations of $\boldsymbol{\mu}$ are now no longer energetically equal. The orientation(s) in the same direction as the magnetic field are energetically favourable (α state(s)), while the orientation(s) in the opposite direction are less favourable (β state(s)). These energy states, also referred to as "Zeeman levels",[166] directly depend on the magnetic field:

$$E_m = -\mu_z \cdot B_0 = -\gamma m\frac{\mathrm{h}}{2\pi}B_0. \tag{3.55}$$

The energy states are populated according to the Boltzmann distribution. For two states, we can therefore formulate:

$$\frac{N_\alpha}{N_\beta} = \exp\left(\frac{\Delta E}{k_\mathrm{B}T}\right), \tag{3.56}$$

with

$$\Delta E = \gamma\frac{\mathrm{h}}{2\pi}B_0, \tag{3.57}$$

and $N_{\alpha/\beta}$ as the population numbers of the states α and β. From eq. 3.57, we see that the

energy difference depends on the type of the nuclei (γ) and the magnetic field strength (B_0).

To change the populations of the energy states, we can use electromagnetic radiation (radiowaves). If we provide the correct energy, transitions are possible from the α- to the β-state, i.e. the nuclei are "in resonance" with the external radiation. From eq. 3.54 and eq. 3.57, we conclude the condition for resonance:

$$\nu = \frac{\Delta E}{\mathrm{h}} = \frac{\gamma B_0}{2\pi}. \tag{3.58}$$

Hence, the resonance of the nuclei can either be achieved varying the frequency ("frequency sweep method") or by variying the magnetic field ("field sweep method").[167]

Usually, when we use NMR spectroscopy, we do not deal with single atoms, but rather with many molecules. Hence, we do not consider single nuclear moments, but the macroscopical magnetisation $\mathbf{M}$, the sum of all μ_z-components of a sample.

In a static field (B_0), i.e. after placing the probe in the NMR spectrometer, we obtain the equilibrium magnetisation M_0. Note that in contrast to the nuclear magnetic moment, M_0 can be collinear to the magnetic field, which is why we do not get a signal from the probe without excitation energy.

If we provide the resonance frequency, we not only change the population of the Zeeman levels, but we also change the orientation of the macromolecular magnetisation. As shown for the nuclear magnetic moment, the macroscopical magnetisation then precesses around the axis of the magnetic field (eq. 3.54). After excitation, $\mathbf{M}$ relaxates back to M_0 and the spectrometer records a "free induction decay", which is then transformed into a spectrum by Fourier transformation.[167]

Regarding the relaxation process, two types are distinguished: (1) "spin-lattice" (T_1) and (2) "spin-spin" relaxation (T_2). In the "spin-lattice" relaxation, the energy from nuclei that return from the excited state, is transferred to other molecules or atoms in the sample. In the "spin-spin" relaxation, this energy is transferred to other precessing nuclei.[167]

Note, the precession of $\mathbf{M}$ after excitation depends on the type of nuclei and the electronical environment of the nuclei in the molecules.[166] The electrons surrounding the nuclei (core and valence electrons) can cause a shielding of the particular nuclei from the magnetic field. Hence, these nuclei are not exposed to the total magnetic field strength:

$$B_{\mathrm{eff}} = (1 - \sigma)B_0, \tag{3.59}$$

where B_{eff} is called the "effective" magnetiv field strength and σ denotes the shielding constant. This has direct consequences for the resonance frequency, with which we would

like to excite these nuclei:

$$\nu_{\text{eff}} = \frac{\gamma\, B_0}{2\pi} \cdot (1 - \sigma). \tag{3.60}$$

In order to compare NMR spectra at different magnetic field strengths, the signals are presented in dependence of the "chemical shift".[166]

$$\delta = \frac{\nu - \nu_{\text{ref}}}{\nu_{\text{ref}}} \cdot 10^6 = (\sigma - \sigma_{\text{ref}}) \cdot 10^6, \tag{3.61}$$

where ν_{ref} denotes the resonance frequency of the precessing nuclei in the reference, e.g. HCD_3 for ^{1}H-NMR spectroscopy. As defined in eq. 3.61, the chemical shift is directly correlated to the shielding of the particular nuclei. Hence, with the chemical shift we can draw conclusions on the shielding of the nuclei influenced by nearby-charges, electronegativity, inductive effects or mesomeric effects.

Let us consider a well-resolved, one-dimensional ^{1}H-NMR spectrum (intensity *vs.* frequency spectrum). For the analysis, the most important aspects are the intensity, the position and the shape of the signals. As described before, the position of a signal is determined by the chemical shift. From the signal's height one can deduce information about the amount of chemically-equivalent protons that possess the same chemical shift. The shape of the signal is determined by the amount and type of NMR-active protons that are in close vicinity, i.e. "reachable" within three covalent bonds. In a simple case, the signal of a proton that couples with n chemically-equivalent protons, would split up $n + 1$-fold distributing the signal's intensity in ratios according to Pascal's triangle. For example, the scalar coupling of a proton towards two chemically-equivalent protons would yield a signal shaped as triplet with intensity rations 1:2:1. The peaks are then separated by J Hz where J denotes the "coupling constant".[168]

For peptides and proteins, however, ^{1}H-NMR spectra are almost impossible to analyse due to the large amount of different protons and the overlapping of their signals. As consequence, it has become common practice to perform higher-dimensional NMR experiments by recording a series of 1D-NMR spectra and combining them to a, e.g. frequency *vs* frequency spectrum, which usually shows the intensities on the third axis as contours ("out of the paper plane"). These 2D-NMR spectra can either be homo- or heteronuclear, and they are particularly helpful to gain insights into the spatial environment of the stimulated nuclei. The reason for that is they provide the chemical shifts of each pair of interacting nuclei as well as the intensities of these interactions.[168] Well-established types are the 2D-NOESY (**N**uclear **O**verhauser **E**nhanced **S**pectroscop**Y**) and the 2D-COSY (**CO**rrelated **S**pectroscop**Y**) experiments.

3.2.2 Combining NMR experiments and MD simulations

2D-NMR techniques are advantageous to MD simulations as they can deliver reference values either usable as restraints (eq. 3.16) or as target values for comparison to the MD ensemble.

A ^{1}H-^{1}H- COSY spectrum shows correlation peaks for protons that are covalently bound over three or less bonds, for example the two H^{α}-protons of glycine will have a correlation peak. Within the context of proteins or peptides, this is especially useful for detecting the backbone amides and measuring the associated coupling constants between the amide proton H^{N} and the H^{α}. The so-obtained coupling constants ${}^3J_{\mathrm{H^N H^\alpha}}$ serve as reference values for MD simulations due to the following relation first formulated by Karplus:[166,170]

$$ {}^3J_{\mathrm{H^N H^\alpha}}(\phi) = A \cdot \cos^2(\phi + \delta) + B \cdot \cos(\phi + \delta) + C. \tag{3.62} $$

A, B, C and δ are parameters derived by correlating observed ${}^3J_{\mathrm{H^N H^\alpha}}$-values with backbone angles measured in, for example, X-ray crystallography. Due to their severe impact on the results from eq. 3.62, these parameters have been studied in great detail.[171–175] Please note, the original formulation of the Karplus' equation is obtained for $\delta = 0$.

Solving eq. 3.62 for the $\phi(t)$ time series, which can directly be extracted from the MD trajectories, one can determine MD ensemble averages according to eq. 3.34 for all ${}^3J_{\mathrm{H^N H^\alpha}}$ constants, and compare them to the experimental ones. In practice, it turns out that the parameter-dependency of eq. 3.62 often leads to differences between MD and NMR despite overall good agreement between the two data sets. In that regard, distances derived from 2D-NOESY experiments are superior, which is why mainly 2D-NOESY experiments were considered in this thesis.

As apparent by its naming, a 2D-NOESY experiment is based on the Nuclear Overhauser Effect (NOE), a phenomenon of cross-relaxation between two magnetically, yet not covalently bound (NMR-active) nuclei within a spatial distance < 0.5 nm.[169] Based on this effect, the signal intensity, more precisely: the cross-peak volume in a 2D-NOESY experiment I^{NOE} can be used to determine the spatial distance d between the coupling nuclei:

$$ \langle I^{\mathrm{NOE}} \rangle_t = \frac{c}{d^6}, \tag{3.63} $$

where $\langle ... \rangle_t$ indicates the average over the measurement time t and c denotes the calibration constant. Often, it is possible to also extract an upper and lower limit for each distance (d_u, d_l) based on the error of the average intensity. Together with these limits, the distance values obtained by 2D-NOESY experiments can then be used to build a much more sophisticated

restraint potential (eq. 3.16) with the following definition ranges:[126]

$$v(d) = k_l(d - d_l)^2 \qquad \forall\ d < d_l, \tag{3.64}$$

$$v(d) = 0 \qquad \forall\ d_l \leq d \leq d_u, \tag{3.65}$$

$$v(d) = k_u(d - d_u)^2 \qquad \forall\ d_u < d. \tag{3.66}$$

If not used for introducing distance restraints in the force field of the MD simulations, the distance values determined by 2D-NOESY experiments may also serve as reference values for the comparison of the MD and NMR ensembles:

After determining the indices of the hydrogen atoms that ought to be in a certain spatial distance, the ensemble averages of the MD simulations D^{MD} are calculated as follows:[176,177]

$$D^{\mathrm{MD}}_{a,b} = \frac{1}{N_{\mathrm{rep}}} \sum_{i=1}^{N_{\mathrm{rep}}} \langle d^{-6}_{a,b} \rangle_i^{-1/6} \tag{3.67}$$

with $\langle ... \rangle_i$ as the time average of replica i for the distance $d_{a,b}$ between the protons a and b assuming a simulation setup that consists of multiple replicas in parallel (total number N_{rep}). Please note, eq. 3.67 holds for small molecules as suggested by Refs.[176, 177] while $\langle d^{-3}_{a,b} \rangle^{-1/3}$ might be more accurate for larger molecules. For the amatoxins, eq. 3.67 was used as stated.

The MD and NMR ensembles are then compared by calculating the "NOE distance violations", i.e. the deviation of the MD ensemble averages from the NMR reference values. Grouping the distances by type allows for conclusions on the structural agreement between the MD simulations and the structures of the molecules in solution. Usually, a distance is concluded 'violated' if the deviation is larger than the experimental uncertainty of 0.1 nm.

Apart from NOESY experiments, HSQC experiments (**H**eteronuclear **S**ingle **Q**uantum **C**orrleation) can be very useful for MD simulations. HSQC experiments, for biomolecules especially: ^{1}H-^{15}N-HSQC or a ^{1}H-^{13}C-HSQC, can show directly bound nuclei, e.g. amide-N and amide-$\mathrm{H^N}$. HSQC spectra are usually two-dimensional showing the chemical shift of one (NMR-active) nucleus on each axis respectively. For instance, ^{15}N-^{1}H-HSQC spectra can directly tell how many amides with attached proton are incorporated in the protein or peptide of interest as each amino acid, except proline and its derivatives, will show a single peak. In addition, ^{15}N-^{1}H-HSQC spectra have proven effective for the detection of binding sites in proteins.

By comparing the HSQC-spectrum of the protein-ligand complex ('holo'-state) with the spectrum of the single protein ('apo'-state), it is possible to observe chemical shift differences for the residues most affected by the binding process. This knowledge is very helpful to scrutinize the intermolecular interactions in the MD simulations of the holo-system. The

combination of ^{15}N-^{1}H-HSQC spectra and the analysis of MD simulations with respect to intermolecular hydrogen bonds is shown in the tWW project.

As a side note: If NOESY and HSQC experiments are combined, then we obtain 3D spectra with one proton axis, one heteronucleus axis and one NOE evolved proton axis. The heteronucleus axis is often displayed as layers of 2D NOESY-spectra.

3.3 Further computational tools

3.3.1 Clustering methods

Once the MD trajectories are generated, ensemble averages (eq. 3.34) are useful to get a general picture of the simulated system. However, especially for peptides or proteins, individual conformations and their relation to the entire conformational ensemble are of special interest. Depending on the research focus, it is crucial to know which stable conformations exist, how they can be differentiated and how they interconvert into each other. To identify these conformations or more general: to identify groups of data points in a given data set, different clustering algorithms have been developed over the last decades.[178–181] In this section, we will discuss selected clustering principles and algorithms in order to contextualise the clustering algorithm that has been used in this thesis. This section is not meant to provide a detailed overview of clustering in general.

Basically, all clustering methods aim for the same target: The data points of a particular data set shall be assorted into groups, in which the members of each group are more related to each other than to other points of the data set. What sounds simple at first, however, proves to be anything but trivial upon closer inspection. All clustering algorithms have to cope with a philosophical problem: What are the criteria according to which points are related to each other?

Comparing the different clustering algorithms available today, the relation between data points of the same cluster can be expressed in very different ways. Assuming a metric input data set, many clustering algorithms however refer to a certain type of pairwise distance to group data points into clusters.

The most straight-forward way is to define a distance between the data points, and to introduce a cut-off for these distances, below which two points belong to the same clusters. "Connectivity"-based clustering algorithms therefore assume that related data points are at least as close as the cut-off. The idea of using the connectivity between data points is the cornerstone of hierarchical clustering, since the variation of the distance cut-off immediately gives a hierarchy of cut-off dependent groupings. An example is classic "single-linkage-" clustering,[182] in which we use the Euklidean distance as central metric. We start at the bottom level where none of the data points is assigned to any cluster. Then, slowly increasing the distance cut-off, always the two closest data points are joined first into new groups or added to existing ones. By its design, single-linkage clustering has an important property: Although we only add points that are connected as defined by our distance cut-off, the single-linkage clustering does not guarantee that the cut-off holds for all pairwise distances in the cluster. Hence, two points of the same cluster, whose pairwise distance exceeds the distance cut-off,

are not expected to be directly related to each other, but indirectly via intermediate points.

Alternatively to the distances between the data points, their closeness to certain pre-defined points (prototypes) can serve as measure for the relation between data points. The most famous clustering algorithm based on this assignment principle is "k-means".[179,183,184] k-means assume that the input data space can be partitioned into k disjoint subsets with centroids $\boldsymbol{\mu}_i$. These centroids are the mean of the clusters, and each point of the data set is assigned to its closest centroid based on the Euclidean distance. One can imagine the clustering result as partitioning of the input data set into Voronoi cells. To find the optimal partitioning of the data set into k clusters, the algorithm searches for the k centroids that fulfil the following condition: Over all clusters, the sum of squared Euclidean distances between points and their centroids should be minimal. Please note, that for Euclidean distances to be computed, the data points must be n-dimensional vectors. We will see later how such vector data points can look like. Here, we face two very important drawbacks of centroid-based algorithms like k-means: First, the number of cluster centres is defined *a priori*, which is problematic if we do not know how many subsets we would expect. Defining too little or too many cluster centres, seriously impairs the significance of the resulting clusters. Second, centroids-based clustering methods like k-means strongly tend to give globular clusters independent of the shape of the data. Especially for separating intertwined or convex clusters k-means, is not suited.[185]

As third alternative, I would like to introduce the "density-based" clustering methods, one of which was used in this thesis. In these clustering methods, we assume that data points in regions of high data point density are related. Data points in sparser regions are usually considered noise which is not considered for ensuing analyses. Density-based clustering algorithms can still refer to the principle of connectivity between the data points of the same cluster. Two points that are assigned to the same cluster have to be "density-reachable", which means that they are not necessarily separated by a distance below a certain cut-off, but by high-enough density. Clusters are groups of points that are density-connected, which means that they are part of the same dense region of the data set. The cut-off is introduced for the density, for which the density-based clustering algorithms have different definitions: The density can be estimated either cell-wise in a grid-based approach (STING[186] or CLIQUE[187]) or by counting neighbours within a certain distance either point-wise (DBSCAN[181]) or pair-wise (Jarvis-Patrick,[180] CommonNN[185,188–190]). The clustering result highly depends on the density-estimation, and each algorithm has its own options to regulate the density-cut-off.

As a side note: for the design of a clustering method, it makes a huge difference whether the algorithm shall assign each data point to a set ("full-partitioning") so that all cluster C_i cover the input data set $\mathcal{D}$ ($\bigcup_i^{n_c} C_i = \mathcal{D}$) or whether it works with the assignment of noise points. Also, it is important to define whether any given data point is to be assigned to exclusively

one cluster ("non-overlapping") or can be member of different clusters ("overlapping").

Since there is no general definition for a "cluster" except the consensus that they contain a subset of the data set, the quality and significance of the clusters is still to be judged by the user of the clustering algorithm. In addition, especially if it comes to MD data, the clustering algorithm is not applied on the entire data set and subdata sets obtained by dimensionality reduction are considered for clustering instead. Apart from the definition problem, any clustering result therefore substantially depends on the informative value of the input data for the original data set.

In this thesis, the density-based, "common-nearest-neighbour" clustering (CommonNN) algorithm has been used.[185,188–190] In order to separate dense from sparse regions, the CommonNN algorithm works as follows: Let $\mathcal{D}$ be the input data set with n (vector) data points $\mathbf{x}_i$ and f features: $\mathcal{D} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and $\mathbf{x}_i = (x_{i,1}, ..., x_{i,f})^{\mathrm{T}}$. $\mathcal{D}$ contains discrete points of the feature space Ω', for which we assume that it is defined in real space ($\Omega' \subset \mathbb{R}^f$), so that we can compute all pairwise distances among the data points. In case of MD data for a system of N atoms, Ω' can be also be a sub-dimensional representation, e.g. in TICs (see 3.1.2), of the conformational space $\Omega \subset \mathbb{R}^{3N}$. Now assume two data points $\mathbf{x}_a = a$ and $\mathbf{x}_b = b$, one of which is assigned to a certain cluster $C_i \subset \mathcal{D}$. To check, whether b can also be assigned to C_i, the CommonNN algorithm uses two cut-offs as proxy for the density: a distance cut-off R and a member cut-off N_c. In three-dimensional space, R can be imagined as radius of a sphere around each point a and b, and all points that lie in these spheres are considered direct neighbours of a and b:

$$\mathcal{G}(a; R) = \{x \in \mathcal{D} \,|\, d(a, x) \leq R\}, \tag{3.68}$$

$$\mathcal{G}(b; R) = \{x \in \mathcal{D} \,|\, d(b, x) \leq R\}, \tag{3.69}$$

where $\mathcal{G}(a; R)$, $\mathcal{G}(b; R)$ denote sets of neighbours of a and b. Whether b can also be assigned to C_i, is then decided based on the intersection of these neighbour sets: $\mathcal{I} = \mathcal{G}(a; R) \cap \mathcal{G}(a; R)$. If $\mathcal{I}$ contains at least N_c data points, a and b belong to the same cluster. In order to revive the terms introduced before: a and b are both "density-reachable" and "density-connected". Imagine a third point $\mathbf{x}_c = c$, which is also part of C_i, but does not share N_c common neighbours with b, c and b are density-connected, but they are not directly density-reachable. In case the point b cannot be assigned to any existing cluster $C_i \subset \mathcal{D}$ and does not form an own one, it is classified as noise and will be removed from the data set for further considerations.

Working with the classification of noise points, makes the CommonNN algorithm very convenient for MD data. Since not all conformations formed by the system are indeed stable or representative, it is useful to remove them as outliers for a better insight on the dominant conformational dynamics. In the implementation of the CommonNN algorithm used in this

thesis (source: GitHub), there is a third parameter M that can be used to influence the sensitivity of the algorithm in the classification of noise. With M, the user can define a minimal number of data points, below which any cluster C_i will be assigned as noise.

In combination, the parameters R and N_c can be understood as density threshold or iso-density line, along which we cut the high-dimensional probability density and potentially observe a splitting into subsets of the data set. By its design, the CommonNN algorithm allows for hierarchical clustering, since we can easily screen different combinations of R and N_c and thereby increase or decrease that iso-density line. Monitoring the splitting caused by each threshold, we can then gain important insights on the relations between the clusters and possible energy barriers between them: The lower the density threshold, the higher is presumably the energy barrier between the two clusters.

At last, let us summarise the advantages of the CommonNN algorithm: In contrast to k-means, the CommonNN algorithm has proven independent of the input data in terms of size, shape or complexity. The CommonNN algorithm is able to cluster convex-shaped or spiral test data sets[185,190] as well as high-dimensional data from MD simulations such as on the C-type lectin receptor langerin[191] or on the tWW of h-FBP21 (see 4.1).[192] The CommonNN algorithm allows for hierarchical clustering very easily. As shown in Ref. [191] and in section 4.4, the clusters provided by density-based clustering with CommonNN are useful candidates for the definition of core sets in core-set Markov state models (see 3.1.3). Finally, the implementation of the CommonNN algorithm has been revised recently.[185,189] Since the CommonNN algorithm is now available as a user-friendly Python package,[190] it has the possibility to become a broadly-used tool for the clustering of all sorts of data.[193]

3.3.2 Docking

As elucidated in the introduction, PPIs are involved in many fundamental cellular functions, which is why they became a prime target for the development of possible drugs.[20,57,194] However, simply combining synthesised drugs (small molecules or peptides) and the targeted receptors by trial and error is not feasible due to the vast amount of possible combinations and modifications. To minimise the efforts for the synthesis and *in vitro* studies, Computer-Aided Drug Design (CADD) has emerged as important tool over the last decades. CADD can be applied for several purposes: On the one hand, it can be used to screen a database of compounds for their binding capacity to a well-known target, so that only the best-binding compounds may be tested *in vitro*. Alternatively, CADD can help to study the structure-activity relationship (SAR) by identifying the key components of the ligand-recognition in a specific receptor.[57] Either way, CADD can be classified as *ligand-based* or *structure-based* depending on how much is known about the target structure. In the ligand-based approach, we already know molecules that bind to a specific target. We use that knowledge to define a set of structural prerequisites for the binding, based on which we can design or search for other pos-

sible binding candidates. In the structure-based approach, we know the three-dimensional structure of the target, either provided by experimental data or homology modelling, and search for a well-fitting ligand.

Molecular docking is the generic term for computational techniques, which aim to propose binding poses with high affinity. They belong to the structure-based methods, since they necessitate the knowledge of the target structure. Essentially, molecular docking comprises three key components: (0) the representation of the system (structures, grids, etc.), (1) the sampling of possible conformations for the ligand in the proximity of the target, and (2) evaluation of the binding poses by scoring functions.[57,195]

Over the years, many docking algorithms have been developed.[20,194]To understand the differences between the docking algorithms, we have to first review the challenges one encounters during a docking experiment. Here, we will discuss three challenges: (1) the degrees of freedom in the receptor and the ligand, (2) the definition of the binding pose and (3) the scoring function as quality measure for the binding poses.

Flexibility. Molecular docking techniques mimic the complex formation, but as apparent from the introduction (see 1.3.2), there are several theories to explain the binding event between a receptor and a ligand: The "lock-and-key", "induced-fit" and "conformational selection" mechanisms mainly differ in the flexibility of the interacting partners. *Rigid* docking approaches follow the idea of "lock-and-key" by defining both the ligand and the receptor as rigid bodies. This is particularly advantageous if large molecules are considered or the binding site of the receptor is unknown. By treating the receptor and the ligand as rigid bodies one can reduce the degrees of freedom to 6 (three translational, three rotational) per interacting partner. This can make the calculation of binding poses computationally very cheap, which allows for exhaustive screening of docking poses. However, the rigid-docking highly depends on the quality of the input structures, since the binding poses are conformationally restricted to the input structures.

Semi-flexible docking approaches improve on that by allowing for flexibility in the ligand. These docking methods work under the assumption that the input structure of the receptor is already the one that is actually binding the ligand. Semi-flexible docking approaches are very useful if a complex structure of the receptor is known which can then be used as a homology model. For receptors with unknown complex structures, however, one should be careful, since the treatment of an entire protein as rigid body is not always verified.[57]

In *flexible* docking, both ligand and receptor are flexible. These docking methods can then either rest on the "induced-fit" or the "conformational selection" binding model.[57]

Finding the binding pose.[20,194] Apart from the treatment of the ligand and the receptors, molecular docking techniques can also differ in the search for the docking poses. Here, "template-based", "local" and "global" docking methods are distinguished. These categories mainly differ in the prediction accuracy, often correlated to the knowledge required prior to the docking experiment.[194] Template-based docking programs search for known complex structures (e.g. in the PDB), whose components share similarities to the receptor and the ligand in the complex of interest. Then, they use the known complex structure as template to build the unknown complex structure. Examples are "GalaxyPepDock",[196] "PepComposer"[197] and "HDock".[198]

In the local docking approach, the program searches for binding modes that involve a user-defined binding site on the receptor. The prediction accuracy of these docking methods therefore highly depends on the available information about the binding site. The local docking methods available today differ in their demands on the users: Programs like "Rosetta FlexPepDock"[199,200] or "PepCrawler"[201] need a modelled complex structure by the user, which will then be optimised during the docking process. Other programs like "HADDOCK",[202,203] which was also used within this thesis, only require a user-defined list of residues ("active residues") that will together be treated as binding site.

If nothing is known about the complex of interest *a priori*, a global docking approach might be advisable. Since they search for both the binding site and the binding pose, these docking programs are usually based on rigid-docking approaches since they optimise on screening as many conformations as possible. Examples are "ClusPro"[204] and "LZerD".[205]

Please note, that these categories are not strictly defined, and many docking programs combine different docking types in order to obtain the best results. The "HADDOCK" protocol used in this thesis (local docking), for instance, combines a rigid-body energy minimisation of the receptor with the ligand, followed by a semi-flexible refinement, in which the active residues (binding sites) are allowed to move, and a final refinement, a short simulation in explicit solvent.[202]

The scoring function[57,206]Scoring functions are sums of energy contributions of different origin. They should primarily be understood as an approximation for the binding affinity, which allows the molecular docking technique to discriminate good from bad binding poses.[57,202]Besides the overall workflow implemented in the docking protocols, scoring functions represent the centrepiece of the different docking techniques. Yet, in the definition and in the usage of scoring functions, the docking protocols differ substantially.

In principle, three types of scoring functions are distinguished:[57,206,207]
(1) Physics-based scoring functions assume that the ligand binding energy can be represented by a combination of potentials for the bonded and non-bonded interactions similar to

the force field of classical MD simulations (see eq. 3.3).

(2) Empirical scoring functions use select energy contributions extended by prefactors that are usually defined in the docking algorithm. In empirical scoring functions, the included energy contributions must not necessarily have a direct physical meaning. The scoring function of the "HADDOCK"-protocol (HS), for instance, is defined as follows:[208,209]

$$\mathrm{HS} = 1.0 \cdot E_{\mathrm{vdw}} + 0.1 \cdot E_{\mathrm{elec}} + 1.0 E_{\mathrm{desol}} + 0.1 E_{\mathrm{air}}, \tag{3.70}$$

where E_{vdw} and E_{elec} denote the van der Waals ("vdw") and Coulomb ("elec") interactions as described for the MD simulations. E_{desolv} is an empirical desolvation term,[209] and E_{air} denotes a soft-square harmonic restraining potential for "ambiguous interaction restraints" (AIR), i.e. the effective distances between the "active" and "passive" residues defined by the user:[202]

$$d_{i,A,B}^{\mathrm{eff}} = \left(\sum_{k_{i,A}=1}^{N_A} \sum_{l_B=1}^{N_{\mathrm{res},B}} \sum_{m_{l_B}=1}^{N_{l_B}} \frac{1}{d_{k_{i,A},m_{l_B}}^6} \right)^{-\frac{1}{6}}. \tag{3.71}$$

Eq. 3.71 is to read as follows: For all atoms N_A of an active residue i in molecule A, the distances to all atoms N_{l_B} of any active or passive residue l_B of molecule B are computed. By its definition, eq. 3.71 resembles the distance average for the NOE-derived inter-proton distances from NMR experiments (cf. eq. 3.63).

(3) There are knowledge-based scoring functions,[57,206,207] in which energy contributions are defined over a statistical analysis of a database of structures such as the PDB. The basic idea in knowledge-based scoring is that a high incidence of a specific atomic protein-ligand distance is directly correlated to favourable interactions. For the structures in the database, the occurring (pairwise) interatomic distances are analysed as distance-dependent number densities $\rho_{ij}(r)$. Then, for each pairwise distance, the "potential of mean force" is derived:

$$w_{ij} = -k_{\mathrm{B}} T \ln \left(\frac{\rho_{ij}(r)}{\rho_{ij}^{\mathrm{ref}}} \right), \tag{3.72}$$

where k_{B} is the Boltzmann constant and T the absolute temperature. ρ_{ij}^{ref} denotes the number density of the pairwise distances ij in a reference state, in which the interatomic interactions are zero. The binding energy for a complex of molecules A and B is then computed by simply summing over the interatomic distances, which are actually present:

$$E_{\mathrm{bind}} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} w_{ij}, \tag{3.73}$$

where N_A and N_B denote the number of atoms in A and B. By its approach, the knowledge-based scoring represents an attractive combination of reduced computational costs despite

predictive accuracy in comparison to the physics-based or empirical scoring functions. The reason for that is that the energy components w_{ij} have to only be computed once and the docking poses for any complex AB are then evaluated based on sums of tabulated values. The major issue of knowledge-based scoring, however, is the definition of the reference state, based on which the energy components are computed (eq. 3.72).

Quite recently, it has been reported that machine-learning can also be used to rank docking poses. These "machine-learning-derived" scoring functions, however, will not be discussed, here, and for further details Ref. [206] is recommended.

At last, please note two important things: First, any docking algorithm might include several definitions for the scoring function. Eq. 3.71, for example, is the HADDOCK score for the final refinement, while the previous steps mentioned above are provided with slightly different combinations of prefactors and energy contributions.

Second, the scoring functions are sometimes accompanied by RMSD-based clustering approaches. In "HADDOCK", the final docking poses are grouped based on the HADDOCK score and the backbone RMSD of the interfaces, defined on the active residues.[202,208,210]

Now that we have discussed molecular docking methods in detail, it should be noticed that despite the differences among the docking protocols, the research community has established a quality control, up to which all docking protocols have to measure. This quality control is called "A Critical Assessment of PRedicted Interactions", or in short: "CAPRI experiment".[211,212] The CAPRI experiments are an ongoing series of rounds, in which researchers working on docking protocols are tasked with predicting protein-protein complexes provided by the reviewers. After collecting all docking proposals, the reviewers publish their experimental data, such as X-ray crystallography data, along with the ranked results from the participating research groups. As such, one can quickly determine the strength and weaknesses of the different docking protocols. When we conducted the docking experiments with HADDOCK protocol, it belonged to the best-performing docking methods.[212] Close contesters were "ClusPro",[204] "HDock",[198] "MDOCKPP",[213] and "LZerd",[205] some of which were also discussed in this section.

4 PUBLICATIONS

4.1 Paper A1

Title:

"Target Recognition in Tandem WW Domains: Complex Structures for Parallel and Antiparallel Ligand Orientation in h-FBP21 Tandem WW"

Biological processes with large structural reorganisations such as signal transduction or RNA splicing, are driven by highly specific, yet reversible PPIs.[28] Often, these PPIs are mediated by adaptor domains.[29,30] One example for these adaptor domains are WW domains, a family of protein modules with 34-40 amino acids.[29,32,33] WW domains consist of a triple-stranded β-sheet with a binding site called *XP-groove*, and selectively recognise proline-rich motifs (PRMs) in the target protein. Based on their selectivity towards consensus PRMs, WW domains can be classified into five groups.[29,38] The complex structures of single WW domains and PRMs have been investigated for many years.[29,37–39]

However, WW domains often occur in pairs, so-called *tandems*. *Tandem* WW domains (tWWs) comprise two WW domains that are connected by a peptide sequence of varying length and flexibility (linker). TWWs can bind PRMs with more specificity and higher affinity. These PRMs can either be linked to one proline-rich sequence or they belong to different ligands.[44,46,47]The elucidation of the complex structure and the binding mechanism of tWW-ligand complexes is challenging because these complexes exhibit a large variety of different structures and because the complexes are hard to crystallize.

In this project, we investigate the tWW of *human*-formin binding protein 21 (h-FBP21) that comprises two stably folded WW domains connected by a flexible linker with twelve amino acids. The tWW of h-FBP21 was identified to bind PRMs of the spliceosomal core Sm protein B/B'. In particular, we are interested in the interaction between h-FBP21 tWW and the amino acid sequence 213-231 of the Sm protein B/B', which we call 'SmB2 ligand'. The SmB2 ligand contains two successive PRMs. Experiments have previously shown that (1) it can be bound by the h-FBP21 tWW in two orientations and (2) the binding to the WW domains might be cooperative.[42] In addition, it is well documented that in principle, also a single WW domain can bind one PRM in two orientations.[29]Knowing the position of the SmB2 ligand relative to h-FBP21 tWW, we therefore cannot predict the binding orientations of the single WW domains.

From complex structures of other tWWs than h-FBP21 tWW, it became apparent that the complex structure is influenced by many factors such as the properties of the individual WW domains or the interdomain region (see 1.3). As consequence, it is impossible to derive the

h-FBP21 tWW-SmB2 ligand complex structure from existing complex structures of other tWWs. Up to now, there had not been any structure elucidation of possible complexes of h-FBP21 tWW. In this project, we therefore aim at the identification of possible structures for the h-FBP21 tWW in complex with the SmB2 ligand.

We used the following techniques:

- unbiased, classical MD simulations with explicit solvent for the isolated h-FBP21 tWW and for the h-FBP21 tWW in complex with the SmB2 ligand
- DSSP analysis[214,215] for the evaluation of the secondary structure of h-FBP21 tWW
- hydrogen bond analyses for the characterisation of intramolecular (intradomain, interdomain) and intermolecular (with ligand) interactions
- density-based clustering on a subdimensional representation of the MD data set of h-FBP21 tWW
- RMSD analysis to prove the structural uniqueness of the h-FBP21 tWW clusters
- docking experiments[202,210] with cluster structures of h-FBP21 tWW and the SmB2 ligand
- distance analyses for the evaluation of h-FBP21 tWW-SmB2 ligand complexes
- mutation studies by replacing arginine residues in the SmB2 ligand with alanine residues, to assess the complex stability

From the MD simulations for the isolated h-FBP21 tWW, we could derive characteristics that are in agreement with experimental evidence: In water, the WW domains of h-FBP21 tWW remain fully-folded. They can adopt a vast variety of spatial arrangements due to the linker being unstructured and flexible. We introduced reaction coordinates that describe the orientation of the C-terminal WW domain (WW2) relative to the N-terminal one (WW1). In the space of these reaction coordinates, we clustered the h-FBP21 tWW-MD data. The clustering yielded 45 structurally unique conformations, for which we performed docking experiments with the SmB2 ligand. The resulting h-FBP21 tWW-SmB2 ligand complex structures substantially differed in the positioning of the PRMs on the surface of h-FBP21 tWW. We selected all complex structures, in which the PRMs of the SmB2 ligand and the XP-grooves of the h-FBP21 tWW were in proximity. To further evaluate these complex structures, we performed MD simulations and analysed them with respect to (1) intermolecular interactions and (2) the distances between the XP-grooves and the PRMs. Two of the final h-FBP21 tWW-SmB2 ligand complexes showed the PRMs stably anchored at the XP-grooves over the entire simulation time. The observed intermolecular interactions were in accordance with literature and our experimental data.[47] This makes the two modelled complex structures

promising candidates for the description of the h-FBP21 tWW-SmB2 ligand complex.

We characterised the two complex structures in great detail to uncover the structural arrangement of the tWW. As suggested by previous studies, the SmB2 ligand can be oriented in two ways relative to h-FBP21 tWW:[29,47] In one complex, the N-terminus of the SmB2 ligand is on the side of the N-terminal WW domain (parallel orientation), whereas in the second complex it is on the side of the C-terminal WW domain (antiparallel orientation). Our results could confirm, however, that the orientation of the SmB2 ligand towards tWW does not allow for assumptions on the binding orientation of the included single WW domains. Using criteria from literature,[29,47] we evaluated the orientation of the PRMs towards the single WW domains in each complex, respectively.

Each WW domain can bind one PRM in a parallel orientation ('p') or an antiparallel orientation ('a'). For our two h-FBP21 tWW-SmB2 ligand complex structures, we observed the following binding pattern (WW1, WW2): ('a', 'p') for the SmB2 ligand in parallel orientation relative to h-FBP21 tWW and ('p', 'a') for the SmB2 ligand in antiparallel orientation. The differences in the binding behaviours of the single WW domains have consequences for the respective tWW structure: In the complex with the SmB2 ligand in parallel orientation, the tWW is stabilised by a hydrogen bond pattern between both WW domains and the linker region. By contrast, in the complex with antiparallel ligand orientation, the tWW is predominantly stabilised by direct hydrophobic contacts between the two WW domains. This brings the WW domains closer together than in the parallel complex.

Taken together, we have found structurally distinct, novel tWW complex structures that exhibit different binding patterns. Both are in good agreement to experimental evidence, and they therefore represent a valid addition to the tWW complex structures that already exist.[44,45] Our h-FBP21 tWW-SmB2 ligand complexes are made publicly available on Zenodo (DOI: 10.5281/zenodo.5680225). Apart from the identification of possible tWW complex structures, we have developed a workflow, which might be applicable to other binding studies. This workflow was published alongside our results in *J. Chem. Inf. Model*.

The tWW project was a collaboration between the groups of Bettina G. Keller (B.G.K., Freie Universität Berlin) and Christian Freund (C.F., Freie Universität Berlin). All experimental evidence was generated by Miriam Bertazzon (M.B., AG Freund), Jana Sticht (J.S., AG Freund / Core Facility BioSupraMol) and Daniela Gjorgjevikj (AG Freund) including protein preparation, protein cloning, NMR spectroscopy and isothermal-titration calorimetry. Stevan Aleksić (AG Keller) performed the MD simulations for the isolated h-FBP21 tWW, and Marius T. Wenz (M.T.W.) analysed and processed them. M.T.W. defined reaction coordinates based on which he clustered the h-FBP21 tWW MD data set. For the clustering, M.T.W. used the "cnnclustering" algorithm, which was developed by B.G.K. and Oliver Lemke (AG Keller) and implemented by Jan-Oliver Kapp-Joswig (AG Keller). For the resulting cluster structures, M.T.W. carried

out the docking studies with a ligand structure provided by AG Freund. Out of the resulting complex structures, M.T.W. selected the most-promising ones, and performed further MD simulations as well as mutation studies. The manuscript for publication was written by B.G.K. and M.T.W. with M.B., J.S. and C.F. proofreading. All authors contributed to the final version of the manuscript.

The presented research was published in: Wenz, M. T. et al. *J. Chem. Inf. Model.* **2022**, DOI: 10.1021/acs.jcim.1c01426.

4.2 Paper B1

Title:

"Iodine-Mediated Tryptathionine Formation Facilitates the Synthesis of Amanitins"

Amatoxins are a family of cyclic peptides with a very special structure: They exhibit an eight-membered ring of amino acids with an additional linkage over the side chains of tryptophan and cystein, called *tryptathionine* bridge. Amatoxins are fungal toxins, which can be lethal already in small doses. α-amanitin is the most famous and one of the deadliest representatives of the amatoxin family. It is the toxin the death-cap mushroom (*Amanita phalloides*). Amatoxins inhibit the RNA polymerase II. This slows down the transcription process and ultimately causes cell death. The tryptathionine bridge organises the amatoxins in a very rigid shape, and that allows for high affinity to RNA polymerase II. Because of their capability to cause cell death, amatoxins might serve as therapeutics for cancer. Currently, they are examined as payload for antibody-drug conjugates.[66,67]

For drug design, it is important to understand the SAR of α-amanitin and its derivatives - both experimentally and computationally. However, the synthesis is extremely challenging. Over the years, there were many publications on the synthesis of amatoxins, yet only a few describe the successful total synthesis of α-amanitin.[68–71] One reason for this might be the complex character of the synthesis of the tryptathionine linkage. For most of the synthetic strategies known today, specific intermediates have to be formed, or dangerous oxidisers have to be used.[68,70] Together, these requirements limit the accessibility of amatoxins in larger amount and variety.

To improve on this, the group of Roderich D. Süssmuth (Technische Universität Berlin) investigated an alternative synthesis route, which was reported to be successful for the total synthesis of phalloidin and derivatives.[71,216,217] Considering linear and cyclic precursors, in which the tryptathionine bridge was not formed yet, they tried to understand the conditions under which the tryptophan-cysteine linkage can be formed in an iodine-mediated reaction. Although they found that the yields differed substantially among the precursors, they could not pinpoint the reason for it. In this project, we therefore scrutinised the structural dynamics of these precursors to draw conclusions on possible pre-organisations, which would favour or disfavour the formation of the tryptathionine linkage.

We used the following methods:

- unbiased, classical MD simulations in explicit solvent for different linear octa-peptides and the cyclic octapeptide (macrolactam), all lacking the tryptathionine bridge
- structural characterisation of the peptides by analysing intramolecular hydrogen bonds, interatomic distances and dihedral angles (backbone and side chain)

- RMSD analyses of the MD trajectories towards reference structure from NMR experiments
- calculation of inter-proton distances for comparison to NOE distance restraints from NMR data

We assumed that in order to facilitate the tryptathionine bridge formation, any precursor had to form a conformation, in which the cysteine and the tryptophan residues are in close vicinity. The linear precursors are more flexible than the macrolactam, but they sometimes form ring-like structures stabilised by internal hydrogen bonds. However, our results revealed that none of the peptides expresses a clear trend to pre-organise in solution. This is even true for the macrolactam. We also calculated different distances between tryptophan and cysteine, and checked the side chain angles of the two residues. In all precursors, the distances between cysteine and tryptophan are broadly distributed, and the side chain angles appear to be uncorrelated. Based on our MD data, we therefore cannot put the differences in the yields down to the pre-organisation of the precursors. Nonetheless, our MD data is in good agreement to the NMR data, which validates the parametrisation for the peptides. In our publication in *J. Am. Chem. Soc.*, we report a novel synthesis and an in-depth analysis of the reaction and the occurring precursors. Because the tryptathionine bridge can now be formed mediated by iodine, a variety of amatoxins are accessible via safer synthetic route. This is helpful to further studying the amatoxin family and to unravel its structure-activity relationship (SAR).

This project was a cooperation between the group of Roderich D. Süssmuth (R.D.S., Technische Universität Berlin), Heidelberg Pharma Research GmbH and the group of Bettina G. Keller (B.G.K., Freie Universität Berlin). All syntheses were carried out by Guiyang Yao (G.Y., Fudan University (China), formerly AG Süssmuth) and Caroline H. Knittel (C.H.K., AG Süssmuth). Bioassays were done by Hendrik Gruß, Alexandra C. Braun, Christian Lutz, Torsten Hechler and Andreas Pahl (Heidelberg Pharma GmbH). NMR spectra were generated and analysed by Simone Kosol (AG Süssmuth). Marius T. Wenz (M.T.W.) performed all MD simulations for the precursors of interest including their construction and parametrisation. M.T.W. analysed the MD simulations and compared the resulting trajectories to experimental data. The manuscript was mainly written by G.Y. and C.H.K with the remaining authors proofreading. All authors contributed to the final version of the manuscript.

The presented research was published in: Yao, G. et al. *J. Am. Chem. Soc.* **2021**, *143*, 14322–14331, DOI: `10.1021/jacs.1c06565`.

4.3 Paper B2

Title:

"The Occurrence of Ansamers in the Synthesis of Cyclic Peptides"

Amatoxins are a family of cyclic peptides with an exceptional structure. They possess an eight-membered peptide ring additionally bridged over a linkage between the side chains of the tryptophan and cysteine residues (tryptathionine bridge). Amatoxins are known as strong inhibitors for RNA polymerase II, causing cell death.[59,63–65] This property makes them very attractive for cancer therapy. Using amatoxins for the construction of antibody drug conjugates,[66,67] however, requires access to a large variety of different amatoxins. As that can only be achieved if the synthesis of amatoxins is fully understood, the investigation of the (total) synthesis of α-amanitin and derivatives has been pursued for many years now.[68–71]

Within this context, it has been discussed that during the synthesis of amatoxins, there is a chance to observe isomerism with respect to the positioning of the tryptathionine bridge relative to the peptide ring. Since the peptide backbone has a direction (C $\rightarrow$ N), the two sides of the macrolactam can be clearly distinguished. In natural amatoxins, the bridge is located exclusively on one side of the marcolactam, which we called "top" position. During the project, the Süssmuth group also synthesized the unnatural variants of the amatoxins in which the tryptathionine bridge is located on the other side of the macrolactam. The two products are formally conformational isomers.

For sterical reasons, a transition between the isomers by flipping the tryptathionine bridge through the peptide chain is impossible. However, we did not have any structural characterisation of the unnatural amatoxin, and we did not know under which reaction conditions they are formed. In addition, the isomers could not be unambiguously classified within the existing nomenclature. In our work, we therefore investigated the synthesis of α-amanitin analogues in order to unravel the occurrence of this isomerism. By studying the pre-conditions for the formation of these isomers, we were able to formulate a protocol for isomer-selective synthesis. Combining NMR spectroscopy, crystallography and MD simulations, we gained an in-depth view on the structures and dynamics of these isomers. Describing them in great detail, we propose the term *ansamers* for them. This terminology can be extended to other cyclic peptides than amatoxins.

In this project, we applied the following techniques:

- unbiased, classical MD simulations in explicit solvent for the natural amanitin derivative and its ansamer as well as for the precursors of interest
- design of a model to describe the relative position of the tryptathionine bridge towards both halves of the macrolactam

- structural characterisation of the modelled peptides involving the analysis of: hydrogen bonds, hydrogen bond sets and transitions between them; interatomic distances; dihedral angles (backbone and side chain); solvent-accessible surface area
- RMSD and RMSF calculations for structural comparison between the peptides
- quantum-chemical calculations to determine the ground state energies of the modelled peptides
- comparison of the MD ensemble with the NMR ensemble by calculating NOE distance violations

We analysed the α-amanitin analogue and its ansamer, and devised a model, which allowed us to monitor the relative positioning of the tryptathionine bridge relative to the macrolactam. In the distribution of the relative position of the tryptathionine linkage, both isomers are clearly distinguishable, and they show no overlap. With very good agreement to experimental evidence, our MD simulations therefore confirm that a transition between the ansamers is not possible. We structurally characterised the α-amanitin and its analogue, and found one dominant conformation respectively. These conformations contain different, but comparably-stable hydrogen bond sets resulting in structures that can be clearly distinguished in the space of the backbone dihedral angles (Ramachandran plots). The hydrogen bond sets as well as calculated solvent exposures of the amides are in great agreement to the crystal structures of the two peptides. In addition, we evaluated specific inter-proton distances and proved that our MD ensemble is in good agreement to our NMR data. We also investigated the ground state energies of the α-amanitin analogue and its ansamer. Our calculations show that the unnatural isomer has a lower ground-state energy than the natural isomer. Besides the ansamers, we simulated the precursors, which showed a clear selectivity for forming only one of the ansamers. We found that the precursors show structural similarities to the final products. This explains the selectivity of the reaction. For the ansamers and the precursors, the highest-probability structures and trajectories with a reduced time-resolution are made publicly available on Zenodo (DOI: 10.5281/zenodo.6974777). Alongside the structures, all information to redo the MD simulations including topology, starting and structure files are given (DOI: 10.5281/zenodo.7125315). Our results and structural data were published in *Nat.Commun.*

This project was a cooperation between the groups of Roderich D. Süssmuth (R.D.S., Technische Universität Berlin) and the group of Bettina G. Keller (B.G.K., Freie Universität Berlin). Guiyang Yao (G.Y., Fudan University (China), formerly AG Süssmuth), Simone Kosol (S.K., AG Süssmuth) and Marius T. Wenz (M.T.W.) are shared first author of the publication and contributed equally to this work. All syntheses were performed by G.Y. NMR experiments were performed and analysed by S.K. Oliver Trapp contributed to the nomenclature of the newly-defined isomers. M.T.W. performed all quantum-mechanical calculations and MD

simulations including the construction and parametrisation of the peptides. M.T.W. analysed the MD simulations with the techniques described above and described all peptides in great detail. For the α-amanitin derivatives, B.G.K. and M.T.W. developed a model to describe the position of the cyclic subunits relative to each other. This model is also applicable to other amatoxins. The manuscript was mainly written by G.Y., S.K., R.D.S., B.G.K. and M.T.W. All authors contributed to the final version of the manuscript.

On the following pages, the publication including supplementary information is provided with permission from "Yao, G. et al. *Nat. Commun.* **2022**, *13*, 6488, DOI: 10.1038/s41467-022-34125-8". Copyright 2022 Springer Nature.

nature communications

Article

https://doi.org/10.1038/s41467-022-34125-8

The occurrence of ansamers in the synthesis of cyclic peptides

Received: 20 May 2022

Accepted: 14 October 2022

Published online: 30 October 2022

Check for updates

Guiyang Yao[1,2,6], Simone Kosol[1,6], Marius T. Wenz[3,6], Elisabeth Irran[1], Bettina G. Keller[3], Oliver Trapp[4,5] & Roderich D. Süssmuth[1]

α-Amanitin is a bicyclic octapeptide composed of a macrolactam with a tryptathionine cross-link forming a handle. Previously, the occurrence of isomers of amanitin, termed atropisomers has been postulated. Although the total synthesis of α-amanitin has been accomplished this aspect still remains unsolved. We perform the synthesis of amanitin analogs, accompanied by in-depth spectroscopic, crystallographic and molecular dynamics studies. The data unambiguously confirms the synthesis of two amatoxin-type isomers, for which we propose the term ansamers. The natural structure of the *P*-ansamer can be ansa-selectively synthesized using an optimized synthetic strategy. We believe that the here described terminology does also have implications for many other peptide structures, e.g. norbornapeptides, lasso peptides, tryptorubins and others, and helps to unambiguously describe conformational isomerism of cyclic peptides.

The chemical synthesis of constrained peptide macrocycles of natural origin or of designed artificial peptides sometimes leads to the occurrence of isomers, which have been designated with various terms. There exist various literature reports: Wareham et al. describe the homeomorphic isomerism of macrobicyclic peptidic compounds which involved a passage of the bridge chain through the macrolactam (Fig. 1a)[1]. Bartoloni et al. investigated the diastereomeric norbornapeptides as potential drug scaffolds which showed bridge-up/down orientations according to the NMR solution structure (Fig. 1b)[2]. The Yudin group reported an unusual tunable atropisomeric peptidyl macrocycle which is made possible by controlling the conformational interconversion[3]. More recently, Baran and co-workers accomplished the reported total synthesis of the peptidic indole alkaloid tryptorubin A and defined non-canonical atropisomers, a family of shape-defined molecules that are distinguished by bridge below/bridge above arrangements (Fig. 1c)[4]. However, the vernacular nomenclature incited some controversy and Crossley and co-workers suggested a composite phenomenon using polytope formalism which is the fundamental of akamptisomerism classification[5]. Ultimately, the existence of atropisomers has also been postulated for the peptide toxins phalloidins and amanitins[6–12].

Phallotoxins and amatoxins are two bicyclic peptide toxin families isolated from the death cap mushroom *Amanita phalloides*. They both belong to the ribosomally synthesized and post-translationally modified peptides (RiPPs) and display high toxicity with low lethal doses in vivo animal experiments. α-Amanitin **1**, a slow acting toxin (LD_{50} = 50 − 100 µg/kg), has been reported to be a selective inhibitor of RNA polymerase II[13,14]. Its bicyclic octapeptide structure contains a 6-hydroxy-tryptathionine-(*R*)-sulfoxide cross-link (Fig. 1d). With the macrolactam ring as an imaginary plane, in a typical presentation, the tryptathionine bridge is located as a handle above the macrolactam, as it can be derived from a previously published X-ray structure (Fig. 1e)[15].

Longstanding questions are whether so-called atropisomers indeed existed and if so, under which circumstances they would occur, and what type of isomerism this would be? Previous studies of surrogate molecules of phalloidin and amanitin reported NMR spectroscopic data accompanied by CD spectroscopic analysis[8] and molecular dynamics (MD) simulations[7]. However, epimerization of sidechain

[1]Institut für Chemie, Technische Universität Berlin, Strasse des 17. Juni 124, 10623 Berlin, Germany. [2]Center for Innovative Drug Discovery, Greater Bay Area Institute of Precision Medicine (Guangzhou), School of Life Sciences, Fudan University, Shanghai, PR China. [3]Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany. [4]Department of Chemistry and Pharmacy, Ludwig-Maximilians-University, Butenandtstr. 5-13, 81377 Munich, Germany. [5]Max-Planck-Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany. [6]These authors contributed equally: Guiyang Yao, Simone Kosol, Marius T. Wenz. e-mail: suessmuth@chem.tu-berlin.de

https://doi.org/10.1038/s41467-022-34125-8

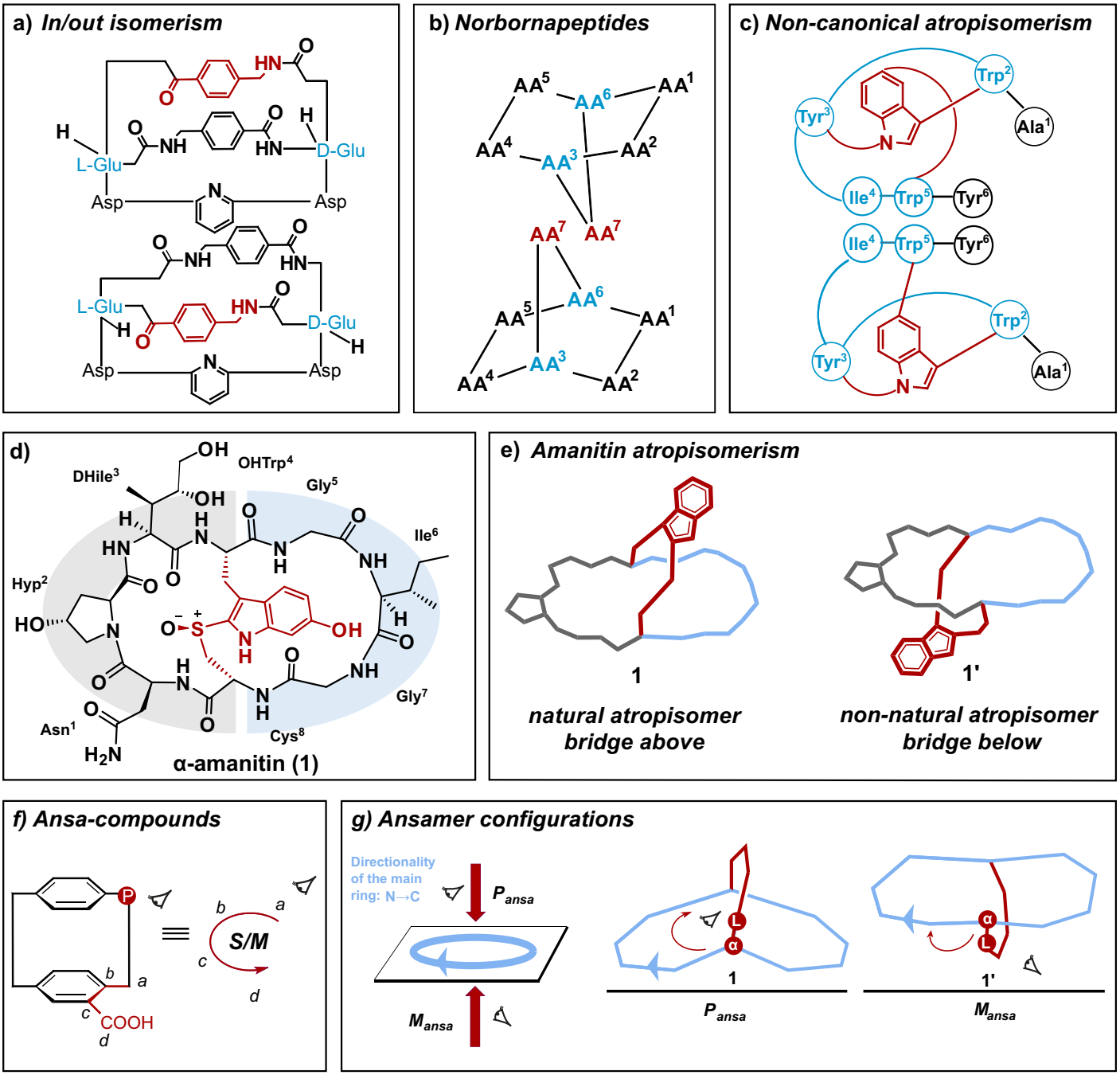


Fig. 1 | Peptide isomerism types and suggested determination of the ansamer configuration of cyclic amanitin-type systems. Several nomenclatures have been proposed to describe the isomerism of peptides displaying ring systems (**a**–**c**). **d** Structure of α-amanitin with the tryptathionine bridge highlighted in red. DHIle (2*S*,3*R*,4*R*-dihydroxyisoleucine), Hyp (4-*trans*-L-hydroxyproline), OHTrp (6-OH-tryptophan). **e** Proposed conformational isomers of α-amanitin. The A and B-rings are colored in gray and blue, respectively. **f** Determination of the configuration for ansa-compounds[24]. **g** Determination of ansamer-configurations: (1) the main cycle and the directionality is identified (light blue): from N- to C-terminus. (2) Bridge-up/bridge-down cases in view of the priority order in the main cycle. Identification of the leading atom/group L of the bridge next to the bridgehead atom α is assigned. (3) The descriptor P_{ansa} or M_{ansa} is assigned according to the directionality (clockwise/counter-clockwise) from the position of the leading atom/group L.

stereocenters as an alternative explanation could not be unambiguously ruled out[6].

Here, in a systematic approach combining various analytical methods (Marfey analytics, X-ray crystallography, NMR, and CD spectroscopy) with MD simulations, we determine the structure and dynamics of this sought isomer. Furthermore, we investigate different macrolactamization sites and show that an optimized strategy can ensure atroposelective synthesis. Finally, we propose the term ansamer to describe and unambiguously assign the configuration of stereoisomers of bridged cyclic systems, which can exist as configurational stereoisomers, depending on the position of the bridge, above or below the main ring. We suggest applying this terminology also to other conformationally restricted cyclic peptides, such as norbonapeptides or lasso peptides.

Results and discussions

Site-dependent macrolactamization and cycloisomer formation

Apart from semi-synthetic attempts, four total syntheses of α-amanitin have been reported to date. These contain three basic synthetic approaches to install the characteristic tryptathionine: An initial route employed the Savige-Fontana reaction via an Hpi (3a-hydroxy-pyrrolo[2,3-b] indole)[16] intermediate by the team from Heidelberg Pharma GmbH. This reaction was also used in a more sophisticated fashion by Perrin and co-workers to accomplish the first total synthesis of

https://doi.org/10.1038/s41467-022-34125-8

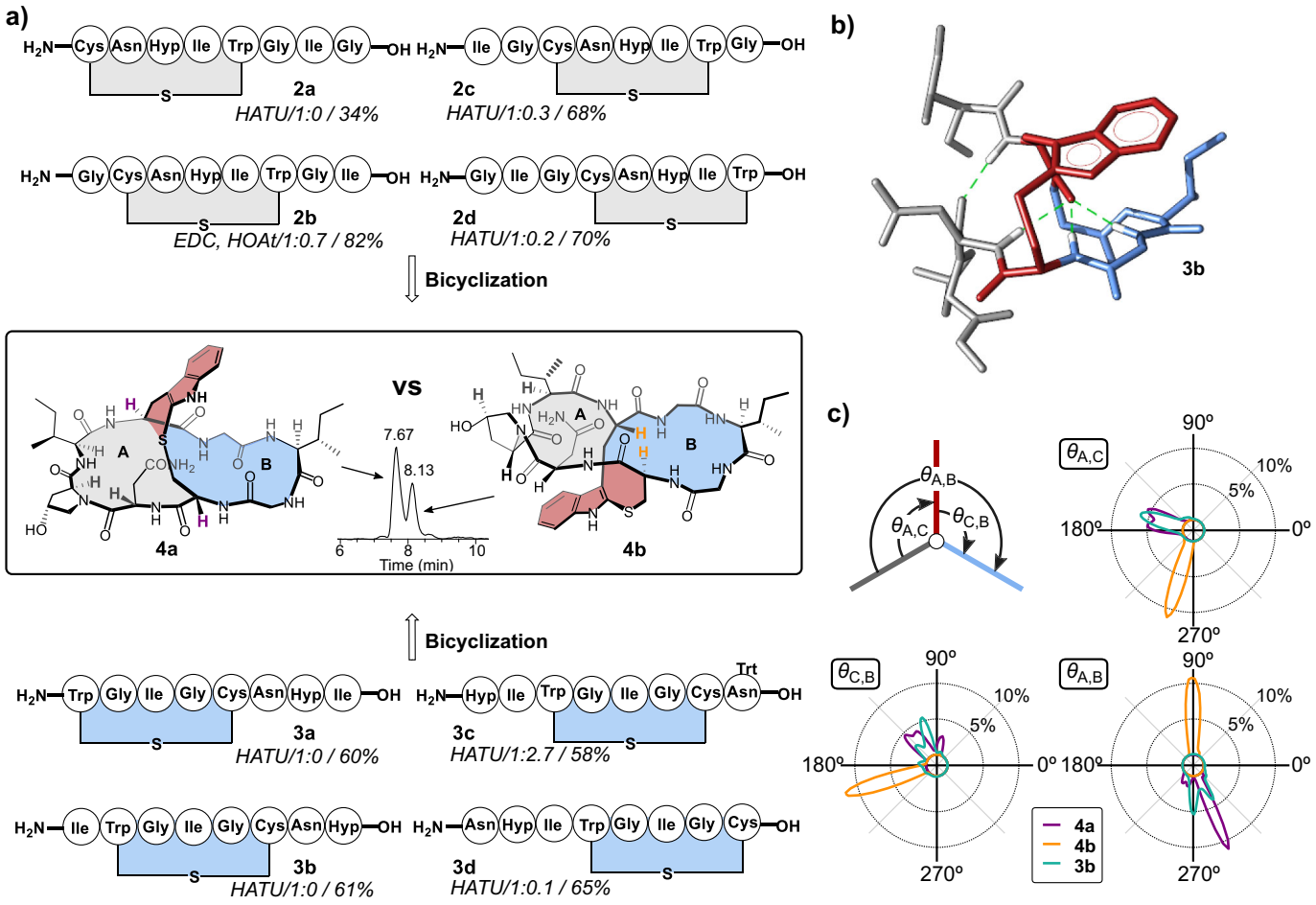


Fig. 2 | Peptide precursor choice impacts isomer yields. a Bicyclization conditions (HATU or EDC/HOAt) and different ring closure sites with preformed A-ring **2a-d** (gray) or B-ring **3a-d** (blue) lead to different isomer ratios (**4a**:**4b**) and yields (**4a**+**4b**). **b** MD-simulated structure of precursor **3b** with highest probability. Hydrogen bonds are shown as green dashed lines. The structure is colored according to the scheme in Fig. 1e. **c** Illustration of the model system to assess the relative orientation of the A-ring, B-ring, and tryptathionine bridge in the amanitin scaffold. The three defined planes (A-ring in gray, B-ring in blue and tryptathionine bridge in red, as in **a**, **b** are shown as lines, their relative orientation is defined by the angles θ as indicated (top left). For each angle ($\theta_{A,C}$, $\theta_{C,B}$, and $\theta_{A,B}$), the average distribution over 20 μs simulation data is presented in a circular plot. The distributions of angles observed in simulations of **4a** (violet), **4b** (orange), and **3b** (cyan) are shown as lines.

amanitin[17]. One approach from our lab was using a preformed tryptathionine from the reaction of indoles with sulfenyl chlorides, in a convergent [5 + 1 + 2] synthesis strategy[18]. Recently, we developed a robust and versatile iodine-mediated tryptathionine formation protocol that enabled us to synthesize various amanitin analogs for detailed SAR studies[19]. When we used the protocol to sample different macrolactamization sites (Fig. 2a), we noticed in some cases the formation of a by-product with an identical molecular mass as the desired amanitin analog. This could be interpreted as the formation of a diastereomer. A preliminary Marfey analysis[20] could however not prove epimerization, which led us to assume another isomer effect.

To establish a simplified model system to further investigate this observation, we decided to use readily available amino acids. Therefore, in the pursuit of the synthetic amatoxin we replaced DHIle[3] with Ile[3] to obtain Ile[3]-*S*-deoxo-amaninamide (Fig. 2). Thus, eight overlapping monocyclic peptides (**2a**–**2d** and **3a**–**3d**) were synthesized following our previous work[19] and the subsequent final macrocyclization was performed by using either HATU or EDC/HOAt as coupling agent, rendering eight corresponding bicyclic peptides. For the bicyclization of **2b**, the coupling reagent HATU was initially tested but significant amounts of the guanidination product were detected. To suppress the formation of the guanidination product, EDC/HOAt was selected as coupling reagent. Since the ratio of **4a** and **4b** is not changed much under different coupling conditions (see Supplementary information section 3.1.8.2 and Supplementary Fig. 19), we concluded that different bicyclization reagents and conditions do not significantly change the ansa-selectivity of the reaction. The crude peptides were carefully analyzed by HPLC-MS (see Supplementary Fig. 1). The synthesis route via monocyclic peptides containing ring A (**2a**–**2d**), repeatedly resulted in two peaks (**4a** and **4b**, R_t = 7.67 min and 8.13 min, respectively; Supplementary Fig. 1). Both peaks had the identical molecular mass of bicyclic Ile[3]-S-deoxo-amaninamide ($[M + H]^+$ = 855.3851 Da), albeit occurring at different ratios (1:0 to 1:0.7, see Fig. 2a). In contrast, for precursor peptides with monocycle B as intermediate (**3a**-**3d**) only one peak was observed, with the exception of **3c** which favored the formation of **4b** (ratio 1:2.7, Fig. 2a). When screening different macrolactamization sites, the ratio of isomer yields did not follow a clear trend. Precursors with preformed A-ring (**2a**-**2d**) tended to result more often in isomeric product mixtures, while precursors with preformed B-rings (**3a**-**3d**) mostly yielded the natural isomer. The NMR spectra of the two isomers formed by different macrolactamization strategies correspond to each other which also excluded an epimerization during bicyclization (see Supplementary Figs. 14 and 15).

In our established iodine-mediated tryptathionine formation protocol, we employed precursor **3b** which exclusively yields **4a** (Fig. 2a). To investigate if **3b** exhibits conformational pre-organization favoring **4a**, we performed classical MD simulations of the precursor as well as of **4a** and **4b**. Indeed, in the simulations, **3b** forms a stable hydrogen bond network (Fig. 2b, Supplementary Fig. 2b), which allows

https://doi.org/10.1038/s41467-022-34125-8

the molecule to adopt a conformation in which the C- and N-termini are spatially close and the tryptathionine bridge is located above the B-ring (Fig. 2b, Supplementary Fig. 3d). To compare the structural organization of **3b** with **4a** and **4b**, we defined three planes and the angles (θ) between them: one plane for ring A and B, respectively, and a third plane for the tryptathionine (Fig. 2a, c). We measured the angles in the MD simulations of **3b**, **4a** and **4b** to compare the orientation of the bridge relative to the macrocycle in the three molecules. Satisfyingly, in **3b**, the tryptathionine and the ring planes are positioned as in **4a** with angles between 90° and 180° (Fig. 2c, Supplementary Fig. 3d). This likely explains the selective reaction of monocyclic **3b** to bicyclic **4a**.

Analytical characterization of bicyclic amaninamide isomers

Since UV absorption of tryptathionines is highly distinctive, it has been previously employed to characterize tryptathionines[21]. Interestingly, the UV maximum absorption of isobaric **4a** and **4b** is slightly different with $\lambda = 289$ nm and $\lambda = 293$ nm, respectively (Fig. 3a). To clearly exclude epimerization during bicyclization, enantiomer analysis of the amino acids by Marfey's method[20] showed identical configuration for every amino acid in both isomers **4a** and **4b** (see Supplementary Fig. 4).

Further analysis revealed that the CD and NMR spectroscopic data of peptide **4a** are fully consistent with Ile[3]-S-deoxo-amaninamide as previously characterized[15]. In contrast, the CD and NMR spectra of **4b** are noticeably different (Fig. 3b, Supplementary Fig. 13) and suggest a different 3D structure. As reported previously for amanitin analogs[8], this could be consistent with the formation of conformational isomers. For compound **4a**, a positive Cotton effect was observed between $\lambda = 210$ nm and 230 nm (in accordance with reported CD data of the natural conformer). In contrast, the potential non-natural conformer **4b** shows a negative Cotton effect at these wavelengths (Fig. 3b).

EXSY NMR analysis indicates that **4a** and **4b** are not readily exchangeable conformers. Interconversion does not occur at elevated temperatures in DMSO (150 °C, 10 h; Supplementary Fig. 17), or water (CD spectroscopy: cycle 20 °C→90 °C→20 °C) which is consistent with VT-NMR measurements (see Supplementary Figs. 5 and 6). Interconversion of the two isomers would require the indole sidechain to pass through the macrolactam ring which appears sterically

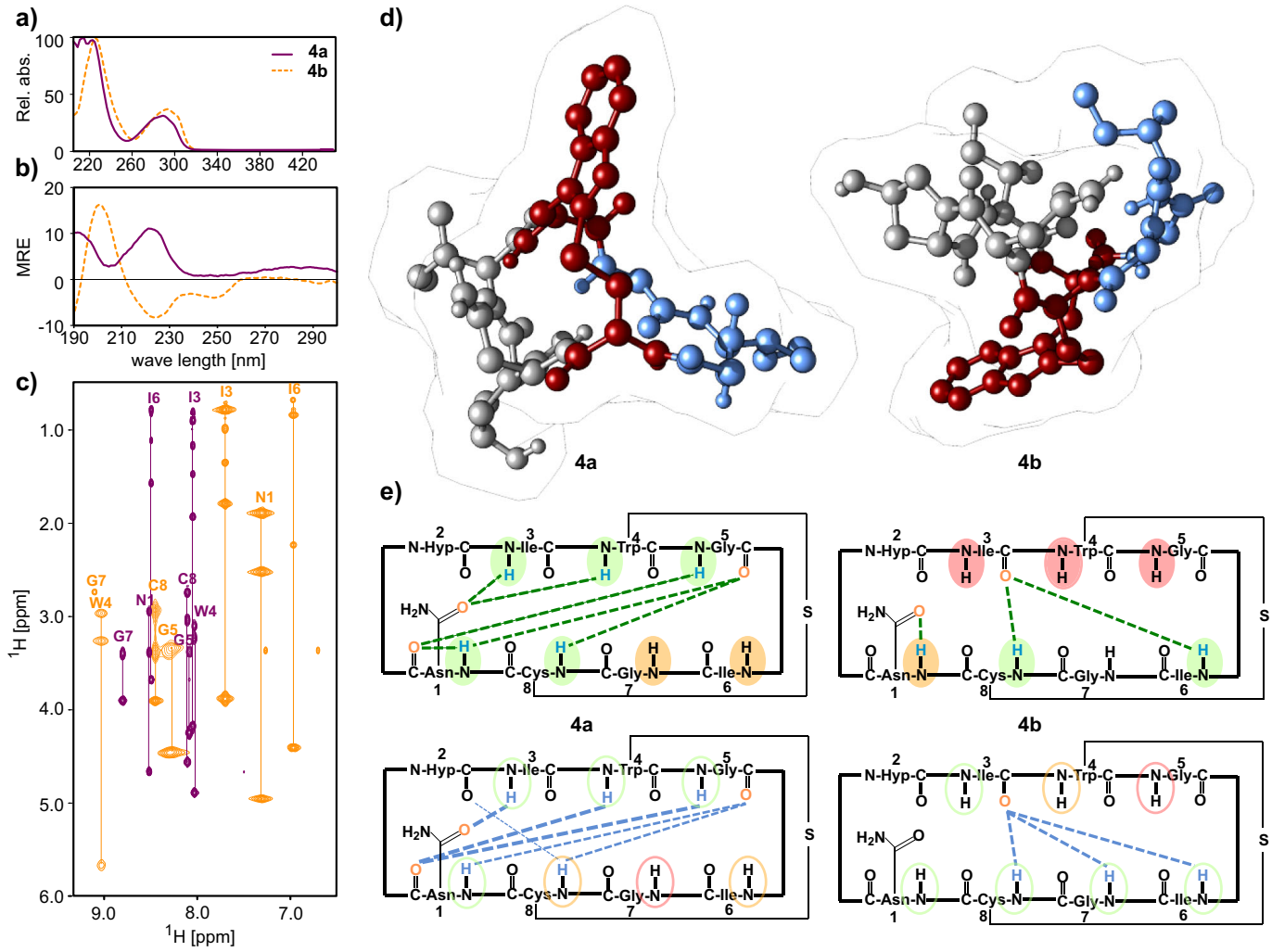


Fig. 3 | Physicochemical differences of amaninamide isomers. UV-absorbance spectra **a** show different maxima at $\lambda = 289$ nm and 293 nm for **4a** (purple) and **4b** (orange). **b** CD spectra of **4a** (purple) and **4b** (orange) show opposite Cotton effects at 225 nm. MRE = mean residue ellipticity in 10^3x [deg*cm^2*dmol^{-1}]. **c** Differences in chemical shifts of **4a** (purple) and **4b** (orange) in the amide region of the ^{1}H-^{1}H-TOCSY NMR spectra. **d** Crystal structures of **4a**[15] and **4b**. The A-ring and B-ring are colored gray and blue, respectively. The tryptathionine bridge is shown in red. **e** H-bonds found in the crystal structures of **4a** (left) and **4b** (right) are indicated by dashed green lines (top schematics). Amides with small chemical shift changes in VT-NMR measurements ($\Delta\delta$HN/ΔT > −3.0 ppbK-1) indicative of H bonding are highlighted in green, solvent-exposed amides with large shift changes ($\Delta\delta$HN/ΔT < −4.6 ppbK-1) are highlighted in red and amides with weak shielding/H-bonding are shaded in orange (top schematics). H-bonds observed in MD simulations are shown as blue dashed lines with the line width indicating the average population over 20 μs simulation time (bottom schematics; only H-bonds that occurred in >10% of the populations are shown, see Supplementary Fig. 12). In the MD simulations, amide groups highlighted with green ellipses showed small calculated solvent accessible surface areas (SASA < 0.02 nm^2), amides with large SASA (> 0.04 nm^2) are circled in red and amides with intermediate accessibility are circled in orange (exact values are listed in Supplementary Table 4).

impossible without the breakage and reformation of covalent bonds (Fig. 3). This is supported by our MD simulations where we also do not observe a transition between **4a** and **4b** (Fig. 2c), even at elevated temperatures.

To further characterize the relationship between isomers **4a** and **4b** both compounds were desulphurized with Raney Nickel at 80 °C to give the corresponding macrolactams (see Supporting Information). LC-MS analysis of the paralleled reactions indicated that the monocycles **5** (R_t = 5.88 min; $[M+H]^+$ = 825.4253 Da) are structurally identical (Supplementary Fig. 7: LC-MS after desulfurization). This was further confirmed by NMR spectroscopy of HPLC-purified **5** (Supplementary Fig. 8: calculated structure), suggesting that the tryptathionine bridge is the key factor of isomer formation.

Final proof for the isomeric nature of **4a**[15] (CCDC deposition number: 1128063) and **4b** was obtained from X-ray crystallography. Crystals of compound **4b** were grown in 10% EtOH aqueous solution and the structure of the peptide was obtained (CCDC: 2153904, Fig. 3d and Supplementary Fig. 9). In the crystal of **4b**, the tryptathionine bridge is clearly located below the plane of the macrolactam. The configuration of all amino acids in **4b** is identical with **4a**, corroborating our results from Marfey's analysis that no epimers were formed during macrocyclization. Remarkably, in the X-ray structure of **4b**, the trans-amide bond between Asn^1 and Hyp^2 is flipped to a cis-amide conformation with the carbonyl-group of Asn^1 facing towards the outside of the macrolactam instead of the inside (the trans character of the hydroxy-group of Hyp^2 is maintained). Overall, the backbone geometry and H-bonding pattern are drastically different from **4a**, giving the molecule a more compact appearance (Fig. 3e, Supplementary Fig. 16). This promotes the formation of a hydrophobic patch by Ile^3 and Ile^6 in **4b**, whereas in **4a** these are oriented in opposite directions. These conformational differences also substantially alter the physicochemical properties of the bicyclic peptides: hence the isomer **4b** is insoluble in water at 2 mM (Supplementary Fig. 10).

Besides the differences between crystal structures of **4a** and **4b**, which are also reflected in the angle populations shown in the Ramachandran plots (Supplementary Fig. 2a), the MD simulations also reveal differences in the dynamic behavior of the two isomers. We calculated the atomic RMSF for the simulated structures of **4a** and **4b** compared to their respective crystal structure (Fig. 3d, Supplementary Fig. 3). The RMSF values suggest that **4b** is generally more rigid than **4a** which is in line with the backbone angle distributions of **4a** being broader than of **4b** (Supplementary Figs. 2a and 3). Interestingly, quantum mechanical calculations (TDDFT level) on 100 optimized structures out of each MD data set showed that **4b** is also energetically favorable compared to **4a** (difference ~30 kJ/mol).

The MD simulations and the crystal structures are in very good agreement, as the all-atom RMSD is <0.4 nm and the backbone RMSD <0.2 nm (Supplementary Fig. 21). However, we noted some differences between the hydrogen bond patterns observed in the crystal structures and the simulated structures (Fig. 3e). For **4a**, in the majority of trajectories, the hydrogen bond pattern matches the crystal structure (Supplementary Fig. 11). Interestingly, in the MD simulations, we observed a subset of structures with a different set of hydrogen bonds compared to the crystal structure. Correlation analysis of the H-bonds in this subset and in the crystal structure (Supplementary Fig. 2b) shows that they are mutually exclusive, suggesting that a different minor conformation could exist which we did not observe in our experiments. Overall, the MD simulations suggest that the structure **4a** adopted in the crystal is likely the most stable conformation of this isomer (Supplementary Fig. 11, Supplementary Table 3), which is supported by the very good agreement in the NOE distances between the NMR and MD ensembles (Supplementary Fig. 13c).

For **4b**, a bifurcated backbone hydrogen bond is present in the crystal between the carbonyl oxygen of Ile^3 and amide hydrogens on the opposite site of the peptide ring (Ile^6 and Cys^8). MD simulations suggest a trifurcated H-bond instead, also involving the amide of Gly^7 (Fig. 3e, Supplementary Fig. 12). Unfortunately, it was not possible to determine if the amide proton of Gly^7 is shielded by a H-bond in solution as it was not resolved in the VT-NMR experiments. However, the calculated solvent-accessible surface areas of the amide groups in the crystal and MD structures agree well with each other (Fig. 3e, Supplementary Fig. 12d, Supplementary Table 4). As in **4a**, the NMR and the MD ensembles are in very good agreement with each other (Supplementary Fig. 13c).

Ansamers – a concept for assigning conformational isomers of cyclic peptides

Our structural and physicochemical characterization of the two isomers shows that the isomers **4a**/**4b** are not enantiomers, but diastereomers, lastly not only because of the pronounced differences in bond geometries, but also due to the chirality of amino acids. Previously, isomers arising from differing bond geometries at the bridgehead in bicyclic peptide systems have been categorized as atropisomers[22], non-canonical atropisomers[4] or akamptisomers[5]. This has led to some controversy and at this time there is no accurate and simple term to describe such a pair of isomers[23]. Atropisomers are clearly defined as stereoisomers, which are interconverted by rotation of a single bond between connecting moieties, which are typically sterically hindered. The herein described bridged cyclic peptide structures **4a** and **4b** are different from such atropisomers, because the interconversion of these stereoisomers is not just attributed to the rotation around a hindered single bond, but a flipping of the bridged cyclic structure. For that matter, the hindered bond-angle inversion that leads to akamptisomerization appears to be a better descriptor of the observed isomerism. But in case of **4a** and **4b** it is planes instead of bonds that undergo an angle inversion (Fig. 2c) and application of heat will not convert one diastereomer to the other. In addition, the flipping of the bridged cyclic structures leads to a conformational change in the cycle. As we find that **4a** and **4b** cannot readily interconvert even under heating we would assign the isomers the same molecular formula and same bond connectivities to configurational rather than conformational isomers. We therefore propose the term ansamer to stereochemically describe the two bridged isomers **4a** and **4b**. Compared to ansa-compounds, which consist of bridged planar chiral phenylene systems[24] (Fig. 1f), in the here presented ansamers, the bridged main cycle is structurally strained (ring strain or a clamp bridge), which increases the interconversion barrier between the two isomers. In contrast to atropisomers, ansa-compounds (lat. ansa = handle), e.g. cyclophanes, are interconverted by rotation of the handle around the planar phenylene. Similar to ansa-compounds[23] the assignment of the stereochemical descriptor can be made as follows in agreement with the CIP rules[25,26]: (1) identification of the main cycle[27–29] with the preferred directionality (Fig. 1g, from N to C terminus), (2) assignment of the leading atom/group of the bridge (Fig. 1g, leading atom/ group L) next to the bridgehead atom (α), followed by (3) assignment of the priority from the position of L: clockwise/counter-clockwise sense of the main cycle (Fig. 1g). (4) The descriptors P_{ansa} or M_{ansa} can be assigned accordingly (Fig. 1g). These assignment rules are unambiguous and correctly describe existing enantiomers, epimers, and diastereomers. This procedure would attribute conformational isomer **4a**, to the P_{ansa} and the non-natural isomer **4b** to the M_{ansa} isomer. Hence, the biosynthesis of α-amanitin **1**, is P_{ansa}-selective in establishing the tryptathionine bridge, as only the isomer with the indole above the ring has been found in peptides from the Amanita mushroom family. A member of the flavoprotein monooxygenase (FMO) family has been suggested to catalyze C-S bond formation[30], however, little is known about this step while the cleavage of the leader peptide and cyclization by a prolyl oligopeptidase (POPB) are well characterized[31,32]. The herein proposed P_{ansa}/M_{ansa} nomenclature could be also applied to previously described norbornapeptides by

Reymond et al.[2] and in an extended version, it may prove helpful to describe the conformation of in/out-isomers[1,33], the lasso peptides[34,35] and even tryptorubin[4] (see Supplementary Fig. 18).

Inspired by the terminology used for ansa-compounds, the term ansa (handle) illustrates the cause of the isomerism and at the same time reflects the planar aspect shared with benzene ansa-compounds. In this sense the nomenclature is consistent. We hope the intended terminological similarity between ansa compound for aromates and ansamer will raise attention and spur discussions in the scientific community about the underlying problem to accurately categorize this type of isomerism.

In summary, we have proven the occurrence of isomers in the synthesis of bicyclic amanitin analogs, which have been previously postulated and termed as atropisomers. The crucial step is the macrolactamization of monocyclic tryptathionine-containing peptides. The resulting isomers which appear fixed in differently bridged conformations, have been thoroughly characterized by spectroscopic and crystallographic methods as well as by MD simulations. For steric reasons, the indole sidechain of the tryptathionine cannot thread through the macrolactam ring. In a stereochemical description of these stereoisomers, we devise the term ansamers which can unambiguously describe these isomers and which can be applied also to various other cyclic peptides.

Methods

General procedure for the monocyclic peptide synthesis

2-CTC resin (1 g, 0.98 mmol/g) was pre-swollen for 20 min in DCM in a manual solid phase peptide synthesis vessel (10 mL). After the solvent was drained, the first amino acid Fmoc-AA^1-OH (0.3 mmol) and DIPEA (0.26 mL, 1.5 mmol) in DCM (5 mL) were added to the resin. The mixture was agitated for 2 h and before the solvent was drained. The resin was rinsed with DMF (4 × 3 mL). Then a mixture of MeOH/DIPEA/DCM (1:1:8) was added to cap the remaining 2-chlorotrityl chloride on the resin. The mixture was agitated for 0.5 h. Then the solvent was drained and the resin was washed with DMF (4 × 3 mL). The resin loading was determined to be 0.30 mmol/g. The Fmoc-group was removed with 20% piperidine in DMF solution. Fmoc-AA^2-OH (4 eq) was coupled to the deprotected resin according to TBTU mediated coupling. The Fmoc-group of the resulting resin was removed employing 20% piperidine in DMF. The following six amino acids were coupled to the deprotected. The tryptathionine formation was carried out on the solid support using I_2-mediated thioether formation. After removal of the Fmoc-group and followed cleavage from the resin, the monocyclic peptide was obtained following subsequent HPLC purification. The synthesis and characterization data of all compounds have been reported in the supplementary information.

General procedure for the bicyclic peptide synthesis

Monocyclic octapeptide **2a**–**2d** and **3a**–**3d** (1.0 eq) was dissolved in DMF (1 mM). Then, DIPEA (2.2 eq) and HATU (2.0 eq) were added at 0 °C. The reaction mixture was allowed to warm to r.t. for 12 h and concentrated under reduced pressure. The crude product was purified using preparative HPLC to afford bicyclic octapeptide as a white powder. Since large amounts of guanidinylation product were detected during macrolactamization of **2b**, the alternative coupling condition EDC (2 eq) and HOAt (2 eq) was employed to cyclize the monocyclic peptide **2b**. All results of LC-MS runs are shown in Supplementary Fig. 1. In addition, the detritylation was performed after macrolactamization of **3c**. The yield and ratio of **4a** and **4b** is shown in Fig. 2.

NMR assignment and structure calculation of desulfurized macrolactam

To obtain resonance assignments for NOE assignment and structure calculations **4a**, **4b**, and desulfurized macrolactam **5** were dissolved in deuterated DMSO-d_6 (~10 mM). TOCSY, COSY, NOESY, and ^{1}H-^{13}C-HSQC spectra were recorded on a Bruker Avance III 700 MHz spectrometer with a TXI 5 mm probe. Standard Bruker pulse programs were used and all spectra were acquired at 298 K. Residual solvent methyl peaks (DMSO-d7 $\delta = 2.502$ for ^{1}H and $\delta = 39.0$ ppm for ^{13}C) were used for chemical shift referencing. 2D homonuclear spectra were measured with acquisition times of 70 and 18 ms for the direct and indirect dimensions, respectively. TOCSY and NOESY spectra were accumulated with 16 or 32 (in case of **4b**) scans and COSY spectra with eight scans. The TOCSY and NOESY mixing times were set to 100 and 300 ms, respectively. Natural abundance ^{1}H-^{13}C-HSQC spectra were measured with 140 scans and acquisition times of 14 and 120 ms for the direct and indirect dimensions. The spectra were processed and analyzed using TopSpin 3.5 (Bruker) and CcpNmr 2.3.1[36]. After shift assignment (Supplementary Table 1), the NOE correlations of **4a** and **4b** were manually assigned and residue interaction matrices of **4a** and **4b** were generated using CcpNmr. For structure determination, the manually assigned chemical shifts of the desulfurized macrolactam and NOESY peak lists were supplied to CYANA for automated NOE assignment and structure calculation (Supplementary Table 2). The program CYLIB[37] was used to generate a CYANA library file for 4-hydroxyproline. A set of 1000 structures was calculated and the 100 best were visually inspected with UCSF Chimera[38].

Molecular dynamics simulations

All-atom classical MD simulations of peptides **4a**, **4b**, and **3b** in explicit dimethylformamide were carried out in GROMACS[39–41]. The force field parameters for the peptides were obtained with ACPYPE[42]. The simulations were conducted at the *NpT* ensemble with $p = 1$ bar and $T = 300$ K or $T = 400$ K. The simulation time was 20 µs for each system at $T = 300$ K, and 0.1 µs for each system at $T = 400$ K. See SI for a detailed protocol.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All processed data that support the findings of this study are available within the article and its Supplementary Information (experimental details; synthetic procedures; X-ray diffraction, NMR, UV/Vis, VT-NMR, MD simulations). All information to redo the MD simulations including topology, starting, and structure files are stored on Zenodo [DOI: 10.5281/zenodo.6974777] together with the highest-probability structures or simulated crystal structures of **4a**, **4b**, **3b**, and **3c** as well as downsampled trajectories with a visualization state for the open-source program VMD. The highest-probability structures for **4a**, **4b**, **3b**, and **3c** are provided as source data file alongside the manuscript. The X-ray crystallographic data for **4b** was deposited at the Cambridge Crystallographic Data Centre (CCDC) under deposition number CCDC 2153904 [DOI: 10.5517/ccdc.csd.cc2b99sr]. The open-source software used in this study is available under: MD simulations and analysis: GROMACS 2019.4 and GROMACS 2020.6 (https://manual.gromacs.org/documentation/), Custom code: Python 3.9.2 (https://www.python.org/downloads/), Jupyter (IPython 8.5.0, notebook version 6.4.12, https://jupyter.org); Visualization: VMD for MACOSXX86_64, version 1.9.4a57 (April 27, 2022, https://www.ks.uiuc.edu/Research/vmd/). MD data is stored together with custom code according to DFG regulations, and they are available from B.G.K. (bettina.keller@fu-berlin.de) upon request. Other data is available from the corresponding author. Source data are provided with this paper.

References

1. Wareham, R. S., Kilburn, J. D., Turner, D. L., Rees, N. H. & Holmes, D. S. Homeomorphic isomerism in a peptidic macrobicycle. *Angew. Chem. Int. Ed. Engl.* **34**, 2660–2662 (1996).

2. Bartoloni, M. et al. Bridged bicyclic peptides as potential drug scaffolds: synthesis, structure, protein binding and stability. *Chem. Sci.* **6**, 5473–5490 (2015).
3. Diaz, D. B. et al. Illuminating the dark conformational space of macrocycles using dominant rotors. *Nat. Chem.* **13**, 218–225 (2021).
4. Reisberg, S. H. et al. Total synthesis reveals atypical atropisomerism in a small-molecule natural product, tryptorubin A. *Science* **367**, 458–463 (2020).
5. Canfield, P. J. et al. A new fundamental type of conformational isomerism. *Nat. Chem.* **10**, 615–624 (2018).
6. May, J. P., Fournier, P., Patrick, B. O. & Perrin, D. M. Synthesis, characterisation, and in vitro evaluation of Pro2-Ile3-S-deoxo-amaninamide and Pro2-D-allo-Ile3-S-deoxo-amaninamide: implications for structure-activity relationships in amanitin conformation and toxicity. *Chemistry* **14**, 3410–3417 (2008).
7. Schmitt, W., Zanotti, G., Wieland, T. & Kessler, H. Conformation of different S -Deoxo-Xaa 3 -amaninamide analogues in DMSO solution as determined by NMR spectroscopy. Strong CD effects induced by βI, βII conformational change. *J. Am. Chem. Soc.* **118**, 4380–4387 (1996).
8. Zanotti, G., Petersen, G. & Wieland, T. Structure-toxicity relationships in the amatoxin series. *Int. J. Pept. Protein Res.* **40**, 551–558 (1992).
9. Anderson, M. O., Shelat, A. A. & Guy, R. K. A solid-phase approach to the phallotoxins: total synthesis of Ala7-phalloidin. *J. Org. Chem.* **70**, 4578–4584 (2005).
10. Falcigno, L. et al. Phalloidin synthetic analogues: structural requirements in the interaction with F-Actin. *Chemistry* **7**, 4665–4673 (2001).
11. Yao, G., Joswig, J.-O., Keller, B. G. & Süssmuth, R. D. Total synthesis of the death cap toxin phalloidin: atropoisomer selectivity explained by molecular-dynamics simulations. *Chemistry* **25**, 8030–8034 (2019).
12. Schuresko, L. A. & Lokey, R. S. A practical solid-phase synthesis of Glu7-phalloidin and entry into fluorescent F-actin-binding reagents. *Angew. Chem.* **46**, 3547–3549 (2007).
13. Brueckner, F. & Cramer, P. Structural basis of transcription inhibition by alpha-amanitin and implications for RNA polymerase II translocation. *Nat. Struct. Mol. Biol.* **15**, 811–818 (2008).
14. Lindell, T. J., Weinberg, F., Morris, P. W., Roeder, R. G. & Rutter, W. J. Specific inhibition of nuclear RNA polymerase II by alpha-amanitin. *Science* **170**, 447–449 (1970).
15. Shoham, G., Rees, D. C., Lipscomb, W. N., Zanotti, G. & Wieland, T. Crystal and molecular structure of S-deoxo[Ile3]amaninamide: a synthetic analog of Amanita toxins. *J. Am. Chem. Soc.* **106**, 4606–4615 (1984).
16. Lutz, C., Simon, W., Werner-Simon, S., Pahl, A. & Müller, C. Total synthesis of α- and β-amanitin. *Angew. Chem. Int. Ed. Engl.* **59**, 11390–11393 (2020).
17. Matinkhoo, K., Pryyma, A., Todorovic, M., Patrick, B. O. & Perrin, D. M. Synthesis of the death-cap mushroom toxin α-amanitin. *J. Am. Chem. Soc.* **140**, 6513–6517 (2018).
18. Siegert, M.-A. J., Knittel, C. H. & Süssmuth, R. D. A convergent total synthesis of the death cap toxin α-amanitin. *Angew. Chem. Int. Ed. Engl.* **59**, 5500–5504 (2020).
19. Yao, G. et al. Iodine-mediated tryptathionine formation facilitates the synthesis of amanitins. *J. Am. Chem. Soc.* **143**, 14322–14331 (2021).
20. Fujii, K., Ikai, Y., Oka, H., Suzuki, M. & Harada, K.-I. A nonempirical method using LC/MS for determination of the absolute configuration of constituent amino acids in a peptide: combination of Marfey's method with mass spectrometry and its practical application. *Anal. Chem.* **69**, 5146–5151 (1997).
21. May, J. P. & Perrin, D. M. Tryptathionine bridges in peptide synthesis. *Biopolymers* **88**, 714–724 (2007).
22. Moss, G. P. Basic terminology of stereochemistry (IUPAC Recommendations 1996). *Pure Appl. Chem.* **68**, 2193–2222 (1996).
23. Gold, V. *The IUPAC Compendium of Chemical Terminology* (International Union of Pure and Applied Chemistry (IUPAC), Research Triangle Park, NC, 2019).
24. Mannancherry, R., Devereux, M., Häussinger, D. & Mayor, M. Molecular ansa-basket: synthesis of inherently chiral all-carbon 12(1,6) pyrenophane. *J. Org. Chem.* **84**, 5271–5276 (2019).
25. Nicolaou, K. C., Boddy, C. N. C. & Siegel, J. S. Does CIP nomenclature adequately handle molecules with multiple stereoelements? A case study of vancomycin and cognates. *Angew. Chem.* **40**, 701–704 (2001).
26. Prelog, V. & Helmchen, G. Basic principles of the CIP-system and proposals for a revision. *Angew. Chem. Int. Ed. Engl.* **21**, 567–583 (1982).
27. Gerlach, H., Haas, G. & Prelog, V. Über zwei cycloisomere Triglycyltri-L-alanyle. *Helv. Chim. Acta*. **49**, 603–607 (1966).
28. Gerlach, H., Owtschinnikow, J. A. & Prelog, V. Cycloenantiomerie und Cyclodiastereomerie 2. Mitteilung. Über cycloenantiomere cyclo-Hexaalanyle und ein cycloenantiomeres cyclo-Diglycyltetraalanyl. *Helv. Chim. Acta*. **47**, 2294–2302 (1964).
29. Perlog, V. & Gerlach, H. Cycloenantiomerie und Cyclodiastereomerie. 1. Mitteilung. *Helv. Chim. Acta*. **47**, 2288–2294 (1964).
30. Walton, J. *The Cyclic Peptide Toxins of Amanita and Other Poisonous Mushrooms* (Springer International Publishing, Cham, 2018).
31. Luo, H. et al. Peptide macrocyclization catalyzed by a prolyl oligopeptidase involved in α-amanitin biosynthesis. *Chem. Biol.* **21**, 1610–1617 (2014).
32. Czekster, C. M., Ludewig, H., McMahon, S. A. & Naismith, J. H. Characterization of a dual function macrocyclase enables design and use of efficient macrocyclization substrates. *Nat. Commun.* **8**, 1045 (2017).
33. Alder, R. W. & East, S. P. In/Out Isomerism. *Chem. Rev.* **96**, 2097–2112 (1996).
34. Schröder, H. V., Zhang, Y. & Link, A. J. Dynamic covalent self-assembly of mechanically interlocked molecules solely made from peptides. *Nat. Chem.* **13**, 850–857 (2021).
35. Knappe, T. A. et al. Isolation and structural characterization of capistruin, a lasso peptide predicted from the genome sequence of Burkholderia thailandensis E264. *J. Am. Chem. Soc.* **130**, 11446–11454 (2008).
36. Vranken, W. F. et al. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696 (2005).
37. Yilmaz, E. M. & Güntert, P. NMR structure calculation for all small molecule ligands and non-standard residues from the PDB chemical component dictionary. *J. Biomol. NMR* **63**, 21–37 (2015).
38. Pettersen, E. F. et al. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
39. Abraham, M. J. et al. GROMACS: high-performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
40. Abraham, M. J., van der Spoel, D., Lindahl, E., Hess, B. & the GROMACS development team. GROMACS User Manual version 2018.8,
41. van der Spoel, D. et al. GROMACS: fast, flexible, and free. *J. Comput. Chem.* **26**, 1701–1718 (2005).
42. Sousa da Silva, A. W. & Vranken, W. F. ACPYPE - AnteChamber PYthon Parser interfacE. *BMC Res. notes* **5**, 367 (2012).

Acknowledgements

Dr. Guiyang Yao thanks the Alexander von Humboldt Foundation for a postdoctoral fellowship and the National Natural Science Foundation of China (82204189). The work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, RTG 2473

Bioactive Peptides to R.D.S., M.T.W., and B.G.K.). M.T.W. and B.G.K. thank the Paderborn Center for Parallel Computing PC2 and ZEDAT of FU Berlin for computing time. Molecular graphics and analyses were performed with UCSF Chimera, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from NIH P41-GM103311. The authors thank Manuel Gemander and Hengshan Wang for helpful discussions and critical suggestions.

Author contributions

G.Y, S.K., and R.D.S. designed the experiments. G.Y. performed the synthesis of all shown compounds. S.K. and G.Y. designed and supervised the CD, NMR, and X-ray studies. E.I. recorded the crystallographic data and solved the structure. O.T. made contributions to the isomery interpretation and to the ansamer nomenclature. S.K., M.T.W., and B.G.K. performed the theoretical calculations of the precursor, ansamers, and macrolactam. All authors wrote, read, discussed, and approved the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-34125-8.

Correspondence and requests for materials should be addressed to Roderich D. Süssmuth.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Supplementary Information

The occurrence of ansamers in the synthesis of cyclic peptides

Guiyang Yao[1,2,‡], Simone Kosol[1,‡] , Marius T. Wenz[3,‡] , Elisabeth Irran,[1] Bettina G. Keller[3], Oliver Trapp[4] and Roderich D. Süssmuth[1*]

1 Institut für Chemie, Technische Universität Berlin, Strasse des 17. Juni 124, 10623 Berlin, Germany.

2 Center for Innovative Drug Discovery, Greater Bay Area Institute of Precision Medicine (Guangzhou), School of Life Sciences, Fudan University, PR China.

3 Department of Biology, Chemistry, Pharmacy, Freie Universität Berlin, Arnimallee 22, 14195 Berlin, Germany.

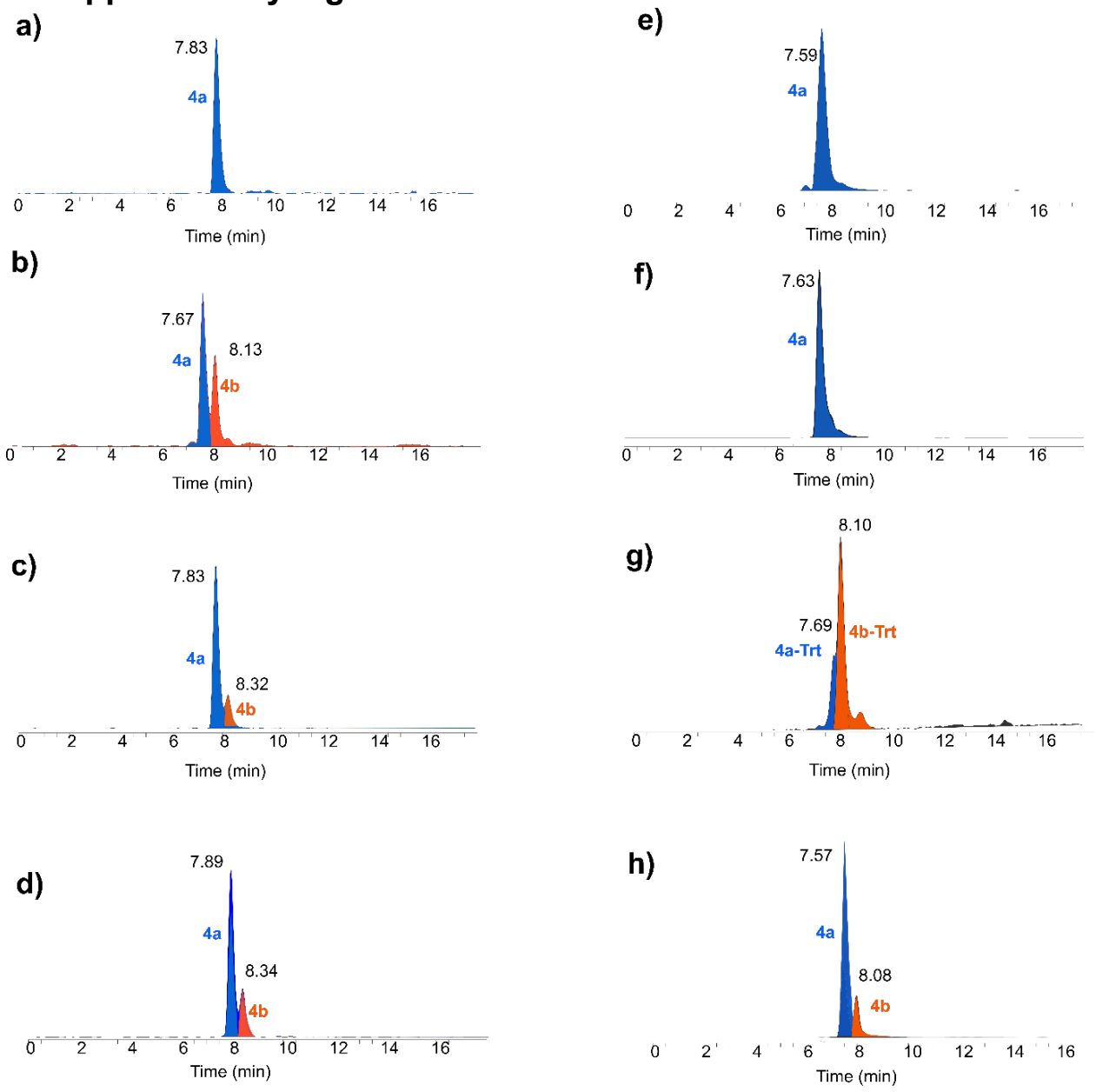
4 Department of Chemistry and Pharmacy, Ludwig-Maximilians-University, Butenandtstr. 5-13, 81377 Munich, Germany, Max-Planck-Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany.

‡ These authors contributed equally to this work.

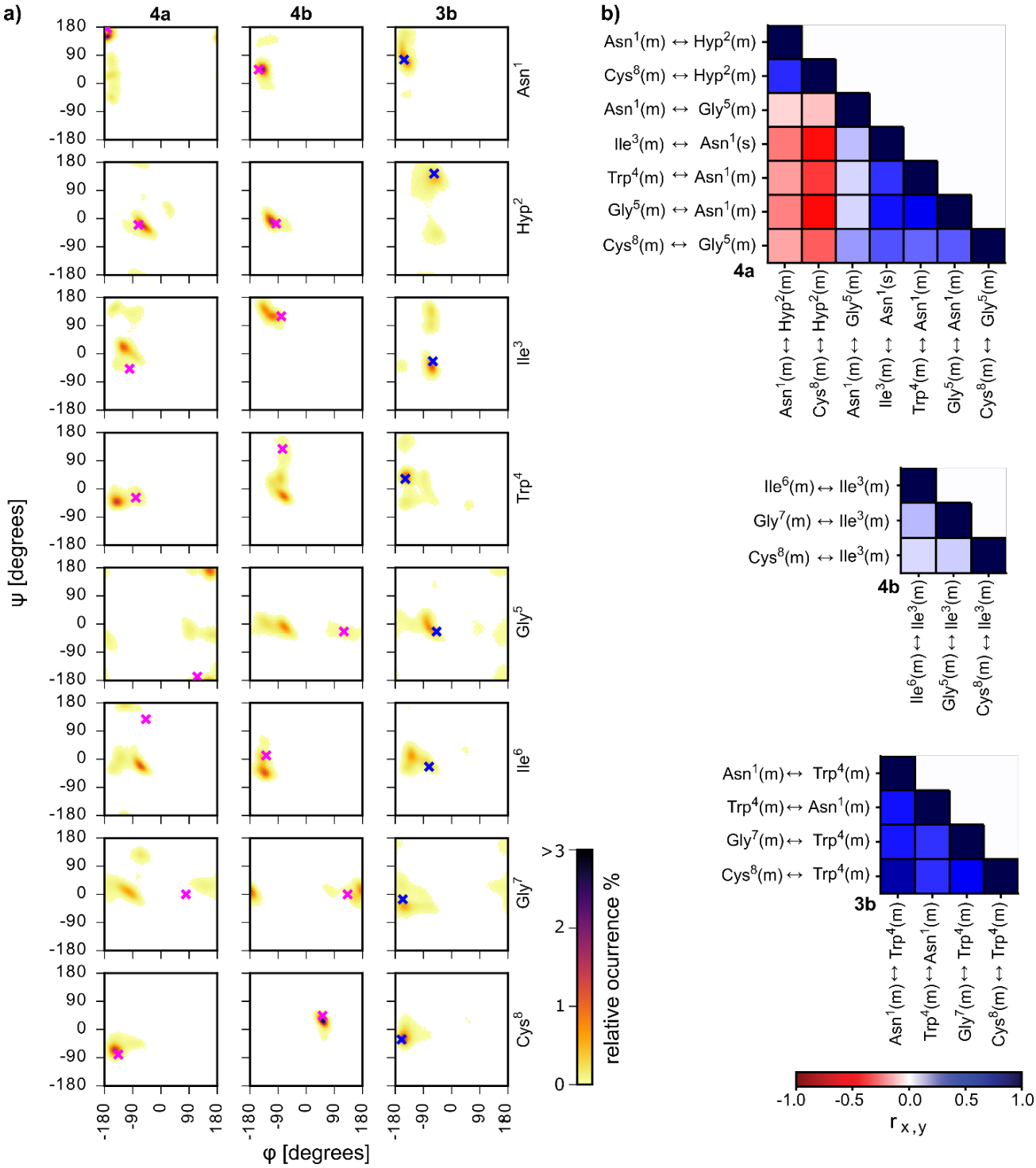
*e-mail: suessmuth@chem.tu-berlin.de

Contents

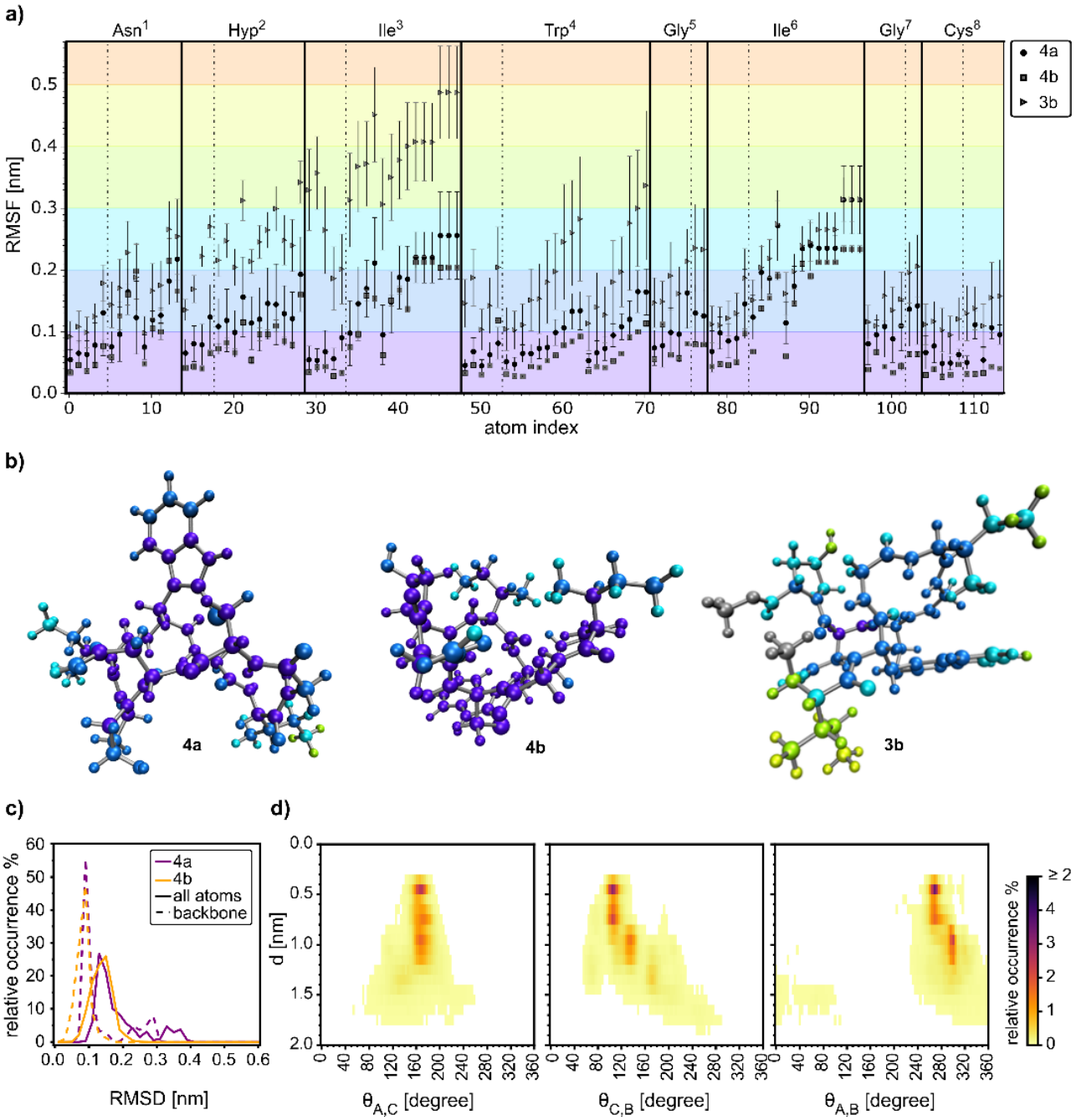
1. Supplementary Figures



Supplementary Figure 1: LC-MS of precursor peptides. LC-MS results of the macrolactamization of monocyclic precursors **2a-2d** and **3a-3d**. The gradient of a)-f) and h) is gradient D, the gradient of g) is gradient B.

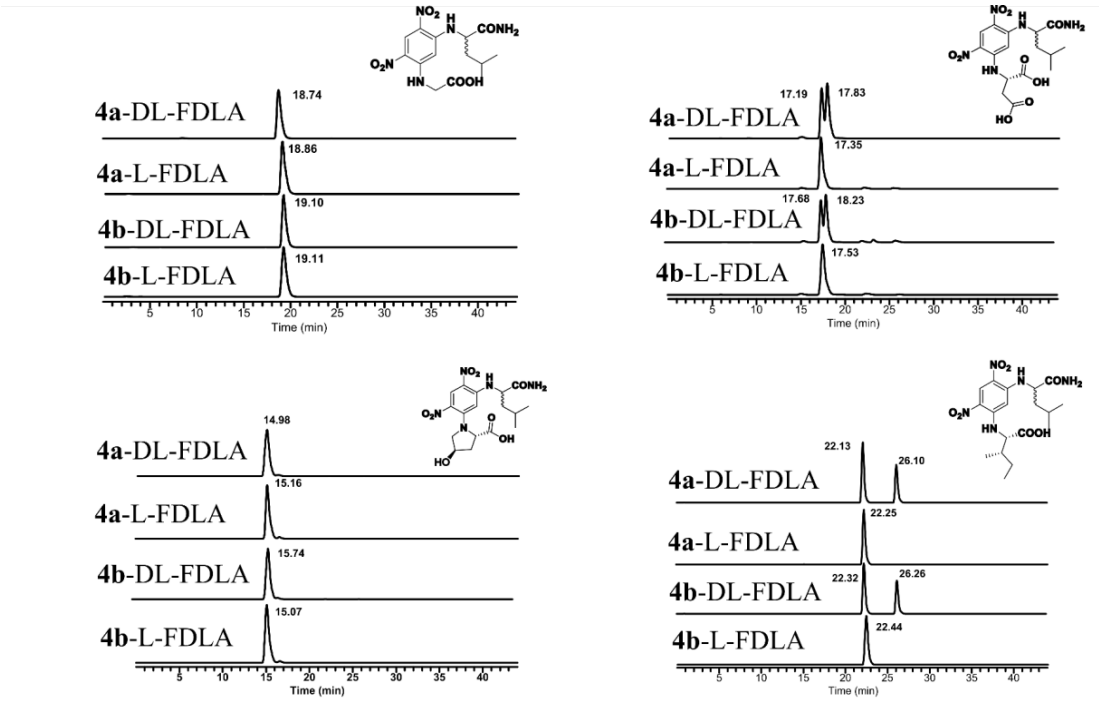


Supplementary Figure 2: Dihedral angle and hydrogen bond analyses of MD structural ensembles. a) Ramachandran plots for the molecules **4a**, **4b** and **3b**. Reference values taken from the crystal structures are indicated as pink crosses. Reference values taken from the highest-probability structure of the MD ensemble of **3b** are indicated as blue crosses (see 3.2.2.9). The backbone angle distributions are normalized referring to the simulation length (20 µs). b) Pearson correlation coefficients ($r_{x,y}$) for the hydrogen bonds with a population greater than 10% on average over all replica (n=20). Values close to |1| indicate a strong correlation (positive, blue) or anti-correlation (negative, red). 'm' denotes the main chain, 's' the side chain of the respective residue.

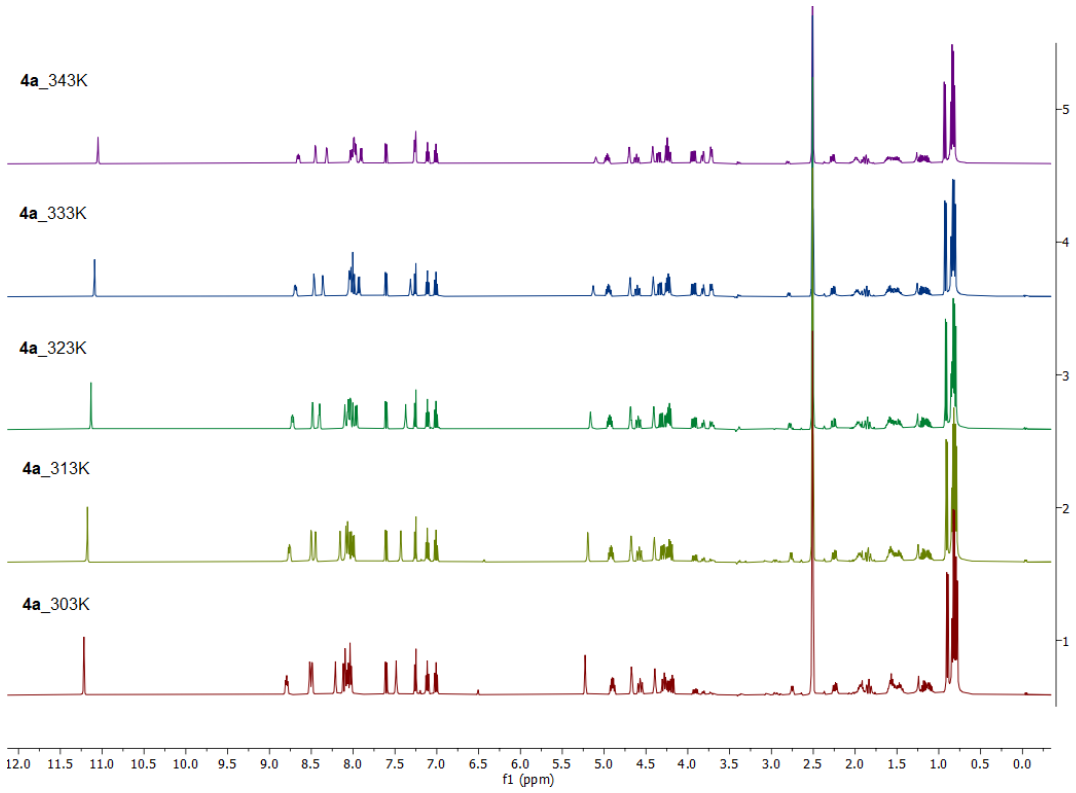


Supplementary Figure 3: Comparison of MD data with crystal structures. a) All-atom RMSF for **4a, 4b** and **3b** with reference to the respective crystal structure (**4a**, **4b**) or one structure of the subset containing the highest probability structure of the MD simulations (**3b**). In each system, the all-atom RMSF was calculated individually for 20 replicas (1 µs simulation time each). The resulting arithmetic mean (points) and standard deviation (error bars) over all replicas (n=20 representing 20 µs simulation time) are shown for each atom grouped according to the amino acid residues. The groups are separated by solid lines. The dotted lines separate 'main-chain' atoms (left) from 'side-chain' atoms (right), where 'main chain' is defined as 'N', 'H' (amide), 'CA', 'C' and 'O' for each residue following the GROMACS convention. For all molecules, the same atom order was applied: N, CA, C, O, for main chain atoms, followed by the carbon atoms along the side chain and hetero-atoms and lastly, the hydrogens. For the atom indices,

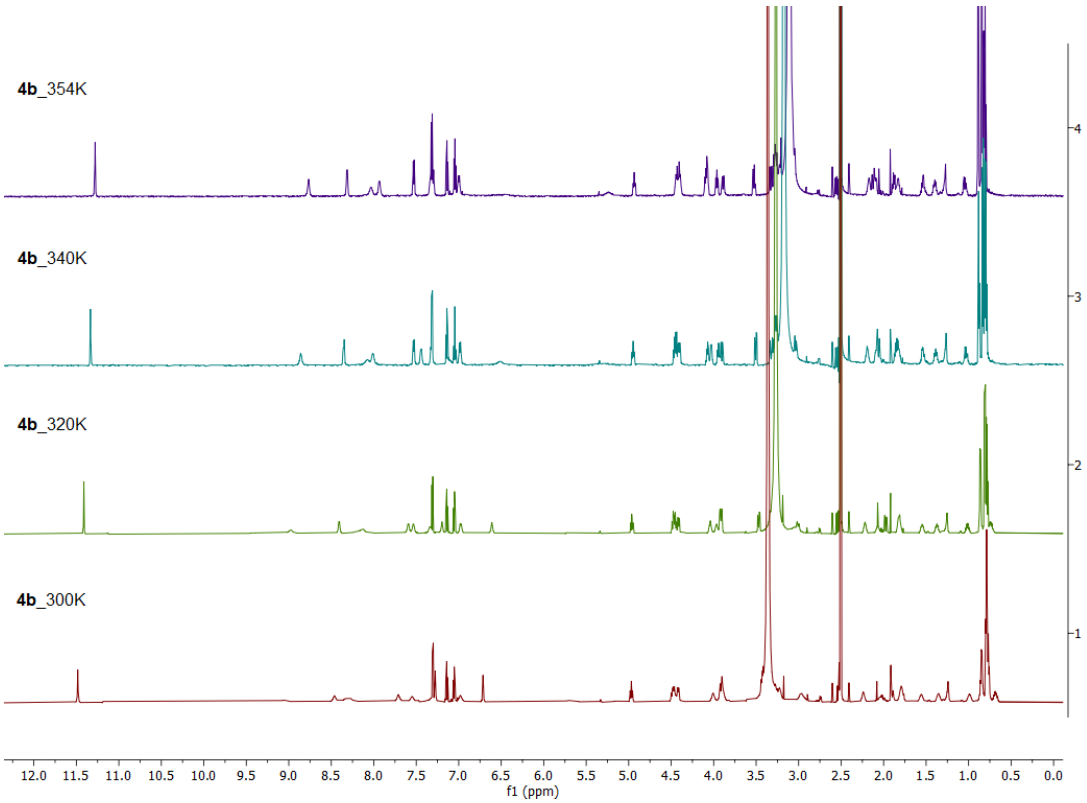
please refer to the methods section and supplied structure files ('RMSF calculations'). Caps were neglected for RMSF calculations. b) Structures of **4a**, **4b** and **3b**. The atoms are colored according to the respective RMSF values: purple: 0.0 - 0.1 nm, blue: 0.1 - 0.2 nm, cyan: 0.2 - 0.3 nm, green: 0.3 - 0.4 nm, yellow: 0.4 - 0.5 nm, orange: 0.5 - 0.6 nm. c) Distribution of the all-atom RMSD (solid line) and the backbone RMSD (dotted line) averaged over all replicas (n=20, 20 µs in total) for **4a** (purple) and **4b** (orange), respectively. d) Joint probability distribution normalized to the simulation length of 20 µs, for the distance between the C- and the N-terminus in **3b** and the angles between the planes $\theta_{A,B}$, $\theta_{A,C}$, $\theta_{C,B}$ defined in our model to assess the spatial orientation of ring A (Cys^8-Asn^1-Hyp^2-Ile^3-Trp^4), ring B (Trp^4-Gly^5-Ile^6-Gly^7-Cys^8) and the tryptathionine bridge. For the definition of the model, please refer to the methods section ('Plane angles'). The C-N-terminus distance was measured between the carbon atom of the carbonyl group in the C-terminus (-**C**O-OMe) and the nitrogen atom in the N-terminus (-**N**H-Me).



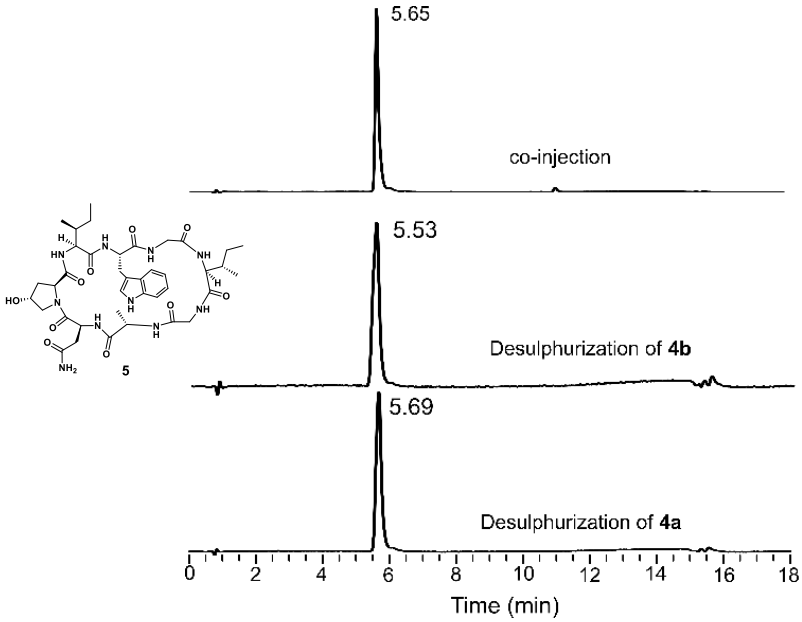
Supplementary Figure 4: Marfey analytics of amino acids in peptides 4a and 4b. HPLC-MS chromatograms of the amino acid analysis of peptides **4a** and **4b** (total hydrolysis followed by modification with Marfey's reagent, treated either with L-FDLA or the racemic DL-FDLA reagent (HPLC gradient E).



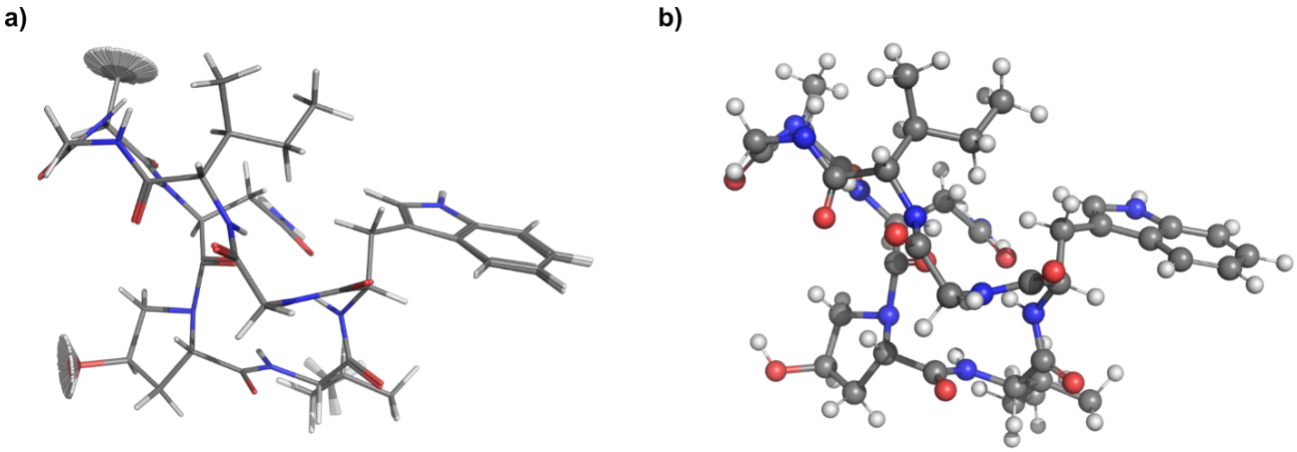
Supplementary Figure 5: VT-NMR of isomer 4a. Variable-temperature NMR spectroscopy study of peptide **4a** at temperatures ranging from 303K to 343K.



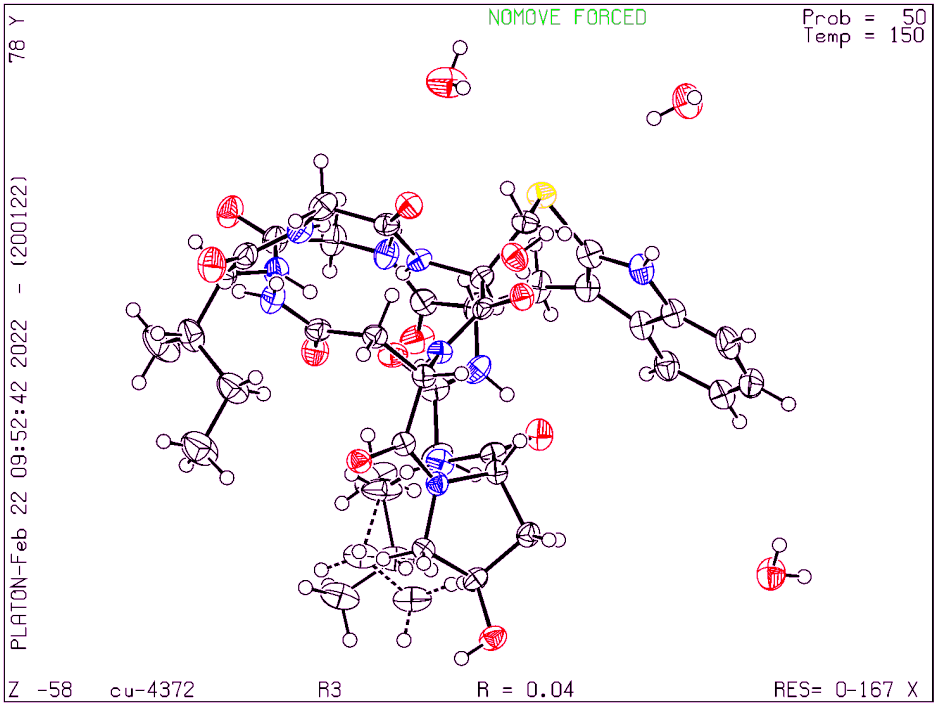
Supplementary Figure 6: VT-NMR of isomer 4b. Variable-temperature NMR spectroscopy study of peptide **4b** at temperatures ranging from 300K to 354K.



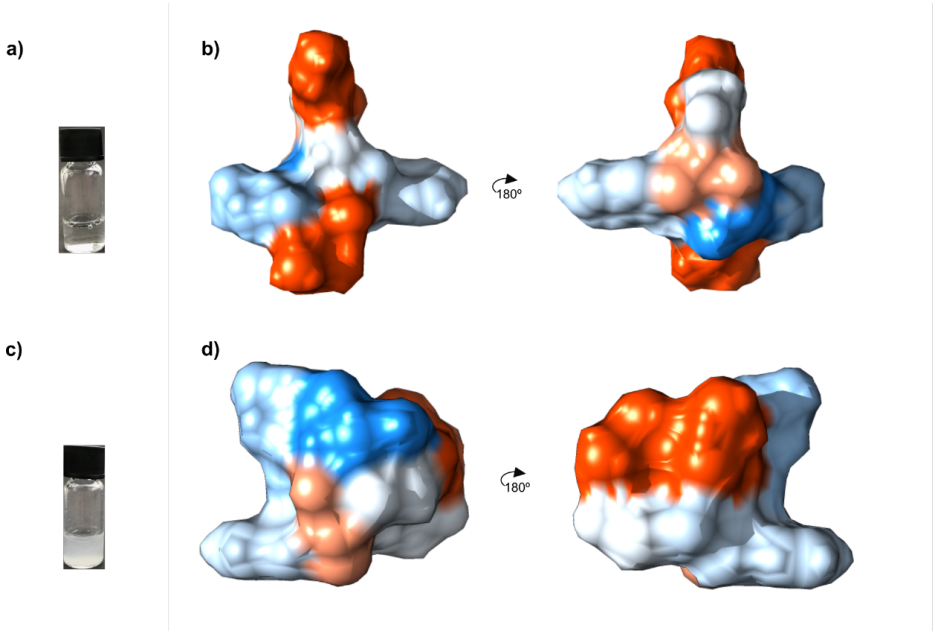
Supplementary Figure 7: Desulfurization of the isomers. LCMS chromatograms of the desulfurization reaction of bicyclic **4a** and **4b** (Raney-Ni in MeOH, 5 h) yielding compound **5**.



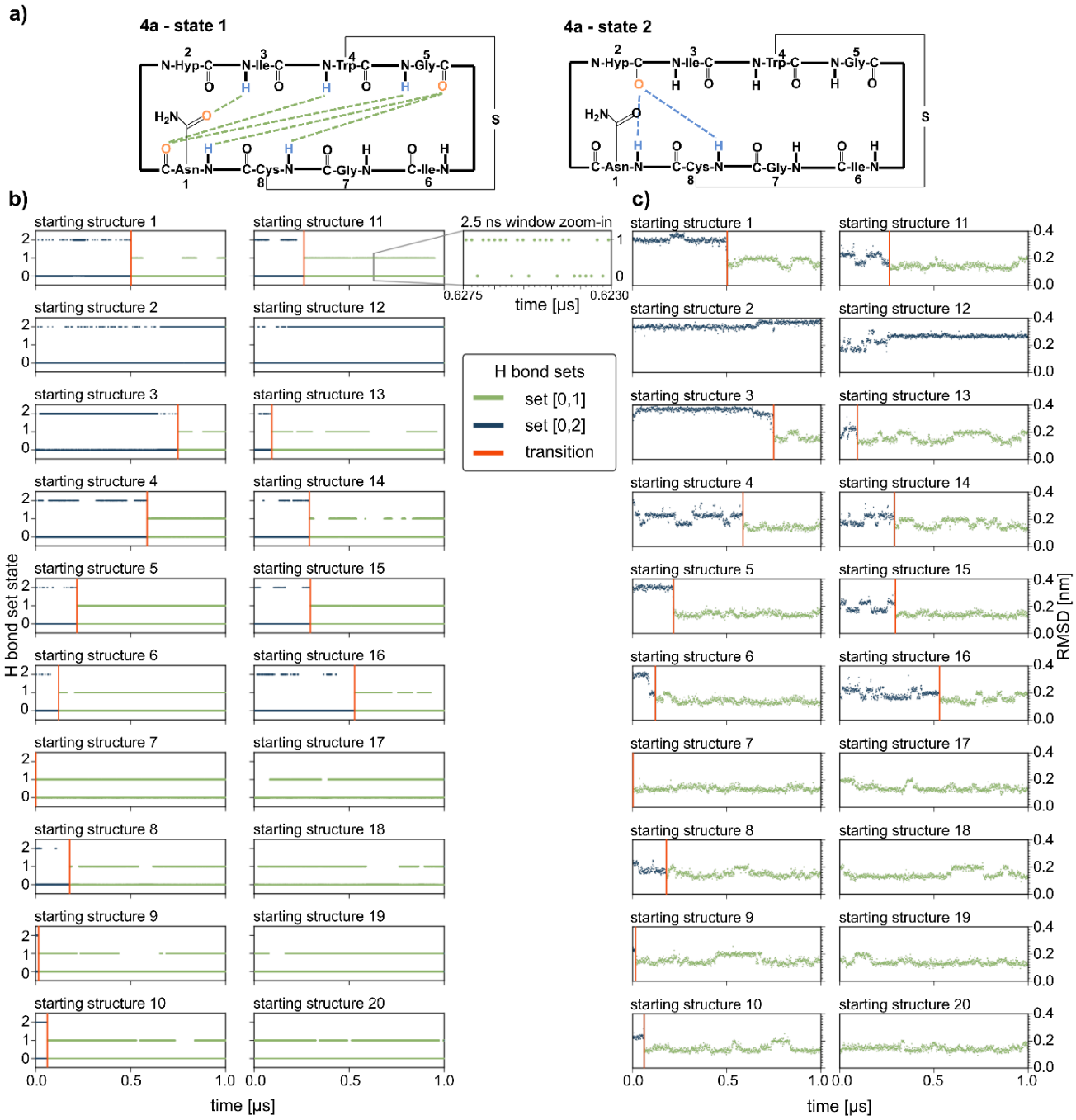
Supplementary Figure 8: Solution structure of desulfurized isomers. NMR-based solution structure of the desulfurized macrolactam **5**. A) The 100 lowest energy structures of desulfurized **4a** and **4b** have an RMSD of 0.01. B) Stick and ball model of the average state. The atoms are colored according to their type: grey: carbon, white: hydrogen, red: oxygen, blue: nitrogen.



Supplementary Figure 9: Crystal structure of 4b. Crystal structure of **4b** visualized by PLATON. The electron density map shows two conformations for the ethyl group of Ile[3] (the second conformation is indicated with dashed bonds). Nitrogen atoms are colored blue, oxygen atoms are shown in red, sulfur atoms are yellow, carbon atoms are depicted as white and black ellipsoids and hydrogens are shown as small white circles.

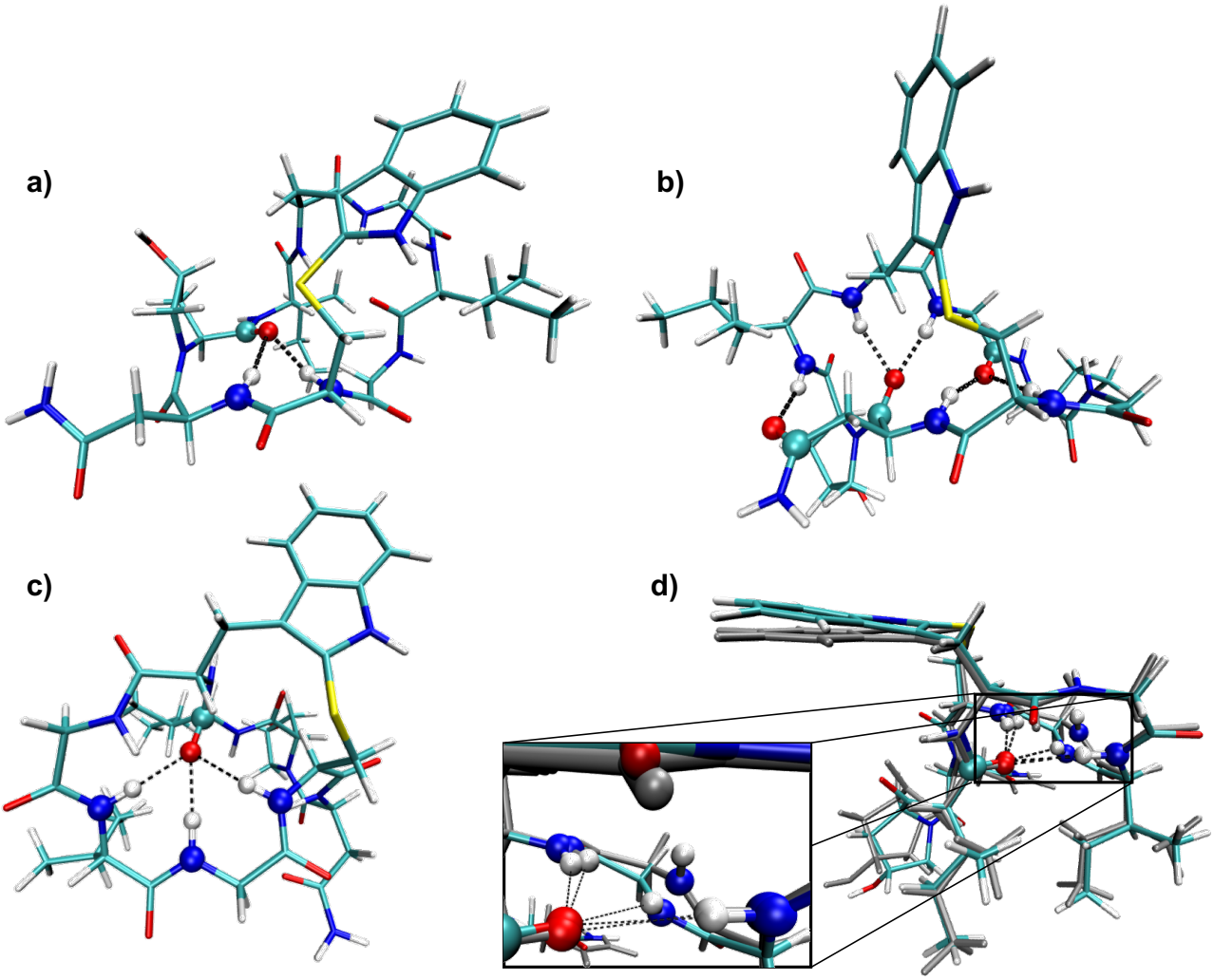


Supplementary Figure 10: Solubility and overall shape of isomers. Compounds a) **4a** and c) **4b** dissolved in water (at 2 mM concentration). Surface representations of b) **4a** and d) **4b**, respectively. Coloring according to the amino acid hydrophobicity following Kyte et al. [1] (hydrophobic residues are colored in orange and polar residues in blue).

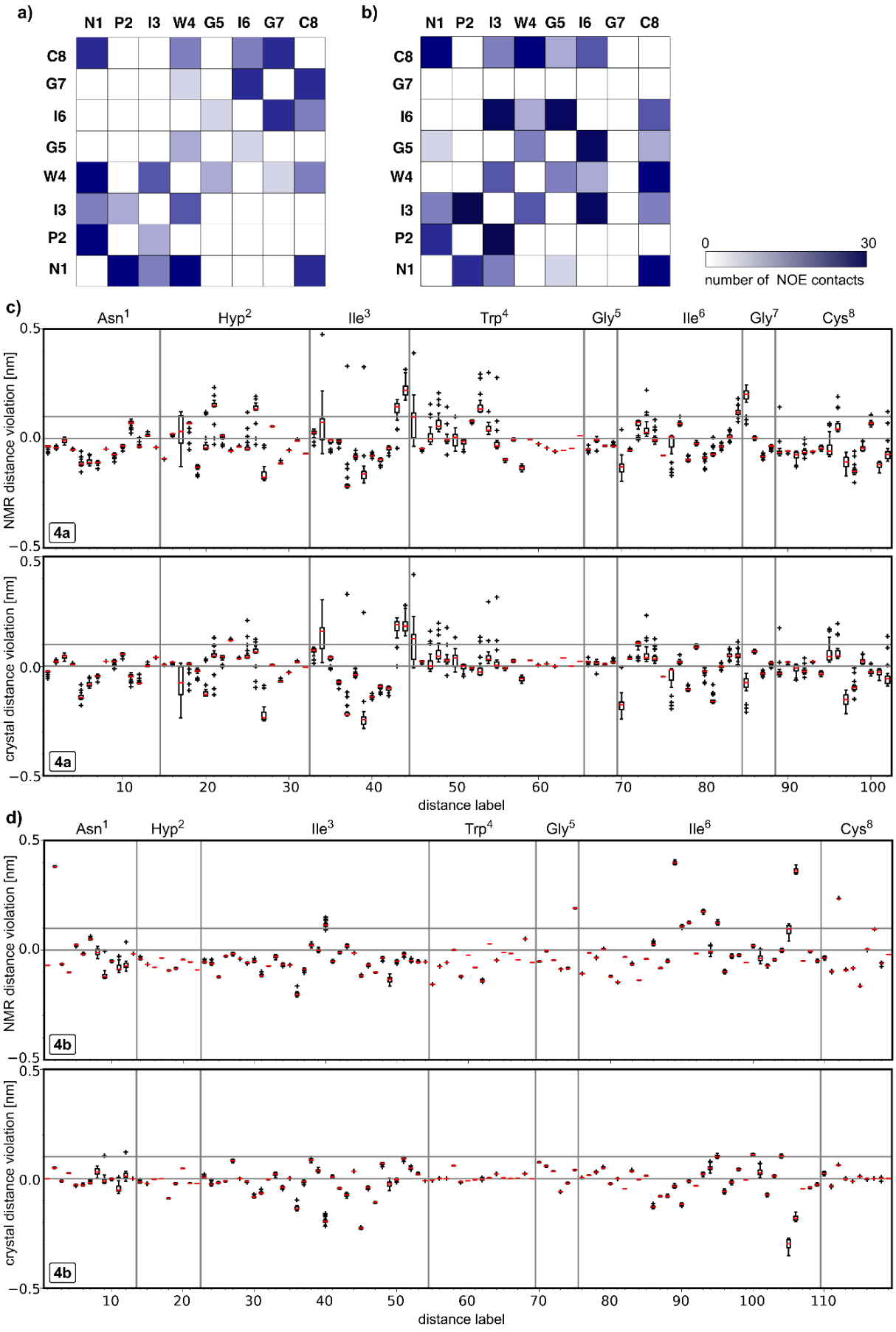


Supplementary Figure 11: Hydrogen bond sets observed in MD simulations. (a) Hydrogen bond sets '1' (olive) and '2' (blue) for **4a**. Hydrogen bond set '1' denotes the hydrogen bonds that agree with the crystal structure of **4a**: Asn^{1}(m)-Gly^{5}(m), Ile^{3}(m)-Asn^{1}(s), Trp^{4}(m)-Asn^{1}(m), Gly^{5}(m)-Asn^{1}(m), Cys^{8}(m)-Gly^{5}(m). Hydrogen bond set '2' denotes the hydrogen bonds that were identified as mutual exclusive to hydrogen set '1' (see Pearson correlations, Supplementary Figure 2b): Asn^{1}(m)-Hyp^{2}(m), Cys^{8}(m)-Hyp^{2}(m). (b,c) Time series for the trajectories of **4a** either expressed in b) hydrogen bond set existences or c) RMSD values. The coloring in (b) and (c) is based on the existence of the hydrogen bond sets: set '1' or set '0': olive, set '2' or set '0' : blue. Hydrogen bond set '0' denotes all short-term transitions out of the two conformations. The inset shows that all states are exclusive to each other. Vertical red lines denote the transition between the hydrogen bond sets '2' and '1' representing the transition

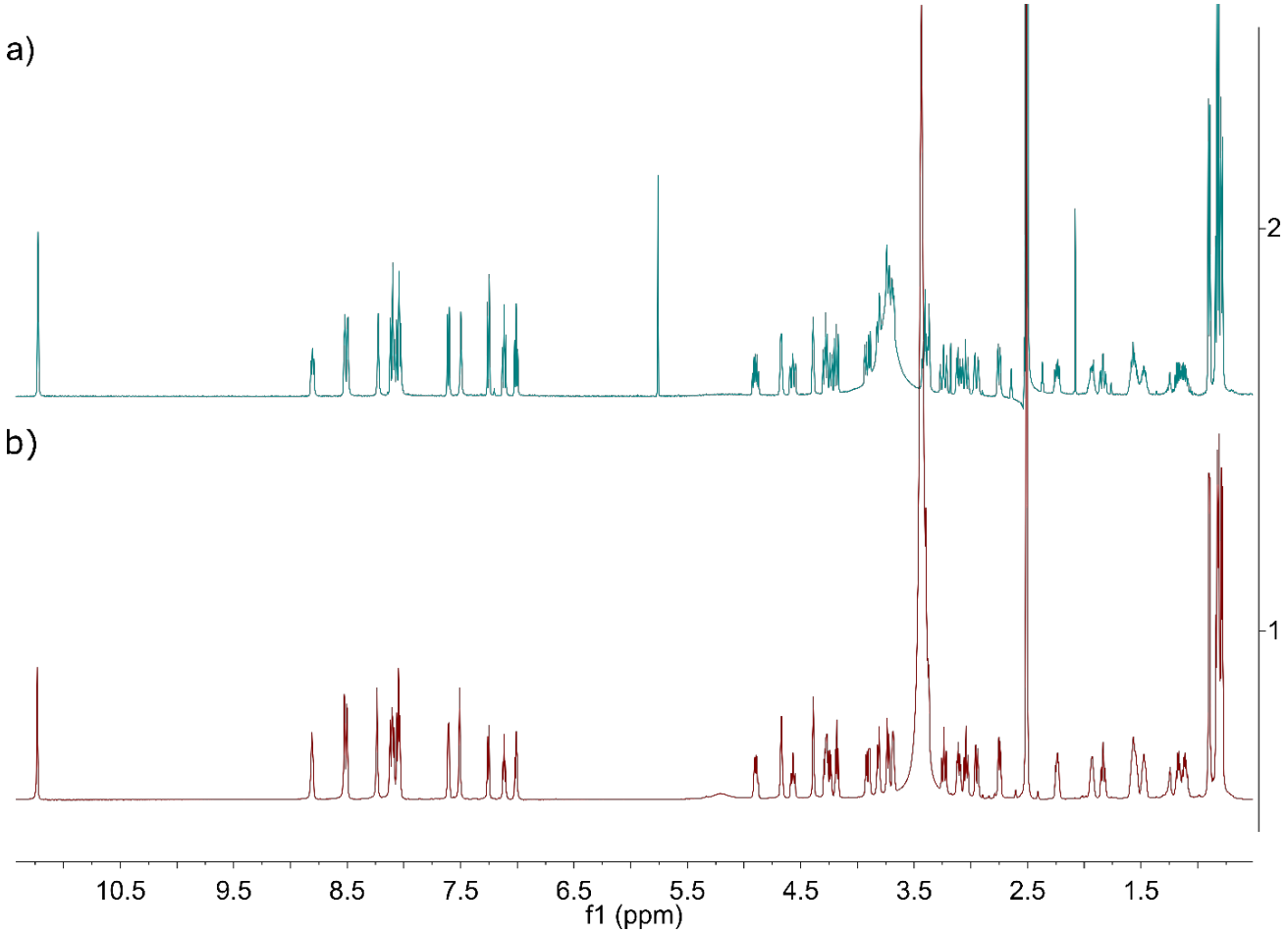
between the starting structure of the MD simulations and the crystal structure. For further details, please refer to section 3.2.2.10.



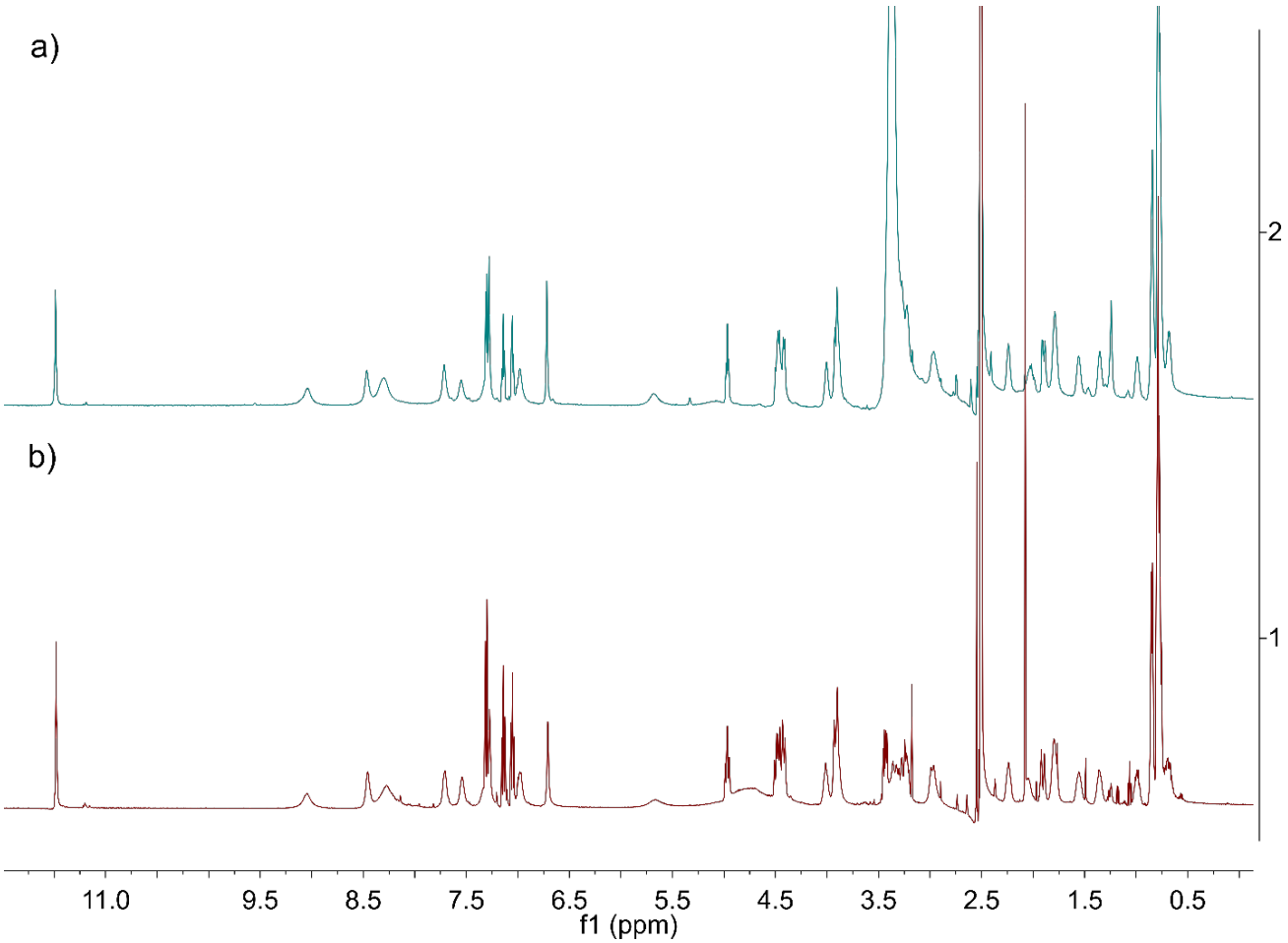
Supplementary Figure 12: MD starting structures. Starting structures of the production MD runs for **4a** with a) worst and b) best agreement to the crystal structure. As measure for the agreement between starting structures and the crystal structure, the RMSD was considered. The RMSD was calculated on all atoms after a least-square fit to the backbone. Important hydrogen bonds are highlighted and correspond to the mutual exclusive hydrogen bond sets 1 and 2. Set 2 (a) denotes the hydrogen bonds: Asn^{1}(m)-Hyp^{2}(m), Cys^{8}(m)-Hyp^{2}(m). Set 1 (b) denotes the hydrogen bonds that agree with the crystal structure of **4a**: Asn^{1}(m)-Gly^{5}(m), Ile^{3}(m)-Asn^{1}(s), Trp^{4}(m)-Asn^{1}(m), Gly^{5}(m)-Asn1(m), Cys^{8}(m)-Gly^{5}(m). 'm' denotes the main chain of the respective residue, 's' denotes the side chain. For the assignment of the atoms, please refer to the methods section ('Hydrogen bonds').
(c) Structure for **4b** taken from the MD ensemble that exhibits the hydrogen bonds Ile^{6}(m)-Ile^{3}(m), Gly^{7}(m)-Ile^{3}(m) and Cys^{8}(m)-Ile^{3}(m). (d) Overlay of the structure shown in (c) with the crystal structure of **4b** (grey). The region of the hydrogen bonds is shown in the details view. Atoms are colored according to their type: red: oxygen, blue: nitrogen, cyan: carbon, white: hydrogen, yellow: sulfur.


a)
N1 P2 I3 W4 G5 I6 G7 C8
C8
G7
I6
G5
W4
I3
P2
N1
b)
N1 P2 I3 W4 G5 I6 G7 C8
C8
G7
I6
G5
W4
I3
P2
N1
0
30
number of NOE contacts
c)
Asn1 Hyp2 Ile3 Trp4 Gly5 Ile6 Gly7 Cys8
NMR distance violation [nm]
4a
crystal distance violation [nm]
4a
distance label
d)
Asn1 Hyp2 Ile3 Trp4 Gly5 Ile6 Cys8
NMR distance violation [nm]
4b
crystal distance violation [nm]
4b
distance label

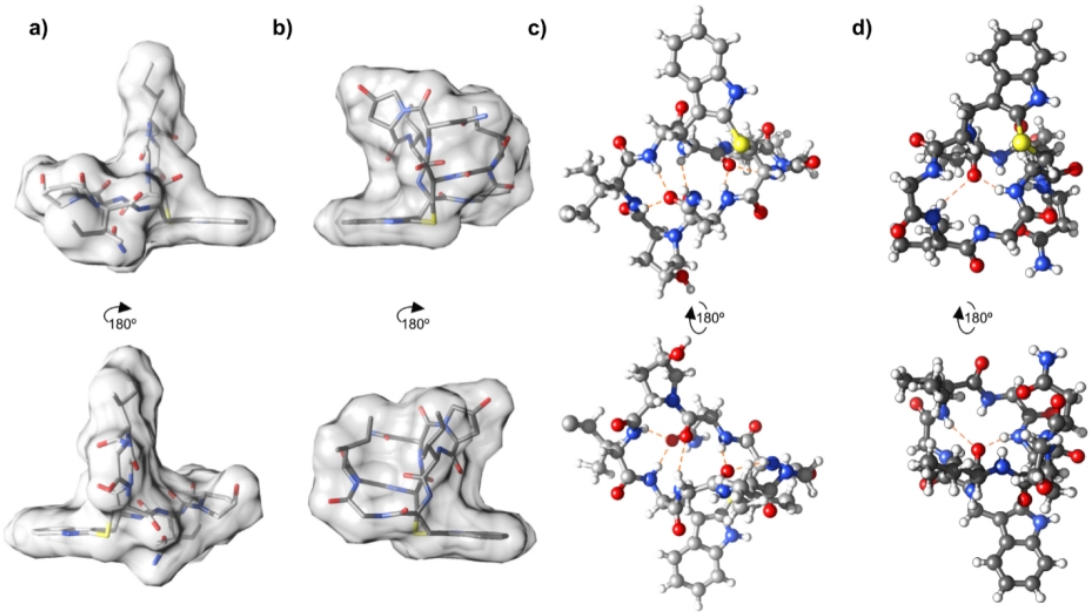
Supplementary Figure 13: Comparison of isomer NOE data and agreement with MD simulations. Distinct NOE patterns in contact maps of a) **4a** and b) **4b**. Distance violations of proton distances in the MD simulations compared to the NMR NOE data or crystal structures of **4a** c) and **4b** d). In each case, the same set of proton-proton distances was derived from NMR experiments and the crystal structures. The distances from MD were calculated as averages (see eq. 7) per replica representing 1 µs of simulation time respectively (see 3.2.2.11). The distribution of the resulting violations over all replicas (n=20) is shown as follows: the box represents 50% of the data points from the first to the third quartile with the median highlighted in red. The whiskers extend from the box by 1.5-fold of the inter-quartile range. Outliers are marked as black '+'.



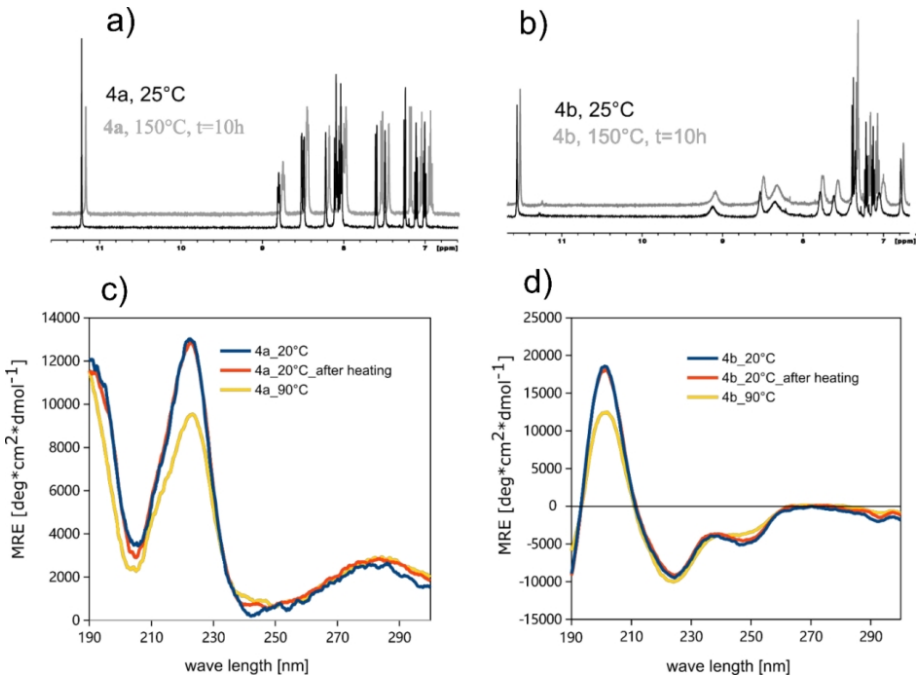
Supplementary Figure 14: Comparison of synthesis product 4a from different precursors. Overlay of ^{1}H NMR spectra of synthetic **4a,** a) **4a** was produced after macrolactamization of **3c** and b) **4a** was produced after macrolactamization of **2b**.



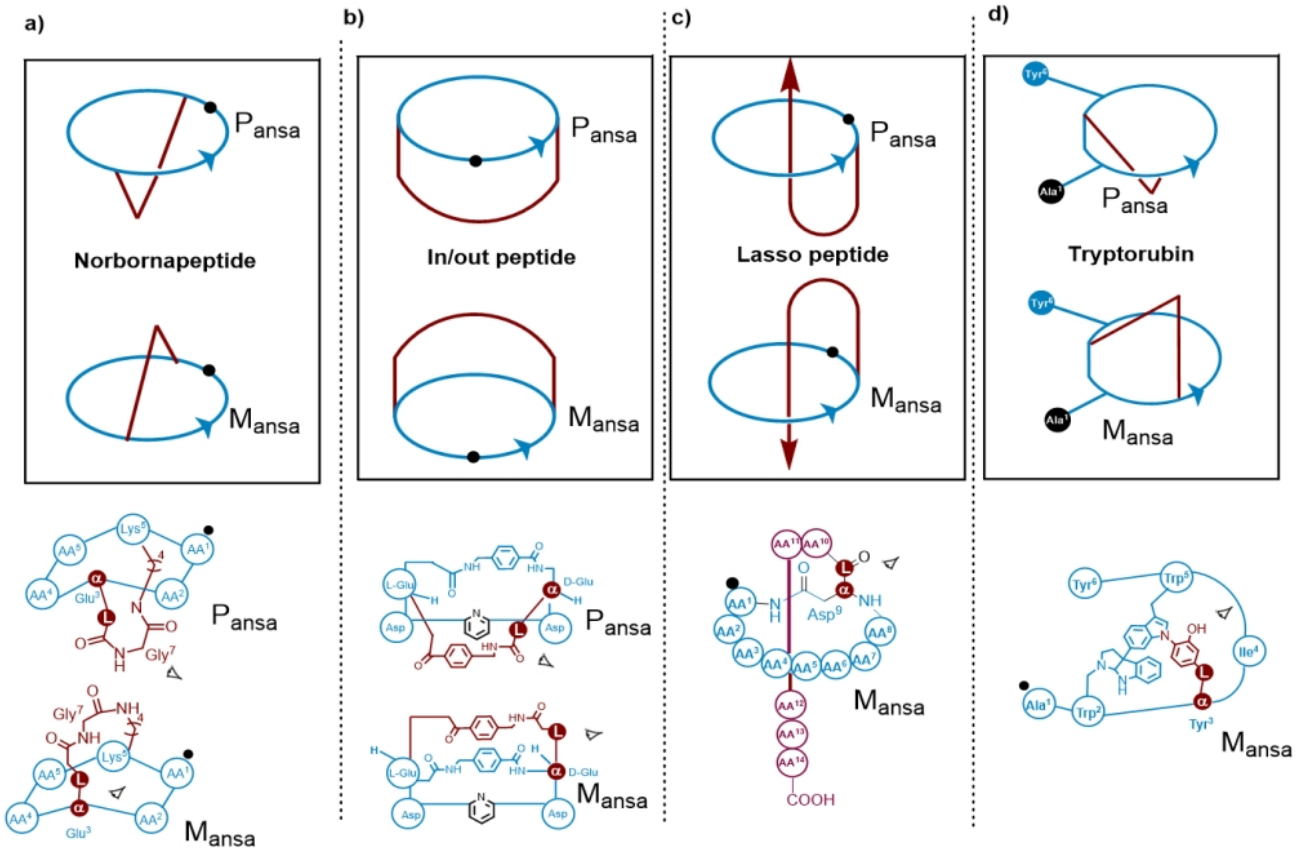
Supplementary Figure 15: Comparison of synthesis product 4b from different precursors. Overlay of ^{1}H NMR spectra of synthetic **4b**, a) **4b** was produced after macrolactamization of **3c** and **4b** was produced after macrolactamization of **2b**).



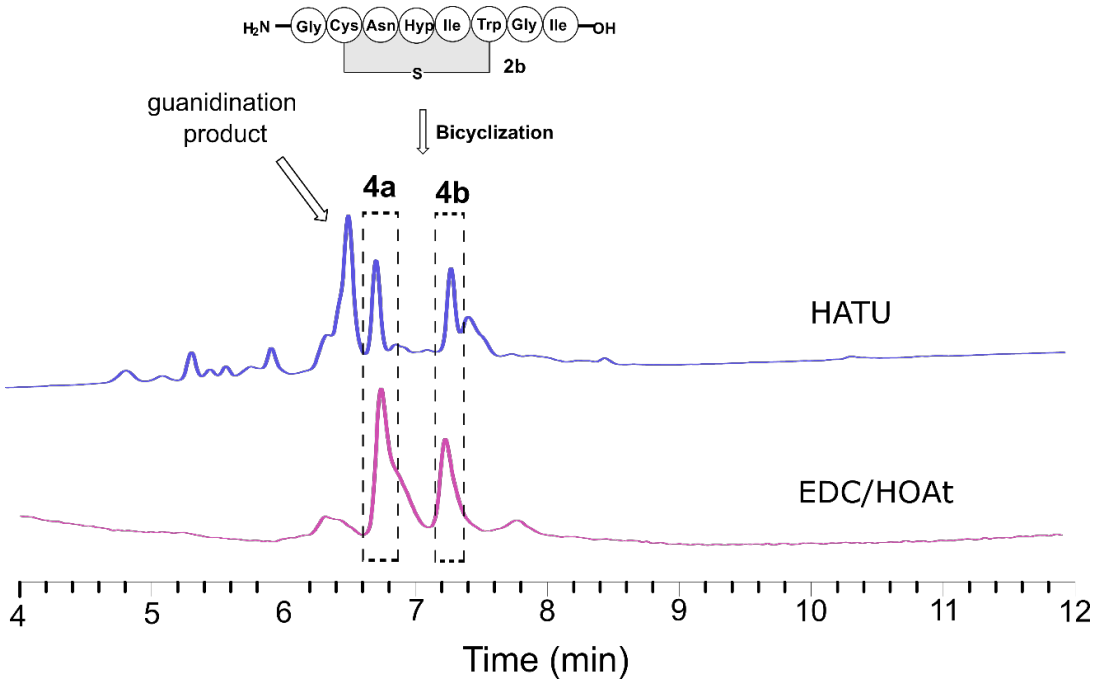
Supplementary Figure 16: Comparison of crystal structures 4a and 4b. Surface representation of the crystal structures of a) **4a** and b) **4b**. **4a** has a volume of 744.1 Å^3 and 598.0 Å^2.**4b** has a volume of 757.6 Å^3 and a surface area of 557.9 Å^2. Ball and stick models of crystal structures of c) **4a** and d) **4b** with hydrogen bonds shown as dashed orange lines. Atoms are colored according to their type: red: oxygen, blue: nitrogen, grey: carbon, white: hydrogen, yellow: sulfur.



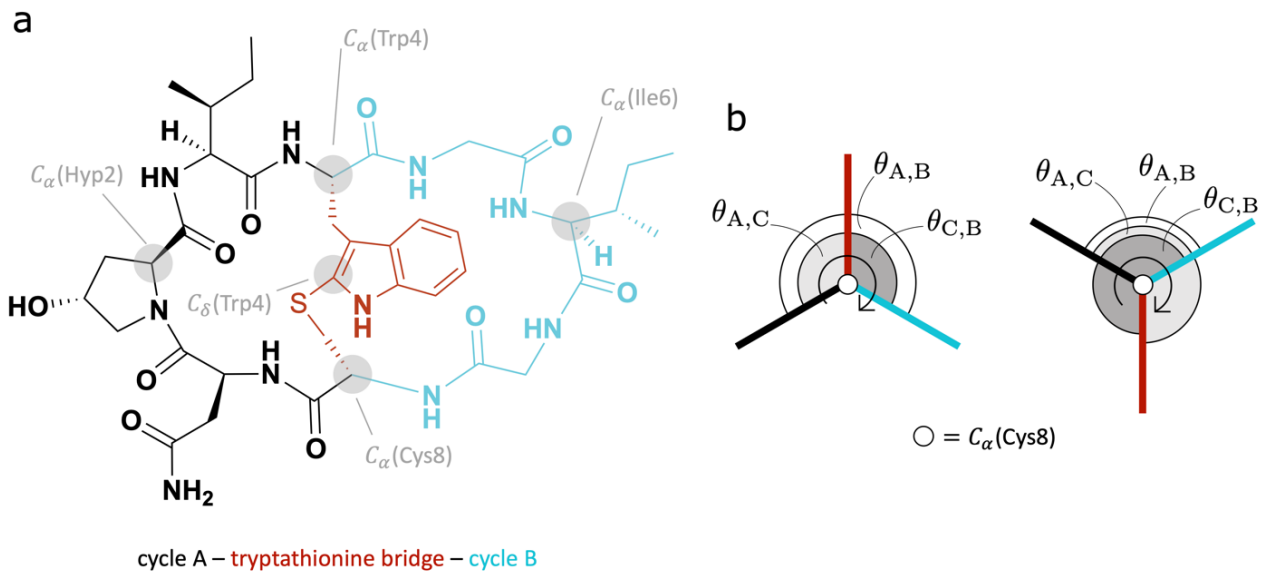
Supplementary Figure 17: Spectroscopic evaluation of heating experiments. a) ^{1}H NMR evidence for the lack of interconversion in **4a** upon heating and b) **4b;** c and d) CD spectroscopy study of peptides **4a** and **4b** at different temperatures.



Supplementary Figure 18: Examples of cyclic peptides and proposed application of the ansamer concept and the *P*$_{ansamer}$/*M*$_{ansamer}$ nomenclature. Main ring of the peptide (blue) with starting moiety/amino acid (black dot) forms the imaginary plane with proposed preferred directionality (blue arrow head). Bridge (red) above or below the main ring. Examples include a) norbornapeptide-type [2], b) in/out peptides [3], c) lasso peptides [4] and d) tryptorubin [5]. a) **Norbornapeptides**: For the drawn example there is an analogous situation as for the amanitin peptides. Main ring: head-to-tail cyclic peptide. Preferred directionality according to the directionality of the N- and C-terminus. b) **in/out peptides**: Determination of the main ring and assignment of the bridges leads to assignment of the P$_{ansa}$/M$_{ansa}$ descriptor. c) **Lasso peptides**: Main ring: cyclopeptide (example herein Gly^{1}-β-Asp^{9}). Preferred directionality according to the N- and C-terminus. Bridge: peptide tail (example herein: Ala^{10} to Asn^{19}). d) **Tryptorubin**: Main ring replaced here by linear peptide from N- to C-terminus (compare a related view on the vancomycin structure [6]) In this example the bridge connects theTyr^{3} with Trp^{5} sidechains of the main peptide.

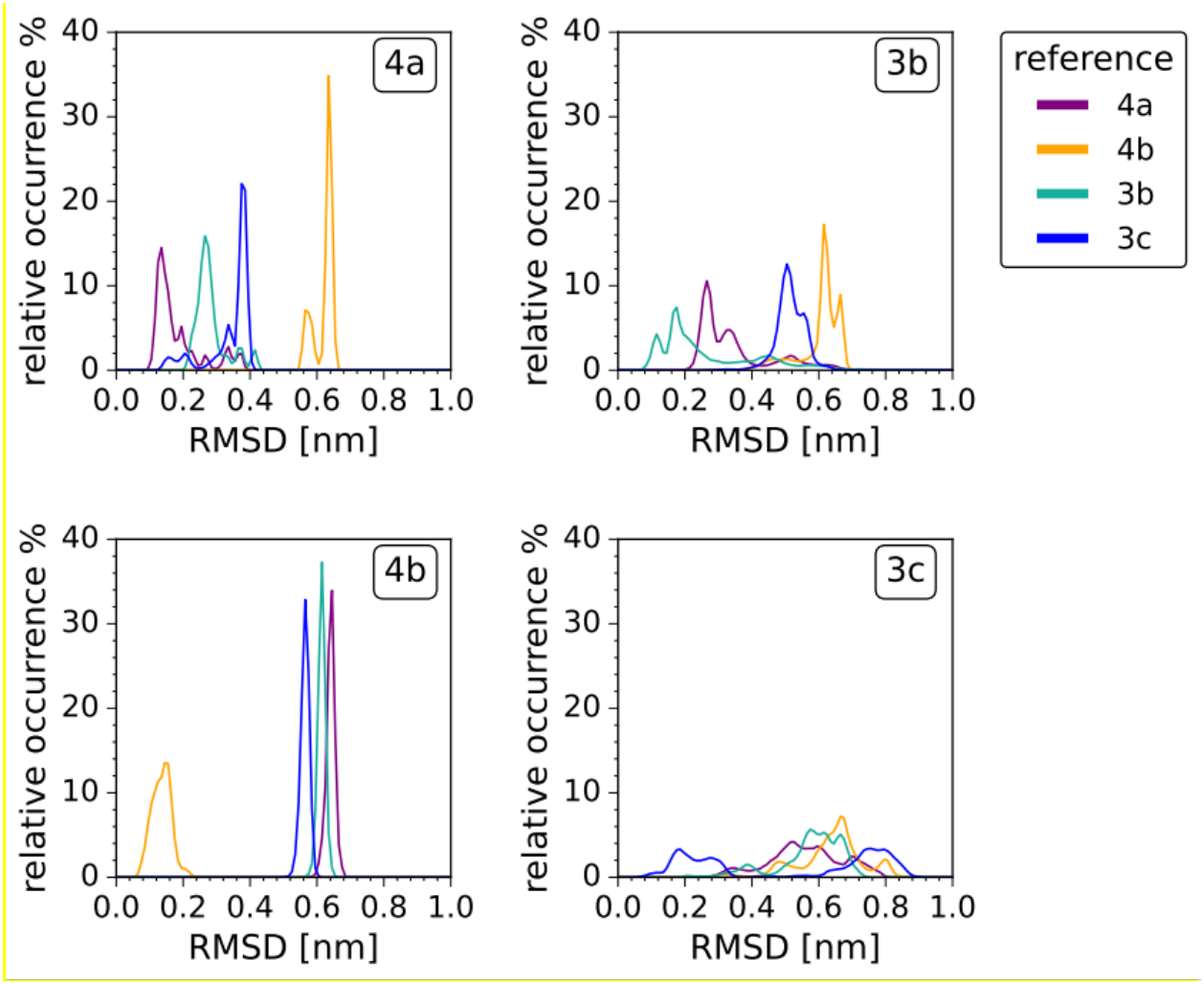


Supplementary Figure 19: Effects of coupling agents. LC-MS of macrolactamization of monocycle **2b** using HATU (blue line, top) or EDC/HOAt (pink line, bottom) as coupling reagent.



Supplementary Figure 20: Assignment of planes in the bicyclic amanitin derivatives. (a) The amanitin derivative molecule is divided into three subunits: ring A (black), ring B (cyan) and tryptathionine bridge (red). C_α atoms considered for the construction of the model shown in (b), are highlighted in gray. (b) Schematic representation of the angles between the planes E_A (black), E_B (cyan) and E_C (red). Each plane represents the subunit in (a) with the respective

color. Please note, in this scheme, the C_α atom of Trp^4 is located behind the C_α atom of Cys^8 and therefore not visible.



Supplementary Figure 21: **RMSD of MD structures to crystal structures.** Average distributions of the all-atom RMSD (after least-square fit on the backbone atoms of the existing ring) over all 20 replicas for each MD simulation data set (as labelled: top left: **4a**, bottom left: **4b**, top right : **3b** and bottom right : **3c**) towards different reference structures: 'purple' : **4a**, 'orange' : **4b**, 'green' : **3b**, 'blue' : **3c**. For **4a** and **4b**, the crystal structures were used as reference structures. For **3b** and **3c** the highest-probability structures were used.

2. Supplementary Tables

Supplementary Table 1: NMR shifts of **4a** and **4b**[a]

Molecule		1H Shift (ppm)										13C Shift (ppm)			
4a	**Residue**	HN	Hα	Hβ	Hγ	Hδ	Hε	Hε3	Hζ2	Hζ3	Hη2	Cα	Cβ	Cγ	Cδ
	1 Asn	8.52	4.67	3.38/2.94		8.23/7.50						50.04	33.08		
	2 Hyp		4.27	1.83/2.23	4.38	3.72/3.81						61.02	37.24	67.78	55.05
	3 Ile	8.04	4.18	1.93	1.46/1.16/0.89*	0.80*						57.26	34.81	24.28/15.14*	9.91
	4 Trp	8.03	4.89	3.23/3.10			11.23	7.6	7.25	7	7.11	52.82	29.13		
	5 Gly	8.08	3.38/4.25									40.43			
	6 Ile	8.5	3.68	1.56	1.56/1.10/0.78*	0.81*						58.28	33.83	24.47/14.12*	9.85
	7 Gly	8.8	3.90/3.40									41.61			
	8 Cys	8.1	4.56	2.74/3.04								51.81	37.69		
4b															
	1 Asn	7.31	4.96	2.52/1.89		7.27/6.71						46.23	36.54		
	2 Hyp		3.89	2.04/1.78	4	3.21/3.42						59.41	49.53	66.14	52.39
	3 Ile	7.71	3.89	1.79	1.35/0.98/0.78*	0.77*						56.33	33.01	23.63/13.85*	8.95
	4 Trp	9.04	5.68	3.27/2.96			11.48	7.54	7.3	7.05	7.14	nd	24.15		
	5 Gly	8.29	4.46/3.35									41.86			
	6 Ile	6.97	4.41	2.23	1.55/0.68/0.84*	0.79*						56.08	34.49	22.32/15.80*	10.98
	7 Gly	9.11	2.74									41.62			
	8 Cys	8.46	2.93	3.90/3.39								nd	34.13		

[a]: Note the large difference in the Cα shifts of the Asn[1] between **4a** and **4b** and the unusual Hα and Hβ shifts for Cys[8] in **4b**. No Cα correlations were observed in ^{1}H-^{13}C-HSQC spectra of **4b** for Trp[4] and Cys[8].

Supplementary Table 2: Structure calculation data CYANA

Cycle	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
Peaks:								
selected	346	346	346	346	346	346	346	
assigned	345	344	339	343	342	342	342	
unassigned	1	2	7	3	4	4	4	
with diagonal assignment	12	12	12	12	12	12	12	
Cross peaks:								
with off-diagonal assignment	333	332	327	331	330	330	330	
with unique assignment	196	245	258	274	285	294	291	
with short-range assignment |i-j|<=1	254	247	243	240	236	231	236	
with medium-range assignment 1<|i-j|<5	48	52	49	57	60	60	59	
with long-range assignment |i-j|>=5	31	33	35	34	34	39	35	
Upper distance limits:								
total	176	158	156	156	154	152	170	170
short-range, |i-j|<=1	114	100	99	95	92	87	95	95
medium-range, 1<|i-j|<5	55	51	35	40	41	41	46	46
long-range, |i-j|>=5	7	7	22	21	21	24	29	29
Average assignments/constraint	2.48	1.92	1.35	1.26	1.18	1.14	1	1
Average target function value	0.37	0.44	1.08	0.28	0.1	0.11	0.09	0.33
RMSD (residues 1..8)								
Average backbone RMSD to mean	0.17	0.22	0.04	0	0.02	0.07	0	0
Average heavy atom RMSD to mean	0.33	0.46	0.1	0.03	0.11	0.23	0.01	0.01

Supplementary Table 3: Approximate time points in µs for the transition from the starting structure to the crystal structure in each replica of **4a** if present. The transition times were determined as described in methods section 3.2.2.10 based on the hydrogen bond sets '1' and '2'. Hydrogen bond set '1' is the majorly-formed one and hydrogen bond set '2' represents a state that either directly (2→1) or indirectly (2→0→1) transitions into hydrogen bond set '1'.

		Hydrogen bond sets	RMSD
Replica	Transition	t [µs]	t [µs]
1	2→1	0.50	0.50
2	–	–	–
3	2→1	0.75	0.75
4	2→1	0.59	0.59
5	2→1	0.22	0.21
6	2→1	0.12	0.09
7	2→1	0.0013	–
8	2→1	0.18	–
9	2→1	0.015	–
10	2→1	0.062	0.062
11	2→1	0.26	0.27
12	–	–	–
13	2→1	0.093	–
14	2→1	0.29	0.29
15	2→1	0.30	0.30
16	2→1	0.53	0.53
17	–	–	–
18	–	–	–
19	–	–	–
20	–	–	–

Supplementary Table 4: Solvent accessible surface area (SASA) and VT-NMR data. SASA values in [nm^2] for the amides calculated as average±standard deviation over the entire simulation time (20 μs). The values were calculated using the 'shrake-rupley algorithm' available as a function of the mdtraj python library. [7] Amide proton shift changes from VT-NMR measurements are shown as ΔδHN/ΔT ($ppbK^{-1}$).

	4a_MD (nm^2)	**4a_cryst**(nm^2)	**4a_VT-NMR** ΔδHN/ΔT ($ppbK^{-1}$)	**4b_MD** (nm^2)	**4b_cryst** (nm^2)	**4b_VT-NMR** ΔδHN/ΔT ($ppbK^{-1}$)
Total	10.039 ± 0.277	10.11		9.728 ± 0.25	9.257	
Asn^1	0.006 ± 0.008	0.005	-1.550	0.006 ± 0.007	0.0	-3.826
Ile^3	0.005 ± 0.006	0.0	-1.793	0.007 ± 0.006	0.001	-6.421
Trp^4	0.006 ± 0.008	0.0	-2.910	0.033 ± 0.017	0.027	-4.807
Gly^5	0.015 ± 0.015	0.0	-1.432	0.072 ± 0.038	0.06	-4.962
Ile^6	0.039 ± 0.028	0.063	-4.171	0.003 ± 0.008	0.0	0.287
Gly^7	0.064 ± 0.03	0.131	-3.403	0.016 ± 0.025	0.036	nd
Cys^8	0.023 ± 0.024	0.031	-2.358	0.001 ± 0.005	0.0	-2.686

3. Supplementary Methods

3.1 Synthesis protocols and characterization data

3.1.1 Reagents, Solvents and Chromatographic Conditions

Commercially available reagents (Carl Roth GmbH and Co. KG, Karlsruhe, Germany; Sigma-Aldrich Taufkirchen, Germany; Iris Biotech GmbH, Marktredwitz, Germany; Orpegen, Heidelberg, Germany; ABCR, Karlsruhe, Germany; Alfa Aesar, Karlsruhe, Germany; Merck, Darmstadt; Germany; TCI, Eschborn, Germany; VWR International GmbH, Darmstadt, Germany; and Acros, Geel, Belgium) and solvents (Fisher Scientific-Acros, Schwerte, Germany) were used without further purification. If necessary, reactions were carried out under an atmosphere of argon or nitrogen and dry solvents. Analytical thin layer chromatography was carried out using aluminium-backed plates coated with silica gel (60, F254; Macherey & Nagel, Düren, Germany). Analysis was performed by visualizing the spots under UV light (λ = 254 nm), and/or by staining with $KMnO_4$ solution (3 g $KMnO_4$, 20 g K_2CO_3, 300 mL dest. H_2O, 5 mL NaOH solution (5 %)) and/or with Ninhydrin solution (0.3 g Ninhydrin, 3 mL AcOH, 100 mL *n*BuOH). Flash chromatography was carried out with silica gel (particle size 40-63 µm, VWR Chemicals, Darmstadt, Germany). Preparative HPLC was carried out on a 1260 Infinity (Agilent Technologies, Waldbronn, Germany) HPLC system with a polymeric reversed phase column (PLRP-S 100A) 300 x 50 mm, particle size 10 µm, Agilent Technologies, Waldbronn, Germany). ^{1}H and ^{13}C spectra were recorded at 298 K using the following spectrometers: Bruker Avance-II 400 MHz, Bruker Avance-III 500 MHz or Bruker Avance III 700 MHz (Bruker, Karlsruhe, Germany). The chemical shifts are reported in ppm using the residual solvent peak as an internal reference (DMSO-d_6, methanol-d_4, $CDCl_3$). Multiplicity (br. s = broad singlet, s = singlet, d = doublet, dd = doublet of doublet, t = triplet, q = quartet, m = multiplet) and coupling constants (J = Hz) are quoted where possible. HPLC-HRMS spectra were recorded on a QTrap LTQ XL (Thermo Fisher Scientific, Waltham, Massachusetts, USA) hyphenated to an Agilent 1200 Series HPLC-System (Agilent Technologies, Waldbronn, Germany) equipped with a C18 column (50 x 2 mm, particle size 3 µm). HPLC-HRMS chromatograms were obtained with a solvent gradient of 0.1% formic acid in water (Solvent A) and 0.1% formic acid in acetonitrile (Solvent B). The solvent gradients were either gradient A or gradient B: gradient A: 0-10 min 10%-50% B, 10-13 min 100% B, 13-16 min 20% B, gradient B: 0-10 min 20%-100% B, 10-13 min 100% B, 13-16 min 20% B Chiral HPLC was performed with a LaChrom system (Hitachi, Tokyo, Japan) equipped with a Chiralpak® Daicel-polysaccharide-column (250 x 4.6 mm, particle size 5 µm, Chiral Technologies Europe - Daicel Group, Illkirch, France).

3.1. 2 Abbreviations

CH_3CN, acetonitrile; COMU, 1-[(1-(cyan-2-ethoxy-2-oxoethylidenaminooxy)-dimethylamino-morpholino)]-uronium-hexafluorophosphate; THF, tetrahydrofuran; DCM, dichloromethane; DIPEA, *N,N'*-diisopropylethylamine; DMF, *N,N'*-dimethylformamide; TFA, trifluoroacetic acid; THF, tetrahydrofuran; HPLC, high-performance liquid chromatography; HATU, *O*-(7-azabenzotriazol-1-yl)-

N,N,N',N'-tetramethyluronium-hexafluorophosphate; 2-CTC, 2-chlorotrityl chloride (resin); TIS, triisopropylsilane; HFIP, hexafluoroisopropanol; D, L-FDLA, 1-fluoro-2,4-dinitrophenyl-5-D,L-leucine-amide; TBS, tert-butyldimethylsilyl; Fmoc, fluorenylmethoxycarbonyl protecting group. The amino acid three letter code was used for the proteinogenic amino acids according to IUPAC standards. If not otherwise stated, L-amino acids were used. Amino acid abbreviations: Hyp, *trans*-4-hydroxy-L-proline.

3.1.3 Variable temperature NMR (VT-NMR)

The temperature dependence of amide chemical shifts is well established and has been widely used to assess the solvent shielding properties of amide protons. Variable temperature NMR was carried out to assess the temperature dependency of NH chemical shifts. ^{1}H-NMR spectra were acquired from 303K to 343K (**4a**) or 300K to 354K (**4b**) in 10 K increments in DMSO-d6. Temperature coefficients are categorized as the following; ΔδHN/ΔT with values less than -4.6 ppb/K indicate solvent-exposed NHs. Intermediate values from –4.6 to –3.0 ppb/K indicate intermediate shielding and potentially weak or strained hydrogen bonding. Whereas ΔδHN/ΔT values greater than –3.0 ppb/K place NHs in the highly shielded and potentially strongly hydrogen bound category.

3.1.4 NMR assignment and structure calculation of desulfurized macrolactam **5**

To obtain resonance assignments for NOE assignment and structure calculations **4a**, **4b** and desulfurized macrolactam **5** were dissolved in deuterated DMSO-d_6 (approx. 10 mM). TOCSY, COSY, NOESY and ^{1}H-^{13}C-HSQC spectra were recorded on a Bruker Avance III 700 MHz spectrometer with a TXI 5 mm probe. Standard Bruker pulse programs were used and all spectra were acquired at 298 K. Residual solvent methyl peaks (DMSO-d7 δ = 2.502 for ^{1}H and δ = 39.0 ppm for ^{13}C) were used for chemical shift referencing. 2D homonuclear spectra were measured with acquisition times of 70 and 18 ms for the direct and indirect dimensions, respectively. TOCSY and NOESY spectra were accumulated with 16 or 32 (in case of **4b**) scans and COSY spectra with 8 scans. The TOCSY and NOESY mixing times were set to 100 and 300 ms, respectively. Natural abundance ^{1}H-^{13}C-HSQC spectra were measured with 140 scans and acquisition times of 14 and 120 ms for the direct and indirect dimensions. The spectra were processed and analyzed using TopSpin 3.5 (Bruker) and CcpNmr 2.3.1 [8] (The CCPN data model for NMR spectroscopy: development of a software pipeline). After shift assignment (Supplementary Table 1), the NOE correlations of **4a** and **4b** were manually assigned and residue interaction matrices of **4a** and **4b** were generated using CcpNmr.

For structure determination the manually assigned chemical shifts of the desulfurized macrolactam and NOESY peak lists were supplied to CYANA for automated NOE assignment and structure calculation (Supplementary Table 2). The program CYLIB [9] was used to generate a CYANA library file for 4-hydroxyproline. A set of 1000 structures was calculated and the 100 best were visually inspected with UCSF Chimera.[10]

3.1.5 Structure desulfurized macrolactam **5**

The calculations show a very similar backbone conformation compared to the Cys-containing macrolactam [11] with a RMSD of 0.806 Å between the two structures. Differences are largely found in the B ring around the two Gly residues and some minor differences in the side-chain orientations of Trp and Ile moieties. Comparison of J coupling constants between the two macrolactams suggests some small differences in phi angles, as observed in the calculated structures. These discrepancies could be due to the different solvents, DMF versus DMSO.

Structural alignments of **4a** and **4b** with the calculated macrolactam structure of **5** show larger differences, with the peptide ring of **4a** showing higher similarities to **5**. The RMSD between **4a** and **5** is 1.422 Å with a very similar backbone geometry of the A-ring and more differences in the B ring. The RMSD between **4b** and **5** is 2.454 Å with very different backbone geometry in rings A and B. In addition, compared to the relatively planar macrolactam, the peptide plane of **4b** shows a strong bend between rings A and B (see crystal structure and MD angle analysis of **4a**).

3.1.6 Analytical Methods

HPLC-MS: HPLC-HRMS spectra were recorded on a QTrap LTQ XL (Thermo Fisher Scientific, Waltham, Massachusetts, USA) hyphenated to an Agilent 1200 Series HPLC-System (Agilent Technologies, Waldbronn, Germany) equipped with a C18 column (50 x 2 mm, particle size 3 µm). HPLC-HRMS chromatograms were obtained with a solvent gradient of 0.1% formic acid in water (Solvent A) and 0.1% formic acid in acetonitrile (Solvent B).

The solvent gradients were shown below:

Gradient A: 0-10 min 10%-50% B, 10-13 min 100% B, 13-16 min 20% B,

Gradient B: 0-10 min 20%-100% B, 10-13 min 100% B, 13-16 min 20% B.

Gradient C: 0-10 min 50%-100% B, 10-13 min 100% B, 13-16 min 20% B.

Gradient D: 0-10 min 5%-100% B, 10-13 min 100% B, 13-16 min 20% B.

Gradient E: 0-30 min 10% to 65% B, 30-33 min 100%B, 33-36 min 20% B.

3.1.7 Experimental

3.1.7.1 General protocol

All linear peptides with different amino acid sequences were synthesized using this protocol with an alternating sequence of Fmoc-deprotections (*Method A*) and amino acid couplings (*Method B/C*).

Method A) Removal of the Fmoc group. A solution of 20% piperidine in DMF (5 mL) was added to the resin (1 g; loading 0.10-0.50 mmol/g) and the resulting suspension was shaken for 10 min. Then the solution was removed from the resin. Again, a solution of 20% piperidine in DMF (5 mL) was added

to the resin and the resulting suspension was shaken for another 10 min. The solution was drained and the resin was washed with DMF (6 x 5 mL).

Method B) Amino acid coupling. Amino acid (4.0 eq) and TBTU (4.0 eq) were dissolved in dry DMF (5 mL). DIPEA (12 eq) was added dropwise to the DMF solution. After activating for 1 min, the resulting solution was added to the Fmoc-deprotected resin (1 g; loading 0.10-0.50 mmol/g). The mixture was shaken until the coupling reaction was completed. Then, the solution was drained and the resin was rinsed with DMF (4 x 5 mL).

Method C) Amino acid coupling. Amino acid (4.0 eq) and HATU (4.0 eq) were dissolved in dry DMF (5 mL). DIPEA (12 eq) was added dropwise to the DMF solution. After activating for 1 min, the resulting solution was added to the Fmoc-deprotected resin (1 g; loading 0.10-0.50 mmol/g). The mixture was shaken until the coupling reaction was completed. Then, the solution was drained and the resin was rinsed with DMF (4 x 5 mL).

3.1.7.2 I_2-mediated thioether formation

The thioether bridge (tryptathionine motif) was obtained from linear resin-bound peptide (1 eq; 500 mg; loading 0.10-0.50 mmol/g) synthesized according to the above method. Formation of the thioether was achieved by adding a freshly prepared solution of iodine in DMF (2 eq, 2 mg/ml) under protecting gas atmosphere (Ar or nitrogen). The mixture was shaken under nitrogen atmosphere for 2.5 h to complete the formation of the thioether. If required also longer reaction times were applied (HPLC-MS control of a test cleavage). Then, the solution was drained and the resin was rinsed with DMF (4 x 3 mL).

3.1.7.3 Cleavage from solid support

Condition A) The resin (1 g; loading 0.3 mmol/g) was treated with 10 mL of a mixture of TFA/TIS/H2O (95:2.5:2.5) for 1 h at room temperature with gentle agitation. The resin was filtered and rinsed with 1% TFA in DCM (2 x5 mL). The rinses and filtrate were combined and evaporated to dryness.

Condition B) The resin (1 g; loading 0.10-0.50 mmol/g) was treated with 10 mL of a mixture of HFIP/DCM (3:7) for 2 h at room temperature with gentle agitation. The resin was filtered and rinsed with DCM (2 x5 mL). The rinses and filtrate were combined and evaporated to dryness.

3.1.7.4 Monitoring of Peptide Coupling and Capping

Chloranil test: During the coupling reaction, a few resin beads were taken out and rinsed with DMF (2 x 1 mL). To the resin were added 2 drops of a 2% solution of acetaldehyde and 2 drops of a 2% solution of chloranil in DMF. The resulting suspension was allowed to stand for 5 min at room temperature. Blue- to green-stained beads indicated the presence of secondary amines.

Kaiser Test: During the coupling reaction, a few resin beads were taken out and rinsed with DMF (2 x 1 mL). To the resin were added 2-3 drops of reagent A (16.5 mg of KCN dissolved in 25 mL of distilled water. 1.0 mL of above solution diluted with 49 mL of pyridine), 2 to 3 drops of reagent B (1.0 g of ninhydrin dissolved in 20 mL of n-butanol.) and 2 to 3 drops of reagent C (40 g of phenol dissolved

in 20 mL of n-butanol). The resulting suspension was allowed to heat at 110°C for 3 min, blue- to green-stained beads indicated the presence of primary amine.

3.1.7.5 Macrolactamization

Condition A) To a solution of monocyclic peptide in solution of DPEA (5 eq) in DMF, HATU (2 eq) was added at 0°C. The solution was stirred for 12 h, followed by preparative HPLC purification. The isolated product was lyophilized to give the white solid.

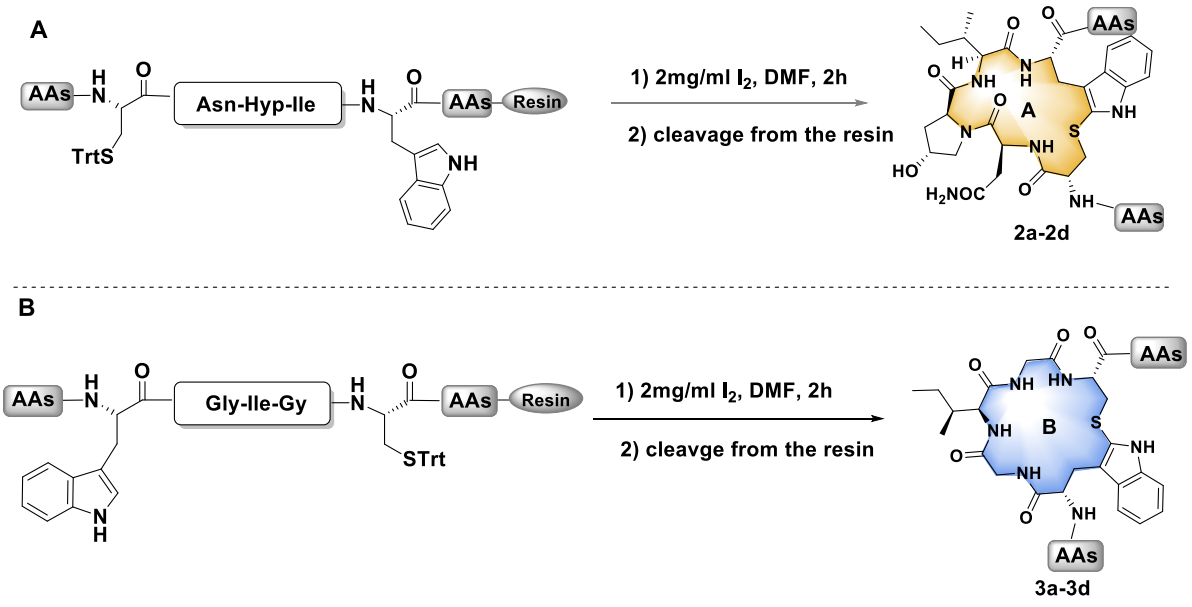
Condition B) To a solution of EDCI (2 eq) and HOAt (2 eq) in solution of DPEA (5 eq) in DMF, monocyclic peptide was added at 0°C. The solution was stirred for 12 h, followed by preparative HPLC purification. The isolated product was lyophilized to give the white solid.

3.1.7.6 Deprotection

Detritylation: The final protected bicyclic peptides were dissolved in DCM (2 ml) and followed by addition of TFA/TIS/H_2O (95:2.5:2.5, 2ml) and were stirred for 30 min. Afterwards, the reaction mixture was evaporated under reduced pressure and the followed crude product was purified by preparative HPLC to give the final compound.

3.1.8 Synthesis and characterization data

3.1.8.1 Synthesis of monocyclic peptides **2a-2d** and **3a-3d**



Supplementary Figure 22. Synthesis scheme of octapeptides. Synthesis of monocyclic octapeptides (**2a-2d** and **3a-3d**) with tryptathionine bridge. A) Synthesis scheme for formation of monocycle A (yellow) in peptides **2a-2d**. B) Synthesis scheme for formation of monocycle B (blue) in peptides **3a**-3d.

General protocol for solid-phase peptide synthesis (SPPS):

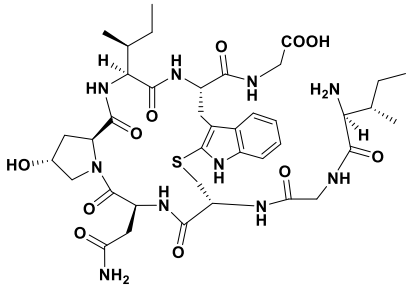
2-CTC resin (1 g, 0.98 mmol/g) was pre-swollen for 20 min in DCM in a manual solid phase peptide synthesis vessel (10 mL). After the solvent was drained, the first amino acid Fmoc-AA[1]-OH (0.3 mmol) and DIPEA (0.26 mL, 1.5 mmol) in DCM (5 mL) were added to the resin. The mixture was agitated for 2 hbefore the solvent was drained. The resin was rinsed with DMF (4 x 3 mL). Then a mixture of MeOH/DIPEA/DCM (1:1:8) was added to cap the remaining 2-chlorotrityl chloride on the resin. The mixture was agitated for 0.5 h. Then the solvent was drained and the resin was washed with DMF (4 x 3 mL). The resin loading was determined to be 0.30 mmol/g. The Fmoc-group was removed according to Method *A*. Fmoc-AA[2]-OH (4 eq) was coupled to the deprotected resin according to Method *B*. The Fmoc-group of the resulting resin was removed according to Method *A*. The following six amino acids were coupled to the deprotected resin according to Method *A* and *B*. The tryptathionine formation was carried out on the solid support. After removal of Fmoc-group using method *A* and followed cleavage from the resin, the monocyclic peptide was obtained following subsequent HPLC purification. The synthesis and characterization data of **2a-2d**, **3a-3b** and **3d** have been reported in our previous paper.[11]

2a: **H_2N-Gly-Ile-Gly-(*cyclo-tryptathionine*)[Cys-Asn-Hyp-Ile-Trp]-OH**

HRMS (ESI): m/z calculated: $C_{39}H_{57}N_{10}O_{11}S^+$ $[M+H]^+$ 873.3923, found 873.3912.

HPLC-MS: Retention time R_t = 8.08 min (Gradient B).

2b: **H_2N-Ile-Gly-(*cyclo-tryptathionine*)[Cys-Asn-Hyp-Ile-Trp]-Gly-OH**



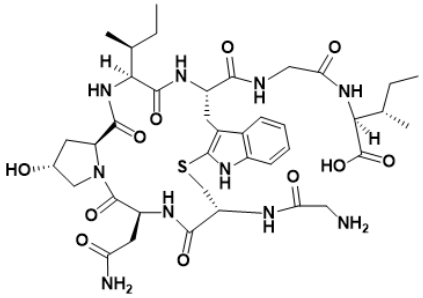
Mixed conformers: ^{1}H NMR (400 MHz, DMSO-d_6) δ 12.3 (s, 1.65H), 11.19 (s, 1H), 11.00 (s, 0.65 H), 8.69 (t, *J* = 5.6 Hz, 1 H), 8.60 (t, *J* = 5.6 Hz, 1H), 8.44 (d, *J* = 8.6 Hz, 0.65 H), 8.32 (d, *J* = 7.9 Hz, 0.65 H), 8.25 (t, *J* = 5.8 Hz, 0.65 H), 8.11 (s, 5H), 7.87 (t, *J* = 9.5 Hz, 2H), 7.70 (s, 1H), 7.60 (d, *J* = 7.9 Hz, 0.65 H), 7.57 (s, 1H), 7.47 (d, *J* = 8.0 Hz, 1H), 7.34 (s, 1H), 7.27 (t, *J* = 8.5 Hz, 2H), 7.17 – 6.92 (m, 4H), 6.83 (s, 0.65 H), 4.77 – 4.53 (m, 5H), 4.32 – 4.25 (m, 2H), 4.16 – 4.11 (m, 1H), 4.05 (dd, *J* = 16.7, 5.8 Hz, 1H), 3.97-3.88 (m, 3H), 3.87 – 3.78 (m, 3H), 3.75 – 3.57 (m, 3H), 3.48 (d, *J* = 11.1 Hz, 1H), 3.43 – 3.27 (m, 1H), 3.22 – 3.08 (m, 5H), 3.04 – 2.97 (m, 1H), 2.92 (d, *J* = 9.8 Hz, 1H), 2.71 – 2.58 (m, 3H), 2.26 (dd, *J* = 15.4, 4.0 Hz, 1H), 2.16 – 1.97 (m, 3H), 1.88 – 1.65 (m, 2H), 1.59 – 1.30 (m, 3H), 1.26 – 1.10 (m, 1.5H), 1.06– 0.97 (m, 1H), 0.96 – 0.82 (m, 10H), 0.77 (t, *J* = 7.3 Hz, 2H), 0.73 (d, *J* = 6.8 Hz, 2H), 0.60 (s, 3H), 0.39 (s, 3H).

^{13}C NMR (101 MHz, DMSO-d_6_HSQC) δ 123.3, 122.6, 122.3, 119.6, 119.4, 119.0, 119.2, 111.3, 69.3, 67.4, 61.7, 60.0, 59.8, 58.0, 56.8, 55.4, 54.0, 53.0, 50.8, 47.4, 41.9, 41.4, 41.2, 37.1, 34.3, 28.0, 27.8.

HRMS (ESI): m/z calculated: $C_{39}H_{57}N_{10}O_{11}S^+$ $[M+H]^+$ 873.3923, found 873.3917.

HPLC-MS: Retention time R_t = 7.40 min (Gradient B).

2c: **H_2N-Gly-(*cyclo-tryptathionine*)[Cys-Asn-Hyp-Ile-Trp]-Gly-Ile-OH**



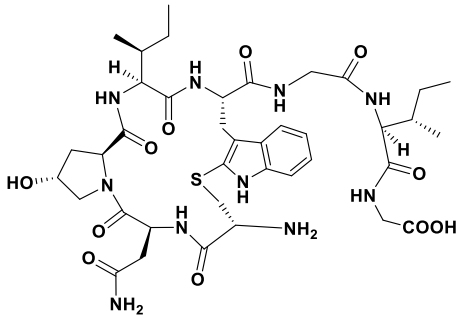
Major conformer: 1H NMR (500 MHz, DMSO-d_6) δ 12.49 (s, 1H), 11.17 (s, 1H), 8.86 (s, 1H), 8.35 (d, J = 7.8 Hz, 1H), 8.07 (t, J = 5.9 Hz, 1H), 8.00 (d, J = 6.0 Hz, 1H), 7.72 (d, J = 6.3 Hz, 1H), 7.51 (s, 1H), 7.48 (d, J = 7.8 Hz, 1H), 7.44 (d, J = 8.0 Hz, 1H), 7.26 (d, J = 8.3 Hz, 1H), 7.12 – 7.08 (m, 1H), 6.99 (t, J = 7.4 Hz, 1H), 4.38 (d, J = 3.9 Hz, 1H), 4.30 – 4.21 (m, 1H), 4.07 (t, J = 7.2 Hz, 1H), 3.82 – 3.73 (m, 2H), 3.21 (dt, J = 14.7, 7.7 Hz, 1H), 3.15 (dd, J = 15.2, 5.5 Hz, 1H), 2.68 – 2.59 (m, 1H), 2.13 – 2.03 (m, 1H), 1.88 – 1.78 (m, 2H), 1.43-1,32 (m, 2H), 1.24-1,15 (m, 1H), 1.04 (s, 1H), 0.89 – 0.83 (m, 10H), 0.81-0.72 (m, 3H).

^{13}C NMR (126 MHz, DMSO-d_6_HSQC) δ 122.5, 119.6, 119.4, 119.2, 111.4, 69.4, 67.4, 62.1, 59.8, 57.9, 57.1, 56.6, 55.7, 53.5, 42.9, 42.7, 40.7, 40.5, 40.3, 39.8, 37.6, 36.7, 36.2, 34.6, 28.0, 25.3, 25.1, 24.9, 15.7, 15.5, 11.6, 11.2.

HRMS (ESI): m/z calculated: $C_{39}H_{57}N_{10}O_{11}S^+$ $[M+H]^+$ 873.3923, found 873.3915.

HPLC-MS: Retention time R_t = 8.67 min (Gradient B).

2d:**H_2N-(*cyclo-tryptathionine*)[Cys-Asn-Hyp-Ile-Trp]-Gly-Ile-Gly-OH**



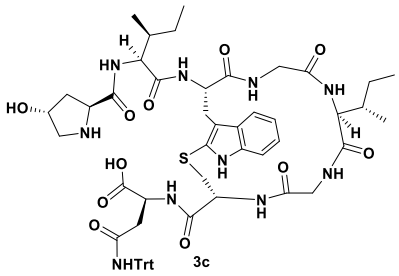
Mixed conformers: ^{1}H NMR (500 MHz, DMSO-d_6) δ 12.47 (s, 1H), 11.29 (s, 0.55 H), 11.15 (s, 0.45H), 8.63 (d, *J* = 8.1 Hz, 0.45 H), 8.38 – 8.32 (m, 3H), 8.15 (s, 0.45 H), 8.03 – 7.95 (m, 1H), 7.93 (d, *J* = 8.9 Hz, 1H), 7.79 (d, *J* = 11.6 Hz, 1H), 7.63 (d, *J* = 8.0 Hz, 0.55 H), 7.50 – 7.29 (m, 3H), 7.15 (t, *J* = 6.9 Hz, 1H), 7.02 (t, *J* = 6.9 Hz, 1H), 6.82 (s, 0.45 H), 4.80 (d, *J* = 3.2 Hz, 0.55 H), 4.74 (dd, *J* = 8.3, 6.3 Hz, 0.45 H), 4.64 – 4.50 (m, 1H), 4.41 (t, *J* = 3.7 Hz, 0.55 H), 4.34 – 4.21 (m, 1H), 4.14 – 4.07 (m, 1H), 3.96 – 3.83 (m, 1H), 3.83 – 3.52 (m, 3H), 3.38 – 3.30 (m, 1H), 3.24 – 3.17 (m, 1H), 3.15 – 3.07 (m, 1H), 2.96 (d, *J* = 6.2 Hz, 1H), 2.78-2.63 (m, 1H), 2.29 – 2.22 (m, 0.55 H), 2.17 – 2.10 (m, 1H), 1.91– 1.81 (m, 2H), 1.76 – 1.66 (m, 1.5H), 1.57 – 1.39 (m, 2.5 H), 1.20 – 1.05 (m, 2.5H), 0.92 – 0.77 (m, 14H).

^{13}C NMR (126 MHz, DMSO-d_6_HSQC) δ 123.0, 119.9, 119.6, 119.4, 111.5, 69.6, 69.2, 67.5, 67.4, 62.6, 59.5, 59.3, 59.1, 58.6, 58.5, 57.8, 56.9, 56.4, 55.9, 55.7, 54.7, 53.5, 53.3, 53.0, 51.4, 49.9, 47.7, 47.5, 43.4, 42.4, 42.2, 41.0, 40.8, 40.7, 40.0, 38.9, 38.8, 37.7, 37.6, 37.2, 36.4, 35.5, 34.8, 34.7, 28.5, 28.3, 26.8, 25.8, 25.6, 24.6, 24.4, 24.2, 16.3, 15.8, 15.7, 15.5, 13.3, 12.2, 11.6, 11.4, 11.0.

HRMS (ESI): m/z calculated: $C_{39}H_{57}N_{10}O_{11}S^{+}$ $[M+H]^{+}$ 873.3923, found 873.3913.

HPLC-MS: Retention time R_t = 8.42 min (Gradient B).

3c: H_2N-Hyp-Ile-(*cyclo-tryptathionine*)[Trp-Gly-Ile-Gly-Cys]-Asn(Trt)-OH



3c

After loading the Fmoc-Asn(Trt)-OH (0.3 mmol) on the resin according the general synthesis, the following peptide sequences Fmoc-Cys(Trt)-OH, Fmoc-Gly-OH (4 eq), Fmoc-L-Ile-OH (4 eq), Fmoc-Gly-OH (4 eq), Fmoc-L-Trp-OH (4 eq), Fmoc-L-Ile-OH (4 eq) and Fmoc-Hyp-OH (4 eq) were coupled to the deprotected resin according to Method *A* and *B*. The tryptathionine formation was carried out on the solid support. After removal of Fmoc group using method *A* and followed cleavage from the resin using cleavage condition B, 150 mg peptide **3c** was obtained as white solid powder with 55% yield following subsequent HPLC purification.

^{1}H NMR (500 MHz, DMSO-d_6) δ 12.61 (s, 1H), 11.23 (s, 1H), 9.53 (s, 1H), 8.61 (s, 1H), 8.58 (s, 1jH), 8.53 (d, *J* = 8.4 Hz, 1H), 8.29 – 8.16 (m, 2H), 8.19 (d, *J* = 7.6 Hz, 1H), 8.13 (d, *J* = 7.7 Hz, 1H), 7.76 (t, *J* = 4.7 Hz, 1H), 7.68 (d, *J* = 5.8 Hz, 1H), 7.65 (t, *J* = 5.5 Hz, 1H), 7.31 – 7.18 (m, 7H), 7.21 – 7.07 (m, 9H), 7.01 – 6.95 (t, *J* = 7.6 Hz, 1H), 5.46 (s, 1H), 4.69 (d, *J* = 2.5 Hz, 1H), 4.53 – 4.40 (m, 2H), 4.33 (d, *J* = 4.5 Hz, 1H), 4.27 – 4.18 (m, 1H), 4.08 (t, *J* = 8.0 Hz, 1H), 4.04 (t, *J* = 7.4 Hz, 1H), 3.92 (dd, *J* = 15.3, 5.9 Hz, 1H), 3.87 (d, *J* = 5.8 Hz, 1H), 3.72 (dd, *J* = 16.8, 4.4 Hz, 1H), 3.65 (dd, *J* = 15.5, 4.7 Hz, 1H), 3.39 (dd, *J* = 13.5, 3.6 Hz, 1H), 3.32 – 3.26 (m, 2H), 2.95 – 2.90 (m, 1H), 2.89 – 2.82 (m, 1H), 2.79 – 2.73 (m, 1H), 2.56 (dd, *J* =

15.6, 7.0 Hz, 1H), 2.14 (dd, *J* = 13.3, 7.3 Hz, 1H), 1.93 (qd, *J* = 6.7, 3.4 Hz, 1H), 1.60 (dd, *J* = 10.2, 3.9 Hz, 1H), 1.30 – 1.22 (m, 1H), 1.20 – 1.10 (m, 1H), 0.99 (m, 1H), 0.91 (d, *J* = 8.8 Hz, 0H), 0.88 (d, *J* = 6.8 Hz, 3H), 0.83 (t, *J* = 7.4 Hz, 3H), 0.76 (t, *J* = 7.4 Hz, 3H), 0.65 (d, *J* = 6.7 Hz, 3H).

^{13}C NMR (101 MHz, DMSO-d_6_HSQC) δ 129.0, 127.9, 126.9, 122.5, 119.5, 119.0, 111.3, 69.3, 59.2, 58.8, 58.0, 57.8, 53.9, 53.5, 52.0, 49.2, 43.2, 42.7, 38.8, 38.1, 37.9, 36.7, 35.7, 35.5, 29.0, 28.7, 28.4, 24.6, 24.3, 24.1, 15.7, 15.2, 11.3, 11.1.

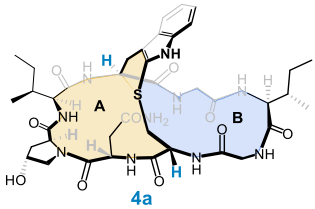
HRMS (ESI): m/z calculated: $C_{58}H_{71}N_{10}O_{11}S^+$ $[M+H]^+$ 1115.5019, found.1115.5009

HPLC-MS: Retention time R_t= 6:98 min (Gradient B)

3.1.8.2 Synthesis of bicyclic peptides **4a** and **4b**

General synthesis: Monocyclic octapeptide **2a-2d** and **3a-3d** (1.0 eq) was dissolved in DMF (1 mM). Then, DIPEA (2.2 eq) and HATU (2.0 eq) was added at 0°C. The reaction mixture was allowed to warm to r.t. for 12 h and concentrated under reduced pressure. The crude product was purified using preparative HPLC to afford bicyclic octapeptide as a white powder. Since large amounts of guanidination product were detected during macrolactamization of **2b** (see supplementary figure 19), the alternative coupling condition EDC (2 eq) and HOAt (2 eq) was employed to cyclize the monocyclic peptide **2b**. All results of LC-MS runs are shown in Supplementary Figure 1. In addition, the detritylation was performed after macrolactamization of **3c**. The yield and ratio of **4a** and **4b** is shown in the table below and in Figure 2.

Natural ansamer: Ile^3-*S*-deoxo-amaninamide (**4a**)



4a

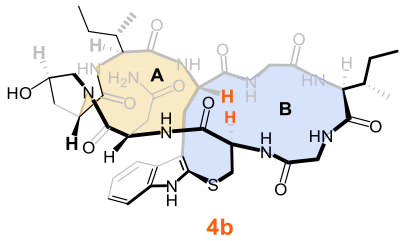
^{1}H NMR (500 MHz, DMSO-d_6) δ 11.22 (s, 1H), 8.80 (dd, *J* = 7.4, 5.1 Hz, 1H), 8.50 (dd, *J* = 13.2, 3.7 Hz, 2H), 8.22 (s, 1H), 8.09 (t, *J* = 8.7 Hz, 2H), 8.04 (t, *J* = 9.0 Hz, 2H), 7.60 (d, *J* = 8.0 Hz, 1H), 7.49 (s, 1H), 7.25 (d, *J* = 8.1 Hz, 1H), 7.11 (t, *J* = 7.4 Hz,1H), 7.00 (t, *J* = 7.3 Hz, 1H), 5.75 (s, 1H), 4.89 (dt, *J* = 13.3, 6.8 Hz, 1H), 4.66 (q, *J* = 4.0 Hz, 1H), 4.56 (ddd, *J* = 12.5, 10.1, 3.7 Hz, 1H), 4.31 – 4.24 (m, 2H), 4.25 – 4.14 (m, 2H), 3.90 (dd, *J* = 17.3, 7.5 Hz, 1H), 3.84 – 3.78 (m, 1H), 3.68 (dd, *J* = 8.3, 4.1 Hz, 1H), 3.39 (ddd, *J* = 17.2, 8.6, 4.8 Hz, 3H), 3.23 (dd, *J* = 14.9, 13.0 Hz, 1H), 3.14 – 3.00 (m, 2H), 2.94 (dd, *J* = 15.7, 4.3 Hz, 1H), 2.74 (dd, *J* = 10.8, 3.6 Hz, 1H), 2.23 (dd, *J* = 12.9, 7.0 Hz, 1H), 1.98 – 1.88 (m, 1H), 1.83 (td, *J* = 12.4, 3.5 Hz, 1H), 1.61 – 1.43 (m, 2H), 1.22 – 1.04 (m, 2H), 0.89 (d, *J* = 6.8 Hz, 3H), 0.85 – 0.75 (m, 10H).

^{13}C NMR (126 MHz, DMSO-d_6_HSQC) δ 122.8, 120.9, 119.0, 111.7, 68.9, 62.1, 59.3, 59.0, 58.3, 55.9, 53.7, 53.0, 51.1, 42.5, 41.9, 41.5, 38.6, 38.1, 36.0, 35.9, 34.7, 34.2, 34.0, 33.8, 30.0, 29.9, 25.3, 25.1, 15.8, 14.8, 12.2, 10.7

HRMS (ESI): m/z calculated: $C_{39}H_{55}N_{10}O_{10}S^+$, 855.3818, found 855.3843

HPLC-MS: Retention time R_t = 5.73 min (Gradient B)

Non-natural atropisomer: Ile^3-*S*-deoxo-amaninamide (**4b**)



4b

^{1}H NMR (500 MHz, DMSO-d_6) δ 11.48 (s, 1H), 9.05 (s, 1H), 8.46 (s, 1H), 8.28 (s, 3H), 7.71 (s, 1H), 7.54 (s, 1H), 7.33 – 7.26 (m, 2H), 7.14 (dd, *J* = 8.1, 6.9 Hz, 1H), 7.13 – 6.97 (m, 2H), 6.71 (s, 1H), 5.67 (s, 1H), 4.97 (td, *J* = 9.8, 3.0 Hz, 1H), 4.75 (m, 2H), 4.48 (dd, *J* = 16.8, 9.7 Hz, 2H), 4.42 (dd, *J* = 10.2, 3.3 Hz, 1H), 4.01 (s, 1H), 3.95 – 3.88 (m, 3H), 3.49 – 3.40 (m, 1H), 3.39 – 3.20 (m, 2H), 2.98 (d, *J* = 15.2 Hz, 2H), 2.24 (s, 1H), 1.91 (d, *J* = 15.2 Hz, 1H), 1.79 (s, 2H), 1.56 (s, 1H), 1.36 (s, 1H), 1.03 – 0.95 (m, 1H), 0.85 (d, *J* = 6.7Hz, 3H), 0.81-0.76 (m, 8H).

^{13}C NMR (176 MHz, DMSO-d_6_HMBC) δ 137.2, 128.4, 128.2, 119.4, 118.7, 111.1, 67.0, 57.3, 57.1, 47.1, 37.6, 35.4, 34.2, 31.7, 23.5, 14.7.

HRMS (ESI): m/z calculated: $C_{39}H_{55}N_{10}O_{10}S^+$, 855.3818, found 855.3837

HPLC-MS: Retention time R_t = 6.47 min (Gradient B)

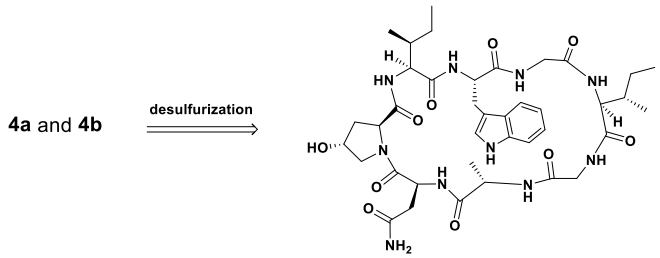
3.1.8.3 Amino acid analysis of **4a** and **4b** (Marfey's reagent)

The absolute configurations of the amino acid units in **4a** and **4b** were determined by the advanced Marfey's method. The hydrolysis products (6 N HCl, 110 °C, 12 h) of **4a** and **4b** were subjected to L-FDLA and L,D-FDLA derivatization and analyzed by LC-MS, which showed that amino acid residues of **4a** were identical with **4b** (see Supplementary Figure 4). Protocol: Isomers **4a** and **4b** (0.1 mg each) were hydrolyzed with stirring in 6 M HCl (200 µL) at 110 °C for 12 h. The residual HCl fumes were removed under a N_2 stream. Acid hydrolysates (suspended in 50 µL of H_2O) were treated with 1 M $NaHCO_3$ (20 µL) and then with L-FDLA and L,D-FDLA (100 µL of a 10 mg/mL solution in acetone), and the mixture was stirred at 37 °C for 1 h. The reaction was quenched with 1 M HCl (20 µL) and then diluted with MeOH for subsequent analysis.

3.1.8.4 Heating experiment of **4a** and **4b**

To investigate the interconversion of **4a** and **4a**, the heating experiment was performed (see Supplementary Figure 17). Protocol: After replacing the air with Argon, the NMR samples of **4a** and **4b** were heated to 150 °C for 10h. The proton NMR spectra were acquired after cooling down to room temperature.

3.1.8.5 Desulfurization of **4a** and **4b**



Raney nickel solution in water was added to the bicyclic peptide **4a** or **4b** dissolved in MeOH, which was then immediately sealed. The reaction was heated to 80 °C for 3h. The extent of desulfurization was monitored by LC-MS, the monocyclic product was obtained following HPLC purification. The results showed that identical product was obtained after desulfurization.

^{1}H NMR (700 MHz, DMSO-d_6) δ 10.71 (d, *J* = 2.4 Hz, 1H), 9.12 (t, *J* = 5.2 Hz, 1H), 8.62 (d, *J* = 8.3 Hz, 1H), 8.07 (s, 1H), 7.92 (t, *J* = 6.2 Hz, 1H), 7.66 (d, *J* = 7.9 Hz, 1H), 7.46 (d, *J* = 7.9 Hz, 1H), 7.42 (s, 1H), 7.35 (d, *J* = 8.0 Hz, 1H), 7.09 (d, *J* = 2.6 Hz, 1H), 7.07 (t, *J* = 7.4 Hz, 2H), 7.01 (t, *J* = 7.4 Hz, 1H), 6.95 (d, *J* = 8.9 Hz, 1H), 4.95 (t, *J* = 3.2 Hz, 1H), 4.45 – 4.39 (m, 2H), 4.38 – 4.33 (m, 1H), 4.35 – 4.29 (m, 1H), 4.24 (t, *J* = 7.6 Hz, 1H), 3.97 (dd, *J* = 10.8, 3.7 Hz, 1H), 3.90 – 3.84 (m, 1H), 3.87 – 3.81 (m, 1H), 3.70 – 3.67 (m, 1H), 3.65 (dd, *J* = 14.7, 4.2 Hz, 1H), 3.49 (dd, *J* = 16.4, 5.6 Hz, 1H), 3.47 – 3.44 (m, 1H), 3.28 – 3.22 (m, 2H), 2.67 (dd, *J* = 14.9, 10.3 Hz, 1H), 2.55 – 2.52 (m, 1H), 2.12 (d, *J* = 4.7 Hz, 1H), 2.00 – 1.87 (m, 2H), 1.58 – 1.53 (m, 1H), 1.47-1.37 (m, 2H), 1.30 (d, *J* = 7.2 Hz, 3H), 1.22 – 1.13 (m, 1H), 1.00 – 0.91 (m, 1H), 0.81-0.75 (m, 10H), 0.56 (d, *J* = 6.9 Hz, 3H).

^{13}C NMR (101 MHz, DMSO-d_6_HSQC) δ 123.5, 121.6, 121.4, 119.2, 118.9, 118.0, 117.8, 112.0, 111.8, 68.4, 61.7, 60.2, 59.8, 55.4, 55.2, 54.4, 48.2, 47.5, 44.3, 44.1, 43.7, 43.6, 39.1, 38.8, 38.4, 37.8, 34.0, 28.5, 28.2, 25.9, 24.2, 23.9, 18.3, 16.7, 16.5, 15.4, 15.2, 13.3, 13.1, 11.3, 11.1, 10.2.

HRMS (ESI): m/z calculated: $C_{39}H_{57}N_{10}O_{10}^{+}$, 825.4257, found 825.4258

HPLC-MS: Retention time R_t = 5.60 min (Gradient B)

3.1.9 CD spectroscopy

The structures of the isomers were evaluated using CD spectroscopy. The measurements were conducted on a J-815 CD spectrometer (Jasco, Groß-Umstadt, Germany). Lyophilized compounds **4a** and **4b** were dissolved in H_2O to reach a concentration of 75 µM. Far-UV spectra were acquired at 20°C 190-300 nm with a path length of 0.1 cm and a bandwidth of 1 nm at a continuous scanning speed of 50 nm/min

in 5 accumulations. The data pitch was set to 0.1 nm. The spectra were processed with Spectra Manager (JASCO) and the mean residue ellipticity (MRE) was calculated

$$MRE = \frac{\theta * 0.1}{l * c * n} \qquad (1)$$

Where θ is the ellipticity and l, c, and n denote the path length, molar concentration and number of amino acids.

3.2 Molecular dynamics simulations

We performed classical, all-atom molecular dynamics simulations for the conformational isomers **4a** and **4b** as well as the precursor **3b**. See table below for the summary of the simulations:

Supplementary Table 5. Overview of the MD simulations for the molecules **3b**, **4a** and **4b** in the named solvent with temperature *T* [K], number of replicas and total simulation time t_{sim} [µs]. For the convenience of the reader, a short description of each molecule is given.

Mol.	**Description**	**Solvent**	T [K]	**Replicas**	t_{sim} [µs]
4a	Derivative of the natural amanitin; tryptathionine bridge above of macrolactam	DMF	300 400	20 1	20 0.1
4b	Conformational isomer of **4a** with the tryptathionine bridge below the macrolactam	DMF	300 400	20 1	20 0.1
3b	Precursor with the cut between Hyp^2 and Ile^3; N-terminus: Ile^3, C-terminus: Hyp^2	DMF	300 400	20 1	20 0.1
3c	Precursor with the cut between Hyp^2 and Asn^1; N-terminus: Hyp^2, C-terminus: Asn^1 Protection group on Asn^1: Trt	DMF	300 400	20 1	20 0.1

3.2.1 Setup of the MD simulations

3.2.1.1 Parametrization of the peptides **4a**, **4b**, **3b** and **3c**

MD simulations for the peptides **4a**, **4b**, **3b** and **3c** were performed with GROMACS 2019.4 [12] simulation package in dimethylformamide (DMF). Both solvent and solute were built with the molecule editor Avogadro [13] and parametrized with ACPYPE [14] setting the molecule's charge *n*=0 and referring to AMBER14SB[15] force field with *a*='amber'. The C- and N-terminus in the precursors **3b** and **3c** were capped with methyl-groups, yielding -CO-OMe and -NH-Me as end groups.

3.2.1.2 Parameters for MD simulation of the peptides **4a**, **4b**, **3b** and **3c**

After their construction with Avogadro [13] and parametrization with ACPYPE [14], the structures of the peptides were solvated in a cubic simulation box (average volume 50 nm^3) using the GROMACS commands 'gmx editconf', 'gmx solvate' and the parametrized DMF molecule as template. The distance of the solute to the box walls was set to 1 nm. On average, 389 solvent molecules were added.

The solvated systems were energy-minimized using the steepest-descent algorithm (emtol=1000 kJ/(mol·nm), nsteps=5000), followed by *NVT* and *NpT* equilibrations at temperature *T*=400 K for 400 ps and 600 ps,

respectively. In both equilibration steps, periodic boundary conditions were applied in all directions and position restraints were used for the peptides.

As pre-production run, the systems were propagated using leap-frog integration [16] with an integration time step dt = 2 fs. All covalent bonds were constrained using LINCS algorithm [17] ('all-bonds', iter = 4, order = 6). Periodic boundary conditions were applied in all three directions. The simulations were conducted in the *NpT* ensemble at temperature T=400 K (velocity-rescale thermostat [18], coupling time τ_T = 0.1 ps) and with pressure p=1 bar (Parrinello-Rahman barostat [19], coupling time τ_p = 2 ps). For Van-der-Waals interactions, Verlet-cutoff scheme [20] was applied using a cut-off radius of r_{vdw}=1 nm and updating the neighborlist every nstlist = 10 integration timesteps. For Coulomb interactions, Particle-Mesh-Ewald algorithm [21] (pme-order = 6, Fourier grid spacing = 0.12 nm, cut-off for short-range electrostatic interaction $r_{Coulomb}$ = 1 nm). The coordinates of all atoms were written to file every 100 ps over a simulation length of 0.1 µs.

To generate starting structures for the production run, we calculated the RMSD of all main chain atoms and all carbon atoms closest to the main chain (within one covalent bond) to the first frame of the simulation (after least-square fit on the backbone) for each prerun using 'gmx rms'. The resulting RMSD distribution was split into 20 equally sized intervals. Out of each interval one starting structure was taken randomly.

These starting structures containing the respective peptide as well as the solvent, were then equilibrated under *NVT* and *NpT* conditions at temperature T=300 K for 200 ps and 300 ps, respectively. In both equilibration steps, periodic boundary conditions were applied in all directions and position restraints were used on the peptide.

Subsequently, the systems were simulated under the same conditions as described for the pre-runs except for the following changes: The temperature was set to T = 300 K. Only the solute's coordinates were written to file every 1 ps. For each structure taken from the prerun, one trajectory with a simulation length of 1 µs was produced, yielding in 20 µs of total simulation time per peptide (see Supplementary Table 5).

3.2.2 Analyses

The analyses performed for the peptides **4a**, **4b**, **3b** and **3c** are described below. To make the visual analysis of the trajectories easier, we centered the peptides in the box in each frame (GROMACS 2019.4 'gmx trjconv' with option '-pbc mol -center') and applied a translational and rotational fit, which included a least-square fit on the C_α atoms ('-fit rot+trans').

3.2.2.1 Plane angle analysis

The structures of the peptides can be described as a macrolactam that consists of two half cyclic subunits, cycle A and B, with cycle A being Cys^8-Asn^1-Hyp^2-Ile^3-Trp^4 and cycle B being Trp^4-Gly^5-Ile^6-Gly^7-Cys^8. Both subunits share Trp^4 and Cys^8 as common amino acids and they are additionally linked via the tryptathionine bridge, which consists of the side chains of Trp^4 and Cys^8 (see Supplementary Figures 3, 12, 20).

For the evaluation of the positioning of the tryptathionine bridge in relation to the macrolactam, a simplified model was created. Based on the C_α atoms of Trp^4 and Cys^8 as well as reference points out of each subunit, three planes were defined each of which represents one subunit of the peptide:

Plane E_A representing cycle A, is defined on C_α (Trp^4), C_α (Cys^8) and C_α (Hyp^2).

Plane E_B representing cycle B, is defined on C_α (Trp^4), C_α (Cys^8) and C_α (Ile^6).

Plane E_C representing the tryptathionine bridge, is defined on C_α (Trp^4), C_α (Cys^8) and C_δ (Trp^4), where C_δ (Trp^4) is the carbon atom directly linked to the sulfur atom of Cys^8.

The positioning of the tryptathione bridge was judged based on the angles between the planes E_A, E_B and E_C in **clockwise** rotation (see Figure 2c and Supplementary Figures 3, 20). In the following, the mathematical proceeding is described:

First, the position $\mathbf{r}_0 = (x_0, y_0, z_0)^T$ of C_α (Cys^8) was defined as reference point for all three planes. All other atom positions $\mathbf{r}_{0,i} = (x_{0,i}, y_{0,i}, z_{0,i})^T$ were described in relation to $\mathbf{r}_0(C_{\alpha,Cys8})$:

$$\mathbf{r'}_i = \mathbf{r}_{0,i} - \mathbf{r}_0(C_{\alpha,Cys8}). \tag{1}$$

The resulting positions $\mathbf{r'}_i$ were used to define the normal vectors of the planes E_A, E_B and E_C:

$$\begin{aligned} E_A&: \quad \mathbf{n}_A = \mathbf{r'}_i(C_{\alpha,Trp4}) \times \mathbf{r'}_i(C_{\alpha,Hyp2}), \\ E_B&: \quad \mathbf{n}_B = \mathbf{r'}_i(C_{\alpha,Trp4}) \times \mathbf{r'}_i(C_{\alpha,Ile6}), \\ E_C&: \quad \mathbf{n}_C = \mathbf{r'}_i(C_{\alpha,Trp4}) \times \mathbf{r'}_i(C_{\delta,Trp4}). \end{aligned} \tag{2}$$

The scalar product of two normal vectors $\mathbf{n}_k$ and $\mathbf{n}_l$ yields the cosine of the angle between the planes E_k and E_l:

$$\cos(\theta_{k,l}) = (\mathbf{n}_k \cdot \mathbf{n}_l) / (|\mathbf{n}_k| \cdot |\mathbf{n}_l|) = {}^x/_y \tag{3}$$

out of which $\theta_{k,l}$ can be calculated using the arccos(x) function and defining limiting cases:

$$\begin{aligned} \theta_{k,l} &= \arccos({}^x/_y) \quad \text{if } 0 \le \theta_{k,l} < \pi, \\ \theta_{k,l} &= 2\pi - \arccos({}^x/_y) \quad \text{if } \pi \le \theta_{k,l} < 2\pi. \end{aligned} \tag{4}$$

with ${}^x/_y$ as defined in eq. (2). k and l represent the circular subunits A, B or C.

All atom positions were extracted from the trajectories using the feature ‘mdtraj.load‘of the python library mdtraj 1.9.3.[15] The normal vectors and angles between the planes were computed using the functions linalg.norm(), dot(), cross() and arccos().[18]

3.2.2.2 Distance calculations

For two atoms *i* and *j*, the Euclidean distance was calculated as follows:

$$\mathbf{d}_{i,j} = | \mathbf{r}_j - \mathbf{r}_i | = [(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2]^{1/2} \quad (5)$$

with $\mathbf{r}_i = (x_i, y_i, z_i)^T$ as the atom position of atom *i* extracted from the MD trajectories using the feature 'mdtraj.load' of the python library mdtraj 1.9.3. [7]

3.2.2.3 Hydrogen bonds

Population analyses for single hydrogen bonds

Hydrogen bonds were analyzed with GROMACS 2019.4[10a] simulation package using 'gmx hbond'. After assigning each atom to either the side or the main chain of its residue with ‚gmx make_ndx', all interactions were classified as side ('s) or main ('m') chain interactions of the involved residues. The hydrogen bonds were evaluated based on their relative occurrence over the total simulation length. Hydrogen bonds with a relative occurrence greater than 50% in at least one simulation were added to a list 'hbond_best'. For the hydrogen bonds of this list, the population average over all replicas of the respective peptide was calculated. Hydrogen bonds of 'hbond_best' with a population of at least 10% were considered significant.

Population analyses for hydrogen bond combinations

The time series of the hydrogen bonds only contain '0' and '1' as values: '1' if the respective hydrogen bond is present, '0' if not. Based on the time series of the individual hydrogen bonds, the populations of all possible combinations of them can be achieved by:

$$X(t) = \Sigma^{N_k} \; 2^k \, t_k(i) \quad (6)$$

where *i* denotes the *i*-th time step of an MD trajectory *X* that is assigned to the states $N_s = \Sigma_k 2^k$ with *N* as the number of different hydrogen bonds. $t_k(i)$ denotes the time series of hydrogen bond *k* at time step *i*.

For **4a**, the time series of the following hydrogen bonds were considered:

A : Asn^1(m) – Gly^5(m),

B : Ile^3(m) – Asn^1(s),

C : Trp^4(m) – Asn^1(m),

D : Gly^5(m) – Asn^1(m),

E : Cys^8(m) – Gly^5(m),

yielding 31 different combinations according to eq. (6):

H-Bonds	A	B	C	D	E
N_S	k_0	k_1	k_2	k_3	k_4
0	0	0	0	0	0
1	1	0	0	0	0
2	0	1	0	0	0
3	1	1	0	0	0
4	0	0	1	0	0
5	1	0	1	0	0
6	0	1	1	0	0
7	1	1	1	0	0
8	0	0	0	1	0
9	1	0	0	1	0
10	0	1	0	1	0
11	1	1	0	1	0
12	0	0	1	1	0
13	1	0	1	1	0
14	0	1	1	1	0
15	1	1	1	1	0
16	0	0	0	0	1
17	1	0	0	0	1
18	0	1	0	0	1
19	1	1	0	0	1
20	0	0	1	0	1
21	1	0	1	0	1
22	0	1	1	0	1
23	1	1	1	0	1
24	0	0	0	1	1
25	1	0	0	1	1
26	0	1	0	1	1
27	1	1	0	1	1
28	0	0	1	1	1
29	1	0	1	1	1
30	0	1	1	1	1
31	1	1	1	1	1

For **4b**, the time series of the following hydrogen bonds were considered:

A : ‘Ile^{6}(m)-Ile^{3}(m)’

B : ‘Gly^{7}(m)-Ile^{3}(m)’

C : ‘Cys^{8}(m)-Ile^{3}(m)’

According to eq. (6) (N = 3), there are 7 different combinations for A, B and C:

H-Bonds	A	B	C
N_S	k_0	k_1	k_2
0	0	0	0
1	1	0	0
2	0	1	0
3	1	1	0
4	0	0	1
5	1	0	1
6	0	1	1
7	1	1	1

For **4a** and **4b**, the populations of the hydrogen bond combinations were then analyzed with reference to the trajectory length (1 µs per trajectory).

3.2.2.4 Linear correlations between time series of observables

For the investigation of the linear correlation between two variables, the Pearson correlation coefficients were calculated using the function corrcoef() of the numpy library.[22]

3.2.2.5 RMSD / RMSF calculations

RMSD and RMSF calculations were done with GROMACS 2019.4 [12a] simulation package using 'gmx rms' and 'gmx rmsf'. Below, the exact input and reference structure selections are explained for the different peptides:

4a and **4b**

The trajectories were fitted (translation and rotation) to the respective crystal structure using 'gmx trjconv -fit rot+trans'. Afterwards, the RMSF was calculated on all atoms towards the crystal structures. Please note, for this step, the crystal structure file and the structure file of the simulations had to be brought into line with each other.

The RMSD was calculated i) on all atoms and ii) on all main chain and carbon atoms closest to the main chain (within one covalent bond) were calculated. For both RMSD calculations, the least-squares fit was applied on the backbone atoms.

3b and **3c**

The trajectories were fitted ('gmx trjconv -fit rot+trans') to the starting structure of the respective simulation focusing on the cycle B (see Figure 2b and Supplementary Figures 3, 20), i.e. the cycle of the molecule that was already formed. As a result, all frames of a trajectory show the same orientation of that cycle in space. The RMSF was calculated on all atoms towards the highest-probability structure of the MD ensemble

(see 3.2.2.9). The RMSD was calculated on i) all atoms and ii) on the existing ring only. In both cases, the least-squares fit was applied on the existing ring.

Indices

To compare the RMSF of the compounds **4a**, **4b**, **3b** and **3c**, the atoms of the structures were re-organized such that the order of the first 114 atoms (excluding caps) was identical in all structure files. The resulting structure files are published on Zenodo (DOI: 10.5281/zenodo.6974777) [23].

3.2.2.6 Backbone and dihedral angles

Time series of dihedral angles were extracted from the simulation data with GROMACS 2019.4 [10a] simulation package using 'gmx gangle'. The backbone angles backbone angles (φ,ψ) of a residue *i* were defined as follows:

φ: $\angle$ (C_{i-1}, N_i, $C_{\alpha,i}$, C_i)

ψ: $\angle$ (N_i, $C_{\alpha,i}$, C_i, N_{i+1})

For non-glycine residues, the dihedral angles of the side chain towards the main chain were defined as:

χ: $\angle$ (N_i, $C_{\alpha,i}$, $C_{\beta,i}$, $C_{\gamma,i}$)

3.2.2.7 Solvent accessible surface area

The solvent accessible surface area of **4b** was calculated with 'md.shrake_rupley' of the python library mdtraj 1.9.3.[7] ,md.shrake_rupley' is a new implementation of an algorithm from Shrake and Rupley from 1973.[24]

As input, the atom indices of the following amide functions (N,H) were considered:

(Residue: N, H)

Asn1: 1, 10

Ile3: 30, 39

Trp4: 49, 68

Gly5: 72, 77

Ile6: 79, 88

Gly7: 98, 103

Cys8: 105, 112

3.2.2.8 Ground state energies

All quantum-chemical calculations were performed using ORCA version 5.0.0[25]. Using Grimme's GFN2-xTB method[26] interfaced with ORCA, single-point energies were calculated for all frames of the MD simulation trajectories of **4a** and **4b**, respectively. The 100 structures with lowest energy were then optimized at the B3LYP-D4(BJ)/def2-TZVP[27] level of density functional theory adding implicit solvation in

water using CPCM[28] (ε=78.4) and using the RIJCOSX approximation[29] ('RIJCOSX def2/J') to accelerate the SCF calculations. For the resulting optimized structures, full TDDFT calculations were performed using the setup of the optimization ('B3LYP DEF2-TZVP D4 CPCM(WATER); 'RIJCOS def2/J'). 25 states were calculated.

3.2.2.9 Extraction of highest-probability structures

Using the hydrogen bond time series described in 3.2.2.3, structures were extracted based on the most-populated, positively correlated hydrogen bonds (Supplementary Figure 2). For **3b**, these hydrogen bonds were (population on average over 20 µs is given in brackets): 'Asn^1(m)-Trp^4(m)' (60%), 'Trp^4(m)-Asn^1(m)' (37%), 'Gly^7(m)-Trp^4(m)' (51%) and 'Cys^8(m)–Trp^4(m)' (65%). For **3c**, these hydrogen bonds were 'Trp^4(m)-Ile^6(m)' (43%), 'Gly^5(m)-Ile^6(m)' (35%) and 'Cys^8(m)-Hyp^2(m)' (43%).

Hence, for **3b** and **3c**, all frames were extracted, in which the respective hydrogen bonds were all present. The resulting sub-data sets comprised 28% (**3b**) and 31% (**3c**) of the respective entire data sets. Both sub-data sets were checked based on the RMSD (see 3.2.2.5) of the backbone from Trp^4 to Asn^1 with a preceding least-squares fit on the existing ring, only. The average RMSD of the sub-data sets amounted to 0.05 nm (**3b**) and 0.03 nm (**3c**). Consequently, the sub-data sets were considered unified and representative of the highest-probability structures of **3b** and **3c**.

The highest probability structures are publicly available on Zenodo (DOI: 10.5281/zenodo.6974777) [23].

3.2.2.10 Determining transition times (solely for **4a**):

The time for the transition between the starting structure and the crystal structure was estimated as follows: The trajectories were discretized into three states according to 3.2.2.3: state '1' comprises the hydrogen bond set that is in agreement with the crystal structure ('Asn^1(m)-Gly^5(m)', 'Ile^3(m)-Asn^1(s)', 'Trp^4(m)-Asn^1(m)', 'Gly^5(m)-Asn^1(m)', 'Cys^8(m)-Gly^5(m)'). State '2' comprises hydrogen bonds that are mutually exclusive to hydrogen bond set '1' ('Asn^1(m)-Hyp^2(m)', 'Cys^8(m)-Hyp^2(m)'). All frames that could not be unambiguously assigned to '1' or '2' were considered as noise ('0', Fig. S11a). These frames denote short-term transitions out of the two conformations. The transition time is the moment when the system reaches state '1' if the system was in '2' before and hasn't visited '1' by that time. Based on the discretized trajectories, the transition times were calculated using python (pseudo-code):

```
> transition = []
>
># Loop over all replicas:
>
>for rep in range(20):
>
>    time = np.arange(0,len(traj),1)
```

```
>   here_set1 = np.where(traj==1)[0]
>   here_set2 = np.where(traj==2)[0]
    # If both states exist:
>   if ((np.size(here_set1)>0) and (np.size(here_set2)>0)):
            m1 = np.min(here_set1)
            m2 = np.max(here_set2)
            if m1 < m2:
                    break
            else:
                    transition.append(m1)
>   else:
            transition.append([ ])
```

3.2.2.11 Comparison between NMR and MD based on NOE distances

Interatomic distances between protons *a* and *b* were calculated with 'gmx distance -oall' using GROMACS 2020.6 simulation package [12,30] and a list of atom indices from NMR experiments.

Considering the time series of the inter-proton distance $r_{a,b}(t)$, the following averages were calculated for each replica:[31,32]

$$R^{MD}_{a,b} = \langle z(t) \rangle^{1/6} = [\ (1/N)\ \Sigma^{N}_{f=1}\ z_f\ ]^{1/6} \quad (7)$$

$$z(t) = (\ r_{a,b}(t)\ )^{-6} \quad (8)$$

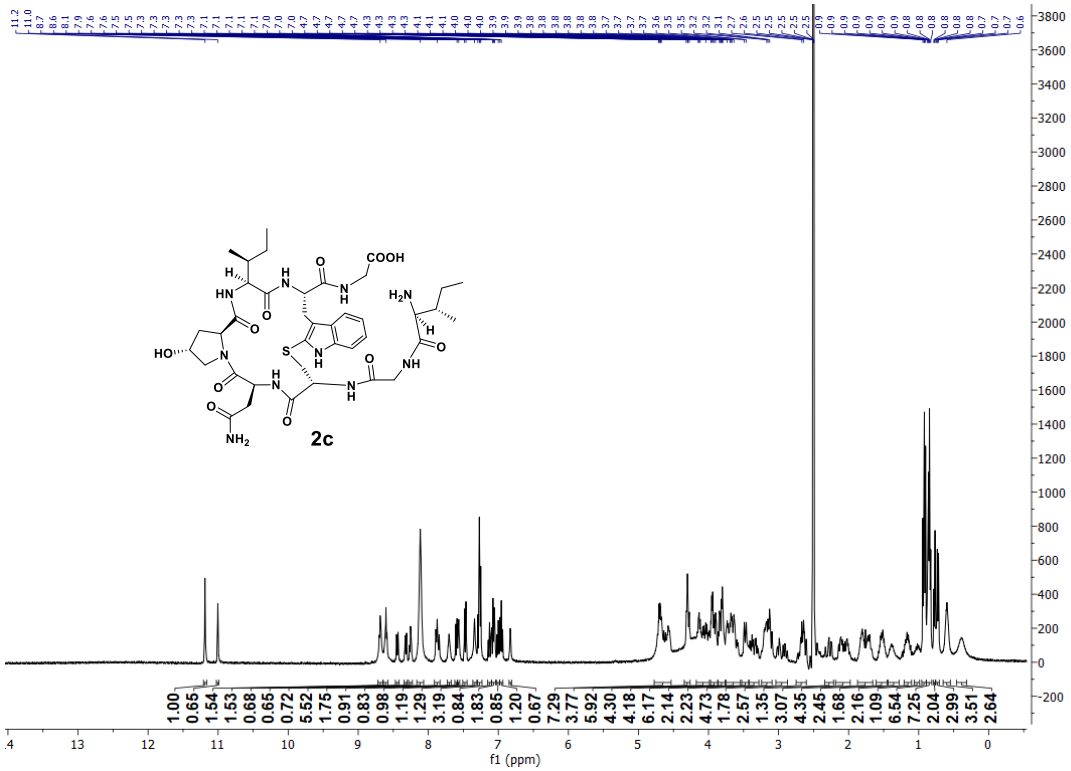
with $N=10^6$ as the number of structures *f* in each replica. The deviations from experimental inter-proton distances (NMR NOE and crystal structure) were then determined by:

$$\Delta(R_{a,b}) = R^{MD}_{a,b} - R^{NMR}_{a,b} \quad (9)$$

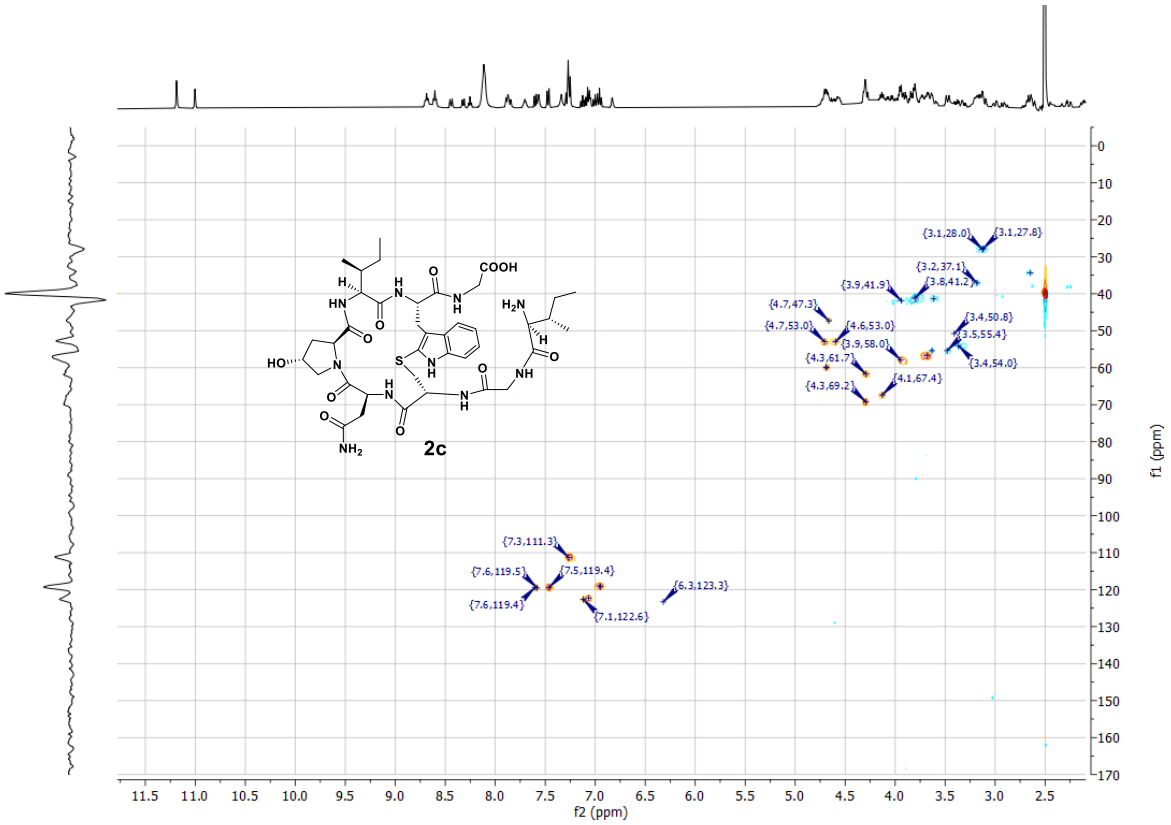
calling $\Delta(R_{a,b})$ 'NOE distance violation'. Any distance with $\Delta(R_{a,b}) > 0.1$ nm was considered 'violated', i.e. the MD trajectory deviates from the NMR ensemble in this distance beyond the experimental uncertainty.

The distribution of these distance violations over all replicas (n=20, 20 µs simulation time in total) was visualized as box-and-whisker plots using 'pyplot.boxplot' from the python library 'matplotlib' (version 3.5.3).

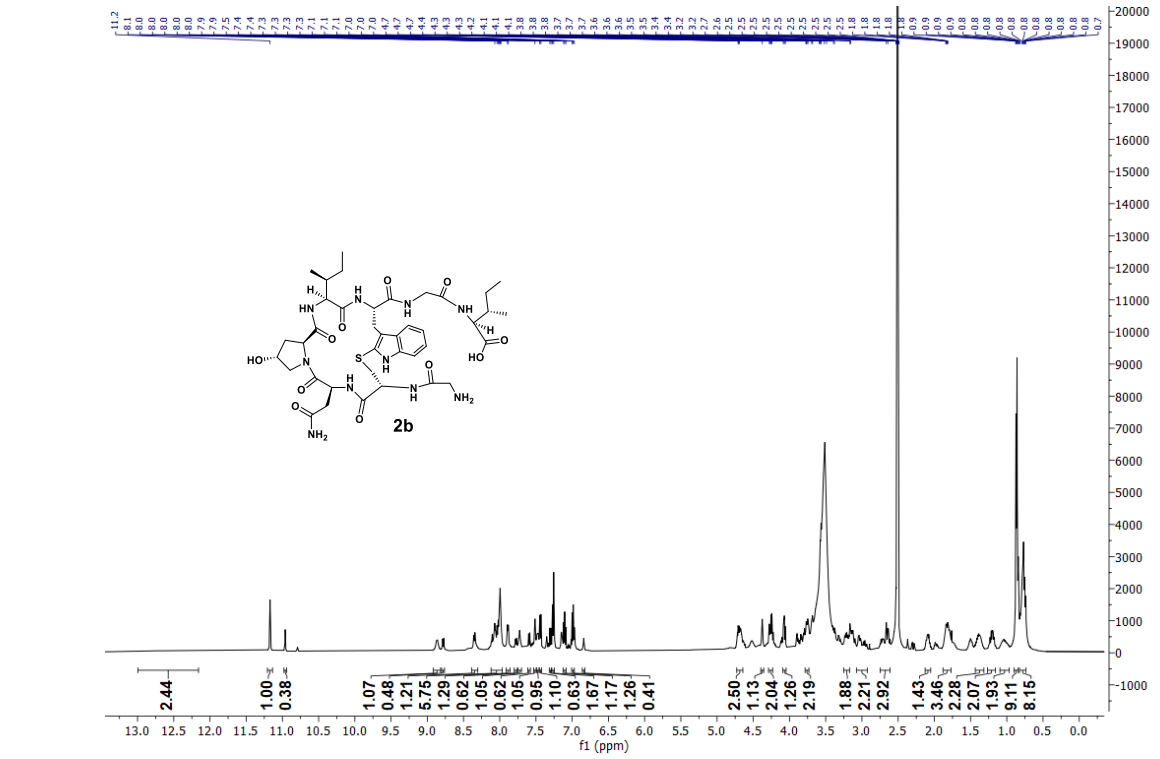
3.3 NMR spectra



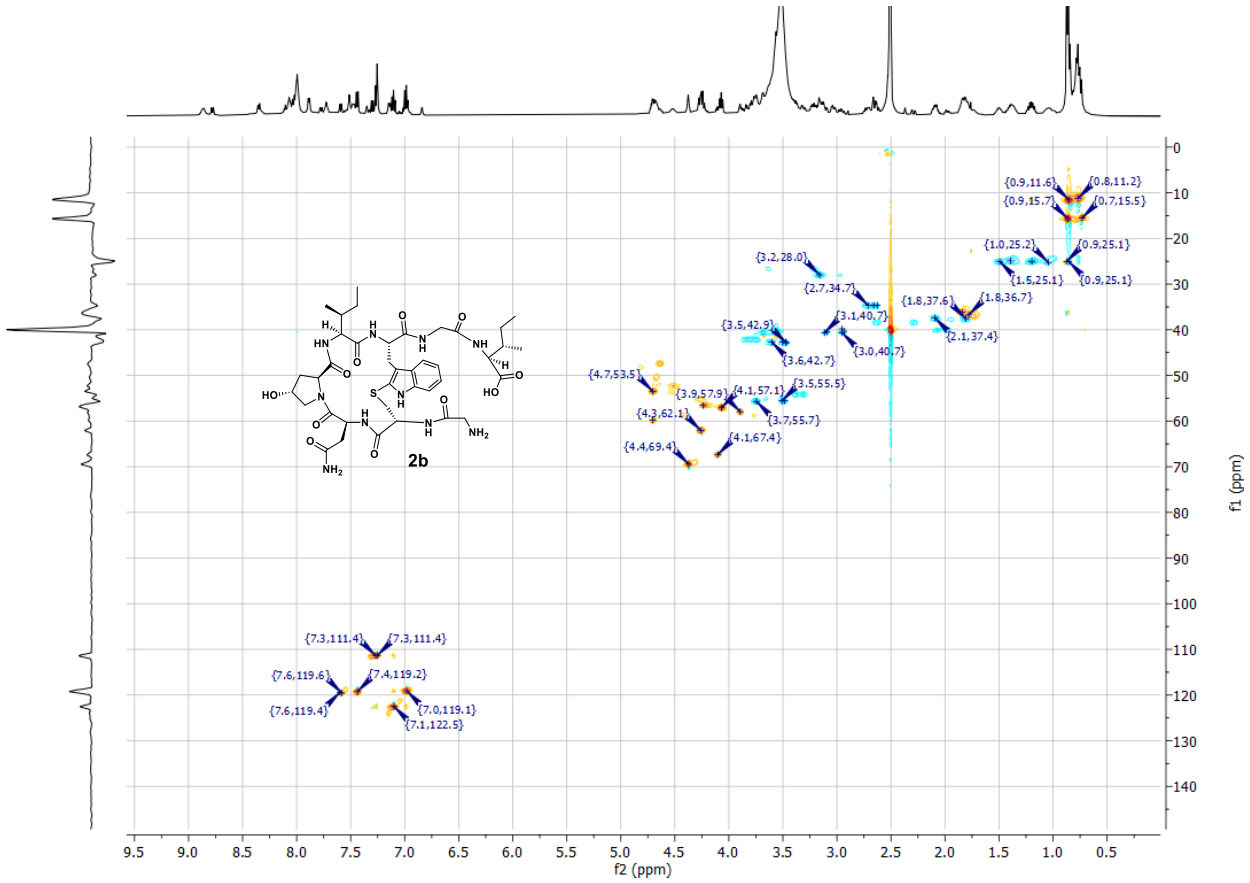
Supplementary Figure 23. ^{1}H NMR Spectrum of peptide **2c**



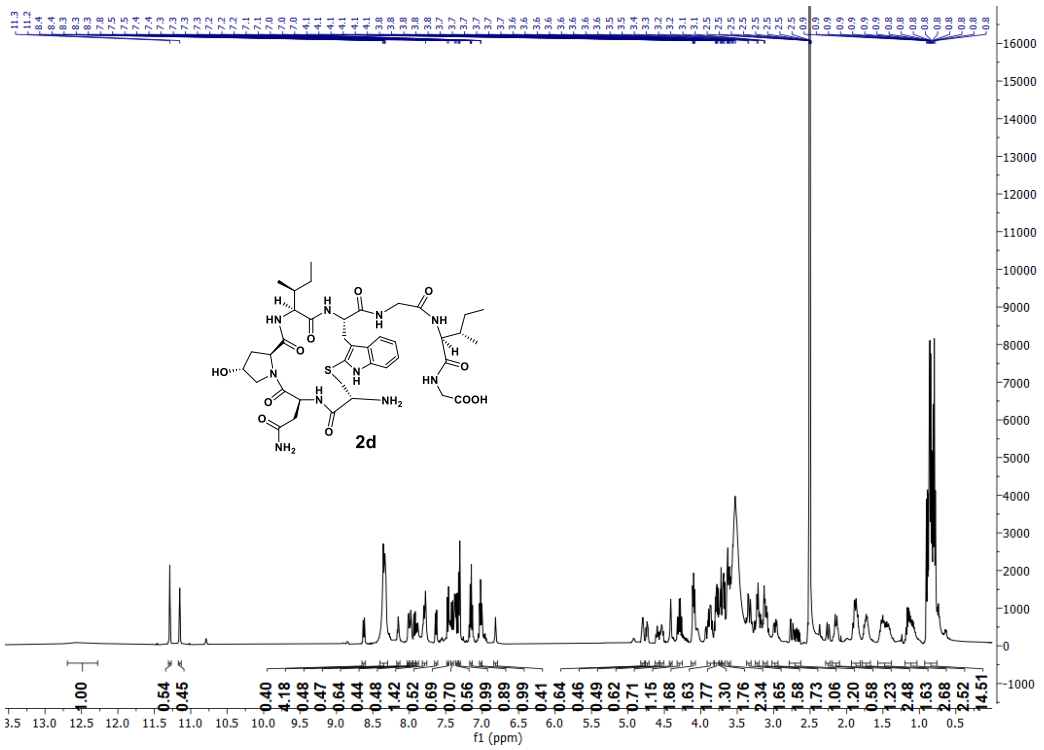
Supplementary Figure 24. HSQC-ed Spectrum of peptide **2c**



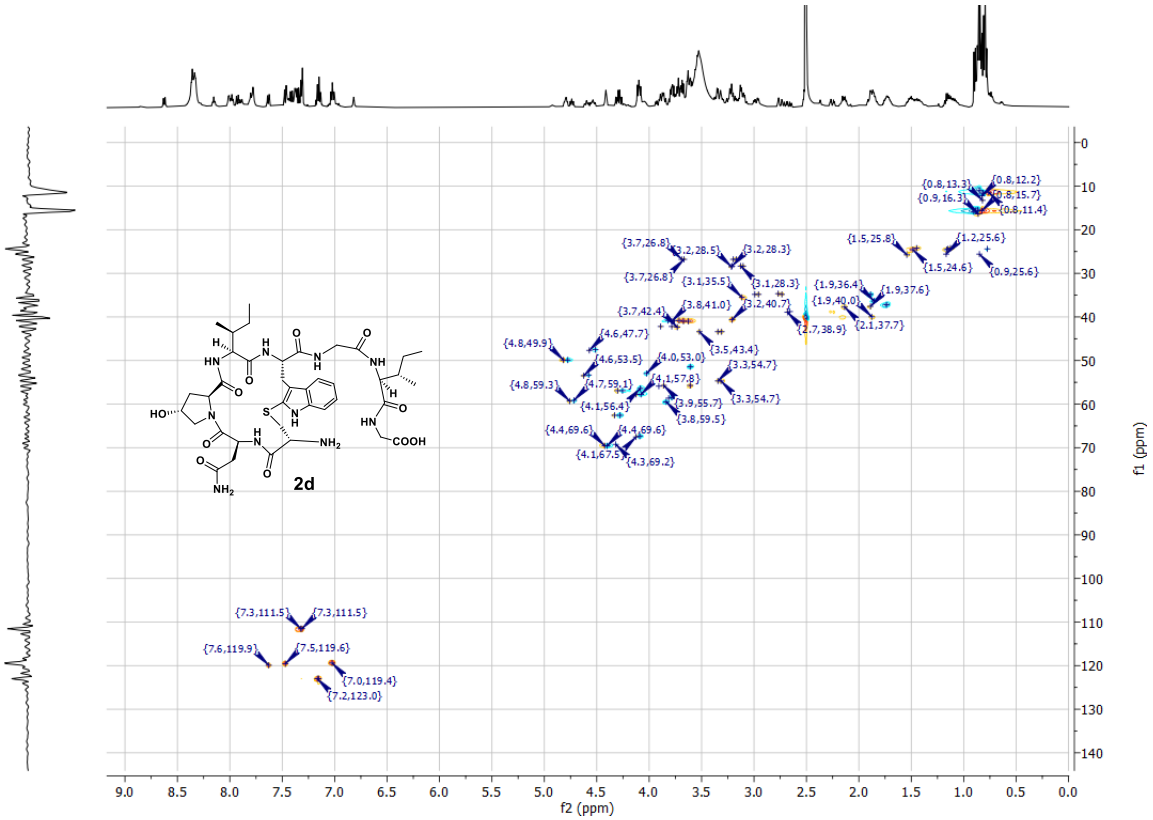
Supplementary Figure 25. ^{1}H NMR Spectrum of peptide **2b**



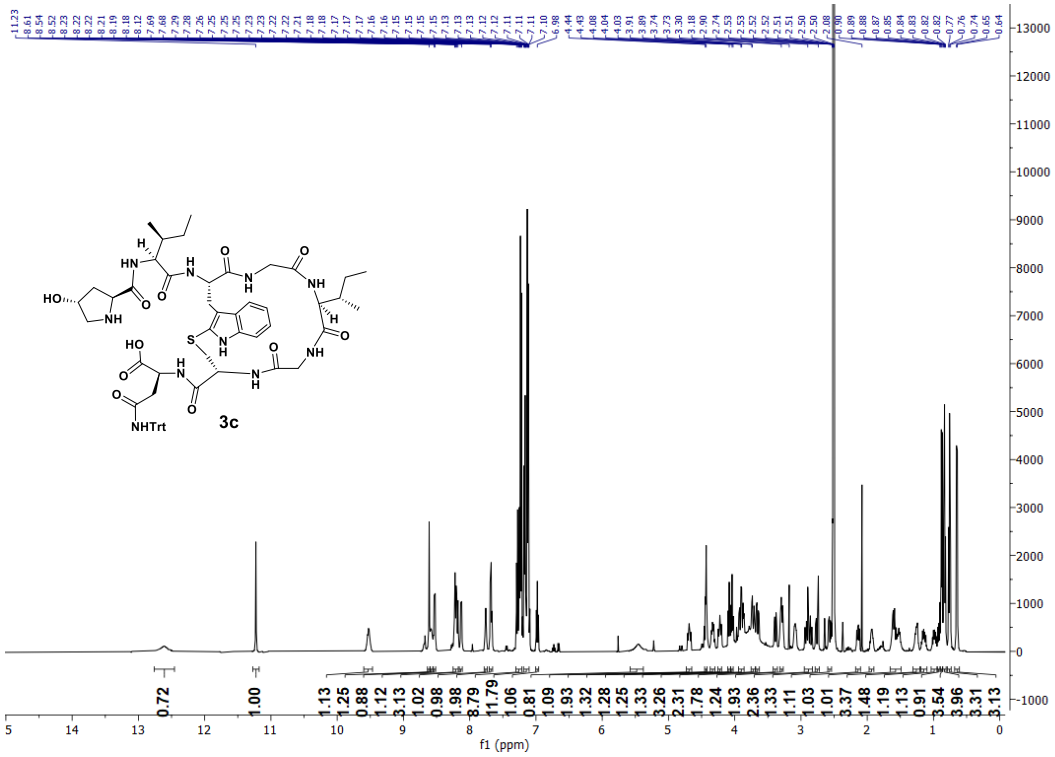
Supplementary Figure 26. HSQC-ed Spectrum of peptide **2b**



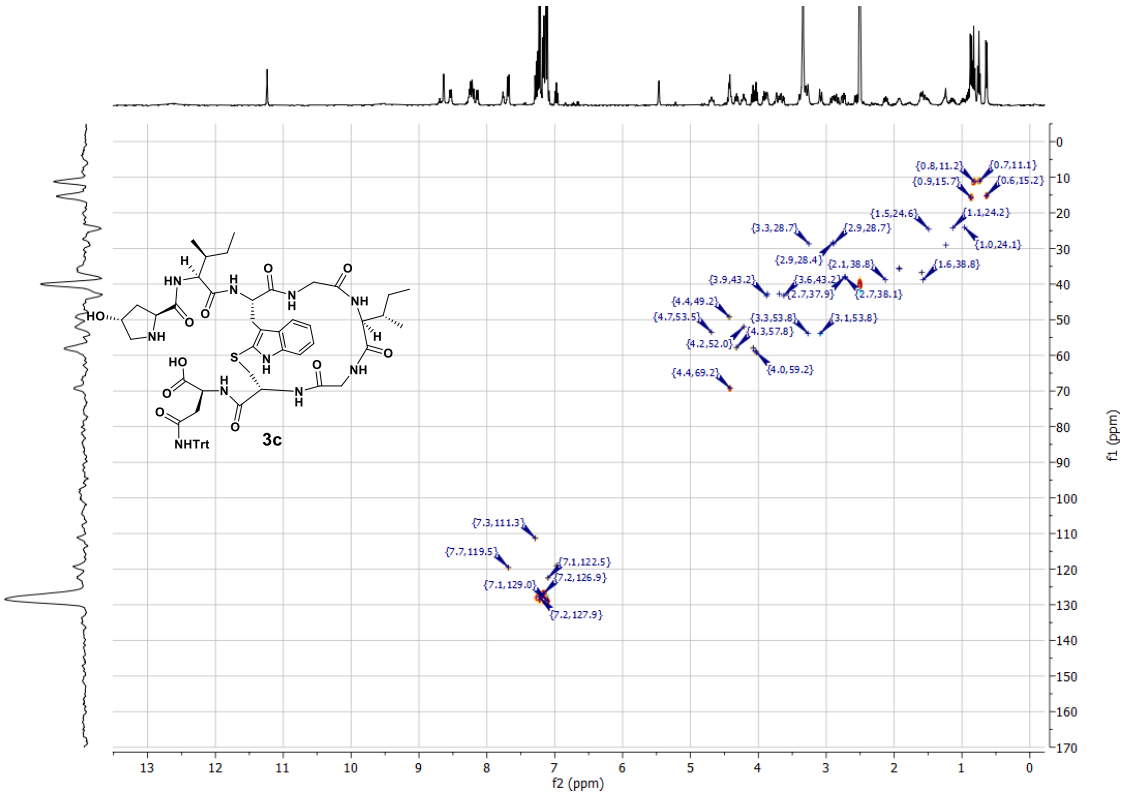
Supplementary Figure 27. ^{1}H NMR Spectrum of peptide **2d**



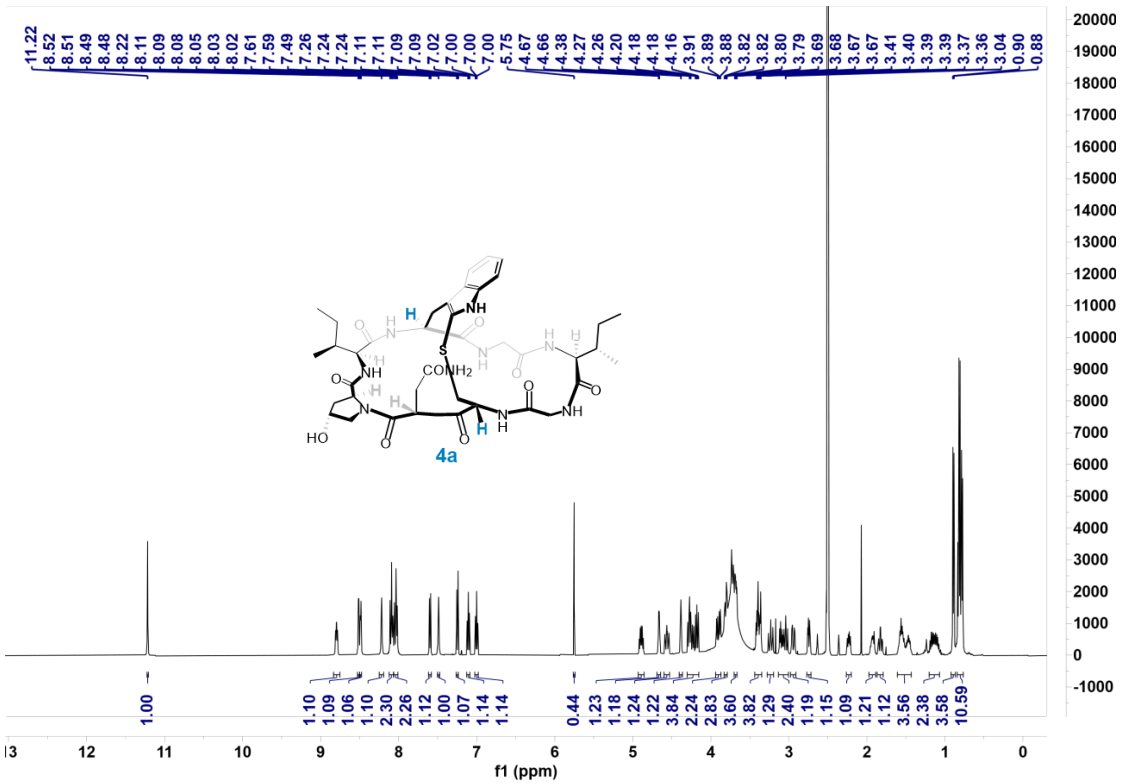
Supplementary Figure 28. HSQC-ed Spectrum of peptide **2d**



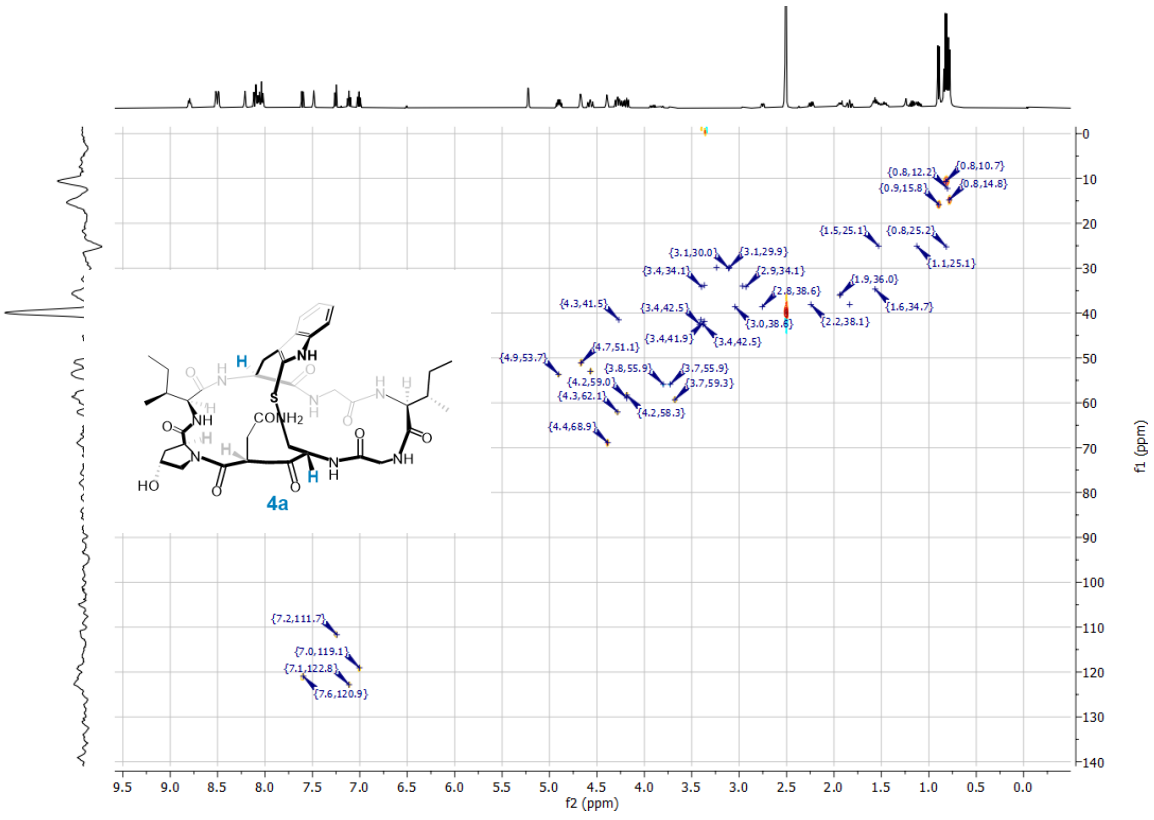
Supplementary Figure 29. ^{1}H NMR Spectrum of peptide **3c**



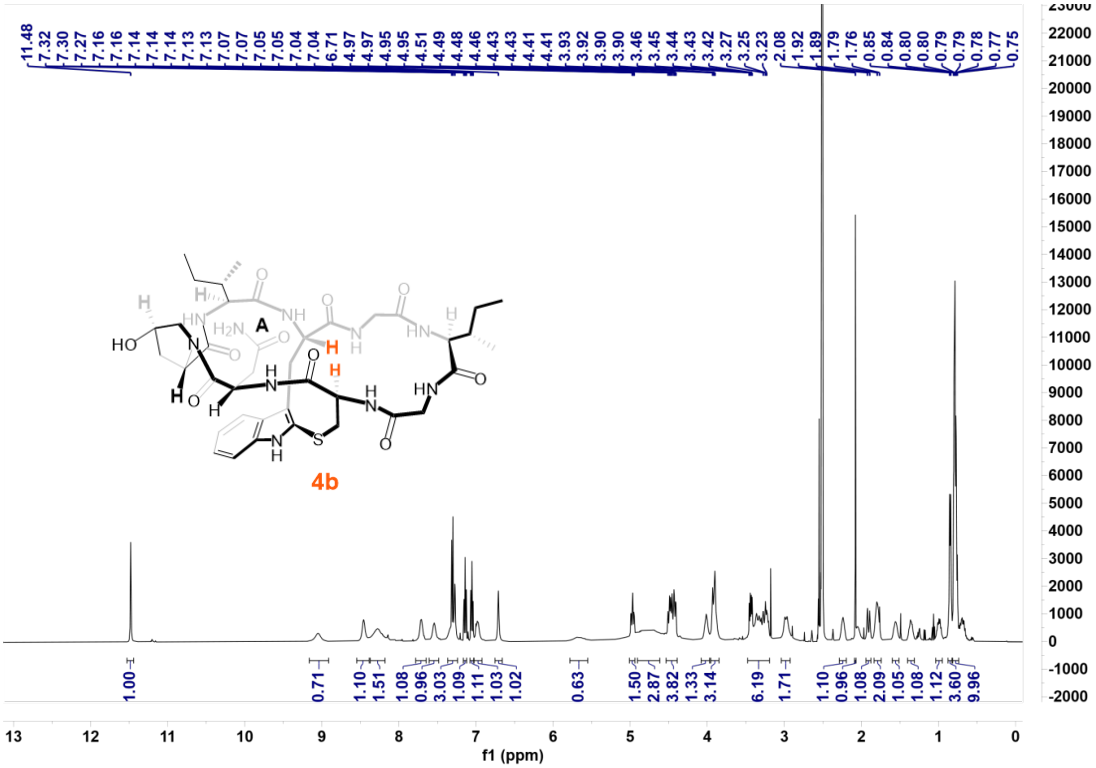
Supplementary Figure 30. HSQC-ed Spectrum of peptide **3c**



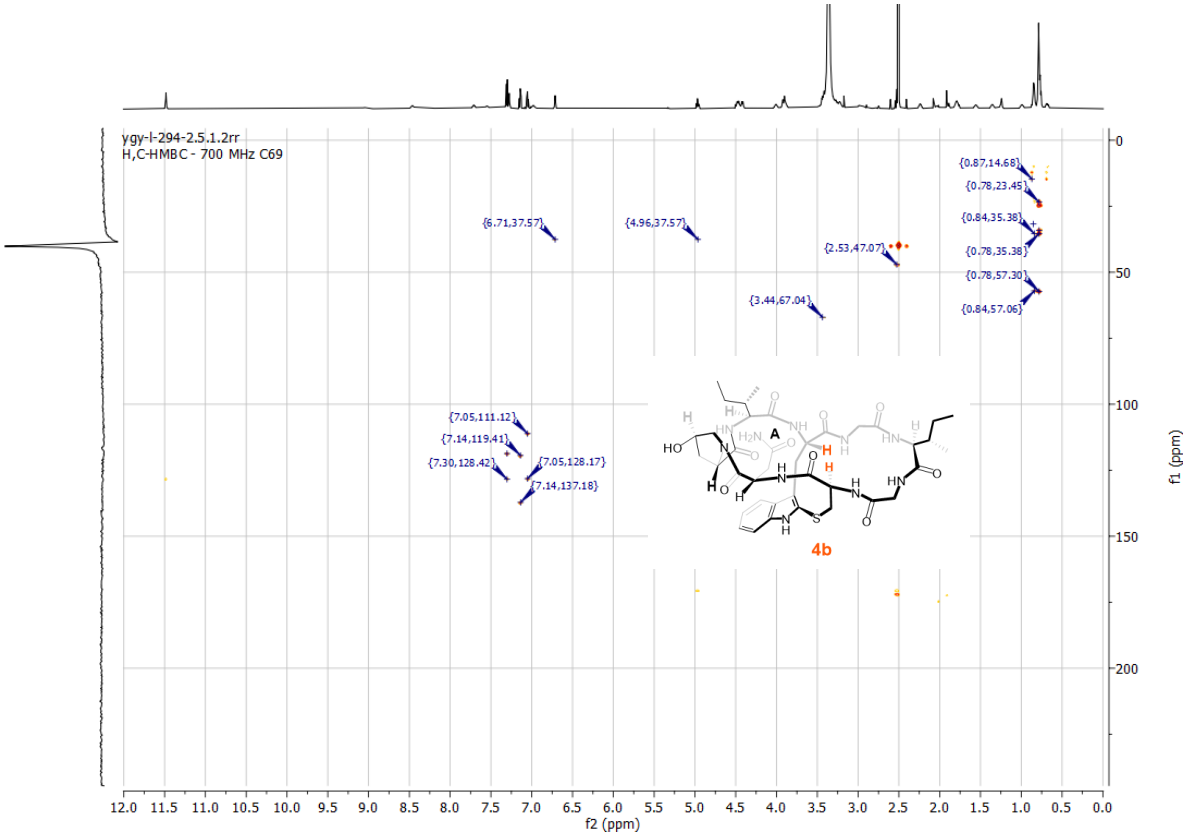
Supplementary Figure 31. ^{1}H NMR Spectrum of peptide **4a**



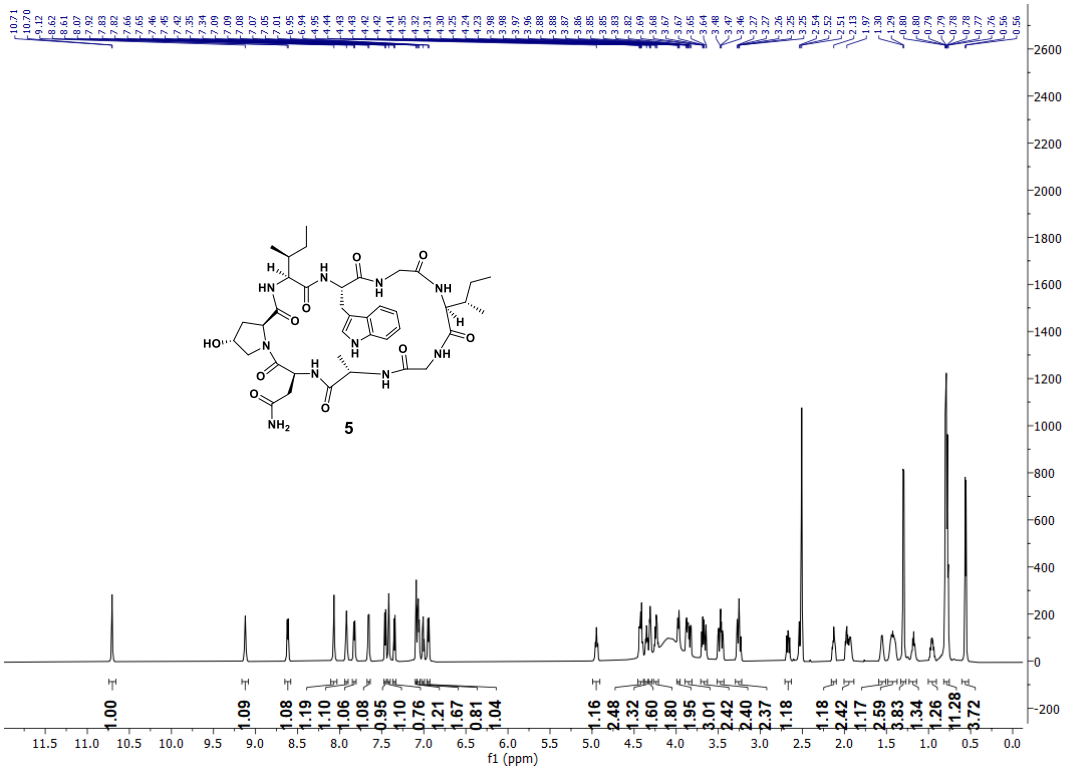
Supplementary Figure 32. HSQC-ed Spectrum of peptide **4a**



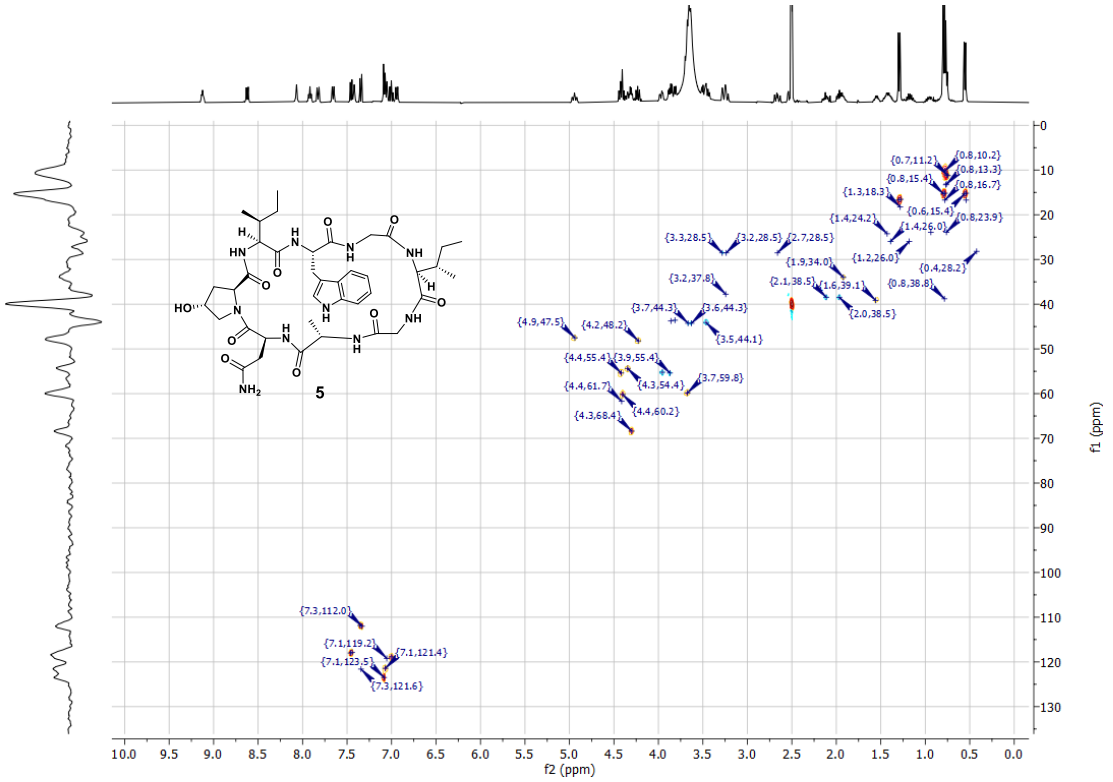
Supplementary Figure 33. ^{1}H NMR Spectrum of peptide **4b**



Supplementary Figure 34. HSQC-ed Spectrum of peptide **4b**



Supplementary Figure 35. ^{1}H NMR Spectrum of peptide **5**



Supplementary Figure 36. HSQC-ed Spectrum of peptide **5**

3.4 X-Ray

Supplementary Table 6: Crystal data and structure refinement for cu-4372.

Identification code	cu-4372	
Empirical formula	C39 H62 N10 O14 S	
Formula weight	927.05	
Temperature	150.01(10) K	
Wavelength	1.54184 Å	
Crystal system	Trigonal	
Space group	R3 (No.146)	
Unit cell dimensions	a = 32.9063(5) Å	a= 90°.
	b = 32.9063(5) Å	b= 90°.
	c = 11.7975(2) Å	g = 120°.
Volume	11063.1(3) $Å^3$	
Z	9	
Density (calculated)	1.252 Mg/m^3	
Absorption coefficient	1.180 mm^{-1}	
F(000)	4446	
Crystal size	0.21 x 0.13 x 0.12 mm^3	
Theta range for data collection	2.69 to 67.44°.	
Index ranges	-39<=h<=38, -38<=k<=39, -14<=l<=13	
Reflections collected	14947	
Independent reflections	6962 [R(int) = 0.0270]	
Completeness to theta = 67.44°	99.9 %	
Absorption correction	Semi-empirical from equivalents	
Max. and min. transmission	0.8714 and 0.7897	

Refinement method	Full-matrix least-squares on F^2
Data / restraints / parameters	6962 / 39 / 621
Goodness-of-fit on F^2	1.054
Final R indices [I>2sigma(I)]	R1 = 0.0399, wR2 = 0.1042
R indices (all data)	R1 = 0.0444, wR2 = 0.1087
Absolute structure parameter	0.041(18)
Largest diff. peak and hole	0.814 and -0.237 e.$Å^{-3}$

Supplementary Table 7: Atomic coordinates (x 10^4) and equivalent isotropic displacement parameters ($Å^2$x 10^3) for cu-4372. U(eq) is defined as one third of the trace of the orthogonalized U^{ij} tensor.

	x	y	z	U(eq)
S(1)	4706(1)	5796(1)	11649(1)	34(1)
O(2)	6043(1)	6329(1)	5008(2)	29(1)
O(3)	5131(1)	5463(1)	6181(2)	39(1)
O(4)	6284(1)	7981(1)	4768(2)	33(1)
O(5)	4288(1)	5842(1)	7781(2)	34(1)
O(6)	5258(1)	5188(1)	9234(2)	33(1)
O(7)	5996(1)	6306(1)	9408(2)	31(1)
O(8)	2594(1)	4132(1)	8032(2)	42(1)
O(10)	5384(1)	7093(1)	8389(2)	37(1)
N(12)	4798(1)	5505(1)	9005(2)	26(1)
N(13)	4096(1)	4678(1)	8003(2)	33(1)
O(14)	3216(1)	5824(1)	8948(2)	48(1)
N(15)	5922(1)	6849(1)	5975(2)	26(1)
O(16)	4101(1)	4404(1)	6260(2)	47(1)
N(18)	4221(1)	6358(1)	8911(2)	34(1)
N(19)	5578(1)	6146(1)	7772(2)	25(1)
N(20)	3328(1)	5218(1)	9434(2)	35(1)
N(21)	5219(1)	6741(1)	11963(2)	33(1)
C(22)	5460(1)	5384(1)	6285(2)	30(1)
C(23)	5425(1)	6951(1)	7456(2)	26(1)

C(24)	5916(1)	7098(1)	7000(2)	25(1)
C(25)	4886(1)	5152(1)	8893(2)	28(1)
N(26)	3323(1)	4733(1)	7594(2)	33(1)
C(27)	5973(1)	6466(1)	5918(2)	25(1)
N(29)	5057(1)	6702(1)	6770(2)	32(1)
C(30)	5619(1)	6144(1)	8911(2)	23(1)
N(31)	5422(1)	4976(1)	6011(2)	41(1)
C(32)	5984(1)	7561(1)	5375(3)	28(1)
C(33)	5926(1)	5742(1)	6788(2)	28(1)
C(34)	5149(1)	5946(1)	9535(2)	24(1)
C(35)	5962(1)	6212(1)	7031(2)	26(1)
C(36)	4530(1)	4703(1)	8362(3)	38(1)
C(37)	3460(1)	5676(1)	9347(3)	34(1)
C(38)	4483(1)	6514(1)	11365(3)	32(1)
C(39)	4593(1)	7351(1)	11559(3)	36(1)
C(40)	5513(1)	7601(1)	12210(3)	41(1)
C(41)	3434(1)	4484(1)	6740(3)	36(1)
C(43)	3908(1)	4523(1)	6977(3)	35(1)
C(44)	5186(1)	7140(1)	11938(2)	32(1)
C(45)	6146(1)	7609(1)	6598(2)	28(1)
C(46)	6001(1)	7131(1)	4950(2)	28(1)
C(47)	3399(1)	4631(1)	5523(3)	43(1)
C(48)	4799(1)	6367(1)	11620(3)	32(1)
C(49)	3990(1)	6226(1)	10934(3)	36(1)
C(50)	2874(1)	4839(1)	9070(3)	36(1)

C(52)	5371(1)	7930(1)	12164(3)	43(1)
C(53)	4362(1)	6244(1)	7961(3)	30(1)
C(54)	4725(1)	7011(1)	11586(2)	31(1)
C(55)	5205(1)	5884(1)	10802(2)	30(1)
C(56)	4592(1)	6627(1)	7070(3)	34(1)
C(57)	3969(1)	6003(1)	9765(3)	32(1)
C(58)	3736(1)	5149(1)	5324(3)	47(1)
C(59)	2919(1)	4536(1)	8182(3)	32(1)
C(60)	2890(1)	4491(2)	5275(4)	67(1)
C(61)	4914(1)	7809(1)	11856(3)	42(1)
C(63)	3778(2)	5296(2)	4088(4)	66(1)
O(1)	6755(1)	6101(1)	9330(2)	36(1)
O(9)	5292(1)	9041(1)	10246(2)	44(1)
O(11)	4964(1)	6076(1)	14786(2)	51(1)
C(7)	3787(1)	6380(1)	6355(4)	62(1)
O(15)	4394(1)	4685(1)	11754(3)	58(1)
C(6)	4280(1)	6499(2)	6027(3)	49(1)
C(1)	4472(2)	6945(2)	5298(5)	47(1)
C(2)	4206(2)	6839(2)	4183(5)	56(1)
C(1X)	4525(4)	6756(4)	4800(9)	45(2)
C(2X)	4642(4)	7259(4)	4975(10)	50(2)

Supplementary Table 8: Bond lengths [Å] and angles [°] for cu-4372.

S(1)-C(48)	1.746(3)
S(1)-C(55)	1.816(3)
O(2)-C(27)	1.229(3)
O(3)-C(22)	1.240(4)
O(4)-C(32)	1.426(3)
O(5)-C(53)	1.238(4)
O(6)-C(25)	1.238(3)
O(7)-C(30)	1.227(3)
O(8)-C(59)	1.233(4)
O(10)-C(23)	1.229(4)
N(12)-C(25)	1.337(4)
N(12)-C(34)	1.468(4)
N(13)-C(43)	1.338(4)
N(13)-C(36)	1.454(4)
O(14)-C(37)	1.220(4)
N(15)-C(27)	1.355(4)
N(15)-C(24)	1.466(3)
N(15)-C(46)	1.467(4)
O(16)-C(43)	1.232(4)
N(18)-C(53)	1.336(4)
N(18)-C(57)	1.449(4)
N(19)-C(30)	1.351(4)
N(19)-C(35)	1.462(3)

N(20)-C(37)	1.346(4)
N(20)-C(50)	1.453(4)
N(21)-C(44)	1.369(4)
N(21)-C(48)	1.372(4)
C(22)-N(31)	1.325(4)
C(22)-C(33)	1.513(4)
C(23)-N(29)	1.342(4)
C(23)-C(24)	1.534(4)
C(24)-C(45)	1.535(4)
C(25)-C(36)	1.488(4)
N(26)-C(59)	1.344(4)
N(26)-C(41)	1.455(4)
C(27)-C(35)	1.547(4)
N(29)-C(56)	1.468(4)
C(30)-C(34)	1.533(4)
C(32)-C(45)	1.519(4)
C(32)-C(46)	1.526(4)
C(33)-C(35)	1.516(4)
C(34)-C(55)	1.531(4)
C(37)-C(57)	1.551(4)
C(38)-C(48)	1.380(4)
C(38)-C(54)	1.440(4)
C(38)-C(49)	1.501(4)
C(39)-C(61)	1.386(5)
C(39)-C(54)	1.389(4)

C(40)-C(52)	1.378(5)
C(40)-C(44)	1.390(5)
C(41)-C(43)	1.527(5)
C(41)-C(47)	1.537(5)
C(44)-C(54)	1.417(4)
C(47)-C(58)	1.516(5)
C(47)-C(60)	1.529(5)
C(49)-C(57)	1.547(4)
C(50)-C(59)	1.507(4)
C(52)-C(61)	1.399(5)
C(53)-C(56)	1.520(4)
C(56)-C(6)	1.521(5)
C(58)-C(63)	1.521(5)
C(7)-C(6)	1.518(5)
C(6)-C(1)	1.540(6)
C(6)-C(1X)	1.667(11)
C(1)-C(2)	1.521(9)
C(1X)-C(2X)	1.517(16)

C(48)-S(1)-C(55)	99.14(13)
C(25)-N(12)-C(34)	119.7(2)
C(43)-N(13)-C(36)	122.6(3)
C(27)-N(15)-C(24)	127.1(2)
C(27)-N(15)-C(46)	118.9(2)
C(24)-N(15)-C(46)	112.0(2)

C(53)-N(18)-C(57)	119.7(2)
C(30)-N(19)-C(35)	120.9(2)
C(37)-N(20)-C(50)	123.6(3)
C(44)-N(21)-C(48)	109.1(2)
O(3)-C(22)-N(31)	122.6(3)
O(3)-C(22)-C(33)	121.8(3)
N(31)-C(22)-C(33)	115.5(3)
O(10)-C(23)-N(29)	122.4(3)
O(10)-C(23)-C(24)	119.6(2)
N(29)-C(23)-C(24)	117.8(2)
N(15)-C(24)-C(23)	114.9(2)
N(15)-C(24)-C(45)	102.7(2)
C(23)-C(24)-C(45)	108.5(2)
O(6)-C(25)-N(12)	121.2(3)
O(6)-C(25)-C(36)	118.8(3)
N(12)-C(25)-C(36)	120.0(3)
C(59)-N(26)-C(41)	123.4(3)
O(2)-C(27)-N(15)	121.2(2)
O(2)-C(27)-C(35)	120.1(2)
N(15)-C(27)-C(35)	118.7(2)
C(23)-N(29)-C(56)	120.1(3)
O(7)-C(30)-N(19)	123.7(3)
O(7)-C(30)-C(34)	122.8(2)
N(19)-C(30)-C(34)	113.4(2)
O(4)-C(32)-C(45)	109.8(2)

O(4)-C(32)-C(46)	114.0(2)
C(45)-C(32)-C(46)	103.7(2)
C(22)-C(33)-C(35)	113.7(2)
N(12)-C(34)-C(55)	111.9(2)
N(12)-C(34)-C(30)	110.2(2)
C(55)-C(34)-C(30)	111.3(2)
N(19)-C(35)-C(33)	109.9(2)
N(19)-C(35)-C(27)	111.9(2)
C(33)-C(35)-C(27)	111.0(2)
N(13)-C(36)-C(25)	116.2(2)
O(14)-C(37)-N(20)	124.6(3)
O(14)-C(37)-C(57)	122.8(3)
N(20)-C(37)-C(57)	112.5(3)
C(48)-C(38)-C(54)	105.8(3)
C(48)-C(38)-C(49)	128.2(3)
C(54)-C(38)-C(49)	126.0(3)
C(61)-C(39)-C(54)	119.8(3)
C(52)-C(40)-C(44)	117.9(3)
N(26)-C(41)-C(43)	111.1(2)
N(26)-C(41)-C(47)	113.1(3)
C(43)-C(41)-C(47)	112.2(3)
O(16)-C(43)-N(13)	121.9(3)
O(16)-C(43)-C(41)	121.0(3)
N(13)-C(43)-C(41)	117.1(3)
N(21)-C(44)-C(40)	130.8(3)

N(21)-C(44)-C(54)	107.7(3)
C(40)-C(44)-C(54)	121.5(3)
C(32)-C(45)-C(24)	102.9(2)
N(15)-C(46)-C(32)	103.8(2)
C(58)-C(47)-C(60)	112.6(3)
C(58)-C(47)-C(41)	111.7(3)
C(60)-C(47)-C(41)	109.0(3)
N(21)-C(48)-C(38)	110.3(3)
N(21)-C(48)-S(1)	121.0(2)
C(38)-C(48)-S(1)	128.5(2)
C(38)-C(49)-C(57)	112.3(2)
N(20)-C(50)-C(59)	111.9(2)
C(40)-C(52)-C(61)	121.7(3)
O(5)-C(53)-N(18)	122.2(3)
O(5)-C(53)-C(56)	121.2(3)
N(18)-C(53)-C(56)	116.5(3)
C(39)-C(54)-C(44)	119.0(3)
C(39)-C(54)-C(38)	133.8(3)
C(44)-C(54)-C(38)	107.1(3)
C(34)-C(55)-S(1)	113.59(19)
N(29)-C(56)-C(53)	110.5(2)
N(29)-C(56)-C(6)	110.7(3)
C(53)-C(56)-C(6)	110.0(2)
N(18)-C(57)-C(49)	111.4(2)
N(18)-C(57)-C(37)	108.9(3)

C(49)-C(57)-C(37)	112.9(2)
C(47)-C(58)-C(63)	113.9(3)
O(8)-C(59)-N(26)	124.3(3)
O(8)-C(59)-C(50)	119.6(3)
N(26)-C(59)-C(50)	116.1(3)
C(39)-C(61)-C(52)	120.0(3)
C(7)-C(6)-C(56)	110.5(4)
C(7)-C(6)-C(1)	105.2(4)
C(56)-C(6)-C(1)	106.8(4)
C(7)-C(6)-C(1X)	122.4(5)
C(56)-C(6)-C(1X)	118.3(5)
C(1)-C(6)-C(1X)	33.6(4)
C(2)-C(1)-C(6)	110.4(5)
C(2X)-C(1X)-C(6)	103.3(9)

Symmetry transformations used to generate equivalent atoms:

Supplementary Table 9: Anisotropic displacement parameters ($Å^2 x\ 10^3$) for cu-4372. The anisotropic displacement factor exponent takes the form: $-2p^2[\ h^2\ a^{*2}U^{11} + ...\ + 2\ h\ k\ a^*\ b^*\ U^{12}]$

	U^{11}	U^{22}	U^{33}	U^{23}	U^{13}	U^{12}
S(1)	33(1)	37(1)	33(1)	5(1)	11(1)	19(1)
O(2)	32(1)	30(1)	25(1)	1(1)	3(1)	16(1)
O(3)	31(1)	39(1)	42(1)	-2(1)	-4(1)	14(1)
O(4)	34(1)	28(1)	35(1)	4(1)	9(1)	13(1)
O(5)	35(1)	35(1)	38(1)	6(1)	5(1)	23(1)
O(6)	37(1)	38(1)	34(1)	-1(1)	-1(1)	26(1)
O(7)	26(1)	34(1)	31(1)	-3(1)	-2(1)	13(1)
O(8)	29(1)	32(1)	47(1)	-5(1)	6(1)	3(1)
O(10)	27(1)	43(1)	38(1)	-5(1)	8(1)	16(1)
N(12)	22(1)	23(1)	33(1)	3(1)	2(1)	12(1)
N(13)	27(1)	31(1)	42(2)	-2(1)	1(1)	14(1)
O(14)	28(1)	40(1)	66(2)	10(1)	-9(1)	11(1)
N(15)	28(1)	26(1)	25(1)	-2(1)	0(1)	14(1)
O(16)	50(1)	51(1)	45(1)	-12(1)	4(1)	29(1)
N(18)	22(1)	24(1)	50(2)	-1(1)	2(1)	6(1)
N(19)	18(1)	27(1)	28(1)	4(1)	2(1)	10(1)
N(20)	23(1)	32(1)	43(2)	-6(1)	-1(1)	10(1)
N(21)	23(1)	42(1)	36(1)	-7(1)	-6(1)	17(1)
C(22)	29(2)	28(1)	29(2)	5(1)	5(1)	12(1)
C(23)	22(1)	25(1)	33(2)	5(1)	3(1)	12(1)

C(24)	21(1)	27(1)	24(1)	-2(1)	2(1)	11(1)
C(25)	30(2)	29(1)	28(1)	3(1)	3(1)	18(1)
N(26)	24(1)	33(1)	35(1)	-2(1)	3(1)	9(1)
C(27)	18(1)	27(1)	28(1)	1(1)	3(1)	10(1)
N(29)	21(1)	39(1)	36(1)	2(1)	-2(1)	15(1)
C(30)	27(1)	19(1)	28(1)	-1(1)	-1(1)	14(1)
N(31)	42(2)	31(1)	45(2)	-9(1)	-2(1)	15(1)
C(32)	23(1)	23(1)	35(2)	2(1)	2(1)	10(1)
C(33)	24(1)	29(1)	27(1)	3(1)	3(1)	12(1)
C(34)	24(1)	24(1)	30(1)	0(1)	3(1)	16(1)
C(35)	22(1)	28(1)	25(1)	1(1)	4(1)	12(1)
C(36)	34(2)	32(2)	50(2)	0(1)	0(1)	19(1)
C(37)	24(1)	37(2)	32(2)	2(1)	4(1)	10(1)
C(38)	25(1)	36(2)	36(2)	-4(1)	-1(1)	15(1)
C(39)	34(2)	42(2)	36(2)	-7(1)	-8(1)	22(1)
C(40)	28(2)	49(2)	33(2)	-9(1)	-5(1)	10(2)
C(41)	28(2)	34(2)	36(2)	-7(1)	3(1)	8(1)
C(43)	33(2)	26(1)	37(2)	0(1)	4(1)	8(1)
C(44)	24(1)	39(2)	32(2)	-4(1)	-1(1)	16(1)
C(45)	26(1)	26(1)	30(2)	-1(1)	3(1)	9(1)
C(46)	30(2)	28(1)	24(1)	1(1)	1(1)	14(1)
C(47)	32(2)	52(2)	32(2)	-6(1)	-1(1)	13(2)
C(48)	26(1)	41(2)	32(2)	-2(1)	6(1)	19(1)
C(49)	22(1)	38(2)	44(2)	-8(1)	4(1)	11(1)
C(50)	20(1)	35(2)	45(2)	-5(1)	2(1)	6(1)

C(52)	45(2)	35(2)	36(2)	-4(1)	-2(1)	11(2)
C(53)	18(1)	34(2)	41(2)	6(1)	3(1)	15(1)
C(54)	26(1)	37(2)	31(2)	-7(1)	-2(1)	17(1)
C(55)	31(2)	34(2)	30(1)	6(1)	4(1)	21(1)
C(56)	21(1)	30(2)	52(2)	9(1)	6(1)	12(1)
C(57)	24(1)	30(2)	39(2)	-6(1)	1(1)	10(1)
C(58)	38(2)	61(2)	34(2)	5(2)	1(1)	18(2)
C(59)	26(2)	31(2)	36(2)	-2(1)	1(1)	11(1)
C(60)	34(2)	93(3)	44(2)	9(2)	-9(2)	10(2)
C(61)	50(2)	43(2)	32(2)	-5(1)	-3(1)	22(2)
C(63)	46(2)	79(3)	46(2)	12(2)	-2(2)	13(2)
O(1)	40(1)	44(1)	29(1)	-3(1)	0(1)	25(1)
O(9)	45(1)	45(1)	50(2)	-8(1)	-8(1)	28(1)
O(11)	39(1)	56(2)	53(2)	-7(1)	-14(1)	20(1)
C(7)	26(2)	44(2)	110(4)	19(2)	-11(2)	15(2)
O(15)	36(1)	68(2)	67(2)	22(1)	0(1)	23(1)
C(6)	25(2)	51(2)	62(2)	32(2)	0(2)	12(2)
C(1)	46(3)	53(3)	47(3)	21(2)	10(2)	28(2)
C(2)	59(3)	67(3)	43(3)	20(2)	10(2)	33(3)
C(1X)	47(3)	51(3)	44(4)	17(3)	9(3)	29(3)
C(2X)	54(4)	51(4)	53(4)	20(3)	10(3)	32(3)

4 Supplementary Reference

1. Kyte, J., Doolittle, R.F. *J. Mol. Biol.*,**157**, 105-132 (1982).
2. Bartoloni, M., Jin. X., Marcaida, M. J., Banha, J., Dibonaventura, I., Bongoni, S., Bartho, K., Gräbner, O., Sefkow, M., Darbre, T., Reymond, J. *Chem. Sci.* **6**, 5473–5490 (2015).
3. Wareham, R. S., Kilburn, J. D., Turner, D. L., Rees, N. H. & Holmes, D. S. *Angew. Chem. Int. Ed. Engl.* **34**, 2660–2662 (1996).
4. Knappe, T. A., Linne, U., Zirah, S., Rebuffat, S., Xie, X., Marahiel, M. *J. Am. Chem. Soc.* **130**, 11446–11454 (2008).
5. Reisberg, S. H., Gao, Y., Walker, A., Helfrich, E. J., Clardy, J., Baran, P. S. *Science* **367**, 458–463 (2020).
6. Nicolaou, K. C., Boddy, C. N. C. & Siegel, J. S. *Angew. Chem. Int. Ed. Engl.* **40**, 701–704 (2001).
7. McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein,C., Swails, J. M., Hernández, C. X., Schwantes, C. R., Wang, L.-P., Lane, T. J., Pande, V. S. *Biophys. J.* **109**, 1528-1532 (2015).
8. Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., Laue, E, D. *Proteins: Structure, Function and Genetics*, **59**(4), 687–696 (2005).
9. Yilmaz, E. M., Güntert, P. *J Biomol NMR*. **63**, 21-37 (2015).
10. Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., Ferrin, T. E. *J Comput Chem*. **25**(13):1605-12 (2004).
11. Yao, G., Knittel, C. H., Kosol, S., Wenz, M. T., Keller, B. G., Gruß, H., Braun, A. C., Lutz, C., Hechler, T., Pahl, A. Suessmuth, R. D. *J. Am. Chem. Soc.***143**, 14322-14331 (2021).
12. [a] Abraham, M. J., van der Spoel, D., Lindahl, E., Hess, B. GROMACS development team, (2019); [b] Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., Berendsen, H. J. *J. Comput. Chem.* **26**, 1701-1718 (2005); [c] Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., Lindahl, E. *SoftwareX* **1**, 19-25 (2015).
13. Hanwell, M. D., Curtis, D. E., Lonie, D. C., Vandermeersch, T., Zurek, E., Hutchison, G. R. *J. Cheminform.* **4**, 17 (2012).
14. Sousa da Silva, A. W., Vranken, W. F. *BMC Res Notes* **5**, 367-367 (2012).
15. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., Simmerling, C. *J. Chem. Theory Comput.* **11**, 3696-3713 (2015).
16. Van Gunsteren, W. F., Berendsen, H. J. *Mol. Simul.* **1**, 173-185 (1988).
17. Hess, B., Bekker, H., Berendsen, H. J., Fraaije, J. G. *J. Comput. Chem.* **18**, 1463-1472 (1997).
18. Bussi, G., Donadio, D., Parrinello, M. *J. Chem. Phys.* **126**, 014101 (2007).
19. Parrinello, M., Rahman, A. *J. Appl. Phys.* **52**, 7182-7190 (1981).
20. Páll, S., Hess, B. *Comput. Phys. Commun.* **184**, 2641-2650 (2013).
21. Darden, T., York, D., Pedersen, L. *J. Chem. Phys.* **98**, 10089-10092 (1993).
22. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J. *Nature* **585**, 357-362 (2020).

23. Yao, G., Kosol, S., Wenz, M.T., Irran, E., Keller, B. G., Trapp, O., Süssmuth, R. D. DOI: 10.5281/zenodo.6974777 (2022).
24. Shrake, A., Rupley, J. A. *J. Mol. Biol.* **79**, 351-371 (1973).
25. Neese, F., Wennmohs, F., Becker, U., Riplinger, C. *J. Chem. Phys.* **152**, 224108 (2020).
26. Bannwarth, C., Ehlert, S., Grimme, S. *J. Chem. Theory Comput.* **15**, 1652-1671 (2019).
27. [a] Becke, A. D. *J. Chem. Phys.* **96**, 2155-2160 (1992); [b] Lee, C., Yang, W., Parr, R. G. *Phys. Rev. B* **37**, 785 (1988); [c] Becke, A. D. *J. Chem. Phys.* **98**, 1372-1377 (1993); [d] Caldeweyher, E., Bannwarth, C., Grimme, S. *J. Chem. Phys.* **147**, 034112 (2017); [e] Weigend, F., Ahlrichs, R. *Phys. Chem. Chem. Phys.* **7**, 3297-3305 (2005).
28. Barone, V., Cossi, M. *J. Phys. Chem. A* **102**, 1995-2001 (1998).
29. Neese, F., Wennmohs, F., Hansen, A., Becker, U. *Chem. Phys.* **356**, 98-109 (2009).
30. Abraham, M. J., van der Spoel, D., Lindahl, E., Hess, B. GROMACS development team (2021).
31. Zagrovic, B., Van Gunsteren, W. F. *Proteins* **63**, 210-218 (2006).
32. Tropp, J. *J. Chem. Phys.* **72**, 6035-6043 (1980).

4.4 Paper B3

Title:

"The influence of N-methylation on the ansamers of an amatoxin: Gly5Sar-amanullin"

The name 'amatoxins' describes a family of fungal toxins that are found in several genera of poisonous mushrooms, most notably *Amanita*.[59,85] The toxicity of the amatoxins comes from their ability to inhibit RNA polymerase II and thereby causing cell death.[63,65,78] Due to this inhibition capacity, amatoxins have gained large interest for cancer research. They are believed to be useful for the development of new antibody drug conjugates (ADCs).[66,67] This application purpose, however, requires (a) a deep understanding of the synthesis to create amatoxins of different kind, and (b) the knowledge of how and where the amatoxins can be tuned in order to obtain or intensify a specific property (SAR studies). Amatoxins comprise eight amino acids that are linked to a cyclic polyamide chain (macrolactam) with an additional cross-link between the side chains of the tryptophan and the cysteine residue (tryptathionine bridge). Due to this bicyclic structure, amatoxins are very rigid. This rigidity significantly contributes to their toxicity, as it makes amatoxins robust against heat or acid degradation.[59,78]

N-methylation is a well-known method to distort peptidic structures,[110,111] and in paper **B2** (see 4.3), we could show that the amatoxin scaffold prinicpally allows for two isomeric forms (M/P-ansamers). In this work, we therefore examined the influence of the N-methylation on the amatoxin scaffold by studying both ansamers of Gly5Sar-amanullin with NMR experiments and MD simulations. We could show that the N-methylation allows the ansamers to form two long-living conformations, respectively. However, both ansamers are differently affected by the N-methylation: In the unmethylated natural ansamer (P), Gly5 is involved in many structure-preserving hydrogen bonds.[77,219] Hence, it was expected that the natural Gly5Sar-amanullin loses these hydrogen bonds due to the N-methylation. However, we do not observe a resulting increased flexibility of the backbone as the loss of the hydrogen bonds is compensated by a new hydrogen bond pattern. In the unnatural Gly5Sar-amanullin (M), the N-methylation enables an interesting cis-trans isomerisation for the peptide bonds directly neighbouring the tryptathionine bridge. Our insight on the effects of the N-methylation could initiate further SAR-studies on the amatoxin scaffold, which is important for the design of better ADCs.

In this project, we made use of the following techniques:

- unbiased, classical MD simulations
- dimensionality reduction of the MD trajectories with time-independent component analysis (TICA)[146,147]

- discretisation of the MD trajectories by density-based clustering with common-nearest-neighbour algorithm (implementation: "CommonNN")[188,190]
- Structural characterisation of the modelled peptides by the analysis of: backbone torsion angles (ϕ, ψ, ω), hydrogen bonds and hydrogen bond sets, RMSD, solvent-accessible surface areas, ${}^{3}J_{\mathrm{H^N H^\alpha}}$-coupling constants, violations of NOE distances from NMR experiments
- core-set Markov state models (cs-MSM) for the kinetic analysis of the clustering results

We synthesised Gly5Sar-amanullin following the synthesis route from paper B1 (see 4.2). Spectroscopical evidence (UV-VIS, CD, NMR NOE) suggested that we have exclusively synthesised the M-ansamer. In the ^{1}H-NMR spectrum, we could identify two peaks for the indole amide of the tryptophan side chain. We concluded that these two peaks indicate two conformers of M_{ansa}-Gly5Sar-amanullin (1_M, 2_M), because all peaks belong to only a single chemical species. From ^{1}H-^{1}H-EXSY spectra we could determine that (1) the two peaks exchanged with a rate $k \approx 1\,\mathrm{s}^{-1}$ at 298 K and (2) the two conformers show populations 40% (1_M) and 60% (2_M).

Since the exchange rate suggests that the interconversion between 1_M and 2_M is not accessible by MD simulations at room temperature, we simulated M_{ansa}-Gly5Sar-amanullin at 500 K starting from two different starting structures. At 500 K, our starting conformations form one joint conformational ensemble. While the (ϕ, ψ)-backbone torsion angles were unremarkable, we observed cis-trans isomerisation for the ω-backbone torsion angles of Trp4 (ω_4), Sar5 (ω_5), Gly7 (ω_7) and Cys8 (ω_8). When we cooled the system down to 300 K, the ω_4 and ω_7 still showed cis and trans configurations, while ω_5 and ω_8 only showed the trans-configuration.

Based on the configurations of ω_4 and ω_7, we could subdivide the 300 K MD data set into four subsets: (trans-ω_4, trans-ω_7), (trans-ω_4, cis-ω_7), (cis-ω_4, trans-ω_7) and (cis-ω_4, cis-ω_7). At 300 K, we do not observe any transition between these configurations, except for short excursions from (trans-ω_4, trans-ω_7) to (cis-ω_4, trans-ω_7). The (trans-ω_4, trans-ω_7)-subset showed almost perfect agreement to NOE distances and amide temperature coefficients of 1_M. We therefore assigned this configuration to 1_M. For 2_M, we had less reference data from NMR, and none of the subsets showed a good agreement to the amide temperature coefficients of 2_M. However, the configurations with trans-ω_7 showed fewer and less significant violations for the NOE distances of 2_M than configurations with cis-ω_7. This might indicate that both 1_M and 2_M are consistent with a trans-ω_7 configuration.

To further investigate the structures in the (trans-ω_4, trans-ω_7)-subset, we performed clustering with CommonNN after a dimensionality reduction with TICA. We could identify five clusters (c1-c5), based on which we constructed a core-set MSM to extract the timescales

for the conformational transitions between them. We identified two slow processes that linked three of the five clusters (c1,c2,c5). The slowest process (c1,c5) belongs to structural fluctuations of the (trans-ω_4,trans-ω_7)-configuration without changes in the configurations of the backbone angles. The second-slowest process (c1,c2) belongs to short excursions to a cis-ω_4-configuration, which is coupled by changes in the ϕ- and ψ-backbone torsion angles of Sar5. We compared the associated cluster with the (cis-ω_4,trans-ω_7)-subset and found a difference of $40°$ in the ψ-angle of Sar5. Based on that difference, we conclude that the second-slowest process might represent a recrossing-event, i.e. a short excursion into the other state without reaching its energetical minimum. The clusters that were not connected on the timescale of our MD simulations do not involve changes in the configurations of ω_4 or ω_7. Although the transitions to these clusters were not sampled sufficiently, we can conclude that the (trans-ω_4,trans-ω_7)-configuration is indeed stable enough to represent 1_M.

We also simulated P_{ansa}-Gly5Sar-amanullin, first at higher temperatures with ensuing cool down to 300 K. We also identified two long-living conformations. In contrast to the M-ansamer, however, we did not observe any cis-trans isomerisation of any backbone torsion angle. Instead, we found two distinct hydrogen bond patterns and the corresponding conformations are stable over several μs.

At last, we compared the (trans-ω_4,trans-ω_7)-M_{ansa}-conformation and both P_{ansa}-conformations with the M_{ansa}- and P_{ansa}-structures of the unmethylated Gly5Sar-amanullin from paper **B2** (see 4.3). We could see that our (trans-ω_4,trans-ω_7)-M_{ansa}-conformation resembles the unmethylated M_{ansa}-amanullin. We can therefore conclude that the N-methylation in Gly5 has little effect on the conformation of the M-ansamer. In the P-ansamer, Gly5 is involved in structure-determining hydrogen bonds. But, in contrast to what one would expect, N-methylating Gly5 does not lead to more flexibility. Instead, two very distinct hydrogen bond patterns compensate the missing interactions. Taken together, our results provide a new perspective on possible tuning options for the amatoxin scaffold, which might be used for further SAR-studies. Our results have been made publicly available on the BioRxiv preprint server (doi:10.1101/2022.12.21.521444).

This project was a cooperation between the group of Roderich D. Süssmuth (R.D.S., Technische Universität Berlin) and the group of Bettina G. Keller (B.G.K., Freie Universität Berlin). The syntheses were performed by Guiyang Yao (G.Y., Fudan University (China), formerly AG Süssmuth). All spectroscopical evidence was generated by Simone Kosol (AG Süssmuth). Marius T. Wenz (M.T.W.) performed all MD simulations including the construction and parametrisation of the peptides. He conducted all clustering approaches including the preceding dimensionality reductions with TICA. Based on the clustered trajectories, M.T.W. constructed core-set Markov state models to extract the kinetic information. For both ansamers, M.T.W. performed structural characterisations in great detail, and proposed final structures for the comparison to the experiments. The manuscript was written by B.G.K. and M.T.W with G.Y.,

S.K. and R.D.S. proofreading. All authors contributed to the final version of the manuscript.

The presented research was published as preprint on the *BioRxiv preprint server*: Wenz, M.T. et al. The influence of N-methylation on the ansamers of an amatoxin: Gly5Sar-amanullin. bioRxiv DOI: 10.1101/2022.12.21.521444.

5 CONCLUSION AND OUTLOOK

Over the years, MD simulations have evolved as a powerful tool to study biological systems.[19,103,104,220] We can work in very high precision: we can describe molecules in atomic positions and their dynamics in multiples of femtoseconds (see eq. 3.10, eq. 3.11). However, our computational microscope has also some limitations that restrict its usability (see 3.1.4). There are sophisticated modifications for our microscope, such as enhanced sampling techniques[113,162] or coarse-grained approaches,[125] which allow us to overcome some of these limitations. Nevertheless, the quality of the insight provided by our computational microscope decisively depends on the hypothesis that is to be addressed.

In this thesis, we have discussed the usage of MD simulations to gain insights into specific aspects of PPIs research. Although both projects are substantially different, the methodology with which we approached them is comparable: the first question before studying any system with MD simulations should therefore be whether MD simulations are indeed the correct method. In other words, are we able to formulate a problem that can be addressed with insights from our computational microscope? Second, we then have to design a strategy to combine MD simulations with data analysis methods, so we can extract the desired features from the system? Third, we have to collate these features and the experimental evidence?

In the following, we will discuss some additional aspects as well as future perspectives for the individual projects, before we then come to a more generalised conclusion. The idea of this chapter is to provide further information about the choices made in the projects and the challenges that remain.

5.1 The tWW project

To study the complex formed during the PPI between h-FBP21 and Sm B/B', we first had to create it. From literature,[42,47,221] it was known that interaction between h-FBP21 and Sm B/B' is based on the recognition event between a *tandem* WW domain (tWW) of h-FBP21 and a bivalent proline-rich peptide sequence from Sm B/B' (SmB2 ligand). However, the versatility of tWWs in terms of structures and binding behaviours (see 1.3.1) made it impossible to predict the tWW-PRMs complex structure from existing complex structures of other tWWs.[45,222] Without prior knowledge about the complex formation, simulating both components together and waiting for the complex formation, certainly is a sparsely expedient approach. So, instead of asking how the complex is formed in solution, we took a detour: based on the properties of the solution structure of h-FBP21 tWW (PDB entry: 2JXW), we assumed a "conformational selection"-mechanism for the formation of possible complex structures (see 1.3.2). This assumption allowed us to explore the conformational ensemble of h-FBP21 tWW,

which is a formidable task for our computational microscope.

Next, we had to face one of the challenges of classical MD simulations: without further processing, a selection of "meaningful" h-FBP21 tWW structures is not possible. We instead considered the relative positioning of the WW domains in space to be pivotal for the ability of the XP grooves to bind to the PRMs of the SmB2 ligand. Hence, we created a sub-dimensional representation of our MD ensemble, which was designed to serve two purposes. (1) By the dimensionality reduction, we were able to consider clustering techniques. (2) The arrangments of the WW domains is reflected in the projection we chose. The user-dependency of the "common-nearest-neighbour" clustering (CommonNN) algorithm might be considered a weak spot of our workflow. However, the CommonNN algorithm has proven suitable to partition other high-dimensional MD data sets effectively.[188,190] Based on the obtained cluster structures, we then mimicked the complex formation with the SmB2 ligand by performing docking experiments. Finally, we characterised the best h-FBP21 tWW-SmB2 ligand complexes by simulating them — again, a task our microscope is fit for.

This project had two important outcomes: (1) a detailed description of the complex structures of h-FBP21 tWW with a proline-rich ligand from Sm B/B' and (2) a workflow, which could be used for other repeats of recognition events. Since the complex structures were solely obtained *in silico*, it would certainly be interesting if future studies could provide further experimental proof, e.g. a crystal structure as verification. Although our complex structures added another perspective on tWWs and their binding behaviours, some challenges remain: we had to mimic the formation of the complex using docking methods, so we still do not know what determines the orientation of the ligand. In the structures provided by the docking experiments, we found an approximate 50:50 ratio for the two orientations described in Ref. [47]. Second, we do not know *how* the complex is formed, either. Based on the stable fold of the WW domains in h-FBP21 tWW, we can exclude a folding-upon-binding mechanism which was reported for other tWWs,[50] and as elaborated in the introduction (see 1.3.2), a pure conformational-selection mechanism is unlikely. Hence, a potential binding mechanism should incorporate multivalency effects, in which one WW domain binding to the target facilitates the binding for the other WW domain. It would therefore be very interesting, if future studies could investigate the formation of the complex in atomistic detail, e.g. by using enhanced sampling methods.

Next, we could not determine the role of Arg6 in the first WW domain. Although it is involved in many hydrogen bonds, our experimental data showed that Arg6 does not contribute to the thermodynamic stability of the h-FBP21tWW-SmB2 ligand complex. Yet, it might be that Arg6 plays an important role for the binding kinetics by favouring the formation of binding-competent structures. Once we have simulation data on the binding equilibrium,

Markov models might be useful to unravel the role of Arg6.

Another interesting focus for future studies might be the arginine residues flanking the PRMs in the SmB2 ligand. We found that they are involved in strong hydrogen bonds to glutamic acids in both complexes. However, mutating the arginine residues with alanine *in silico* did not affect the complex structures much. If the arginine residues are not crucial for binding, they still can be important for the ligand orientation. The involvement of these arginine residues in the orientation of the ligand relative to the h-FBP21 tWW might be another feature, simulations of the binding equilibrium might provide.

Last, we claimed that our workflow might also be applicable for other recognition domains in tandem repeats. Future studies on other domains such as the SH3 domain[223] might therefore benefit from our protocol.

5.2 The amatoxins

The small size of the amatoxins could give a false sense of simplicity. By contrast, their complex structure made it impossible to refer to force field parameters of regular amino acids in a single polyamide chain. Instead, we had to parametrise the molecules of interest, which added another dependency to our microscope.

5.2.1 The tryptathionine bridge formation

In the first subproject (publication **B1**, see 4.2), our computational microscope had more of an advisory role. For different precursors (linear and macrocyclic octapeptides), our colleagues found different yields for the formation of the tryptathionine bridge. Yet, it was unclear what favours the reaction. Bond forming or breaking cannot be visualised with our computational microscope. So, instead of asking how the tryptathionine linkage is established, we wondered whether the peptides can adopt conformations more frequently in which the side chains of tryptophan and cysteine are in close vicinity. Assuming proper sampling, MD simulations are the perfect tool for a probability analysis of the occurring conformations. We found that none of the precursors, linear and monocyclic, expresses a clear trend to pre-organise in solution, and the side chains of the relevant residues (Trp4, Cys8) are independent of each other. Hence, we concluded that the selectivities cannot be explained by pre-organisation of the precursors.

The most important outcome of this project was the novel synthetic strategy, which enables us to synthesise a library of amatoxins. However, there are two more important aspects. (1) We could exclude the formation of a particular conformation, which favours the reaction and (2) we could show that the parametrisation of our precursors yields structural dynamics in agreement to NMR data (NOE distances). Since our results indicate that the reaction might

not be driven by the formation of a certain conformation, future studies should definitely be dedicated to the investigation of the reaction mechanism. Here, quantum mechanical calculations would be better suited than classical MD simulations to calculate possible transition states. Our colleagues found that the macrocyclic precursor involving *D*-Pro instead of hydroxyl-proline gives higher yields for the formation of the tryptathionine bridge. It might therefore be interesting to also study the impact of *D*-amino acids on both the reaction and the structural dynamics. Last, we performed unbiased classical MD simulations of the precursors in solution. In the lab, the precursors are however linked to a resin plate. For modelling, the physical interactions to the resin plate are too complex, but perhaps future studies could run simulations with a position restraint on the residue, which is bound to the resin.

5.2.2 The ansamers

In the second subproject (publication **B2**, see 4.3), we investigated the occurrence of isomers (ansamers) in the synthesis of amatoxins. Our results revealed that the amatoxin scaffold allows for two conformers, which cannot interconvert into each other. We performed classical MD simulations for the two isomers as well as their precursors. The atomistic resolution of our computational microscope enabled us to examine the structural conditions of the isomers in solution, and to compare them to the crystal structures. Knowing the exact position of each single atom in each time step has some key-advantages: first, we can screen the conformational ensembles of the precursors for structural similarities to the two ansamers. Second, we can apply basic linear algebra to define vectors or planes, and track their change over time. This came in very handy to develop a model, with which we could monitor the position of the tryptathionine bridge relative to the cyclic peptide chain. In our model, we divide the amatoxin scaffold in three parts, which we then represented by planes. Although this representation is not perfectly accurate, it was sufficient to exemplify the fundamental difference between the isomers and their incapability to convert into each other.

In this project, the most important outcomes were (1) the synthesis of the ansamers, (2) the structural characterisation of the ansamers by crystallography and MD simulations and (3) the novel nomenclature “ansamers” to unambiguously describe the conformations of cyclic peptides. As reported in the introduction (see 1.4.5), there also are other bicyclic peptides that possess a tryptathionine linkage. Hence, it is conceivable that these peptides can also form ansamers. Here, future studies both experimentally and theoretically could be performed to isolate and identify ansameric structures in these peptides. Subproject **B2** can be seen as perfect example for the symbiosis of experimental evidence and insights from our computational microscope. Future studies will show whether the nomenclature becomes accepted by the research community.

5.2.3 N-methylation in amatoxins

In the third subproject (publication **B3**, see 4.4), our task was to characterise the two major conformations of the M_{ansa}-Gly5Sar-amanullin (1_M, 2_M). In this subproject, we were confronted with several difficulties: first, we could experimentally determine that the two conformations interconvert on timescales, which are not accessible with classical MD simulations. Second, we only had little experimental evidence on the structures of the two conformations. Consequently, the choice of the starting structure and the analysis of the MD ensemble were difficult. From the previous projects, we knew that amatoxins are highly organised in solution, and there is little flexibility in the structures. Hence, there might be very high energy barriers between different conformations and consequently only a few transitions if any might be observable with our computational microscope. Given the slow exchange rate between 1_M and 2_M, we therefore performed high-temperature simulations to allow the system to overcome these barriers and prevent the system from oversampling structures. Although MD simulations at temperatures greater than 400 K do not represent proper representations of the conformational ensemble at room temperature, they are useful to explore the structural possibilities of the Gly5Sar-amanullin. Since we knew from project **B2** that there are ansamers, we also simulated P_{ansa}-Gly5Sar-amanullin although this ansamer had not been synthesised.

The most important outcomes of our study were: (1) the two ansamers are differently affected by the N-methylation and (2) we observe two long-living conformations for both ansamers, respectively. We could identify an interesting cis-trans isomerisation for the ω-backbone torsion angles of Trp4 and Gly7. Although we could find four different configurations (see 4.4), we were only able to assign one configuration to the experimental reference data. One issue was that the transitions between these configurations were not sampled sufficiently. Here, future studies could make use of enhanced sampling methods like replica-exchange to further sample these transitions. Also, to rule out that we missed conformations, it would be interesting to sample conformations via Monte-Carlo simulations or other methods that focus on generating the equilibrium distribution and not the dynamics. The so-sampled conformations could then be used for ensuing MD simulations. Since amatoxins are promising antibody drug conjugates,[66,67] the elucidation of the structure-activity relationship (SAR) is very important. Here, future studies will show whether the N-methylation can indeed be a new starting point for SAR-studies.

5.3 Closing remarks

Once more, MD simulations have proven effective to provide a different perspective on biological systems. Knowing its limitations and asking the correct questions, we are able to use our computational microscope to unravel the structural and dynamical conditions beyond

experimental resolution. Of course, MD simulations are no panacea, but as we have seen, MD simulations represent a beneficial compromise between accuracy and computational costs. Especially if they are well-adjusted to experiments, they can be an integral part of research on interdisciplinary questions. The projects presented in this thesis are proof for the significance of our computational microscope and its valid claim to either complete experiments (**B1**, **B2**) or produce results that stand on their own (**A**, **B3**). At the same time, the success of the projects is evidence for the fruitful cooperations with the group of Christian Freund (Freie Universität Berlin) and the group of Roderich D. Süssmuth (Technische Universität Berlin). As such, our success emphasises the usefulness of interdisciplinary research associations such as graduate schools to explore complex problems in chemistry.

REFERENCES

(1) Alberts, B. *Cell* **1998**, *92*, 291–294, DOI: 10.1016/S0092-8674(00)80922-8.

(2) Kinbara, K.; Aida, T. *Chem. Rev.* **2005**, *105*, 1377–1400, DOI: 10.1021/cr030071r.

(3) Kolomeisky, A. B.; Fisher, M. E. *Annu. Rev. Phys. Chem.* **2007**, *58*, 675–695, DOI: 10.1146/annurev.physchem.58.032806.104532.

(4) Spirin, A. S. *FEBS Lett.* **2002**, *514*, 2–10, DOI: 10.1016/s0014-5793(02)02309-8.

(5) Chen, H.-C.; Cheng, S.-C. *Biosci. Rep.* **2012**, *32*, 345–359, DOI: 10.1042/bsr20120007.

(6) Bren, A.; Eisenbach, M. *J. Bacteriol.* **2000**, *182*, 6865–6873, DOI: 10.1128/jb.182.24.6865-6873.2000.

(7) Westermarck, J.; Ivaska, J.; Corthals, G. L. *Mol. Cell. Proteom.* **2013**, *12*, 1752–1763, DOI: 10.1074/mcp.R113.027771.

(8) Jones, S.; Thornton, J. M. *Proc. Natl. Acad. Sci.* **1996**, *93*, 13–20, DOI: 10.1073/pnas.93.1.13.

(9) Nooren, I. M.; Thornton, J. M. *EMBO J.* **2003**, *22*, 3486–3492, DOI: 10.1093/emboj/cdg359.

(10) Sheinerman, F. B.; Norel, R.; Honig, B. *Curr. Opin. Struct. Biol.* **2000**, *10*, 153–159, DOI: 10.1016/s0959-440x(00)00065-8.

(11) Petta, I.; Lievens, S.; Libert, C.; Tavernier, J.; De Bosscher, K. *Mol. Ther.* **2016**, *24*, 707–718, DOI: 10.1038/mt.2015.214.

(12) Lu, H.; Zhou, Q.; He, J.; Jiang, Z.; Peng, C.; Tong, R.; Shi, J. *Sig. Transduct. Target. Ther.* **2020**, *5*, 213, DOI: 10.1038/s41392-020-00315-3.

(13) Ryan, D. P.; Matthews, J. M. *Curr. Opin. Struct. Biol.* **2005**, *15*, 441–446, DOI: https://doi.org/10.1016/j.sbi.2005.06.001.

(14) LaCount, D. J.; Vignali, M.; Chettier, R.; Phansalkar, A.; Bell, R.; Hesselberth, J. R.; Schoenfeld, L. W.; Ota, I.; Sahasrabudhe, S.; Kurschner, C.; Fields, S.; Hughes, R. E. *Nature* **2005**, *438*, 103–107, DOI: 10.1038/nature04104.

(15) Wells, J. A.; McClendon, C. L. *Nature* **2007**, *450*, 1001–1009, DOI: 10.1038/nature06526.

(16) Arkin, M. R.; Wells, J. A. *Nat. Rev. Drug Discov.* **2004**, *3*, 301–317, DOI: 10.1038/nrd1343.

(17) De Las Rivas, J.; Fontanillo, C. *PLoS Comput Biol* **2010**, *6*, e1000807, DOI: 10.1371/journal.pcbi.1000807.

(18) Nero, T. L.; Morton, C. J.; Holien, J. K.; Wielens, J.; Parker, M. W. *Nat. Rev. Cancer* **2014**, *14*, 248–262, DOI: 10.1038/nrc3690.

(19) Rakers, C.; Bermudez, M.; Keller, B. G.; Mortier, J.; Wolber, G. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **2015**, *5*, 345–359, DOI: 10.1002/wcms.1222.

(20) Lee, A. C.-L.; Harris, J. L.; Khanna, K. K.; Hong, J.-H. *Int. J. Mol. Sci.* **2019**, *20*, 2383, DOI: 10.3390/ijms20102383.

(21) Beck-Sickinger, A. *Nachrichten aus der Chemie* **2019**, *67*, 48–48, DOI: https://doi.org/10.1002/nadc.20194086786.

(22) Els-Heindl, S.; Bellmann-Sickert, K. *Nachrichten aus der Chemie* **2019**, *67*, 48–52, DOI: https://doi.org/10.1002/nadc.20194089042.

(23) Lau, J. L.; Dunn, M. K. *Bioorg. Med. Chem.* **2018**, *26*, 2700–2707, DOI: 10.1016/j.bmc.2017.06.052.

(24) Loregian, A.; Palù, G. *J. Cell. Physiol.* **2005**, *204*, 750–762, DOI: https://doi.org/10.1002/jcp.20356.

(25) Hamley, I. W. *Chem. Rev.* **2017**, *117*, 14015–14041, DOI: 10.1021/acs.chemrev.7b00522.

(26) Kastin, A., *Handbook of Biologically Active Peptides*; Elsevier Science & Technology: San Diego, 2013, DOI: 10.1016/C2010-0-66490-X.

(27) Wang, L.; Wang, N.; Zhang, W.; Cheng, X.; Yan, Z.; Shao, G.; Wang, X.; Wang, R.; Fu, C. *Sig. Transduct. Target Ther.* **2022**, *7*, 48, DOI: 10.1038/s41392-022-00904-4.

(28) Hegele, A.; Kamburov, A.; Grossmann, A.; Sourlis, C.; Wowro, S.; Weimann, M.; Will, C. L.; Pena, V.; Lührmann, R.; Stelzl, U. *Mol. Cell* **2012**, *45*, 567–580, DOI: 10.1016/j.molcel.2011.12.034.

(29) Ball, L. J.; Kühne, R.; Schneider-Mergener, J.; Oschkinat, H. *Angew. Chem. Int. Ed.* **2005**, *44*, 2852–2869, DOI: 10.1002/anie.200400618.

(30) Cohen, G. B.; Ren, R.; Baltimore, D. *Cell* **1995**, *80*, 237–248, DOI: 10.1016/0092-8674(95)90406-9.

(31) Pawson, T.; Scott, J. D. *Science* **1997**, *278*, 2075–2080, DOI: 10.1126/science.278.5346.2075.

(32) Sudol, M.; Bork, P.; Einbond, A.; Kastury, K.; Druck, T.; Negrini, M.; Huebner, K.; Lehman, D. *J. Biol. Chem.* **1995**, *270*, 14733–14741, DOI: 10.1074/jbc.270.24.14733.

(33) Sudol, M. *Prog. Biophys. Mol. Biol.* **1996**, *65*, 113–132, DOI: 10.1016/s0079-6107(96)00008-9.

(34) Huang, X.; Poy, F.; Zhang, R.; Joachimiak, A.; Sudol, M.; Eck, M. J. *Nat. Struct. Mol. Biol.* **2000**, *7*, 634–638, DOI: 10.1038/77923.

(35) Kanelis, V.; Rotin, D.; Forman-Kay, J. D. *Nat. Struct. Mol. Biol.* **2001**, *8*, 407–412, DOI: 10.1038/87562.

(36) Macias, M. J.; Gervais, V.; Civera, C.; Oschkinat, H. *Nat. Struct. Mol. Biol.* **2000**, *7*, 375–379, DOI: 10.1038/75144.

(37) Bedford, M. T.; Sarbassova, D.; Xu, J.; Leder, P.; Yaffe, M. B. *J. Biol. Chem.* **2000**, *275*, 10359–10369.

(38) Otte, L.; Wiedemann, U.; Schlegel, B.; Pires, J. R.; Beyermann, M.; Schmieder, P.; Krause, G.; Volkmer-Engert, R.; Schneider-Mergener, J.; Oschkinat, H. *Protein Sci.* **2003**, *12*, 491–500, DOI: 10.1110/ps.0233203.

(39) Sudol, M.; Hunter, T. *Cell* **2000**, *103*, 1001–1004, DOI: 10.1016/s0092-8674(00)00203-8.

(40) Wiesner, S.; Stier, G.; Sattler, M.; Macias, M. J. *J. Mol. Biol.* **2002**, *324*, 807–822, DOI: 10.1016/s0022-2836(02)01145-2.

(41) Fedoroff, O. Y.; Townson, S. A.; Golovanov, A. P.; Baron, M.; Avis, J. M. *J. Biol. Chem.* **2004**, *279*, 34991–35000, DOI: 10.1074/jbc.m404987200.

(42) Huang, X.; Beullens, M.; Zhang, J.; Zhou, Y.; Nicolaescu, E.; Lesage, B.; Hu, Q.; Wu, J.; Bollen, M.; Shi, Y. *J. Biol. Chem.* **2009**, *284*, 25375–25387, DOI: 10.1074/jbc.m109.024828.

(43) Lin, Z.; Yang, Z.; Xie, R.; Ji, Z.; Guan, K.; Zhang, M. *eLife* **2019**, *8*, DOI: 10.7554/elife.49439.

(44) Ji, Z.; Li, H.; Yang, Z.; Huang, X.; Ke, X.; Ma, S.; Lin, Z.; Lu, Y.; Zhang, M. *Cell Rep.* **2019**, *26*, 2064–2077. e7, DOI: 10.1016/j.celrep.2019.01.097.

(45) Lin, Z.; Xie, R.; Guan, K.; Zhang, M. *Cell Rep.* **2020**, *32*, 108118, DOI: 10.1016/j.celrep.2020.108118.

(46) Sudol, M.; Recinos, C. C.; Abraczinskas, J.; Humbert, J.; Farooq, A. *IUBMB Life* **2005**, *57*, 773–778, DOI: 10.1080/15216540500389039.

(47) Klippel, S.; Wieczorek, M.; Schümann, M.; Krause, E.; Marg, B.; Seidel, T.; Meyer, T.; Knapp, E.-W.; Freund, C. *J. Biol. Chem.* **2011**, *286*, 38478–38487, DOI: 10.1074/jbc.m111.265710.

(48) Webb, C.; Upadhyay, A.; Giuntini, F.; Eggleston, I.; Furutani-Seiki, M.; Ishima, R.; Bagby, S. *Biochemistry* **2011**, *50*, 3300–3309, DOI: 10.1021/bi2001888.

(49) Schuchardt, B. J.; Mikles, D. C.; Hoang, L. M.; Bhat, V.; McDonald, C. B.; Sudol, M.; Farooq, A. *FEBS J.* **2014**, *281*, 5532–5551, DOI: 10.1111/febs.13095.

(50) Salah, Z.; Alian, A.; Aqeilan, R. I. *Front. Biosci.* **2012**, *17*, 331–348.

(51) Fischer, E. *Ber. Dtsch. Chemischen Ges.* **1894**, *27*, 2985–2993, DOI: https://doi.org/10.1002/cber.18940270364.

(52) Koshland Jr, D. *Proc. Natl. Acad. Sci.* **1958**, *44*, 98, DOI: 10.1073/pnas.44.2.98.

(53) Hammes, G. G.; Chang, Y.-C.; Oas, T. G. *Proc. Natl. Acad. Sci.* **2009**, *106*, 13737–13741, DOI: 10.1073/pnas.0907195106.

(54) Boehr, D. D.; Nussinov, R.; Wright, P. E. *Nat. Chem. Biol.* **2009**, *5*, 789–796, DOI: 10.1038/nchembio.232.

(55) Kumar, S.; Ma, B.; Tsai, C.-J.; Sinha, N.; Nussinov, R. *Protein Sci.* **2000**, *9*, 10–19, DOI: 10.1110/ps.9.1.10.

(56) Changeux, J.-P.; Edelstein, S. *F1000 Biol. Rep.* **2011**, *3*, 19–19, DOI: 10.3410/B3-19.

(57) Salmaso, V.; Moro, S. *Front. Pharmacol.* **2018**, *9*, DOI: 10.3389/fphar.2018.00923.

(58) Vetter, J. *Toxicon* **1998**, *36*, 13–24, DOI: 10.1016/s0041-0101(97)00074-3.

(59) Wieland, T.; Faulstich, H. *Experientia* **1991**, *47*, 1186–1193, DOI: 10.1007/bf01918382.

(60) Wieland, H.; Hallermayer, R. *Liebigs Ann. Chem.* **1941**, *548*, 1–18, DOI: 10.1002/jlac.19415480102.

(61) Hallen, H. E.; Luo, H.; Scott-Craig, J. S.; Walton, J. D. *Proc. Natl. Acad. Sci.* **2007**, *104*, 19097–19101, DOI: 10.1073/pnas.0707340104.

(62) Walton, J. D.; Hallen-Adams, H. E.; Luo, H. *Biopolymers* **2010**, *94*, 659–664, DOI: 10.1002/bip.21416.

(63) Brueckner, F.; Cramer, P. *Nat. Struct. Mol. Biol.* **2008**, *15*, 811–818, DOI: 10.1038/nsmb.1458.

(64) Chafin, D. R.; Guo, H.; Price, D. H. *J. Biol. Chem.* **1995**, *270*, 19114–19119, DOI: https://doi.org/10.1074/jbc.270.32.19114.

(65) Rudd, M. D.; Luse, D. S. *J. Biol. Chem.* **1996**, *271*, 21549–21558, DOI: 10.1074/jbc.271.35.21549.

(66) Pahl, A.; Lutz, C.; Hechler, T. *Drug Discov. Today: Technologies* **2018**, *30*, 85–89, DOI: 10.1016/j.ddtec.2018.08.005.

(67) Anderl, J.; Faulstich, H.; Hechler, T.; Kulke, M. In *Antibody-Drug Conjugates*, Ducry, L., Ed.; Humana Press: Totowa, NJ, 2013, pp 51–70, DOI: 10.1007/978-1-62703-541-5_4.

(68) Matinkhoo, K.; Pryyma, A.; Todorovic, M.; Patrick, B. O.; Perrin, D. M. *J. Am. Chem. Soc.* **2018**, *140*, 6513–6517, DOI: 10.1021/jacs.7b12698.

(69) Siegert, M.-A. J.; Knittel, C. H.; Süssmuth, R. *Angew. Chem. Int. Ed.* **2019**, DOI: 10.1002/anie.201914620.

(70) Lutz, C.; Simon, W.; Werner-Simon, S.; Pahl, A.; Müller, C. *Angew. Chem. Int. Ed.* **2020**, *59*, 11390–11393, DOI: 10.1002/anie.201914935.

(71) Yao, G.; Knittel, C. H.; Kosol, S.; Wenz, M. T.; Keller, B. G.; Gruß, H.; Braun, A. C.; Lutz, C.; Hechler, T.; Pahl, A. *J. Am. Chem. Soc.* **2021**, *143*, 14322–14331, DOI: 10.1021/jacs.1c06565.

(72) Wieland, T.; Götzendörfer, C.; Vaisius, A. C.; Zanotti, G. *Eur. J. Biochem.* **1981**, *117*, 161–164, DOI: 10.1111/j.1432-1033.1981.tb06315.x.

(73) Zanotti, G.; Wieland, T.; Benedetti, E.; Di Blasio, B.; Pavone, V.; Pedone, C. *Int. J. Pept. Protein Res.* **1989**, *34*, 222–228, DOI: 10.1111/j.1399-3011.1989.tb00234.x.

(74) Zanotti, G.; Petersen, G.; Wieland, T. *Int. J. Pept. Protein Res.* **1992**, *40*, 551–558, DOI: 10.1111/j.1399-3011.1992.tb00440.x.

(75) May, J. P.; Fournier, P.; Patrick, B. O.; Perrin, D. M. *Chem. Eur. J.* **2008**, *14*, 3410–3417, DOI: 10.1002/chem.200701297.

(76) Matinkhoo, K.; Wong, A. A. W. L.; Hambira, C. M.; Kato, B.; Wei, C.; Müller, C.; Hechler, T.; Braun, A.; Gallo, F.; Pahl, A.; Perrin, D. M. *Chem. Eur. J.* **2021**, *27*, 10282–10292, DOI: 10.1002/chem.202101373.

(77) Shoham, G.; Rees, D. C.; Lipscomb, W. N.; Zanotti, G.; Wieland, T. *J. Am. Chem. Soc.* **1984**, *106*, 4606–4615, DOI: 10.1021/ja00328a051.

(78) Le Daré, B.; Ferron, P.-J.; Gicquel, T. *Toxins* **2021**, *13*, 417, DOI: 10.3390/toxins13060417.

(79) Enjalbert, F.; Rapior, S.; Nouguier-Soulé, J.; Guillon, S.; Amouroux, N.; Cabot, C. *J. Toxicol., Clin. Toxicol.* **2002**, *40*, 715–757, DOI: 10.1081/clt-120014646.

(80) Meldolesi, J.; Pelosi, G.; Brunelli, A.; Genovese, E. *Virchows Arch. path Anat.* **1967**, *342*, 221–235, DOI: 10.1007/BF00960591.

(81) Kröncke, K.; Fricker, G.; Meier, P.; Gerok, W.; Wieland, T.; Kurz, G. *J. Biol. Chem.* **1986**, *261*, 12562–12567, DOI: 10.1016/s0021-9258(18)67125-x.

(82) Gundala, S.; Wells, L. D.; Milliano, M. T.; Talkad, V.; Luxon, B. A.; Neuschwander-Tetri, B. A. *Arch. Toxicol.* **2004**, *78*, 68–73, DOI: 10.1007/s00204-003-0527-y.

(83) Letschert, K.; Faulstich, H.; Keller, D.; Keppler, D. *Toxicol. Sci.* **2006**, *91*, 140–149, DOI: 10.1093/toxsci/kfj141.

(84) Bushnell, D. A.; Cramer, P.; Kornberg, R. D. *Proc. Natl. Acad. Sci.* **2002**, *99*, 1218–1222, DOI: 10.1073/pnas.251664698.

(85) Wieland, T. *Int. J. Pept. Protein Res.* **1983**, *22*, 257–276, DOI: 10.1111/j.1399-3011.1983.tb02093.x.

(86) Wang, M.; Chen, Y.; Guo, Z.; Yang, C.; Qi, J.; Fu, Y.; Chen, Z.; Chen, P.; Wang, Y. *Toxicon* **2018**, *156*, 34–40, DOI: 10.1016/j.toxicon.2018.11.002.

(87) Blanc, A.; Dietrich, D. J.; Perrin, D. M. *Peptide Science* **2019**, *111*, e24050, DOI: 10.1002/pep2.24050.

(88) Cramer, P.; Bushnell, D. A.; Kornberg, R. D. *Science* **2001**, *292*, 1863–1876, DOI: 10.1126/science.1059493.

(89) Liu, X.; Farnung, L.; Wigge, C.; Cramer, P. *J. Biol. Chem.* **2018**, *293*, 7189–7194, DOI: 10.1074/jbc.ra118.002545.

(90) Shoham, G.; Lipscomb, W. N.; Wieland, T. *J. Am. Chem. Soc.* **1989**, *111*, 4791–4809, DOI: 10.1021/ja00195a036.

(91) May, J. P.; Perrin, D. M. *Biopolymers* **2007**, *88*, 714–724, DOI: 10.1002/bip.20807.

(92) Wieland, T.; Faulstich, H.; Fiume, L. *Crit. Rev. Biochem. Mol. Biol.* **1978**, *5*, 185–260, DOI: 10.3109/10409237809149870.

(93) Buku, A.; Wieland, T.; Bodenmüller, H.; Faulstich, H. *Experientia* **1980**, *36*, 33–34, DOI: 10.1007/bf02003953.

(94) Zanotti, G.; Möhringer, C.; Wieland, T. *Int. J. Pept. Protein Res.* **1987**, *30*, 450–459, DOI: 10.1111/j.1399-3011.1987.tb03353.x.

(95) Todorovic, M.; Rivollier, P.; Wong, A. A.; Wang, Z.; Pryyma, A.; Nguyen, T. T.; Newell, K. C.; Froelich, J.; Perrin, D. M. *J. Med. Chem.* **2022**, DOI: 10.1021/acs.jmedchem.1c02226.

(96) May, J. P.; Perrin, D. M. *Chem. Eur. J.* **2008**, *14*, 3404–3409, DOI: https://doi.org/10.1002/chem.200701088.

(97) Lynen, F.; Wieland, U. *Liebigs Ann. Chem.* **1938**, *533*, 93–117, DOI: 10.1002/jlac.19385330105.

(98) Wieland, T.; Govindan, V. *FEBS Lett.* **1974**, *46*, 351–353.

(99) Faulstich, H.; Buku, A.; Bodenmueller, H.; Wieland, T. *Biochemistry* **1980**, *19*, 3334–3343.

(100) Tebbett, I.; Caddy, B. *Experientia* **1984**, *40*, 441–446, DOI: 10.1007/bf01952379.

(101) Swain, J. A.; Walker, S. R.; Calvert, M. B.; Brimble, M. A. *Nat. Prod. Rep.* **2022**, *39*, 410–443, DOI: 10.1039/d1np00043h.

(102) Field, M. J. *Arch. Biochem. Biophys.* **2015**, *582*, 3–9, DOI: 10.1016/j.abb.2015.03.005.

(103) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. *J. Med. Chem.* **2016**, *59*, 4035–4061, DOI: 10.1021/acs.jmedchem.5b01684.

(104) Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. *Annu. Rev. Biophys* **2012**, *41*, 429–452, DOI: 10.1146/annurev-biophys-042910-155245.

(105) Karplus, M.; McCammon, J. A. *Nat. Struct. Mol. Biol.* **2002**, *9*, 646–652, DOI: 10.1038/nsb0902-646.

(106) Lee, E. H.; Hsin, J.; Sotomayor, M.; Comellas, G.; Schulten, K. *Structure* **2009**, *17*, 1295–1306, DOI: 10.1016/j.str.2009.09.001.

(107) Schulten, K. *Biophys. J.* **2015**, *108*, 209a, DOI: 10.1016/j.bpj.2014.11.1158.

(108) Tajkhorshid, E.; Chipot, C. *J. Phys. Chem. B* **2017**, *121*, 3203–3205, DOI: 10.1021/acs.jpcb.7b02140.

(109) Schmitt, W.; Zanotti, G.; Wieland, T.; Kessler, H. *J. Am. Chem. Soc.* **1996**, *118*, 4380–4387, DOI: 10.1021/ja9529706.

(110) Fosgerau, K.; Hoffmann, T. *Drug Discov. Today* **2015**, *20*, 122–128, DOI: 10.1016/j.drudis.2014.10.003.

(111) Hilimire, T. A.; Bennett, R. P.; Stewart, R. A.; Garcia-Miranda, P.; Blume, A.; Becker, J.; Sherer, N.; Helms, E. D.; Butcher, S. E.; Smith, H. C.; Miller, B. L. *ACS Chem. Biol.* **2016**, *11*, 88–94, DOI: 10.1021/acschembio.5b00682.

(112) Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I.; Prasad, S.; Zuckerman, D. M. *Living J. Comput. Mol. Sci* **2019**, *1*, 1–28, DOI: 10.33011/livecoms.1.1.5957.

(113) Leach, A. R., *Molecular Modelling: Principles and Applications*, 2nd ed.; Pearson Education Limited: Essex, UK, 2001.

(114) Van Gunsteren, W. F. *Computer simulation of biomolecular system: theoretical and experimental applications* **1993**, *2*, 3–36.

(115) Born, M.; Oppenheimer, J. R. *Ann. Phys.* **1927**, *389*, 457–484.

(116) Ponder, J. W.; Case, D. A. In *Adv. Protein Chem.* Academic Press: 2003, pp 27–85, DOI: 10.1016/S0065-3233(03)66002-X.

(117) Ehrenfest, P. *Zeitschrift für Phys.* **1927**, *45*, 455–457, DOI: 10.1007/BF01329203.

(118) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins* **2010**, *78*, 1950–1958, DOI: 10.1002/prot.22711.

(119) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197, DOI: 10.1021/ja00124a002.

(120) Páll, S.; Hess, B. *Comput. Phys. Commun.* **2013**, *184*, 2641–2650, DOI: 10.1016/j.cpc.2013.06.003.

(121) Van Gunsteren, W. F.; Berendsen, H. J. *Mol. Simul.* **1988**, *1*, 173–185, DOI: 10.1080/08927028808080941.

(122) Hess, B.; Bekker, H.; Berendsen, H. J.; Fraaije, J. G. *J. Comput. Chem.* **1997**, *18*, 1463–1472, DOI: 10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h.

(123) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101, DOI: 10.1063/1.2408420.

(124) Parrinello, M.; Rahman, A. *J. Appl. Phys.* **1981**, *52*, 7182–7190, DOI: 10.1063/1.328693.

(125) Van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberger, P. H. *Angew. Chem. Int. Ed.* **2006**, *45*, 4064–4092, DOI: 10.1002/anie.200502655.

(126) Abraham, M. J.; van der Spoel, D.; Lindahl, E.; Hess, B.; team, G. d. GROMACS 2019 Manual, 2019, DOI: 10.5281/zenodo.2424486.

(127) Sagui, C.; Darden, T. A. *Annu. Rev. Biophys. Biomol. Struct.* **1999**, *28*, 155–179.

(128) Cisneros, G. A.; Karttunen, M.; Ren, P.; Sagui, C. *Chem. Rev.* **2014**, *114*, 779–814, DOI: 10.1021/cr300461d.

(129) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089–10092, DOI: 10.1063/1.464397.

(130) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593, DOI: 10.1063/1.470117.

(131) Frenkel, D.; Smit, B., *Understanding molecular simulation: from algorithms to applications*; Elsevier: 2001; Vol. 1.

(132) Van Gunsteren, W. F.; Berendsen, H. J. *Angew. Chem. Int. Ed.* **1990**, *29*, 992–1023, DOI: 10.1002/anie.199009921.

(133) Roe, D. R.; Brooks, B. R. *J. Chem. Phys.* **2020**, *153*, 054123, DOI: 10.1063/5.0013849.

(134) Sousa da Silva, A. W.; Vranken, W. F. *BMC Res Notes* **2012**, *5*, 367–367, DOI: 10.1186/1756-0500-5-367.

(135) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins* **2006**, *65*, 712–725, DOI: 10.1002/prot.21123.

(136) Wang, J.; Wang, W.; Kollman, P. A.; Case, D. A. *J. Mol. Graphics Modell.* **2006**, *25*, 247–260, DOI: 10.1016/j.jmgm.2005.12.005.

(137) Berman, H. M. *Nucleic Acids Res.* **2000**, *28*, 235–242, DOI: 10.1093/nar/28.1.235.

(138) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S. *Acta Crystallogr. Sect. D. Biol. Crystallogr.* **2002**, *58*, 899–907, DOI: 10.1107/s0907444902003451.

(139) Berman, H.; Henrick, K.; Nakamura, H. *Nat. Struct. Mol. Biol.* **2003**, *10*, 980–980, DOI: 10.1038/nsb1203-980.

(140) Karplus, M. *Fold Des.* **1997**, *2*, S69–S75, DOI: https://doi.org/10.1016/S1359-0278(97)00067-9.

(141) Levinthal, C. *Mossbauer spectroscopy in biological systems* **1969**, *67*, 22–24.

(142) Bryngelson, J. D.; Wolynes, P. G. *J. Phys. Chem.* **1989**, *93*.

(143) Robert, S. *Science* **2020**, DOI: 10.1126/science.abf9367.

(144) Jumper, J. et al. *Nature* **2021**, *596*, 583–589, DOI: 10.1038/s41586-021-03819-2.

(145) Keller, B.; Hunenberger, P.; van Gunsteren, W. F. *J. Chem. Theory Comput.* **2011**, *7*, 1032–1044, DOI: 10.1021/ct200069c.

(146) Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. *J. Chem. Phys.* **2013**, *139*, 07B604_1, DOI: 10.1063/1.4811489.

(147) Schwantes, C. R.; Pande, V. S. *J. Chem. Theory Comput.* **2013**, *9*, 2000–2009, DOI: 10.1021/ct300878a.

(148) Wu, C.; Lei, H.; Duan, Y. *Biophys. J.* **2004**, *87*, 3000–3009, DOI: 10.1529/biophysj.104.047076.

(149) Tong, L.; Soon, V. C.; Huang, Y. F.; Liu, R. In *IEEE International Symposium on Circuits and Systems*, 1784–1787 vol.3, DOI: 10.1109/ISCAS.1990.111981.

(150) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542, DOI: 10.1021/acs.jctc.5b00743.

(151) Hotelling, H. *J. Edu. Psych.* **1933**, *24*, 417, DOI: 10.1037/h0071325.

(152) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52, DOI: 10.1016/0169-7439(87)80084-9.

(153) Swope, W. C.; Pitera, J. W.; Suits, F. *J. Phys. Chem. B* **2004**, *108*, 6571–6581, DOI: 10.1021/jp037421y.

(154) Bowman, G. R.; Noé, F.; Pande, V. S., *An introduction to Markov state models and their application to long timescale molecular simulation*; Advances in experimental medicine and biology 797; Springer: Dordrecht, 2014.

(155) G. Keller, B.; Aleksić, S.; Donati, L. In *Biomolecular Simulations in Structure-Based Drug Discovery*, 2018, pp 67–86, DOI: https://doi.org/10.1002/9783527806836.ch4.

(156) Plattner, N.; Noé, F. *Nat. Commun.* **2015**, *6*, 7653, DOI: 10.1038/ncomms8653.

(157) Bowman, G. R.; Bolin, E. R.; Hart, K. M.; Maguire, B. C.; Marqusee, S. *Proc. Natl. Acad. Sci.* **2015**, *112*, 2734–2739, DOI: doi:10.1073/pnas.1417811112.

(158) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. *J. Chem. Phys.* **2015**, *143*, 11B601_1, DOI: 10.1063/1.4934536.

(159) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. *J. Chem. Phys.* **2011**, *134*, 174105, DOI: 10.1063/1.3565032.

(160) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. *J. Comput. Chem.* **2005**, *26*, 1701–1718, DOI: 10.1002/jcc.20291.

(161) Lane, T. J.; Shukla, D.; Beauchamp, K. A.; Pande, V. S. *Curr. Opin. Struct. Biol.* **2013**, *23*, 58–65, DOI: 10.1016/j.sbi.2012.11.002.

(162) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. *Living J. Comp. Mol. Sci.* **2022**, *4*, 1583, DOI: 10.33011/livecoms.4.1.1583.

(163) MacKerell Jr, A. D. *J. Comput. Chem.* **2004**, *25*, 1584–1604, DOI: 10.1002/jcc.20082.

(164) Warshel, A.; Levitt, M. *J. Mol. Biol.* **1976**, *103*, 227–249, DOI: https://doi.org/10.1016/0022-2836(76)90311-9.

(165) Kessler, H. *Angew. Chem. Int. Ed.* **1982**, *21*, 512–523, DOI: 10.1002/anie.198205121.

(166) Cavanagh, J.; Fairbrother, W. J.; Palmer III, A. G.; Skelton, N. J., *Protein NMR Spectroscopy: Principles and Practice*, 2nd ed.; Elsevier Academic Press: Burlington, US, 2007.

(167) Balci, M., *Basic 1H-and 13C-NMR Spectroscopy*, 1st ed.; Elsevier: Amsterdam, NL, 2005.

(168) Jacobsen, N. E., *NMR Spectroscopy Explained: Simplified Theory, Applications and Examples for Organic Chemistry and Structural Biology*; John Wiley & Sons: New Jersey, US, 2007.

(169) Gauglitz, G.; Moore, D. S., *Handbook of Spectroscopy*; Wiley-Vch Weinheim: Weinheim, DE, 2014; Vol. 1-4.

(170) Karplus, M. *J. Chem. Phys.* **1959**, *30*, 11–15, DOI: 10.1063/1.1729860.

(171) Bystrov, V. F. *Prog. Nucl. Magn. Reson. Spectrosc.* **1976**, *10*, 41–82, DOI: 10.1016/0079-6565(76)80001-5.

(172) Case, D. A.; Scheurer, C.; Brüschweiler, R. *J. Am. Chem. Soc.* **2000**, *122*, 10390–10397, DOI: 10.1021/ja001798p.

(173) Habeck, M.; Rieping, W.; Nilges, M. *J. Magn. Reson.* **2005**, *177*, 160–165, DOI: 10.1016/j.jmr.2005.06.016.

(174) Lee, J. H.; Li, F.; Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2015**, *137*, 1432–1435, DOI: 10.1021/ja512593s.

(175) Wang, A. C.; Bax, A. *J. Am. Chem. Soc.* **1996**, *118*, 2483–2494, DOI: 10.1021/ja9535524.

(176) Tropp, J. *J. Chem. Phys.* **1980**, *72*, 6035–6043, DOI: 10.1063/1.439059.

(177) Zagrovic, B.; van Gunsteren, W. F. *Proteins* **2006**, *63*, 210–218, DOI: 10.1002/prot.20872.

(178) Estivill-Castro, V. *ACM SIGKDD explorations newsletter* **2002**, *4*, 65–75, DOI: 10.1145/568574.568575.

(179) Hartigan, J. A.; Wong, M. A. *J. Royal Stat. Soc.* **1979**, *28*, 100–108, DOI: 10.2307/2346830.

(180) Jarvis, R. A.; Patrick, E. A. *IEEE Trans. Comput.* **1973**, *C-22*, 1025–1034, DOI: 10.1109/T-C.1973.223640.

(181) Schubert, E.; Sander, J.; Ester, M.; Kriegel, H. P.; Xu, X. *ACM Trans. Database Syst.* **2017**, *42*, 1–21, DOI: 10.1145/3068335.

(182) Graham, R. L.; Hell, P. *IEEE Ann. Hist. Comput.* **1985**, *7*, 43–57, DOI: 10.1109/MAHC.1985.10011.

(183) Lloyd, S. *IEEE Trans. Inf. Theory* **1982**, *28*, 129–137, DOI: 10.1109/tit.1982.1056489.

(184) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892, DOI: 10.1109/tpami.2002.1017616.

(185) Lemke, O.; Keller, B. G. *Algorithms* **2018**, *11*, 19, DOI: 10.3390/a11020019.

(186) Wang, W.; Yang, J.; Muntz, R. In *Vldb*; Vol. 97, pp 186–195.

(187) Agrawal, R.; Gehrke, J.; Gunopulos, D.; Raghavan, P. *SIGMOD Rec.* **1998**, *27*, 94–105, DOI: 10.1145/276305.276314.

(188) Keller, B. G.; Daura, X.; van Gunsteren, W. F. *J. Chem. Phys.* **2010**, *132*, 02B610, DOI: 10.1063/1.3301140.

(189) Lemke, O.; Keller, B. G. *J. Chem. Phys.* **2016**, *145*, 164104, DOI: 10.1063/1.4965440.

(190) Kapp-Joswig, J.-O.; Keller, B. G. *bioRxiv* **2022**, 2022.11.28.518169, DOI: 10.1101/2022.11.28.518169.

(191) Joswig, J.-O.; Anders, J.; Zhang, H.; Rademacher, C.; Keller, B. G. *J. Biol. Chem.* **2021**, 100718, DOI: 10.1016/j.jbc.2021.100718.

(192) Wenz, M. T.; Bertazzon, M.; Sticht, J.; Aleksić, S.; Gjorgjevikj, D.; Freund, C.; Keller, B. G. *J. Chem. Inf. Model.* **2022**, DOI: 10.1021/acs.jcim.1c01426.

(193) Han, J.; Kamber, M.; Mining, D. *Morgan Kaufmann* **2006**, *340*, 94104–3205.

(194) Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. *Drug Discov. Today* **2018**, *23*, 1530–1537, DOI: 10.1016/j.drudis.2018.05.006.

(195) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. *Proteins* **2002**, *47*, 409–443, DOI: https://doi.org/10.1002/prot.10115.

(196) Lee, H.; Heo, L.; Lee, M. S.; Seok, C. *Nucleic Acids Res.* **2015**, *43*, W431–W435, DOI: 10.1093/nar/gkv495.

(197) Obarska-Kosinska, A.; Iacoangeli, A.; Lepore, R.; Tramontano, A. *Nucleic Acids Res.* **2016**, *44*, W522–W528, DOI: 10.1093/nar/gkw366.

(198) Yan, Y.; Tao, H.; He, J.; Huang, S.-Y. *Nat. Protoc.* **2020**, *15*, 1829–1852, DOI: 10.1038/s41596-020-0312-x.

(199) Raveh, B.; London, N.; Schueler-Furman, O. *Proteins* **2010**, *78*, 2029–2040, DOI: 10.1002/prot.22716.

(200) London, N.; Raveh, B.; Cohen, E.; Fathi, G.; Schueler-Furman, O. *Nucleic Acids Res.* **2011**, *39*, W249–W253, DOI: 10.1093/nar/gkr431.

(201) Donsky, E.; Wolfson, H. J. *Bioinformatics* **2011**, *27*, 2836–2842, DOI: 10.1093/bioinformatics/btr498.

(202) De Vries, S. J.; Van Dijk, M.; Bonvin, A. M. *Nat. Protoc.* **2010**, *5*, 883, DOI: 10.1038/nprot.2010.32.

(203) Van Zundert, G.; Rodrigues, J.; Trellet, M.; Schmitz, C.; Kastritis, P.; Karaca, E.; Melquiond, A.; van Dijk, M.; De Vries, S.; Bonvin, A. *J. Mol. Biol.* **2016**, *428*, 720–725, DOI: 10.1016/j.jmb.2015.09.014.

(204) Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. *Nat. Protoc.* **2017**, *12*, 255–278, DOI: 10.1038/nprot.2016.169.

(205) Venkatraman, V.; Yang, Y. D.; Sael, L.; Kihara, D. *BMC Bioinform.* **2009**, *10*, 1–21, DOI: 10.1186/1471-2105-10-407.

(206) Li, J.; Fu, A.; Zhang, L. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11*, 320–328, DOI: 10.1007/s12539-019-00327-w.

(207) Huang, S.-Y.; Zou, X. *Int. J. Mol. Sci.* **2010**, *11*, 3016–3034, DOI: 10.3390/ijms11083016.

(208) De Vries, S. J.; Van Dijk, A. D.; Krzeminski, M.; van Dijk, M.; Thureau, A.; Hsu, V.; Wassenaar, T.; Bonvin, A. M. *Proteins* **2007**, *69*, 726–733, DOI: 10.1002/prot.21723.

(209) Vangone, A.; Rodrigues, J.; Xue, L.; van Zundert, G.; Geng, C.; Kurkcuoglu, Z.; Nellen, M.; Narasimhan, S.; Karaca, E.; van Dijk, M. *Proteins* **2017**, *85*, 417–423, DOI: 10.1002/prot.25198.

(210) Dominguez, C.; Boelens, R.; Bonvin, A. M. *J. Am. Chem. Soc.* **2003**, *125*, 1731–1737, DOI: 10.1021/ja026939x.

(211) Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J.; Vajda, S.; Vakser, I.; Wodak, S. J. *Proteins* **2003**, *52*, 2–9, DOI: 10.1002/prot.10381.

(212) Lensink, M. F.; Nadzirin, N.; Velankar, S.; Wodak, S. J. *Proteins* **2020**, *88*, 916–938, DOI: 10.1002/prot.25870.

(213) Xu, X.; Yan, C.; Zou, X. *J. Comput. Chem.* **2018**, *39*, 2409–2413, DOI: 10.1002/jcc.25555.

(214) Touw, W. G.; Baakman, C.; Black, J.; Te Beek, T. A.; Krieger, E.; Joosten, R. P.; Vriend, G. *Nucleic Acids Res.* **2015**, *43*, D364–D368, DOI: 10.1093/nar/gku1028.

(215) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577–2637, DOI: 10.1002/bip.360221211.

(216) Yao, G.; Joswig, J.-O.; Keller, B. G.; Süssmuth, R. D. *Chem. Eur. J.* **2019**, DOI: 10.1002/chem.201901888.

(217) Schuresko, L. A.; Lokey, R. S. *Angew. Chem. Int. Ed.* **2007**, *46*, 3547–3549, DOI: 10.1002/anie.200700017.

(218) Yao, G.; Kosol, S.; Wenz, M. T.; Irran, E.; Keller, B. G.; Trapp, O.; Süssmuth, R. D. *Nat. Commun.* **2022**, *13*, 6488, DOI: 10.1038/s41467-022-34125-8.

(219) Tonelli, A. E.; Patel, D. J.; Wieland, T.; Faulstich, H. *Biopolymers* **1978**, *17*, 1973–1986, DOI: 10.1002/bip.1978.360170812.

(220) Böckmann, R. A.; Grubmüller, H. *Nat. Struct. Mol. Biol.* **2002**, *9*, 198–202, DOI: 10.1038/nsb760.

(221) Bedford, M. T.; Reed, R.; Leder, P. *Proc. Natl. Acad. Sci.* **1998**, *95*, 10602–10607, DOI: 10.1073/pnas.95.18.10602.

(222) Chong, P. A.; Lin, H.; Wrana, J. L.; Forman-Kay, J. D. *Proc. Natl. Acad. Sci.* **2010**, *107*, 18404–18409, DOI: 10.1073/pnas.1003023107.

(223) Zarrinpar, A.; Bhattacharyya, R. P.; Lim, W. A. *Sci. Stke* **2003**, *2003*, re8–re8, DOI: 10.1126/stke.2003.179.re8.